Title
ADVANCES IN SAMPLING AND STOCHASTIC OPTIMIZATION

Author
He, Ye

Publication Date
2023

ADVANCES IN SAMPLING AND STOCHASTIC OPTIMIZATION

By

YE HE
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Krishna Balasubramanian, Chair

Miles Lopes

Alexander Soshnikov

Committee in Charge

2023

Contents

Abstract

This dissertation explores various aspects of sampling algorithms and stochastic optimization algorithms. We investigate the efficiency and behavior of various sampling methods for target distributions, particularly those with heavy-tails, through the analysis of different discretization techniques, functional inequalities, and asymptotic limits. In Chapter 2, we study specific diffusion-based sampling algorithms, the randomized midpoint method, for simulating continuous-time Langevin diffusions, establishing its asymptotic normality and providing insights into its behavior. In Chapter 3, we introduce two algorithms to sample heavy-tail targets. In Section 3.2, we study the oracle complexity of sampling from polynomially decaying heavy-tailed target densities using the Transformed Unadjusted Langevin Algorithm (TULA), highlighting connections to functional inequalities. In Section 3.3, by discretizing a class of Itô diffusions associated with weighted Poincaré inequalities, we examine the complexity of sampling from heavy-tailed distributions and provide iteration complexity estimates in terms of the Wasserstein-2 distance. In Chapter 4, we propose the Regularized Stein Variational Gradient Flow, which interpolates between the Stein Variational Gradient Flow and the Wasserstein Gradient Flow, and establish its theoretical properties. We also introduce a particle-based algorithm based the Regularized Stein Variational Gradient Flow and provide preliminary numerical evidence on its improved performance. In Chapter 5, we derive high-dimensional scaling limits and fluctuations for online least-squares Stochastic Gradient Descent (SGD) algorithm by treating the iterates as an interacting particle system, characterizing the limiting mean-square estimation or prediction errors and their fluctuations.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Krishna Balasubramanian, for his unwavering support, guidance, and mentorship throughout my PhD journey. Your expertise, patience, and encouragement have been invaluable in shaping my research and helping me grow as a scholar.

I am grateful to my thesis committee members, Professor Krishna Balasubramanian, Professor Alexander Soshnikov, and Professor Miles Lopes, for their insightful feedback, constructive criticism, and valuable suggestions that have contributed to the improvement of my dissertation.

I would also like to acknowledge the Mathematics Department in University of California, Davis for providing an excellent academic environment, resources, and support throughout my time as a doctoral candidate. A special mention goes to Tina and Diana for their assistance and unwavering kindness.

I gratefully acknowledge the financial support received from NSF TRIPODS grant CCF-1934568, which enabled me to complete my research successfully.

Last but not least, I am eternally grateful to my family for their love, encouragement, and belief in my abilities. To my parents, Kedong He and Yuzhen Ye, thank you for instilling in me the values of hard work, perseverance, and the importance of education.

I dedicate this dissertation to all those who have supported me on this journey, and to those who continue to believe in the power of knowledge and research to make the world a better place.

CHAPTER 1

Introduction

Sampling and optimization problems are two prevalent and intertwined challenges faced in various scientific, engineering, and computational disciplines. Both problem classes deal with the essential tasks of exploring and exploiting complex spaces to make informed decisions. Sampling problems focus on generating representative samples from a given target density, while optimization problems revolve around finding the best possible solution within a specific problem domain. The study of these problems has led to the development of a rich tapestry of methods and algorithms, such as Markov Chain Monte Carlo, linear programming, and evolutionary algorithms. In recent years, the intersection of these fields has given rise to innovative approaches that leverage the strengths of both sampling and optimization techniques, enabling researchers and practitioners to tackle a diverse range of real-world challenges, such as resource allocation, machine learning, and network design. By understanding the underlying principles and strategies, we can continue to refine our methods and develop new insights to drive progress in these critical areas.

Sampling Problem. The problem of sampling from a given target density

$$\pi(x) := \frac{1}{Z} e^{-f(x)}, \qquad \forall x \in \mathbb{R}^d \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is the potential function and $Z = \int_{\mathbb{R}^d} e^{-f(x)} dx$ is an unknown normalization constant. This problem frequently arises in statistics and machine learning with numerous applications to high-dimensional Bayesian inference [WT11, LCCC16, MHB17, DM19], numerical integration [LP02, HLW06], volume computation [Vem10], optimization and learning [RRT17, EMS18, MPM$^+$20], graphical models [KF09], and molecular dynamics [MT13, LM16].

There are two widely-used approaches for sampling:

(1) diffusion-based *randomized* algorithms, which are based on discretizations of certain diffusion processes.

(2) particle-based *deterministic* algorithms, which are discretizations of certain *approximate* gradient flows.

The diffusion-based randomized algorithms are MCMC algorithms and they have received a lot of attention recently. The fundamental idea behind such algorithms is that a continuous-time diffusion with its invariant measure as the target π is approximately simulated via a numerical sampler. The intuition behind the success of these methods is that by appropriately selecting the step-size parameter, the discrete approximation resulting from the numerical sampler tracks the continuous-time diffusion. Thus, rapid convergence properties of the diffusion process (see, for example, [RT96, LS16, Ebe16, EGZ19, LBBG19, DMS19]) is inherited by the discrete algorithm with an invariant measure that is close to that of the diffusion, which is the target π. While a variety of diffusion processes can lead to a rich class of MCMC samplers, algorithms that are based on discretizing Langevin dynamics have been the primary focus of research due to their simplicity, accuracy, and well-understood theoretical guarantees in high-dimensional settings [Dal17a, CB18a, CCAY^{+}18, DMM19, VW19, MCC^{+}19, CB18a, DMP18, DM17, EH20].

The particle-based sampling algorithms use a set of interacting particles to approximate the target distribution. A canonical particle-based sampling algorithm is the Stein Variational Gradient Descent (SVGD) introduced in [LW16]. However, unlike the diffusion-based sampling algorithms, SVGD introduces a complicated particle system and how to characterize the computation complexity of SVGD remains to be an open question.

Optimization Problem. Optimization lies at the heart of many problems in machine learning, operations research, and engineering. In this rapidly evolving field, stochastic optimization has emerged as a powerful and versatile technique for solving complex problems. We provide a gentle yet comprehensive overview of stochastic optimization and one of its most popular methods: stochastic gradient descent (SGD).

Stochastic optimization is a class of optimization algorithms designed to handle uncertainties in the objective function and/or constraints. The word "stochastic" refers to the presence of randomness or probabilistic behavior in these problems. In many real-world scenarios, the objective function and/or constraints are affected by noise or uncertainty, making traditional optimization methods less effective or even inapplicable. Stochastic optimization provides a framework for tackling such problems, enabling efficient solutions even in the presence of uncertainty.

SGD is a particular instance of stochastic optimization that has gained widespread use in machine learning, especially for training large-scale models such as neural networks. The main idea behind SGD is to update the model parameters iteratively based on random samples or "mini-batches" of data, rather than the entire dataset. This enables the algorithm to converge more quickly and reduce the computational burden, which is essential when dealing with massive datasets or high-dimensional models. One of the main challenges in understanding SGD is comprehending its convergence properties. In the case of fixed-dimensional problems, the learning theory and optimization communities have focused on providing non-asymptotic bounds, either in expectation or with high-probability, over the past two decades. However, such bounds often tend to be overly conservative in predicting the actual behaviour of the SGD algorithm on large-scale statistical problems occurring in practice that are invariably based on specific data generating models.

Sampling as Optimization over measures. The viewpoint of sampling as optimization can explain both the diffusion-based sampling algorithms and the particle-based sampling algorithms. The seminal work [JKO98] provides a variational interpretation of the Langevin diffusion as the gradient flow of a Kullback-Leibler (KL) divergence over the Wasserstein space of probability measures. From this optimization viewpoint, the diffusion-based sampling algorithms, the ULA, can be viewed as a discretization of the Wasserstein gradient flow (WGF). While for particle-based sampling algorithm, SVGD, can't be viewed as a discretization of the WGF as all the discretizations of the WGF have to be random. In fact, SVGD is a deterministic discretization of a constant-order approximation to the WGF due to the kernel integral operator.

Inspired by the above observations, in [HBSL22], I propose a regularized version of SVGD that provides a deterministic discretization of ϵ-approximate WGF for any $\epsilon \in (0, 1]$. When $\epsilon = 1$, the regularized SVGD is exactly the SVGD. When $\epsilon \in (0, 1)$, the regularized SVGD admits similar convergence properties as the SVGD. As we decrease ϵ to zero, the regularized SVGD will behave

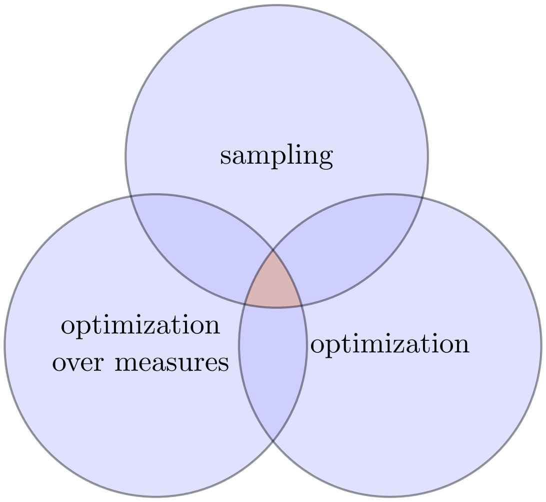

Figure 1.1. Venn Diagram

more and more likely to the discretization to the WGF, i.e. the ULA. I derive the mean-field partial differential equation (PDE) that describes the mean-field limit of the interacting particle system and characterize the uniqueness and existence of the solution to the mean-field PDE. Furthermore, in the population limit of the regularized version of SVGD, I provide rates of convergence to the equilibrium density in two cases: under log-Sobolev inequality (LSI) assumption in the KL divergence metric and under no assumptions in the Fisher information metric.

Organizations. As summarized in the figure 1.1, we explore various aspects of sampling and optimization problems in this dissertation. For sampling problems, we study the computational and statistical properties of existing sampling algorithms, such as the randomized midpoint method [SL19], in Chapter 2. We also develop and analyze new sampling methods, the Transformed Unadjusted Langevin Algorithm (TULA) and the Itô discretization, when the target density is heavy-tailed in Chapter 3. Regarding the optimization problems, we provide a fine-grained analysis on the high-dimensional scaling limits and fluctuations of the online least-squares SGD in Chapter 5. Motivated by the viewpoint of understanding sampling problem as optimization in the space of probability measures, in Chapter 5, we introduce and analyze the Regularized Stein Variational Gradient Flow (RSVGF), which can interpolate between the particle-based sampling algorithm, SVGD, and the diffusion based sampling algorithm, ULA.

CHAPTER 2

Randomized Midpoint Method

We consider the problem of computing the following expectation

$$\mathbb{E}_\pi[\varphi(x)] \quad \text{where} \quad \pi(x) = \tfrac{1}{Z_f} e^{-f(x)}, \tag{2.1}$$

for a potential function $f : \mathbb{R}^d \to \mathbb{R}$ and a test function $\varphi : \mathbb{R}^d \to \mathbb{R}$, when the normalization constant $Z_f = \int e^{-f(x)} dx$ is unknown. This problem frequently arises in statistics and machine learning with numerous applications to high-dimensional Bayesian inference [WT11, LCCC16, MHB17, DM19], numerical integration [LP02, HLW06], volume computation [Vem10], optimization and learning [RRT17, EMS18, MPM$^+$20], graphical models [KF09], and molecular dynamics [MT13, LM16]. Markov chain Monte Carlo (MCMC) methods provide a powerful framework for computing the integral in (2.1), and have been successfully deployed in various scientific fields [Liu08].

In particular, MCMC algorithms that are based on diffusion processes have received a lot of attention recently. The fundamental idea behind such algorithms is that a continuous-time diffusion with its invariant measure as the target π is approximately simulated via a numerical sampler. The intuition behind the success of these methods is that by appropriately selecting the step-size parameter, the discrete approximation resulting from the numerical sampler tracks the continuous-time diffusion. Thus, rapid convergence properties of the diffusion process (see, for example, [RT96, LS16, Ebe16, EGZ19, LBBG19, DMS19]) is inherited by the discrete algorithm with an invariant measure that is close to that of the diffusion, which is the target π. While a variety of diffusion processes can lead to a rich class of MCMC samplers, algorithms that are based on discretizing Langevin dynamics have been the primary focus of research due to their simplicity, accuracy, and well-understood theoretical guarantees in high-dimensional settings [Dal17a, CB18a, CCAY$^+$18, DMM19, VW19, MCC$^+$19, CB18a, DMP18, DM17, EH20].

Although motivated by the problem of computing the integral in (2.1), much of the theoretical focus on analyzing sampling methods in the recent literature has been on providing guarantees for the sampling problem itself (see [TTV16] for an exception), i.e., the number of iterations needed to reach ϵ-neighborhood of a d-dimensional target distribution in some probability metric. The choice of step-size of the sampler is crucial to obtain such theoretical guarantees. While the problem of estimating expectations such as in (2.1) is based on sampling from the target π itself, the theoretical guarantees established for the sampling problem can provide very little to no information on computing the expectation in (2.1) based on the sampler. The main reason for this is, the step-size choice of the sampler required to obtain optimal theoretical guarantees for numerical integration of (2.1) turns out to be different from that of sampling. Furthermore, if the ultimate task is to perform inference on the quantity $\mathbb{E}_\pi[\varphi(x)]$, confidence intervals are required. Thus, one needs central limit theorems (CLT) to quantify the fluctuations of the estimator of the expectation in (2.1), depending on a specific numerical integrator being used.

The randomized midpoint method, a numerical sampler proposed by [SL19], has emerged as an optimal algorithm for sampling from strongly log-concave densities, achieving the information theoretical lower bound for this problem in terms of both dimension and tolerance dependency [CLW20]. In lieu of this optimality result, one anticipates a superior performance from the randomized midpoint method in other fundamental problems that relies on a MCMC sampler as the main computation tool, e.g. estimating expectations of the form (2.1). However, properties of this sampler for the purpose of numerical integration, in particular its inferential properties, are not well-studied. In this chapter, we explore various probabilistic properties of the randomized midpoint discretization method, when used as a numerical integrator. Towards that, we examine several results for the randomized midpoint method considering both the overdamped and underdamped Langevin diffusions. Our first contribution is the explicit characterization of the bias of the randomized midpoint numerical scheme, namely the difference between its stationary distribution and the target distribution π. We show that asymptotic unbiasedness, a desired property in general, can be achieved under a decreasing step-size sequence. As our principal contribution, we establish the ergodicity of the randomized midpoint method and prove a central limit theorem which can be leveraged for inference on the expectation (2.1). We compute the bias and the variance of the asymptotic normal distribution for various step size choices, and show that different step-size sequences are suitable for making inference in different settings.

Our Contributions. We summarize our contributions as follows:

(1) We show the ergodicity of constant step-size (denoted as h) randomized midpoint discretization of the overdamped and underdamped Langevin diffusions in Theorems 1 and 3, respectively. For both cases, the stationary distribution π_h of the resulting discretized Markov chain is unique and is biased away from the target distribution π.

(2) The choice of a constant step-size for the randomized midpoint discretization causes bias in sampling. We characterize this bias explicitly in Propositions 2 and 4 for the overdamped and underdamped Langevin diffusions, respectively. We show that Wasserstein-2 distance between π_h and π is of order $\mathcal{O}(h^{0.5})$ and $\mathcal{O}(h^{1.5})$ respectively for the overdamped and underdamped Langevin diffusions.

(3) The established order of bias points toward using particular choices of decreasing step-size sequence for the sake of inference. Specifically, we prove a CLT for numerical integration using the randomized midpoint discretization of the overdamped and underdamped Langevin diffusions in Theorems 2 and 4 respectively, for various choices of decreasing step-size. Depending on the specific choice of step-size sequence, the CLT is either unbiased or biased. When discretizing the overdamped Langevin diffusion with polynomially decreasing step-size choices, the rate of unbiased CLT turns out to be $\mathcal{O}(n^{(1/3)-\epsilon})$ for any $\epsilon > 0$. But the optimal rate turns out to be $\mathcal{O}(n^{1/3})$ for which one can only obtain a biased CLT. When discretizing underdamped Langevin diffusions with polynomially decreasing step-size choices, we show that the optimal rate can be improved to $\mathcal{O}(n^{5/8})$ under a certain condition, which is satisfied only by the class of constant test functions.

2.1. Notations and Preliminaries

We denote an ℓ-th order symmetric tensor of dimension d by $A \in \mathbb{R}^{d\otimes\ell}$. For a given vector $u \in \mathbb{R}^d$, we use $\|u\|$ to denote the Euclidean-norm of the vector. We define the ℓ-th order rank-1 tensor formed from $u \in \mathbb{R}^d$ as $u^{\otimes\ell}$. In addition, let A and B be two ℓ-th order tensors, we define the inner product between A and B as $\langle A, B\rangle = \sum_{j_1=1}^{d} \cdots \sum_{j_\ell=1}^{d} A_{j_1 j_2 \cdots j_\ell} \cdot B_{j_1 j_2 \cdots j_k}$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, $\nabla f \in \mathbb{R}^d$ and $D^\ell \in \mathbb{R}^{d \otimes \ell}$ represents the gradient, and ℓ-th order derivative tensor (for $\ell > 1$). We let $(\Omega, \mathcal{F}, P)$ represent a probability space, and denote by $\mathcal{B}(\mathbb{R}^d)$, the Borel σ-field of $\mathbb{R}^d$. We use $\overset{d}{\to}$ and $\overset{p}{\to}$ to denote convergence in distribution and probability respectively. The set of all twice continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ is denoted as $\mathcal{C}^2(\mathbb{R}^d)$. We use I_d to represent the $d \times d$ identity matrix. Let $x_0, x_1, \ldots$ be a d-dimensional Markov chain. The transition probability of the chain, at the k-th step is defined as $P^k(x, A) := P(x_k \in A | x_0 = x)$, for some $x \in \mathbb{R}^d$ and represents the probability that the chain is in set A at time n given the starting point was $x \in \mathbb{R}^d$. We use $\tilde{\mathcal{O}}$ to hide log factors. Finally, for a sequence γ_k and positive integer ℓ, we define $\Gamma_n^{(\ell)} := \sum_{k=1}^{n} \gamma_k^\ell$. We also make the following assumption on the potential function.

Assumption 2.1.1. *The potential function $f \in \mathcal{C}^2(\mathbb{R}^d)$ satisfies the following properties. For some $0 < m \leq M < \infty$: (a) f has a M-Lipschitz gradient; that is, $D^2 f \preceq M I_d$, and (b) f is m-strongly convex; that is, $m I_d \preceq D^2 f$. We also define the condition number as $\kappa := M/m$.*

2.2. Results for the Overdamped Langevin Diffusion

The overdamped Langevin diffusion is described by the following stochastic differential equation:

$$dx(t) = -\nabla f(x(t))dt + \sqrt{2}dW(t), \tag{2.2}$$

where $W(t)$ is a d-dimensional Brownian motion. It is well-known that this diffusion has $\pi(x) \propto e^{-f(x)}$ as its stationary distribution under mild regularity conditions [MT12]. In general, simulating a continuous-time diffusion such as (2.2) is impractical; thus, a numerical integration scheme is needed.

We now describe the *randomized midpoint discretization* of the above diffusion in (2.2), which we denote as RLMC. Denoting the n-th iteration of the algorithm with x_n, the integral formulation of the diffusion with x_n as the initial value would then be $x_n^*(t) = x_n - \int_0^t \nabla f(x_n^*(s))ds + \sqrt{2}W(t)$. Let $h > 0$ be the choice of step size for the discretization and, let (α_n) be an i.i.d. sequence of random variables following uniform distribution on $[0, 1]$, i.e. $\alpha_n \sim U[0, 1]$. The fundamental idea behind the randomized midpoint technique is to use $h\nabla f(x_n^*(\alpha_{n+1}h))$ to approximate the integral $\int_0^h \nabla f(x_n^*(s))ds$. Indeed, notice that $\mathbb{E}[h\nabla f(x_n^*(\alpha_{n+1}h))] = h\int_0^1 \nabla f(x_n^*(\alpha h))d\alpha = \int_0^h \nabla f(x_n^*(s))ds$. RLMC proceeds by approximating $x_n^*(\alpha_{n+1}h)$ with the Euler discretization, which ultimately yields an explicit numerical integration step. Although [SL19] considered this discretization only for the constant step-size choice and the *underdamped* Langevin diffusion (which we discuss in Section 2.3), below we present a single iteration of the RLMC algorithm with the choice of variable step-size γ_{n+1} for the overdamped diffusion in (2.2):

$$\text{(RLMC)}\qquad \begin{aligned} x_{n+\frac{1}{2}} &= x_n - \alpha_{n+1}\gamma_{n+1}\nabla f(x_n) + \sqrt{2\alpha_{n+1}\gamma_{n+1}}U'_{n+1}, \\ x_{n+1} &= x_n - \gamma_{n+1}\nabla f(x_{n+\frac{1}{2}}) + \sqrt{2\gamma_{n+1}}U_{n+1}, \end{aligned}$$

where (U_n) and (U'_n) are sequences of i.i.d d-dimensional standard Gaussian vectors with cross-covariance matrix $\sqrt{\alpha_n}I_d$ for each n and the initial point x_0. We briefly digress now to make the following remark. If instead of $\alpha_n \sim U[0,1]$, one uses $\alpha_n = 1$ for all n deterministically, then the iterates of (RLMC) algorithm is reminiscent of the extra-gradient descent algorithm from the optimization literature [LT93], perturbed by Gaussian noise in each step. Furthermore, its noteworthy that with the deterministic choice of $\alpha_n = 1$, one cannot obtain the improved rates that the uniformly random α_n provides. Lastly, the filtration $(\mathcal{F}_n)$ is defined by $\mathcal{F}_n := \sigma(\alpha_k, U_k, U'_k; 1 \leq k \leq n)$, the smallest σ-algebra generated by the noise sequence and uniform random variables that are used in the first n iterations.

2.2.1. Wasserstein-2 Rates for Constant Step-size RLMC. Before, we state our main result, we investigate a few important characteristics of the (RLMC) algorithm that are not explored yet. We start with its rate of convergence in Wasserstein-2 distance (see [Vil09] for definition) for the (RLMC) algorithm. The proof of the proposition below essentially follows from a similar idea of the more general result for the underdamped Langevin dynamics in [SL19]. We include the result with its proof for the sake of completeness.

Proposition 1. Suppose f satisfies Assumption 2.1.1. Set $x_0 = \arg\min_x f(x)$, $\gamma_n := h = \mathcal{O}(\epsilon^{2/3}/\kappa^{1/3}M)$ when $\kappa hM > 1$, and $\gamma_n := h = \mathcal{O}(\epsilon/M)$ when $\kappa hM \leq 1$ with $Mh < \frac{1}{4}$. After running the (RLMC) algorithm for

$$K = \tilde{\mathcal{O}}\left(\frac{\kappa^{4/3}}{\epsilon^{2/3}} + \frac{\kappa}{\epsilon}\right) \quad \text{steps,}$$

we have $W_2(\nu_K, \pi) \leq \epsilon\sqrt{d/m}$, where ν_K is the probability distribution of x_K.

When κ is of constant order, we see that W_2 rate is of order $\tilde{\mathcal{O}}(1/\epsilon)$. Notably, with the randomized midpoint technique, we obtain this particular ϵ-dependency by discretizing just the overdamped Langevin diffusion with only the Lipschitz gradient condition on the potential function f. Prior works require Euler-discretization of higher-order Langevin diffusions to obtain a W_2 rate of order $\tilde{\mathcal{O}}(1/\epsilon)$ [DK19, MMW$^+$19] or require higher-order smoothness assumption along with other specialized discretization methods [SZ19, LWME19, DM19, DK19].

2.2.2. Analysis of the Markov Chain Generated by Constant Step-size RLMC. Using the randomized midpoint technique, we obtain an improved dependency on ϵ for the W_2 rate under weaker assumptions while discretizing the Langevin diffusion in (2.2). Although not explicit from the proof of Proposition 1, the rate improvement is obtained by a careful balancing of bias and variance through the choice of step-size parameter h. In this section, in Theorem 1, we first show that the (RLMC) Markov chain is ergodic and has a unique stationary distribution, denoted by π_h. Due to the choice of constant step-size h, it is not hard to see that the stationary distribution of the (RLMC) is different from the stationary distribution π of the Lanvegin diffusion in (2.2), i.e $\pi_h \neq \pi$. Hence, in Proposition 2, we characterize the Wasserstein-2 distance between π and π_h.

Firstly, if $f \in \mathcal{C}^2(\mathbb{R}^d)$ and f has a Lipschitz gradient with parameter M, then we can immediately see that the transition kernel of chain (x_n), $P(x, y) \in \mathcal{C}(\mathbb{R}^d \times \mathbb{R}^d)$ is positive everywhere. Therefore, it's easy to obtain that the chain (x_n) is μ^{Leb}-irreducible and aperiodic. Given all this information, we can give a sufficient condition to make sure that the chain has a unique invariant probability measure, and it is ergodic.

Theorem 1. Let the potential function f satisfy part (a) of Assumption 2.1.1, and let $\gamma_n := h$ be small enough. Then the (RLMC) Markov chain (x_n) has a unique stationary probability measure

π_h, and for every $x \in \mathbb{R}^d$, we have

$$\sup_{A \in \mathcal{B}(\mathbb{R}^d)} |P^n(x, A) - \pi_h(A)| \to 0 \qquad \text{as} \qquad n \to \infty.$$

We next address the question: how far is π_h from π? This question can be typically answered by a careful inspection on the proof of Proposition 1. However, for (RLMC), this is not the case, and requires using a different technique. Towards that, we derive an upper bound of $W_2(\pi, \pi_h)$ under the same assumptions in the previous theorem and the additional assumption that f is also strongly convex with parameter m.

Proposition 2. Let the potential function satisfy Assumption 2.1.1, and let $\gamma_n := h \in (0, \frac{2}{m+M})$ in the (RLMC) algorithm. Then, we have

$$W_2(\pi, \pi_h) \leq 3\sqrt{dh}\frac{(1+2Mh)^2}{\kappa^{-1} - Mh/\sqrt{3}}. \tag{2.3}$$

Remark 1. The above proposition shows that the order of the bias between the stationary distribution of the Langevin diffusion and that of the (RLMC) chain is of the order $\mathcal{O}(\sqrt{h})$.

2.2.3. Wasserstein-2 rates and CLT with Decreasing Step-size. In this part, we consider the (RLMC) algorithm with a fast decreasing time step sequence (γ_n) and establish a convergence rate in W_2 distance as well as a CLT for the numerical integration (2.1).

Proposition 3. Suppose f satisfies Assumption 2.1.1. Let $x_0 := \arg\min_x f(x)$ and $\gamma_{n+1} \leq \frac{m}{m^2+M^2(33+n)}$. After running (RLMC) algorithm for $K = \mathcal{O}\left(\kappa^{1.5}/\epsilon\right)$ steps, we obtain $W_2(\nu_K, \pi) \leq \epsilon\sqrt{d/m}$, where ν_K is the probability distribution of x_K.

Remark 2. There are two aspects of this result. The first aspect is rather standard; there is no logarithmic factor in $1/\epsilon$ compared to the result in Proposition 1. Similar phenomenon has been previously observed for the LMC algorithm [DK19]. The second aspect is that we never obtain the $\mathcal{O}(1/\epsilon^{2/3})$ term as in Proposition 1, with the constant step-size choice. This is not an artifact of our analysis. This is due to the fact that with this choice of decreasing step-size, we reduce the bias much more at the expense of slightly increased variance. However, as we demonstrate next, this choice of decreasing step-size is crucial for obtaining an unbiased CLT for numerical integration.

