Statistics and Computing (2022) 32:48
https://doi.org/10.1007/s11222-022-10109-y

Complexity of zigzag sampling algorithm for strongly log-concave distributions

Jianfeng Lu[1] · **Lihan Wang**[2]

Received: 2 June 2021 / Accepted: 26 April 2022 / Published online: 3 June 2022

Abstract
We study the computational complexity of zigzag sampling algorithm for strongly log-concave distributions. The zigzag process has the advantage of not requiring time discretization for implementation, and that each proposed bouncing event requires only one evaluation of partial derivative of the potential, while its convergence rate is dimension independent. Using these properties, we prove that the zigzag sampling algorithm achieves ε error in chi-square divergence with a computational cost equivalent to $O\left(\kappa^2 d^{\frac{1}{2}} (\log \frac{1}{\varepsilon})^{\frac{3}{2}}\right)$ gradient evaluations in the regime $\kappa \ll \frac{d}{\log d}$ under a warm start assumption, where κ is the condition number and d is the dimension.

Keywords Monte Carlo sampling · Zigzag sampler · Log-concave distribution · Computational complexity

1 Introduction and Main Results

Monte Carlo sampling from a high-dimensional probability distribution is a fundamental problem with applications in various areas including Bayesian statistics, machine learning, and statistical physics. Many sampling algorithms, especially those for continuous state space like $\mathbb{R}^d$, are based on continuous time Markov processes. Examples of these processes include the overdamped Langevin dynamics, whose invariant measure is the target measure, the underdamped Langevin dynamics and Hamiltonian Monte Carlo (HMC) Duane et al. (1987), both augment the state space with a velocity variable v, and have the x-marginal distribution of the invariant measure as the target measure. For strongly log-concave distributions, all these processes converge to the equilibrium exponentially fast with rates independent of the dimension, making them suitable for sampling purposes. On the other hand, all of these processes require time discretizations for implementation, which not only induces further numerical errors but requires the time step to be small as well, requiring higher computational complexity if a small bias is desired. To remove such bias due to discretization, the conventional procedure is to introduce the Metropolis-Hastings acceptance-rejection step, but rejections indicate waste of computational resources.

A very different line of sampling algorithms have been recently developed in statistical physics and statistics literature Peters and de With (2012), which are based on piecewise deterministic Markov processes (PDMPs) Davis (1984). These processes are non-reversible, which may mix faster than reversible MCMC methods (Diaconis et al. 2000; Turitsyn et al. 2011). Examples of such samplers include the randomized Hamiltonian Monte Carlo Bou-Rabee and Sanz-Serna (2017), the zigzag process Bierkens et al. (2019), the bouncy particle sampler (Peters and de With 2012; Bouchard-Côté et al. 2018), and some others (Vanetti et al. 2017; Michel et al. 2014; Bierkens et al. 2020). The zigzag and bouncy particle samplers are appealing for big data applications, as they can be unbiased even if stochastic gradient is used (Bouchard-Côté et al. 2018; Bierkens et al. 2019). These algorithms, as they are still relatively new, have not yet been thoroughly analyzed. In particular, no non-asymptotic computational complexity bounds on these algorithms have been established yet, to the best of our knowledge. Our previous work Lu et al. (2020) gives explicit exponential convergence rates for the PDMPs with log-concave potentials, which opens the possi-

✉ Lihan Wang
lihanw@andrew.cmu.edu

Jianfeng Lu
jianfeng@math.duke.edu

[1] Department of Mathematics, Department of Physics, and Department of Chemistry, Duke University, Durham, NC 27708, USA

[2] Department of Mathematical Sciences, Carnegie Mellon University, 311 Hamerschlag Drive, Pittsburgh, PA 15213, USA

bility of deriving such complexity bounds for PDMPs, and provides the foundation of this work.

1.1 Algorithm and assumptions

Let x denote the state variable in $\mathbb{R}^d$ where d is the dimension. The target distribution we want to sample from is denoted by

$$\mathrm{d}\mu(x) = Z^{-1}\exp(-U(x))\,\mathrm{d}x,$$

where $U(x)$ is the potential and $Z = \int_{\mathbb{R}^d}\exp(-U(x))\,\mathrm{d}x$ is the normalizing constant. Although the zigzag process can also be applied to sample non log-concave distributions, we will restrict our analysis to strongly log-concave distributions, namely, we make the following assumption throughout:

Assumption 1 The potential function $U(x)$ satisfies

$$m\mathrm{Id} \le \nabla^2 U(x) \le L\mathrm{Id}, \tag{1}$$

for some $0 < m \le 1 \le L$. Moreover, $U(x)$ has a unique minimizer at $x = 0$, and $U(0) = 0$.

For any random variable X, we use $\rho(X)$ to denote its law. In this paper, we use chi-square divergence to measure the difference between two probability measures: for probability measures ρ_1, ρ_2 that $\rho_1 \ll \rho_2$, it is defined as

$$\chi^2(\rho_1 \,\|\, \rho_2) := \int_{\mathbb{R}^d}\Big(\frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2} - 1\Big)^2\,\mathrm{d}\rho_2.$$

The zigzag sampling algorithm is based on a piecewise deterministic Markov process, called zigzag process. Besides the variable x, we augment the state space by an auxiliary velocity variable taking value in $\mathbb{R}^d$. A trajectory of the zigzag process, denoted by (X_t, V_t), can be described as follows. Given some initial (X_0, V_0), the position X_t always evolves according to $\frac{\mathrm{d}}{\mathrm{d}t}X_t = V_t$, while the velocity V_t is piecewise constant which only changes when bouncing or refreshing events occur at some random time following Poisson clocks. Bouncing events on the j-th direction occur with rate $(V_t^{(j)}\partial_{x_j}U(X_t))_+$, and at such an event the velocity V_t changes by flipping its j-th component to $-V_t^{(j)}$. Refreshing events occur with rate λ for some fixed $\lambda > 0$, when the velocity V_t is completely redrawn from the standard normal $\mathcal{N}(0, \mathrm{Id})$.

It has been established (Andrieu et al. 2021; Bierkens et al. 2019; Lu et al. 2020) that under Assumption 1, $\rho(X_t, V_t)$ converges to the invariant measure of the zigzag process, which is a product measure of the target measure in x and the standard Gaussian in v:

$$\mathrm{d}\bar{\mu}(x, v) = \mathrm{d}\mu(x)\,\mathrm{d}\nu(v) \quad \text{where } \mathrm{d}\nu(v) = (2\pi)^{-\frac{d}{2}}\exp\Big(-\frac{|v|^2}{2}\Big)\,\mathrm{d}v.$$

Our analysis relies on the following more quantitative convergence result for zigzag process proved in Lu et al. (2020), which also specifies the optimal choice of refreshing rate λ.[1] We would like to comment here that the choice of $\lambda = \sqrt{L}$ is completely technical since it optimizes the theoretical convergence rate (up to a universal constant) of the zigzag process established in Lu et al. (2020). The zigzag process is ergodic even if $\lambda = 0$ and in practice the choice $\lambda = 0$ is common.

Proposition 1 *(Lu et al. 2020, Theorem 1) Under Assumption 1, there exists a universal constant K independent of all parameters, such that for any initial density $\bar{\mu}_0$, the zigzag process with friction parameter $\lambda = \sqrt{L}$ satisfies*

$$\chi^2(\rho(X_T, V_T) \,\|\, \bar{\mu}) \le K\exp\Big(-\frac{m}{K\sqrt{L}}T\Big)\chi^2(\bar{\mu}_0 \,\|\, \bar{\mu}). \tag{2}$$

The left-hand side of (2) controls desired divergence of $\rho(X)$ with respect to the target measure μ, as we have

$$\begin{aligned}
&\chi^2(\rho(X_T, V_T) \,\|\, \bar{\mu}) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d}\Big(\frac{\mathrm{d}\rho(X_T, V_T)}{\mathrm{d}\bar{\mu}}\Big)^2\,\mathrm{d}\bar{\mu}(x, v) - 1 \\
&= \int_{\mathbb{R}^d}\Big(\frac{\mathrm{d}\rho(X_T)}{\mathrm{d}\mu}\Big)^2\Big(\int_{\mathbb{R}^d}\Big(\frac{\mathrm{d}\rho(V_T \mid X_T)}{\mathrm{d}\nu(v)}\Big)^2\,\mathrm{d}\nu(v)\Big)\,\mathrm{d}\mu(x) - 1 \\
&= \int_{\mathbb{R}^d}\Big(\frac{\mathrm{d}\rho(X_T)}{\mathrm{d}\mu}\Big)^2\Big(1 + \chi^2\big(\rho(V_T \mid X_T) \,\|\, \nu\big)\Big)\,\mathrm{d}\mu(x) - 1 \\
&\ge \int_{\mathbb{R}^d}\Big(\frac{\mathrm{d}\rho(X_T)}{\mathrm{d}\mu}\Big)^2\,\mathrm{d}\mu(x) - 1 = \chi^2(\rho(X_T) \,\|\, \mu).
\end{aligned}$$

Moreover, we would take initial condition in the form of

$$(X_0, V_0) \sim \bar{\mu}_0(x, v) = \mu_0(x)\nu(v), \tag{3}$$

which implies that $\chi^2(\bar{\mu}_0 \,\|\, \bar{\mu}) = \chi^2(\mu_0 \,\|\, \mu)$. Therefore, we get

$$\chi^2(\rho(X_T) \,\|\, \mu) \le K\exp\Big(-\frac{m}{K\sqrt{L}}T\Big)\chi^2(\mu_0 \,\|\, \mu), \tag{4}$$

which suggests the total time T needed to achieve control of chi-square divergence.

