Journal of Machine Learning Research 25 (2024) 1-44 Submitted 10/22; Revised 2/24; Published 2/24

Mean-Square Analysis of Discretized Itô Diffusions for Heavy-tailed Sampling

Ye He LEOHE@UCDAVIS.EDU
Department of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332, USA

Tyler Farghly FARGHLY@STATS.OX.AC.UK
Department of Statistics
University of Oxford
Oxford, OX1 2JD, United Kingdom

Krishnakumar Balasubramanian KBALA@UCDAVIS.EDU
Department of Statistics
University of California
Davis, CA 95616, USA

Murat A. Erdogdu ERDOGDU@CS.TORONTO.EDU
Department of Computer Science & Department of Statistics
University of Toronto
Toronto, ON M5S 3G4, Canada

Editor: Alexandre Bouchard

Abstract

We analyze the complexity of sampling from a class of heavy-tailed distributions by discretizing a natural class of Itô diffusions associated with weighted Poincaré inequalities. Based on a mean-square analysis, we establish the iteration complexity for obtaining a sample whose distribution is ϵ close to the target distribution in the Wasserstein-2 metric. In this paper, our results take the mean-square analysis to its limits, i.e., we invariably only require that the target density has finite variance, the minimal requirement for a mean-square analysis. To obtain explicit estimates, we compute upper bounds on certain moments associated with heavy-tailed targets under various assumptions. We also provide similar iteration complexity results for the case where only function evaluations of the unnormalized target density are available by estimating the gradients using a Gaussian smoothing technique. We provide illustrative examples based on the multivariate t-distribution.

Keywords: Weighted Poincaré inequalities, Itô diffusion, Euler-Marayama discretization, multivariate t-distribution, Complexity of Sampling.

1. Introduction

The problem of sampling from a given target density $\pi : \mathbb{R}^d \to \mathbb{R}$ arises in a wide variety of problems in statistics, machine learning, operations research and applied mathematics. Markov chain Monte Carlo (MCMC) algorithms are a popular class of algorithms for sampling (Robert and Casella, 1999; Andrieu et al., 2003; Hairer et al., 2006; Brooks et al., 2011;

Meyn and Tweedie, 2012; Leimkuhler and Matthews, 2016; Douc et al., 2018); a widely used approach in this domain is to discretize an Itô diffusion that has the target as its stationary density. A popular choice of diffusion is the overdamped Langevin diffusion,

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dB_t, \tag{1}$$

where B_t is a d-dimensional Brownian motion. For example, the Unadjusted Langevin Algorithm[1] (Rossky et al., 1978), the Metropolis Adjusted Langevin Algorithm (Roberts and Tweedie, 1996; Roberts and Rosenthal, 1998) and the proximal sampler (Titsias and Papaspiliopoulos, 2018; Lee et al., 2021; Vono et al., 2022) arise as different discretizations of (1). Under light-tailed assumptions, i.e. when the density π has exponentially fast decaying tails, the diffusion X_t in (1) converges exponentially fast to π as its stationary density, which motivates the use of discretizations of (1) as practical algorithms for sampling. In the last decade, the non-asymptotic iteration complexity of various discretizations have been well-explored, thereby providing a relatively comprehensive story of sampling from light-tailed densities.

Motivated by applications in robust statistics (Kotz and Nadarajah, 2004; Jarner and Roberts, 2007; Kamatani, 2018), multiple comparison procedures (Genz et al., 2004; Genz and Bretz, 2009), Bayesian statistics (Gelman et al., 2008; Ghosh et al., 2018), and statistical machine learning (Balcan and Zhang, 2017; Nguyen et al., 2019; Şimşekli et al., 2020; Diakonikolas et al., 2020), in this work, we are interested in sampling from densities that have heavy-tails, for example, those with tails that are polynomially decaying. When the target density π is heavy-tailed, the solution to (1) does not converge exponentially to its stationary density in various metrics of interest. Indeed, Theorem 2.4 by Roberts and Tweedie (1996) shows that if $|\nabla \log \pi(x)| \to 0$ as $|x| \to \infty$, then the solution to (1) is *not* exponentially ergodic. In the other direction, standard results in the literature, for example Wang (2006); Bakry et al. (2014) show that the solution to (1) converging exponentially fast to its equilibrium density in the χ^2 metric, is equivalent to the density π satisfying the Poincaré inequality, which in turn requires π having exponentially decaying tails. Furthermore, when π has polynomially decaying tails, the convergence is only sub-exponential or polynomial (Wang, 2006, Chapter 4).

Note that the above results are predominantly for the Langevin diffusion. To verify this phenomenon for ULA, in Figure 1, we plot instantiations of sample paths of ULA with three different initializations for sampling from the standard multivariate t-distribution with 4 degrees of freedom. For comparison, we also plot the proposed Itô discretization (introduced later in (10)). The observed results show empirical evidence of a slow-start phenomenon associated with ULA for heavy-tailed targets. Recently, Mousavi-Hosseini et al. (2023) characterized the above phenomenon for ULA theoretically by obtaining upper and lower-bounds on the iteration complexity of the ULA when the target density satisfies weak-Poincaré inequalities. A canonical heavy-tailed density that satisfies the weak-Poincaré inequality is the heavy-tailed multivariate t-distribution. Specifically, Mousavi-Hosseini et al. (2023) showed that as the tail of the target density get heavier, the ULA exhibits an exponential dependence on the initial density used. In particular, for the multivariate t-distribution, the exponential dependence on the initializer is unavoidable unless there

1. We refer to it as ULA in this draft.

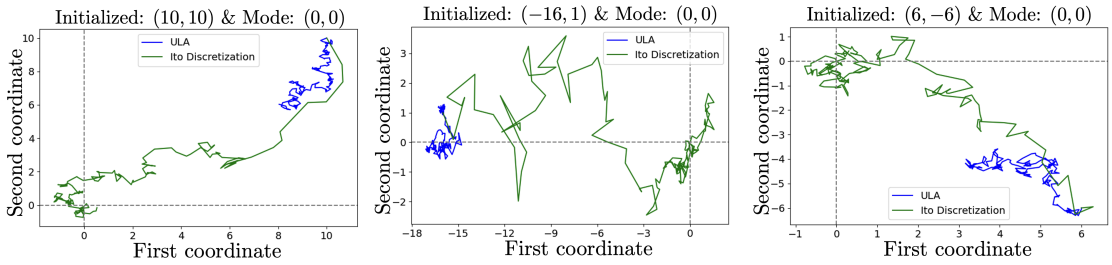

Figure 1: Instantiations of the sample paths of ULA and Itô discretization in (10) for sampling from two-dimensional standard t-distribution with 4 degrees of freedom. The *left*, *middle* and *right* panel corresponds to initialization being $(10, 10)$, $(-16, 1)$ and $(6, -6)$ respectively. For all the plots, both algorithms were run for 150 iterations with step-size set to 0.03. While we provide three illustrative instantiations here, we observed this general behavior for various other choices of (relatively smaller) step-sizes and initialization as well. We also refer to Erdogdu et al. (2018) for an empirical demonstration of divergence of ULA with relatively larger step-sizes. Additional experimental results are provided in Section 7.

is a good initialization. Hence, algorithms like the ULA obtained as discretizations of the Langevin diffusion in (1) are suited to sampling only from light-tailed exponentially decaying densities, and are rather inefficient for sampling from heavy-tailed densities.

Our approach to heavy-tailed sampling is hence based on discretizing certain natural Itô diffusions that arise in the context of the following Weighted Poincaré inequality (Blanchet et al., 2009; Bobkov and Ledoux, 2009). Such inequalities could be considered generalizations of the Brascamp-Lieb inequality (established for the class of log-concave densities) to a class of heavy-tailed densities. We emphasize here that typical analyses of Langevin diffusions and ULA assumes densities of the form e^{-V}, where V is a potential function. Below, we consider densities of the form $V^{-\beta}$ as they are natural in the context of Weighted Poincaré inequalities.

Theorem 1 (Weighted Poincaré Inequality; Bobkov and Ledoux (2009)) *Let the target density be of the form $\pi_\beta \propto V^{-\beta}$ with $\beta > d$ and $V \in \mathcal{C}^2(\mathbb{R}^d)$ positive, convex and with $(\nabla^2 V)^{-1}(x)$ well-defined for all $x \in \mathbb{R}^d$. For any smooth and π_β-integrable function g on $\mathbb{R}^d$ and $G = Vg$,*

$$(\beta+1) Var_{\pi_\beta}(g) \leq \int_{\mathbb{R}^d} \frac{\langle (\nabla^2 V)^{-1} \nabla G, \nabla G \rangle}{V} d\pi_\beta + \frac{d}{\beta - d} \left(\int_{\mathbb{R}^d} g d\pi_\beta \right)^2. \tag{2}$$

A canonical example of a heavy-tailed density that satisfies the conditions in Theorem 1, and hence (2), is the multivariate t-distribution. In particular, we consider the following Itô diffusion process

$$dX_t = -(\beta - 1)\nabla V(X_t)dt + \sqrt{2V(X_t)}dB_t, \tag{3}$$

where $(B_t)_{t \geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. The Itô diffusion in (3) converges exponentially fast to the target π_β in the χ^2-divergence as long as it satisfies the Weighted Poincaré inequality and additional mild assumptions; see Proposition 4 for details. Hence,

we study the oracle complexity of the Euler-Maruyama discretization of (3), for sampling from heavy-tailed densities. Our proofs are based on *mean-square analysis* techniques, a popular technique to analyze numerical discretizations of stochastic differential equations; see, for example, Milstein and Tretyakov (2004) for an overview. Our results in this paper pushes mean-square analysis to its limits; the heavy-tailed densities we consider invariably need to have only finite variance, which is the minimum requirement when using this technique.

1.1 Our Contributions

In this work, we make the following contributions:

- In Theorem 5, we provide upper bounds on the number of iterations required by the Euler-Maruyama discretization of (3) to obtain a sample that is ϵ-close in the Wasserstein-2 metric to the target density. The established bounds are in terms of certain (first and second-order) moments of the target density π. Our proof technique is based on a mean-squared analysis; we demonstrate that for the case of multivariate t-distributions, our analysis is non-vacuous as long as the density has finite variance, a necessary condition to carry out the mean-squared analysis.
- While the result in Theorem 5 assumes access to the exact gradient of the unnormalized target density function (referred to as the first-order setting), in Theorem 16, we analyze the case when the gradient is estimated based on function evaluations (the zeroth-order setting) based on a Gaussian smoothing technique.
- We provide several illustrative examples highlighting the differences between the results in the first and the zeroth-order setting. Specifically, in Section 5 we show that for the multivariate t-distribution with smaller degrees of freedom, (and hence the truly heavy-tailed case) the gradient estimation error is dominated by the discretization error. Whereas, in the case with larger degrees of freedom (and hence the comparatively moderately heavy-tailed case), the discretization error is of comparable order to the gradient estimation error. Hence, the zeroth-order algorithm matches the iteration complexity of the first-order algorithm by using mini-batch gradient estimators.

1.2 Related Work

Non-asymptotic iteration complexity of different discretizations of (1) have been analyzed extensively in the last decade. The analysis of the Unadjusted Langevin Algorithm (ULA) under various light-tailed assumptions was carried out, for example, in Dalalyan (2017); Durmus and Moulines (2017); Dalalyan and Karagulyan (2019); Durmus et al. (2019); Lee et al. (2020); Shen and Lee (2019); He et al. (2020); Chen et al. (2020); Durmus et al. (2019); Dalalyan et al. (2022); Li and Erdogdu (2023); Chen et al. (2020); Chewi et al. (2022) and references therein. In particular, Vempala and Wibisono (2019); Erdogdu and Hosseinzadeh (2021); Chewi et al. (2022) analyzed the performance of ULA under various functional inequalities suited to light-tailed densities. Furthermore, the recent work of Balasubramanian et al. (2022) analyzed the performance of (averaged) ULA for target densities that are only Hölder continuous, albeit in the weaker Fisher information metric.

Several works, for example, Dwivedi et al. (2019); Chewi et al. (2021); Wu et al. (2022), analyzed the Metropolis-Adjusted Langevin Algorithm (MALA) in light-tailed settings. The

proximal sampler algorithm was analyzed under various light-tailed assumptions in Lee et al. (2021); Chen et al. (2022); Liang and Chen (2022); Gopi et al. (2022); Fan et al. (2023); Gopi et al. (2023). The iteration complexity of the widely used Hamiltonian Monte Carlo algorithm and discretizations of underdamped Langevin diffusions were analyzed, for example, in Dalalyan and Riou-Durand (2020); Bou-Rabee et al. (2020); Chen et al. (2020); Ma et al. (2021); Monmarché (2021); Cao et al. (2021); Wang and Wibisono (2022); Chen and Vempala (2022). We also refer interested readers to Lu and Wang (2022); Ding and Li (2021) for non-asymptotic analyses of other MCMC algorithms used in practice in light-tailed settings.

In the context of heavy-tailed sampling, Kamatani (2018) considered the scaling limits of appropriately modified Metropolis random walk in an asymptotic setting. Johnson and Geyer (2012) proposed a variable transformation method in the context of Metropolis Random Walk algorithms. Here, the heavy-tailed density is converted into a light-tailed one based on certain invertible transformations so that one can leverage the rich literature on light-tailed sampling algorithms. Similar ideas were also examined recently in Yang et al. (2022). It is also worth highlighting that Deligiannidis et al. (2019); Durmus et al. (2020) and Bierkens et al. (2019) used the transformation approach for proving asymptotic exponential ergodicity of bouncy particle and zig-zag samplers respectively, in the heavy-tailed setting. We also point out the recent works of Andrieu et al. (2021) and Andrieu et al. (2022) that establish similar sub-exponential ergodicity results for other sampling methods such as the piecewise deterministic Markov process Monte Carlo, independent Metropolis-Hastings sampler and pseudo-marginal methods in the polynomially heavy-tailed setting. The works of Şimşekli et al. (2020); Huang et al. (2021) and Zhang and Zhang (2023) established exponential ergodicity results for diffusions driven by α-stable processes with heavy-tailed densities as its equilibrium in the continuous-time setting. However, the problem of obtaining convergence results for practical discretizations of these diffusions is still largely open.

The literature on non-asymptotic oracle complexity analysis of heavy-tailed sampling is extremely limited. Chandrasekaran et al. (2009) considered the iteration complexity of Metropolis random walk algorithm for sampling from s-concave distributions. He et al. (2023) considered ULA on a class of transformed densities (i.e., the heavy-tailed density is transformed to a light-tailed one with an invertible transformation, similar to Johnson and Geyer (2012)) and established non-asymptotic oracle complexity results. However, they focused mainly on the case of isotropic densities. Li et al. (2019) analyzed a class of discretizations of general Itô diffusions that admit heavy-tailed equilibrium densities. A detailed comparison to Li et al. (2019) is provided in Section 5.

The recent works by Hsieh et al. (2018); Zhang et al. (2020); Chewi et al. (2020); Ahn and Chewi (2021); Jiang (2021); Li et al. (2022) also considered sampling based on discretizations of the Mirror Langevin diffusions. The above-mentioned works mainly focus on sampling from constrained densities. The continuous-time convergence is analyzed typically under the so-called mirror Poincaré inequalities, which are generalizations of the Brascamp-Lieb inequalities in a different direction compared to the Weighted Poincaré inequalities. The discretization analysis by Li et al. (2022) is based on mean-squared analysis.

As mentioned previously, our work leverages the literature on weighted functional inequalities, that are satisfied by heavy-tailed densities. The weighted Poincare inequality was

introduced in Blanchet et al. (2009) and Bobkov and Ledoux (2009), and using an extension of the Brascamp-Lieb inequality, is shown to hold for the class of s-concave densities. We also refer the interested reader to Cattiaux et al. (2010, 2011); Bonnefont et al. (2016); Cordero-Erausquin and Gozlan (2017); Cattiaux et al. (2019) for various extensions and improvements of the works of Blanchet et al. (2009) and Bobkov and Ledoux (2009).

1.3 Notation

We use the following notation throughout the rest of the paper.

- $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product and $|\cdot|$ denotes the Euclidean norm.
- For two matrices A and B, $A \preceq B$ means that $B - A$ is positive semi-definite. The 2-norm of any $d \times d$ matrix A is denoted as $\|A\|_2$. I_d is the $d \times d$ identity matrix.
- Δ denotes the Laplacian, and ∇ denotes the gradient of a given function.
- $\mathcal{C}^2(\mathbb{R}^d)$ refers to the set of all real functions on $\mathbb{R}^d$ that are twice continuously differentiable. $\mathcal{C}_c^2(\mathbb{R}^d)$ refers to the set of all functions in $\mathcal{C}^2(\mathbb{R}^d)$ with compact support.
- The Wasserstein-2 distance between two probability measures μ and ν on $\mathbb{R}^d$ is given by

$$W_2(\mu, \nu) := \inf_{\zeta \in C(\mu,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \zeta(dx, dy) \right)^{\frac{1}{2}}.$$

  where $C(\mu, \nu)$ is the set of all measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are μ and ν respectively.
- The χ^2 divergence from a probability measure ν to a probability measure μ is defined as

$$\chi^2(\nu|\mu) := \int_{\mathbb{R}^d} \left(\frac{\nu(dx)}{\mu(dx)} - 1 \right)^2 \mu(dx).$$

- The gamma and beta functions are given by:

$$\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} dt, \quad \forall\ z > 0, \quad \text{and} \quad B(x, y) := \int_0^1 t^{x-1}(1-t)^{y-1} dt, \quad \forall\ x, y > 0.$$

- For two positive quantities $f(d), g(d)$ depending on d, we define $f(d) = O(g(d))$ if there exists a constant $C > 0$ such that $f(d) \leq Cg(d)$ for all $d > 1$. We define $f(d) = \Theta(g(d))$ if there exist constants $C_1, C_2 > 0$ such that $C_1 g(d) \leq f(d) \leq C_2 g(d)$ for all $d > 1$. We use $\tilde{O}$ to hide log factors in the O notation.