As the main contribution of this section, we characterize the fluctuations of (RLMC) when it is used for computing the integral $\int_{\mathbb{R}^d} \varphi d\pi$ for a π-integrable function φ. Choosing the Langevin diffusion in (2.2) with the stationary distribution π, we have by Theorem 1 that it is ergodic, and $\lim_{t\to+\infty} \frac{1}{t}\int_0^t \varphi(X(s))ds = \int_{\mathbb{R}^d} \varphi d\pi := \pi(\varphi)$, almost surely. Motivated by this, we first discretize the diffusion using (RLMC) and then compute a discrete analogue of the average. The procedure consists of two successive phases:

(a) **Discretization:** The (RLMC) algorithm is run with a step size sequence (γ_n) satisfying for all n, $\gamma_n > 0$, $\lim_{n\to+\infty} \gamma_n = 0$, and $\lim_{n\to+\infty} \Gamma_n = +\infty$, where $\Gamma_n := \sum_{k=1}^n \gamma_k$.

(b) **Averaging:** Using the (RLMC) iterates (x_n), construct a weighted empirical measure via the same weight sequence $\gamma := (\gamma_n)$: For every $n \geq 1$ and every $\omega \in \Omega$, set

$$\pi_n^\gamma(\omega, dx) := \frac{\gamma_1 \delta_{x_0(\omega)} + \cdots + \gamma_{k+1}\delta_{x_k(\omega)} + \cdots + \gamma_n \delta_{x_{n-1}(\omega)}}{\gamma_1 + \cdots + \gamma_n},$$

and use $\pi_n^\gamma(\omega, \varphi) := \int_{\mathbb{R}^d} \varphi \pi_n^\gamma(\omega, dx) = \frac{1}{\Gamma_n}\sum_{k=1}^n \gamma_k \varphi(x_{k-1}(\omega))$ to estimate the expectation (2.1).

For numerical purposes, for a fixed function φ, $\pi_n^\gamma(\omega, \varphi)$ can be recursively computed as follows:

$$\pi_{n+1}^\gamma(\omega, \varphi) = \pi_n^\gamma(\omega, \varphi) + \tilde{\gamma}_{n+1}\left(\varphi(x_n(\omega)) - \pi_n^\gamma(\omega, \varphi)\right) \quad \text{with } \tilde{\gamma}_{n+1} := \frac{\gamma_{n+1}}{\Gamma_{n+1}}.$$

We now provide the main result of this section, a central limit theorem for the algorithm (RLMC) when it is used to compute integrals of the form in (2.1).

Theorem 2. Let π be such that its potential f satisfies Assumption 2.1.1. Consider a test function $\varphi : \mathbb{R}^d \to \mathbb{R}$ of the form $\varphi = \mathcal{A}\phi$ for some function $\phi : \mathbb{R}^d \to \mathbb{R}$, where $\mathcal{A}$ denotes the generator of the diffusion (2.2), i.e., $\mathcal{A}\phi := -\langle \nabla f, \nabla \phi \rangle + \Delta\phi$. Define $\hat{\gamma}_n := \frac{1}{\sqrt{\Gamma_n}} \sum_{k=1}^n \gamma_k^2$ and let $\hat{\gamma}_\infty = \lim_{n\to\infty} \hat{\gamma}_n$. Then for all $\phi \in \mathcal{C}^4(\mathbb{R}^d)$ with $D^2\phi$, $D^3\phi$ being bounded, and $D^4\phi$ being bounded and Lipschitz, and

$\sup_{x\in\mathbb{R}^d} \|\nabla\phi(x)\|^2/(1+\|x\|^2) < +\infty$, we have the following central limit theorem for the numerical integration computed via (RLMC):

(i) If $\hat{\gamma}_\infty = 0$, then $\sqrt{\Gamma_n}\pi_n^\gamma(\varphi) \xrightarrow{d} \mathcal{N}(0, 2\int_{\mathbb{R}^d} \|\nabla\phi(x)\|^2\pi(dx))$,

(ii) If $\hat{\gamma}_\infty \in (0,+\infty)$, then $\sqrt{\Gamma_n}\pi_n^\gamma(\varphi) \xrightarrow{d} \mathcal{N}(\varrho\, \hat{\gamma}_\infty, 2\int_{\mathbb{R}^d} \|\nabla\phi(x)\|^2\pi(dx))$,

(iii) If $\hat{\gamma}_\infty = +\infty$, then $\frac{\sqrt{\Gamma_n}}{\hat{\gamma}_n}\pi_n^\gamma(\varphi) \xrightarrow{p} \varrho$,

where the mean ϱ is given as

$$\begin{aligned}\varrho = &\int\int\langle D^3\phi(x), \nabla f(x)\otimes u\otimes u\rangle\mu(du)\pi(dx) - \tfrac{1}{2}\int\langle D^2 f(x), \nabla\phi(x)\otimes\nabla f(x)\rangle\pi(dx)\\ &+ \tfrac{1}{2}\int\int\langle D^3 f(x), \nabla\phi(x)\otimes u\otimes u\rangle\mu(du)\pi(dx) - \tfrac{1}{2}\int\langle D^2\phi(x), \nabla f(x)\otimes\nabla f(x)\rangle\pi(dx)\\ &+ \int_{\mathbb{R}^d} trace(D^2\phi(x)^2)\pi(dx) - \tfrac{1}{6}\int\int\langle D^4\phi(x), u^{\otimes 4}\rangle\mu(du)\pi(dx),\end{aligned}$$

and μ is the distribution for a d-dimensional standard Gaussian measure.

Remark 3. First note that a CLT for the Euler discretization of Langevin diffusion follows from [LP02, Thm. 10]. The rates of the CLT established in Theorem 2 are similar to that case, with only the bias term ρ being different. Specifically, following the same computation in [LP02], we see that the optimal rate with polynomially decaying step-size choice $\gamma_k = k^{-\alpha}$, for some $\alpha > 0$, is $\mathcal{O}(n^{1/3})$. But in this case, the established CLT is biased. However, for any $0 < \alpha < 1/3$, we obtain an unbiased CLT as well. Hence, although the (RLMC) chain provides rate improvements for sampling (with respect to W_2 distance), as demonstrated in [SL19] and in Proposition 1, it does not seem to provide any improvements for CLT. In retrospect, this is expected as the rate improvements for sampling is achieved by the choice of constant step-size for which it is not possible to establish even a nearly unbiased CLT.

The class of test functions that the above CLT can cover is intimately related to the solution of the *Stein equation* (or Poisson equation) $\varphi = \mathcal{A}\phi$. Given φ, there is an explicit characterization of ϕ that solves the Stein's equation, and various properties of φ are translated to ϕ [GDVM16, EMS18].

2.3. Results for the Underdamped Langevin Diffusion

The underdamped Langevin diffusion is given by

$$d\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} v(t) \\ -(\beta v(t) + u\nabla f(x(t))) \end{bmatrix} dt + \sqrt{2\beta u}\begin{bmatrix} 0_d \\ I_d \end{bmatrix} dW(t), \tag{2.4}$$

where $\beta > 0$ is the friction coefficient and $u > 0$ is the inverse mass. For simplicity, we will consider $\beta = 2$ in the later text. Under mild conditions, it is well-known that the continuous-time Markov process $(x(t), v(t))$ is positive recurrent, and its invariant distribution is given by $\nu(x, v) \propto \exp\left\{-f(x) - \frac{1}{2u}\|v\|^2\right\}$, $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$. This diffusion, with an additional Hamiltonian component, has gathered a lot of attention recently due to its improved convergence properties [DRD20a, CCBJ17, SL19, LBBG19, DMS19] and empirical performance [Nea11, CFG14].

The *randomized midpoint discretization* of the underdamped Langevin diffusion (2.4) is given as:

$$\begin{aligned} x_{n+\frac{1}{2}} &= x_n + \tfrac{1}{2}(1-e^{-2\alpha_{n+1}\gamma_{n+1}})v_n - \tfrac{u}{2}\left(\alpha_{n+1}\gamma_{n+1} - \tfrac{1}{2}(1-e^{-2\alpha_{n+1}\gamma_{n+1}})\right)\nabla f(x_n) + \sqrt{u}\sigma_{n+1}^{(1)}U_{n+1}^{(1)}, \\ x_{n+1} &= x_n + \tfrac{1}{2}(1-e^{-2\gamma_{n+1}})v_n - \tfrac{u}{2}\gamma_{n+1}(1-e^{-2(1-\alpha_{n+1})\gamma_{n+1}})\nabla f(x_{n+\frac{1}{2}}) + \sqrt{u}\sigma_{n+1}^{(2)}U_{n+1}^{(2)}, \\ v_{n+1} &= v_n e^{-2\gamma_{n+1}} - u\gamma_{n+1}e^{-2(1-\alpha_{n+1})\gamma_{n+1}}\nabla f(x_{n+\frac{1}{2}}) + 2\sqrt{u}\sigma_{n+1}^{(3)}U_{n+1}^{(3)}, \end{aligned} \tag{RULMC}$$

where (γ_n) is the sequence of time steps, $\sigma_n^{(1)}$, $\sigma_n^{(2)}$ and $\sigma_n^{(3)}$ are positive with $(\sigma_n^{(1)})^2 = \alpha_n\gamma_n + \frac{1-e^{-4\alpha_n\gamma_n}}{4} - (1 - e^{-2\alpha_n\gamma_n})$, $(\sigma_n^{(2)})^2 = \gamma_n + \frac{1-e^{-4\gamma_n}}{4} - (1 - e^{-2\gamma_n})$ and $(\sigma_n^{(3)})^2 = \frac{1-e^{-4\gamma_n}}{4}$, and (α_n) is a sequence of identically distributed random variables following the distribution $\alpha_n \sim U[0, 1]$. $(U_n^{(1)}, U_n^{(2)}, U_n^{(3)})$ are independent centered Gaussian random vectors in $\mathbb{R}^{3d}$, also independent of (α_n) and initial point (x_0, v_0), having the following pairwise covariances:

$$\begin{aligned} \mathrm{cov}(\sigma_n^{(1)}U_n^{(1)}, \sigma_n^{(2)}U_n^{(2)}) &= \left(\alpha_n\gamma_n - \left(e^{-\alpha_n\gamma_n} + e^{-2\gamma_n}\sinh(\alpha_n\gamma_n)\right)\sinh(\alpha_n\gamma_n)\right) I_{d\times d}, \\ \mathrm{cov}(\sigma_n^{(2)}U_n^{(2)}, \sigma_n^{(3)}U_n^{(3)}) &= \left(e^{-2\gamma_n}\sinh(\gamma_n)^2\right) I_{d\times d}, \\ \mathrm{cov}(\sigma_n^{(1)}U_n^{(1)}, \sigma_n^{(3)}U_n^{(3)}) &= \left(e^{-2\gamma_n}\sinh(\alpha_n\gamma_n)^2\right) I_{d\times d}. \end{aligned}$$

The (RULMC) algorithm has emerged as an optimal sampling algorithm in the sense that it achieves the information theoretical lower bound in both tolerance ϵ and dimension d for sampling from a strongly log-concave densities [CLW20, SL19]. Therefore, it is interesting to examine if (RULMC) based numerical integrator have any benefits in other MCMC-based tasks such as (2.1). Towards that, we characterize the order of bias with a constant step-size choice for (RULMC) iterates as proposed in [SL19]. Compared to the bias result in Proposition 2 for the (RLMC) discretization, we note that order of bias is increased (i.e. smaller bias). Next, in Theorem 4 we provide a CLT for numerical integration with (RULMC). Our results show that when it comes to computing expectations of the form in (2.1) using (RULMC) and characterizing its fluctuations, the (RULMC) discretization obtains rate improvements only for a class of constant test functions (as described in Remark 6).

2.3.1. Analysis of the Markov Chain generated by Constant Step-size RULMC. Recall that $\pi(x)$ is the marginal density function of $\nu(x,v)$ with respect to x. Similarly $\nu_h(x,v)$ be the stationary density function of the Markov chain generated by (RULMC) chain and $\pi_h(x)$ be the marginal density function of $\nu_h(x,v)$, with respect to x. Furthermore, the filtration $(\mathcal{F}_n)$ is defined as $\mathcal{F}_n := \sigma(\alpha_k, U_k^{(i)}; 1 \leq k \leq n, i = 1,2,3)$. When $f \in \mathcal{C}^2(\mathbb{R}^d)$ and is gradient Lipschitz with parameter M, then we can immediately see that the transition kernel of chain (x_n, v_n): $P((x,v),(x',v')) \in \mathcal{C}(\mathbb{R}^{2d} \times \mathbb{R}^{2d})$ is positive everywhere. Therefore, it's easy to obtain that the chain (x_n, v_n) is μ^{Leb}-irreducible and aperiodic. Given all this information, we can give a sufficient condition to make sure that the chain has a unique invariant probability measure and is ergodic.

Theorem 3. Let the potential function f satisfy part (a) of Assumption 2.1.1, and let $\gamma_n := h$ be small enough. Then if $u \in (0, \frac{4}{2M-m})$, the (RULMC) Markov chain (x_n, v_n) has a unique stationary probability measure ν_h and for every $(x,v) \in \mathbb{R}^{2d}$, we have

$$\sup_{A \in \mathcal{B}(\mathbb{R}^{2d})} |P^n((x,v),A) - \nu_h(A)| \to 0 \qquad \text{as} \qquad n \to \infty.$$

We next derive an upper bound on the bias $W_2(\pi, \pi_h)$ of (RULMC) algorithm, under the additional strong convexity assumption on the potential function f.

Proposition 4. Suppose that f satisfies Assumption 2.1.1. If we run the (RULMC) algorithm with $u = 1/M$ and $\gamma_n := h$, for universal constants $C_1, C_2 > 0$, we have

$$W_2^2(\pi, \pi_h) \leq \frac{C_1 h^3(\kappa h^3 + 1)d}{1 - \frac{h}{4\kappa} - C_2 h^3 \kappa(1 + \kappa h^3)}.$$

Remark 4. Note that we have $W_2(\pi, \pi_h) \to 0$ as $h \to 0$. Furthermore, as $h \to 0$, $W_2(\pi, \pi_h) < \mathcal{O}(h^{\frac{3}{2}})$. Hence, the bias order is increased for the underdamped Langevin diffusion compared to the overdamped case (cf. Proposition 2), providing a smaller bias for the same step-size.

2.3.2. Wasserstein-2 rates and CLT with Decreasing Step-size. We now provide the rate of convergence in Wasserstein-2 metric with decreasing step-size for (RULMC). The specific choice for the decreasing step-size that we consider below, also is satisfied for our CLT result in Remark 6.

Proposition 5. Suppose f satisfies Assumption 2.1.1. Fix $u = 1/M$. Let $x_0 := \arg\min_x f(x)$ and choose $\gamma_n = \frac{16\kappa}{32\kappa^{\frac{5}{3}} + (n - K_1)^+}$, for a $K_1 \in (0, \infty)$ (where $(a)^+ := \max(0, a)$). After running (RULMC) for $K = \tilde{\mathcal{O}}\left(\kappa^{3/2}/\epsilon^{2/3}\right)$ steps, we obtain $W_2(\nu_K, \pi) \leq \epsilon\sqrt{d/m}$, where ν_K is the probability distribution of x_K.

Remark 5. Similar to the result in Proposition 3, there are two aspects of this result. The first aspect is again removing the logarithmic factor in $1/\epsilon$ compared to the result in Theorem 3 in [SL19], which is quite standard in the literature. The second aspect is that we never obtain the $\mathcal{O}(1/\epsilon^{1/3})$ part, as in Theorem 3 in [SL19] with the constant step-size choice.

Similar to the previous case, we now describe the numerical integration procedure using the (RULMC) discretization. We denote the n-th iterate as (x_n, v_n). The time-step we use is (γ_n) such that $\forall n \in \mathbb{N}^*, \gamma_n \geq 0, \lim_n \gamma_n = 0$ and $\lim_n \Gamma_n^{(1)} = +\infty$, where $\Gamma_n^{(\ell)} := \sum_{i=1}^n \gamma_i^\ell$. Our averaging is a weighted empirical measure with $Y_n = (x_n, v_n)$ using the step size sequence $\gamma := (\gamma_n)$ as the weights. Let δ_x denote the Dirac mass at x. Then for every $n \geq 1$, set

$$\nu_n^\gamma(\omega, dx) := \frac{\gamma_1 \delta_{Y_0(\omega)} + \cdots + \gamma_{k+1}\delta_{Y_k(\omega)} + \cdots + \gamma_n \delta_{Y_{n-1}(\omega)}}{\gamma_1 + \cdots + \gamma_n}$$

and we can use $\nu_n^\gamma(\omega, \varphi)$ to approximate $\nu(\varphi) = \mathbb{E}_\nu[\varphi'(Y)]$, where $\varphi' : \mathbb{R}^{2d} \to \mathbb{R}$.

If we assume $g : \mathbb{R}^{2d} \to \mathbb{R}$ such that $\mathcal{L}g = \varphi'$, we can establish the following theorem, in which we state only the unbiased CLT result for simplicity.

Theorem 4. Let π be such that its potential function f satisfies Assumption 2.1.1. Assume $u \in (0, \frac{4}{2M-m})$. Consider a test function $\varphi' = \mathcal{L}g$, for some function $g : \mathbb{R}^{2d} \to \mathbb{R}$, where $\mathcal{L} = 2u\Delta_v - 2\langle v, \nabla_v\rangle - u\langle \nabla f(x), \nabla_v\rangle + \langle v, \nabla_x\rangle$ denotes the generator of the diffusion (2.4). Suppose the step-size (γ_k) is non-increasing, $\lim_{n\to+\infty}(1/\sqrt{\Gamma_n})\sum_{k=1}^n \gamma_k^{3/2} = +\infty$. Then, if $\lim_{n\to+\infty}(1/\sqrt{\Gamma_n})\sum_{k=1}^n \gamma_k^2 = 0$, for every $g \in \mathcal{C}^4(\mathbb{R}^{2d})$ function with $D^2 g$ bounded, $D^3 g$ bounded and Lipschitz, and if the condition $\sup_{(x,v)\in\mathbb{R}^{2d}} \|\nabla g(x,v)\|/(1 + \|x\|^2 + \|v\|^2) < +\infty$ holds, we have the following central limit theorem for the numerical integration computed using the (RULMC) iterates:

$$\sqrt{\Gamma_n}\nu_n^\gamma(\mathcal{L}g) \xrightarrow{d} \mathcal{N}\big(0, 4u \textstyle\int \|\nabla_v g(x,v)\|^2 \nu(dx, dv)\big).$$

The rate of convergence of the CLT in Theorem 4 follows exactly the same behavior in Theorem 2. Hence, for the class of general test functions, Theorem 4 does not exhibit a rate improvement. Towards that, we make the following remarks under a carefully constructed condition for the class of test functions.

Remark 6. Let π be such that its potential function f satisfies Assumption 2.1.1. Assume $u \in (0, \frac{4}{2M-m})$. Consider a test function $\varphi = \mathcal{L}g$ which could be written as $\mathcal{L}g(v, \phi(x)) = \langle v, \nabla\phi(x)\rangle$, for some function $\phi : \mathbb{R}^d \to \mathbb{R}$, where $\mathcal{L} = 2u\Delta_v - 2\langle v, \nabla_v\rangle - u\langle \nabla f(x), \nabla_v\rangle + \langle v, \nabla_x\rangle$ denotes the generator of the diffusion (2.4). Suppose the time step-size (γ_k) is non-increasing, and satisfies $\lim_{n\to\infty}(\gamma_{n-1} - \gamma_n)/\gamma_n^4 = 0$ and $\lim_{n\to\infty}\Gamma_n^{(4)} = +\infty$. Define $\hat{\gamma}_n := \Gamma_n^{(4)}/\sqrt{\Gamma_n^{(3)}}$ and let $\hat{\gamma}_\infty = \lim_{n\to\infty}\hat{\gamma}_n$. Then, for all $\phi \in \mathcal{C}^4(\mathbb{R}^d)$ with $D^2\phi$, $D^3\phi$ and $D^4\phi$ bounded and Lipschitz and $\sup_{(x,v)\in\mathbb{R}^{2d}} \|\nabla\phi(x)\|^2/(1 + \|x\|^2 + \|v\|^2) < +\infty$, we obtain the following central limit theorem for numerical integration computed using the (RULMC) algorithm:

(i) If $\hat{\gamma}_\infty = 0$, we have $\frac{\Gamma_n}{\sqrt{\Gamma_n^{(3)}}}\nu_n^\gamma(\mathcal{L}\phi) \xrightarrow{d} \mathcal{N}(0, \frac{10}{3}u \int_{\mathbb{R}^d} \|\nabla\phi(x)\|\pi(dx))$,

(ii) If $\hat{\gamma}_\infty \in (0, +\infty)$, we have $\frac{\Gamma_n}{\Gamma_n^{(4)}}\nu_n^\gamma(\mathcal{L}\phi) \xrightarrow{d} \mathcal{N}(\rho, \frac{10}{3}u\hat{\gamma}_\infty^{-2} \int_{\mathbb{R}^d} \|\nabla\phi(x)\|\pi(dx))$,

(iii) If $\hat{\gamma}_\infty = +\infty$, we have $\frac{\Gamma_n}{\Gamma_n^{(4)}} \nu_n^\gamma(\mathcal{L}\phi) \xrightarrow{p} \rho$,

where,

$$\rho = \tfrac{5u}{12} \int\int \langle D^3\phi(x), \nabla f(x) \otimes v \otimes v\rangle \nu(dx, dv) + \tfrac{u}{24} \int\int \langle D^3 f(x), \nabla\phi(x) \otimes v \otimes v\rangle \nu(dx, dv)$$

$$+ \tfrac{7u}{12} \int\int (D^2\phi D^2 f)(x) v^{\otimes 2} \nu(dx, dv) - \tfrac{u^2}{4} \int \langle D^2\phi(x), \nabla f(x)^{\otimes 2}\rangle \pi(dx)$$

$$- \tfrac{u^2}{24} \int \langle D^2 f(x), \nabla\phi(x) \otimes \nabla f(x)\rangle \pi(dx).$$

Remark 7. For polynomial time steps $\gamma_k := k^{-\alpha}$, since we require that $\Gamma_n^{(4)} \to +\infty$ as $n \to +\infty$, we need $0 < \alpha \leq \frac{1}{4}$. Using L'Hospitals rule, it is straightforward to check that the condition $\lim_{n\to+\infty} \frac{\gamma_{n-1}-\gamma_n}{\gamma_n^4} = 0$ is satisfied when $\alpha \in (0, \frac{1}{4}]$. We then have the following order estimates:

$$\Gamma_n \sim \frac{n^{1-\alpha}}{1-\alpha}, \qquad \sqrt{\Gamma_n^{(3)}} \sim \frac{n^{\frac{1}{2}-\frac{3}{2}\alpha}}{\sqrt{1-3\alpha}}, \qquad \Gamma_n^{(4)} \sim \begin{cases} \dfrac{n^{1-4\alpha}}{1-4\alpha}, & \text{if } \alpha \in (0, \frac{1}{4}), \\ \sqrt{\ln n}, & \text{if } \alpha = \frac{1}{4}. \end{cases}$$

Hence, as $n \to +\infty$,

$$\frac{\Gamma_n^{(4)}}{\sqrt{\Gamma_n^{(3)}}} \to \hat{\gamma}_\infty = \begin{cases} 0 & \text{if } \alpha \in (\frac{1}{5}, \frac{1}{4}], \\ \sqrt{10} & \text{if } \alpha = \frac{1}{5}, \\ +\infty & \text{if } \alpha \in (0, \frac{1}{5}). \end{cases}$$

If $\alpha \in (\frac{1}{5}, \frac{1}{4}]$, the unbiased CLT holds at rate $\Gamma_n/\sqrt{\Gamma_n^{(3)}} = \mathcal{O}(n^{\frac{1}{2}(1+\alpha)}) \leq \mathcal{O}(n^{\frac{5}{8}})$. The optimal rate is achieved when $\alpha = \frac{1}{4}$. If $\alpha = \frac{1}{5}$, the biased CLT holds at rate $\Gamma_n/\sqrt{\Gamma_n^{(3)}} = \mathcal{O}(n^{3\alpha}) = \mathcal{O}(n^{\frac{3}{5}})$. If $\alpha \in (0, \frac{1}{5})$, the rate of the convergence in probability is $\Gamma_n/\sqrt{\Gamma_n^{(3)}} = \mathcal{O}(n^{3\alpha}) < \mathcal{O}(n^{\frac{3}{5}})$. Therefore the optimal convergence rate $\mathcal{O}(n^{\frac{5}{8}})$ is obtained when an unbiased CLT holds. While the rate of this CLT is faster than the one obtained in Theorem 2, the test functions that satisfy this condition is severely restricted.

2.3.3. Discussion. In this work, we present several probabilistic properties of the randomized midpoint discretization technique, focussing our attention on overdamped and underdamped

Langevin diffusion. Our results could be biased as follows: To obtain optimal rates for sampling (in W_2 distance), one needs to have a constant choice of step-size. With such a constant step-size choice, the Markov chain generated by the discretization process is biased. This suggest that a decreasing step-size choice is required for using the randomized midpoint method for sampling and the related task of numerical integration. For several decreasing choices of step-sizes, we establish CLTs and highlight the relative merits and disadvantages of using randomized midpoint technique for numerical integration. In particular, our results have interesting consequence for computing confidence interval for numerical integration.

2.4. Additional Notations

We also use the following notations for the proofs. Due to the ease of presentation, whenever it is clear in the proof, we refer to the inner product between two compatible vectors $\langle a, b\rangle$ simply by $a \cdot b$. For any random variable X, $\|X\|_{L^2} := \mathbb{E}[\|X\|^2]^{\frac{1}{2}}$ where the expectation is taken over all randomness of X.