Of course, in practice, we cannot simulate the zigzag process directly, as simulating the Poisson process associated with the bouncing event would require integrating $(V_t^{(j)}\partial_{x_j}U(X_t))_+$ along the trajectory. To turn the zigzag

[1] Lu et al. (2020) shows exponential convergence for the backward equation. By duality the exponential convergence of the backward equation in $L^2(\bar{\mu})$ is equivalent to the exponential convergence of the forward equation in χ^2 with the same rate.

process into an efficient and practical sampling algorithm, the Poisson process for the bouncing events are usually simulated using the Poisson thinning trick (see e.g., discussions in Bierkens et al. (2019)*Section 3). Under Assumption 1, we will use the following upper bound estimate for the rate:

$$\begin{aligned}(v_i\partial_{x_i}U(x+vt))_+ &\le |v_i\partial_{x_i}U(x+vt)|\\ &\le |v_i||\partial_{x_i}U(x+vt)|\\ &\le L|v_i|(|x|+t|v|).\end{aligned} \tag{5}$$

This upper bound has the advantage of not involving evaluations of U and its partial derivatives, which greatly reduces the computational cost, compared with using numerical quadrature for d Poisson clocks. The price to pay is the increased frequency of potential bouncing events, which scales like $O(\sqrt{d})$ since the pessimistic bound for the partial derivative $|\partial_{x_i}U(x)| \le |\nabla U(x)| \le L|x|$ typically sacrifices a factor of $O(\sqrt{d})$ in the first inequality.

Following the above discussions, the zigzag sampling algorithm is described in Algorithm 1, where Step 12 uses the upper bound estimate in (5), while Steps 19–23 correspond to the Poisson thinning step. Note that for each potential bouncing event, the algorithm requires one evaluation of $\partial_{x_i}U$ in Step 19. In practice, typically accessing the partial derivatives of U is the most time consuming step, therefore, in our complexity analysis, we focus on the number of access to partial derivatives.

We also need the following assumption for technical purposes, as will be discussed after stating the main results:

Assumption 2 The initial distribution $\mu_0(x)$ satisfies a *warm-start* condition:

$$\chi^2(\mu_0 \,\|\, \mu) \le \exp\Big(\frac{d}{8K\kappa\log d}\Big), \tag{6}$$

where $\kappa := L/m$ is the condition number, and K is the same universal constant as in (2). Furthermore, the initial distribution is concentrated in the sense of

$$\eta := \mathbb{P}_{\mu_0}\Big(|x| > \sqrt{\frac{2d}{m}}\Big) < \frac{1}{4}. \tag{7}$$

Remark 1 The concentration condition (7) can be easily satisfied. By Gaussian Annulus Theorem, if we pick $\mu_0 = \mathcal{N}(0, \frac{1}{m}\mathrm{Id})$, then $\mathbb{P}_{\mu_0}\big(|x| > \sqrt{\frac{2d}{m}}\big) \le 3e^{-cd}$ for some universal constant c. The failure probability gets smaller if we take $\mu_0 = \mathcal{N}(0, \frac{1}{L}\mathrm{Id})$ or $\mu_0 = \mu$. The warm start condition (6) is more stringent but can be achieved by first running Langevin Monte Carlo (LMC). We will discuss that after presenting our main result.

Algorithm 1 The zigzag sampling algorithm

Input: Terminal time T, initial distribution μ_0.

1: Draw $x \sim \mu_0$.
2: Set $t \leftarrow 0$.
3: Set refr $\leftarrow$ true.
4: **while** $t < T$ **do**
5: **if** refr **then**
6: Draw $v \sim \mathcal{N}(0, \mathrm{Id})$.
7: Draw $t_{\mathrm{refr}} \sim \mathrm{Exp}(\sqrt{L})$.
8: $t_{\mathrm{refr}} \leftarrow \min\{t_{\mathrm{refr}}, T-t\}$.
9: refr $\leftarrow$ false.
10: **end if**
11: **for** $i = 1, \cdots, d$ **do**
12: Draw τ_i such that $\mathbb{P}(\tau_i \ge s) = \exp\big(-sL|v_i||x| - \frac{s^2}{2}|v_i||v|\big)$.
13: **end for**
14: Pick $j = \arg\min_{i=1,\cdots,d} \tau_i$.
15: $\Lambda_j \leftarrow L|v_j|(|x| + \tau_j|v|)$.
16: $t \leftarrow t + \min\{\tau_j, t_{\mathrm{refr}}\}$.
17: $x \leftarrow x + v\min\{\tau_j, t_{\mathrm{refr}}\}$.
18: **if** $\tau_j < t_{\mathrm{refr}}$ **then**
19: $\lambda_j \leftarrow (v_j\partial_{x_j}U(x))_+$.
20: Draw $\alpha \sim \mathrm{Unif}(0,1)$;
21: **if** $\alpha < \frac{\lambda_j}{\Lambda_j}$ **then**
22: $v_j \leftarrow -v_j$.
23: **end if**
24: $t_{\mathrm{refr}} \leftarrow t_{\mathrm{refr}} - \tau_j$.
25: **else**
26: refr $\leftarrow$ true.
27: **end if**
28: **end while**
29: **return** x.

1.2 Main results

Theorem 1 *Under Assumption 1, for any prescribed accuracy $\varepsilon > 0$, Algorithm 1 outputs a random variable X such that*

$$\chi^2(\rho(X) \,\|\, \mu) \le \varepsilon, \tag{8}$$

for terminal time T chosen as

$$T = K\Big(\frac{\sqrt{L}}{m}\big(\log\frac{1}{\varepsilon} + \log\chi^2(\mu_0 \,\|\, \mu) + \log K\big)\Big), \tag{9}$$

where K is the universal constant in (2).

Moreover, if $\varepsilon \ge \exp\big(-\frac{d}{8K\kappa\log d}\big)$, then, under Assumption 2, with probability $1 - \frac{C}{\sqrt{LT}} - C\log^{-\frac{3}{2}} d - \eta$, Algorithm 1 returns an output with a computational cost of

$$O\Big(d^{\frac{3}{2}}\kappa^2\big(\log^{\frac{3}{2}}\frac{1}{\varepsilon} + \log^{\frac{3}{2}}\chi^2(\mu_0 \,\|\, \mu)\big)\Big)$$

evaluations of partial derivatives of U, where η is defined in (7) and C is a universal constant.

Remark 2 By repeated trials, the theorem implies that for any $\delta \in (0, \frac{1}{4})$, with probability $1-\delta$, Algorithm 1 returns the desired output with a computational cost of

$$O\Big(d^{\frac{3}{2}}\kappa^2\Big(\log^{\frac{3}{2}}\frac{1}{\varepsilon}+\log^{\frac{3}{2}}\chi^2(\mu_0\,\|\,\mu)\Big)\log\frac{1}{\delta}\Big|\log^{-1}\Big(\frac{1}{\sqrt{LT}}+\log^{-\frac{3}{2}}d+\eta\Big)\Big|\Big),$$

that is $\widetilde{O}(d^{\frac{3}{2}}\kappa^2)$ evaluations of partial derivatives of U, where $\widetilde{O}(\cdot)$ hides logarithmic factors.

With the common computational model that d evaluations of partial derivatives of U is equivalent to one evaluation of ∇U in complexity, the complexity of zigzag is equivalent to $\widetilde{O}(d^{\frac{1}{2}}\kappa^2)$ evaluations of ∇U.

Let us explain the choice of T in (9): For the zigzag sampling algorithm to reach the target ε accuracy according to (4), the terminal time T needs to be large enough. Meanwhile, the Assumption 2 guarantees that T is not too large, as otherwise we cannot effectively control the number of bouncing events either due to a very large V drawn from a velocity refreshing event or the trajectory reaching regions with large gradient. These motivate our previous Assumption 2 on the initial distribution μ_0, as well as the restriction on ε that it cannot be too small compared to d. We remark that the assumption on ε is not prohibitive as we are interested in high dimensional cases and the error threshold is exponentially small in d.