1.4 Organization

In Section 2, we first establish the exponential ergodicity of the Itô diffusion in (3) under certain assumptions that are favorable for the discretization analysis. We next provide our main results on the non-asymptotic oracle complexity of the Euler-Maruyama discretization of (3). In Section 3, we provide moment computations in the heavy-tailed setting that are required to obtain explicit rates from the results in Section 2. In Section 4, we provide an extension of our results to the zeroth-order setting. In Section 5 we provide several illustrative examples. We discuss further implications of our assumptions in Section 6. The proofs are provided in Section 8 and in Appendices A, B and C.

2. Itô Discretizations and Weighted Poincare inequalities

In this section, our goal is to analyze the Itô diffusion in (3) which admits a specific class of heavy-tailed densities as its stationary density. Let X_0 follow distribution ρ_0 and denote the distribution of X_t by ρ_t for all $t \geq 0$. For any function $\psi \in \mathcal{C}_c^2(\mathbb{R}^d)$, the infinitesimal generator of (3) is given by

$$\mathcal{L}\psi = -(\beta - 1)\langle \nabla V, \nabla \psi \rangle + V \Delta \psi. \tag{4}$$

Hence, the Fokker-Planck equation corresponding to (3) is

$$\partial_t \rho_t = \nabla \cdot (\beta \rho_t \nabla V + V \nabla \rho_t) = \nabla \cdot \left(\rho_t V \nabla \log \frac{\rho_t}{\pi_\beta} \right). \tag{5}$$

It follows that, under the conditions in Theorem 1, $\pi_\beta \propto V^{-\beta}$ is the unique stationary density of (3). We next examine the convergence properties of (3) to its stationary density. To do so, we introduce the following assumption.

Assumption 1 *There exists a positive constant C_V such that, for all $x \in \mathbb{R}^d$,*

$$\frac{\langle (\nabla^2 V)^{-1}(x) \nabla V(x), \nabla V(x) \rangle}{V(x)} \leq C_V.$$

When V is radially symmetric, i.e., when $V(x) := \phi(|x|)$ for some $\phi \in \mathcal{C}^2(\mathbb{R}_+)$, the condition in Assumption 1 simplifies as follows. Note that

$$\nabla V(x) = \frac{\phi'(|x|)}{|x|} x, \qquad \text{and} \qquad \nabla^2 V = \left(\phi''(|x|) - \frac{\phi'(|x|)}{|x|} \right) \frac{x \otimes x}{|x|^2} + \frac{\phi'(|x|)}{|x|} I_d,$$

where $\otimes$ denotes outer-product. Hence, it follows that it is sufficient for ϕ to satisfy

$$\phi'(r) \leq (\phi''(r) r) \wedge (C_V \phi(r)/r), \text{ for all } r \geq 0.$$

For example, this property holds with $C_V = p$ if ϕ is a p-order polynomial with $p \geq 2$ and non-negative coefficients. For convenience, we also define the following quantity,

$$\delta := \frac{\beta - 1 - \frac{1}{4} C_V d}{\frac{1}{4} C_V d}. \tag{6}$$

The condition $\delta > 0$ will be used in Theorem 5. The meaning behind the constant δ and the positivity condition will be made clear in Remark 6.

We next provide the following corollary to Theorem 1, motivated by the discussion in Section 2 of Bobkov and Ledoux (2009).

Corollary 2 *Consider the setting of Theorem 1 and suppose further that Assumption 1 holds with $C_V \in (0, \beta + 1)$, then for any smooth, π_β-integrable function, ϕ on $\mathbb{R}^d$,*

$$Var_{\pi_\beta}(\phi) \leq \left(\sqrt{\beta + 1} - \sqrt{C_V} \right)^{-2} \int_{\mathbb{R}^d} \langle V(x) (\nabla^2 V)^{-1}(x) \nabla \phi(x), \nabla \phi(x) \rangle \pi_\beta(x) dx. \tag{7}$$

Proof [Proof] We start from (2), assume that $\int_{\mathbb{R}^d} g d\pi_\beta = 0$. Then (2) could be rewritten as

$$(\beta+1)\int_{\mathbb{R}^d} g(x)^2\pi_\beta(x)dx \le \int_{\mathbb{R}^d} \frac{\langle(\nabla^2 V)^{-1}(x)\nabla(gV)(x), \nabla(gV)(x)\rangle}{V(x)}\pi_\beta(x)dx.$$

Now, note that we have the following elementary bound

$$\langle A(u+v),(u+v)\rangle \le r\langle Au,u\rangle + \frac{r}{r-1}\langle Av,v\rangle, \quad u,v\in\mathbb{R}^d, r>1,$$

for any arbitrary positive definite symmetric matrix $A\in\mathbb{R}^{d\times d}$. Hence, we obtain

$$\begin{aligned}(\beta+1)\int_{\mathbb{R}^d} g(x)^2\pi_\beta(x)dx \le r\int_{\mathbb{R}^d} \frac{\langle(\nabla^2 V)^{-1}(x)g(x)\nabla V(x), g(x)\nabla V(x)\rangle}{V(x)}\pi_\beta(x)dx \\ + \frac{r}{r-1}\int_{\mathbb{R}^d} \frac{\langle(\nabla^2 V)^{-1}(x)V(x)\nabla g(x), V(x)\nabla g(x)\rangle}{V(x)}\pi_\beta(x)dx.\end{aligned}$$

Invoking the condition in Assumption 1, we further obtain

$$\begin{aligned}(\beta+1)\int_{\mathbb{R}^d} g(x)^2\pi_\beta(x)dx \le rC_V\int_{\mathbb{R}^d} g(x)^2\pi_\beta(x)dx \\ + \frac{r}{r-1}\int_{\mathbb{R}^d}\langle V(x)(\nabla^2 V)^{-1}(x)\nabla g(x), \nabla g(x)\rangle\pi_\beta(x)dx,\end{aligned}$$

which then implies that, for any $r\in(1,(\beta+1)/C_V)$,

$$\int_{\mathbb{R}^d} g(x)^2\pi_\beta(x)dx \le \frac{r}{(r-1)(\beta+1-rC_V)}\int_{\mathbb{R}^d}\langle V(x)(\nabla^2 V)^{-1}(x)\nabla g(x), \nabla g(x)\rangle\pi_\beta(x)dx.$$

With the choice of $r \coloneqq \sqrt{\frac{\beta+1}{C_V}} > 1$, we get that for all g such that $\int g d\pi_\beta = 0$, and

$$\int_{\mathbb{R}^d} g(x)^2\pi_\beta(x)dx \le \left(\sqrt{\beta+1}-\sqrt{C_V}\right)^{-2}\int_{\mathbb{R}^d}\langle V(x)(\nabla^2 V)^{-1}(x)\nabla g(x), \nabla g(x)\rangle\pi_\beta(x)dx.$$

For all general ϕ, letting $g=\phi-\int\phi d\pi_\beta$, we get

$$Var_{\pi_\beta}(\phi) \le \left(\sqrt{\beta+1}-\sqrt{C_V}\right)^{-2}\int_{\mathbb{R}^d}\langle V(x)(\nabla^2 V)^{-1}(x)\nabla\phi(x), \nabla\phi(x)\rangle\pi_\beta(x)dx.$$

■

When V is strongly convex, Assumption 1 holds under the following sufficient condition.

Assumption 2 *The function $V:\mathbb{R}^d\to(0,\infty)$ is twice continuously differentiable and V satisfies*

(1) V is α-strongly convex, i.e. $\nabla^2 V(x)\succeq\alpha I_d$ for all $x\in\mathbb{R}^d$.

(2) There exists a positive constant C_V such that, for all $x\in\mathbb{R}^d$,

$$\frac{\langle\nabla V(x),\nabla V(x)\rangle}{V(x)} \le \alpha C_V.$$

The following result follows immediately from Assumption 2.

Lemma 3 *Let $\beta > d$. If Assumption 2 holds with $C_V \in (0, \beta + 1)$, then for any smooth, π_β integrable function ϕ on $\mathbb{R}^d$, we have*

$$Var_{\pi_\beta}(\phi) \leq \alpha^{-1} \left(\sqrt{\beta+1} - \sqrt{C_V}\right)^{-2} \int_{\mathbb{R}^d} V(x)|\nabla\phi(x)|^2 \pi_\beta(x)dx. \tag{8}$$

With (8), we can show the exponential decay in χ^2-divergence along (3). The proof of the following proposition is standard, and we include it here for completeness.

Proposition 4 *Under the conditions in Lemma 3, for (X_t) following diffusion (3) with ρ_t being the distribution of X_t, we have*

$$\chi^2(\rho_t|\pi_\beta) \leq \exp\left(-2\alpha \left(\sqrt{\beta+1} - \sqrt{C_V}\right)^2 t\right) \chi^2(\rho_0|\pi_\beta). \tag{9}$$

Proof [Proof of Proposition 3] First we can calculate the derivative of $\chi^2(\rho_t|\pi)$ via (5),

$$\begin{aligned}
\frac{d}{dt}\chi^2(\rho_t|\pi_\beta) &= \frac{d}{dt}\int_{\mathbb{R}^d} \left(\frac{\rho_t(x)}{\pi_\beta(x)} - 1\right)^2 \pi_\beta(x)dx \\
&= 2\int_{\mathbb{R}^d} \partial_t\rho_t(x) \left(\frac{\rho_t(x)}{\pi_\beta(x)} - 1\right) dx \\
&= -2\int_{\mathbb{R}^d} \left\langle \nabla\left(\frac{\rho_t}{\pi_\beta}\right)(x), \nabla\log\left(\frac{\rho_t}{\pi_\beta}\right)(x) \right\rangle V(x)\rho_t(x)dx \\
&= -2\int_{\mathbb{R}^d} V(x) \left|\nabla\left(\frac{\rho_t}{\pi_\beta}\right)(x)\right|^2 \pi_\beta(x)dx.
\end{aligned}$$

According to (8), we get

$$\begin{aligned}
\frac{d}{dt}\chi^2(\rho_t|\pi_\beta) &\leq -2\alpha \left(\sqrt{\beta+1} - \sqrt{C_V}\right)^2 \mathrm{Var}_{\pi_\beta}\left(\frac{\rho_t}{\pi_\beta}\right) \\
&= -2\alpha \left(\sqrt{\beta+1} - \sqrt{C_V}\right)^2 \chi^2(\rho_t|\pi_\beta).
\end{aligned}$$

Finally, (9) follows from Gronwall's inequality. ■

The above result shows that for the class of π_β satisfying Assumption 2, the Itô diffusion in (3), converges exponentially fast to its stationary density. Hence, time-discretizations of (3) provide a practical way of sampling from that class of densities. The Euler-Maruyama discretization to (3) is given by

$$x_{k+1} = x_k - h(\beta - 1)\nabla V(x_k) + \sqrt{2hV(x_k)}\xi_{k+1}, \tag{10}$$

where $h > 0$ is the step size and $\{\xi_k\}_{k=1}^\infty$ is a sequence of i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$. We now present our main result on the iteration complexity of (10) for sampling from π_β. We state our discretization result, based on a mean-square analysis, in the W_2 metric. In particular, we highlight that Proposition 4 requires that condition that $\beta > d$, in addition to Assumption 2, whereas Theorem 5 below, does not. In Section 6, we revisit these conditions and provide additional insights. Obtaining convergence results in the stronger χ^2-divergence is left for future work.

Theorem 5 *Let V be gradient-Lipschitz with parameter $L > 0$, satisfying Assumption 2. Recall the definition of δ in (6) and assume that $\delta > 0$. Let $(x_k)_{k=0}^{\infty}$ be generated from (10) with ν_k denoting the distribution of x_k, for all $k \geq 0$. Then with the step-size,*

$$h < \min\left(\frac{1}{4(\beta-1)L}, \frac{2\delta}{3(1+\delta)\alpha(\beta-1)}\right),$$

the decay of Wasserstein-2 distance along the Markov chain $(x_k)_{k=0}^{\infty}$ can be described by the following equation: For all $k \geq 1$,

$$W_2(\nu_k, \pi_\beta) \leq (1-A)^k W_2(\nu_0, \pi_\beta) + \frac{C}{A} + \frac{B}{\sqrt{A(2-A)}}. \tag{11}$$

with A, B and C given respectively in (50), (51) and (52).

Remark 6 (Constant δ) *We now motivate the definition and the condition on the constant δ based on exponential contractivity arguments. Let X_t, Y_t be two different solutions to the same stochastic differential equation (SDE) with initial conditions x, y respectively. We say the SDE is W_2-exponential contractive if there exists a constant $\kappa > 0$, such that*

$$W_2(L(X_t), L(Y_t)) \leq e^{-\kappa t}\ |x - y|,$$

where, by $L(X)$, we refer to the law of X.

Uniform dissipativity is a sufficient condition for exponential contractivity (Gorham et al., 2019, Theorem 10). The uniform dissipativity condition for (3) can be represented as

$$-(\beta-1)\langle \nabla V(x) - \nabla V(y), x - y\rangle + \frac{1}{2}\left\|\sqrt{2V(x)}I_d - \sqrt{2V(y)}I_d\right\|_F^2 \leq -\kappa|x-y|^2,$$

or equivalently as

$$-(\beta-1)\langle \nabla V(x) - \nabla V(y), x - y\rangle + d|\sqrt{V(x)} - \sqrt{V(y)}|^2 \leq -\kappa|x-y|^2.$$

When V satisfies Assumption 2, a sufficient condition for the above uniform dissipativity condition is given by

$$-\alpha(\beta-1)|x-y|^2 + \frac{d}{4}\alpha C_V|x-y|^2 \leq -\kappa|x-y|^2,$$

or equivalently,

$$\alpha\left(\beta - 1 - \frac{d}{4}C_V\right) \leq \kappa.$$

The sufficient condition coincides with the condition that $\delta > 0$ in Theorem 5, which also motivates the assumption in Theorem 5.

Remark 7 (Iteration complexity) *With Theorem 5, we can calculate the order of the iteration complexity to reach an ϵ-accuracy in Wasserstein-2 distance. With* (50),(51),(52), *we have*

$$\frac{C}{A} = \frac{9(\delta+1)L}{\alpha\delta} d^{\frac{1}{2}} h^{\frac{1}{2}} \mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + \frac{6(\delta+1)L}{\alpha\delta}(\beta-1)h\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}},$$
$$\frac{B}{\sqrt{A(2-A)}} \leq \frac{8(\delta+3)}{\delta} d^{\frac{1}{2}} h^{\frac{1}{2}} \mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + \frac{8(\delta+3)}{\delta}(\beta-1)h\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}.$$

The above display implies that

$$\frac{C}{A} + \frac{B}{\sqrt{A(2-A)}} \leq \frac{9(\delta+3)}{\delta}\left(1+\frac{L}{\alpha}\right)\left(d^{\frac{1}{2}} h^{\frac{1}{2}} \mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + (\beta-1)h\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}\right).$$

Hence, we get $\frac{C}{A} + \frac{B}{\sqrt{A(2-A)}} < \epsilon/2$ if the step-size h satisfies

$$h < \min\left\{ \frac{\delta^2 \mathbb{E}_{\pi_\beta}\left[V(X)\right]^{-1}\epsilon^2}{81d(\delta+3)^2(1+\frac{L}{\alpha})^2}, \frac{\delta\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{-\frac{1}{2}}\epsilon}{81(\beta-1)(\delta+3)(1+\frac{L}{\alpha})} \right\}. \tag{12}$$

Defining $K_\epsilon = \log\left(2W_2(\nu_0,\pi_\beta)/\epsilon\right)$, we have $W_2(\nu_k,\pi_\beta) < \epsilon$ for all $k \geq K$ with

$$\begin{aligned} K &= \frac{3(1+\delta)}{\alpha(\beta-1)\delta h^*} K_\epsilon \\ &\leq 273 \max\left\{ \frac{(\delta+3)^3(1+\frac{L}{\alpha})^2 d\mathbb{E}_{\pi_\beta}\left[V(X)\right]}{\alpha\delta^3(\beta-1)\epsilon^2}, \frac{(\delta+3)^2(1+\frac{L}{\alpha})\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}}{\alpha\delta^2\epsilon} \right\} K_\epsilon. \end{aligned} \tag{13}$$

Recall the definition of δ in (6). *The order of K depends on the order of δ. That is, we have the following two cases:*

- *If $\delta = O(1)$ and $\beta = O(d)$, we have that*

$$K = \tilde{O}\left(\frac{1}{\alpha\epsilon^2}\left(1+\frac{L}{\alpha}\right)^2 \mathbb{E}_{\pi_\beta}\left[V(X)\right] + \frac{1}{\alpha\epsilon}\left(1+\frac{L}{\alpha}\right)\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}} \right).$$

- *If $\delta = O(1/d)$ and $\beta = O(d)$, we have that*

$$K = \tilde{O}\left(\frac{d^3}{\alpha\epsilon^2}\left(1+\frac{L}{\alpha}\right)^2 \mathbb{E}_{\pi_\beta}\left[V(X)\right] + \frac{d^2}{\alpha\epsilon}\left(1+\frac{L}{\alpha}\right)\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}} \right).$$

In order to obtain more explicit iteration complexity bounds from Remark 7, it is required to compute bounds on the following two quantities: $\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]$ and $\mathbb{E}_{\pi_\beta}\left[V(X)\right]$.

3. Moment Bounds

In this section, we compute moment bounds under the conditions in Theorem 5.