2.5. Proofs for Section 2.2

We now define the following condition, which is a consequence of Assumption 2.1.1

Assumption 2.5.1. *There exists a twice differentiable function* $V: \mathbb{R}^d \to [1, \infty)$ *such that:* *(i)* $\lim_{\|x\|\to\infty} V(x) = +\infty$, *(ii) there exists* $\alpha > 0$ *and* $\beta > 0$*:* $\langle \nabla V(x), \nabla f(x)\rangle \geq \alpha V(x) - \beta$ *for every* x, *(iii) there exists* $c_V > 0$*:* $\|\nabla V(x)\|^2 + \|\nabla f(x)\|^2 \leq c_V V(x)$ *for every* x, *and (iv)* $\left\|D^2 V\right\|_\infty := \sup_{x\in\mathbb{R}^d} \|D^2 V\|_{op} < \infty$ *(where* $\|\cdot\|_{op}$ *denotes the operator norm).*

Lemma 2.5.1. *Assumption 2.1.1 implies Assumption 2.5.1.*

Proof. Since $f \in \mathcal{C}^2(\mathbb{R}^d)$ is strongly convex, $\lim_{|x|\to+\infty} f(x) = +\infty$ and f has a unique global minimizer $x^* \in \mathbb{R}^d$. It's easy to observe that $\nabla f(x^*) = 0$. We consider our $V(x) = f(x) - f(x^*) + 1$. Then it's easy to see (i) is satisfied. (iv) is also satisfied because f is gradient Lipschitz. (iii) is equivalent to that there exists a $C > 0$ such that

$$\frac{|\nabla f(x)|^2}{f(x) - f(x^*) + 1} \leq C \qquad \text{for } \forall x \in \mathbb{R}^d.$$

We Taylor expand the numerator and denominator:

$$\begin{aligned}
|\nabla f(x)|^2 &= \sum_{i=1}^{d} \left(f_i(x^*) + \nabla f_i(\xi)^T (x - x^*)\right)^2 \\
&\leq \sum_{i,j=1}^{d} |f_{ij}(\xi)|^2 |x - x^*|^2 = \left\|D^2 f(\xi)\right\|_F^2 |x - x^*|^2 \\
&\leq d^2 M^2 |x - x^*|^2,
\end{aligned}$$

and

$$\begin{aligned}
f(x) - f(x^*) + 1 &= \nabla f(x^*)^T (x - x^*) + \frac{1}{2} D^2 f(\xi)(x - x^*)^{\otimes 2} + 1 \\
&= \frac{1}{2} D^2 f(\xi)(x - x^*)^{\otimes 2} + 1 \\
&\geq \frac{m}{2} |x - x^*|^2.
\end{aligned}$$

Then

$$\frac{|\nabla f(x)|^2}{f(x) - f(x^*) + 1} \leq \frac{2d^2 M^2}{m} \qquad \text{for } \forall x \in \mathbb{R}^d.$$

(ii) is equivalent to that there exists $\alpha, \beta > 0$ such that

$$|\nabla f(x)|^2 \geq \alpha (f(x) - f(x^*) + 1) - \beta \qquad \text{for } \forall x \in \mathbb{R}^d.$$

According to the strongly convexity of f, we have

$$\begin{aligned}
f(x^*) - f(x) &\geq \nabla f(x)^T (x^* - x) + \frac{m}{2} |x^* - x|^2 \\
&= \frac{m}{2} |x^* - x + \frac{1}{m} \nabla f(x)|^2 - \frac{1}{2m} |\nabla f(x)|^2,
\end{aligned}$$

which then implies

$$|\nabla f(x)|^2 \geq 2m \left(f(x) - f(x^*) + 1\right) - 2m \qquad \text{for } \forall x \in \mathbb{R}^d.$$

(ii) is satisfied by choosing $\alpha = \beta = 2m > 0$. ■

Remark 8. For the $V(x)$ we choose in the proof, under assumption 2.1.1, we can verify that: $V(x) = O(|x|^2)$ when $|x| \to +\infty$. We will use this fact later in the proof when we establish the CLT statement.

2.5.1. Proofs for Section 2.2.1.

LEMMA 2.5.2. *Let $x(t)$ be the solution to Langevin dynamics SDE with initial condition x_0 and $y(t)$ be the solution to Langevin dynamics SDE with initial condition y_0. Then we have the following estimates for Langevin dynamics when f satisfies Assumption 2.1.1 and $Mh < \frac{1}{2}$:*

$$\mathbb{E}[\sup_{t\in[0,h]} \|\nabla f(x(t))\|^2] \leq 4\|\nabla f(x_0)\|^2 + 8M^2dh,$$

$$\mathbb{E}[\sup_{t\in[0,h]} \|x(t)-x_0\|^2] \leq O(h^2\|\nabla f(x_0)\|^2 + M^2h^3d + 2dh),$$

$$\mathbb{E}[\|x(t)-y(t)\|^2] \leq e^{-2mt}\|x_0-y_0\|^2 .$$

PROOF. By triangle inequality we have

$$\mathbb{E}[\sup_{t\in[0,h]} \|\nabla f(x(t))\|^2] \leq 2\|\nabla f(x_0))\|^2 + 2M^2\mathbb{E}[\sup_{t\in[0,h]} \|x(t)-x_0\|^2].$$

Furthermore, we have

$$\begin{aligned}\mathbb{E}[\sup_{t\in[0,h]} \|x(t)-x_0\|^2] &= \mathbb{E}[\sup_{t\in[0,h]} \left\|-\int_0^t \nabla f(x(s))ds + \sqrt{2}W_t\right\|^2]\\ &\leq h^2\mathbb{E}[\sup_{t\in[0,h]} \|\nabla f(x(t))\|^2] + 2dh.\end{aligned}$$

Combining the two inequalities and $Mh < \frac{1}{2}$, we can obtain the first two estimates. The last estimate could be easily obtained by energy method. ■

PROOF OF PROPOSITON 1. We denote $x_n = x_n(0)$ to be the algorithm iterate points, y_n to be the n-th step of Langevin diffusion with $y_0 \sim \exp(-f(y))$, $x^*_{n+1} = x_n(h)$ to be one step solution of Langevin dynamics with initial values x_n. When $Mh < \frac{1}{2}$, apply lemma 2.5.2 and we get:

$$\mathbb{E}[\sup_{t\in[0,h]} \|x_{n-1}(\alpha_n h) - x_{n-1}(t)\|^2] \leq O(h^2\|\nabla f(x_{n-1})\|^2_{L^2} + M^2h^3d + 2dh),$$

$$\begin{aligned}\mathbb{E}[\left\|\nabla f(x_{n-\frac{1}{2}}) - \nabla f(x_{n-1}(\alpha_n h))\right\|^2] &\leq M^2\mathbb{E}\left\|\int_0^{\alpha_n h} \nabla f(x_{n-1}(s)) - \nabla f(x_{n-1}(0))ds\right\|^2\\ &\leq M4h^2\mathbb{E}[\alpha_n^2 \sup_{t\in[0,\alpha_n h]} \|x_{n-1}(t)-x_{n-1}(0)\|^2]\\ &\leq O(M^4h^4\|\nabla f(x_{n-1})\|^2_{L^2} + dM^4h^3 + dM^6h^5).\end{aligned}$$

Consider the distance between our iterates and the continuous process:

$$\begin{aligned}
\mathbb{E}_{\alpha_K}[\|x_K - y_K\|^2] &= \mathbb{E}_{\alpha_K}[\|x_K - x_K^* + x_K^* - y_K\|^2] \\
&\leq \|y_K - x_K^*\|^2 + \mathbb{E}_{\alpha_K}[\|x_K - x_K^*\|^2] - 2(y_K - x_K^*)^T(\mathbb{E}_{\alpha_K} x_K - x_K^*) \\
&\leq (1 + hm)\left\|y_K - x_K^*\right\|^2 + \frac{1}{hm}\left\|\mathbb{E}_{\alpha_K} x_K - x_K^*\right\|^2 + \mathbb{E}_{\alpha_K}[\|x_K - x_K^*\|^2].
\end{aligned}$$

Taking expectations over $\{\alpha_k, U_l, U_l'; 1 \leq k \leq K-1, 1 \leq l \leq K\}$, applying lemma 2.5.2 again and using induction, we have

$$\begin{aligned}
\|x_K - y_K\|_{L^2}^2 &\leq (1 + hm)\left\|y_K - x_K^*\right\|_{L^2}^2 + \frac{1}{hm}\mathbb{E}\left\|\mathbb{E}_{\alpha_K} x_K - x_K^*\right\|^2 + \|x_K - x_K^*\|_{L^2}^2 \\
&\leq (1 + hm)e^{-2mh}\left\|x_{K-1} - y_{K-1}\right\|_{L^2}^2 + \frac{1}{hm}\mathbb{E}\left\|\mathbb{E}_{\alpha_K} x_K - x_K^*\right\|^2 + \|x_K - x_K^*\|_{L^2}^2 \\
&\leq (1 + hm)e^{-2mKh}\left\|x_0 - y_0\right\|_{L^2}^2 + \sum_{n=1}^{K}\frac{1}{hm}\mathbb{E}\left\|\mathbb{E}_{\alpha_n} x_n - x_n^*\right\|^2 + \sum_{n=1}^{K}\|x_n - x_n^*\|_{L^2}^2 \\
&\leq e^{-mKh}\left\|x_0 - y_0\right\|_{L^2}^2 + A + B.
\end{aligned}$$

Next we bound part A and part B. For part A:

$$\begin{aligned}
\|\mathbb{E}_{\alpha_n} x_n - x_n^*\|^2 &= \left\|\mathbb{E}_{\alpha_n}[h\nabla f(x_{n-\frac{1}{2}})] - \int_0^h \nabla f(x_{n-1}(s))ds\right\|^2 \\
&\leq 2\mathbb{E}_{\alpha_n}\left\|h\nabla f(x_{n-\frac{1}{2}}) - h\nabla f(x_{n-1}(\alpha_n h))\right\|^2 + 2\left\|\mathbb{E}_{\alpha_n}[h\nabla f(x_{n-1}(\alpha_n h))] - \int_0^h \nabla f(x_{n-1}(s))ds\right\|^2 \\
&\leq 2h^2\mathbb{E}_{\alpha_n}\left\|\nabla f(x_{n-\frac{1}{2}}) - \nabla f(x_{n-1}(\alpha_n h))\right\|^2 + 0.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}\left\|\mathbb{E}_{\alpha_n} x_n - x_n^*\right\|^2 &\leq 2h^2\mathbb{E}[\left\|\nabla f(x_{n-\frac{1}{2}}) - \nabla f(x_{n-1}(\alpha_n h))\right\|^2] \\
&\leq O(M^4h^6\left\|\nabla f(x_{n-1})\right\|_{L^2}^2 + dM^4h^5).
\end{aligned}$$

For part B, use our previous estimates:

$$\begin{aligned}\|x_n - x_n^*\|_{L^2}^2 &= \left\| h\nabla f(x_{n-\frac{1}{2}}) - \int_0^h \nabla f(x_{n-1}(s))ds \right\|_{L^2}^2 \\ &\leq 2\left\| h\nabla f(x_{n-\frac{1}{2}}) - h\nabla f(x_{n-1}(\alpha_n h)) \right\|_{L^2}^2 + 2\left\| \int_0^h \nabla f(x_{n-1}(s)) - \nabla f(x_{n-1}(\alpha_n h))ds \right\|_{L^2}^2 \\ &\leq 2h^2 \left\| \nabla f(x_{n-\frac{1}{2}}) - \nabla f(x_{n-1}(\alpha_n h)) \right\|_{L^2}^2 + 2M^2h^2\mathbb{E}[\sup_{t\in[0,h]} \|x_{n-1}(\alpha_n h) - x_{n-1}(t)\|^2] \\ &\leq O(M^2h^4 \|\nabla f(x_{n-1})\|_{L^2}^2 + dM^2h^3).\end{aligned}$$

Plug the estimates on A and B into the inequality we have

$$\begin{aligned}\|x_K - y_K\|_{L^2}^2 \leq e^{-mKh}\|x_0 - y_0\|_{L^2}^2 + O(m^{-1}M^4h^5 \sum_{n=0}^{K-1} \|\nabla f(x_n)\|_{L^2}^2 + dm^{-1}M^4Kh^4) \\ + O(M^2h^4 \sum_{n=0}^{K-1} \|\nabla f(x_n)\|_{L^2}^2 + dM^2Kh^3).\end{aligned}$$

Next we need to estimate $\sum_{n=0}^{K-1} \|\nabla f(x_n)\|_{L^2}^2$. Since

$$\begin{aligned}f(x_n(h)) &= f(x_n(0)) + \int_0^h df(x_n(t)) \\ &= f(x_n(0)) - \int_0^h |\nabla f(x_n(t))|^2 dt + \sqrt{2}\int_0^h \nabla f(x_n(t))dW(t) + \int_0^h \Delta f(x_n(t))dt.\end{aligned}$$

we have

$$\mathbb{E}[f(x_{n+1}(0))] - \mathbb{E}[f(x_n(h))] = \mathbb{E}[f(x_{n+1}(0)) - f(x_n(0))] + \mathbb{E}[\int_0^h |\nabla f(x_n(t))|^2 dt] - \mathbb{E}[\int_0^t \Delta f(x_n(t))dt].$$

When $Mh < \frac{1}{4}$,

$$\begin{aligned}\mathbb{E}[\inf_{t\in[0,h]} \|\nabla f(x(t))\|^2] &\geq \frac{1}{2}\|\nabla f(x(0))\|_{L^2}^2 - \mathbb{E}[\sup_{t\in[0,h]} \|\nabla f(x(t)) - \nabla f(x(0))\|^2] \\ &\geq \frac{1}{2}\|\nabla f(x(0))\|_{L^2}^2 - M^2\mathbb{E}[\sup_{t\in[0,h]} \|x(t) - x(0)\|^2] \\ &\geq \frac{1}{4}\|\nabla f(x(0))\|_{L^2}^2 + O(dM^2h)\end{aligned}$$

$$\text{and} \qquad |\Delta f(x_n(t))| \leq d\left\|\nabla^2 f(x_n(t))\right\| \leq Md.$$

Plug these two estimates into our previous identity and we obtain,

$$\mathbb{E}[f(x_{n+1}(0)) - f(x_n(h))] \geq \mathbb{E}[f(x_{n+1}) - f(x_n)] + \frac{h}{4}\left\|\nabla f(x_n)\right\|_{L^2}^2 - dMh + O(dM^2h^2).$$

Next we consider that

$$\begin{aligned}\mathbb{E}_{\alpha_{n+1}}[f(x_{n+1}(0))] &\leq f(x_n(h)) + \nabla f(x_n(h))^T(\mathbb{E}_{\alpha_{n+1}}[x_{n+1}(0)] - x_n(h)) + \frac{M}{2}\mathbb{E}_{\alpha_{n+1}}[\|x_{n+1}(0) - x_n(h)\|^2] \\ &\leq f(x_n(h)) + Mh^2\left\|\nabla f(x_n(h))\right\|_{L^2}^2 + M^{-1}h^{-2}\left\|\mathbb{E}_{\alpha_{n+1}}[x_{n+1}(0)] - x_n(h)\right\|^2 \\ &+ \frac{M}{2}\mathbb{E}_{\alpha_{n+1}}[\|x_{n+1}(0) - x_n(h)\|^2],\end{aligned}$$

where

$$\begin{aligned}Mh^2\mathbb{E}[\|\nabla f(x_n(h))\|^2] &\leq O(Mh^4\left\|\nabla f(x_n)\right\|_{L^2}^2 + dMh^3), \\ M^{-1}h^{-2}\left\|\mathbb{E}_{\alpha_{n+1}}[x_{n+1}(0)] - x_n(h)\right\|^2 &\leq O(M^3h^4\left\|\nabla f(x_n)\right\|_{L^2}^2 + dM^3h^3), \\ \frac{M}{2}\mathbb{E}[\|x_{n+1} - x_n(h)\|^2] &\leq O(M^3h^4\left\|\nabla f(x_n)\right\|_{L^2}^2 + dM^3h^3).\end{aligned}$$

Hence we have

$$\mathbb{E}[f(x_{n+1}(0)) - f(x_n(h))] \leq O(M^3h^4\left\|\nabla f(x_n)\right\|_{L^2}^2 + dM^3h^3)$$

and

$$O(M^3h^4\left\|\nabla f(x_n)\right\|_{L^2}^2 + dM^3h^3) \geq \mathbb{E}[f(x_{n+1}) - f(x_n)] + \frac{h}{4}\left\|\nabla f(x_n)\right\|_{L^2}^2 + O(dM^2h^2) - dMh.$$

Sum up over k from 0 to $K-1$:

$$O(M^3h^4\sum_{k=0}^{K-1}\left\|\nabla f(x_n)\right\|_{L^2}^2 + dM^3Kh^3) \geq \mathbb{E}[f(x_K) - f(x_0)] + \frac{h}{4}\sum_{k=1}^{K-1}\left\|\nabla f(x_n)\right\|_{L^2}^2 + O(dM^2Nh^2) - dMKh.$$

Picking $x_0 = \text{argmin} f(x)$, we can ensure $\mathbb{E}[f(x_K) - f(x_0)] \geq 0$, when $Mh < \frac{1}{2}$, we have

$$\begin{aligned}&\frac{h}{8}\sum_{k=0}^{K-1}\left\|\nabla f(x_n)\right\|_{L^2}^2 \leq dKMh - O(dKM^2h^2) + O(dKM^3h^3) \\ \Longrightarrow\quad &\sum_{k=0}^{K-1}\left\|\nabla f(x_n)\right\|_{L^2}^2 \leq O(dKM).\end{aligned}$$

Therefore

$$\begin{aligned}\|x_K - y_K\|_{L^2}^2 &\leq e^{-mKh} \|x_0 - y_0\|_{L^2}^2 + O(m^{-1}M^5h^5Kd + m^{-1}M^4h^4Kd) + O(M^3h^4Kd + M^2h^3Kd) \\ &\leq e^{-mKh} \|x_0 - y_0\|_{L^2}^2 + O(\kappa M^3h^4Kd) + O(M^2h^3Kd).\end{aligned}$$

Hence we have

$$W_2(\nu_K, \pi)^2 \leq e^{-mKh} \|x_0 - y_0\|_{L^2}^2 + O(M^3h^4Kd) \max\{\kappa, \frac{1}{Mh}\}$$

a) When $\kappa > \frac{1}{Mh}$, by choosing $h \sim O(\frac{\epsilon^{2/3}}{\kappa^{1/3}M})$, we can ensure $W_2(\nu_K, \pi)^2 \leq \epsilon^2 d/m$ after K steps when $K \sim \tilde{O}(\frac{\kappa^{4/3}}{\epsilon^{2/3}})$.

b) When $\kappa \leq \frac{1}{Mh}$, by choosing $h \sim O(\frac{\epsilon}{M})$, we can ensure $W_2(\nu_K, \pi)^2 \leq \epsilon^2 d/m$ after K steps when $K \sim \tilde{O}(\frac{\kappa}{\epsilon})$.

■

2.5.2. Proofs for Section 2.2.2.

Proof of Theorem 1. Under the assumption 2.5.1, we can show that the following Lyapunov condition is satisfied for small h.

(Lyapunov Condition): There exists a function $V: \mathbb{R}^d \to [1, \infty)$ such that:

0) $\lim_{|x| \to \infty} V(x) = +\infty$,

1) There exists $\hat{\alpha} \in (0, 1)$ and $\hat{\beta} \geq 0$: $\mathbb{E}[V(x_{n+1})|\mathcal{F}_n] \leq \hat{\alpha} V(x_n) + \hat{\beta}$.

Proof: To show that Assumption 2.5.1 implies Lyapunov condition, we first do Taylor expansion of $V(x_{n+1})$ at x_n:

$$\begin{aligned}V(x_{n+1}) = V(x_n) &- h\langle \nabla V(x_n), \nabla f(x_n)\rangle + \alpha_{n+1}h^2 \langle D^2 f(x_n); \nabla f(x_n), \nabla V(x_n)\rangle \\ &- \sqrt{2\alpha_{n+1}}h^{\frac{3}{2}} \langle D^2 f(x_n); \nabla V(x_n), U'_{n+1}\rangle + \sqrt{2h}\nabla V(x_n) \cdot U_{n+1} \\ &+ \frac{1}{2} D^2 V(\theta_n)(-h\nabla f(x_n) + \alpha_{n+1}h^2 D^2 f(x_n) \nabla f(x_n) - \sqrt{2\alpha_{n+1}}h^{\frac{3}{2}} U'_{n+1} + \sqrt{2h}U_{n+1})^{\otimes 2},\end{aligned}$$

where θ_n is a random point on the line segment joining x_n and x_{n+1}. Using the fact that f is M-gradient Lipschitz, we have:

$$\begin{aligned}\mathbb{E}[V(x_{n+1})|\mathcal{F}_n] &\leq V(x_n) - h\langle \nabla V(x_n), \nabla f(x_n)\rangle + \frac{1}{4}Mh^2(|\nabla f(x_n)|^2 + |\nabla V(x_n)|^2) \\ &+ 2\left\|D^2V\right\|_\infty (h^2|\nabla f(x_n)|^2 + \frac{1}{3}M^2h^4|\nabla f(x_n)|^2 + h^3d + 2hd) \\ &\leq (1 - \alpha h + \frac{1}{4}Mh^2c_V + 2\left\|D^2V\right\|_\infty h^2c_V + \frac{2}{3}c_V\left\|D^2V\right\|_\infty M^2h^4c_V)V(x_n) \\ &+ \beta h + 2d\left\|D^2V\right\|_\infty h^3 + 4d\left\|D^2V\right\|_\infty h \\ &\leq \hat{\alpha}V(x_n) + \hat{\beta},\end{aligned}$$

for some $\hat{\alpha} \in (0,1)$ and $\hat{\beta} \geq 0$ when h is small. ■

Once we have the Lyapunov condition, we can define the stopping time $\tau_C = \inf\{n > 0 : x_n \in C\}$ and show that $\sup_{x\in C}\mathbb{E}_x[\tau_C] \leq M_C < \infty$ for all small set C. Then uniqueness of stationary probability measure and ergodicity all follow by Theorem 1.3.1 in [MT12]. Next we prove that $\sup_{x\in C}\mathbb{E}_x[\tau_C] \leq M_C < \infty$ given Lyapunov condition. To do so, note that we have

$$\mathbb{E}_x[\tau_C] = \sum_{k=1}^{\infty} k\mathbb{P}(\tau_C = k) = \sum_{k\geq 1}\mathbb{P}(\tau_C > k-1).$$

Under Lyapunov condition, for any stopping time N, according to Lemma A.3 and Corollary A.4 in [MSH02], we have

$$\mathbb{P}(\tau_C > k-1) \leq \mathbb{E}[V(x_n)1_{\tau_C > k-1}] \leq \frac{\kappa[\gamma^{k-1}V(x_0) + 1]}{1-\gamma} \leq \kappa\gamma^{n-1}[V(x_0)+1],$$

for some $\gamma \in (\hat{\alpha}, 1)$ and constant κ. Therefore we have

$$\mathbb{E}_x[\tau_C] \leq \sum_{k\geq 1}\kappa\gamma^{n-1}[V(x_0)+1] = \frac{\kappa[V(x)+1]}{1-\gamma}$$

and

$$\sup_{x\in C}\mathbb{E}_x[\tau_C] \leq \frac{\kappa}{1-\gamma}\sup_{x\in C}V(x) + \frac{\kappa}{1-\gamma} \leq M_C < \infty.$$

So as a conclusion, the statement of the theorem follows. ■

Proof of Proposition 2. Consider that $x_n \sim \pi_h$ and $x_n^* \sim \pi$ are two independent random variables. Define x_{n+1} to be the one step RLMC result starting from x_n and $x_n^*(h)$ to be the solution of Langevin dynamics with initial value x_n^*. Therefore, $x_{n+1} \sim \pi_h$ and $x_n^*(h) \sim \pi$ are also independent and $\|x_n^* - x_n\|_{L^2} = \|x_n^*(h) - x_{n+1}\|_{L^2}$. We can compute the diffenrence between x_{n+1} and $x_n^*(h)$:

$$x_n^*(h) - x_{n+1} = (x_n^* - x_n) - \int_0^h \nabla f(x_n^*(s))ds + h\nabla f(x_n^*(\alpha_{n+1}h)) - h(-\nabla f(x_{n+\frac{1}{2}}) + \nabla f(x_n^*(\alpha_{n+1}h))).$$

It's easy to see that $\mathbb{E}_{\alpha_{n+1}}[\int_0^h \nabla f(x_n^*(s))ds - h\nabla f(x_n^*(\alpha_{n+1}h))] = 0$. And we can rewrite the last term as

$$\begin{aligned} h(-\nabla f(x_{n+\frac{1}{2}}) + \nabla f(x_n^*(\alpha_{n+1}h))) = {} & h(\nabla f(x_{n+\frac{1}{2}} + x_n^* - x_n) - \nabla f(x_{n+\frac{1}{2}})) \\ & + h\nabla f(x_n^* - \int_0^{\alpha_{n+1}h} \nabla f(x_n^*(s))ds + \sqrt{2}W_{\alpha_{n+1}h}) \\ & - h\nabla f(x_n^* - \alpha_{n+1}h\nabla f(x_n) + \sqrt{2\alpha_{n+1}h}U'_{n+1}). \end{aligned}$$

Take L_2-norm on other randomness, we have

$$\begin{aligned} & \|x_n^*(h) - x_{n+1}\|_{L^2} \\ \le {} & \left\|(x_n^* - x_n) - h(\nabla f(x_{n+\frac{1}{2}} + x_n^* - x_n) - \nabla f(x_{n+\frac{1}{2}}))\right\|_{L^2} \\ & + h\left\|\nabla f(x_n^* - \int_0^{\alpha_{n+1}h} \nabla f(x_n^*(s))ds + \sqrt{2}W_{\alpha_{n+1}h}) - \nabla f(x_n^* - \alpha_{n+1}h\nabla f(x_n) + \sqrt{2\alpha_{n+1}h}U'_{n+1})\right\|_{L^2} \\ & + \left\|\int_0^h \nabla f(x_n^*(s))ds - h\nabla f(x_n^*(\alpha_{n+1}h))\right\|_{L^2}. \end{aligned}$$

Since f is twice differentiable and f is also M-gradient Lipschitz and strongly convex with parameter m,

$$\left\|(x_n^* - x_n) - h(\nabla f(x_{n+\frac{1}{2}} + x_n^* - x_n) - \nabla f(x_{n+\frac{1}{2}}))\right\|_{L^2} \le \rho \left\|x_n^* - x_n\right\|_{L^2},$$

where $\rho = \max(1 - mh, Mh - 1) = 1 - mh$.