The warm start condition (6) can be achieved if we start with a Gaussian distribution in x and run Langevin Monte Carlo

$$X_{n+1}=X_n-h\nabla U(X_n)+\sqrt{2h}\,\xi_n \tag{10}$$

where h is the step size, and ξ_n are i.i.d. $\mathcal{N}(0,\mathrm{Id})$ random variables. This leads to the following corollary:

Corollary 1 *Let $d\gg 1$. Suppose the potential U satisfies Assumption 1 for some $\kappa\geq 1$ such that $\kappa^{\frac{9}{5}}\leq\frac{d^{\frac{4}{5}}}{C\log^3 d}$ for some computable (from Erdogdu et al. (2020)) universal constant C. Then, for any prescribed accuracy $\varepsilon>0$, if we initialize $X_0\sim\mathcal{N}(0,\frac{1}{2L}\mathrm{Id})$, the hybrid algorithm by first running LMC (10) for $N=d^{4/5}\kappa^{16/5}$ steps with step size $h=\frac{4}{5}d^{-4/5}\kappa^{-16/5}m^{-1}\log\frac{d}{\kappa}$ and then Algorithm 1 up to time $T=K\Big(\frac{\sqrt{L}}{m}\Big(\log\frac{1}{\varepsilon}+d^{\frac{1}{5}}\kappa^{\frac{4}{5}}\log^2\frac{d}{\kappa}+\log K\Big)\Big)$ outputs a random variable X such that*

$$\chi^2(\rho(X)\,\|\,\mu)\leq\varepsilon. \tag{11}$$

Moreover, if $\varepsilon\geq\exp\big(-\frac{d}{8K\kappa\log d}\big)$, then there exists some universal constant $c>0$ such that, with probability $1-\frac{C}{\sqrt{LT}}-C\log^{-\frac{3}{2}}d-C\exp(Cd^{\frac{1}{5}}\kappa^{\frac{4}{5}}\log^2\frac{d}{\kappa}-cd)$, Algorithm 1 returns an output with a computational cost of

$$O\Big(d^{\frac{1}{2}}\kappa^2\log^{\frac{3}{2}}\frac{1}{\varepsilon}+d^{\frac{4}{5}}\kappa^{\frac{16}{5}}\log^3\frac{d}{\kappa}\Big)$$

evaluations of partial derivatives of U.

Proof It is easy to verify that our choice of N,h satisfies $h\leq\frac{m}{4L^2}$ and $Nh^2\leq\frac{1}{196c\kappa^2L^2}$ (where c satisfies Erdogdu et al. (2020)*Lemma 14). Therefore, we may appeal Lemmas 2, 14, 25, 26 of Erdogdu et al. (2020), so that the random variable X_N produced in (10) satisfies

$$\begin{aligned}&\chi^2(\rho(X_N)\|\mu)\\&\leq\exp\Big(Cd\exp(-Nhm)+CNh^2\kappa^2L^2(d+\log N)\Big)\\&=\exp\Big(Cd^{\frac{1}{5}}\kappa^{\frac{4}{5}}\log^2\frac{d}{\kappa}\Big)\end{aligned}\tag{12}$$

for some universal constant C. This, combined with our assumption on κ, guarantees that (6) holds with $\rho(X_N)$ playing the role of μ_0. We can also check the validity of (7) by

$$\begin{aligned}&\mathbb{P}_{\rho(X_N)}\big(|x|>\sqrt{\frac{2d}{m}}\big)\\&\leq\Big(1+\chi^2(\rho(X_N)\|\mu)\Big)^{\frac{1}{2}}\Big(\mathbb{P}_\mu\big(|x|>\sqrt{\frac{2d}{m}}\big)\Big)^{\frac{1}{2}}\\&\leq C\exp\Big(Cd^{\frac{1}{5}}\kappa^{\frac{4}{5}}\log^2\frac{d}{\kappa}-cd\Big)\ll 1.\end{aligned}$$

Therefore we may apply Theorem 1 with $\mu_0=\rho(X_N)$, and derive that the total computational cost (in terms of number of evaluations of ∇U) equals to

$$\begin{aligned}&O\Big(N+d^{\frac{1}{2}}\kappa^2\big(\log^{\frac{3}{2}}\frac{1}{\varepsilon}\\&\quad+\log^{\frac{3}{2}}\chi^2(\rho(X_N)\|\mu)\big)\Big)\\&=O\Big(d^{\frac{4}{5}}\kappa^{\frac{16}{5}}\log^3\frac{d}{\kappa}+d^{\frac{1}{2}}\kappa^2\log^{\frac{3}{2}}\frac{1}{\varepsilon}\Big).\end{aligned}$$

□

Theorem 1 guarantees that the zigzag sampling algorithm (Algorithm 1) outputs a sample from a distribution with χ^2-divergence at most ε away from the target density for a computational complexity equivalent to $\widetilde{O}(d^{\frac{3}{2}}\kappa^2)$ partial derivative evaluations (i.e., amounts to $\widetilde{O}(d^{\frac{1}{2}}\kappa^2)$ gradient evaluations), in the regime $\max\{\kappa,\log\frac{1}{\varepsilon}\}\ll\frac{d}{\log d}$ with a warm-start condition. Corollary 1 establishes that the hybrid LMC-zigzag algorithm outputs a sample for a computational complexity $\widetilde{O}(d^{\frac{4}{5}}\kappa^{\frac{16}{5}})$ gradient evaluations. The initialization using LMC is added only for technical reasons as we currently do not have complexity guarantees otherwise with an explicit initial distribution, nor is it necessary for actual implementations. We would also like to comment that our goal is to obtain the best possible scaling in d, and the scaling in κ might be possibly improved by a more careful analysis.

Our analysis is based on the quantitative convergence rate of the zigzag process established in Lu et al. (2020), which

is $O(\frac{m}{\sqrt{L}})$ for m-convex and L-smooth potentials. The rest of our proof is based on estimating $\sup|X_t|$ along a single trajectory of the zigzag process and subsequently turn this into an estimate on the number of potential bouncing events, and hence number of partial derivative evaluations. Our analysis utilizes the two important and desirable features of the zigzag sampling process:

- The implementation of the zigzag process does not need time discretization, as the velocity in deterministic portion of the trajectory remains constant, which makes it possible to simulate the exact trajectories of the zigzag process while eliminating an important source of error. This is the reason that the complexity of the zigzag process only has logarithmic dependence on $\frac{1}{\varepsilon}$, without Metropolis acceptance/rejection.
- Moreover, for each potential bouncing event of zigzag, only one evaluation of a *partial derivative* of the potential is required, which is $O(d)$ cheaper than a full gradient evaluation in computational cost for usual model of computation.

We would also remark that we quantify the error of distribution in terms of χ^2-divergence, which provides stronger guarantee than total variation, KL divergence or 2-Wasserstein distance. While χ^2-divergence is relatively convenient for obtaining convergence rates of continuous processes based on Poincarè inequality (Cao et al. 2019; Lu et al. 2020), it does not seem easy to use for analyzing discretization error of SDEs. The work Vempala and Wibisono (2019) made assumptions of Poincarè inequality for the discrete invariant measure, which is difficult to verify. We are fortunate to avoid such problem for zigzag sampler, thanks to the fact that zigzag does not need time discretization. After the first version of this work appears online, Erdogdu et al. (2020) established convergence of LMC in χ^2- and Rényi divergence, using the exponential convergence of continuous time overdamped Langevin dynamics in Rényi divergence (Cao et al. 2019; Vempala and Wibisono 2019).

1.3 Previous works

Here we focus on results on non-asymptotic analysis of sampling algorithms, which has been a focused research area in recent years. Many sampling algorithms have been analyzed including algorithms based on overdamped Langevin dynamics (Dalalyan 2017; Durmus and Moulines 2019; Durmus et al. 2019; Vempala and Wibisono 2019; Li et al. 2019; Ding et al. 2021), underdamped Langevin dynamics (Cheng et al. 2018; Dalalyan and Riou-Durand 2020; Ma et al. 2021; Shen and Lee 2019; Ding et al. 2021; Monmarché 2021), Hamiltonian Monte Carlo (Mangoubi and Smith 2019; Lee et al. 2018; Chen et al. 2019; Mangoubi and Vishnoi 2018; Bou-Rabee et al. 2020), or high order Langevin dynamics Mou et al. (2021), among others. These methods involve discretization of ODEs or SDEs, which yields an error that scales polynomially with step size. Thus the complexity of these algorithms has polynomial dependence on ε^{-1}, where ε is the desired accuracy threshold.