3.1 An Example: Multivariate t-distribution

We first start with the isotropic case.

Proposition 8 *Let $\pi_\beta = Z_\beta^{-1} V^{-\beta}$ with $\beta > d/2 + 1$, $V(x) = 1 + |x|^2$ and $Z_\beta = \int_{\mathbb{R}^d} (1 + |x|^2)^{-\beta} dx$. We have*

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = \frac{\beta - 1}{\beta - 1 - \frac{d}{2}} \quad \text{and} \quad \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] = \frac{2d}{\beta - 1 - \frac{d}{2}}. \tag{14}$$

Proof Let $A_d(1)$ denote the surface area of the unit sphere in d dimensions. By a standard calculation, we have that, for all $\beta > \frac{d}{2}$,

$$\begin{aligned} Z_\beta &= \int_{\mathbb{R}^d} (1 + |x|^2)^{-\beta} dx = \int_0^\infty (1 + r^2)^{-\beta} r^{d-1} dr A_d(1) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_0^\infty (1 + R)^{-\beta} R^{\frac{d}{2}-1} dR \\ &= \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_0^1 u^{\frac{d}{2}-1} (1 - u)^{\beta - \frac{d}{2} - 1} du = \frac{\pi^{\frac{d}{2}} B(\frac{d}{2}, \beta - \frac{d}{2})}{\Gamma(\frac{d}{2})}, \end{aligned}$$

where B is the beta function. In the above calculation, the second identity follows from a change to polar coordinates. The third identity follows from a substitution with $R = r^2$ and the fourth identity follows from a substitution $u = R/(1 + R)$. Therefore for all $\beta > d/2+1$, we have that

$$\begin{aligned} \mathbb{E}_{\pi_\beta}\left[V(X)\right] &= Z_\beta^{-1} \int_{\mathbb{R}^d} (1 + |x|^2)(1 + |x|^2)^{-\beta} dx = \frac{Z_{\beta-1}}{Z_\beta} = \frac{\pi^{\frac{d}{2}} B(\frac{d}{2}, \beta - 1 - \frac{d}{2})}{\Gamma(\frac{d}{2})} \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}} B(\frac{d}{2}, \beta - \frac{d}{2})} \\ &= \frac{B(\frac{d}{2}, \beta - 1 - \frac{d}{2})}{B(\frac{d}{2}, \beta - \frac{d}{2})} = \frac{\Gamma(\frac{d}{2})\Gamma(\beta - 1 - \frac{d}{2})}{\Gamma(\beta - 1)} \frac{\Gamma(\beta)}{\Gamma(\frac{d}{2})\Gamma(\beta - \frac{d}{2})} = \frac{\beta - 1}{\beta - 1 - \frac{d}{2}}. \end{aligned}$$

where the fourth identity follows from the property of Beta function, $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ and the fifth identity follows from the property of Γ function, $\Gamma(1 + z) = z\Gamma(z)$. For the other expectation, we have

$$\begin{aligned} \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] &= Z_\beta^{-1} \int_{\mathbb{R}^d} |2x|^2 (1 + |x|^2)^{-\beta} dx = 4 Z_\beta^{-1} A_{d-1}(1) \int_0^\infty r^2 (1 + r^2)^{-\beta} r^{d-1} dr \\ &= \frac{4\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}) Z_\beta} \int_0^\infty R^{\frac{d}{2}} (1 + R)^{-\beta} dR = \frac{4\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}) Z_\beta} \int_0^1 u^{\frac{d}{2}} (1 - u)^{\beta - \frac{d}{2} - 2} du \\ &= \frac{4\pi^{\frac{d}{2}} B(\frac{d}{2} + 1, \beta - \frac{d}{2} - 1)}{\Gamma(\frac{d}{2})} \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}} B(\frac{d}{2}, \beta - \frac{d}{2})} = \frac{4B(\frac{d}{2} + 1, \beta - \frac{d}{2} - 1)}{B(\frac{d}{2}, \beta - \frac{d}{2})} \\ &= 4 \frac{\Gamma(\frac{d}{2} + 1)\Gamma(\beta - \frac{d}{2} - 1)}{\Gamma(\beta)} \frac{\Gamma(\beta)}{\Gamma(\frac{d}{2})\Gamma(\beta - \frac{d}{2})} = \frac{2d}{\beta - \frac{d}{2} - 1}, \end{aligned}$$

where we apply the same substitutions and properties of Beta functions and Gamma functions in the above calculation. ■

Remark 9 *If π_β is the class of isotropic multivariate t-distributions, with the results in Proposition 8, the order of the two expectations in terms of the dimension parameter d is given as follows,*

- *when $\beta > \frac{d}{2} + 1$ and $\beta - 1 - \frac{d}{2} = O(d)$, we have*
$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = O(1), \qquad \text{and} \qquad \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] = O(1).$$
- *when $\beta > \frac{d}{2} + 1$ and $\beta - 1 - \frac{d}{2} = O(1)$, we have*
$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = O(d), \qquad \text{and} \qquad \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] = O(d).$$

For a general class of non-isotropic multivariate t-distribution, we consider $\pi_\beta = Z_\beta^{-1} V^{-\beta}$ with $V(x) = 1 + x^T \Sigma x$ where Σ is a strictly positive-definite $d \times d$ matrix. In Roth (2012), it's been shown that for any $\beta > \frac{d}{2}$, the normalization constant is

$$Z_\beta = \frac{\Gamma(\frac{\nu}{2})\pi^{\frac{d}{2}}\sqrt{\det(\Sigma)}}{\Gamma(\frac{\nu+d}{2})} = \frac{\Gamma(\beta - \frac{d}{2})\pi^{\frac{d}{2}}\sqrt{\det(\Sigma)}}{\Gamma(\beta)}.$$

Therefore for any $\beta > \frac{d}{2} + 1$, we have

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = \frac{Z_{\beta-1}}{Z_\beta} = \frac{\Gamma(\beta)\Gamma(\beta - 1 - \frac{d}{2})}{\Gamma(\beta-1)\Gamma(\beta - \frac{d}{2})} = \frac{\beta - 1}{\beta - 1 - \frac{d}{2}},$$

and

$$\begin{aligned}
\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] &= Z_\beta^{-1} \int_{\mathbb{R}^d} \langle \nabla V(x), V(x)^{-\beta} \nabla V(x) \rangle dx \\
&= -Z_\beta^{-1} \int_{\mathbb{R}^d} V(x) \nabla \cdot \left(V(x)^{-\beta} \nabla V(x) \right) dx \\
&= \beta \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] - Z_\beta^{-1} \int_{\mathbb{R}^d} \Delta V(x) V(x)^{-(\beta-1)} dx.
\end{aligned}$$

The above identity implies

$$\begin{aligned}
\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] &= (\beta-1)^{-1} Z_\beta^{-1} \int_{\mathbb{R}^d} \Delta V(x) V(x)^{-(\beta-1)} dx \\
&\leq (\beta-1)^{-1} Z_\beta^{-1} \int_{\mathbb{R}^d} \operatorname{trace}(\nabla^2 V(x)) V(x)^{-(\beta-1)} dx \\
&\leq \frac{\operatorname{trace}(\Sigma)}{\beta - 1} \mathbb{E}_{\pi_\beta}\left[V(X)\right] \\
&\leq \frac{\operatorname{trace}(\Sigma)}{\beta - 1 - \frac{d}{2}}
\end{aligned}$$

where the second inequality follows from the fact that $\nabla^2 V(x) = \Sigma$.

Remark 10 *If π_β is in the class of non-isotropic multivariate t-distributions, the order of the two expectations in terms of the dimension parameter d is as follows,*

- *when $\beta > \frac{d}{2} + 1$ and $\beta - 1 - \frac{d}{2} = O(d)$, we have*

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = O(1), \qquad \text{and} \qquad \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] = O(d^{-1}\text{trace}(\Sigma)).$$

- *when $\beta > \frac{d}{2} + 1$ and $\beta - 1 - \frac{d}{2} = O(1)$, we have*

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = O(d), \qquad \text{and} \qquad \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] = O(\text{trace}(\Sigma)).$$

3.2 Non-isotropic densities with quadratic-like V outside of a ball

In this section, we estimate the expectations for a class of non-isotropic densities in the form of $\pi_\beta \propto V^{-\beta}$ with V satisfying the following Lyapunov condition:

$$\exists\ \varepsilon, R > 0 \text{ such that } \Delta V(x) - (\beta - 1)\frac{|\nabla V(x)|^2}{V(x)} \leq -\varepsilon \qquad \forall\ |x| \geq R. \tag{15}$$

The above Lyapunov condition characterizes the class of V that are 'quadratic-like' outside a ball of radius R. If we assume that V has Lipschitz gradients, then when β is sufficiently large, the above assumption is satisfied if V satisfies the PL inequality $|\nabla V(x)|^2 \geq a^2 V(x)$ wherever $|x| \geq R$ with some $a > 0$ and it is from this inequality that quadratic growth follows. In particular, if V satisfies the gradient Lipschitz assumption with parameter L, we have that for all $\beta \geq 1 + a^{-2}(dL + \varepsilon)$,

$$\Delta V(x) - (\beta - 1)\frac{|\nabla V(x)|^2}{V(x)} \leq dL - (\beta - 1)a^2 \leq -\varepsilon \qquad \forall\ |x| \geq R,$$

thereby leading to the Lyapunov condition in (15).

Proposition 11 *If $V \in \mathcal{C}^2(\mathbb{R}^d)$ is positive, L-gradient Lipschitz and satisfies* (15)*, then we have*

$$\mathbb{E}_{\pi_\beta}\left[V(x)\right] \leq (dL + \varepsilon)\max_{|x| \leq R} V(x), \quad \text{and} \quad \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] \leq \frac{dL\,(dL + \varepsilon)}{(\beta - 1)}\max_{|x| \leq R} V(X). \tag{16}$$

Proof Since $\mathcal{L}$ is ergodic with stationary distribution π_β, we have

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = \lim_{t \to \infty} \mathbb{E}\left[V(X_t)\right],$$

with $(X_t)_{t \geq 0}$ being the solution to (3) with initial condition $X_0 = x$. We will first bound $\mathbb{E}\left[V(X_t)\right]$ and then take $t \to \infty$. Let $(P_t)_{t \geq 0}$ be the Markov semigroup of (3), then

$$\frac{d}{dt}\mathbb{E}_{\pi_\beta}\left[V(X_t)\right] = \frac{d}{dt}P_t V(x) = P_t \mathcal{L} V(x).$$

With (4), we have

$$\mathcal{L}V(x) = V(x)\left[\Delta V(x) - (\beta - 1)\frac{|\nabla V(x)|^2}{V(x)}\right]$$

$$\begin{aligned}
&\leq V(x)\left(-\varepsilon 1_{|x|\geq R} + dL 1_{|x|<R}\right) \\
&\leq -\varepsilon V(x) + (dL+\varepsilon)\max_{|x|\leq R} V(x),
\end{aligned}$$

where the first inequality follows from (15) and the fact that $\Delta V \leq d\left\|\nabla^2 V\right\|_2$. Therefore we obtain

$$\frac{d}{dt}P_t V(x) \leq -\varepsilon P_t V(x) + (dL+\varepsilon)\max_{|x|\leq R} V(x),$$

and it follows from Gronwall's inequality that

$$\mathbb{E}_{\pi_\beta}\left[V(X_t)\right] = P_t V(x) \leq V(x)e^{-\varepsilon t} + \left(1-e^{-\varepsilon t}\right)(dL+\varepsilon)\max_{|x|\leq R} V(x).$$

We hence have that $\mathbb{E}_{\pi_\beta}\left[V(X)\right] \leq (dL+\varepsilon)\max_{|x|\leq R} V(x)$ by taking $t\to\infty$. For the other expectation, we have

$$\begin{aligned}
\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] &= Z_\beta^{-1}\int_{\mathbb{R}^d}\langle \nabla V(x), V(x)^{-\beta}\nabla V(x)\rangle dx \\
&= -Z_\beta^{-1}\int_{\mathbb{R}^d} V(x)\nabla\cdot\left(V(x)^{-\beta}\nabla V(x)\right)dx \\
&= \beta\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] - Z_\beta^{-1}\int_{\mathbb{R}^d}\Delta V(x)V(x)^{-(\beta-1)}dx.
\end{aligned}$$

The above identity implies

$$\begin{aligned}
\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] &= (\beta-1)^{-1}Z_\beta^{-1}\int_{\mathbb{R}^d}\Delta V(x)V(x)^{-(\beta-1)}dx \\
&\leq (\beta-1)^{-1}Z_\beta^{-1}\int_{\mathbb{R}^d}\mathrm{trace}(\nabla^2 V(x))V(x)^{-(\beta-1)}dx \\
&\leq (\beta-1)^{-1}Z_\beta^{-1}dL\int_{\mathbb{R}^d} V(x)^{-(\beta-1)}dx \\
&= \frac{dL}{\beta-1}\mathbb{E}_{\pi_\beta}\left[V(X)\right] \\
&\leq \frac{dL\,(dL+\varepsilon)}{\beta-1}\max_{|x|\leq R} V(x).
\end{aligned}$$

■

3.3 General Case

Next we discuss the general case where $\pi_\beta = Z_\beta^{-1}V^\beta$ and $V\in\mathcal{C}^2(\mathbb{R}^d)$ is positive such that there exist constants $\alpha, L>0$ and $\alpha I_d \preceq \nabla^2 V(x) \preceq L I_d$ for all $x\in\mathbb{R}^d$. Since V is strongly convex, there is a unique $x^*\in\mathbb{R}^d$ such that $V(x)\geq V(x^*)>0$ for all $x\in\mathbb{R}^d$ and $\nabla V(x^*)=0$. Without loss of generality, we assume $x^*=0$.

Proposition 12 *Let $\beta > \frac{d}{2}+1$. If $V \in \mathcal{C}^2(\mathbb{R}^d)$ is positive, α-strongly convex and L-gradient Lipschitz, we have for any $r \in (0, \beta - \frac{d}{2} - 1)$,*

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] \leq \left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}} V(0) \left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})}\right)^{\frac{1}{\beta-\frac{d}{2}-r}}, \tag{17}$$

$$\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] \leq \frac{dL}{\beta-1}\left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}} V(0) \left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})}\right)^{\frac{1}{\beta-\frac{d}{2}-r}}. \tag{18}$$

Proof For any $r \in (0, \beta - \frac{d}{2} - 1)$, we have

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = \frac{\int_{\mathbb{R}^d} V(x)V(x)^{-\beta}dx}{Z_\beta} = \frac{Z_{\beta-1}}{Z_\beta} \leq \left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}} V(0) \left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})}\right)^{\frac{1}{\beta-\frac{d}{2}-r}}.$$

where the last inequality follows from Lemma 25. For the other expectation, we have

$$\begin{aligned}
\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] &= Z_\beta^{-1}\int_{\mathbb{R}^d} \langle \nabla V(x), V(x)^{-\beta}\nabla V(x)\rangle dx \\
&= -Z_\beta^{-1}\int_{\mathbb{R}^d} V(x)\nabla\cdot\left(V(x)^{-\beta}\nabla V(x)\right)dx \\
&= \beta\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] - Z_\beta^{-1}\int_{\mathbb{R}^d}\Delta V(x)V(x)^{-(\beta-1)}dx.
\end{aligned}$$

The above identity implies

$$\begin{aligned}
\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] &= (\beta-1)^{-1}Z_\beta^{-1}\int_{\mathbb{R}^d}\Delta V(x)V(x)^{-(\beta-1)}dx \\
&\leq (\beta-1)^{-1}Z_\beta^{-1}\int_{\mathbb{R}^d}\mathrm{trace}(\nabla^2 V(x))V(x)^{-(\beta-1)}dx \\
&\leq (\beta-1)^{-1}Z_\beta^{-1}dL\int_{\mathbb{R}^d} V(x)^{-(\beta-1)}dx \\
&= \frac{dL}{\beta-1}\frac{Z_{\beta-1}}{Z_\beta} \\
&\leq \frac{dL}{\beta-1}\left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}} V(0)\left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})}\right)^{\frac{1}{\beta-\frac{d}{2}-r}}.
\end{aligned}$$

where the last inequality also follows from Lemma 25. ■

Remark 13 *A ratio between Gamma functions appears in* (17) *and* (18)*. The ratio can be written explicitly via properties of Gamma functions.*

- *When d is an even number and $d = 2k$ for some integer k,*

$$\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})} = \frac{\Gamma(r)}{\Gamma(\frac{d}{2}+r)}\frac{\Gamma(\beta)}{\Gamma(\beta-\frac{d}{2})} = \frac{\Gamma(r)}{\Gamma(r)\prod_{i=1}^k(\frac{d}{2}+r-i)}\frac{\Gamma(\beta-\frac{d}{2})\prod_{i=1}^k(\beta-i)}{\Gamma(\beta-\frac{d}{2})}$$

$$= \frac{\prod_{i=1}^{k}(\beta - i)}{\prod_{i=1}^{k}(\frac{d}{2} + r - i)} \leq \left(\frac{\beta - \frac{d}{2}}{r}\right)^{\frac{d}{2}},$$

- *When d is an odd number with $d = 2k - 1$ for some integer k,*

$$\begin{aligned}
\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})} &= \frac{\Gamma(r)}{\Gamma(\frac{d}{2}+r)}\frac{\Gamma(\beta)}{\Gamma(\beta-\frac{d}{2})} = \frac{\Gamma(r)}{\Gamma(\frac{1}{2}+r)\prod_{i=1}^{k-1}(\frac{d}{2}+r-i)}\frac{\Gamma(\beta-\frac{d}{2}+\frac{1}{2})\prod_{i=1}^{k-1}(\beta-i)}{\Gamma(\beta-\frac{d}{2})} \\
&= \frac{\prod_{i=1}^{k-1}(\beta-i)}{\prod_{i=1}^{k-1}(\frac{d}{2}+r-i)}\frac{r^{-1}\Gamma(r+1)}{\Gamma(\frac{1}{2}+r)}\frac{\Gamma(\beta-\frac{d}{2}+\frac{1}{2})}{\Gamma(\beta-\frac{d}{2})} \\
&\leq \left(\frac{\beta-\frac{d}{2}+\frac{1}{2}}{r+\frac{1}{2}}\right)^{k-1} r^{-1}(1+r)^{\frac{1}{2}}\left(\beta-\frac{d}{2}+\frac{1}{2}\right)^{\frac{1}{2}} \\
&\leq \sqrt{\frac{1+r}{r}}\left(\frac{\beta-\frac{d}{2}}{r}\right)^{\frac{d}{2}},
\end{aligned}$$

where the first inequality follows from Gautschi's inequality (Ismail and Muldoon, 1994).