For the second term:

$$\begin{aligned}
&h\left\|\nabla f(x_n^* - \int_0^{\alpha_{n+1}h} \nabla f(x_n^*(s))ds + \sqrt{2}W_{\alpha_{n+1}h}) - \nabla f(x_n^* - \alpha_{n+1}h\nabla f(x_n) + \sqrt{2\alpha_{n+1}h}U'_{n+1})\right\|_{L^2} \\
&\le Mh\left\|\int_0^{\alpha_{n+1}h} \nabla f(x_n^*(s)) - \nabla f(x_n)ds\right\|_{L^2} \\
&\le \frac{\sqrt{3}}{3}M^2h^2\left\|x_n^* - x_n\right\|_{L^2} + \frac{\sqrt{3}}{3}M^2h^2 \sup_{0<s<h}\left\|x_n^*(s) - x_n^*\right\|_{L^2} \\
&\le \frac{\sqrt{3}}{3}M^2h^2\left\|x_n^* - x_n\right\|_{L^2} + \frac{\sqrt{3}}{3}M^2h^2(4h^2\left\|\nabla f(x_n^*)\right\|^2 + 8M^2dh^3 + 2dh)^{\frac{1}{2}} \\
&\le \frac{\sqrt{3}}{3}M^2h^2\left\|x_n^* - x_n\right\|_{L^2} + \frac{\sqrt{3}}{3}M^2h^2(2dh + 4Mdh^2 + 8M^2dh^3)^{\frac{1}{2}}.
\end{aligned}$$

For the third term:

$$\begin{aligned}
\left\|\int_0^h \nabla f(x_n^*(s))ds - h\nabla f(x_n^*(\alpha_{n+1}h))\right\|_{L^2} &= \{\mathbb{E}\mathbb{E}_{\alpha_{n+1}}[(\int_0^h \nabla f(x_n^*(s))ds - h\nabla f(x_n^*(\alpha_{n+1}h)))^2]\}^{\frac{1}{2}} \\
&= \{\mathbb{E}[h\int_0^h |\nabla f(x_n^*(s))|^2 ds - (\int_0^h \nabla f(x_n^*(s))ds)]\}^{\frac{1}{2}} \\
&= \{\mathbb{E}[h\int_0^h |\nabla f(x_n^*(s)) - \frac{1}{h}\int_0^h \nabla f(x_n^*(s'))ds'|^2 ds]\}^{1/2} \\
&\le \{\mathbb{E}[h\int_0^h \frac{1}{h}\int_0^h \left\|\nabla f(x_n^*(s)) - \nabla f(x_n^*(s'))\right\|^2 ds'ds]\}^{1/2} \\
&\le 2Mh\{\sup_{s\in(0,h)} \left\|x_n^*(s) - x_n^*\right\|^2\}^{1/2} \\
&\le 2Mh(4h^2\left\|\nabla f(x_n^*)\right\|^2 + 8M^2dh^3 + 2dh)^{1/2} \\
&\le 2Mh(2dh + 4Mdh^2 + 8M^2dh^3)^{\frac{1}{2}}.
\end{aligned}$$

Combine all the bounds:

$$\left\|x_n^* - x_n\right\|_{L^2} \le \frac{\frac{\sqrt{3}}{3}M^2h^2(2dh + 4Mdh^2 + 8M^2dh^3)^{\frac{1}{2}} + 2Mh(2dh + 4Mdh^2 + 8M^2dh^3)^{\frac{1}{2}}}{mh - \frac{\sqrt{3}}{3}M^2h^2}.$$

The final statement follows by the fact that $W_2(\pi, \pi_h) \le \left\|x_n - x_n^*\right\|_{L^2}$. ■

2.5.3. Proofs for Section 2.2.3.

PROOF OF PROPOSITION 3. From previous analysis, if we keep track of the coefficients in all those bounds and assume that $M\gamma_n \leq \frac{1}{2}$ for all n, we have:

$$\begin{aligned}
&\mathbb{E}[\|x_{n+1} - y_{n+1}\|^2] \\
&\leq (1 + m\gamma_{n+1})\mathbb{E}[\left\|y_{n+1} - x^*_{n+1}\right\|^2] + \frac{1}{m\gamma_{n+1}}\mathbb{E}\left\|\mathbb{E}_{\alpha_{n+1}} x_{n+1} - x^*_{n+1}\right\|^2 + \mathbb{E}\left\|x_{n+1} - x^*_{n+1}\right\|^2 \\
&\leq (1 + m\gamma_{n+1})e^{-2m\gamma_{n+1}}\mathbb{E}\left\|x_n - y_n\right\|^2 + \frac{1}{m\gamma_{n+1}}\mathbb{E}\left\|\mathbb{E}_{\alpha_{n+1}} x_{n+1} - x^*_{n+1}\right\|^2 + \mathbb{E}\left\|x_{n+1} - x^*_{n+1}\right\|^2 \\
&\leq (1 + m\gamma_{n+1})e^{-2m\gamma_{n+1}}\mathbb{E}\left\|x_n - y_n\right\|^2 + \frac{2\gamma^2_{n+1}}{m\gamma_{n+1}}\mathbb{E}\left\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n(\alpha_{n+1}\gamma_{n+1}))\right\|^2 \\
&\quad + 2\gamma^2_{n+1}\mathbb{E}\left\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n(\alpha_{n+1}\gamma_{n+1}))\right\|^2 + 2M^2\gamma^2_{n+1}\mathbb{E}\sup_{t\in[0,\gamma_{n+1}]}\left\|x_n(\alpha_{n+1}\gamma_{n+1}) - x_n(t)\right\|^2 \\
&\leq (1 + m\gamma_{n+1})e^{-2m\gamma_{n+1}}\mathbb{E}\left\|x_n - y_n\right\|^2 \\
&\quad + 2\gamma^4_{n+1}(1 + \frac{1}{m\gamma_{n+1}})M^4(\frac{1}{5}\gamma^2_{n+1}\left\|\nabla f(x_n)\right\|^2_{L^2} + \frac{1}{6}M^2 d\gamma^3_{n+1} + \frac{2}{3}d\gamma_{n+1}) \\
&\quad + 4M^2\gamma^2_{n+1}\left(\frac{4\gamma^2_{n+1}}{1 - 2M^2\gamma_{n+1}}\left\|\nabla f(x_n)\right\|^2_{L^2} + \frac{8M^2 d\gamma^3_{n+1}}{1 - 2M^2\gamma^2_{n+1}} + 4d\gamma_{n+1}\right) \\
&\leq (1 + m\gamma_{n+1})e^{-2m\gamma_{n+1}}\mathbb{E}\left\|x_n - y_n\right\|^2 + (33 + \kappa)M^2\gamma^4_{n+1}\left\|\nabla f(x_n)\right\|^2_{L^2} + (33 + \kappa)M^2 d\gamma^3_{n+1}.
\end{aligned}$$

We can further bound $\left\|\nabla f(x_n)\right\|^2_{L^2}$:

$$\begin{aligned}
\left\|\nabla f(x_n)\right\|^2_{L^2} &\leq 2\left\|\nabla f(y_n)\right\|^2_{L^2} + 2\left\|\nabla f(y_n) - \nabla f(x_n)\right\|^2_{L^2} \\
&\leq 2\left\|\nabla f(y_n)\right\|^2_{L^2} + 2M^2\left\|x_n - y_n\right\|^2_{L^2} \\
&\leq 2Md + 2M^2\left\|x_n - y_n\right\|^2_{L^2}.
\end{aligned}$$

Therefore we have the following iterative inequality:

$$\begin{aligned}
\mathbb{E}[\|x_{n+1} - y_{n+1}\|^2] &\leq (1 + m\gamma_{n+1})e^{-2m\gamma_{n+1}}\mathbb{E}\left\|x_n - y_n\right\|^2 + 2(33 + \kappa)M^2 d\gamma^3_{n+1} + 2(33 + \kappa)M^4\gamma^4_{n+1}\mathbb{E}\left\|x_n - y_n\right\|^2 \\
&\leq \left[1 - m\gamma_{n+1} + (\frac{m^2}{2} + \frac{M^2(33 + \kappa)}{2})\gamma^2_{n+1}\right]\mathbb{E}\left\|x_n - y_n\right\|^2 + 2(33 + \kappa)M^2 d\gamma^3_{n+1}.
\end{aligned}$$

Since (γ_n) is fast decreasing, we can assume that $\gamma_{n+1} \leq \frac{m}{m^2 + M^2(33+\kappa)} \leq \frac{1}{m+34M}$ for large n, and for those n we have

$$\mathbb{E}[\|x_{n+1} - y_{n+1}\|^2] \leq (1 - \frac{1}{2}m\gamma_{n+1})\mathbb{E}\left\|x_n - y_n\right\|^2 + 2(33 + \kappa)M^2 d\gamma^3_{n+1}.$$

Our strategy of choosing (γ_n): for the first K_1 steps, we choose constant step size $h = \frac{1}{m+34M}$, K_1 is the first time so that $\mathbb{E}[\|x_{K_1} - y_{K_1}\|^2] \leq 5\kappa(\kappa+33)M(\frac{d^{\frac{1}{2}}}{m+34M})^2$. such K_1 exists because

$$\begin{aligned}\mathbb{E}[\|x_{K_1} - y_{K_1}\|^2] &\leq (1 - \frac{m}{2m+68M})^{K_1}\mathbb{E}[\|x_0 - y_0\|^2] + \frac{2M^2(\kappa+33)d}{(m+34M)^3}\frac{2(m+34M)}{m} \\ &= (1 - \frac{m}{2m+68M})^{K_1}\mathbb{E}[\|x_0 - y_0\|^2] + 4\kappa(\kappa+33)M(\frac{d^{\frac{1}{2}}}{m+34M})^2.\end{aligned}$$

Claim: There exists $\lambda > 0$ such that if we choose $\gamma_{n+1} = \frac{1}{m+34M+\lambda(n-K_1)}$ for all $n \geq K_1$, we can ensure that $\mathbb{E}[\|x_k - y_k\|^2] \leq 5\kappa(\kappa+33)M(\frac{d^{\frac{1}{2}}}{m+34M+\lambda(n-K_1)})^2$ for all $n \geq K_1$.

Proof of Claim: Simply use induction:

$$\begin{aligned}\mathbb{E}[\|x_{n+1} - y_{n+1}\|^2] &\leq (1 - \frac{1}{2}m\gamma_{n+1})5\kappa(\kappa+33)Md\gamma_{n+1}^2 + 2M^2(\kappa+33)d\gamma_{n+1}^3 \\ &= 5\kappa(\kappa+33)Md\gamma_{n+1}^2(1 - \frac{m}{10}\gamma_{n+1}).\end{aligned}$$

Our goal is to ensure $5\kappa(\kappa+33)Md\gamma_{n+1}^2(1 - \frac{m}{10}\gamma_{n+1}) < 5\kappa(\kappa+33)M(\frac{d^{\frac{1}{2}}}{m+34M+\lambda(n+1-K_1)})^2$. It boils down to discuss the following polynomial inequality relates to λ:

$$G(\lambda) = (K - \frac{1}{10}m(K+1)^2)\lambda^2 + (X - \frac{1}{5}mX(K+1))\lambda - \frac{1}{10}mX^2 \leq 0,$$

where $X = m + 34M$ and $K = n - K_1 > 0$. It's not hard to see that there's always positive λ satisfying the inequality.

At last to get small error, we require $\mathbb{E}\|x_n - y_n\|^2 \leq \frac{d\epsilon^2}{m}$, i.e

$$5\kappa(\kappa+33)M\frac{d}{(m+34M+\lambda(n-K_1))^2} \leq \frac{d\epsilon^2}{m}.$$

Then we have

$$n \geq K_1 + \lambda^{-1}m^{\frac{1}{2}}M^{\frac{1}{2}}\kappa^{\frac{1}{2}}(\kappa+33)^{\frac{1}{2}}/\epsilon - \lambda^{-1}(m+34M) \sim O(\kappa^{\frac{3}{2}}/\epsilon).$$

■

2.5.4. Proof of Theorem 2. Before we prove Theorem 2, we need several intermediate results on the tightness of the (RLMC) chain.

Lemma 2.5.3. *Under Assumption 2.5.1, for every continuous function φ satisfying $\varphi(x) = o(V^k(x))$ for some $k \in \mathbb{N}$, $\lim_n \pi_n^\gamma(\varphi) = \pi(\varphi)$.*

Proof of Lemma 2.5.3. The proof is divided into three step:
1) For all $p \geq 1$, there exists $\tilde{\alpha} \in (0,1)$ and $\tilde{\beta}, n_0 \in \mathbb{N}$ such that $\mathbb{E}[V^p(x_{n+1})|\mathcal{F}_n] \leq V^p(x_n) + \gamma_{n+1}V^{p-1}(x_n)(\tilde{\beta} - \tilde{\alpha}V(x_n))$ for all $n \geq n_0$.
When $p = 1$, the statement follows from Assumption 2.5.1.
When $p > 1$, first we Taylor expand $V^p(x_{n+1})$ at x_n:

$$\begin{aligned}
V^p(x_{n+1}) &= V^p(x_n) + pV^{p-1}(x_n)\nabla V(x_n) \cdot (x_{n+1} - x_n) + \frac{1}{2}D^2(V^p)(\xi_{n+1})(x_{n+1} - x_n)^{\otimes 2} \\
&= V^p(x_n) - \gamma_{n+1}pV^{p-1}(x_n)\nabla V(x_n) \cdot \nabla f(x_{n+\frac{1}{2}}) + \sqrt{2\gamma_{n+1}}pV^{p-1}\nabla V(x_n) \cdot U_{n+1} \\
&\quad + \frac{1}{2}D^2(V^p)(\xi_{n+1})\left(-\gamma_{n+1}\nabla f(x_{n+\frac{1}{2}}) + \sqrt{2\gamma_{n+1}}U_{n+1}\right)^{\otimes 2} \\
&\leq V^p(x_n) - \gamma_{n+1}pV^{p-1}(x_n)\nabla V(x_n) \cdot \nabla f(x_{n+\frac{1}{2}}) + \sqrt{2\gamma_{n+1}}pV^{p-1}(x_n)\nabla V(x_n) \cdot U_{n+1} \\
&\quad + p\lambda_p V^{p-1}(\xi_{n+1})|-\gamma_{n+1}\nabla f(x_{n+\frac{1}{2}}) + \sqrt{2\gamma_{n+1}}U_{n+1}|^2,
\end{aligned}$$

where ξ_{n+1} is a point on the line segment joining x_n and x_{n+1} and $\lambda_p := \frac{1}{2}\lambda_{D^2V+(p-1)(\nabla V \otimes \nabla V)/V} < +\infty$. Due to $\nabla(\sqrt{V}) = \frac{\nabla V}{2V}$ and $|\nabla V|^2 \leq c_V V$, we have $\sqrt{V}$ is Lipschitz continuous and the Lipschitz constant $[\sqrt{V}]_1 = \frac{1}{4}c_V < +\infty$. Hence for a point ξ_{n+1} on the line segment between x_n and x_{n+1},

$$\begin{aligned}
V^{p-1}(\xi_{n+1}) &= (\sqrt{V})^{2(p-1)}(\xi_{n+1}) \leq \left(\sqrt{V}(x_n) + [\sqrt{V}]_1|x_{n+1} - x_n|\right)^{2(p-1)} \\
&\leq \begin{cases} V^{p-1}(x_n) + [\sqrt{V}]_1^{2(p-1)}|x_{n+1} - x_n|^{2(p-1)}, & 2(p-1) \leq 1 \\ V^{p-1}(x_n) + c\left(V^{(2p-3)/2}(x_n)|x_{n+1} - x_n| + |x_{n+1} - x_n|^{2(p-1)}\right), & 2(p-1) > 1 \end{cases}
\end{aligned}$$

We can further bound

$$\begin{aligned}
|x_{n+1} - x_n| &= |-\gamma_{n+1}\nabla f(x_n) + \sqrt{2\gamma_{n+1}}U_{n+1} - \gamma_{n+1}\left(\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n)\right)| \\
&\leq \gamma_{n+1}|\nabla f(x_n)| + \sqrt{2\gamma_{n+1}}|U_{n+1}| + M\gamma_{n+1}|-\tilde{\gamma}_{n+1}\nabla f(x_n) + \sqrt{2\alpha_{n+1}\gamma_{n+1}}U'_{n+1}| \\
&\leq \gamma_{n+1}(1 + M\alpha_{n+1}\gamma_{n+1})|\nabla f(x_n)| + \sqrt{2\gamma_{n+1}}|U_{n+1}| + \sqrt{2}M\gamma_{n+1}\alpha_{n+1}^{\frac{1}{2}}\gamma_{n+1}^{\frac{1}{2}}|U'_{n+1}| \\
&\leq C\sqrt{V}(x_n)\gamma_{n+1}^{\frac{1}{2}}(1 + |U_{n+1}| + |U'_{n+1}|).
\end{aligned}$$

Plug these results into the last term in the first inequality we obtained from Taylor expansion:

$$\begin{aligned}
p\lambda_p V^{p-1}(\xi_{n+1})|x_{n+1}-x_n|^2 &\le p\lambda_p V^{p-1}(x_n)|x_{n+1}-x_n|^2 \\
&\quad + Cp\lambda_p \begin{cases} |x_{n+1}-x_n|^{2p}, & 2p \le 3 \\ V^{(2p-3)/2}(x_n)|x_{n+1}-x_n|^3 + |x_{n+1}-x_n|^{2p}, & 2p > 3 \end{cases} \\
&\le p\lambda_p V^{p-1}(x_n)|x_{n+1}-x_n|^2 + C\gamma_{n+1}^{p\wedge\frac{3}{2}} V^p(x_n)(1+|U_{n+1}|^{2p}+|U'_{n+1}|^{2p}).
\end{aligned}$$

We then take conditional expectation, there exists $\alpha > 0$ and $\beta \ge 0$ such that for all $n \ge n_0$:

$$\begin{aligned}
\mathbb{E}[V^p(x_{n+1})|\mathcal{F}_n] &\le V^p(x_n) - pV^{p-1}(x_n)(\alpha V(x_n) - \beta) \\
&\quad - p\gamma_{n+1}V^{p-1}(x_n)\mathbb{E}[\nabla V(x_n)\cdot\left(\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n)\right)|\mathcal{F}_n] \\
&\quad + 2p\lambda_p V^{p-1}(x_n)\mathbb{E}[\gamma_{n+1}^2|\nabla f(x_{n+\frac{1}{2}})|^2 + 2\gamma_{n+1}|U_{n+1}|^2|\mathcal{F}_n] \\
&\quad + CV^p(x_n)(1+\mathbb{E}|U_{n+1}|^{2p}+\mathbb{E}|U'_{n+1}|^{2p})\gamma_{n+1}^{p\wedge\frac{3}{2}} \\
&\le V^p(x_n) - pV^{p-1}(x_n)(\alpha V(x_n)-\beta) + 2p\lambda_p\mathbb{E}|U_{n+1}|^2\gamma_{n+1}V^{p-1}(x_n) \\
&\quad + CV^p(x_n)(1+\mathbb{E}|U_{n+1}|^{2p}+\mathbb{E}|U'_{n+1}|^{2p})\gamma_{n+1}^{p\wedge\frac{3}{2}} \\
&\quad + c_V Mp\gamma_{n+1}^2 V^p(X_n) + \sqrt{2}c_V Mp\gamma_{n+1}^{\frac{3}{2}}\mathbb{E}|U'_{n+1}|V^{p-1/2}(x_n) \\
&\quad + c_V p\lambda_p\gamma_{n+1}^2 V^{p-1}(x_n)\mathbb{E}[V(x_{n+\frac{1}{2}})|\mathcal{F}_n].
\end{aligned}$$

From x_n to $x_{n+\frac{1}{2}}$, it's simply the Euler discretization with time step $\alpha_{n+1}\gamma_{n+1}$, we could use the result in [LP02]: there exists a $\bar{\alpha} > 0$ and $\bar{\beta} \in \mathbb{R}$ such that for all $n \ge n_0$:

$$\mathbb{E}[V(x_{n+\frac{1}{2}})|\mathcal{F}_n] \le V(x_n)(1-\bar{\alpha}\tilde{\gamma}_{n+1}) + \bar{\beta}\tilde{\gamma}_{n+1}.$$

Therefore we have

$$\mathbb{E}[V^{x_{n+1}}|\mathcal{F}_n] \le (1-\alpha p\gamma_{n+1}+o(\gamma_{n+1}))V^p(x_n)+\gamma_{n+1}V^{p-1}(x_n)(p\beta+2p\lambda_p\mathbb{E}|U_{n+1}|^2+c_V Mp\mathbb{E}|U'_{n+1}|^2).$$

There exists $\hat{\alpha} > 0$ and $\hat{\beta} \in \mathbb{R}$ such that for all $n \ge n_0$:

$$\mathbb{E}[V^p(x_{n+1})|\mathcal{F}_n] \le V^p(x_n) + \gamma_{n+1}V^{p-1}(x_n)\left(\hat{\beta} - \hat{\alpha}V(x_n)\right).$$

2) From step 1), we derive

$$\sup_{n\geq n_0} \mathbb{E}[V^p(x_n)] \leq (\frac{\hat{\beta}}{\hat{\alpha}})^p \vee \mathbb{E}[V^p(x_{n_0})].$$

Hence $\sup_n \mathbb{E}[V^p(x_n)] < +\infty$ for all $p \geq 1$. Therefore $\sup_n \pi_n^\gamma(\omega, V^p) < +\infty$ $\mathbb{P}$-a.s. for all $p \geq 1$.

3) Identification of the weak limit: To identify the limit, we essentially follow the same steps in [LP02] and hence we omit the proof.

(1) **(Echeverría-Weiss Theorem)** Let E be a locally compact Polish space and A a linear operator satisfying the positive maximum principle. Assume that its domain $\mathcal{D}(A)$ is an algebra everywhere dense in $(\mathcal{C}_0(E), \| \ \|_\infty)$ containing a sequence $(f_n)_{n\in\mathbb{N}}$ satisfying

$$\sup_{n\in\mathbb{N}} (\|f_n\|_\infty + \|\mathcal{L}f_n\|_\infty) < +\infty, \quad \forall x \in E, \quad f_n(x) \to 1 \ \text{ and } \ Af_n(x) \to 0.$$

If a distribution on $(E, \mathcal{B}(E))$ satisfies $\int_E Af d\nu = 0$ for every $f \in \mathcal{D}(A)$, then there exists a stationary solution for the martingale problem (A, ν) (this means that there exists a stationary continuous-time homogeneous Markov process with infinitesimal generator A and invariant distribution ν).

(2) The generator of the Langevin dynamics, $\mathcal{A}$, satisfies the assumptions of the Echeverría-Weiss theorem.

(3) Under assumption 2.5.1, for every bounded Lipschitz continuous function $\varphi : \mathbb{R}^d \to \mathbb{R}$, $\lim_n \frac{1}{\Gamma_n} \sum_{k=1}^n \mathbb{E}[\varphi(x_k) - \varphi(x_{k-1}) | \mathcal{F}_{k-1}] = 0$ $\mathbb{P}$-a.s.

(4) Under assumption 2.5.1, for every twice continuously differentiable function φ with compact support, $\lim_n \left(\frac{1}{\Gamma_n} \sum_{k=1}^n \mathbb{E}[\varphi(x_k) - \varphi(x_{k-1}) | \mathcal{F}_{k-1}] - \pi_n^\gamma(\mathcal{A}\varphi) \right) = 0$ $\mathbb{P}$-a.s.

$a), b), c), d)$ together imply that the weak limit of the empirical distribution π_n^γ is π, i.e the stationary distribution of the Langevin dynamics. ■

Proof of Theorem 2. Since f satisfies Assumption 2.1.1, we can show that the Langevin dynamics satisfies Assumption 2.5.1. Therefore lemma 2.5.3 is true. Then we may use the following method to discuss the CLT of (RLMC).

$$\begin{aligned}
x_k - x_{k-1} &= -\gamma_k \left(\nabla f(x_{k-1}) + D^2 f(x_{k-1})(x_{k-\frac{1}{2}} - x_{k-1}) + r_2(x_{k-\frac{1}{2}}, x_{k-1})\right) + \sqrt{2\gamma_k} U_k \\
&= -\gamma_k \nabla f(x_{k-1}) + \sqrt{2\gamma_k} U_k - \gamma_k D^2 f(x_{k-1})(x_{k-\frac{1}{2}} - x_{k-1}) - \gamma_k r_2(x_{k-\frac{1}{2}}, x_{k-1}) \\
&= -\gamma_k \nabla f(x_{k-1}) + \sqrt{2\gamma_k} U_k + \alpha_k \gamma_k^2 D^2 f(x_{k-1}) \nabla f(x_{k-1}) - \sqrt{2\alpha_k} \gamma_k^{\frac{3}{2}} \nabla^2 f(x_{k-1}) U_k' - \gamma_k r_2(x_{k-\frac{1}{2}}, x_{k-1})
\end{aligned}$$

where

$$\begin{aligned}
r_2(x_{k-\frac{1}{2}}, x_{k-1}) &= \nabla f(x_{k-\frac{1}{2}}) - \nabla f(x_{k-1}) - D^2 f(x_{k-1})(x_{k-1} - x_{k-\frac{1}{2}}) \\
&= \frac{1}{2} D^3 f(x_{k-1})(x_{k-\frac{1}{2}} - x_{k-1})^{\otimes 2} + \frac{1}{6} D^4 f(x_{k-1})(x_{k-\frac{1}{2}} - x_{k-1})^{\otimes 3} + O(\gamma_k^2) \\
&= \alpha_k \gamma_k D^3 f(x_{k-1}) U_k'^{\otimes 2} - \sqrt{2} \alpha_k^{\frac{3}{2}} \gamma_k^{\frac{3}{2}} \langle D^3 f(x_{k-1}); \nabla f(x_{k-1}), U_k' \rangle \\
&\quad + \frac{\sqrt{2}}{3} \alpha_k^{\frac{3}{2}} D^4 f(x_{k-1}) U_k'^{\otimes 4} + O(\gamma_k^2).
\end{aligned}$$

Then

$$\begin{aligned}
x_k - x_{k-1} = -\gamma_k \nabla f(x_{k-1}) &+ \sqrt{2\gamma_k} U_k - \sqrt{2\alpha_k} \gamma_k^{\frac{3}{2}} \nabla^2 f(x_{k-1}) U_k' \\
&+ \alpha_k \gamma_k^2 D^2 f(x_{k-1}) \nabla f(x_{k-1}) - \alpha_k \gamma_k^2 D^3 f(x_{k-1}) U_k'^{\otimes 2} + O(\gamma_k^{\frac{5}{2}}).
\end{aligned}$$

We can decompose $\phi(x_k)$:

$$\begin{aligned}
\phi(x_k) - \phi(x_{k-1}) &= \nabla \phi(x_{k-1})(x_k - x_{k-1}) + \frac{1}{2} D^2 \phi(x_{k-1})(x_k - x_{k-1})^{\otimes 2} + \frac{1}{6} D^3 \phi(x_{k-1})(x_k - x_{k-1})^{\otimes 3} \\
&\quad + \frac{1}{24} D^4 \phi(x_{k-1})(x_k - x_{k-1})^{\otimes 4} + O(\gamma_k^{\frac{5}{2}}) \\
&= \nabla \phi(x_{k-1})(\sqrt{2} \gamma_k^{\frac{1}{2}} U_k - \gamma_k \nabla f(x_{k-1}) - \sqrt{2\alpha_k} \gamma_k^{\frac{3}{2}} D^2 f(x_{k-1}) U_k' \\
&\quad + \alpha_k \gamma_k^2 D^2 f(x_{k-1}) \nabla f(x_{k-1}) - \alpha_k \gamma_k^2 D^3 f(x_{k-1}) U_k'^{\otimes 2}) \\
&\quad + \frac{1}{2} D^2 \phi(x_{k-1}) \left(\sqrt{2} \gamma_k^{\frac{1}{2}} U_k - \gamma_k \nabla f(x_{k-1}) - \sqrt{2\alpha_k} \gamma_k^{\frac{3}{2}} D^2 f(x_{k-1}) U_k'\right)^{\otimes 2} \\
&\quad + \frac{1}{6} D^3 \phi(x_{k-1}) \left(\sqrt{2} \gamma_k^{\frac{1}{2}} U_k - \gamma_k \nabla f(x_{k-1})\right)^{\otimes 3} + \frac{1}{24} D^4 \phi(x_{k-1})(\sqrt{2} \gamma_k^{\frac{1}{2}} U_k)^{\otimes 4} + O(\gamma_k^{\frac{5}{2}}).
\end{aligned}$$