Metropolized variants of sampling algorithms, including Metropolized HMC and Metropolis Adjusted Langevin Algorithm (MALA), have also been studied in (Dwivedi et al. 2018; Chen et al. 2020; Lee et al. 2020), the complexities of which have only logarithmic dependence on ε^{-1}, similar to the zigzag sampling process analyzed here. In Dwivedi et al. (2018) the complexity upper bound for MALA is established as $\widetilde{O}(\kappa d + \kappa^{\frac{3}{2}} d^{\frac{1}{2}})$ under warm start condition, and $\widetilde{O}(\kappa d^2 + \kappa^{\frac{3}{2}} d^{\frac{3}{2}})$ with a feasible start. In Chen et al. (2020) the complexity upper bound for MALA is improved to $\widetilde{O}(\kappa d + \kappa^{\frac{3}{2}} d^{\frac{1}{2}})$ with feasible start (where $\mu_0 = \mathcal{N}(0, \frac{1}{L}\mathrm{Id})$). The work Chen et al. (2020) also established bounds for Metropolized HMC, which is $\widetilde{O}(\kappa d^{\frac{11}{12}})$ with warm start (which is in fact more stringent than our Assumption 2) in the regime $\kappa = O(d^{\frac{2}{3}})$, and $\widetilde{O}(\kappa^{\frac{3}{4}} d + \kappa^{\frac{7}{4}} d^{\frac{1}{2}})$ with feasible start if the target potential function has a bounded Hessian. The complexity upper bound has been improved in Lee et al. (2020) to $\widetilde{O}(\kappa d)$ for both Metropolized HMC and MALA with a feasible start, based on a refined analysis using concentration of gradient norm. In comparison, our result for zigzag relies on a warm start (which is achievable by LMC), while the complexity upper bound has better dependence in d. The issue of feasible start will be further discussed in Sect. 3.

Regarding asymptotic analysis for the convergence of zigzag process, the ergodicity was first established in Bierkens et al. (2019). Exponential convergence of the zigzag process is established in (Fontbona et al. 2016; Bierkens and Roberts 2017) using a Lyapunov function argument. A central limit theorem of the zigzag process is established in Bierkens and Duncan (2017), and a large deviation principle is established for the empirical measure in Bierkens et al. (2021). The spectrum of the zigzag process has been studied in (Bierkens et al. 2019; Guillin and Nectoux 2020). A dimension independent exponential convergence rate for the zigzag process is established in Andrieu et al. (2021), using the hypocoercivity framework developed in Dolbeault et al. (2015). Finally, a more quantitative convergence estimate was established in Lu et al. (2020), for which our analysis of the sampling algorithm is based on.

2 Strategy of the proof

Since Algorithm 1 always simulates exact trajectories of the zigzag process, we see that (8) is guaranteed with the correct choice of T. Therefore we only need to estimate the computational complexity. The strategy of the proof is to first give an estimate on $\sup_{t\in[0,T]} U(X_t)$ (Lemma 1), which directly controls $\sup_{t\in[0,T]} |X_t|$. The upper bound on $|X_t|$ in turn provides us an estimate of upper bound on the number of partial derivative evaluations of U. The complexity upper bound we derive holds with high probability, while it does not always hold (for example, the number of proposed bouncing events from the Poisson clock might be atypically high), such events only occur with very small probability, which will be controlled in the proof.

Let $N+1$ be the total number of velocity refreshments (including the initial refreshment), therefore N is a Poisson random variable such that

$$\mathbb{P}(N=n) = \frac{(\sqrt{LT})^n}{n!} e^{-\sqrt{LT}}. \tag{13}$$

Let $0 = T_0 < T_1 < T_2 < \cdots < T_N \le T < T_{N+1}$ be the refresh times, and V_{T_k} be the velocity variable after refreshment at time T_k. For $k = 1, \cdots, N$, we use $t_k = T_k - T_{k-1}$ to denote the time duration between refreshments. For convenience, we will also denote $t_{N+1} = T - T_N$.

The first step of the proof is the following lemma which controls $\sup_{t\in[0,T]} U(X_t)$ condition on some high probability events. The proof will be deferred to the appendix.

Lemma 1 *Under Assumptions 1 and 2, suppose the following conditions hold:*

$$\frac{1}{2}\sqrt{LT} \le N \le \frac{3}{2}\sqrt{LT}; \tag{14a}$$

$$|V_{T_k} \cdot \nabla U(X_{T_k})| \le \left(\frac{d}{\sqrt{LT}}\right)^{1/2} |\nabla U(X_{T_k})|, \quad \forall k = 1, \cdots, N; \tag{14b}$$

$$|V_{T_k}| \le 2\sqrt{d}, \quad \forall k = 1, \cdots, N \tag{14c}$$

$$U(X_0) \le \kappa d; \tag{14d}$$

$$\sum_{k=1}^{N+1} t_k^2 \le \frac{4T}{\sqrt{L}}. \tag{14e}$$

Then there exists a universal constant C such that

$$\sup_{t\in[0,T]} U(X_t) \le C\sqrt{LT}d. \tag{15}$$

The next element in the proof is to control the failure event that (14) does not hold. The control of the first four events are relatively straightforward and will thus be directly carried out in the proof of theorem below; we state the probability for the event (14e) to hold as the following lemma, which will also be proved in the appendix.

Lemma 2 *There exists a universal constant C such that, if $\sqrt{LT} > C$, then with probability $1 - \frac{2}{\sqrt{LT}}$, condition (14e) holds.*

The final component of the proof is to turn the estimate for $\sup_{t\in[0,T]} U(X_t)$ to an upper bound for the number of proposed bouncing events.

Proof of Theorem 1 Let p_i be the probability that condition i in (14) of Lemma 1 fails. We start with condition (14a) of Lemma 1. For Poisson process with t_i as the arrival times, we may estimate the first failure probability (here and for the rest of the proofs C denotes a universal constant that may change from line to line)

$$p_a \le \exp(-\frac{1}{C}\sqrt{LT}) \le \frac{C}{\sqrt{LT}}. \tag{16}$$

We now check the conditions (14b) and (14c) of Lemma 1. By Gaussian Annulus Theorem, for each refreshment, we have

$$\mathbb{P}(|V_{T_k}| \ge 2\sqrt{d}) \le 3e^{-cd}, \tag{17}$$

where $c > 0$ is some universal constant. We also require V_{T_k} to satisfy $|V_{T_k} \cdot n(X_{T_k})| \le \left(\frac{d}{\sqrt{LT}}\right)^{1/2}$, where $n(X_{T_k}) = \frac{\nabla U(X_{T_k})}{|\nabla U(X_{T_k})|}$, which has failure probability

$$\begin{aligned}
&\mathbb{P}(|V \cdot n(X)| \ge \left(\frac{d}{\sqrt{LT}}\right)^{1/2}) \\
&= \frac{1}{\sqrt{2\pi}} \int_{\left(\frac{d}{\sqrt{LT}}\right)^{1/2}}^{\infty} \exp(-\frac{r^2}{2})\, \mathrm{d}r \\
&\le \frac{1}{\sqrt{2\pi}} \int_{\left(\frac{d}{\sqrt{LT}}\right)^{1/2}}^{\infty} \exp\Big(-\frac{r}{2}\big(\frac{d}{\sqrt{LT}}\big)^{1/2}\Big)\, \mathrm{d}r \\
&\le \sqrt{\frac{2}{\pi}} \Big(\frac{\sqrt{LT}}{d}\Big)^{1/2} \exp(-\frac{d}{2\sqrt{LT}}).
\end{aligned}$$

Since we have to draw V for N times, cumulatively this yields a failure probability

$$p_b + p_c \le C\left(e^{-cd} + \left(\frac{\sqrt{LT}}{d}\right)^{1/2} \exp\Big(-\frac{d}{2\sqrt{LT}}\Big)\right)\mathbb{E}N. \tag{18}$$

Recall the assumption $\varepsilon \ge \exp(-\frac{d}{8K\kappa \log d})$ as well as (6) (and that $\kappa K \log K \le \frac{d}{4\log d}$), which implies that $\sqrt{LT} \le \frac{d}{2\log d}$ for our choice of T as in (9). Together with condition (14a), we derive (neglecting the obviously smaller term e^{-cd})

$$p_b + p_c \le C\sqrt{LT}\Big(\frac{\sqrt{LT}}{d}\Big)^{1/2} \exp\Big(-\frac{d}{2\sqrt{LT}}\Big) \le C\log^{-\frac{3}{2}} d.$$

The failure probability for condition (14d) is straightforward to estimate. Using Assumption 1, we have

$$U(X_0) \le \frac{L}{2}|X_0|^2,$$

which indicates

$$p_d \le \eta = \mathbb{P}\left(|X_0| \ge \sqrt{\frac{2d}{m}}\right).$$

Finally, p_e is already estimated in Lemma 2, which yields $p_e \le \frac{2}{\sqrt{LT}}$. In summary, the total failure probability of (14) can be bounded as

$$p_a + p_b + p_c + p_d + p_e \le \frac{C}{\sqrt{LT}} + C\log^{-\frac{3}{2}} d + \eta. \quad (19)$$

We now assume that condition (14) holds. Thus, Lemma 1 together with Assumption 1 implies that

$$\sup_{t\in[0,T]} |X_t| \le \Big(\frac{2}{m}\sup_{t\in[0,T]} U(X_t)\Big)^{1/2} \le C\Big(\frac{\sqrt{L}}{m}Td\Big)^{1/2}. \quad (20)$$