Remark 14 *With Theorem 12 and the upper bounds in Remark 13, we can get the estimations for $\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]$ and $\mathbb{E}_{\pi_\beta}\left[V(X)\right]$: for any $r \in (0, \beta - \frac{d}{2} - 1)$,*

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] \leq V(0)\left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}}\left(\frac{1+r}{r}\right)^{\frac{1}{2(\beta-\frac{d}{2}-r)}}\left(\frac{\beta-\frac{d}{2}}{r}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}}, \tag{19}$$

$$\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] \leq \frac{V(0)dL}{\beta-1}\left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}}\left(\frac{1+r}{r}\right)^{\frac{1}{2(\beta-\frac{d}{2}-r)}}\left(\frac{\beta-\frac{d}{2}}{r}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}}. \tag{20}$$

4. Zeroth-Order Itô Discretization

While previously we consider the case when the gradient of the function V is analytically available to us, we now consider the case when we have access only to the function evaluations. This setting is called the zeroth-order setting and has been recently examined in the context of complexity of sampling in the works of Dwivedi et al. (2019); Lee et al. (2021); Roy et al. (2022). In this setting, we construct an approximation to the gradient via zeroth-order information, i.e., function evaluations. For simplicity, we consider the case of obtaining exact function evaluations. Based on the Gaussian smoothing technique (Nesterov and Spokoiny, 2017; Roy et al., 2022), for any $x \in \mathbb{R}^d$, we define the zeroth order gradient estimator $g_{\sigma,,m}(x)$ as

$$g_{\sigma,m}(x) := \frac{1}{m}\sum_{i=1}^{m}\frac{V(x+\sigma u_i) - V(x)}{\sigma}u_i \tag{21}$$

where $u_i \sim \mathcal{N}(0, I_d)$ are assumed to be independent and identically distributed. The parameter m is called the batch size parameter. Then the zeroth order algorithm to sample

π_β is given by

$$x_{k+1} = x_k - h(\beta - 1)g_{\sigma,m}(x_k) + \sqrt{2V(x_k)}\xi_{k+1} \tag{22}$$

where $h > 0$ is the step size and $\{\xi_{k+1}\}_{k=0}^{\infty}$ is a sequence of independent identically distributed standard Gaussian random vectors in $\mathbb{R}^d$. From Balasubramanian and Ghadimi (2022) and Roy et al. (2022), we recall the following property of $g_{\sigma,m}$.

Proposition 15 *(Roy et al., 2022, Section 8.1) Assume V is L-gradient Lipschitz. Define $\zeta_k = g_{\sigma,m}(x_k) - \nabla V(x_k)$ with $g_{\sigma,m}$ defined in (21) and $\{x_k\}_{k=0}^{\infty}$ generated by (22). We have for any $k \geq 0$,*

$$\mathbb{E}\left[\left|\mathbb{E}\left[\zeta_k | x_k\right]\right|^2\right] \leq L^2\sigma^2 d, \tag{23}$$

and

$$\mathbb{E}\left[\left|\zeta_k - \mathbb{E}\left[\zeta_k | x_k\right]\right|^2\right] \leq \frac{\sigma^2}{2m}L^2(d+3)^3 + \frac{2(d+5)}{m}\mathbb{E}\left[|\nabla V(x_k)|^2\right]. \tag{24}$$

Theorem 16 *Suppose V is gradient-Lipschitz with parameter $L > 0$ and satisfies Assumption 2 with δ in (6). Let $g_{\sigma,m}$ be as defined in (21) and $(x_k)_{k=0}^{\infty}$ be generated from (22) with $x_k \sim \nu_k$ for all $k \geq 0$. Then with the time step size*

$$h < \min\left\{\frac{2\delta}{3(1+\delta)\alpha(\beta-1)}, \frac{\alpha m\delta}{24(1+\delta)(\beta-1)(d+5)L^2}, \frac{1}{4(\beta-1)L}\right\}, \tag{25}$$

the decay of Wasserstein-2 distance along the Markov chain $(x_k)_{k=0}^{\infty}$ can be described by the following equation. For all $k \geq 1$,

$$W_2(\nu_k, \pi_\beta) \leq (1 - A')^k W_2(\nu_0, \pi_\beta) + \frac{C'}{A'} + \frac{B'}{\sqrt{A'(2-A')}}. \tag{26}$$

with A', B' and C' given respectively in (59), (60) and (61).

Remark 17 *With Theorem 16, we can study the iteration complexity to reach an ε-accuracy in Wasserstein-2 distance. In the following discussion, we focus on the dimension dependence and ε dependence in the iteration complexity. When $\beta = \Theta(d)$ and $\alpha, L = \Theta(1)$, and when h satisfies (25), we have*

$$A' = O\left(\delta dh\right), \qquad \frac{C'}{A'} = O\left(\frac{(dh\mathbb{E}_{\pi_\beta}\left[V(X)\right])^{\frac{1}{2}} + dh\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}} + \sigma d^{\frac{1}{2}}}{\delta}\right),$$

$$\frac{B'}{\sqrt{A'(2-A')}} = O\left(\left(\frac{dh}{\delta} + \frac{dh^{\frac{1}{2}}}{(\delta m)^{\frac{1}{2}}}\right)\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}} + \frac{(dh)^{\frac{1}{2}}}{\delta}\mathbb{E}_{\pi_\beta}\left[V(X)\right] + \frac{\sigma d^2 h^{\frac{1}{2}}}{(\delta m)^{\frac{1}{2}}}\right).$$

To ensure $W_2(\nu_K, \pi_\beta) < \varepsilon$, we require that each of

$$(1 - A')^K W_2(\nu_0, \pi_\beta), \qquad \frac{C'}{A'}, \qquad \frac{B'}{\sqrt{A'(2-A')}},$$

is smaller than $\varepsilon/3$. Setting $\sigma = \varepsilon\delta/\sqrt{d}$, and

$$h = O\left(\min\left\{\frac{(\varepsilon\delta)^2}{d}\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{-1}, \frac{\varepsilon\delta}{d}\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{-\frac{1}{2}}, \frac{\varepsilon^2\delta m}{d^2}\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{-1}\right\}\right),$$

we hence obtain that the iteration complexity K is of order

$$K = \tilde{O}\left(\max\left\{\frac{1}{\varepsilon^2\delta^3}\mathbb{E}_{\pi_\beta}\left[V(X)\right], \frac{1}{\varepsilon\delta^2}\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}, \frac{d}{\varepsilon^2\delta^2 m}\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]\right\}\right). \quad (27)$$

The number of function evaluations is hence mK.

5. Illustrative Examples

We now provide illustrative examples to highlight the implications of our results. Before we proceed, we remark that using Mousavi-Hosseini et al. (2023, Corallary 8) and the inequality $\ln\left(1 + \frac{W_2^4(\rho,\pi)}{4\mathbb{E}_\pi[\|x\|^4]}\right) \leq R_2(\rho|\pi)$, we can obtain upper bounds on the oracle complexity of ULA, with *favorable initializations*, for sampling from multivariate t-distributions with finite 4-th moments, in Wasserstein-2 distance. For this class of densities, when the degrees of freedom ν is small (constant order), the oracle complexity of ULA is of order $O(d^{8+12/\nu}\epsilon^{-4(1+4/\nu)})$. When the degree of freedom ν is large (i.e., order $O(d)$), the oracle complexity of ULA is of order $O(d^7\epsilon^{-4})$. Below, we show that our algorithm admits significantly better dependencies on the dimension (d) and the inverse accuracy ($1/\epsilon$).

5.1 Multivariate t-distribution: Large Degree of Freedom

We first consider the isotropic multivariate t-distribution with the degrees of freedom being $d+2$. We choose $V(x) = 1 + |x|^2$, $\beta = d+1$ and $\pi_\beta(x) \propto V(x)^{-\beta} = (1+|x|^2)^{-(d+1)}$. With this choice of V and β, V satisfies Assumption 2 with $\alpha = 2$, $C_V = 2$, and V is L-Lipschitz gradient with $L = 2$. The constant δ in Theorem 5 becomes $\delta = 1$. Furthermore, according to proposition 8, $\mathbb{E}_{\pi_\beta}[V(X)] = 2$ and $\mathbb{E}_{\pi_\beta}[|\nabla V(X)|^2] = 4$.

5.1.1 first order algorithm

According to Theorem 5 and (13), to obtain ϵ-accuracy in Wasserstein-2 distance, the iteration complexity is of order $\tilde{O}(1/\epsilon^2)$. With the same choice of V and β, we check the conditions of Theorem 1 in Li et al. (2019). The diffusion (3) is α'-uniformly dissipative with $\alpha' = d$ and the Euler discretization given in (10) has local deviation with order $(p_1, p_2) = (1, 3/2)$ and $(\lambda_1, \lambda_2) = (\Theta(d^5), \Theta(d^4))$. The detailed calculation for deriving the constants above is provided in Appendix B. Hence, by Theorem 1 in Li et al. (2019), to reach an ϵ-accuracy in Wasserstein-2 distance, the iteration complexity is of order $\tilde{O}(d^3/\epsilon^2)$. Hence, in comparison with the result in Li et al. (2019), we obtain a dimension-free iteration complexity to ensure an ϵ-accuracy in Wasserstein-2 distance.

5.1.2 zeroth order algorithm

According to Theorem 16 and (27), to obtain ε-accuracy in Wasserstein-2 distance, the iteration complexity is of order $\tilde{O}\left((1 \vee d/m)/\varepsilon^2\right)$. When $m = 1$, the iteration complexity

$K \sim \tilde{O}(d/\varepsilon^2)$ and the number of functions evaluations mK is also of the same order $\tilde{O}(d/\varepsilon^2)$. If we choose the batch size $m = d$, we get a dimension independent iteration complexity $K \sim \tilde{O}(1/\varepsilon^2)$ but the number of function evaluations is of order $\tilde{O}(d/\varepsilon^2)$. Hence, we notice that in the case of multivariate t-distribution distributions with large degrees of freedom, the cost of estimating the gradient has an effect on the sampling complexities.

5.2 Multivariate t-distribution: Small Degrees of Freedom

We now consider the isotropic multivariate t-distribution with the degrees of freedom being 3. We denote the corresponding density function by π_β. The exact number of 3 is chosen just for convenience; the results of this example apply to all cases where the degrees of freedom is *strictly* larger than 2 which corresponds to the setting where the variance is finite. We choose $V(x) = 1 + |x|^2$, $\beta = (d+3)/2$ and $\pi_\beta(x) \propto V(x)^{-\beta} = (1+|x|^2)^{-(d+3)/2}$. With the above choice of V and β, V satisfies Assumption 2 with $\alpha = 2$, $C_V = 2$ and V is L-Lipschitz gradient with $L = 2$. Hence, the constant δ in Theorem 5 is given by $\delta = 1/d$. According to Proposition 8, $\mathbb{E}_{\pi_\beta}[V(X)] = d + 1$ and $\mathbb{E}_{\pi_\beta}[|\nabla V(X)|^2] = 4d$.

5.2.1 first order algorithm

According to Theorem 5 and (13), to obtain ϵ-accuracy in Wasserstein-2 distance, the iteration complexity is of order $\tilde{O}(d^4/\epsilon^2)$. With the same choice of V and β, we check the conditions of Theorem 1 in Li et al. (2019). The diffusion (3) is α'-uniformly dissipative with $\alpha' = 1$ and the Euler discretization given in (10) has local deviation with order $(p_1, p_2) = (1, 3/2)$ and $(\lambda_1, \lambda_2) = (\Theta(d^5), \Theta(d^4))$. The detailed calculation for deriving the constants is provided in Appendix B. Hence, according to Theorem 1 in Li et al. (2019), to reach an ϵ-accuracy in Wasserstein-2 distance, the iteration complexity is of order $\tilde{O}(d^6/\epsilon^2)$. Even in this extremely heavy-tail case (i.e., only the variance exists), to ensure an ϵ-accuracy in Wasserstein-2 distance, we can obtain an iteration complexity with polynomial dimension dependence. Furthermore, in comparison to Li et al. (2019), our analysis helps to decrease the dimension exponent by a factor of 2.

5.2.2 zeroth order algorithm

According to Theorem 16 and (27), to obtain ε-accuracy in Wasserstein-2 distance, the iteration complexity is of order $\tilde{O}\left(\max\{d^4/\varepsilon^2,\ d^{\frac{5}{2}}/\varepsilon,\ d^4/\varepsilon^2 m\}\right)$. Hence, we have that for any batch size m, the iteration complexity $K = \tilde{O}(d^4/\varepsilon^2)$. Picking $m = 1$, the number of function evaluations are of the same order, i.e., $mK = \tilde{O}(d^4/\varepsilon^2)$.

Remark 18 *The example discussed in Section 5.2.2 highlights the following important observation: Choosing a large batch size does not improve the iteration complexity. To explain this, we understand both* (10) *and* (22) *as approximation to the continuous dynamics* (3)*. For the first-order algorithm, the error of the approximation only comes from the Euler-Maruyama discretization. For the zeroth-order algorithm, the error of the approximation comes from both the Euler-Maruyama discretization and the zeroth-order gradient estimate=. When the error from the Euler-Maruyama discretization dominates, the optimal batch size is always* 1 *and the oracle complexity of the zeroth order algorithm is the same as the iteration complexity for the first-order algorithm. When the error from the zeroth-order*

gradient estimate dominates, we need to choose a large batch size depending on d so that the iteration complexity for the zeroth-order algorithm is the same as the iteration complexity for the first-order algorithm while the zeroth-order oracle complexity is of order m-times larger.

6. Further Results and Additional Insights on Assumptions

In Section 2, we provide sufficient conditions on V such that when $\beta > d$, $\pi_\beta \propto V^{-\beta}$ satisfies the weighted Poincaré inequality with weight V. In this section, we relax the conditions in Section 2 by introducing the following assumptions.

Assumption 3 *The function $V : \mathbb{R}^d \to (0, \infty)$ is twice continuously differentiable and V satisfies*

(1) $\nabla^2 V(x)$ is invertible for all $x \in \mathbb{R}^d$.
(2) There exists $\gamma \in \left(0, \frac{\beta}{d+2}\right]$, such that

$$\sup_{x\in\mathbb{R}^d} \left\| V(x)^{\gamma-1} \left(\nabla^2 V_\gamma\right)^{-1}(x) \right\|_2 \le C_V(\gamma),$$

where $V_\gamma := V^\gamma$ and $C_V(\gamma)$ is a positive constant depending on γ.