If $\mathcal{A}$ is the generator of Langevin dynamics and summing up over k:

$$\begin{aligned}\sum_{k=1}^n \gamma_k \mathcal{A}\phi(x_{k-1}) &= \phi(x_n) - \phi(x_0) - \sqrt{2}\sum_{k=1}^n \gamma_k^{\frac{1}{2}} \nabla\phi(x_{k-1})U_k - \sum_{k=1}^n \gamma_k \left(D^2\phi(x_{k-1})U_k^{\otimes 2} - \mathbb{E}[D^2\phi(x_{k-1})U_k^{\otimes 2}|\mathcal{F}_{k-1}]\right) \\ &\quad + \sqrt{2}\sum_{k=1}^n \gamma_k^{\frac{3}{2}} \langle D^2\phi(x_{k-1}); \nabla f(x_{k-1}), U_k\rangle - \frac{\sqrt{2}}{3}\sum_{k=1}^n \gamma_k^{\frac{3}{2}} D^3\phi(x_{k-1})U_k^{\otimes 3} \\ &\quad + \sum_{k=1}^n \sqrt{2\alpha_k}\gamma_k^{\frac{3}{2}} \langle D^2 f(x_{k-1}); \nabla\phi(x_{k-1}), U_k'\rangle + \sum_{k=1}^n \gamma_k^2 \langle D^3\phi(x_{k-1}); \nabla f(x_{k-1}), U_k^{\otimes 2}\rangle \\ &\quad - \sum_{k=1}^n \alpha_k\gamma_k^2 \langle D^2 f(x_{k-1}); \nabla\phi(x_{k-1}), \nabla f(x_{k-1})\rangle + \sum_{k=1}^n \alpha_k\gamma_k^2 \langle D^3 f(x_{k-1}); \nabla\phi(x_{k-1}), U_k'^{\otimes 2}\rangle \\ &\quad - \frac{1}{2}\sum_{k=1}^n \gamma_k^2 D^2\phi(x_{k-1})\nabla f(x_{k-1})^{\otimes 2} + \sum_{k=1}^n 2\alpha_k^{\frac{1}{2}}\gamma_k^2 \langle D^2\phi(x_{k-1}); D^2\phi(x_{k-1})U_k', U_k\rangle \\ &\quad - \frac{1}{6}\sum_{k=1}^n \gamma_k^2 D^4\phi(x_{k-1})U_k^{\otimes 4} + \sum_{k=1}^n O(\gamma_k^{\frac{5}{2}}) \\ &:= N_n^{(0)} + N_n^{(\frac{1}{2})} + N_n^{(1)} + N_n^{(\frac{3}{2})} + N_n^{(2)} + N_n^{(\frac{5}{2})}.\end{aligned}$$

In the fast decreasing time step situation($\sum_{k=1}^n \gamma_k^2/\sqrt{\Gamma_n} \to 0$), the CLT for (RLMC) is the same as that of LMC. In the slowly decreasing time step situation, when $\sum_{k=1}^n \gamma_k^2/\sqrt{\Gamma_n} \to \hat{\gamma} \in (0, +\infty]$:

a) $\frac{\phi(x_n)-\phi(x_0)}{\Gamma_n^{(2)}} \to 0$ because (x_n) is tight and ϕ is continuous.

b) $\frac{-\sqrt{2}\sum_{k=1}^n \gamma_k^{\frac{1}{2}} \nabla\phi(x_{k-1})U_k}{\sqrt{\Gamma_n}} \implies \mathcal{N}(0, 2\int_{\mathbb{R}^d} |\nabla\phi(x)|^2\pi(dx))$. Therefore,

$$\frac{-\sqrt{2}\sum_{k=1}^n \gamma_k^{\frac{1}{2}} \nabla\phi(x_{k-1})U_k}{\Gamma_n^{(2)}} \implies \begin{cases} \mathcal{N}(0, 2\hat{\gamma}^{-2}\displaystyle\int_{\mathbb{R}^d} |\nabla\phi(x)|^2\pi(dx)), & \text{when } \hat{\gamma} < +\infty \\ \qquad\qquad 0 \qquad , & \text{when } \hat{\gamma} = +\infty \end{cases}$$

c) $\frac{-\sum_{k=1}^n \gamma_k\left(D^2\phi(x_{k-1})U_k^{\otimes 2} - \mathbb{E}[D^2\phi(x_{k-1})U_k^{\otimes 2}|\mathcal{F}_{k-1}]\right)}{\sqrt{\Gamma_n}} \to 0$ in L^2.

d) $\frac{\sqrt{2}\sum_{k=1}^n \gamma_k^{\frac{3}{2}} \langle D^2\phi(x_{k-1});\nabla f(x_{k-1}),U_k\rangle}{\sqrt{\Gamma_n}} \to 0$ in L^2.
$\frac{-\frac{\sqrt{2}}{3}\sum_{k=1}^n \gamma_k^{\frac{3}{2}} D^3\phi(x_{k-1})U_k^{\otimes 3}}{\sqrt{\Gamma_n}} \to 0$ in probability because $\mathbb{E}[U_k^{\otimes 3}] = 0$.
$\frac{\sum_{k=1}^n \sqrt{2\alpha_k}\gamma_k^{\frac{3}{2}} \langle D^2 f(x_{k-1});\nabla\phi(x_{k-1}),U_k'\rangle}{\sqrt{\Gamma_n}} \to 0$ in L^2.
Therefore $\frac{N_n^{(\frac{3}{2})}}{\Gamma_n^{(2)}} \to 0$ in probability.

e) $\frac{\sum_{k=1}^{n}\gamma_k^2\langle D^3\phi(x_{k-1});\nabla f(x_{k-1}),U_k^{\otimes 2}\rangle}{\Gamma_n^{(2)}} \to \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\langle D^3\phi(x);\nabla f(x),u^{\otimes 2}\rangle\mu(du)\pi(dx)$ in probability.
$\frac{-\sum_{k=1}^{n}\alpha_k\gamma_k^2\langle D^2 f(x_{k-1});\nabla\phi(x_{k-1}),\nabla f(x_{k-1})\rangle}{\Gamma_n^{(2)}} \to -\frac{1}{2}\int_{\mathbb{R}^d}\langle D^2 f(x);\nabla\phi(x),\nabla f(x)\rangle\pi(dx)$ in probability.
$\frac{\sum_{k=1}^{n}\alpha_k\gamma_k^2\langle D^3 f(x_{k-1});\nabla\phi(x_{k-1}),U_k'^{\otimes 2}\rangle}{\Gamma_n^{(2)}} \to \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\langle D^3 f(x);\nabla\phi(x),u^{\otimes 2}\rangle\mu(du)\pi(dx)$ in probability.
$\frac{-\frac{1}{2}\sum_{k=1}^{n}\gamma_k^2 D^2\phi(x_{k-1})\nabla f(x_{k-1})^{\otimes 2}}{\Gamma_n^{(2)}} \to -\frac{1}{2}\int_{\mathbb{R}^d} D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx)$ in probability.
$\frac{\sum_{k=1}^{n}2\alpha_k^{\frac{1}{2}}\gamma_k^2\langle D^2\phi(x_{k-1});D^2\phi(x_{k-1})U_k',U_k\rangle}{\Gamma_n^{(2)}} \to \int_{\mathbb{R}^d}\int_0^1\int_{\mathbb{R}^{2d}}2\alpha^{\frac{1}{2}}\langle D^2\phi(x),D^2\phi(x)u',u\rangle\mu_\alpha(du,du')d\alpha\pi(dx)$ in probability, where $\mu_\alpha(du,du')$ is the joint measure of (U_n,U_n') for all n conditioned on $\alpha_n=\alpha$. With some calculation, we can simplify the limit as $\int_{\mathbb{R}^d}trace(D^2\phi(x)^2)\pi(dx)$. Note that in deriving the above limit, we used the fact that the cross-covariance matrix between (U_n) and (U_n') is $\sqrt{\alpha_n}I_d$.
$\frac{-\frac{1}{6}\sum_{k=1}^{n}\gamma_k^2 D^4\phi(x_{k-1})U_k^{\otimes 4}}{\Gamma_n^{(2)}} \to -\frac{1}{6}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}D^4\phi(x)u^{\otimes 4}\mu(du)\pi(dx)$ in probability.
Therefore

$$\frac{N_n^{(2)}}{\Gamma_n^{(2)}} \to \varrho \qquad \text{in probability,}$$

where

$$\begin{aligned}\varrho = &\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\langle D^3\phi(x);\nabla f(x),u^{\otimes 2}\rangle\mu(du)\pi(dx) - \frac{1}{2}\int_{\mathbb{R}^d}\langle D^2 f(x);\nabla\phi(x),\nabla f(x)\rangle\pi(dx)\\ &+\frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^2}\langle D^3 f(x);\nabla\phi(x),u^{\otimes 2}\rangle\mu(du)\pi(dx) - \frac{1}{2}\int_{\mathbb{R}^d}D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx)\\ &+\int_{\mathbb{R}^d}trace(D^2\phi(x)^2)\pi(dx) - \frac{1}{6}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}D^4\phi(x)u^{\otimes 4}\mu(du)\pi(dx)\end{aligned}$$

and μ is the distribution for a d-dimensional standard Gaussian random variable.

f) $\frac{N_n^{\frac{5}{2}}}{\Gamma_n^{(2)}} \to 0$ in L^1.

As a conclusion, we obtain the proof of part (1) of the theorem:

$$\frac{\sum_{k=1}^{n}\gamma_k\mathcal{A}\phi(x_{k-1})}{\Gamma_n^{(2)}} \to \begin{cases}\mathcal{N}(\varrho, 2\hat{\gamma}^{-2}\int_{\mathbb{R}^d}|\nabla\phi(x)|^2\pi(dx)), & \text{when } \hat{\gamma}<+\infty\\ \qquad\qquad\varrho \quad , & \text{when } \hat{\gamma}=+\infty\end{cases}$$

For the fast decreasing step, i.e., part (2) of the theorem, the proof follows by the same arguments in the corresponding part of Theorem 10 in [LP02] and hence we omit it. ■

2.6. Proofs for Section 2.3

In this section, we would denote the drift function that appears in 2.4 as $b(x,v)$, i.e.

$$b(x,v) = \begin{bmatrix} v \\ -2v - u\nabla f(x) \end{bmatrix}.$$

ASSUMPTION 2.6.1. *There exists a twice differentiable function $V: \mathbb{R}^{2d} \to [1,\infty)$ such that: (0) $\lim_{\|(x,v)\|\to\infty} V(x,v) = +\infty$, (1) there exists $\alpha > 0$ and $\beta > 0$: $\langle \nabla V(x,v), b(x,v)\rangle \leq -\alpha V(x,v) + \beta$ for every (x,v), (2) there exists $c_V > 0$: $\|\nabla V(x,v)\|^2 + \|b(x,v)\|^2 \leq c_V V(x,v)$ for every (x,v), and (3) $\left\|D^2 V\right\|_\infty := \sup_{(x,v)\in\mathbb{R}^{2d}} \|D^2 V\|_{op} < \infty$.*

LEMMA 2.6.1. *Assumption 2.1.1 implies Assumption 2.6.1 when $u \in (0, \frac{4}{2M-m})$.*

PROOF OF LEMMA 2.6.1. For simplicity, We choose $V(x,v) = \|x - x_*\|^2 + \|x - x_* + v\|^1 + 1$ with $f(x_*) = \min f(x)$. Now we check conditions 0), 1), 2), 3) in $(\mathcal{L}_{V,\infty})$ are satisfied.

0) It's obvious that $\lim_{|(x,v)|\to+\infty} V(x,v) = +\infty$ and $V(x,v) \geq 1$ for all $(x,v) \in \mathbb{R}^d$.

3) The Hessian of V we choose is

$$D^2 V(x,v) = \begin{bmatrix} 4I_d & 2I_d \\ 2I_d & 2I_d \end{bmatrix}.$$

For arbitrary $(x,v)^T, (y,w)^T \in \mathbb{R}^{2d}$:

$$\left\|D^2 V(x,v)(y,w)^T\right\|^2 = \left\| \begin{bmatrix} 4y + 2w \\ 2y + 2w \end{bmatrix} \right\|^2 \leq 40 \left\|(y,w)^T\right\|^2.$$

Therefore $\left\|D^2V\right\|_\infty < \infty$.

2) Take gradient of the V we choose:

$$\nabla V(x,v) = \begin{bmatrix} 2(x-x_*) + 2(x-x_*+v) \\ 2(x-x_*+v) \end{bmatrix}.$$

Then for all $(x,v) \in \mathbb{R}^{2d}$,

$$\begin{aligned}
|\nabla V(x,v)|^2 + |b(x,v)|^2 &\leq 2(4\left\|x-x_*\right\|^2 + 4\left\|x-x_*+v\right\|^2) + 4\left\|x-x_*+v\right\|^2 \\
&\quad + \left\|v\right\|^2 + 2(4\left\|v\right\|^2 + u^2\left\|\nabla f(x)\right\|^2) \\
&\leq 8\left\|x-x_*\right\|^2 + 12\left\|x-x_*+v\right\|^2 + 9\left\|v\right\|^2 + 2u^2M^2\left\|x-x_*\right\|^2 \\
&\leq \max\{26+2u^2M^2, 30\}V(x,v).
\end{aligned}$$

1) Last we consider

$$\begin{aligned}
\langle \nabla V(x,v), b(x,v)\rangle &= 2(x-x_*)\cdot v + 2(x-x_*+v)\cdot v - 4(x-x_*+v)\cdot v \\
&\quad - 2u(x-x_*+v)\cdot\nabla f(x) \\
&\leq -2\left\|v\right\|^2 - 2u\left[f(x) - f(x_*-v) + \frac{m}{2}\left\|x-x_*+v\right\|^2\right] \\
&\leq -2\left\|v\right\|^2 - um\left\|x-x_*+v\right\|^2 - 2u\left(f(x_*) + \frac{m}{2}\left\|x-x_*\right\|^2\right) \\
&\quad + 2u\left(f(x_*) + \frac{M}{2}\left\|v\right\|^2\right) \\
&= -um\left\|x-x_*+v\right\|^2 - um\left\|x-x_*\right\|^2 - (2-uM)\left\|v\right\|^2.
\end{aligned}$$

The second inequality follows from the fact that f is m-strongly convex.

When $u \in (0, \frac{2}{M}]$, $\langle \nabla V(x,v), b(x,v)\rangle \leq -umV(x,v) + um$ for all $(x,v) \in \mathbb{R}^{2d}$. Therefore 1) is satisfied.

When $u > \frac{2}{M}$, we can use triangle inequality to further bound our result:

$$\begin{aligned}\langle \nabla V(x,v), b(x,v)\rangle &\leq -um \left\|x - x_* + v\right\|^2 - um \left\|x - x_*\right\|^2 + (uM-2)\left\|v\right\|^2 \\ &\leq [-um + 2(uM-2)](\|x - x_* + v\|^2 + \|x - x_*\|^2) \\ &\leq -[4 - u(2M-m)]V(x,v) - [4 - u(2M-m)].\end{aligned}$$

When $u \in (\frac{2}{M}, \frac{4}{2M-m})$, 1) is satisfied because $4 - u(2M-m) > 0$. Therefore, 1) holds when $u \in (0, \frac{4}{2M-m})$. ■

Remark 9. For the $V(x,v)$ we choose in the proof, under assumption 2.1.1, we can verify that: $V(x,v) = O(|x|^2 + |v|^2)$ when $|(x,v)| \to +\infty$. We will use this fact later in the proof when we establish the CLT statement.

2.6.1. Proofs for Section 2.3.1.

Proof of Theorem 3. Under the assumption 2.6.1, we can show that the following Lyapunov condition is satisfied for small h.

(Lyapunov Condition): There exists a function $V: \mathbb{R}^{2d} \to [1,\infty)$ such that:

0) $\lim_{|(x,v)|\to\infty} V(x,v) = +\infty$,

1) There exists $\hat{\alpha} \in (0,1)$ and $\hat{\beta} \geq 0$: $\mathbb{E}[V(x_{n+1}, v_{n+1})|\mathcal{F}_n] \leq \hat{\alpha} V(x_n, v_n) + \hat{\beta}$.

Proof: To show that assumption 2.6.1 implies Lyapunov condition, we first do Taylor expansion of $V(x_{n+1}, v_{n+1})$ at (x_n, v_n):

$$V(x_{n+1}, v_{n+1}) = V(x_n, v_n) + \nabla V(x_n, v_n) \cdot (x_{n+1} - x_n, v_{n+1} - v_n)^T + \frac{1}{2} D^2 V(\theta_n)[(x_{n+1} - x_n, v_{n+1} - v_n)^T]^{\otimes 2}$$

where θ_n is a random point on the line segment joining (x_n, v_n) and (x_{n+1}, v_{n+1}). Use the RULMC algorithm and part (a) of Assumption 2.1.1:

$$\begin{aligned}
\mathbb{E}[V(x_{n+1}, v_{n+1})|\mathcal{F}_n] \leq V(x_n, v_n) + \nabla V(x_n, v_n) \cdot &\begin{bmatrix} \frac{1-e^{-2h}}{2} v_n - \frac{u}{2}(h - \frac{1-e^{-2h}}{2})\nabla f(x_n) \\ -2\frac{1-e^{-2h}}{2} v_n - u\frac{1-e^{-2h}}{2} \nabla f(x_n) \end{bmatrix} \\
- \nabla V(x_n . v_n) \cdot &\begin{bmatrix} \frac{u}{2}(h - \frac{1-e^{-2h}}{2})\mathbb{E}[\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n)|\mathcal{F}_n] \\ u\frac{1-e^{-2h}}{2}\mathbb{E}[\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n)|\mathcal{F}_n] \end{bmatrix} \\
+ \frac{3M}{2} &\left[5(\frac{1-e^{-2h}}{2})^2 |v_n|^2 + u^2h^2 |\nabla f(x_n)^2| + ({\sigma_{n+1}^{(2)}}^2 + 4{\sigma_{n+1}^{(3)}}^2)ud \right] \\
+ \frac{3M}{2} &u^2h^2 \mathbb{E}[|\nabla f(x_{n+\frac{1}{2}})|^2 - |\nabla f(x_n)|^2 |\mathcal{F}_n],
\end{aligned}$$

where we can further estimate

$$\begin{aligned}
\mathbb{E}[\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n)|\mathcal{F}_n] &\leq M\mathbb{E}[x_{n+\frac{1}{2}} - x_n|\mathcal{F}_n] \\
&\leq M\frac{1}{2h}(h - \frac{1-e^{-2h}}{2})|v_n| + \sqrt{ud}M\sigma_{n+1}^{(1)} \\
&\quad + \frac{u}{2}(\frac{h}{2} - \frac{h - \frac{1-e^{-2h}}{2}}{2h})|\nabla f(x_n)|,
\end{aligned}$$

and there exists ξ_n such that $|\nabla f(x_{n+\frac{1}{2}})|^2 - |\nabla f(x_n)|^2 = 2(x_{n+\frac{1}{2}} - x_n)^T D^2 f(\xi_n) \nabla f(\xi_n)$ and ξ_n is on the line segment joining x_n and $x_{n+\frac{1}{2}}$. Therefore $|\xi_n - x_n| \leq |x_{n+\frac{1}{2}} - x_n|$. then we have

$$\begin{aligned}
\mathbb{E}[|\nabla f(x_{n+\frac{1}{2}})|^2 - |\nabla f(x_n)|^2|\mathcal{F}_n] &\leq 2M\mathbb{E}[|\nabla f(\xi_n)||x_{n+\frac{1}{2}} - x_n||\mathcal{F}_n] \\
&\leq 2M|\nabla f(x_n)|\mathbb{E}[|x_{n+\frac{1}{2}} - x_n||\mathcal{F}_n] + 2M^2\mathbb{E}[|x_{n+\frac{1}{2}} - x_n|^2|\mathcal{F}_n] \\
&\leq |\nabla f(x_n)|^2 + 3M^2\mathbb{E}[|x_{n+\frac{1}{2}} - x_n|^2|\mathcal{F}_n] \\
&\leq |\nabla f(x_n)|^2 + 6M^2(\frac{h^2}{3}|v_n|^2 + \frac{u^2h^4}{20}|\nabla f(x_n)|^2 + ud{\sigma_{n+1}^{(1)}}^2).
\end{aligned}$$

When h is small, we can use polynomials of h to bound those exponential coefficients. We can obtain that there exists $C > 0$:

$$\mathbb{E}[V(x_{n+1}, v_{n+1})|\mathcal{F}_n] \leq V(x_n, v_n) + h\nabla V(x_n, v_n) \cdot b(x_n, v_n)^T + Ch^2(d + |v_n|^2 + |\nabla f(x_n)|^2).$$

then assumption 2.6.1 implies that there exists $\alpha > 0, \beta > 0$ such that

$$\mathbb{E}[V(x_{n+1}, v_{n+1})|\mathcal{F}_n] \leq (1 - \alpha h + Cc_V h^2)V(x_n, v_n) + Ch^2 d + \beta.$$

When h is small, there exists $\hat{\alpha} = 1 - \alpha h + Cc_V h^2 \in (0, 1)$ and $\hat{\beta} = Ch^2 d + \beta > 0$ such that $\mathbb{E}[V(x_{n+1}, v_{n+1})|\mathcal{F}_n] \leq \hat{\alpha} V(x_n, v_n) + \hat{\beta}$. ■

Once we have the Lyapunov condition, we can define the stopping time $\tau_C = \inf\{n > 0 : (x_n, v_n) \in C\}$ and show that $\sup_{(x,v)\in C} \mathbb{E}_{(x,v)}[\tau_C] \leq M_C < \infty$ for all small set C. Then uniqueness of stationary probability measure and ergodicity all follow by Theorem 1.3.1 in [MT12]. Next we prove that $\sup_{(x,v)\in C} \mathbb{E}_{(x,v)}[\tau_C] \leq M_C < \infty$ given Lyapunov condition. To do so, note that we have

$$\mathbb{E}_{(x,v)}[\tau_C] = \sum_{n=1}^{\infty} n\mathbb{P}(\tau_C = n) = \sum_{n\geq 1} \mathbb{P}(\tau_C > n - 1).$$

Under Lyapunov condition, for any stopping time N, according to Lemma A.3 and Corollary A.4 in [MSH02], we have

$$\mathbb{P}(\tau_C > n - 1) \leq \mathbb{E}[V(x_n, v_n)1_{\tau_C > n-1}] \leq \frac{\kappa[\gamma^{n-1} V(x_0, v_0) + 1]}{1 - \gamma} \leq \kappa\gamma^{n-1}[V(x_0, v_0) + 1]$$

for some $\gamma \in (\hat{\alpha}, 1)$ and constant κ. Therefore, we have

$$\mathbb{E}_{(x,v)}[\tau_C] \leq \sum_{k\geq 1} \kappa\gamma^{n-1}[V(x_0, v_0) + 1] = \frac{\kappa[V(x, v) + 1]}{1 - \gamma},$$

and

$$\sup_{(x,v)\in C} \mathbb{E}_{(x,v)}[\tau_C] \leq \frac{\kappa}{1 - \gamma} \sup_{(x,v)\in C} V(x, v) + \frac{\kappa}{1 - \gamma} \leq M_C < \infty.$$

So as a conclusion, the statement of the theorem follows. ■

Before proving Proposition 4, we require some preliminary estimtes from [SL19], that we present below. First, let (y_n, w_n) be the solution of Underdamped Langevin dynamics evaluated at $t = \sum_{k=1}^{n} \gamma_k$ with initial value (y_0, w_0). (x_n, v_n) is the nth iterates in the (RULMC) algorithm with initial value (x_0, v_0). $(x_n^*(t), v_n^*(t))$ is the solution of Underdamped Langevin dynamics with initial value (x_{n-1}, v_{n-1}) and $(x_n^*, v_n^*) = (x_{n-1}^*(\gamma_n), v_{n-1}^*(\gamma_n))$. Then, we have the following results from

Lemma 2 in [SL19]. When $\gamma_{n+1} < \frac{1}{2}$ and $u = \frac{1}{M}$, we have:

$$\mathbb{E}\left\|\mathbb{E}_\alpha x_{n+1} - x^*_{n+1}\right\|^2 \leq 45(\gamma^{10}_{n+1}\mathbb{E}\left\|v_n\right\|^2 + M^{-2}\gamma^{12}_{n+1}\mathbb{E}\left\|\nabla f(x_n)\right\|^2 + M^{-1}d\gamma^{11}_{n+1}),$$

$$\mathbb{E}\left\|x_{n+1} - x^*_{n+1}\right\|^2 \leq 1800(\gamma^{6}_{n+1}\mathbb{E}\left\|v_n\right\|^2 + M^{-2}\gamma^{4}_{n+1}\mathbb{E}\left\|\nabla f(x_n)\right\|^2 + M^{-1}d\gamma^{7}_{n+1}),$$

$$\mathbb{E}\left\|\mathbb{E}_\alpha v_{n+1} - v^*_{n+1}\right\|^2 \leq 45(\gamma^{8}_{n+1}\mathbb{E}\left\|v_n\right\|^2 + M^{-2}\gamma^{10}_{n+1}\mathbb{E}\left\|\nabla f(x_n)\right\|^2 + M^{-1}d\gamma^{9}_{n+1}),$$

$$\mathbb{E}\left\|v_{n+1} - v^*_{n+1}\right\|^2 \leq 1300(\gamma^{4}_{n+1}\mathbb{E}\left\|v_n\right\|^2 + M^{-2}\gamma^{4}_{n+1}\mathbb{E}\left\|\nabla f(x_n)\right\|^2 + M^{-1}d\gamma^{5}_{n+1}).$$

Proof of Proposition 4 . Denote $A_n^2 = \mathbb{E}[\|x_n - y_n\|^2 + \|(x_n + v_n) - (y_n + w_n)\|^2]$. Using triangle inequality we have

$$\begin{aligned}\mathbb{E}_\alpha[\|x_n - y_n\|^2 + \|(x_n + v_n) - (y_n + w_n)\|^2] \leq (1 + \frac{h}{2\kappa})(\|x^*_k - y_n\|^2 + \|(x^*_k + v^*_k) - (y_n + w_n)\|^2) \\ + \frac{2\kappa}{h}(\|\mathbb{E}_\alpha[x_n] - x^*_k\|^2 + \|\mathbb{E}_\alpha[x_n + v_n] - (x^*_n + v^*_n)\|^2) \\ + \mathbb{E}_\alpha\|x_n - x^*_n\|^2 + \mathbb{E}_\alpha\|(x_n + v_n) - (x^*_n + v^*_n)\|^2 .\end{aligned}$$

Furthermore, we can take expectation on ω and use the contraction of Underdamped Langevin dynamics:

$$\begin{aligned}A_n^2 &\leq (1 + \frac{h}{2\kappa})e^{-\frac{h}{\kappa}}A^2_{n-1} + \frac{2\kappa}{h}(\mathbb{E}\|\mathbb{E}_\alpha x_n - x^*_n\|^2 + \mathbb{E}\|\mathbb{E}_\alpha[x_n + v_n] - (x^*_n + v^*_n)\|^2) \\ &\quad + \mathbb{E}\|x^*_n - x_n\|^2 + \mathbb{E}\|(x_n + v_n) - (x^*_n + v^*_n)\|^2 \\ &\leq e^{-\frac{h}{2\kappa}}A^2_{n-1} + \frac{2\kappa}{h}(3\mathbb{E}\|\mathbb{E}_\alpha x_n - x^*_n\|^2 + 2\mathbb{E}\|\mathbb{E}_\alpha v_n - v^*_n\|^2) \\ &\quad + 3\mathbb{E}\|x_n - x^*_n\|^2 + 2\mathbb{E}\|v_n - v^*_n\|^2 .\end{aligned}$$

When $h < \frac{1}{2}$, $u = \frac{1}{M}$ and $m = 1$:

$$A_n^2 \leq e^{-\frac{h}{2\kappa}}A^2_{n-1} + 8250\left[(\kappa h^7 + h^4)\mathbb{E}\|v_{n-1}\|^2 + (\kappa^{-1}h^8 + \kappa^{-2}h^4)\mathbb{E}\|\nabla f(x_{n-1})\|^2 + (\kappa^{-1}h^5 + h^7)\right].$$

Our next step is to bound $\mathbb{E}\left\|v_{n-1}\right\|^2$ and $\mathbb{E}\left\|\nabla f(x_{n-1})\right\|^2$. First for Underdamped Langevin dynamics with f satisfying Assumption 2.1.1, it's easy to compute that:

$$\begin{aligned}
\mathbb{E}\left\|w_{n-1}\right\|^2 &= d/M, \\
\mathbb{E}\left\|\nabla f(y_{n-1})\right\|^2 &= \frac{1}{\int e^{-f(x)}dx}\int |\nabla f(x)|^2 e^{-f(x)}dx \\
&= -\frac{1}{\int e^{-f(x)}dx}\int (\nabla f(x))^T \nabla e^{-f(x)}dx \\
&= \frac{1}{\int e^{-f(x)}dx}\int \Delta f(x) e^{-f(x)}dx \\
&\leq \left\|\Delta f(x)\right\|_\infty \leq Md.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left\|v_{n-1}\right\|^2 \leq 2d/M + 2\mathbb{E}\left\|v_{n-1} - w_{n-1}\right\|^2 \leq 2d/M + 4A_{n-1}^2, \\
\mathbb{E}\left\|\nabla f(x_{n-1})\right\|^2 \leq 2Md + 2M^2\mathbb{E}\left\|x_{n-1} - y_{n-1}\right\|^2 \leq 2Md + 2M^2A_{n-1}^2.
\end{aligned}$$

Plug the upper bounds into our previous result:

$$\begin{aligned}
A_n^2 &\leq e^{-\frac{h}{2\kappa}}A_{n-1}^2 + 8250\left[(\kappa h^7 + h^4)(2d/M + 4A_{n-1}^2) + (\kappa^{-1}h^8 + \kappa^{-2}h^4)(2Md + 2M^2A_{n-1}^2) + (\kappa^{-1}h^5 + h^7)\right] \\
&\leq \left[1 - \frac{h}{2\kappa} + \frac{h^2}{8\kappa^2} + 49500(h^4 + \kappa h^7)\right]A_{n-1}^2 + 41250d(h^7 + \kappa^{-1}h^4).
\end{aligned}$$

If we choose $(x_{n-1}, v_{n-1}) \sim \pi_h^*(x,v)$ and $(y_{n-1}, w_{n-1}) \sim \pi^*(x,v)$ such that

$$A_{n-1}^2 = \min_{X\sim\pi_h^*,\ Y\sim\pi^*} \mathbb{E}\left\|X - Y\right\|^2,$$

then we have

$$W_2(\pi, \pi_h)^2 \leq A_{n-1}^2 \leq \frac{82500h^3(\kappa h^3 + 1)d}{1 - \frac{h}{4\kappa} - 99000h^3\kappa(1 + \kappa h^3)}.$$

We can see that $W_2(\pi, \pi_h) \to 0$ as $h \to 0$. Furthermore, as $h \to 0$, $W_2(\pi, \pi_h) < O(h^{\frac{3}{2}})$. ■

2.6.2. Proofs for Section 2.3.2.

PROOF OF THEOREM 5. Define $A_n^2 = \mathbb{E}[\|x_n - y_n\|^2 + \|(x_n + v_n) - (y_n + w_n)\|^2]$. From the proof of proposition 4, we know that

$$A_n^2 \leq \left[1 - \frac{\gamma_n}{2\kappa} + \frac{\gamma_n^2}{8\kappa^2} + 49500(\gamma_n^4 + \kappa\gamma_n^7)\right] A_{n-1}^2 + 41250d(\gamma_n^7 + \kappa^{-1}\gamma_n^4).$$

When time step h is a constant, apply the inequality repeatedly to get

$$A_n^2 \leq \left[1 - \frac{h}{2\kappa} + \frac{h^2}{8\kappa^2} + 49500(h^4 + \kappa h^7)\right]^k A_0^2 + \frac{82500h^3(\kappa h^3 + 1)d}{1 - \frac{h}{4\kappa} - 99000h^3\kappa(1 + \kappa h^3)}.$$

Denote ν_n to be the density function of x_n, then $W_2(\nu_n, \pi) \leq A_n$. By choosing $\gamma_n = h \sim O(\epsilon^{\frac{2}{3}})$, we can guarantee that $W_2(\nu_n, \pi) < \epsilon\sqrt{\frac{d}{m}}$ for all $n > K \sim \tilde{O}(\epsilon^{-\frac{2}{3}})$.