After each refreshment or bouncing event, Algorithm 1 runs d independent Poisson clocks $\{\tau_i\}_{i=1,\cdots,d}$ defined in Step 12 where, noticing $\sum_i |V_i| \le \sqrt{d}|V| \le 2d$,

$$\begin{aligned}\mathbb{P}(\min \tau_i \ge t) &\ge \exp\Big(-tL|X|\sum_i |V_i| - \frac{t^2}{2}|V|\sum_i |V_i|\Big)\\ &\ge \exp\big(-Cd^{\frac{3}{2}}(L^{\frac{5}{4}}m^{-\frac{1}{2}}T^{\frac{1}{2}}t + t^2)\big). \quad (21)\end{aligned}$$

This motivates us to consider the following counting process $\tilde{N}_t$: suppose $\tilde{t}_1, \cdots$ are i.i.d. random variables with $\mathbb{P}(\tilde{t}_i \ge s) = \exp(-As - Bs^2)$ where $A = Cd^{\frac{3}{2}}L^{\frac{5}{4}}m^{-\frac{1}{2}}T^{\frac{1}{2}}$ and $B = Cd^{\frac{3}{2}}$, and let $\tilde{N}_t = \inf_n\{\sum_{i=1}^n \tilde{t}_i > t\}$. By construction, the probability of $N > 8AT$ under condition (14) is controlled by $\mathbb{P}(\tilde{N}_T > 8AT)$. Therefore, it suffices to estimate $\mathbb{P}(\tilde{N}_T > 8AT)$.

We compute the expectation of $\tilde{t}_1$ (here notice $A \gg B \gg 1$):

$$\begin{aligned}\mathbb{E}\tilde{t}_1 &= \int_0^\infty s(A + 2Bs)\exp(-As - Bs^2)\,\mathrm{d}s\\ &\ge \int_0^{\frac{A}{B}} s(A + 2Bs)\exp(-2As)\,\mathrm{d}s\\ &= \frac{1}{4A} + \frac{B}{2A^3} - \Big(\frac{3A}{2B} + \frac{5}{4A}\\ &\quad + \frac{B}{2A^3}\Big)e^{-\frac{2A^2}{B}} \ge \frac{1}{4A}. \quad (22)\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}\tilde{t}_1^2 &\le \int_0^\infty s^2(A + 2Bs)\exp(-As)\,\mathrm{d}s\\ &= \frac{2}{A^2} + \frac{12B}{A^4} \le \frac{33}{16A^2}.\end{aligned}$$

Therefore we may appeal to Kolmogorov's inequality (Durrett 2019, Theorem 2.5.2) (here S_n denotes $\sum_{i=1}^n \tilde{t}_i$):

$$\begin{aligned}\mathbb{P}(\tilde{N}_T > 8AT) &= \mathbb{P}(S_{8AT} < T)\\ &= \mathbb{P}(S_{8AT} - \mathbb{E}S_{8AT}\\ &< T - \mathbb{E}S_{8AT}) \overset{16}{\le} \mathbb{P}(S_{8AT} - \mathbb{E}S_{8AT} < -T)\\ &\le \frac{1}{T^2}\operatorname{Var} S_{8AT}\\ &= \frac{8A}{T}\operatorname{Var}\tilde{t}_1 \le \frac{16}{AT} \le \frac{C}{\sqrt{LT}}.\end{aligned}$$

To sum up, we have established that with high probability the number of partial derivative evaluations is bounded by

$$\begin{aligned}O(AT) &= O(d^{\frac{3}{2}}L^{\frac{5}{4}}m^{-\frac{1}{2}}T^{\frac{3}{2}})\\ &= O\Big(d^{\frac{3}{2}}\kappa^2\big(\log^{\frac{3}{2}}\frac{1}{\varepsilon} + \log^{\frac{3}{2}}\chi^2(\mu_0\,\|\,\mu)\big)\Big).\end{aligned}$$

□

3 Discussion

We establish non-asymptotic complexity bounds for the zigzag sampling algorithm. While we focus on zigzag sampler in this work, we expect that similar analysis for other PDMPs (Bouchard-Côté et al. 2018; Vanetti et al. 2017; Michel et al. 2014; Bierkens et al. 2020) can be carried out. We leave these for future research.

We admit that our warm-start requirement (6) may be stringent. We observe that (6) implicitly requires the condition number κ to be much smaller than d, as otherwise, if $\kappa \sim d$, (6) requires $\chi^2(\mu_0\,\|\,\mu) = O(1)$ which is unrealistic. Corollary 1 essentially requires $\kappa \ll d^{\frac{4}{9}}$ for the analysis to hold. This restriction on condition number is not completely unexpected since the zigzag sampler does perform poorly for highly anisotropic densities (see for example numerical results in Michel et al. (2014)).

A major issue of the warm-start assumption comes from our choice of χ^2 divergence, rather than total variation, 2-Wasserstein distance, or KL divergence as in previous works for non-asymptotic analysis of sampling algorithms. In particular, if we choose the initial condition

$$\mathrm{d}\mu_0(x) = (\frac{L}{2\pi})^{\frac{d}{2}}\exp(-\frac{L|x|^2}{2})\,\mathrm{d}x, \quad (23)$$

as in previous works, then for $U(x) = \frac{m|x|^2}{2}$, we have

$$\begin{aligned}\chi^2(\mu_0 \,\|\, \mu) &= Z(\frac{L}{2\pi})^d \int_{\mathbb{R}^d} \exp(-L|x|^2 + U(x))\,\mathrm{d}x - 1 \\ &= \kappa^{\frac{d}{2}}(\frac{L}{2\pi})^{\frac{d}{2}} \int_{\mathbb{R}^d} \exp\big(-(L - \frac{m}{2})|x|^2\big)\,\mathrm{d}x - 1 \\ &= \kappa^{\frac{d}{2}} \left(\frac{L}{2L-m}\right)^{\frac{d}{2}} - 1,\end{aligned}$$

which violates (6). On the other hand, for the same choice of μ_0, as long as U satisfies Assumption 1, one can estimate

$$\begin{aligned}&\mathrm{KL}(\mu_0 \,\|\, \mu) \\ &= (\frac{L}{2\pi})^{\frac{d}{2}} \int_{\mathbb{R}^d} \Big(\frac{d}{2}\log\frac{L}{2\pi} \\ &\quad + \log Z - \frac{L}{2}|x|^2 + U(x)\Big) \exp(-\frac{L}{2}|x|^2)\,\mathrm{d}x \\ &\le \frac{d}{2}\log\kappa.\end{aligned}$$

This means $\log \mathrm{KL}(\mu_0 \,\|\, \mu)$, and consequently the logarithm of total variation or 2-Wasserstein distances are much smaller than any algebraic power of d, making it suitable for initialization. We hope the following conjecture is true:

Conjecture 1 *Under Assumption 1, there exists a universal constant K independent of all parameters, such that for any initial density $\bar{\mu}_0$, the zigzag process with friction parameter $\lambda = \sqrt{L}$ satisfies*

$$\mathrm{KL}(\rho(X_T, V_T) \,\|\, \bar{\mu}) \le K \exp\big(-\frac{m}{K\sqrt{L}}T\big)\,\mathrm{KL}(\bar{\mu}_0 \,\|\, \bar{\mu}).$$

If this is indeed true, we can establish the convergence in KL divergence of the pure zigzag sampler using a feasible start, without using LMC for initialization.

Another interesting open question is whether one can find a tighter upper bound than Step 12 of Algorithm 1 in order to reduce the computational complexity, since it magnifies the proposed bouncing rates by $O(\sqrt{d})$. The following lemma, which might be of independent interest, provides a concentration bound for $|\partial_{x_i} U|$ so that we might be able to give up a small probability to obtain a much sharper bouncing rate control.

Lemma 3 *Let $U(x)$ satisfy Assumption 1, then for any $c > 0$,*

$$\mathbb{P}_\mu\Big(|\partial_{x_i} U| \ge 2\sqrt{L} + 2c\sqrt{L}\log d\Big) \le 3d^{-c}. \tag{24}$$

The proof of this lemma, deferred to the appendix, is inspired by Lee et al. (2020), which uses the following Brascamp-Lieb inequality Brascamp and Lieb (1976):

Lemma 4 *Let $U(x)$ satisfy Assumption 1, then for any $g \in H^1(\mu)$,*

$$\mathrm{Var}_\mu\, g \le \int_{\mathbb{R}^d} \nabla g (\nabla^2 U)^{-1} \nabla g\,\mathrm{d}\mu. \tag{25}$$

With Lemma 3, it might be possible to improve Algorithm 1 while surrendering a small probability by replacing Step 12 with $\mathbb{P}(\tau_i \ge s) = \exp(-cs\sqrt{L}|v_i|\log d)$ since $(v_i \partial_{x_i} U(x + vs))_+ \le c\sqrt{L}|v_i|\log d$ with high probability. This motivates the following conjecture:

Conjecture 2 *Under the Assumption 1, for any κ and $\log\frac{1}{\varepsilon}$ that are both smaller than some algebraic power of d, there exists an algorithm that gives a random variable X such that*

$$\chi^2(\rho(X) \,\|\, \mu) \le \varepsilon. \tag{26}$$

Moreover, with high probability, the algorithm requires $O\Big(d\kappa \log d\big(\log\frac{1}{\varepsilon} + \log\chi^2(\mu_0 \,\|\, \mu)\big)\Big)$ evaluations of partial derivatives of U.