Lemma 19 *Under Assumption 3, for any smooth function $\phi \in L^2(\pi_\beta)$,*

$$Var_{\pi_\beta}(\phi) \le C_{WPI} \int_{\mathbb{R}^d} |\nabla\phi(x)|^2 V(x)\pi_\beta(x)dx, \quad \text{with} \quad C_{WPI} = C_V(\gamma)\left(\frac{\beta}{\gamma} - 1\right)^{-1}. \tag{28}$$

Proof First we define $V_\gamma := V^\gamma$. Choose $\beta' = \beta - 2\gamma$. For $\pi_{\beta'} \propto V^{-\beta'}$, we can write it as $\pi_{\beta'} \propto {V_\gamma}^{-a}$ with

$$a = \frac{\beta'}{\gamma} = \frac{\beta - 2\gamma}{\gamma} \ge d,$$

where the inequality follows from the fact that $\gamma \in \left(0, \frac{\beta}{d+2}\right]$. Therefore we can apply Theorem 1 to $\pi_{\beta'} \propto {V_\gamma}^{-a}$ and get for any smooth, $\pi_{\beta'}$-square integrable function g with $\mathbb{E}_{\pi_{\beta'}}[g(X)] = 0$ and $G = V_\gamma g$,

$$(a+1)\int_{\mathbb{R}^d} g(x)^2\pi_{\beta'}(x)dx \le \int_{\mathbb{R}^d} \frac{\langle(\nabla^2 V_\gamma)^{-1}(x)\nabla G(x), \nabla G(x)\rangle}{V_\gamma(x)}\pi_{\beta'}(x)dx. \tag{29}$$

Since $\beta' = \beta - 2\gamma$, (29) is equivalent to

$$(a+1)\int_{\mathbb{R}^d} \frac{|G(x)|^2}{V(x)} V(x)^{-(\beta-1)}dx \le \int_{\mathbb{R}^d} \langle(\nabla^2 V_\gamma)^{-1}(x)\nabla G(x), \nabla G(x)\rangle V(x)^{-(\beta'+\gamma)}dx. \tag{30}$$

Under Assumption 3, we have

$$\int_{\mathbb{R}^d} \langle(\nabla^2 V_\gamma)^{-1}(x)\nabla G(x), \nabla G(x)\rangle V(x)^{-(\beta'+\gamma)}dx$$

$$\leq C_V(\gamma) \int_{\mathbb{R}^d} |\nabla G(x)|^2 V(x)^{1-\gamma} V(x)^{-(\beta'+\gamma)} dx$$
$$= C_V(\gamma) \int_{\mathbb{R}^d} |\nabla G(x)|^2 V(x)^{-(\beta-1)} dx,$$

where the last identity follows from the fact that $\beta' = \beta - 2\gamma$. Along with (30), we get

$$(a+1) \int_{\mathbb{R}^d} \frac{|G(x)|^2}{V(x)} V(x)^{-(\beta-1)} dx \leq C_V(\gamma) \int_{\mathbb{R}^d} |\nabla G(x)|^2 V(x)^{-(\beta-1)} dx. \tag{31}$$

Since $G = V^\gamma g$, G is smooth, π_β-square integrable and $\mathbb{E}_{\pi_{\beta-\gamma}}[G(X)] = 0$. For any π_β-square integrable ϕ, let $G = \phi - \mathbb{E}_{\pi_{\beta-\gamma}}[\phi(X)]$ and we get

$$\int_{\mathbb{R}^d} |\phi(x) - \mathbb{E}_{\pi_{\beta-\gamma}}[\phi(X)]|^2 \pi_\beta(x) dx \leq \frac{C_V(\gamma)}{a+1} \int_{\mathbb{R}^d} |\nabla \phi(x)|^2 V(x) \pi_\beta(x) dx. \tag{32}$$

Therefore for any smooth, π_β-square integrable ϕ,

$$Var_{\pi_\beta}(\phi) = \inf_{c \in \mathbb{R}} \int_{\mathbb{R}^d} |\phi(x) - c|^2 \pi_\beta(x) dx \leq \frac{C_V(\gamma)}{a+1} \int_{\mathbb{R}^d} |\nabla \phi(x)|^2 V(x) \pi_\beta(x) dx,$$

which is equivalent to (28) with $C_{\mathrm{WPI}} = \frac{C_V(\gamma)}{a+1} = C_V(\gamma) \left(\frac{\beta}{\gamma} - 1\right)^{-1}$. ■

Remark 20 *Lemma 19 can be applied to the class of multivariate t-distributions with* $V(x) = 1 + |x|^2$. *When* $\beta \in \left(\frac{d+2}{2}, d\right]$, *with the choice of* $\gamma = \frac{\beta}{d+2}$, *Assumption 3 holds with*

$$C_V(\gamma) = \frac{(d+2)^2}{2\beta(2\beta - d - 2)}.$$

Hence, Lemma 19 implies that the multivariate t-distribution with degree of freedom $\nu \in (2, d]$ *satisfies the weighted Poincaré inequality with weight* $1 + |x|^2$ *and with*

$$C_{WPI} = \frac{(d+2)^2}{\nu(d+1)(d+\nu)}.$$

The detailed calculation for deriving the above mentioned constants is provided in Appendix C.

As an immediate consequence of Lemma 19, we have the following χ^2 convergence result for (3).

Proposition 21 *Under Assumption 3, with* (X_t) *satisfying* (3) *with* ρ_t *being the distribution of* X_t, *we have*

$$\chi^2(\rho_t|\pi_\beta) \leq \exp\left(-C_V(\gamma)^{-1} \left(\frac{\beta}{\gamma} - 1\right) t\right) \chi^2(\rho_0|\pi_\beta). \tag{33}$$

For the case of multivariate t-distributions, Proposition 21 allows us to show exponential convergence of (3) in the χ^2 divergence with smaller degrees of freedom (and hence heavier tails) compared to Proposition 4.

6.1 Relationship between Lemma 3 and Lemma 19

The result in Lemma 19 complements that in Lemma 3. It can be used to study the WPI for π_β when $\beta \leq d$. In particular, when $\beta \leq d$, if $\pi_\beta \propto V^{-\beta}$ and V satisfies Assumption 2 with $C_V \in (0, \frac{d+2}{d+2-\beta})$, then V satisfies Assumption 3. Therefore π_β satisfies the WPI. In Proposition 22, this relation is proved formally.

Proposition 22 *When $\beta \leq d$, if Assumption 2 holds with $C_V \in (0, \frac{d+2}{d+2-\beta})$, then Assumption 3 holds.*

Proof First $\nabla^2 V$ is invertible because $\nabla^2 V \succeq \alpha I_d$. Next we show that there exists $\gamma \in (0, \frac{\beta}{d+2}]$ such that $\left\|V(x)^{\gamma-1}(\nabla^2 V_\gamma)^{-1}(x)\right\|_2 \leq C_V(\gamma)$ for all $x \in \mathbb{R}^d$. It is equivalent to showing that there exists $\gamma \in (0, \frac{\beta}{d+2}]$ such that $\left\|V(x)^{1-\gamma}(\nabla^2 V_\gamma)(x)\right\|_2 > 0$ for all $x \in \mathbb{R}^d$. From the calculations in Section C, we have

$$\nabla^2 V_\gamma(x) = \gamma V(x)^{\gamma-1}\left((\gamma-1)V(x)^{-1}\nabla V(x)^T \nabla V(x) + \nabla^2 V(x)\right).$$

Therefore

$$\begin{aligned} V(x)^{1-\gamma}(\nabla^2 V_\gamma)(x) &= \gamma\left(\nabla^2 V(x) - (1-\gamma)V(x)^{-1}\nabla V(x)^T \nabla V(x)\right) \\ &\succeq \alpha\gamma\left(1-(1-\gamma)C_V\right) I_d, \end{aligned}$$

where the inequality follows from Assumption 2. Last we show that there exists $\gamma \in (0, \frac{\beta}{d+2}]$ such that $1-(1-\gamma)C_V > 0$. Note that

$$1-(1-\gamma)C_V > 0 \implies \gamma > 1 - \frac{1}{C_V}.$$

Since $C_V \in \left(0, \frac{d+2}{d+2-\beta}\right)$, we have that

$$1 - \frac{1}{C_V} < \frac{\beta}{d+2}$$

Therefore there exists a constant $\gamma \in \left(0, \frac{\beta}{d+2}\right]$ such that $\left\|V(x)^{1-\gamma}(\nabla^2 V_\gamma)(x)\right\|_2 > 0$ for all $x \in \mathbb{R}^d$. ■

6.2 Relation between assumptions in Theorem 5 and Proposition 21

Proposition 21 studies the convergence of the continuous dynamics (3) while Theorem 5 studies the convergence of the discretization (10). The assumptions in Theorem 5 can be shown to imply assumptions in Proposition 21. In Proposition 21 we only assume Assumption 3. In Theorem 5, we assume (i) Assumption 2, (ii) $\delta = \frac{\beta-1-\frac{1}{4}C_V d}{\frac{1}{4}C_V d} > 0$, and (iii) V is gradient Lipschitz. In the following proposition, we show that these three assumptions together imply Assumption 3.

Proposition 23 *If Assumption 2 holds such that* $\delta = \frac{\beta - 1 - \frac{1}{4}C_V d}{\frac{1}{4}C_V d} > 0$ *and* V *is* L*-gradient Lipschitz, then Assumption 3 holds.*

Proof [Proof of Proposition 23] Under Assumption 2 and L-gradient Lipschitzness assumption, we have that V is 'essential quadratic'. That is, assuming V attains its global minimum at x^*, for all $x \in \mathbb{R}^d$,

$$V(x^*) + \frac{\alpha}{2}|x - x^*|^2 \leq V(x) \leq V(x^*) + \frac{L}{2}|x - x^*|^2.$$

Therefore for all $x \in \mathbb{R}^d$,

$$\frac{|\nabla V(x)|^2}{V(x)} \leq \frac{L^2|x - x^*|^2}{V(x^*) + \frac{\alpha}{2}|x - x^*|^2} \leq \frac{2L^2}{\alpha},$$

which implies that Assumption 2-(2) is satisfied with $C_V = \frac{2L^2}{\alpha^2}$. Furthermore,

$$V(x)^{1-\gamma}(\nabla^2 V_\gamma)(x) \succeq \alpha\gamma\left(1 - (1-\gamma)C_V\right) I_d = \alpha\gamma\left(1 - 2(1-\gamma)\frac{L^2}{\alpha^2}\right) I_d.$$

The condition $\delta = \frac{\beta - 1 - \frac{1}{4}C_V d}{\frac{1}{4}C_V d} > 0$ is equivalent to the condition $\beta > \frac{L^2}{2\alpha^2}d + 1$. Notice that for all $d \geq 1$, we have

$$\left(1 - \frac{\alpha^2}{2L^2}\right)(d+2) < \frac{L^2}{2\alpha^2}d + 1$$

Therefore for any

$$\beta > \frac{L^2}{2\alpha^2}d + 1 > \left(1 - \frac{\alpha^2}{2L^2}\right)(d+2),$$

we can choose $\gamma = \frac{\beta}{d+2}$ and obtain

$$\begin{aligned} V(x)^{1-\gamma}(\nabla^2 V_\gamma)(x) &\succeq \frac{2L^2\beta}{\alpha(d+2)}\left(\frac{\alpha^2}{2L^2} + \frac{\beta}{d+2} - 1\right) I_d \\ &= \frac{2L^2\beta}{\alpha(d+2)^2}\left(\beta - \left(1 - \frac{\alpha^2}{2L^2}\right)(d+2)\right) I_d \end{aligned}$$

Therefore Assumption 3-(2) is satisfied with $\gamma = \beta/(d+2)$ and

$$C_V(\gamma) = \frac{\alpha(d+2)^2}{2L^2\beta}\left(\beta - \left(1 - \frac{\alpha^2}{2L^2}\right)(d+2)\right)^{-1} > 0.$$

The proof is now complete because Assumption 3-(1) is automatically satisfied under Assumption 2.

■

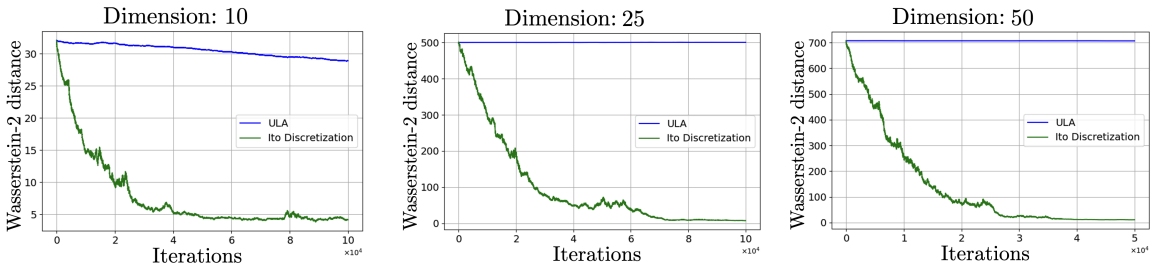

Figure 2: Decay of the Wasserstein-2 distance versus iterations for sampling from Multivariate Student-t targets with 4 degrees of freedom.

7. Numerical Experiments

We now provide simulation results comparing the performance of the ULA and the Itô discretization in (10). We use the *Python Optimal Transport* package (Flamary et al., 2021) for the experiments.

In our first set of experiments, we use the algorithms to sample from multivariate student-t targets with 4 degrees of freedom and the scaling matrix being the identity matrix. In Figure 2, we examine the cases of dimension being 10, 25 and 50 respectively from left to right. The initializations are set to $(10, \cdots, 10)$ for the case of 10 dimensions and to $(100, \cdots, 100)$ for the other two cases. For the case of 10 and 25 dimensions, both algorithms are run for 100000 iterations in parallel for 100 times, with step-size set to 0.0001. For the case of 50 dimensions, the algorithm is run for 50000 iterations in parallel for 100 times, with step-size set to 0.0002. We plot the Wasserstein-2 distance along the trajectories of ULA (blue curve) and the Itô discretization (green curve) for these high-dimensional student-t targets. The Wasserstein-2 distance is approximated by the empirical Wasserstein-2 distances between the generated samples and the target samples, which are numerically computed using the Sinkhorn algorithm. *We emphasize here that the restriction on the dimension is not a consequence of our algorithm (or ULA). It is due to the statistical and computational inefficiency in computing the Wasserstein distance (Weed and Bach, 2019).* The plots in Figure 2 show that the Itô discretization is efficient in higher dimensions while ULA is not. Specifically, for the case of 25 and 50 dimensions, the Wasserstein-2 distance decays fast along the Itô trajectory, while it almost does not decay along the ULA trajectory.

In our second set of experiments, we provide further details for experiments introduced in Section 1, i.e., sampling from a two-dimensional multivariate student-t target with 4 degrees of freedom and the scaling matrix being the identity matrix. In this case, we compute the *exact* Wasserstein-2 distance, instead of the approximate distance used in the previous set of experiments. Both algorithms are run for 10000 iterations in parallel for 500 times, with step-size set to 0.001. The first row, second row and third row are with different initializations, $(10, 10)$, $(-16, 1)$ and $(6, -6)$ respectively. In the left two columns of Figure 3, we plot the averaged first two coordinates' trajectories along ULA (blue curve) and the Itô discretization (green curve). The first column and the second column plot the first coordinate and the second coordinate respectively. The shaded regions characterize the standard errors over the 500 runs. In the third column of Figure 3, we plot the Wasserstein-2 distance along the trajectories of ULA (blue curve) and the Itô discretization (green curve).

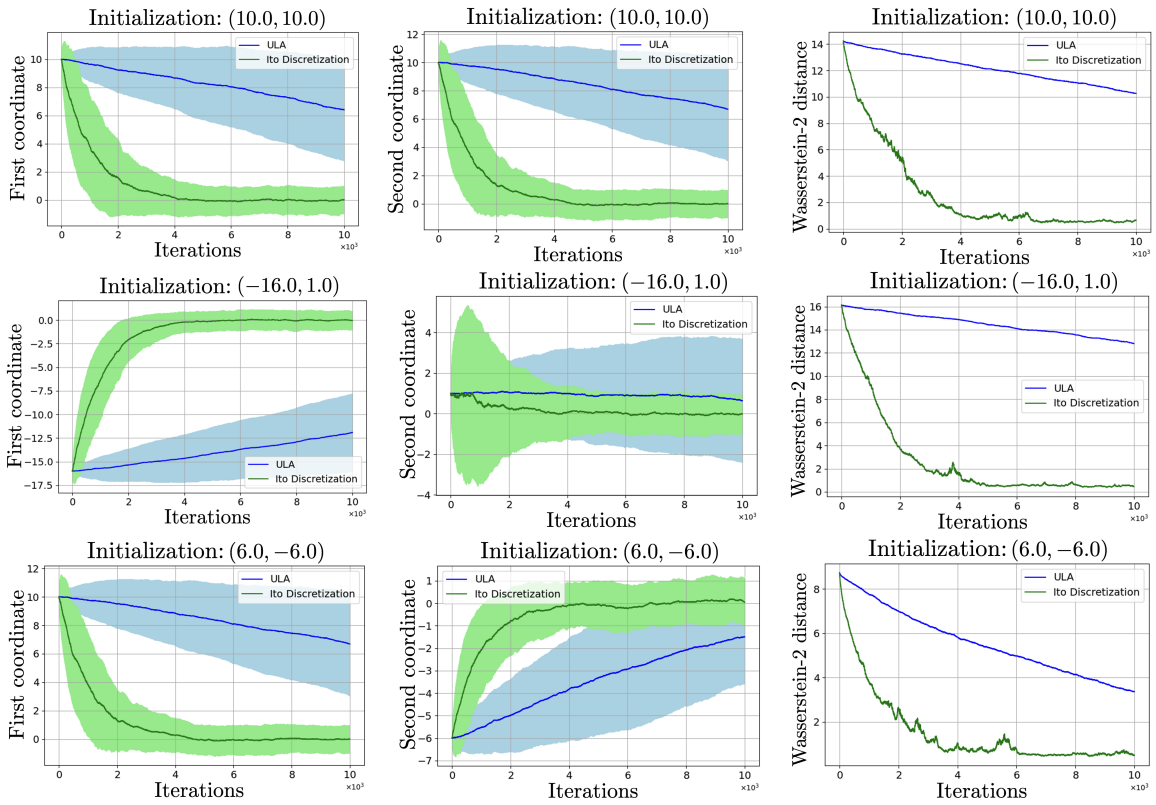

Figure 3: Convergence of the coordinate-wise trajectories and the decay of the Wasserstein-2 distance for various initializations when sampling from a two-dimensional Student-t distribution.

From the left two columns of Figure 3, we conclude that, with different initializations, the averaged first two coordinates along the Itô discretization trajectories always converge to the true first two coordinates' mean faster than those along the ULA trajectories. The Itô discretization samples also have smaller standard deviations. The third column in Figure 3 indicates directly that the Itô discretization outperforms ULA in terms of the Wasserstein-2 distance decay for the same experimental setup.

8. Proofs of the Main Results

8.1 Proofs of Theorem 5 and Theorem 16

In this section, we provide the proof of Theorem 5 and Theorem 16 via mean square analysis. We first start with the following intermediate result.