When the time step γ_n is variant, the inequality we correspondingly have

$$A_n^2 \leq \left[1 - \frac{\gamma_n}{2\kappa} + \frac{\gamma_n^2}{8\kappa^2} + 49500(\gamma_n^4 + \kappa\gamma_n^7)\right] A_{n-1}^2 + 41250d(\gamma_n^7 + \kappa^{-1}\gamma_n^4).$$

When $\gamma_n < 1$, $\frac{\gamma_n^2}{8\kappa^2} < \frac{\gamma_n}{8\kappa}$. When $\gamma_n < (\frac{99000}{8\kappa^2}) < 24\kappa^{-\frac{2}{3}}$, we have $49500(\gamma_n^4 + \kappa\gamma_n^7) < \frac{\gamma_n}{8\kappa}$. Similarly, when $\gamma_n < 1$, we have $41250d(\gamma_n^7 + \kappa^{-1}\gamma_n^4) < 82500d\gamma_n^4$. Therefore, when $\gamma_n < \min\{1/2, 24\kappa^{-\frac{2}{3}}\}$, we have

$$A_n^2 < (1 - \frac{\gamma_n}{4\kappa})A_{n-1}^2 + 82500d\gamma_n^4.$$

If we choose $\gamma_n = \frac{16\kappa}{32\kappa^{\frac{5}{3}} + (n-K_1)^+}$, where K_1 is the smallest integer such that

$$A_{K_1}^2 < (1 - \frac{4}{\kappa^{\frac{5}{3}}})^{K_1} A_0^2 + (82500)d\frac{1}{2\kappa} < 2\frac{82500d}{\kappa},$$

then we claim that for all $n \geq K_1$, we have

$$A_n^2 < \frac{82500(16)^4 d\kappa^4}{(32\kappa^{\frac{5}{3}} + n - K_1)^3}.$$

The claim can be proved by induction: Assume that the claim hold for A_n^2 and denote $b = 32\kappa^{\frac{5}{3}} + n - K_1$, then

$$\begin{aligned}
A_{n+1}^2 &< (1 - \frac{4}{1+b})\frac{82500(16)^4 d\kappa^4}{b^3} + \frac{82500d(16)^4\kappa^4}{(b+1)^4} \\
&= \frac{82500(16)^4 d\kappa^4}{(b+1)^3}\left[\frac{(b-3)(b+1)^2}{b^3} + \frac{1}{b+1}\right] \\
&< \frac{82500(16)^4 d\kappa^4}{(b+1)^3} \\
&= \frac{82500(16)^4 d\kappa^4}{(32\kappa^{\frac{5}{3}} + n + 1 - K_1)^3}.
\end{aligned}$$

Therefore, under our choice of time step (γ_n), we can guarantee $W_2(\nu_n, \pi) < \epsilon\sqrt{\frac{d}{m}}$ for all $n > K \sim O(\epsilon^{-\frac{2}{3}})$. Compared to the running time of constant step size RULMC, vanishing step size help reduce the factor $\log(\frac{1}{\epsilon})$ in the guarantees. ■

Now we introduce the CLT statement for another sampling algorithm related to (RULMC) and give a complete proof of the statement. The proof of Remark 6 can be done in the same way. In the following theorem, we give a central limit result with specific choice of weights and time step-size. The Euler-discretization of the underdamped Langevin diffusion (which we call as KLMC, following [DRD20a]) is given by the following algorithm:

$$\text{(KLMC)} \quad \begin{aligned}
x_{n+1} &= x_n + \frac{1-e^{-2\gamma_{n+1}}}{2}v_n - \frac{u}{2}(\gamma_{n+1} - \frac{1-e^{-2\gamma_{n+1}}}{2})\nabla f(x_n) + \sqrt{u}\sigma_{n+1}^{(1)}U_{n+1}^{(1)}, \\
v_{n+1} &= v_n e^{-2\gamma_{n+1}} - u\frac{1-e^{-2\gamma_{n+1}}}{2}\nabla f(x_n) + 2\sqrt{u}\sigma_{n+1}^{(2)}U_{n+1}^{(2)}.
\end{aligned}$$

where $\{\gamma_n\}$ are the time steps. $\sigma_n^{(1)}$ and $\sigma_n^{(2)}$ are positive with ${\sigma_n^{(1)}}^2 = \gamma_n + \frac{1-e^{-4\gamma_n}}{4} - (1 - e^{-2\gamma_n})$, ${\sigma_n^{(2)}}^2 = \frac{1-e^{-4\gamma_n}}{4}$. $\{(U_n^{(1)}, U_n^{(2)})\}_n$ are independent Centered Gaussian random vectors in $\mathbb{R}^{2d}$ with $(U_n^{(1)}, U_n^{(2)}) \sim \mathcal{N}(0, \sigma_n^2 I_d)$ and $\sigma_n^2 = \frac{1+e^{-4\gamma_n}-2e^{-2\gamma_n}}{4\sigma_n^{(1)}\sigma_n^{(2)}}$. Numerical integration with the above sampler follows the same steps as described in Section 2.3.2. We now provide the following CLT.

THEOREM 5. Assume potential function f satisfies Assumption 2.1.1. Let $\{(x_k, v_k)\}$ and $\{(U_k^{(1)}, U_k^{(2)})\}$ be the same as what we have in the (KLMC) algorithm and the time step-size $\{\gamma_k\}$ is non-increasing and $\lim_k(\gamma_{k-1} - \gamma_k)/\gamma_k^4 = 0$. If $\lim_n(1/\sqrt{\Gamma_n^{(3)}})\sum_{k=1}^n \gamma_k^4 = \hat{\gamma} \in (0, +\infty]$ and $\lim_n \Gamma_n^{(4)} = +\infty$, then for all $\phi \in \mathcal{C}^3$ with $D^2\phi$, $D^3\phi$ and $D^4\phi$ bounded and Lipschitz and $\sup_{(x,v)\in\mathbb{R}^{2d}} |\nabla\phi(x)|^2/V(x,v) < +\infty$, we have

$$\frac{\Gamma_n}{\Gamma_n^{(4)}}\nu_n^{\gamma}(\mathcal{L}\phi) \to \mathcal{N}(\rho, \frac{10}{3}u\hat{\gamma}^{-2}\int_{\mathbb{R}^d}|\nabla\phi(x)|\pi(dx)) \qquad \text{if } \hat{\gamma} < +\infty,$$
$$\frac{\Gamma_n}{\Gamma_n^{(4)}}\nu_n^{\gamma}(\mathcal{L}\phi) \to \rho \qquad \text{if } \hat{\gamma} = +\infty,$$

where

$$\rho = \tfrac{u}{6}\int\int\langle D^3\phi(x);\nabla f(x), v^{\otimes 2}\rangle\nu(dx,dv) + \tfrac{u}{24}\int\int\langle D^3 f(x);\nabla\phi(x), v^{\otimes 2}\rangle\nu(dx,dv)$$
$$+ \tfrac{u}{12}\int\int(D^2\phi D^2 f)(x)v^{\otimes 2}\nu(dx,dv) - \tfrac{1}{12}\int\int D^4\phi(x)v^{\otimes 4}\nu(dx,dv)$$
$$- \tfrac{u^2}{24}\int\langle D^2 f(x);\nabla\phi(x),\nabla f(x)\rangle\pi(dx).$$

In the following context we'll discuss the weak convergence of empirical measure ν_n^{η} and build a central limit theorem under certain assumptions.

1) **(Lyapunov Conditions)** The underdamped Langevin dynamics can be rewritten as

$$dY_t = b(Y_t)dt + \sigma(Y_t)dW_t$$

where $Y_t = [X_t, V_t]^T$, $b(y) = b(x,v) = [v, -2v - u\nabla f(x)]^T$, $\sigma(y) = 2\sqrt{u}[0_d, I_d]^T$ for all $x, v \in \mathbb{R}^d$. $\{W_t\}$ is a 2d-dimensional Brownian motion.

The Lyapunov condition is similar to the one that's introduced in[LP02].

Assumption $(\mathcal{L}_{V,\infty})$: There's a $\mathcal{C}^2$ function $V : \mathbb{R}^{2d} \to [v_*, +\infty)$ for some $v_* > 0$ satisfying the following conditions:

a) $\left\|D^2V\right\|_\infty = \sup_{(x,v)^T\in\mathbb{R}^{2d}}\left\|D^2V(x,v)\right\|_{op} < +\infty$ and $\lim_{|(x,v)|\to+\infty} V(x,v) = +\infty$;

b) $|\nabla V(x,v)|^2 + |b(x,v)|^2 \le c_V V(x,v)$ for all $(x,v)^T \in \mathbb{R}^{2d}$ and some $c_V > 0$;

c) $\langle\nabla V(x,v), b(x,v)\rangle \le -\alpha V(x,v) + \beta$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$.

Assumption $(\mathcal{L}_{V,p})$: There's a $\mathcal{C}^2$ function $V : \mathbb{R}^{2d} \to [v_*, +\infty)$ for some $v_* > 0$ satisfying for some $p \ge 1$:

a) $\left\|D^2V\right\|_\infty = \sup_{(x,v)^T\in\mathbb{R}^{2d}}\left\|D^2V(x,v)\right\|_{op} < +\infty$ and $\lim_{|(x,v)|\to+\infty} V(x,v) = +\infty$;

b) $|\nabla V(x,v)|^2 + |b(x,v)|^2 + \mathrm{Tr}(\sigma(x,v)\sigma(x,v)^T) \leq c_V V(x,v)$ for all $(x,v)^T \in \mathbb{R}^{2d}$ and some $c_V > 0$;

c) $\langle \nabla V(x,v), b(x,v)\rangle + \lambda_p \mathrm{Tr}(\sigma(x,v)\sigma(x,v)^T) \leq -\alpha V(x,v) + \beta$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$, where $\lambda_p = \frac{1}{2}\lambda_{D^2V+(p-1)(\nabla V \otimes \nabla V)/V}$.

Remark 10. (1) We can show that: $(\mathcal{L}_{V,p'}) \implies (\mathcal{L}_{V,p})$ if $p' \geq p \geq 1$. Especially $(\mathcal{L}_{V,\infty}) \implies (\mathcal{L}_{V,p})$ for all $p \geq 1$.

(2) If we choose b and σ the same as those in the Underdamped Langevin dynamics, then $(\mathcal{L}_{V,\infty})$ is almost the same as assumption 2.6.1. We can instantly obtain that assumption 2.6.1 implies $(\mathcal{L}_{V,\infty})$. Therefore, according to lemma 2.6.1, assumption 2.1.1 implies $(\mathcal{L}_{V,\infty})$.

2) **(Tightness Result)** We now establish the almost sure tightness of the weighted empirical measures. The filtration $\{\mathcal{F}_n\}$ we consider is $\mathcal{F}_n = \sigma(Y_0, (U_1^{(1)}, U_1^{(2)}), \cdots, (U_n^{(1)}, U_n^{(2)}))$.

Lemma 2.6.2. (a) *If $(\mathcal{L}_{V,1})$ holds, then for every $a \geq \frac{1}{2}$,*

$$|V^a(Y_{n+1}) - V^a(Y_n)| \leq c_a \sqrt{\gamma_{n+1}} V^a(Y_n)(1 + |U_{n+1}^{(1)}|^{2a} + |U_{n+1}^{(2)}|^{2a}).$$

(b) *If $(\mathcal{L}_{V,p})$ holds for some $p \geq 1$, then there exists real numbers $\tilde{\alpha} > 0$ and $\tilde{\beta}$ and $n_0 \in \mathbb{N}$ such that*

$$\mathbb{E}[V^p(Y_{n+1})|\mathcal{F}_n] \leq V^(Y_n) + \gamma_{n+1} V^{p-1}(Y_n)(\tilde{\beta} - \tilde{\alpha} V(Y_n)), \qquad \forall\ n \geq n_0$$

and furthermore

$$\sup_{n\in\mathbb{N}} \mathbb{E}[V^p(Y_n)] < +\infty.$$

Proof of Lemma 2.6.2. (a) Using mean value theorem and $(\mathcal{L}_{V,1})$:

$$
\begin{aligned}
|V^a(Y_{n+1}) - V^a(Y_n)| &= a|V^{a-1}(\xi_{n+1})\langle \nabla V(\xi_{n+1}), Y_{n+1} - Y_n\rangle| \\
&\leq CV^{a-\frac{1}{2}}(\xi_{n+1})|Y_{n+1} - Y_n|.
\end{aligned}
$$

From $(\mathcal{L}_{V,1})$-b) we get that $\nabla\sqrt{V}$ is bounded, i.e $\sqrt{V}$ is Lipschitz with parameter $[\sqrt{V}]_1$. Hence

$$
\begin{aligned}
V^{a-\frac{1}{2}}(\xi_{n+1}) &\leq (\sqrt{V}(Y_n) + [\sqrt{V}]_1|Y_{n+1} - Y_n|)^{2a-1} \\
&\leq 2^{2a-1}\left(V^{a-\frac{1}{2}}(Y_n) + [\sqrt{V}]_1^{2a-1}|Y_{n+1} - Y_n|^{2a-1}\right).
\end{aligned}
$$

Meanwhile,

$$
\begin{aligned}
|Y_{n+1} - Y_n|^2 &= \left| \begin{bmatrix} \frac{1-e^{-2\gamma_{n+1}}}{2}v_n - \frac{u}{2}(\gamma_{n+1} - \frac{1-e^{-2\gamma_{n+1}}}{2})\nabla f(x_n) + \sqrt{u}\sigma^{(1)}_{n+1}U^{(1)}_{n+1} \\ -2\frac{1-e^{-2\gamma_{n+1}}}{2}v_n - u\frac{1-e^{-2\gamma_{n+1}}}{2}\nabla f(x_n) + 2\sqrt{u}\sigma^{(2)}_{n+1}U^{(2)}_{n+1} \end{bmatrix} \right|^2 \\
&\leq 15(\frac{1-e^{-2\gamma_{n+1}}}{2})^2|v_n^2| + [\frac{3u^2}{4}(\gamma_{n+1} - \frac{1-e^{-2\gamma_{n+1}}}{2})^2 + 3u^2(\frac{1-e^{-2\gamma_{n+1}}}{2})^2]|\nabla f(x_n)|^2 \\
&\quad + 3u{\sigma^{(1)}_{n+1}}^2|U^{(1)}_{n+1}|^2 + 12u{\sigma^{(2)}_{n+1}}^2|U^{(2)}_{n+1}|^2.
\end{aligned}
$$

Since $\gamma_n \to 0$ as $n \to \infty$ and $\frac{1-e^{-2\gamma_n}}{2} \sim O(\gamma_n)$, $\gamma_n - \frac{1-e^{-2\gamma_n}}{2} \sim O(\gamma_n^2)$, $\sigma_n^{(1)} \sim O(\gamma_n^{\frac{3}{2}})$ and $\sigma_n^{(2)} \sim O(\gamma_n^{\frac{1}{2}})$, there exist $C_1, C_2, C_3 > 0$ such that

$$
\begin{aligned}
|Y_{n+1} - Y_n|^2 &\leq C_1\left[\gamma_{n+1}^2(|v_n|^2 + |\nabla f(x_n)|^2) + \gamma_{n+1}(|U^{(1)}_{n+1}|^2 + |U^{(2)}_{n+1}|^2)\right] \\
&\leq C_2\left[\gamma_{n+1}^2 V(Y_n) + \gamma_{n+1}(|U^{(1)}_{n+1}|^2 + |U^{(2)}_{n+1}|^2 + 1)\right] \\
\implies \quad |Y_{n+1} - Y_n| &\leq C_3\sqrt{\gamma_{n+1}}\sqrt{V(Y_n)}(|U^{(1)}_{n+1}| + |U^{(2)}_{n+1}| + 1).
\end{aligned}
$$

Combining our estimations, since $a \geq 1/2$. we get

$$
\begin{aligned}
|V^a(Y_{n+1}) - V^a(Y_n)| &\leq C2^{2a-1}\left(V^{a-\frac{1}{2}}(Y_n) + [\sqrt{V}]_1^{2a-1}|Y_{n+1} - Y_n|^{2a-1}\right)|Y_{n+1} - Y_n| \\
&\leq c'_a\left(\sqrt{\gamma_{n+1}}V^a(Y_n)(|U^{(1)}_{n+1}| + |U^{(2)}_{n+1}| + 1) + \gamma_{n+1}^a V^a(Y_n)(|U^{(1)}_{n+1}| + |U^{(2)}_{n+1}| + 1)^{2a}\right) \\
&\leq c_a\sqrt{\gamma_{n+1}}V^a(Y_n)(|U^{(1)}_{n+1}|^{2a} + |U^{(2)}_{n+1}|^{2a} + 1).
\end{aligned}
$$

(b) We Taylor expand $V^p(Y_{n+1})$ at Y_n:

$$V^p(Y_{n+1}) = V^p(Y_n) + pV^{p-1}(Y_n)\langle \nabla V(Y_n), Y_{n+1} - Y_n\rangle + \frac{1}{2}D^2(V^p)(\xi_{n+1})(Y_{n+1} - Y_n)^{\otimes 2}.$$

Since $D^2(V^p) = pV^{p-1}D^2V + p(p-1)V^{p-1}\nabla V \nabla V^T$, by the definition of λ_p:

$$D^2(V^p)(\xi_{n+1})(Y_{n+1} - Y_n)^{\otimes 2} \leq 2p\lambda_p V^{p-1}(\xi_{n+1})|Y_{n+1} - Y_n|^2.$$

Therefore

$$V^p(Y_{n+1}) \leq V^p(Y_n) + pV^{p-1}(Y_n)\langle \nabla V(Y_n), Y_{n+1} - Y_n\rangle + p\lambda_p V^{p-1}(\xi_{n+1})|Y_{n+1} - Y_n|.$$

When $p = 1$, take conditional expectation on $\mathcal{F}_n$:

$$\begin{aligned}
\mathbb{E}[V(Y_{n+1})|\mathcal{F}_n] \leq V(Y_n) &+ \frac{1 - e^{-2\gamma_{n+1}}}{2}\langle \nabla V(x_n, v_n), b(x_n, v_n)\rangle \\
&- \frac{u}{2}(\gamma_{n+1} - \frac{1 - e^{-2\gamma_{n+1}}}{2})\nabla_x V(x_n, v_n) \cdot \nabla f(x_n) \\
&+ \lambda_1(\frac{1 - e^{-2\gamma_{n+1}}}{2})^2[5|v_n|^2 + u^2|\nabla f(x_n)|^2 + 4u\nabla f(x_n) \cdot v_n] \\
&- \frac{u}{2}\frac{1 - e^{-2\gamma_{n+1}}}{2}(\gamma_{n+1} - \frac{1 - e^{-2\gamma_{n+1}}}{2})\nabla f(x_n) \cdot v_n \\
&+ \frac{u^2}{4}(\gamma_{n+1} - \frac{1 - e^{-2\gamma_{n+1}}}{2})^2|\nabla f(x_n)|^2 + u({\sigma_{n+1}^{(1)}}^2 + 4{\sigma_{n+1}^{(2)}}^2)d.
\end{aligned}$$

There exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$

$$\begin{aligned}
&\frac{1 - e^{-2\gamma_{n+1}}}{2}\langle \nabla V(x_n, v_n), b(x_n, v_n)\rangle \leq \gamma_{n+1}(-\alpha V(Y_n) + \beta), \qquad \text{for some } \alpha > 0, \beta \in \mathbb{R}; \\
&-\frac{u}{2}(\gamma_{n+1} - \frac{1 - e^{-2\gamma_{n+1}}}{2})\nabla_x V(x_n, v_n) \cdot \nabla f(x_n) \leq C\gamma_{n+1}^2(|\nabla V(Y_n)|^2 + |b(Y_n)|^2) \leq C\gamma_{n+1}^2 V(Y_n); \\
&\lambda_1(\frac{1 - e^{-2\gamma_{n+1}}}{2})^2[5|v_n|^2 + u^2|\nabla f(x_n)|^2 + 4u\nabla f(x_n) \cdot v_n] \leq C\gamma_{n+1}^2|b(Y_n)|^2 \leq C\gamma_{n+1}^2 V(Y_n); \\
&-\frac{u}{2}\frac{1 - e^{-2\gamma_{n+1}}}{2}(\gamma_{n+1} - \frac{1 - e^{-2\gamma_{n+1}}}{2})\nabla f(x_n) \cdot v_n \leq C\gamma_{n+1}^3|b(Y_n)|^2 \leq C\gamma_{n+1}^3 V(Y_n); \\
&\frac{u^2}{4}(\gamma_{n+1} - \frac{1 - e^{-2\gamma_{n+1}}}{2})^2|\nabla f(x_n)|^2 \leq C\gamma_{n+1}^4|b(Y_n)|^2 \leq C\gamma_{n+1}^4 V(Y_n); \\
&u({\sigma_{n+1}^{(1)}}^2 + 4{\sigma_{n+1}^{(2)}}^2)d \leq C\gamma_{n+1}.
\end{aligned}$$

Therefore, for all $n \geq n_0$, there exist $\tilde{\alpha} > 0$, $\tilde{\beta} \in \mathbb{R}$ such that

$$\begin{aligned}\mathbb{E}[V(Y_{n+1})|\mathcal{F}_n] &\leq V(Y_n)(1 - \alpha\gamma_{n+1} + C(2\gamma_{n+1}^2 + \gamma_{n+1}^3 + \gamma_{n+1}^4)) + \gamma_{n+1}(\beta + C) \\ &\leq V(Y_n)(1 - \tilde{\alpha}\gamma_{n+1}) + \tilde{\beta}\gamma_{n+1},\end{aligned}$$

and $1 - \tilde{\alpha}\gamma_{n+1} > 0$. This leads to

$$\mathbb{E}[V(Y_{n+1})] \leq \mathbb{E}[V(Y_n)](1 - \tilde{\alpha}\gamma_{n+1}) + \tilde{\beta}\gamma_{n+1}.$$

We could use induction to prove:

$$\sup_{n \geq n_0} \mathbb{E}[V(Y_n)] \leq \frac{\tilde{\beta}}{\tilde{\alpha}} \vee \mathbb{E}[V(Y_{n_0})].$$

Assume now $p > 1$. Due to $(\mathcal{L}_{V,p})$-b), we derive that $\sqrt{V}$ is Lipschitz with parameter $[\sqrt{V}]_1$. Consequently,

$$\begin{aligned}V^{p-1}(\xi_{n+1}) = \sqrt{V}^{2(p-1)}(\xi_{n+1}) &\leq \left(\sqrt{V}(Y_n) + [\sqrt{V}]_1|Y_{n+1} - Y_n|\right)^{2(p-1)} \\ &\leq \begin{cases} V^{p-1}(Y_n) + ([\sqrt{V}]_1|Y_{n+1} - Y_n|)^{2(p-1)} & \text{if } 2(p-1) \leq 1, \\ V^{p-1}(Y_n) + C\left(V^{(2p-3)/2}(Y_n)|Y_{n+1} - Y_n| + |Y_{n+1} - Y_n|^{2(p-1)}\right) & \text{if } 2(p-1) > 1. \end{cases}\end{aligned}$$

Using the fact we've proved in part a):

$$|Y_{n+1} - Y_n|^2 \leq C_2\left[\gamma_{n+1}^2 V(Y_n) + \gamma_{n+1}(|U_{n+1}^{(1)}|^2 + |U_{n+1}^{(2)}|^2 + 1)\right].$$

We derive

$$V^{p-1}(\xi_{n+1})|Y_{n+1} - Y_n|^2 \leq V^{p-1}(Y_n)|Y_{n+1} - Y_n|^2 + C\gamma_{n+1}^{p\wedge\frac{3}{2}} V^p(Y_n)(1 + |U_{n+1}^{(1)}|^{2p} + |U_{n+1}^{(2)}|^{2p}).$$

Then we take conditional expectation:

$$\begin{aligned}\mathbb{E}[V^p(Y_{n+1})|\mathcal{F}_n] \leq V^p(Y_n) &+ pV^{p-1}\frac{1-e^{-2\gamma_{n+1}}}{2}\langle \nabla V(x_n, v_n), b(x_n, v_n)\rangle \\ &- pV^{p-1}(Y_n)\frac{u}{2}(\gamma_{n+1} - \frac{1-e^{-2\gamma_{n+1}}}{2})\nabla_x V(x_n, v_n)\cdot \nabla f(x_n) \\ &+ CV^{p-1}(Y_n)\left[\gamma_{n+1}^2 V(Y_n) + \gamma_{n+1}(|U_{n+1}^{(1)}|^2 + |U_{n+1}^{(2)}|^2 + 1)\right] \\ &+ C\gamma_{n+1}^{p\wedge\frac{3}{2}} V^p(Y_n)(1 + |U_{n+1}^{(1)}|^{2p} + |U_{n+1}^{(2)}|^{2p}).\end{aligned}$$

There exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$

$$\begin{aligned}&\frac{1-e^{-2\gamma_{n+1}}}{2}\langle \nabla V(x_n, v_n), b(x_n, v_n)\rangle \leq \gamma_{n+1}(-\alpha V(Y_n) + \beta), \qquad \text{for some } \alpha > 0, \beta \in \mathbb{R}; \\ &-\frac{u}{2}(\gamma_{n+1} - \frac{1-e^{-2\gamma_{n+1}}}{2})\nabla_x V(x_n, v_n)\cdot \nabla f(x_n) \leq C\gamma_{n+1}^2(|\nabla V(Y_n)|^2 + |b(Y_n)|^2) \leq C\gamma_{n+1}^2 V(Y_n).\end{aligned}$$

Since $\gamma_n^{p\wedge\frac{3}{2}}, \gamma_n^2 \sim o(\gamma_n)$, there exists $\tilde{\alpha} > 0, \tilde{\beta} \in \mathbb{R}$, such that for all $n \geq n_0$:

$$\mathbb{E}[V^p(Y_n)|\mathcal{F}_n] \leq V^p(Y_n) + \gamma_{n+1}V^{p-1}(Y_n)(\tilde{\beta} - \tilde{\alpha}V(Y_n)).$$

Same as the proof for $p = 1$, we can show

$$\sup_{n\in\mathbb{N}} \mathbb{E}[V^p(Y_n)] < +\infty.$$

■

THEOREM 6. Let $p \in [0, +\infty)$. Assume $(\mathcal{L}_{V,p})$, If there exists $s \in (0, 1]$ such that

$$\sum_{n\geq 1} \frac{1}{H_n}(\Delta\frac{\eta_n}{\gamma_n})_+ < +\infty \;\text{ and }\; \sum_{n\geq 1}(\frac{\eta_n}{H_n\sqrt{\gamma_n}})^{1+s} < +\infty,$$

then

$$\mathbb{P}(d\omega)\text{-a.s.} \quad \sup_{n\in\mathbb{N}} \nu_n^\eta(\omega, V^{p/(1+s)}) < +\infty.$$

Based on Lemma 2.6.2, the proof of Theorem 6 immediately follows, by using the same steps in the proof of Theorem 4 in [LP02]. Hence we don't replicate the proof here.