Unfortunately there are several difficulties for proving the conjecture. One is that although $\partial_{x_i} U$ does not exceed $O(\log d)$ with high probability, we are unable to control the partial derivatives for a trajectory of the zigzag process. Another issue is that since some trajectories of the zigzag process may go to regions with partial derivatives exceeding $O(\log d)$, we do not always simulate the exact trajectories, which introduces bias in the sampling.

Acknowledgements This work is supported in part by National Science Foundation via grants CCF-1910571 and DMS-2012286. We would like to thank Murat Erdogdu for pointing us to their complexity analysis of Langevin Monte Carlo in chi-square divergence Erdogdu et al. (2020) to remove the warm start assumptions.

Appendix

Proof of lemma 1

Proof Let $\lambda(t) = V_t \cdot \nabla_x U(X_t)$. If no bouncing happens, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda(t) = V_t^\top \nabla_x^2 U(X_t) V_t \le L|V_t|^2.$$

In addition, $\lambda(t)$ decreases when bouncing happens, since there is some positive $V_t^{(i)}\partial_{x_i} U(X_t)$ being changed to $-V_t^{(i)}\partial_{x_i} U(X_t)$ while X_t and other $V_t^{(j)}$'s remain unchanged. Therefore, since $|V_t|$ does not change between refreshments,

we have for any $t \in (0, T_{k+1} - T_k)$, [2]

$$\lambda(T_k + t) \le \lambda(T_k) + tL|V_{T_k}|^2. \tag{27}$$

Notice for a convex function $U(x)$ that satisfy Assumption 1, we have by co-coercivity

$$|\nabla U(x)|^2 \le 2LU(x),$$

therefore for any $t \in [0, T_{k+1} - T_k)$, and any $\alpha > 0$,

$$\begin{aligned} U(X_{T_k+t}) &= U(X_{T_k}) + \int_0^t \lambda(T_k + \tau)\,\mathrm{d}\tau \\ &\le U(X_{T_k}) + t\lambda(T_k) + \frac{Lt^2}{2}|V_{T_k}|^2 \\ &\overset{(14b),(14c)}{\le} U(X_{T_k}) + t\Big(\frac{d}{\sqrt{L}T}\Big)^{1/2}|\nabla U(X_{T_k})| + 2Lt^2 d \\ &\le U(X_{T_k}) + t\Big(\frac{2d\sqrt{L}}{T}\Big)^{1/2}\sqrt{U(X_{T_k})} + 2Lt^2 d \\ &\le (1+\alpha)U(X_{T_k}) + d\sqrt{L}t^2\Big(\frac{1}{\sqrt{2}T\alpha} + 2\sqrt{L}\Big). \end{aligned} \tag{28}$$

In particular,

$$U(X_{T_{k+1}}) \le (1+\alpha)U(X_{T_k}) + d\sqrt{L}t_{k+1}^2\Big(\frac{1}{\sqrt{2}T\alpha} + 2\sqrt{L}\Big).$$

Choosing $\alpha = \frac{1}{\sqrt{L}T}$, we have

$$U(X_{T_{k+1}}) \le (1+\alpha)U(X_{T_k}) + CdLt_{k+1}^2.$$

Now we apply the above formula iteratively and derive

$$\begin{aligned} U(X_T) &\le (1+\alpha)^{N+1}U(X_0) + CLd\sum_{k=1}^{N+1}(1+\alpha)^{N-k+1}t_k^2 \\ &\le (1+\alpha)^{N+1}\Big(U(X_0) + CLd\sum_{k=1}^{N+1}t_k^2\Big) \\ &\overset{(14d),(14e)}{\le} C\sqrt{L}Td. \end{aligned}$$

Here we used $\alpha = \frac{1}{\sqrt{L}T} = O(\frac{1}{N})$ so $(1+\alpha)^{N+1} = O(1)$, which is true due to (14a), and that $\kappa \le \sqrt{L}T$, which is true with our choice of T in (9). □

[2] We remark here that $\lambda(t)$ is not well-defined at the bouncing times. Nevertheless, (27) still makes sense since $\lambda(t)$ decreases at the bouncing events, and since we only use (27) in the time integral sense, this will not cause any problem.

Proof of lemma 2

Proof Let $\Xi = \sum_{k=1}^{N+1} t_k^2$. By properties of the Poisson process Durrett (1999), if we condition on N, the distribution of $T_1, T_2, \cdots, T_N$ has the same joint distribution as that of N i.i.d. random variables uniformly distributed in $(0, T)$. This means

$$\mathbb{E}(\Xi \mid N) = \frac{N!}{T^N}\int_{t_1+\cdots+t_N<T}\Big(\sum_{k=1}^N t_k^2 + \big(T - \sum_{k=1}^N t_k\big)^2\Big)\,\mathrm{d}t_N\cdots\mathrm{d}t_1. \tag{29}$$

To calculate $\mathbb{E}(\Xi \mid N)$, let us define

$$I_1(N,T) = \int_{t_1+\cdots+t_N<T}\Big(\sum_{k=1}^N t_k^2 + \big(T - \sum_{k=1}^N t_k\big)^2\Big)\,\mathrm{d}t_N\cdots\mathrm{d}t_1$$

and compute $I_1(N,T)$ by induction in N. For $N = 0$, as the sum contains only one term, $I_1(0,T) = T^2$. An easy calculation shows that $I_1(1,T) = \frac{2}{3}T^3$. We will show in general

$$I_1(N,T) = \frac{2(N+1)}{(N+2)!}T^{N+2}. \tag{30}$$

Indeed, suppose (30) holds for $N-1$, we want to prove (30) for N, the starting point of which is the following observation:

$$\begin{aligned} I_1(N,T) &= \int_{t_1+\cdots+t_N<T} t_1^2\,\mathrm{d}t_N\cdots\mathrm{d}t_1 \\ &\quad + \int_0^T I_1(N-1, T-t_1)\,\mathrm{d}t_1. \end{aligned}$$

The first integral can be treated by integrating the variables one by one, from t_N to t_{N-1} and then t_{N-2}, etc.

$$\begin{aligned} &\int_{t_1+\cdots+t_N<T} t_1^2\,\mathrm{d}t_N\cdots\mathrm{d}t_1 \\ &= \int_{t_1+\cdots+t_{N-1}<T} t_1^2(T - t_1 - \cdots - t_{N-1})\,\mathrm{d}t_{N-1}\cdots\mathrm{d}t_1 \\ &= \frac{1}{2}\int_{t_1+\cdots+t_{N-2}<T} t_1^2(T - t_1 - \cdots - t_{N-2})^2\,\mathrm{d}t_{N-2}\cdots\mathrm{d}t_1 \\ &= \cdots \\ &= \frac{1}{(N-1)!}\int_0^T t_1^2(T-t_1)^{N-1}\,\mathrm{d}t_1 = \frac{2}{(N+2)!}T^{N+2}. \end{aligned} \tag{31}$$

By the induction assumption (30) for $N-1$ we have

$$\begin{aligned} &\int_0^T I_1(N-1, T-t_1)\,\mathrm{d}t_1 \\ &= \int_0^T \frac{2N}{(N+1)!}(T-t_1)^{N+1}\,\mathrm{d}t_1 = \frac{2N}{(N+2)!}T^{N+2}. \end{aligned}$$

Combining above with (31) we finish the proof for N. Therefore

$$\mathbb{E}(\Xi \mid N) = \frac{N!}{T^N}\frac{2(N+1)T^{N+2}}{(N+2)!} = \frac{2T^2}{N+2}.$$

The full expectation $\mathbb{E}\Xi$ follows as N is a Poisson random variable

$$\begin{aligned}\mathbb{E}\Xi &= \sum_{n=0}^{\infty}\mathbb{E}(\Xi \mid N=n)\mathbb{P}(N=n)\\ &= \sum_{n=0}^{\infty}\frac{2T^2}{n+2}\frac{(\sqrt{L}T)^n}{n!}e^{-\sqrt{L}T}\\ &= 2T^2e^{-\sqrt{L}T}\sum_{n=0}^{\infty}\Big(\frac{(\sqrt{L}T)^n}{(n+1)!}-\frac{(\sqrt{L}T)^n}{(n+2)!}\Big)\\ &= \frac{2T}{\sqrt{L}}-\frac{2}{L}+\frac{2e^{-\sqrt{L}T}}{L}\le\frac{2T}{\sqrt{L}}.\end{aligned}$$