Proposition 24 *Let $(X_t)_{t\geq 0}$ follow (3) with $X_t \sim \rho_t$ for all $t \geq 0$. If V is gradient Lipschitz with parameter L, then we have*

$$\mathbb{E}\left[|X_t - X_0|^2\right] \leq 4\left[(\beta-1)^2 t^2 \mathbb{E}\left[|\nabla V(X_0)|^2\right] + td\mathbb{E}\left[V(X_0)\right]\right] \exp\left(4(\beta-1)^2 L^2 t^2 + d(\beta-1)Lt^2 + 2dLt\right). \tag{34}$$

Proof [Proof of Proposition 24] According to (3), we have

$$\mathbb{E}[|X_t - X_0|^2] \le 2(\beta-1)^2\mathbb{E}\left[\left|\int_0^t \nabla V(X_s)ds\right|^2\right] + 4d\mathbb{E}\left[\int_0^t V(X_s)ds\right],$$

where

$$\begin{aligned}
\mathbb{E}\left[\left|\int_0^t \nabla V(X_s)ds\right|^2\right] &\le 2\mathbb{E}\left[\left(\int_0^t |\nabla V(X_s) - \nabla V(X_0)|ds\right)^2\right] + 2\mathbb{E}\left[\left(\int_0^t |\nabla V(X_0)|ds\right)^2\right] \\
&\le 2t\mathbb{E}\left[\int_0^t |\nabla V(X_s) - \nabla V(X_0)|^2 ds\right] + 2t\mathbb{E}\left[\int_0^t |\nabla V(X_0)|^2 ds\right] \\
&\le 2L^2 t\int_0^t \mathbb{E}\left[|X_s - X_0|^2\right] ds + 2t^2\mathbb{E}\left[|\nabla V(X_0)|^2\right], \qquad (35)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\int_0^t V(X_s)ds\right] \\
&\le \mathbb{E}\left[\int_0^t V(X_0) + \langle \nabla V(X_0), X_s - X_0\rangle + \frac{L}{2}|X_s - X_0|^2 ds\right] \\
&= t\mathbb{E}\left[V(X_0)\right] + \frac{L}{2}\mathbb{E}\left[\int_0^t |X_s - X_0|^2 ds\right] - (\beta-1)\mathbb{E}\left[\int_0^t\int_0^s \langle \nabla V(X_0), \nabla V(X_u)\rangle du ds\right] \\
&= t\mathbb{E}\left[V(X_0)\right] + \frac{L}{2}\mathbb{E}\left[\int_0^t |X_s - X_0|^2 ds\right] - \frac{(\beta-1)t^2}{2}\mathbb{E}\left[|\nabla V(X_0)|^2\right] \\
&\quad - (\beta-1)\mathbb{E}\left[\int_0^t\int_0^s \langle \nabla V(X_0), \nabla V(X_u) - \nabla V(X_0)\rangle du ds\right] \\
&\le t\mathbb{E}\left[V(X_0)\right] + \frac{L}{2}\mathbb{E}\left[\int_0^t |X_s - X_0|^2 ds\right] - \frac{(\beta-1)t^2}{2}\mathbb{E}\left[|\nabla V(X_0)|^2\right] \\
&\quad + \frac{(\beta-1)t^2}{2}\mathbb{E}\left[|\nabla V(X_0)|^2\right] + \frac{\beta-1}{4}\mathbb{E}\left[\int_0^t\int_0^s |\nabla V(X_u) - \nabla V(X_0)|^2 du ds\right] \\
&\le t\mathbb{E}\left[V(X_0)\right] + \frac{L}{2}\mathbb{E}\left[\int_0^t |X_s - X_0|^2 ds\right] + \frac{(\beta-1)L^2}{4}\mathbb{E}\left[\int_0^t\int_0^s |X_u - X_0|^2 du ds\right] \\
&\le t\mathbb{E}\left[V(X_0)\right] + \left(\frac{L}{2} + \frac{(\beta-1)L^2 t}{4}\right)\mathbb{E}\left[\int_0^t |X_s - X_0|^2 ds\right]. \qquad (36)
\end{aligned}$$

In the above, the first and the second last inequalities use that V is gradient-Lipschitz with parameter L and the second inequality follows from the Young's inequality, $ab \le 2a^2 + \frac{1}{4}b^2$, for any $a, b \in \mathbb{R}$. With (35) and (36), we get

$$\begin{aligned}
\mathbb{E}[|X_t - X_0|^2] \le \int_0^t &\left[4(\beta-1)^2L^2t + 2dL + d(\beta-1)L^2t\right]\mathbb{E}\left[|X_s - X_0|^2\right] ds + 4dt\mathbb{E}\left[V(X_0)\right] \\
&+ 4(\beta-1)^2t^2\mathbb{E}\left[|\nabla V(X_0)|^2\right].
\end{aligned}$$

We define

$$a(t) = 4(\beta-1)^2L^2t + 2dL + d(\beta-1)L^2t,$$

$$b(t) = 4dt\mathbb{E}\left[V(X_0)\right] + 4(\beta-1)^2t^2\mathbb{E}\left[|\nabla V(X_0)|^2\right],$$
$$h(t) = \mathbb{E}[|X_t - X_0|^2],$$

and note that for any $t \geq 0$ we have,

$$h(t) \leq b(t) + a(t)\int_0^t h(s)ds.$$

If we further define $H(t) = \int_0^t h(s)ds$ for any $t \geq 0$, then for any $t \geq 0$,

$$H'(t) \leq b(t) + a(t)H(t),$$

which implies

$$\frac{d}{dt}\left(H(t)\exp(-\int_0^t a(s)ds)\right) \leq b(t)\exp(-\int_0^t a(s)ds).$$

Since $H(0) = 0$, integrate both sides and we get

$$H(t) \leq \int_0^t b(s)\exp(\int_s^t a(u)du)ds.$$

Next since $h(t) \leq b(t) + a(t)H(t)$, we have

$$H(t) \leq b(t) + a(t)\int_0^t b(s)\exp\left(\int_s^t a(t)du\right)ds.$$

Last since a, b are both positive increasing functions, we have

$$\begin{aligned} b(t) + a(t)\int_0^t b(s)\exp\left(\int_s^t a(t)du\right)ds &\leq b(t) + b(t)\int_0^t a(t)\exp\left(\int_s^t a(t)du\right)ds \\ &\leq b(t) + b(t)\int_0^t d\left(-\exp\left(\int_s^t a(t)du\right)\right) \\ &= b(t)\exp(\int_0^t a(t)dt) = b(t)\exp(ta(t)). \end{aligned}$$

Therefore, the proof is completed. ■

Based on the above proposition, we now prove Theorem 5 below.

Proof [Proof of theorem 5] We perform mean square analysis to (10). Let $(X_t)_{t\geq 0}$ follow (3) with $X_0 \sim \pi_\beta$. Since π_β is the unique stationary distribution to (3), $X_t \sim \pi_\beta$ for all $t \geq 0$. With (10), we can calculate the difference between X_h and x_1,

$$\begin{aligned} &X_h - x_1 \\ =&X_0 - \int_0^h (\beta-1)\nabla V(X_t)dt + \int_0^t \sqrt{2V(X_t)}dB_t - \left(x_0 - (\beta-1)h\nabla V(x_0) + \sqrt{2hV(x_0)}\xi_1\right) \\ =&(X_0 - x_0) - (\beta-1)h\left(\nabla V(X_0) - \nabla V(x_0)\right) - \int_0^h (\beta-1)\left(\nabla V(X_t) - \nabla V(X_0)\right)dt \end{aligned}$$

$$\int_0^h \left(\sqrt{2V(X_t)} - \sqrt{2V(x_0)}\right) dB_t$$
$$:= U_1 + U_2 + U_3,$$

where

$$U_1 := (X_0 - x_0) - (\beta - 1)h \left(\nabla V(X_0) - \nabla V(x_0)\right), \tag{37}$$
$$U_2 := -\int_0^h (\beta - 1) \left(\nabla V(X_t) - \nabla V(X_0)\right) dt, \tag{38}$$
$$U_3 := \int_0^h \left(\sqrt{2V(X_t)} - \sqrt{2V(x_0)}\right) dB_t. \tag{39}$$

Therefore according to the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\mathbb{E}[|X_h - x_1|^2|]^{\frac{1}{2}} &= \mathbb{E}[|U_1 + U_2 + U_3|^2]^{\frac{1}{2}} \\
&= \left(\mathbb{E}[|U_1 + U_3|^2] + \mathbb{E}[|U_2|^2] + 2\mathbb{E}[(U_1 + U_3)U_2]\right)^{\frac{1}{2}} \\
&\le \left(\mathbb{E}[|U_1 + U_3|^2] + \mathbb{E}[|U_2|^2] + 2\mathbb{E}[|U_1 + U_3|^2]^{\frac{1}{2}}\mathbb{E}[|U_2|^2]^{\frac{1}{2}}\right)^{\frac{1}{2}} \\
&= \mathbb{E}[|U_1 + U_3|^2]^{\frac{1}{2}} + \mathbb{E}[|U_2|^2]^{\frac{1}{2}}.
\end{aligned}$$

Let $\mathcal{F}_0$ be the σ-algebra generated by x_0, X_0, B_0. Since U_1 is adapted to $\mathcal{F}_0$ and $\mathbb{E}[U_3|\mathcal{F}_0] = 0$, we get

$$\begin{aligned}
\mathbb{E}[|U_1 + U_3|^2|\mathcal{F}_0] &= |U_1|^2 + \mathbb{E}[|U_3|^2|\mathcal{F}_0] \\
&= |(X_0 - x_0) - (\beta - 1)h \left(\nabla V(X_0) - \nabla V(x_0)\right)|^2 \\
&\quad + \mathbb{E}\left[\int_0^h \left\|\sqrt{2V(X_t)}I_d - \sqrt{2V(x_0)}I_d\right\|_F^2 dt|\mathcal{F}_0\right].
\end{aligned}$$

Since V is α-strongly convex and L-gradient Lipschitz, $\bar{V}(x) := V(x) - \alpha|x|^2/2$ is convex and $L - \alpha$ gradient-Lipschitz. and it satisfies the following co-coercivity condition,

$$\frac{1}{L - \alpha}|\nabla\bar{V}(x) - \nabla\bar{V}(y)|^2 \le \langle x - y, \bar{V}(x) - \nabla\bar{V}(y)\rangle, \qquad \forall\ x, y \in \mathbb{R}^d.$$

Letting $x = X_0$, $y = x_0$ and rewriting the above condition in terms of V, we have

$$\langle X_0 - x_0, \nabla V(X_0) - \nabla V(x_0)\rangle \ge \frac{\alpha L}{\alpha + L}|X_0 - x_0|^2 + \frac{1}{\alpha + L}|\nabla V(X_0) - \nabla V(x_0)|^2.$$

Therefore when $h \le \frac{2}{(\beta-1)(\alpha+L)}$,

$$\begin{aligned}
&|(X_0 - x_0) - (\beta - 1)h \left(\nabla V(X_0) - \nabla V(x_0)\right)|^2 \\
=&|X_0 - x_0|^2 - 2(\beta - 1)h\langle X_0 - x_0, \nabla V(X_0) - \nabla V(x_0)\rangle + (\beta - 1)^2h^2|\nabla V(X_0) - \nabla V(x_0)|^2 \\
\le& \left(1 - \frac{2(\beta - 1)\alpha Lh}{\alpha + L}\right)|X_0 - x_0|^2 + (\beta - 1)h\left((\beta - 1)h - \frac{2}{\alpha + L}\right)|\nabla V(X_0) - \nabla V(x_0)|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(1 - \frac{2(\beta-1)\alpha L h}{\alpha + L}\right) |X_0 - x_0|^2 + (\beta-1)h\left((\beta-1)h - \frac{2}{\alpha+L}\right)\alpha^2 |X_0 - x_0|^2 \\
&= \left(1 - (\beta-1)\alpha h\right)^2 |X_0 - x_0|^2, \qquad (40)
\end{aligned}
$$

where the second inequality follows from the fact that $h \leq \frac{2}{(\beta-1)(\alpha+L)}$ and V is α-strongly convex. Meanwhile, for arbitrary $r > 0$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\int_0^h \left\|\sqrt{2V(X_t)} - \sqrt{2V(x_0)}\right\|_F^2 dt\right] \\
=&d\mathbb{E}\left[\int_0^h |\sqrt{2V(X_t)} - \sqrt{2V(x_0)}|^2 dt\right] \\
\leq& d\left(h\left(\sqrt{2V(X_0)} - \sqrt{2V(x_0)}\right)^2 + \mathbb{E}\left[\int_0^h \left|\sqrt{2V(X_t)} - \sqrt{2V(X_0)}\right|^2 dt\right]\right) \\
&+ 2d|\sqrt{2V(X_0)} - \sqrt{2V(x_0)}|h^{\frac{1}{2}}\mathbb{E}\left[\int_0^h \left|\sqrt{2V(X_t)} - \sqrt{2V(X_0)}\right|^2 dt\right]^{\frac{1}{2}} \\
\leq& d(1+r)h\left(\sqrt{2V(X_0)} - \sqrt{2V(x_0)}\right)^2 + d(1+r^{-1})\mathbb{E}\left[\int_0^h \left|\sqrt{2V(X_t)} - \sqrt{2V(X_0)}\right|^2 dt\right],
\end{aligned}
$$

where the first inequality is a result of Holder's inequality and the last inequality follows from Young's inequality. Notice that under Assumption 2, we have

$$
|\nabla(\sqrt{2V(x)})| = \frac{\sqrt{2}|\nabla V(x)|}{2\sqrt{V(x)}} \leq \frac{\sqrt{2\alpha C_V}}{2},
$$

for all $x \in \mathbb{R}^d$. Therefore

$$
(\sqrt{2V(X_0)} - \sqrt{2V(x_0)})^2 \leq \frac{\alpha C_V}{2}|X_0 - x_0|^2, \qquad (41)
$$

and

$$
\int_0^h |\sqrt{2V(X_t)} - \sqrt{2V(X_0)}|^2 dt \leq \frac{\alpha C_V}{2}\int_0^h |X_t - X_0|^2 dt. \qquad (42)
$$

With (41) and (42), we get

$$
\begin{aligned}
\mathbb{E}[\int_0^h \left\|\sqrt{2V(X_t)} - \sqrt{2V(x_0)}\right\|_F^2 dt] \leq& \frac{\alpha C_V d h(1+r)}{2}\mathbb{E}[|X_0 - x_0|^2] \\
&+ \frac{\alpha C_V d(1+r^{-1})}{2}\int_0^h \mathbb{E}[|X_t - X_0|^2]dt.
\end{aligned} \qquad (43)
$$

Next we apply Proposition 24 to $\mathbb{E}[|X_t - X_0|^2]$. In particular, when

$$
t \in [0, h] \quad \text{and} \quad h < \frac{1}{4(\beta-1)L},
$$

we have

$$
\mathbb{E}[|X_t - X_0|^2] \leq \left(4dt\mathbb{E}\left[V(X_0)\right] + 4(\beta-1)^2 t^2 \mathbb{E}\left[|\nabla V(X_0)|^2\right]\right)\exp(1)
$$

$$\leq 12dt\mathbb{E}\left[V(X_0)\right] + 12(\beta-1)^2t^2\mathbb{E}\left[|\nabla V(X_0)|^2\right]. \tag{44}$$

Combining (43) and (44), when $h < \frac{1}{4(\beta-1)L}$, we have that

$$\begin{aligned}
&\mathbb{E}[\int_0^h \left\| \sqrt{2V(X_t)} - \sqrt{2V(x_0)} \right\|_F^2 dt] \\
\leq &\frac{1}{2}\alpha C_V d(1+r)h\mathbb{E}[|X_0 - x_0|^2] \qquad (45)\\
&+ 6\alpha C_V d(1+r^{-1}) \int_0^h \left(dt\mathbb{E}\left[V(X_0)\right] + (\beta-1)^2t^2\mathbb{E}\left[|\nabla V(X_0)|^2\right]\right) dt \\
= &\frac{1}{2}\alpha C_V d(1+r)h\mathbb{E}[|X_0 - x_0|^2] \qquad (46)\\
&+ 3\alpha C_V d^2(1+r^{-1})h^2\mathbb{E}\left[V(X_0)\right] + 2\alpha C_V d(\beta-1)^2(1+r^{-1})h^3\mathbb{E}\left[|\nabla V(X_0)|^2\right].
\end{aligned}$$

With (40) and (45), we get

$$\begin{aligned}
&\mathbb{E}[|U_1 + U_3|^2] \\
\leq &\left(1 - 2(\beta-1)\alpha h + (\beta-1)^2\alpha^2h^2 + \frac{1}{2}\alpha C_V d(1+r)h\right)\mathbb{E}\left[|X_0 - x_0|^2\right] \\
&+ 3\alpha C_V d^2(1+r^{-1})h^2\mathbb{E}\left[V(X_0)\right] + 2\alpha C_V d(\beta-1)^2(1+r^{-1})h^3\mathbb{E}\left[|\nabla V(X_0)|^2\right] \\
\leq &\left(1 - 2(\beta-1)\alpha h + (\beta-1)^2\alpha^2h^2 + \frac{1}{2}\alpha C_V d(1+r)h\right)\mathbb{E}\left[|X_0 - x_0|^2\right] \\
&+ 2\alpha C_V d(1+r^{-1})h^2\left(3d\mathbb{E}\left[V(X_0)\right] + 2(\beta-1)^2h\mathbb{E}\left[|\nabla V(X_0)|^2\right]\right). \qquad (47)
\end{aligned}$$

Since $C_V < \frac{4(\beta-1)}{d}$, denote $\delta = \frac{(\beta-1)-\frac{1}{4}C_V d}{\frac{1}{4}C_V d} > 0$. We have

$$\begin{aligned}
&1 - 2(\beta-1)\alpha h + (\beta-1)^2\alpha^2h^2 + \frac{1}{2}\alpha C_V d(1+r)h \\
= &1 - 2(\beta-1)\alpha h + (\beta-1)^2\alpha^2h^2 + 2(\beta-1)\alpha\frac{1+r}{1+\delta}h \\
= &\left[1 - \alpha(\beta-1)(1 - \frac{1+2r}{1+\delta})h\right]^2 + \alpha^2(\beta-1)^2h^2 \\
&- 2\alpha(\beta-1)\frac{r}{1+\delta}h - \alpha^2(\beta-1)^2h^2\left(\frac{\delta-2r}{1+\delta}\right)^2.
\end{aligned}$$