3) **(Identification of the limit)**

Theorem 7 (Echeverría-Weiss Theorem). Let E be a locally compact Polish space and $\mathcal{L}$ a linear operator satisfying the positive maximum principle. Assume that its domain $\mathcal{D}(A)$ is an algebra everywhere dense in $(\mathcal{C}_0(E), \| \ \|_\infty)$ containing a sequence $(f_n)_{n\in\mathbb{N}}$ satisfying

$$\sup_{n\in\mathbb{N}} (\|f_n\|_\infty + \|\mathcal{L}f_n\|_\infty) < +\infty, \quad \forall x \in E, \quad f_n(x) \to 1 \text{ and } \mathcal{L}f_n(x) \to 0.$$

If a distribution on $(E, \mathcal{B}(E))$ satisfies $\int_E \mathcal{L}f d\nu = 0$ for every $f \in \mathcal{D}(A)$, then there exists a stationary solution for the martingale problem $(\mathcal{L}, \nu)$ (this means that there exists a stationary continuous-time homogeneous Markov process with infinitesimal generator $\mathcal{L}$ and invariant distribution ν).

Lemma 2.6.3. *If the potential function f is Gradient Lipschitz and strongly convex, then the generator of kinetic, $\mathcal{L}$, satisfies the assumptions of the Echeverría-Weiss theorem.*

Proof of lemma 2.6.3. First it's well-known that the infinitesimal generator of a Fellerian semigroup satisfies the maximum principle. We can choose our $f_n(y) = \phi(y/n)$ for any $y \in \mathbb{R}^{2d}$ where ϕ is $\mathcal{C}^2$ with compact support and $\phi(0) = 1$. It's easy to check that $\forall y \in \mathbb{R}^{2d}$, $f_n(y) \to 0$ and $\mathcal{L}f_n(y) \to 0$. It's also straightforward that $\sup_{n\in\mathbb{N}} \|f_n\|_\infty < +\infty$. The last thing to check is $\sup_{n\in\mathbb{N}} \|\mathcal{L}f_n\|_\infty < +\infty$. Since $\mathcal{L}$ can also be written as $b(x,v) \cdot [\nabla_x, \ \nabla_v]^T + 2u\Delta_v$ and we've shown that under our assumptions on f, $(\mathcal{L}_{V,\infty})$ is satisfied, we have the Lyapunov function $V(y) = O(|y|^2)$ and $|b(x,v)| \leq C(1 + |(x,v)|)$. Therefore we get $\sup_{n\in\mathbb{N}} \|\mathcal{L}f_n\|_\infty < +\infty$. ∎

Theorem 8. Assume that f is gradient Lipschitz and strongly convex. Assume also

$$\lim_n \frac{1}{H_n} \sum_{k=1}^{n} |\Delta \frac{\eta_n}{\gamma_n}| = 0 \text{ and } \sum_{n\geq 1} (\frac{\eta_n}{\sqrt{\gamma_n} H_n})^2 < +\infty.$$

Let $a \geq \frac{1}{2}$. Assume $\sup_n \nu_n^\eta(V^a) < +\infty$ $\mathbb{P}$-a.s. If $a < 1$, assume also that $\sum_{n\geq 1} \eta_n \gamma_n / H_n < +\infty$. Then $\mathbb{P}$-a.s, every limiting distribution $\nu_\infty(\omega, dx)$ of the sequence $(\nu_n^\eta(\omega, dx))$ is an invariant distribution of the underdamped Langevin dynamics introduced in the previous section.

The proof of Theorem 8 follows immediately from Theorem 7, Lemma 2.6.3, Lemma 2.6.4 and Lemma 2.6.5.

LEMMA 2.6.4. *Under the assumptions in Theorem 8, then for every bounded Lipschitz continuous function* $g : \mathbb{R}^{2d} \to \mathbb{R}$*,*

$$\mathbb{P}\text{-}a.s. \quad \lim_n \frac{1}{H_n} \sum_{k=1}^{n} \frac{\eta_k}{\gamma_k} \mathbb{E}[g(Y_k) - g(Y_{k-1}) | \mathcal{F}_{k-1}] = 0.$$

PROOF OF LEMMA 2.6.4. Setting $\eta_0/\gamma_0 = 0$ gives

$$\frac{1}{H_n} \sum_{k=1}^{n} \mathbb{E}[g(Y_k) - g(Y_{k-1}) | \mathcal{F}_{k-1}] = \frac{1}{H_n} \sum_{k=1}^{n} \frac{\eta_k}{\gamma_k} (g(Y_k) - g(Y_{k-1})) - \frac{1}{H_n} \sum_{k=1}^{n} \frac{\eta_k}{\gamma_k} \left(g(Y_k) - \mathbb{E}[g(Y_k) | \mathcal{F}_{k-1}]\right).$$

As g is bounded, it follows by lemma 3-b) in[LP02] that

$$\mathbb{P}\text{-a.s.} \quad \lim_n \frac{1}{H_n} \sum_{k=1}^{n} (g(Y_k) - g(Y_{k-1})) = 0.$$

Then

$$\frac{1}{H_n} \sum_{k=1}^{n} \frac{\eta_k}{\gamma_k} \left(g(Y_k) - \mathbb{E}[g(Y_k) | \mathcal{F}_{k-1}]\right)$$

will converge to 0 once the martingale

$$M_n^g := \sum_{k=1}^{n} \frac{\eta_k}{\gamma_k H_k} \left(g(Y_k) - \mathbb{E}[g(Y_k) | \mathcal{F}_{k-1}]\right)$$

converge a.s. in $\mathbb{R}$.

$$\begin{aligned} \mathbb{E}\langle M_n^g \rangle_\infty &= \sum_{n \geq 1} (\frac{\eta_n}{\gamma_n H_n})^2 \left\| g(Y_n) - \mathbb{E}[g(Y_n) | \mathcal{F}_{n-1}] \right\|_2^2 \leq \sum_{n \geq 1} (\frac{\eta_n}{\gamma_n H_n})^2 \left\| g(Y_n) - g(Y_{n-1}) \right\|_2^2 \\ &\leq [f]_1^2 \sum_{n \geq 1} (\frac{\eta_n}{\gamma_n H_n})^2 \left\| Y_n - Y_{n-1} \right\|_2^2. \end{aligned}$$

Since $(\mathcal{L}_{V,1})$ holds under our assumptions on f and by Lemma 2-b)

$$\| Y_n - Y_{n-1} \|_2^2 \leq C' \mathbb{E}[\gamma_n^2 V(Y_{n-1}) + (2d+1)\gamma_n] \leq C\gamma_n.$$

Therefore

$$\mathbb{E}\langle M_n^g \rangle_\infty \leq C \sum_{n \geq 1} (\frac{\eta_n}{\sqrt{\gamma_n} H_n})^2 < +\infty.$$

■

LEMMA 2.6.5. *Under the assumptions in Theorem 8, then for every $g \in \mathcal{C}^2(\mathbb{R}^{2d})$ with compact support,*

$$\lim_n \left(\frac{1}{H_n} \sum_{k=1}^{n} \frac{\eta_k}{\gamma_k} \mathbb{E}[g(Y_k) - g(Y_{k-1}) | \mathcal{F}_{k-1}] - \nu_n^{\eta}(\mathcal{L}g) \right) = 0 \quad \boldsymbol{a.s.}$$

PROOF OF LEMMA 2.6.5. Setting $R_2(y_1, y_2) := g(y_2) - g(y_1) - \langle \nabla g(y_1), y_2 - y_1 \rangle - \frac{1}{2} D^2 g(y_1)(y_2 - y_1)^{\otimes 2}$, we obtain for every $k \in \mathbb{N}$,

$$\begin{aligned}
& g(Y_k) - g(Y_{k-1}) \\
&= \langle \nabla g(Y_{k-1}), Y_k - Y_{k-1} \rangle + \frac{1}{2} D^2 g(Y_{k-1})(Y_k - Y_{k-1})^{\otimes 2} + R_2(Y_{k-1}, Y_k) \\
&= \nabla_x g(x_{k-1}, v_{k-1}) \cdot [\frac{1 - e^{-2\gamma_k}}{2} v_{k-1} - \frac{u}{2}(\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \nabla f(x_{k-1}) + \sqrt{u} \sigma_k^{(1)} U_k^{(1)}] \\
&+ \nabla_v g(x_{k-1}, v_{k-1}) \cdot [-2 \frac{1 - e^{-2\gamma_k}}{2} v_{k-1} - u \frac{1 - e^{-2\gamma_k}}{2} \nabla f(x_{k-1}) + 2\sqrt{u} \sigma_k^{(2)} U_k^{(2)}] \\
&+ \frac{1}{2} D_x^2 g(x_{k-1}, v_{k-1}) [\frac{1 - e^{-2\gamma_k}}{2} v_{k-1} - \frac{u}{2}(\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \nabla f(x_{k-1}) + \sqrt{u} \sigma_k^{(1)} U_k^{(1)}]^{\otimes 2} \\
&+ \frac{1}{2} D_v^2 g(x_{k-1}, v_{k-1}) [-2 \frac{1 - e^{-2\gamma_k}}{2} v_{k-1} - u \frac{1 - e^{-2\gamma_k}}{2} \nabla f(x_{k-1}) + 2\sqrt{u} \sigma_k^{(2)} U_k^{(2)}]^{\otimes 2} \\
&+ \langle D_{xv} g(x_{k-1}, v_{k-1}); \frac{1 - e^{-2\gamma_k}}{2} v_{k-1} - \frac{u}{2}(\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \nabla f(x_{k-1}) + \sqrt{u} \sigma_k^{(1)} U_k^{(1)}, \\
&\qquad\qquad - 2 \frac{1 - e^{-2\gamma_k}}{2} v_{k-1} - u \frac{1 - e^{-2\gamma_k}}{2} \nabla f(x_{k-1}) + 2\sqrt{u} \sigma_k^{(2)} U_k^{(2)} \rangle + R_2(Y_{k-1}, Y_k) \\
&= \gamma_k \mathcal{L} g(Y_{k-1}) - (\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \nabla_x g(Y_{k-1}) \cdot v_{k-1} - \frac{u}{2}(\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \nabla_x g(Y_{k-1}) \cdot \nabla f(x_{k-1}) \\
&+ 2(\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \nabla_v g(Y_{k-1}) \cdot v_{k-1} + u(\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \nabla_v g(Y_{k-1}) \cdot \nabla f(x_{k-1}) \\
&+ \sqrt{u} \sigma_k^{(1)} \nabla g(Y_{k-1}) \cdot U_k^{(1)} + 2\sqrt{u} \sigma_k^{(2)} \nabla g(Y_{k-1}) \cdot U_k^{(2)} \\
&+ \frac{1}{2}(\frac{1 - e^{-2\gamma_k}}{2})^2 D_x^2 g(Y_{k-1}) v_{k-1}^{\otimes 2} + \frac{u^2}{8}(\gamma_k - \frac{1 - e^{-2\gamma_k}}{2})^2 D_x^2 g(Y_{k-1}) \nabla f(x_{k-1})^{\otimes 2} \\
&+ \frac{u}{2} {\sigma_k^{(1)}}^2 D_x^2 g(Y_{k-1}) {U_k^{(1)}}^{\otimes 2} - \frac{u}{2} \frac{1 - e^{-2\gamma_k}}{2} (\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \langle D_x^2 g(Y_{k-1}); v_{k-1}, \nabla f(x_{k-1}) \rangle \\
&+ \sqrt{u} \frac{1 - e^{-2\gamma_k}}{2} \sigma_k^{(1)} \langle D_x^2 g(Y_{k-1}); v_{k-1}, U_k^{(1)} \rangle - \frac{u^{3/2}}{2}(\gamma_k - \frac{1 - e^{-2\gamma_k}}{2}) \sigma_k^{(1)} \langle D_x^2 g(Y_{k-1}); \nabla f(x_{k-1}), U_k^{(1)} \rangle \\
&+ 2(\frac{1 - e^{-2\gamma_k}}{2})^2 D_v^2 g(Y_{k-1}) v_{k-1}^{\otimes 2} + \frac{u^2}{2}(\frac{1 - e^{-2\gamma_k}}{2})^2 D_v^2 g(Y_{k-1}) \nabla f(x_{k-1})^{\otimes 2} \\
&+ 2u \left({\sigma_k^{(2)}}^2 D_v^2 g(Y_{k-1}) {U_k^{(2)}}^{\otimes 2} - \gamma_k \mathbb{E}[D_v^2 g(Y_{k-1}) {U_k^{(2)}}^{\otimes 2} | \mathcal{F}_{k-1}] \right) \\
&+ 2u(\frac{1 - e^{-2\gamma_k}}{2})^2 \langle D_v^2 g(Y_{k-1}); v_{k-1}, \nabla f(x_{k-1}) \rangle - 4\sqrt{u} \frac{1 - e^{-2\gamma_k}}{2} \sigma_k^{(2)} \langle D_v^2 g(Y_{k-1}); v_{k-1}, U_k^{(2)} \rangle \\
&- 2u^{3/2} \frac{1 - e^{-2\gamma_k}}{2} \sigma_k^{(2)} \langle D_v^2 g(Y_{k-1}); \nabla f(x_{k-1}), U_k^{(2)} \rangle + R_2(Y_{k-1}, Y_k).
\end{aligned}$$

Take conditional expectation:

$$\begin{aligned}
&\mathbb{E}[g(Y_k) - g(Y_{k-1})|\mathcal{F}_{k-1}] - \gamma_k \mathcal{L} g(Y_{k-1}) \\
=& -(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\nabla_x g(Y_{k-1}) \cdot v_{k-1} - \frac{u}{2}(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\nabla_x g(Y_{k-1}) \cdot \nabla f(x_{k-1}) \\
&+ 2(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\nabla_v g(Y_{k-1}) \cdot v_{k-1} + u(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\nabla_v g(Y_{k-1}) \cdot \nabla f(x_{k-1}) \\
&+ \frac{1}{2}(\frac{1-e^{-2\gamma_k}}{2})^2 D_x^2 g(Y_{k-1}) v_{k-1}^{\otimes 2} + \frac{u^2}{8}(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})^2 D_x^2 g(Y_{k-1}) \nabla f(x_{k-1})^{\otimes 2} \\
&+ \frac{u}{2}{\sigma_k^{(1)}}^2 \Delta_x g(Y_{k-1}) - \frac{u}{2}\frac{1-e^{-2\gamma_k}}{2}(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\langle D_x^2 g(Y_{k-1}); v_{k-1}, \nabla f(x_{k-1})\rangle \\
&+ 2(\frac{1-e^{-2\gamma_k}}{2})^2 D_v^2 g(Y_{k-1}) v_{k-1}^{\otimes 2} + \frac{u^2}{2}(\frac{1-e^{-2\gamma_k}}{2})^2 D_v^2 g(Y_{k-1}) \nabla f(x_{k-1})^{\otimes 2} \\
&+ 2u\left({\sigma_k^{(2)}}^2 - \gamma_k\right)\Delta_v g(Y_{k-1}) + 2u(\frac{1-e^{-2\gamma_k}}{2})^2 \langle D_v^2 g(Y_{k-1}); v_{k-1}, \nabla f(x_{k-1})\rangle \\
&+ \mathbb{E}[R_2(Y_{k-1}, Y_k)|\mathcal{F}_{k-1}].
\end{aligned}$$

Observe that for all the terms, except for $R_2(Y_{k-1}, Y_k)$, on the right hand side of the equation, their coefficients are of order $O(\gamma_k^2)$ or $o(\gamma_k^2)$. Furthermore, ∇g and $D^2 g$ are bounded because g is $\mathcal{C}^2$ and compact supported. Since $(\mathcal{L}_{V,\infty})$ is satisfied under our assumptions, $\sup_{n\in\mathbb{N}} \mathbb{E}[|v_n|^2 + |\nabla f(x_n)|^2] < C \sup_{n\in\mathbb{N}} \mathbb{E}[V(Y_n))] < +\infty$. Therefore, we obtain that as $n \to 0$,

$$\frac{1}{H_n}\sum_{k=1}^{n} \frac{\eta_k}{\gamma_k}\mathbb{E}[g(Y_k) - g(Y_{k-1})|\mathcal{F}_{k-1}] - \eta_k \mathcal{L} g(Y_{k-1}) - \frac{\eta_k}{\gamma_k}\mathbb{E}[R_2(Y_{k-1}, Y_k)|\mathcal{F}_{k-1}] \to 0$$

because $\frac{1}{H_n}\sum_{k=1}^{n} \eta_k \gamma_k \to 0$ as $n \to 0$. Now we deal with $\mathbb{E}[R_2(Y_{k-1}, Y_k)|\mathcal{F}_{k-1}]$. For any $x, y \in \mathbb{R}^{2d}$, define

$$r_2(x, y) := \frac{1}{2}\sup_{t\in(0,1)} \left\| D^2 g(x + t(y-x)) - D^2 g(x) \right\|.$$

It's easy to see that r_2 is a bounded continuous function on $\mathbb{R}^d \times \mathbb{R}^d$, $r_2(x,x) = 0$ and

$$|R_2(x, y)| \le r_2(x, y)|x - y|^2.$$

Therefore we obtain

$$\frac{\eta_k}{\gamma_k}|\mathbb{E}[R_2(Y_{k-1}, Y_k)|\mathcal{F}_{k-1}]| \le C\left(\eta_k\gamma_k \|r_2\|_\infty V(Y_{k-1}) + (2d+1)\eta_k \mathbb{E}[r_2(Y_{k-1}, Y_k)(|U_k^{(1)}|^2 + |U_k^{(2)}|^2)|\mathcal{F}_{k-1}]\right).$$

If $a \geq 1$, $\mathbb{P}$-a.s.

$$\frac{1}{H_n}\sum_{k=1}^{n} C\eta_k\gamma_k \left\|r_2\right\|_\infty V(Y_{k-1}) \leq C'\frac{1}{H_n}\sum_{k=1}^{n}\eta_k\gamma_k V(Y_{k-1}) \to 0 \quad \text{as } \sup_{n\in\mathbb{N}}\nu_n^\eta(V) < +\infty \text{ and } \gamma_n \to 0.$$

If $a \in [1/2, 1)$, the same limit follows from the Kronecker lemma mentioned in[LP02] and

$$\sum_{n\geq 1}\eta_n\gamma_n/H_n < +\infty.$$

Meanwhile, we also have

$$J(\gamma, x, v) = \int_{\mathbb{R}^d\times\mathbb{R}^d} r_2((x,v),(x',v'))(|r_1|^2+|r_2|^2)\mu(dr_1, dr_2)$$

where

$$(x',v') = (x + \frac{1-e^{-2\gamma}}{2}v - \frac{u}{2}(\gamma - \frac{1-e^{-2\gamma}}{2})\nabla f(x) + \sqrt{u}\sigma^{(1)}r_1, e^{-2\gamma}v - u\frac{1-e^{-2\gamma}}{2}\nabla f(x) + 2\sqrt{u}\sigma^{(2)}r_2)$$

$$\sigma^{(1)} = \left(\gamma + \frac{1-e^{-4\gamma}}{4} - (1-e^{-2\gamma})\right)^{1/2}, \quad \sigma^{(2)} = \left(\frac{1-e^{-4\gamma}}{4}\right)^{1/2},$$

$$(U^{(1)}, U^{(2)}) \sim \mu = \mathcal{N}(0, \frac{1+e^{-4\gamma}-2e^{-2\gamma}}{4\sigma^{(1)}\sigma^{(2)}}I_{2d}).$$

We can see that J is a bounded continuous function on $\mathbb{R}_+\times\mathbb{R}^d\times\mathbb{R}^d$ and $J(0,x,v) = 0$. Since $\lim_{|y|\to\infty} V(y) = +\infty$. We can also write

$$(2d+1)\eta_k\mathbb{E}[r_2(Y_{k-1}, Y_k)(|U_k^{(1)}|^2 + |U_k^{(2)}|^2)|\mathcal{F}_{k-1}] = \eta_k V^a((x_{k-1}, v_{k-1}))\theta((x_{k-1}, v_{k-1}))J(\gamma_k, x_{k-1}, v_{k-1})$$

where $\lim_{|(x_{k-1},v_{k-1})|\to\infty}\theta((x_{k-1}, v_{k-1})) = 0$. It remains to show that

$$\mathbb{P}\text{-a.s.} \quad \lim_n \frac{1}{H_n}\sum_{k=1}^{n}\eta_k V^a((x_{k-1}, v_{k-1}))\theta((x_{k-1}, v_{k-1}))J(\gamma_k, x_{k-1}, v_{k-1}) = 0.$$

For a fixed number $A > 0$, J is uniformly continuous on $[0, \sup_n\gamma_n]\times\bar{B}_{2d}(0, A)$, then

$$J(\gamma_k, x_{k-1}, v_{k-1})1_{|(x_{k-1},v_{k-1})|\leq A} \to 0 \;\; \mathbb{P}\text{-a.s.}$$

and $V^a((x_{k-1}, v_{k-1}))\theta((x_{k-1}, v_{k-1}))$ is bounded on $\bar{B}_{2d}(0, A)$. Therefore

$$\mathbb{P} - a, s \quad \lim_n \frac{1}{H_n} \sum_{k=1}^{n} \eta_k V^a((x_{k-1}, v_{k-1}))\theta((x_{k-1}, v_{k-1}))J(\gamma_k, x_{k-1}, v_{k-1})1_{|(x_{k-1}, v_{k-1})| \leq A} = 0.$$

On the other hand side

$$\begin{aligned}\limsup_n \frac{1}{H_n} \sum_{k=1}^{n} \eta_k V^a((x_{k-1}, v_{k-1}))\theta((x_{k-1}, v_{k-1}))J(\gamma_k, x_{k-1}, v_{k-1})1_{|(x_{k-1}, v_{k-1})| > A} \\ \leq \sup_{|(x,v)| > A} |\theta(x, v)| \, \|J\|_\infty \sup_n \nu_n^\eta(V^a) \to 0 \quad \textbf{as } A \to +\infty.\end{aligned}$$

So taking $A \to +\infty$ completes the proof. ■

Theorem 9. Let $p \in [1, +\infty)$. Assume $(\mathcal{L}_{V,p})$. Let $s \in (0, 1]$. Assume that

$$\sum_{n \geq 1} \frac{1}{H_n} \left(\Delta \frac{\eta_n}{\gamma_n} \right)_+ < +\infty. \quad \lim_n \frac{1}{H_n} \sum_{k=1}^{n} |\Delta \frac{\eta_k}{\gamma_k}| = 0 \text{ and } \sum_{n \geq 1} \left(\frac{\eta_n}{H_n \sqrt{\gamma_n}} \right)^{1+s} < +\infty$$

(a) Then

$$\mathbb{P}\text{-a.s.} \quad \sup_{n \in \mathbb{N}} \nu_n^\eta(\omega, V^{p/(1+s)}) < +\infty.$$

(b) When $p \leq 1 + s$, assume also $\sum_{n \geq 1} \eta_n \gamma_n / H_n < +\infty$. Then with probability 1, any weak limit of the sequence (ν_m^η) is an invariant distribution of the underdamped Langevin dynamics.

Theorem 9 follows directly from theorem 6 and theorem 8.