To get the desired estimate, we apply Chebyshev's inequality using the second moment. By the same arguments leading towards (29), we have

$$\mathbb{E}(\Xi^2 \mid N) = \frac{N!}{T^N}\int_{t_1+\cdots+t_N<T}\Big(\sum_{k=1}^{N}t_k^2+(T-\sum_{k=1}^{N}t_k)^2\Big)^2.$$

Denote

$$I_2(N,T) = \int_{t_1+\cdots+t_N<T}\Big(\sum_{k=1}^{N}t_k^2+(T-\sum_{k=1}^{N}t_k)^2\Big)^2.$$

Using the same induction argument as the proof of (30), we can prove

$$I_2(N,T) = \frac{4(N+1)(N+6)}{(N+4)!}T^{N+4}.$$

This can be easily verified for $N=0,1$ and the induction follows form the calculation:

$$\begin{aligned}I_2(N,T) &= \int_{t_1+\cdots+t_N<T}t_1^4\\ &\quad+\int_0^T I_2(N-1,T-t_1)\,\mathrm{d}t_1\\ &\quad+2\int_0^T t_1^2 I_1(N-1,T-t_1)\,\mathrm{d}t_1\\ &= \frac{1}{(N-1)!}\int_0^T t_1^4(T-t_1)^{N-1}\,\mathrm{d}t_1\\ &\quad+\frac{4N(N+5)}{(N+3)!}\int_0^T(T-t_1)^{N+3}\,\mathrm{d}t_1\\ &\quad+\frac{4N}{(N+1)!}\int_0^T t_1^2(T-t_1)^{N+1}\,\mathrm{d}t_1\\ &= \frac{4(N+1)(N+6)}{(N+4)!}T^{N+4}.\end{aligned}$$

This shows $\mathbb{E}(\Xi^2 \mid N) = \frac{N!}{T^N}I_2(N,T) = \frac{4(N+6)}{(N+2)(N+3)(N+4)}T^4$, and therefore

$$\begin{aligned}\mathbb{E}\Xi^2 &= \sum_{n=0}^{\infty}\mathbb{E}(\Xi^2 \mid N=n)\mathbb{P}(N=n)\\ &= \sum_{n=0}^{\infty}\frac{4T^4(n+6)}{(n+2)(n+3)(n+4)}\frac{(\sqrt{L}T)^n}{n!}e^{-\sqrt{L}T}\\ &= 4T^4e^{-\sqrt{L}T}\sum_{n=0}^{\infty}\Big(\frac{(\sqrt{L}T)^n}{(n+2)!}-6\frac{(\sqrt{L}T)^n}{(n+4)!}\Big)\\ &= \frac{4T^2}{L}-\frac{24}{L^2}+8e^{-\sqrt{L}T}\Big(\frac{T^2}{L}\\ &\quad+\frac{3T}{L^{\frac{3}{2}}}+\frac{3}{L^2}\Big).\end{aligned}$$

This means

$$\begin{aligned}\mathbb{E}(\Xi-\mathbb{E}\Xi)^2 &= \frac{8T}{L^{\frac{3}{2}}}-\frac{28}{L^2}+8e^{-\sqrt{L}T}\Big(\frac{T^2}{L}\\ &\quad+\frac{2T}{L^{\frac{3}{2}}}+\frac{4}{L^2}-\frac{4e^{-\sqrt{L}T}}{L^2}\Big)\le\frac{8T}{L^{\frac{3}{2}}},\end{aligned}$$

where the inequality above holds for $\sqrt{L}T$ larger than some universal constant (which we would assume as it is the interesting parameter regime).

Finally, to conclude the proof, we apply Chebyshev inequality to estimate the failure probability as

$$\begin{aligned}\mathbb{P}(\Xi\ge\frac{4T}{\sqrt{L}}) &\le \mathbb{P}(\Xi-\mathbb{E}\Xi\ge\frac{2T}{\sqrt{L}})\le\frac{L\mathbb{E}(\Xi-\mathbb{E}\Xi)^2}{4T^2}\\ &\le\frac{2}{\sqrt{L}T}.\end{aligned}$$

□

Proof of lemma 3

Proof The first step is to show that

$$\mathbb{E}_\mu|\partial_{x_i}U|\le\sqrt{L}. \tag{32}$$

This is straightforward, since using integration by parts,

$$\mathbb{E}_\mu|\partial_{x_i}U|^2 = \int_{\mathbb{R}^d}(\partial_{x_i}U)^2\,\mathrm{d}\mu = \int_{\mathbb{R}^d}\partial_{x_ix_i}U\,\mathrm{d}\mu\le L, \tag{33}$$

and (32) then follows from Cauchy-Schwarz inequality.

The next step is to establish a concentration bound. Let $G(x) = \psi(\partial_{x_i} U)$, where $\psi(a) = \psi(|a|)$ is a smooth non-negative increasing function satisfying

$$\psi(0) = \psi'(0) = 0,\ \psi(a) = |a| \ \text{ for } |a| \geq 1, \ \text{and } |\psi'(a)| \leq 2,$$

and $g(x) = \exp(\frac{1}{2}\lambda G(x))$. By the construction of G, we have

$$\mathbb{E}_\mu G = \mathbb{E}_\mu \psi(\partial_{x_i} U) \leq 2\mathbb{E}_\mu |\partial_{x_i} U| \leq 2\sqrt{L}. \tag{34}$$

Then $\nabla g(x) = \frac{\lambda}{2}\psi'(\partial_{x_i} U)\nabla(\partial_{x_i} U)g(x)$. By Lemma 4 for $g(x)$, we have

$$\begin{aligned}
&\mathbb{E}_\mu \exp(\lambda G) - \big(\mathbb{E}_\mu \exp(\frac{\lambda G}{2})\big)^2 = \mathrm{Var}_\mu\, g(x) \\
&\leq \frac{\lambda^2}{4}\int_{\mathbb{R}^d} (\psi'(\partial_{x_i} U))^2 \nabla(\partial_{x_i} U)(\nabla^2 U)^{-1}\nabla(\partial_{x_i} U) g^2(x)\,\mathrm{d}\mu \\
&\leq \lambda^2 \int_{\mathbb{R}^d} \nabla(\partial_{x_i} U)(\nabla^2 U)^{-1}\nabla(\partial_{x_i} U) g^2(x)\,\mathrm{d}\mu \\
&= \lambda^2 \int_{\mathbb{R}^d} \partial_{x_i x_i} U g^2(x)\,\mathrm{d}\mu \leq \lambda^2 L \mathbb{E}_\mu \exp(\lambda G).
\end{aligned}$$

Thus for $\lambda \leq \frac{1}{2\sqrt{L}}$ we have

$$\mathbb{E}_\mu \exp(\lambda G) \leq \frac{1}{1-\lambda^2 L}\big(\mathbb{E}_\mu \exp(\frac{\lambda G}{2})\big)^2. \tag{35}$$

Now we use (35) recursively, and we obtain for $H(\lambda) := \mathbb{E}_\mu \exp(\lambda G)$,

$$H(\lambda) \leq \prod_{k=0}^{\infty}\Big(\frac{1}{1-\frac{\lambda^2 L}{4^k}}\Big)^{2^k} \lim_{\ell\to\infty} H(\frac{\lambda}{\ell})^\ell. \tag{36}$$

Notice

$$\begin{aligned}
\lim_{\ell\to\infty} H(\frac{\lambda}{\ell})^\ell &= \lim_{\ell\to\infty}\Big(\mathbb{E}_\mu \exp(\frac{\lambda G}{\ell})\Big)^\ell \\
&= \lim_{\ell\to\infty}\Big(1 + \mathbb{E}_\mu \frac{\lambda G}{\ell}\Big)^\ell = \exp(\lambda \mathbb{E}_\mu G).
\end{aligned} \tag{37}$$

Moreover, by Bobkov and Ledoux (1997)*Proposition 4.1,

$$\begin{aligned}
&\prod_{k=0}^{\infty}\Big(\frac{1}{1-\frac{\lambda^2 L}{4^k}}\Big)^{2^k} \\
&\leq \frac{1+\lambda\sqrt{L}}{1-\lambda\sqrt{L}}.
\end{aligned} \tag{38}$$

Substituting (37) and (38) into (36), we obtain

$$H(\lambda) \leq \frac{1+\lambda\sqrt{L}}{1-\lambda\sqrt{L}} \exp(\lambda \mathbb{E}_\mu G).$$

Finally, combining the above exponential moment bound of G with Chebyshev inequality, we get

$$\mathbb{P}_\mu\Big(G(x) \geq \mathbb{E}_\mu G + r\Big) \leq \exp(-\lambda r)\frac{1+\lambda\sqrt{L}}{1-\lambda\sqrt{L}}.$$