By picking $r = \frac{\delta}{3}$, we get for any $h \in \left(0, \frac{2\delta}{3(1+\delta)\alpha(\beta-1)}\right)$ that

$$\begin{aligned}
&1 - 2(\beta-1)\alpha h + (\beta-1)^2\alpha^2h^2 + \frac{1}{2}\alpha C_V d(1+r)h \\
&\leq \left[1 - \alpha(\beta-1)\frac{\delta}{3(1+\delta)}h\right]^2 + \alpha^2(\beta-1)^2h\left(h - \frac{2\delta}{3(1+\delta)}\alpha^{-1}(\beta-1)^{-1}\right) \\
&\leq \left[1 - \alpha(\beta-1)\frac{\delta}{3(1+\delta)}h\right]^2.
\end{aligned}$$

With the choice of $r = \delta/3$, (47) could be rewritten as

$$\mathbb{E}[|U_1 + U_3|^2] \leq \left(1 - \frac{\alpha(\beta-1)\delta}{3(1+\delta)}h\right)^2 \mathbb{E}[|X_0 - x_0|^2] + \frac{8\alpha(\beta-1)(3+\delta)h^2}{(1+\delta)\delta}\left(3d\mathbb{E}\left[V(X_0)\right] + 2(\beta-1)^2 h\mathbb{E}\left[|\nabla V(X_0)|^2\right]\right). \quad (48)$$

Next, with the bound in (44), we get when $h < \frac{1}{4(\beta-1)L}$,

$$\begin{aligned}\mathbb{E}[|U_2|^2] &\leq (\beta-1)^2 L^2 \mathbb{E}\left[\left(\int_0^h |X_t - X_0| dt\right)^2\right] \\ &\leq (\beta-1)^2 L^2 h \int_0^h \mathbb{E}\left[|X_t - X_0|^2\right] dt \\ &\leq 6d(\beta-1)^2 L^2 h^3 \mathbb{E}\left[V(X_0)\right] + 4(\beta-1)^4 L^2 h^4 \mathbb{E}\left[|\nabla V(X_0)|^2\right]. \end{aligned} \quad (49)$$

With (48) and (49), we get when $h < \min\left(\frac{1}{4(\beta-1)L}, \frac{2\delta}{3(1+\delta)\alpha(\beta-1)}\right)$,

$$\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} \leq \left[(1-A)^2 \mathbb{E}\left[|X_0 - x_0|^2\right] + B^2\right]^{\frac{1}{2}} + C,$$

with

$$A = \frac{\alpha(\beta-1)\delta}{3(1+\delta)}h, \quad (50)$$

$$B = \frac{5\alpha^{\frac{1}{2}}(\beta-1)^{\frac{1}{2}}(3+\delta)^{\frac{1}{2}}h}{(1+\delta)^{\frac{1}{2}}\delta^{\frac{1}{2}}}\left(d^{\frac{1}{2}}\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + (\beta-1)h^{\frac{1}{2}}\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}\right), \quad (51)$$

$$C = 3d^{\frac{1}{2}}(\beta-1)Lh^{\frac{3}{2}}\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + 2(\beta-1)^2 Lh^2 \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}. \quad (52)$$

The above analysis works for each step, therefore we get for all $k \geq 1$,

$$\mathbb{E}\left[|X_{kh} - x_k|^2\right]^{\frac{1}{2}} \leq \left[(1-A)^2 \mathbb{E}\left[|X_{(k-1)h} - x_{k-1}|^2\right] + B^2\right]^{\frac{1}{2}} + C.$$

According to (Dalalyan and Karagulyan, 2019, Lemma 9), with A, B, C given in (50),(51),(52), for all $k \geq 1$,

$$\mathbb{E}\left[|X_{kh} - x_k|^2\right]^{\frac{1}{2}} \leq (1-A)^k \mathbb{E}\left[|X_0 - x_0|^2\right]^{\frac{1}{2}} + \frac{C}{A} + \frac{B}{\sqrt{A(2-A)}}.$$

Choosing X_0 such that $W_2(\nu_0, \pi_\beta) = \mathbb{E}\left[|X_0 - x_0|^2\right]^{\frac{1}{2}}$, we get (11).
■

We now prove Theorem 16.

Proof [Proof of Theorem 16] Following the same strategy and notation in the proof of Theorem 5, we have

$$X_h - x_1 = U_1 + U_2 + U_3 + (\beta-1)h\mathbb{E}[\zeta_0|x_0] + (\beta-1)h\left(\zeta_0 - \mathbb{E}[\zeta_0|x_0]\right), \quad (53)$$

where U_1, U_2, U_3 are defined in (37),(38),(39) respectively and $\zeta_0 = g_{\sigma,m}(x_0) - \nabla V(x_0)$. Therefore we have

$$\begin{aligned}\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} &\leq \mathbb{E}\left[|U_1 + U_3 + (\beta-1)h\left(\zeta_0 - \mathbb{E}[\zeta_0|x_0]\right)|^2\right]^{\frac{1}{2}} \\ &\quad + \mathbb{E}\left[|U_2|^2\right]^{\frac{1}{2}} + (\beta-1)h\mathbb{E}\left[|\mathbb{E}[\zeta_0|x_0]|^2\right]^{\frac{1}{2}} \\ &= \left\{\mathbb{E}\left[|U_1+U_3|^2\right] + (\beta-1)^2h^2\mathbb{E}\left[|\zeta_0 - \mathbb{E}[\zeta_0|x_0]|^2\right]\right\}^{\frac{1}{2}} \\ &\quad + \mathbb{E}\left[|U_2|^2\right]^{\frac{1}{2}} + (\beta-1)h\mathbb{E}\left[|\mathbb{E}[\zeta_0|x_0]|^2\right]^{\frac{1}{2}}.\end{aligned} \tag{54}$$

From the proof of Theorem 5 and Proposition 15, when

$$h < \min\left(\frac{1}{4(\beta-1)h}, \frac{2\delta}{3(1+\delta)\alpha(\beta-1)}\right),$$

we have that

$$\begin{aligned}\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} \leq \Bigg\{&(1-A)^2\mathbb{E}\left[|X_0 - x_0|^2\right] + B^2 + \frac{\sigma^2}{2m}L^2(\beta-1)^2(d+3)^3h^2 \\ &+ \frac{2(d+5)(\beta-1)^2h^2}{m}\mathbb{E}\left[|\nabla V(x_0)|^2\right]\Bigg\}^{\frac{1}{2}} + C + L\sigma(\beta-1)d^{\frac{1}{2}}h,\end{aligned} \tag{55}$$

where A, B, C are defined in (50),(51),(52). Using the fact that V is gradient Lipshcitz, we have

$$\begin{aligned}\mathbb{E}\left[|\nabla V(x_0)|^2\right] &\leq \mathbb{E}\left[\left(|\nabla V(X_0)| + L|X_0 - x_0|\right)^2\right] \\ &\leq 2\mathbb{E}\left[|\nabla V(X_0)|^2\right] + 2L^2\mathbb{E}\left[|X_0 - x_0|^2\right].\end{aligned} \tag{56}$$

Plugging (56) in (55), we get

$$\begin{aligned}\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} \leq \Bigg\{&(1-A)^2\mathbb{E}\left[|X_0 - x_0|^2\right] + \frac{4(d+5)(\beta-1)^2L^2h^2}{m}\mathbb{E}\left[|X_0 - x_0|^2\right] + B^2 \\ &+ \frac{\sigma^2}{2m}L^2(\beta-1)^2(d+3)^3h^2 + \frac{4(d+5)(\beta-1)^2h^2}{m}\mathbb{E}\left[|\nabla V(X_0)|^2\right]\Bigg\}^{\frac{1}{2}} \\ &+ C + L\sigma(\beta-1)d^{\frac{1}{2}}h.\end{aligned} \tag{57}$$

When we pick the step-size such that

$$h < \min\left\{\frac{2(1+\delta)}{\alpha(\beta-1)\delta}, \frac{\alpha m\delta}{24(1+\delta)(\beta-1)(d+5)L^2}\right\},$$

we have

$$(1-A)^2 + \frac{4(d+5)(\beta-1)^2L^2h^2}{m} \leq \left(1 - \frac{A}{2}\right)^2.$$

Therefore we have

$$\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} \leq \left\{(1-A')^2\mathbb{E}\left[|X_0 - x_0|^2\right] + B'^2\right\}^{\frac{1}{2}} + C', \tag{58}$$

where

$$A' = \frac{\alpha(\beta-1)\delta}{6(1+\delta)}h, \tag{59}$$

$$B' = \left(5\frac{\alpha^{\frac{1}{2}}(\beta-1)^{\frac{3}{2}}(3+\delta)^{\frac{1}{2}}h^{\frac{3}{2}}}{(1+\delta)^{\frac{1}{2}}\delta^{\frac{1}{2}}} + \frac{2(\beta-1)(d+5)^{\frac{1}{2}}h}{m^{\frac{1}{2}}}\right)\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}$$
$$+ \frac{5\alpha^{\frac{1}{2}}(\beta-1)^{\frac{1}{2}}d^{\frac{1}{2}}(3+\delta)^{\frac{1}{2}}h}{(1+\delta)^{\frac{1}{2}}\delta^{\frac{1}{2}}}\,\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + \frac{\sigma L(\beta-1)(d+3)^{\frac{3}{2}}}{m^{\frac{1}{2}}}h, \tag{60}$$

$$C' = 3L(\beta-1)d^{\frac{1}{2}}h^{\frac{3}{2}}\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + 2L(\beta-1)^2h^2\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}} + \sigma L(\beta-1)d^{\frac{1}{2}}h. \tag{61}$$

The rest of the proof is the same as the proof of Theorem 5, and hence we get (26). ■

Acknowledgements

Part of the work was done when YH affiliated with UC Davis and was supported in part by NSF TRIPODS grant CCF-1934568. TF was supported by the Engineering and Physical Sciences Research Council (EP/T5178) and by the DeepMind scholarship. KB was supported in part by NSF grant DMS-2053918. MAE was supported by NSERC Grant [2019-06167], Connaught New Researcher Award, CIFAR AI Chairs program, and CIFAR AI Catalyst grant. This work was initiated when YH, KB and MAE visited the Simons Institute for the Theory of Computing as a part of the "Geometric Methods in Optimization and Sampling" program during Fall 2021.

References

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. *Advances in Neural Information Processing Systems*, 34:28405–28418, 2021.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1):5–43, 2003.

Christophe Andrieu, Paul Dobson, and Andi Q. Wang. Subgeometric hypocoercivity for piecewise-deterministic Markov process Monte carlo methods. *Electronic Journal of Probability*, 26, 2021.

Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q Wang. Comparison of markov chains via weak poincaré inequalities with application to pseudo-marginal mcmc. *The Annals of Statistics*, 50(6):3592–3618, 2022.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022.

Maria-Florina F Balcan and Hongyang Zhang. Sample and computationally efficient learning algorithms under s-concave distributions. *Advances in Neural Information Processing Systems*, 30, 2017.

Joris Bierkens, Gareth Roberts, and Pierre-André Zitt. Ergodicity of the zigzag process. *The Annals of Applied Probability*, 29(4):2266–2301, 2019.

Adrien Blanchet, Matteo Bonforte, Jean Dolbeault, Gabriele Grillo, and Juan Luis Vázquez. Asymptotics of the fast diffusion equation via entropy estimates. *Archive for Rational Mechanics and Analysis*, 191(2):347–385, 2009.

Sergey Bobkov and Michel Ledoux. Weighted Poincaré-type inequalities for Cauchy and other convex measures. *The Annals of Probability*, 37(2):403–427, 2009.

Michel Bonnefont, Aldéric Joulin, and Yutao Ma. Spectral gap for spherically symmetric log-concave probability measures, and beyond. *Journal of Functional Analysis*, 270(7): 2456–2482, 2016.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of applied probability*, 30(3):1209–1250, 2020.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

Yu Cao, Jianfeng Lu, and Lihan Wang. Complexity of randomized algorithms for underdamped Langevin dynamics. *Communications in Mathematical Sciences*, 19(7):1827–1853, 2021.

Patrick Cattiaux, Nathael Gozlan, Arnaud Guillin, and Cyril Roberto. Functional inequalities for heavy tailed distributions and application to isoperimetry. *Electronic Journal of Probability*, 15:346–385, 2010.

Patrick Cattiaux, Arnaud Guillin, and Li-Ming Wu. Some remarks on Weighted Logarithmic Sobolev Inequality. *Indiana University Mathematics Journal*, pages 1885–1904, 2011.

Patrick Cattiaux, Arnaud Guillin, Pierre Monmarché, and Chaoen Zhang. Entropic multipliers method for Langevin diffusion and Weighted Log-Sobolev Inequalities. *Journal of Functional Analysis*, 277(11):108288, 2019.

Karthekeyan Chandrasekaran, Amit Deshpande, and Santosh Vempala. Sampling s-concave functions: The limit of convexity based isoperimetry. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 420–433. Springer, 2009.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21: 92–1, 2020.

Zongchen Chen and Santosh S Vempala. Optimal Convergence Rate of Hamiltonian Monte Carlo for Strongly Logconcave Distributions. *Theory of Computing*, 18(1):1–18, 2022.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-langevin diffusions. *Advances in Neural Information Processing Systems*, 33:19573–19585, 2020.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-Adjusted Langevin Algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. In *Proceedings of Thirty Fifth Conference on Learning Theory (COLT)*, pages 1–2, 2022.

Dario Cordero-Erausquin and Nathael Gozlan. Transport proofs of weighted poincaré inequalities for log-concave distributions. *Bernoulli*, 23(1):134–158, 2017.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B*, 79(3):651–676, 2017.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

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

Arnak S Dalalyan, Avetik Karagulyan, and Lionel Riou-Durand. Bounding the error of discretized langevin algorithms for non-strongly log-concave targets. *The Journal of Machine Learning Research*, 23(1):10720–10757, 2022.

George Deligiannidis, Alexandre Bouchard-Côté, and Arnaud Doucet. Exponential ergodicity of the bouncy particle sampler. *The Annals of Statistics*, 47(3):1268–1287, 2019.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory*, pages 1486–1513. PMLR, 2020.

Zhiyan Ding and Qin Li. Langevin Monte Carlo: Random coordinate descent and variance reduction. *J. Mach. Learn. Res.*, 22:205–1, 2021.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the Unadjusted Langevin Algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Alain Durmus, Arnaud Guillin, and Pierre Monmarché. Geometric ergodicity of the bouncy particle sampler. *The Annals of Applied Probability*, 30(5):2069–2098, 2020.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20:1–42, 2019.

Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. In *Conference on Learning Theory*, pages 1776–1822. PMLR, 2021.

Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. *Advances in Neural Information Processing Systems*, 31, 2018.

Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 1473–1521, 2023.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, 2(4):1360–1383, 2008.

Alan Genz and Frank Bretz. *Computation of multivariate normal and t-probabilities*, volume 195. Springer Science & Business Media, 2009.

Alan Genz, Frank Bretz, and Yosef Hochberg. Approximations to multivariate t integrals with application to multiple comparison procedures. In *Recent Developments in Multiple Comparison Procedures*, pages 24–32. Institute of Mathematical Statistics, 2004.

Joyee Ghosh, Yingbo Li, and Robin Mitra. On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383, 2018.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *Conference on Learning Theory*, pages 1948–1989. PMLR, 2022.

Sivakanth Gopi, Yin Tat Lee, Daogao Liu, Ruoqi Shen, and Kevin Tian. Algorithmic aspects of the log-laplace transform and a non-euclidean proximal sampler. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 2399–2439, 2023.

Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. On the Ergodicity, Bias and Asymptotic Normality of Randomized Midpoint Sampling Method. *Advances in Neural Information Processing Systems*, 33, 2020.

Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. An analysis of transformed unadjusted langevin algorithm for heavy-tailed sampling. *IEEE Transactions on Information Theory*, 2023.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin Dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

Lu-Jing Huang, Mateusz B Majka, and Jian Wang. Approximation of heavy-tailed distributions via stable-driven SDEs. *Bernoulli*, 27(3):2040–2068, 2021.

Mourad Ismail and Martin E Muldoon. Inequalities and monotonicity properties for Gamma and q-Gamma functions. In *Approximation and Computation: A Festschrift in Honor of Walter Gautschi*, pages 309–323. Springer, 1994.

Søren Jarner and Gareth Roberts. Convergence of heavy-tailed Monte Carlo Markov Chain algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815, 2007.

Qijia Jiang. Mirror Langevin Monte Carlo: the Case Under Isoperimetry. *Advances in Neural Information Processing Systems*, 34:715–725, 2021.

Leif T Johnson and Charles J Geyer. Variable transformation to obtain geometric ergodicity in the Random-Walk Metropolis algorithm. *The Annals of Statistics*, 40(6):3050–3076, 2012.

Kengo Kamatani. Efficient strategy for the Markov chain Monte Carlo in high-dimension with heavy-tailed target probability distribution. *Bernoulli*, 24(4B):3711–3750, 2018.

Samuel Kotz and Saralees Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter run-times for Metropolized Hamiltonian Monte Carlo. In *Conference on Learning Theory*, pages 2565–2597, 2020.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured Log-Concave sampling with a Restricted Gaussian Oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Springer, 2016.

Mufan Li and Murat A Erdogdu. Riemannian langevin algorithm for solving semidefinite programs. *Bernoulli*, 29(4):3093–3113, 2023.

Ruilin Li, Molei Tao, Santosh S Vempala, and Andre Wibisono. The mirror langevin algorithm converges with vanishing bias. In *International Conference on Algorithmic Learning Theory*, pages 718–742. PMLR, 2022.

Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. *Advances in neural information processing systems*, 32, 2019.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-smooth potentials. In *2022 Winter Simulation Conference (WSC)*, pages 3229–3240. IEEE, 2022.

Jianfeng Lu and Lihan Wang. Complexity of zigzag sampling algorithm for strongly log-concave distributions. *Statistics and Computing*, 32(3):1–12, 2022.

Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942–1992, 2021.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Grigori N Milstein and Michael V Tretyakov. *Stochastic numerics for mathematical physics*, volume 456. Springer, 2004.

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

Alireza Mousavi-Hosseini, Tyler Farghly, Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. Towards a Complete Analysis of Langevin Monte Carlo: Beyond Poincaré Inequality. In *Conference on Learning Theory*, 2023.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Than Huy Nguyen, Umut Şimşekli, and Gaël Richard. Non-asymptotic analysis of Fractional Langevin Monte Carlo for non-convex optimization. In *International Conference on Machine Learning*, pages 4810–4819, 2019.

Christian Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Gareth Roberts and Jeffrey Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Gareth Roberts and Richard Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Peter J Rossky, Jimmie D Doll, and Harold L Friedman. Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

Michael Roth. *On the multivariate t distribution*. Linköping University Electronic Press, 2012.

Abhishek Roy, Lingqing Shen, Krishnakumar Balasubramanian, and Saeed Ghadimi. Stochastic zeroth-order discretizations of Langevin diffusions for Bayesian inference. *Bernoulli*, 28(3):1810–1834, 2022.

Ruoqi Shen and Yin Tat Lee. The Randomized Midpoint Method for log-concave sampling. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2100–2111, 2019.

Umut Şimşekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. Fractional underdamped Langevin dynamics: Retargeting SGD with momentum under heavy-tailed gradient noise. In *International Conference on Machine Learning*, pages 8970–8980, 2020.

Michalis K Titsias and Omiros Papaspiliopoulos. Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80 (4):749–767, 2018.

Santosh Vempala and Andre Wibisono. Rapid convergence of the Unadjusted Langevin Algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Maxime Vono, Daniel Paulin, and Arnaud Doucet. Efficient MCMC Sampling with Dimension-Free Convergence Rate using ADMM-type Splitting. *Journal of Machine Learning Research*, 23:1–69, 2022.

Feng-Yu Wang. *Functional inequalities Markov semigroups and spectral theory*. Elsevier, 2006.

Jun-Kun Wang and Andre Wibisono. Accelerating hamiltonian monte carlo via chebyshev integration time. In *The Eleventh International Conference on Learning Representations*, 2022.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4 A):2620–2648, 2019.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax Mixing Time of the Metropolis-Adjusted Langevin Algorithm for Log-Concave Sampling. *Journal of Machine Learning Research*, 23(270):1–63, 2022.

Jun Yang, Krzysztof Łatuszyński, and Gareth Roberts. Stereographic Markov Chain Monte Carlo. *arXiv preprint arXiv:2205.12112*, 2022.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror Langevin Monte Carlo. In *Conference on Learning Theory*, pages 3814–3841. PMLR, 2020.

Xiaolong Zhang and Xicheng Zhang. Ergodicity of supercritical sdes driven by α-stable processes and heavy-tailed sampling. *Bernoulli*, 29(3):1933–1958, 2023.

Appendix A. Computations for Section 3.3

Lemma 25 *Let $\beta > \frac{d}{2} + 1$. If $V \in \mathcal{C}^2(\mathbb{R}^d)$ is positive, α-strongly convex and L-gradient Lipschitz, we have for any $r \in (0, \beta - \frac{d}{2} - 1)$,*

$$\frac{Z_{\beta-1}}{Z_\beta} \leq \left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}} V(0) \left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})}\right)^{\frac{1}{\beta-\frac{d}{2}-r}}. \tag{62}$$

Proof Since $V(x) \leq V(0) + \frac{L}{2}|x|^2$, we know that for any $r \in (0, \beta - \frac{d}{2} - 1)$, $Z_{\frac{d}{2}+r}$ is finite and $\pi_{\frac{d}{2}+r}$ is a probability measure. Therefore

$$\begin{aligned}
\frac{Z_{\beta-1}}{Z_\beta} &= \frac{\int_{\mathbb{R}^d} V(x)^{-(\beta-1)} dx}{\int_{\mathbb{R}^d} V(x)^{-\beta} dx} \\
&= \frac{Z_{\frac{d}{2}+r} \int_{\mathbb{R}^d} V(x)^{-(\beta-\frac{d}{2}-1-r)} \pi_{\frac{d}{2}+r}(x) dx}{Z_{\frac{d}{2}+r} \int_{\mathbb{R}^d} V(x)^{-(\beta-\frac{d}{2}-r)} \pi_{\frac{d}{2}+r}(x) dx} \\
&\leq \frac{\left(\int_{\mathbb{R}^d} V(x)^{-(\beta-\frac{d}{2}-r)} \pi_{\frac{d}{2}+r}(x) dx\right)^{\frac{\beta-\frac{d}{2}-1-r}{\beta-\frac{d}{2}-r}}}{\int_{\mathbb{R}^d} V(x)^{-(\beta-\frac{d}{2}-r)} \pi_{\frac{d}{2}+r}(x) dx} \\
&= \left(\int_{\mathbb{R}^d} V(x)^{-(\beta-\frac{d}{2}-r)} \pi_{\frac{d}{2}+r}(x) dx\right)^{-\frac{1}{\beta-\frac{d}{2}-r}} \\
&= \left(Z_{\frac{d}{2}+r}\right)^{\frac{1}{\beta-\frac{d}{2}-r}} \left(\int_{\mathbb{R}^d} V(x)^{-\beta} dx\right)^{-\frac{1}{\beta-\frac{d}{2}-r}} \\
&\leq \left(Z_{\frac{d}{2}+r}\right)^{\frac{1}{\beta-\frac{d}{2}-r}} \left(\int_{\mathbb{R}^d} (V(0) + \frac{L}{2}|x|^2)^{-\beta} dx\right)^{-\frac{1}{\beta-\frac{d}{2}-r}}.
\end{aligned}$$

For the integral $\int_{\mathbb{R}^d} (V(0) + \frac{L}{2}|x|^2)^{-\beta} dx$, we can calculate it via change of polar coordinates and substitutions,

$$\int_{\mathbb{R}^d} (V(0) + \frac{L}{2}|x|^2)^{-\beta} dx = A_{d-1}(1) \int_0^\infty (V(0) + \frac{L}{2}R^2)^{-\beta} R^{d-1} dR$$

$$
\begin{aligned}
&= \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_0^\infty (V(0) + V(0)R_L)^{-\beta} (\frac{2V(0)}{L})^{\frac{d}{2}-1} {R_L}^{\frac{d}{2}-1} \frac{2V(0)}{L} dR_L \\
&= \frac{2^{\frac{d}{2}}\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}) L^{\frac{d}{2}} V(0)^{\beta-\frac{d}{2}}} \int_0^\infty (1+R_L)^{-\beta} R_L^{\frac{d}{2}-1} dR_L \\
&= \frac{2^{\frac{d}{2}}\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}) L^{\frac{d}{2}} V(0)^{\beta-\frac{d}{2}}} \int_0^1 u^{\frac{d}{2}-1} (1-u)^{\beta-\frac{d}{2}-1} du \\
&= \frac{2^{\frac{d}{2}}\pi^{\frac{d}{2}} B(\frac{d}{2}, \beta-\frac{d}{2})}{\Gamma(\frac{d}{2}) L^{\frac{d}{2}} V(0)^{\beta-\frac{d}{2}}},
\end{aligned}
$$

where the second identity follows from a substitution with $R_L = LR^2/(2V(0))$ and the fourth identity follows from a substitution with $u = \frac{R_L}{1+R_L}$. For $Z_{\frac{d}{2}+r}$, we have

$$
\begin{aligned}
Z_{\frac{d}{2}+r} &= \int_{\mathbb{R}^d} V(x)^{-\frac{d}{2}-r} dx \\
&\le \int_{\mathbb{R}^d} \left(V(0) + \frac{\alpha}{2}|x|^2 \right)^{-\frac{d}{2}-r} dx \\
&= \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_0^\infty \left(V(0) + \frac{\alpha}{2} R^2 \right)^{-\frac{d}{2}-r} R^{d-1} dR \\
&= \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_0^\infty (V(0) + V(0)R_\alpha)^{-\frac{d}{2}-r} \left(\frac{2V(0)}{\alpha} \right)^{\frac{d}{2}-1} {R_\alpha}^{\frac{d}{2}-1} \frac{2V(0)}{\alpha} dR_\alpha \\
&= \frac{2^{\frac{d}{2}}\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}) \alpha^{\frac{d}{2}} V(0)^r} \int_0^\infty (1+R_\alpha)^{-\frac{d}{2}-r} {R_\alpha}^{\frac{d}{2}-1} dR_\alpha \\
&= \frac{2^{\frac{d}{2}}\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}) \alpha^{\frac{d}{2}}} \int_0^1 u^{\frac{d}{2}-1} (1-u)^{r-1} du \\
&= \frac{2^{\frac{d}{2}}\pi^{\frac{d}{2}} B(\frac{d}{2}, r)}{\Gamma(\frac{d}{2}) \alpha^{\frac{d}{2}} V(0)^r}.
\end{aligned}
$$

Therefore, we can further get

$$
\begin{aligned}
\frac{Z_{\beta-1}}{Z_\beta} &\le \left(\frac{2^{\frac{d}{2}}\pi^{\frac{d}{2}} B(\frac{d}{2}, r)}{\Gamma(\frac{d}{2}) \alpha^{\frac{d}{2}} V(0)^r} \frac{\Gamma(\frac{d}{2}) L^{\frac{d}{2}} V(0)^{\beta-\frac{d}{2}}}{2^{\frac{d}{2}}\pi^{\frac{d}{2}} B(\frac{d}{2}, \beta-\frac{d}{2})} \right)^{\frac{1}{\beta-\frac{d}{2}-r}} \\
&= \left(\frac{L^{\frac{d}{2}} V(0)^{\beta-\frac{d}{2}-r}}{\alpha^{\frac{d}{2}}} \frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})} \right)^{\frac{1}{\beta-\frac{d}{2}-r}} \\
&= \left(\frac{L}{\alpha} \right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}} V(0) \left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})} \right)^{\frac{1}{\beta-\frac{d}{2}-r}}.
\end{aligned}
$$

■

Appendix B. Computations for Sections 5.1 and 5.2

Let $\pi_\beta(x) \propto V(x)^{-\beta} = (1+|x|^2)^{-\beta}$ with $\beta > \frac{d+2}{2}$. The gradient and Hessian of V is

$$\nabla V(x) = 2x, \qquad \nabla^2 V(x) = 2I_d.$$

Therefore V is α-strongly convex with $\alpha = 2$ and L-gradient Lipschitz with $L = 2$. (3) reduces to

$$dX_t = b(x)dt + \sigma(X_t)dB_t, \tag{63}$$

with $b(x) = -2(\beta-1)x$ and $\sigma(x) = \sqrt{2}(1+|x|^2)^{\frac{1}{2}} I_d$.

Next we look at the uniform dissipativity condition:

$$\begin{aligned}
&\langle b(x) - b(y), x-y\rangle + \frac{1}{2}\left\|(1+|x|^2)^{\frac{1}{2}} I_d - (1+|y|^2)^{\frac{1}{2}} I_d\right\|_F^2 \\
=& -2(\beta-1)|x-y|^2 + d|(1+|x|^2)^{\frac{1}{2}} - (1+|y|^2)^{\frac{1}{2}}|^2 \\
\le& -2(\beta - 1 - \frac{d}{2})|x-y|^2,
\end{aligned} \tag{64}$$

where the inequality follows from the fact that $x \mapsto (1+|x|^2)^{\frac{1}{2}}$ is 1-Lipschitz. Therefore diffusion (63) is α'-uniform dissipative with $\alpha' = 2(\beta - 1 - \frac{d}{2})$. In particular, $\alpha' = d$ when $\beta = d+1$ and $\alpha' = 1$ when $\beta = \frac{d+3}{2}$.

Last we look at the local deviation for the Euler discretization to (63). We use the same notations in Li et al. (2019). According to (Li et al., 2019, lemma 29), $p_1 = 1$ and

$$\lambda_1 = 2\left(\mu_1(b)^2 + \mu_1^F(\sigma)^2\right)\left(\pi_{1,2}(b) + \pi_{1,2}^F(\sigma)\right)(1 + \mathbb{E}[|\tilde{X}_0|^2] + 2\pi_{1,2}(b)\alpha'^{-1}).$$

According to (Li et al., 2019, lemma 29), $p_2 = \frac{3}{2}$ and

$$\lambda_2 = \mu_1(b)\left(\pi_{1,2}(b) + \pi_{1,2}^F(\sigma)\right)(1 + \mathbb{E}[|\tilde{X}_0|^2] + 2\pi_{1,2}(b)\alpha'^{-1}),$$

with

$$\begin{aligned}
\mu_1(b) &:= \sup_{x,y\in\mathbb{R}^d, x\neq y} \frac{|b(x)-b(y)|}{|x-y|} = 2(\beta-1), \\
\mu_1^F(\sigma) &:= \sup_{x,y\in\mathbb{R}^d, x\neq y} \frac{\|\sigma(x)-\sigma(y)\|_F}{|x-y|} = \sqrt{2d}, \\
\pi_{1,2}(b) &:= \sup_{x\in\mathbb{R}^d} \frac{|b(x)|^2}{1+|x|^2} = 4(\beta-1)^2, \\
\pi_{1,2}^F(\sigma) &:= \sup_{x\in\mathbb{R}^d} \frac{\|\sigma(x)\|_F^2}{1+|x|^2} = 2d.
\end{aligned}$$

The order of λ_1 and λ_2 in dimension parameter d is given by:

$$\begin{aligned}
\lambda_1 &= \Theta\left(\left((\beta-1)^2 + d\right)\left((\beta-1)^2 + 2d\right)\left(1 + (\beta-1)^2\alpha'^{-1}\right)\right), \\
\lambda_2 &= \Theta\left((\beta-1)\left((\beta-1)^2 + 2d\right)\left(1 + (\beta-1)^2\alpha'^{-1}\right)\right).
\end{aligned}$$

Therefore, we have that

- when $\beta = d+1$, $(\lambda_1, \lambda_2) = (\Theta(d^5), \Theta(d^4))$,
- when $\beta = \frac{d+3}{2}$, $(\lambda_1, \lambda_2) = (\Theta(d^5), \Theta(d^4))$.

Appendix C. Computations for Remark 20

In the example of Cauchy class distributions, $V(x) = 1 + |x|^2$ and $V_\gamma := V^\gamma$. When $\gamma > \frac{1}{2}$,

$$\begin{aligned}
\nabla V_\gamma(x) &= \gamma V(x)^{\gamma-1} \nabla V(x) \\
\nabla^2 V_\gamma(x) &= \gamma(\gamma-1) V(x)^{\gamma-2} \nabla V(x)^T \nabla V(x) + \gamma V(x)^{\gamma-1} \nabla^2 V(x) \\
&= \gamma V(x)^{\gamma-1} \left((\gamma-1) V(x)^{-1} \nabla V(x)^T \nabla V(x) + \nabla^2 V(x) \right).
\end{aligned}$$

Plug in $V(x) = 1 + |x|^2$, we get

$$\begin{aligned}
\nabla V_\gamma(x) &= 2\gamma(1+|x|^2)^{\gamma-1} x \\
\nabla^2 V_\gamma(x) &= 2\gamma(1+|x|^2)^{\gamma-1} \left(I_d + 2(\gamma-1) \frac{|x|^2}{1+|x|^2} \frac{x^T x}{|x|^2} \right) \\
&= 2\gamma(1+|x|^2)^{\gamma-1} \left((I_d - \frac{x^T x}{|x|^2}) + \left(1 - 2(1-\gamma) \frac{|x|^2}{1+|x|^2} \right) \frac{x^T x}{|x|^2} \right),
\end{aligned}$$

and

$$(\nabla^2 V_\gamma)^{-1}(x) = \frac{1}{2\gamma} (1+|x|^2)^{1-\gamma} \left((I_d - \frac{x^T x}{|x|^2}) + \frac{1+|x|^2}{1+(2\gamma-1)|x|^2} \frac{x^T x}{|x|^2} \right).$$

When $\beta \in \left(\frac{d+2}{2}, d\right]$, $\gamma = \frac{\beta}{d+2} \in \left(\frac{1}{2}, 1\right]$,

$$(\nabla^2 V_\gamma)^{-1}(x) \preceq \frac{1}{2\gamma(2\gamma-1)} (1+|x|^2)^{1-\gamma} I_d = \frac{(d+2)^2}{2\beta(2\beta-d-2)} (1+|x|^2)^{1-\gamma} I_d.$$

Therefore Assumption 3 holds with $C_V(\gamma) = \frac{(d+2)^2}{2\beta(2\beta-d-2)}$. For the Cauchy distribution $\pi_\beta \propto (1+|x|^2)^{-\beta} = (1+|x|^2)^{-\frac{d+\nu}{2}}$ with $\beta \in (\frac{d+2}{2}, d]$, i.e. $\nu \in (2, d]$, according to lemma 19, π_β satisfies the weighted Poincaré inequality with weight $1+|x|^2$ with weighted Poincaré constant

$$C_{\mathrm{WPI}} = C_V(\gamma) \left(\frac{\beta}{\gamma} - 1 \right)^{-1} = \frac{(d+2)^2}{2(d+1)\beta(2\beta-d-2)} = \frac{(d+2)^2}{\nu(d+1)(d+\nu)}.$$