Now take $\lambda = 1/2\sqrt{L}$, and $r = 2c\sqrt{L}\log d$, and using (34) (noticing $G(x) = |\partial_{x_i} U|$ when $G(x) \geq r$ since $r \geq 1$), we arrive at

$$\mathbb{P}_\mu\Big(|\partial_{x_i} U| \geq 2\sqrt{L} + 2c\sqrt{L}\log d\Big) \leq 3d^{-c}.$$

□

References

Andrieu, C., Durmus, A., Nüsken, N., Roussel, J.: Hypocoercivity of piecewise deterministic Markov process-Monte Carlo. Ann. Appl. Prob. **31**(5), 2478–2517 (2021)

Bierkens, J., Grazzi, S., Kamatani, K., Roberts, G.: The boomerang sampler, Int. Con. Mach. Learn. 908–918, (2020)

Bierkens, J., Lunel, S. M.V.: Spectral analysis of the zigzag process. arXiv preprint arXiv:1905.01691, (2019)

Bierkens, J., Duncan, A.: Limit theorems for the zig-zag process. Advances Appl. Prob. **49**(3), 791–825 (2017)

Bierkens, J., Roberts, G.: A piecewise deterministic scaling limit of lifted Metropolis-Hastings in the Curie-Weiss model. Ann. Appl. Prob. **27**(2), 846–882 (2017)

Bierkens, J., Fearnhead, P., Roberts, G.: The zig-zag process and super-efficient sampling for Bayesian analysis of big data. Ann. Stat. **47**(3), 1288–1320 (2019)

Bierkens, J., Roberts, G.O., Zitt, P.-A.: Ergodicity of the zigzag process. Ann. Appl. Prob. **29**(4), 2266–2301 (2019)

Bierkens, J., Nyquist, P., Schlottke, M.C.: Large deviations for the empirical measure of the zig-zag process. Ann. Appl. Prob. **31**(6), 2811–2843 (2021)

Bobkov, S., Ledoux, M.: Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential distribution. Prob. Theory Relate. Fields. **107**(3), 383–400 (1997)

Bouchard-Côté, A., Vollmer, S.J., Doucet, A.: The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. J. Am. Stat. Ass. **113**(522), 855–867 (2018)

Bou-Rabee, N., Sanz-Serna, J.M.: Randomized Hamiltonian Monte Carlo. Ann. Appl. Prob. **27**(4), 2159–2194 (2017)

Bou-Rabee, N., Eberle, A., Zimmer, R.: Coupling and convergence for Hamiltonian Monte Carlo. Ann. Appl. Prob. **30**(3), 1209–1250 (2020)

Brascamp, H.J., Lieb, E.H.: On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. J. Func. Anal. **22**(4), 366–389 (1976)

Cao, Y., Lu, J., Wang, L.: On explicit L^2-convergence rate estimate for underdamped Langevin dynamics, arXiv preprint arXiv:1908.04746, (2019)

Cao, Y., Lu, J., Lu, Y.: Exponential decay of rényi divergence under fokker-planck equations. J. Stat. Phys. **176**(5), 1172–1184 (2019)

Chen, Z., Vempala, S.S.: Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019), **145**,64 (2019)

Chen, Y., Dwivedi, R., Wainwright, M.J., Yu, B.: Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. J. Mach. Lear. Res. **21**(92), 1–72 (2020)

Cheng, X., Chatterji, N.S., Bartlett, P.L., Jordan, M.I.: Underdamped Langevin MCMC: A non-asymptotic analysis, PMLR, Conference on learning theory, pp. 300–323, (2018)

Dalalyan, A.S.: Theoretical guarantees for approximate sampling from smooth and log-concave densities. J. Royal Stat. Soc.: Series B (Stat. Meth.) **3**(79), 651–676 (2017)

Dalalyan, A.S., Riou-Durand, L.: On sampling from a log-concave density using kinetic Langevin diffusions. Bernoulli. **26**(3), 1956–1988 (2020)

Davis, M.H.A.: Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models. J. Royal Stat. Soc: Series B (Meth.) **46**(3), 353–376 (1984)

Diaconis, P., Holmes, S., Neal, R.M.: Analysis of a nonreversible Markov Chain sampler, Ann. Appl. Prob. 726–752 (2000)

Ding, Z., Li, Q., Lu, J., Wright, S.J.: Random coordinate Langevin Monte Carlo, PMLR, Conference on learning theory, 1683–1710 (2021)

Ding, Z., Li, Q., Lu, J., Wright, S.J.: Random coordinate underdamped Langevin Monte Carlo, PMLR, International conference on artificial intelligence and statistics, 2701–2709 (2021)

Dolbeault, J., Mouhot, C., Schmeiser, C.: Hypocoercivity for linear kinetic equations conserving mass. Trans. Am. Math. Soc. **367**(6), 3807–3828 (2015)

Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. letters B. **195**(2), 216–222 (1987)

Durmus, A., Moulines, E.: High-dimensional Bayesian inference via the unadjusted Langevin algorithm. Bernoulli **25**(4A), 2854–2882 (2019)

Durmus, A., Majewski, S., Miasojedow, B.: Analysis of Langevin Monte Carlo via convex optimization. J. Mach. Learn. Res. **20**, 73–1 (2019)

Durrett, R.: Essentials of stochastic processes, Springer, Vol. 1 (1999)

Durrett, R.: Probability: theory and examples, Cambridge university press, Vol. 49 (2019)

Dwivedi, R., Chen, Y., Wainwright, M.J., Yu, B.: Log-concave sampling: Metropolis-Hastings algorithms are fast!, PMLR, Conference on learning theory, 793–797 (2018)

Erdogdu, M.A., Hosseinzadeh, R., Zhang, M.S.: Convergence of Langevin Monte Carlo in chi-squared and Renyi divergence, arXiv preprint arXiv:2007.11612 (2020)

Fontbona, J., Guérin, H., Malrieu, F.: Long time behavior of telegraph processes under convex potentials. Stoch. Proc. their Appl. **126**(10), 3077–3101 (2016)

Guillin, A., Nectoux, B.: Low-lying eigenvalues and convergence to the equilibrium of some piecewise deterministic Markov processes generators in the small temperature regime, organization=Springer. Ann. Henri Poincaré. **21**(11), 3575–3608 (2020)

Lee, Y.T., Shen, R., Tian, K.: Logsmooth gradient concentration and tighter runtimes for metropolized Hamiltonian Monte Carlo, PMLR, Conference on learning theory, 2565–2597, (2020)

Lee, Y.T., Song, Z., Vempala, S.S.: Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities, arXiv preprint arXiv:1812.06243 (2018)

Li, X., Wu, Y., Mackey, L., Erdogdu, M.A.: Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond, Advances in neural information processing systems, 7748–7760 (2019)

Lu, Jianfeng, Wang, Lihan,: On explicit L^2-convergence rate estimate for piecewise deterministic Markov processes in MCMC algorithms, arXiv preprint arXiv:2007.14927 (2020)

Ma, Y.-A., Chatterji, N.S., Cheng, X., Flammarion, N., Bartlett, P.L., Jordan, M.I.: Is there an analog of Nesterov acceleration for gradient-based MCMC? Bernoulli. **27**(3), 1942–1992 (2021)

Mangoubi, O., Smith, A.: Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators, The 22nd international conference on artificial intelligence and statistics. 586–595 (2019)

Mangoubi, O., Vishnoi, N.: Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. Adva. Neural Info. Proc. Syst. **31**, 6027–6037 (2018)

Michel, M., Kapfer, S.C., Krauth, W.: Generalized event-chain Monte Carlo: Constructing rejection-free global-balance algorithms from infinitesimal steps. J. Chem. Phys. **140**(5), 054116 (2014)

Monmarché, P.: High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. Electronic J. Stat. **15**(2), 4117–4166 (2021)

Mou, W., Ma, Y.-A., Wainwright, M.J., Bartlett, P.L., Jordan, M.I.: High-order Langevin diffusion yields an accelerated MCMC algorithm. J. Mach. Learn. Res. **22**, 42–1 (2021)

Peters, E.A.J.F., de With, G.: Rejection-free Monte Carlo sampling for general potentials. Phys. Rev. E. **85**(2), 026703 (2012)

Shen, R., Lee, Y.T.: The randomized midpoint method for log-concave sampling, Adv. Neural Inf. Process. Sys. 2100–2111, (2019)

Turitsyn, K.S., Chertkov, M., Vucelja, M.: Irreversible Monte Carlo algorithms for efficient sampling. Phys. D: Nonlinear Phenom. **240**(4–5), 410–414 (2011)

Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., Doucet, A.: Piecewise-deterministic Markov chain Monte Carlo, arXiv preprint arXiv:1707.05296, (2017)

Vempala, S., Wibisono, A.: Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices, Advances in neural information processing systems, 8094–8106, (2019)