Proof of Theorem 5. First we try to decompose $\sum_{k=1}^{n} \gamma_k \mathcal{L}\phi(x_{k-1})$ using Taylor expansion.

$$\phi(x_k) = \phi(x_{k-1}) + \nabla\phi(x_{k-1}) \cdot (x_k - x_{k-1}) + \frac{1}{2} D^2\phi(x_{k-1})(x_k - x_{k-1})^{\otimes 2} + R_2^{(k)}$$

where $R_2^{(k)} = \phi(x_k) - \phi(x_{k-1}) - \nabla\phi(x_{k-1}) \cdot (x_k - x_{k-1}) - \frac{1}{2}D^2\phi(x_{k-1})(x_k - x_{k-1})^{\otimes 2}$. We can plug our discretization into the equation and obtain:

$$\begin{aligned}\phi(x_k) - \phi(x_{k-1}) = \gamma_k \mathcal{L}\phi(x_{k-1}) &- (\gamma_k - \frac{1-e^{-2\gamma_k}}{2})v_{k-1} \cdot \nabla\phi(x_{k-1}) \\ &- \frac{u}{2}(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\nabla f(x_{k-1}) \cdot \nabla\phi(x_{k-1}) + \sqrt{u}\sigma_k^{(1)}\nabla\phi(x_{k-1}) \cdot U_k^{(1)} \\ &+ \frac{1}{2}(\frac{1-e^{-2\gamma_k}}{2})^2 D^2\phi(x_{k-1})v_{k-1}^{\otimes 2} + \frac{u}{2}\sigma_k^{(1)^2}D^2\phi(x_{k-1})U_k^{(1)^{\otimes 2}} \\ &+ \frac{u^2}{8}(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})^2 D^2\phi(x_{k-1})\nabla f(x_{k-1})^{\otimes 2} \\ &- \frac{u}{2}\frac{1-e^{-2\gamma_k}}{2}(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\langle D^2\phi(x_{k-1}); v_{k-1}, \nabla f(x_{k-1})\rangle \\ &+ \sqrt{u}\sigma_k^{(1)}\frac{1-e^{-2\gamma_k}}{2}\langle D^2\phi(x_{k-1}); v_{k-1}, U_k^{(1)}\rangle \\ &- \frac{u^{3/2}}{2}\sigma_k^{(1)}(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\langle D^2\phi(x_{k-1}); \nabla f(x_{k-1}), U_k^{(1)}\rangle \\ &+ R_2^{(k)}\end{aligned}$$

where

$$\begin{aligned}R_2^{(k)} = &\frac{1}{6}(\frac{1-e^{-2\gamma_k}}{2})^3 D^3\phi(x_{k-1})v_{k-1}^{\otimes 3} - \frac{u}{4}(\frac{1-e^{-2\gamma_k}}{2})^2(\gamma_k - \frac{1-e^{-2\gamma_k}}{2})\langle D^3\phi(x_{k-1}); v_{k-1}^{\otimes 2}, \nabla f(x_{k-1})\rangle \\ &+ \frac{\sqrt{u}}{2}\sigma_k^{(1)}(\frac{1-e^{-2\gamma_k}}{2})^2\langle D^3\phi(x_{k-1}); v_{k-1}^{\otimes 2}, U_k^{(1)}\rangle + \frac{u}{2}\sigma_k^{(1)2}\frac{1-e^{-2\gamma_k}}{2}\langle D^3\phi(x_{k-1}); v_{k-1}, U_k^{(1)^{\otimes 2}}\rangle \\ &+ \frac{1}{24}(\frac{1-e^{-2\gamma_k}}{2})^4 D^4\phi(x_{k-1})v_{k-1}^{\otimes 4} + r^{(k)}.\end{aligned}$$

Since f is gradient Lipschitz and strongly convex, we've shown $(\mathcal{L}_{V,\infty})$ holds. Using $(\mathcal{L}_{V,\infty})$ the fact that $D^4\phi$ is bounded and Lipschitz, we can show there exists a constant $C > 0$ such that

$$|r_k| \le C\gamma_k^{9/2}V^2(x_{k-1}, v_{k-1}).$$

Apply theorem 6 for $p = 4$ and $s = 1$, we have $\sup_n \nu_n^\gamma(V^2) < +\infty \quad \mathbb{P}$-a.s. Therefore

$$\frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n r^{(k)} \to 0 \qquad \text{in } \mathbb{L}^1.$$

In the following proof, we will use $o(\gamma_k^4)$ to denote the sum of those terms b_k such that $\frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n b_k \to 0 \quad \mathbb{P}$-a.s. According to our decomposition, we can pull out polynomials of γ_k from factors $\frac{1-e^{-2\gamma_k}}{2}$, $\gamma_k - \frac{1-e^{-2\gamma_k}}{2}$ and $\sigma_k^{(1)}$ so that the terms left could be included in $o(\gamma_k^4)$. Then we obtain

$$
\begin{aligned}
\sum_{k=1}^n \gamma_k \mathcal{L}\phi(x_{k-1}) = \sum_{k=1}^n \Big\{ & [\phi(x_k) - \phi(x_{k-1})] + (\gamma_k^2 - \frac{2}{3}\gamma_k^3 + \frac{1}{3}\gamma_k^4) v_{k-1} \cdot \nabla\phi(x_{k-1}) \\
& + \frac{u}{2}(\gamma_k^2 - \frac{2}{3}\gamma_k^3 + \frac{1}{3}\gamma_k^4)\nabla f(x_{k-1}) \cdot \nabla\phi(x_{k-1}) \\
& - \frac{2\sqrt{3u}}{3}\gamma_k^{\frac{3}{2}} \nabla\phi(x_{k-1}) \cdot U_k^{(1)} \\
& - \frac{1}{2}(\gamma_k^2 - 2\gamma_k^3 + \frac{7}{3}\gamma_k^4) D^2\phi(x_{k-1}) v_{k-1}^{\otimes 2} \\
& - \frac{u^2}{8}\gamma_k^4 D^2\phi(x_{k-1}) \nabla f(x_{k-1})^{\otimes 2} \\
& - \frac{u}{2}(\frac{4}{3}\gamma_k^3 - 2\gamma_k^4) D^2\phi(x_{k-1}) {U_k^{(1)}}^{\otimes 2} \\
& + \frac{u}{2}(\gamma_k^3 - \frac{5}{3}\gamma_k^4)\langle D^2\phi(x_{k-1}); v_{k-1}, \nabla f(x_{k-1})\rangle \\
& - \frac{1}{6}(\gamma_k^3 - 3\gamma_k^4) D^3\phi(x_{k-1}) v_{k-1}^{\otimes 3} \\
& + \frac{u}{4}\gamma_k^4 \langle D^3\phi(x_{k-1}); v_{k-1}^{\otimes 2}, \nabla f(x_{k-1})\rangle \\
& - \frac{2u}{3}\gamma_k^4 \langle D^3\phi(x_{k-1}); v_{k-1}, {U_k^{(1)}}^{\otimes 2}\rangle \\
& - \frac{1}{24}\gamma_k^4 D^4\phi(x_{k-1}) v_{k-1}^{\otimes 4} + o(\gamma_k^4) \Big\} \\
:= {} & Z_n^{(0)} + Z_n^{(2)} + Z_n^{(3)} + Z_n^{(4)} + N_n + r_n
\end{aligned}
$$

where

$$
\begin{aligned}
Z_n^{(0)} &= \phi(x_n) - \phi(x_0), \\
Z_n^{(2)} &= \sum_{k=1}^n \gamma_k^2 \left[v_{k-1} \cdot \nabla\phi(x_{k-1}) + \frac{u}{2} \nabla f(x_{k-1}) \cdot \nabla\phi(x_{k-1}) - \frac{1}{2} D^2\phi(x_{k-1}) v_{k-1}^{\otimes 2} \right] \\
&= \sum_{k=1}^n \gamma_k^2 z_{k-1}^{(2)}, \\
Z_n^{(3)} &= \sum_{k=1}^n \gamma_k^3 \Big[-\frac{2}{3} v_{k-1} \cdot \nabla\phi(x_{k-1}) - \frac{u}{3} \nabla f(x_{k-1}) \cdot \nabla\phi(x_{k-1}) + D^2\phi(x_{k-1}) v_{k-1}^2 \\
&\quad - \frac{2u}{3} D^2\phi(x_{k-1}) {U_k^{(1)}}^{\otimes 2} + \frac{u}{2} \langle D^2\phi(x_{k-1}); v_{k-1}, \nabla f(x_{k-1}) \rangle - \frac{1}{6} D^3\phi(x_{k-1}) v_{k-1}^{\otimes 3} \Big] \\
&= -\sum_{k=1}^n \gamma_k^3 z_{k-1}^{(3)}, \\
Z_n^{(4)} &= \sum_{k=1}^n \gamma_k^4 \Big[\frac{1}{3} v_{k-1} \cdot \nabla\phi(x_{k-1}) + \frac{u}{6} \nabla f(x_{k-1}) \cdot \nabla\phi(x_{k-1}) - \frac{7}{6} D^2\phi(x_{k-1}) v_{k-1}^{\otimes 2} \\
&\quad - \frac{u^2}{8} D^2\phi(x_{k-1}) \nabla f(x_{k-1})^{\otimes 2} + u D^2\phi(x_{k-1}) {U_k^{(1)}}^{\otimes 2} - \frac{5u}{6} \langle D^2\phi(x_{k-1}); v_{k-1}, \nabla f(x_{k-1}) \rangle \\
&\quad + \frac{1}{2} D^3\phi(x_{k-1}) v_{k-1}^{\otimes 3} + + \frac{u}{4} \langle D^3\phi(x_{k-1}); v_{k-1}^{\otimes 2}, \nabla f(x_{k-1}) \rangle - \frac{2u}{3} \langle D^3\phi(x_{k-1}); v_{k-1}, {U_k^{(1)}}^{\otimes 2} \\
&\quad - \frac{1}{24} D^4\phi(x_{k-1}) v_{k-1}^{\otimes 4} \Big] = \sum_{k=1}^n \gamma_k^4 z_{k-1}^{(4)}, \\
N_n &= \sum_{k=1}^n \frac{2\sqrt{3u}}{3} \gamma_k^{\frac{3}{2}} \nabla\phi(x_{k-1}) \cdot U_k^{(1)}, \\
r_n &= \sum_{k=1}^n o(\gamma_k^4).
\end{aligned}
$$

First, it's easy to see that $r_n/\Gamma_n^{(4)} \to 0$ $\mathbb{P}$-a.s. as $n \to +\infty$. Apply lemma 2.6.2 and we obtain $\sup_n \mathbb{E}[V(x_n, v_n)] < +\infty$. Therefore we can further obtain the tightness of sequence $\{x_n\}$ and it follows from the continuity of ϕ that $\{\phi(x_n)\}$ is also tight. According to the tightness, $Z_n^{(0)}/\Gamma_n^{(4)} \to 0$ $\mathbb{P}$-a.s. For $Z_n^{(4)}$, under our assumptions on ϕ and f, we can show that

$$\lim_{|(x_n, v_n)| \to +\infty} z_n^{(4)}/V^4(x_n, v_n) = 0.$$

Therefor apply theorem 9 with $p = 8,\ s = 1$ and we obtain:

$$\begin{aligned}\mathbb{P} - a.s \qquad Z_n^{(4)}/\Gamma_n^{(4)} \to & \frac{u}{4}\int_{\mathbb{R}^{2d}} \langle D^3\phi(x); \nabla f(x), v^{\otimes 2}\rangle \nu(dx, dv) - \frac{u^2}{8}\int_{\mathbb{R}^d} D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx) \\ & - \frac{1}{24}\int_{\mathbb{R}^{2d}} D^4\phi(x)v^{\otimes 4}\nu(dx, dv).\end{aligned}$$

To consider the limit of $Z_n^{(i)}/\Gamma_n^{(4)}$ for $i = 2, 3$, We first Taylor expand $\mathcal{L}\phi(x_{k-1})$ at x_{k-2}:

$$\begin{aligned}\mathcal{L}\phi(x_{k-1}) = & v_{k-2}\cdot\nabla\phi(x_{k-2}) + \langle D^2\phi(x_{k-2}); v_{k-2}, x_{k-1} - x_{k-2}\rangle + \nabla\phi(x_{k-2})\cdot(v_{k-1} - v_{k-2}) \\ & + \frac{1}{2}\langle D^3\phi(x_{k-2}); v_{k-2}, (x_{k-1} - x_{k-2})^{\otimes 2}\rangle + \langle D^2\phi(x_{k-2}); v_{k-1} - v_{k-2}, x_{k-1} - x_{k-2}\rangle \\ & + \frac{1}{6}\langle D^4\phi(x_{k-2}); v_{k-2}, (x_{k-1} - x_{k-2})^{\otimes 3}\rangle \\ & + \frac{1}{2}\langle D^3\phi(x_{k-2}); v_{k-1} - v_{k-2}, (x_{k-1} - x_{k-2})^{\otimes 2}\rangle \\ & + o(\gamma_k^3).\end{aligned}$$

Plug the discretization into the Taylor expansions and preserve the “large” terms, then we obtain:

$$\begin{aligned}\mathcal{L}\phi(x_{k-1}) = & \mathcal{L}\phi(x_{k-2}) + (\gamma_{k-1} - \gamma_{k-1}^2 + \frac{2}{3}\gamma_{k-1}^3)D^2\phi(x_{k-2})v_{k-2}^{\otimes 2} \\ & - \frac{u}{2}(\gamma_{k-1}^2 - \frac{2}{3}\gamma_{k-1}^3)\langle D^2\phi(x_{k-2}); v_{k-2}, \nabla f(x_{k-2})\rangle \\ & + \frac{2\sqrt{3u}}{3}\gamma_{k-1}^{\frac{3}{2}}\langle D^2\phi(x_{k-2}); v_{k-2}, U_{k-1}^{(1)}\rangle \\ & - (2\gamma_{k-1} - 2\gamma_{k-1}^2 + \frac{4}{3}\gamma_{k-1}^3)v_{k-2}\cdot\nabla\phi(x_{k-2}) \\ & - u(\gamma_{k-1} - \gamma_{k-1}^2 + \frac{2}{3}\gamma_{k-1}^3)\nabla f(x_{k-2})\cdot\nabla\phi(x_{k-2}) \\ & + 2\sqrt{u}\gamma_{k-1}^{\frac{1}{2}}\nabla\phi(x_{k-2})\cdot U_{k-1}^{(2)} \\ & + \frac{1}{2}(\gamma_{k-1}^2 - 2\gamma_{k-1}^3)D^3\phi(x_{k-2})v_{k-2}^{\otimes 3} \\ & + \frac{2u}{3}\gamma_{k-1}^3\langle D^3\phi(x_{k-2}); v_{k-2}, {U_{k-1}^{(1)}}^{\otimes 2}\rangle \\ & + \sqrt{u}\gamma_{k-1}^{\frac{5}{2}}\langle D^3\phi(x_{k-2}); v_{k-2}^{\otimes 2}, U_{k-1}^{(1)}\rangle \\ & - \frac{u}{2}\gamma_{k-1}^3\langle D^3\phi(x_{k-2}); v_{k-2}^{\otimes 2}, \nabla f(x_{k-2})\rangle \\ & - 2(\gamma_{k-2}^2 - 2\gamma_{k-1}^3)D^2\phi(x_{k-2})v_{k-2}^{\otimes 2}\end{aligned}$$

$$
\begin{aligned}
&- u(\gamma_{k-1}^2 - 2\gamma_{k-1}^3)\langle D^2\phi(x_{k-2}); v_{k-2}, \nabla f(x_{k-2})\rangle \\
&+ 2\sqrt{u}\gamma_{k-1}^{\frac{3}{2}}\langle D^2\phi(x_{k-2}); v_{k-2}, U_{k-1}^{(2)}\rangle \\
&+ u\gamma_{k-1}^3\langle D^2\phi(x_{k-2}); v_{k-2}, \nabla f(x_{k-2})\rangle \\
&+ \frac{u^2}{2}\gamma_{k-1}^3 D^2\phi(x_{k-2})\nabla f(x_{k-2})^{\otimes 2} \\
&- u^{\frac{3}{2}}\gamma_{k-1}^{\frac{5}{2}}\langle D^2\phi(x_{k-2}); \nabla f(x_{k-2}), U_{k-1}^{(2)}\rangle \\
&- 2\sqrt{u}\gamma_{k-1}^{\frac{5}{2}}\langle D^2\phi(x_{k-2}); v_{k-2}, U_{k-1}^{(1)}\rangle \\
&- u^{\frac{3}{2}}\gamma_{k-1}^{\frac{5}{2}}\langle D^2\phi(x_{k-2}); \nabla f(x_{k-2}), U_{k-1}^{(1)}\rangle \\
&+ \langle D^2\phi(x_{k-2}); \sqrt{u}\sigma_{k-1}^{(1)}U_{k-1}^{(1)}, 2\sqrt{u}\sigma_{k-1}^{(2)}U_{k-1}^{(2)}\rangle + o(\gamma_{k-1}^3).
\end{aligned}
$$

Apply theorem 9 with $p = 4, s = 1$ to the terms of order $o(V^2(x_{k-2}, v_{k-2}))$ in the decomposition. We obtain

$$
\begin{aligned}
\lim_n \frac{\sum_{k=2}^n \gamma_k \mathcal{L}\phi(x_{k-1})}{\Gamma_n^{(4)}} = \lim_n \frac{1}{\Gamma_n^{(4)}} \Bigg[&\sum_{k=2}^n \gamma_k \mathcal{L}\phi(x_{k-2}) + \sum_{k=2}^n \gamma_k(\gamma_{k-1} - 3\gamma_{k-1}^2) D^2\phi(x_{k-2}) v_{k-2}^{\otimes 2} \\
&- \sum_{k=2}^n \frac{3u}{2}\gamma_k\gamma_{k-1}^2 \langle D^2\phi(x_{k-2}); \nabla f(x_{k-2}), v_{k-2}\rangle \\
&- \sum_{k=2}^n \gamma_k(2\gamma_{k-1} - 2\gamma_{k-1}^2)\nabla\phi(x_{k-2}) \cdot v_{k-2} \\
&- \sum_{k=2}^n u\gamma_k(\gamma_{k-1} - \gamma_{k-1}^2)\nabla f(x_{k-2}) \cdot \nabla\phi(x_{k-2}) \\
&+ \sum_{k=2}^n \frac{1}{2}\gamma_k\gamma_{k-1}^2 D^3\phi(x_{k-2}) v_{k-2}^{\otimes 3} \\
&+ \sum_{k=2}^n 2\sqrt{u}\gamma_k\gamma_{k-1}^{\frac{1}{2}}\nabla\phi(x_{k-2}) \cdot U_{k-1}^{(2)} \\
&+ \sum_{k=2}^n \gamma_k \langle D^2\phi(x_{k-2}); \sqrt{u}\sigma_{k-1}^{(1)}U_{k-1}^{(1)}, 2\sqrt{u}\sigma_{k-1}^{(2)}U_{k-1}^{(2)}\rangle \Bigg] \\
&+ 4u\int_{\mathbb{R}^d} \Delta\phi(x)\pi(dx) - \frac{u}{2}\int_{\mathbb{R}^{2d}} \langle D^3\phi(x); v^{\otimes 2}, \nabla f(x)\rangle \nu(dx, dv) \\
&+ \frac{u^2}{2}\int_{\mathbb{R}^d} D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx).
\end{aligned}
$$

Since $\gamma_{k-1} - \gamma_k = o(\gamma_k^4)$, we can substitute all the γ_k on the right hand side with γ_{k-1} and it won't change the limits. For the last term inside the square bracket, notice that $\mathrm{Var}(\sqrt{u}\sigma_{k-1}^{(1)}U_{k-1}^{(1)}, 2\sqrt{u}\sigma_{k-1}^{(2)}U_{k-1}^{(2)}) = \frac{u}{2}(1 + e^{-4\gamma_{k-1}} - 2e^{-2\gamma_{k-1}})I_d \sim u(2\gamma_{k-1}^2 - 4\gamma_{k-1}^3)I_d$. Therefore

$$\begin{aligned}
&\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_k \langle D^2\phi(x_{k-2}); \sqrt{u}\sigma_{k-1}^{(1)}U_{k-1}^{(1)}, 2\sqrt{u}\sigma_{k-1}^{(2)}U_{k-1}^{(2)}\rangle \\
=&\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n 2u\gamma_{k-1}^3 \Delta\phi(x_{k-2}) - 4u \int_{\mathbb{R}^d} \Delta\phi(x)\pi(dx).
\end{aligned}$$

We can rewrite the equation as

$$\begin{aligned}
\lim_n \frac{\sum_{k=2}^n \gamma_k \mathcal{L}\phi(x_{k-1})}{\Gamma_n^{(4)}} &= \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_k \mathcal{L}\phi(x_{k-2}) + \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n 2\sqrt{u}\gamma_{k-1}^{\frac{3}{2}} \nabla\phi(x_{k-2}) \cdot U_{k-1}^{(2)} \\
&+ \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^2 [D^2\phi(x_{k-2})v_{k-2}^{\otimes 2} - 2\nabla\phi(x_{k-2}) \cdot v_{k-2} - u\nabla f(x_{k-2}) \cdot \nabla\phi(x_{k-2})] \\
&+ \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^3 [-3D^{\phi}(x_{k-2})v_{k-2}^{\otimes 2} - \frac{3u}{2}\langle D^2\phi(x_{k-2}); \nabla f(x_{k-2}), v_{k-2}\rangle \\
&+ 2\nabla\phi(x_{k-2}) \cdot v_{k-2} + u\nabla f(x_{k-2}) \cdot \nabla\phi(x_{k-2}) + \frac{1}{2}D^3\phi(x_{k-2})v_{k-2}^{\otimes 3} + 2u\Delta\phi(x_{k-2})] \\
&- \frac{u}{2}\int_{\mathbb{R}^{2d}} \langle D^3\phi(x); v^{\otimes 2}, \nabla f(x)\rangle \nu(dx, dv) + \frac{u^2}{2}\int_{\mathbb{R}^d} D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx) \\
&= \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_k \mathcal{L}\phi(x_{k-2}) + \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n 2\sqrt{u}\gamma_{k-1}^{\frac{3}{2}} \nabla\phi(x_{k-2}) \cdot U_{k-1}^{(2)} \\
&+ \lim_n \frac{1}{\Gamma_n^{(4)}}(-2Z_n^{(2)} - 3Z_n^{(3)}) \\
&- \frac{u}{2}\int_{\mathbb{R}^{2d}} \langle D^3\phi(x); v^{\otimes 2}, \nabla f(x)\rangle \nu(dx, dv) + \frac{u^2}{2}\int_{\mathbb{R}^d} D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx).
\end{aligned}$$

We can instantly get that

$$\begin{aligned}
\lim_n \frac{1}{\Gamma_n^{(4)}}(2Z_2^{(n)} + 3Z_n^{(3)}) = \lim_n \frac{1}{\Gamma_n^{(4)}} &\sum_{k=2}^n 2\sqrt{u}\gamma_{k-1}^{\frac{3}{2}} \nabla\phi(x_{k-2}) \cdot U_{k-1}^{(2)} \\
&+ \frac{u^2}{2}\int_{\mathbb{R}^d} D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx) \\
&- \frac{u}{2}\int_{\mathbb{R}^{2d}} \langle D^3\phi(x); \nabla f(x), v^{\otimes 2}\rangle \nu(dx, dv).
\end{aligned}$$

Similarly, apply Taylor expansion to $z_{k-1}^{(2)}$ at x_{k-2}, we achieve:

$$\begin{aligned}
\nabla f(x_{k-1}) \cdot \nabla \phi(x_{k-1}) = {} & \nabla f(x_{k-2}) \cdot \nabla \phi(x_{k-2}) + \langle D^2 f(x_{k-2}); \nabla \phi(x_{k-2}), x_{k-1} - x_{k-2} \rangle \\
& + \langle D^2 \phi(x_{k-2}); \nabla f(x_{k-2}), x_{k-1} - x_{k-2} \rangle \\
& + \frac{1}{2} D^3 (\nabla f \cdot \nabla \phi)(x_{k-2})(x_{k-1} - x_{k-2})^{\otimes 2} + o(\gamma_k^2),
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{2} D^2 \phi(x_{k-1}) v_{k-1}^{\otimes 2} = {} & \frac{1}{2} D^2 \phi(x_{k-2}) v_{k-2}^{\otimes 2} + \frac{1}{2} \langle D^3 \phi(x_{k-1}); v_{k-2}^{\otimes 2}, x_{k-1} - x_{k-2} \rangle \\
& + \langle D^2 \phi(x_{k-2}); v_{k-2}, v_{k-1} - v_{k-2} \rangle + \frac{1}{2} D^2 \phi(x_{k-2})(v_{k-1} - v_{k-2})^{\otimes 2} \\
& + \frac{1}{4} (\frac{1 - e^{-2\gamma_{k-1}}}{2})^2 D^4 \phi(x_{k-2}) v_{k-2}^{\otimes 4} \\
& + \frac{1}{2} \langle D^3 \phi(x_{k-2}); v_{k-2}, x_{k-1} - x_{k-2}, v_{k-1} - v_{k-2} \rangle \\
& + \frac{1}{6} \langle D^3 \phi(x_{k-2}); x_{k-1} - x_{k-2}, (v_{k-1} - v_{k-2})^{\otimes 2} \rangle + o(\gamma_k^2).
\end{aligned}$$

Simplifying the coefficients lead us to

$$\begin{aligned}
\nabla f(x_{k-1}) \cdot \nabla \phi(x_{k-1}) = {} & \nabla f(x_{k-2}) \cdot \nabla \phi(x_{k-2}) + (\gamma_{k-1} - \gamma_{k-1}^2) \langle D^2 f(x_{k-2}); \nabla \phi(x_{k-2}), v_{k-2} \rangle \\
& - \frac{u}{2} \gamma_{k-1}^2 \langle D^2 f(x_{k-2}); \nabla \phi(x_{k-2}), \nabla f(x_{k-2}) \rangle \\
& + \frac{2\sqrt{3u}}{3} \gamma_{k-1}^{\frac{3}{2}} \langle D^2 f(x_{k-2}); \nabla \phi(x_{k-2}), U_{k-1}^{(1)} \rangle \\
& + (\gamma_{k-1} - \gamma_{k-1}^2) \langle D^2 \phi(x_{k-2}); \nabla f(x_{k-2}), v_{k-2} \rangle \\
& - \frac{u}{2} \gamma_{k-1}^2 D^2 \phi(x_{k-2}) \nabla f(x_{k-2})^{\otimes 2} + \sqrt{u} \gamma_{k-1}^{\frac{3}{2}} \langle D^2 \phi(x_{k-2}); \nabla f(x_{k-2}), U_{k-1}^{(1)} \rangle \\
& + \frac{1}{2} \gamma_{k-1}^2 (D^3 f \nabla \phi + 2 D^2 \phi D^2 f + D^3 \phi \nabla f)(x_{k-2}) v_{k-2}^{\otimes 2} + o(\gamma_{k-1}^2),
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{2}D^2\phi(x_{k-1})v_{k-1}^{\otimes 2} = {} & \frac{1}{2}D^2\phi(x_{k-2})v_{k-2}^{\otimes 2} + \frac{1}{2}(\gamma_{k-1} - \gamma_{k-1}^2)D^3\phi(x_{k-2})v_{k-2}^{\otimes 3} \\
& - \frac{u}{4}\gamma_{k-1}^2\langle D^3\phi(x_{k-2}); v_{k-2}^{\otimes 2}, \nabla f(x_{k-2})\rangle + \frac{\sqrt{u}}{2}\gamma_{k-1}^{\frac{3}{2}}\langle D^3\phi(x_{k-2}); v_{k-2}^{\otimes 2}, U_{k-1}^{(1)}\rangle \\
& - 2(\gamma_{k-1} - \gamma_{k-1}^2)D^2\phi(x_{k-2})v_{k-2}^{\otimes 2} - u(\gamma_{k-1} - \gamma_{k-1}^2)\langle D^2\phi(x_{k-2}); \nabla f(x_{k-2}), v_{k-2}\rangle \\
& + 2\sqrt{u}\gamma_{k-1}^{\frac{1}{2}}\langle D^2\phi(x_{k-2}); v_{k-2}, U_{k-1}^{(2)}\rangle + 2\gamma_{k-1}^2 D^2\phi(x_{k-2})v_{k-2}^{\otimes 2} \\
& + \frac{u^2}{2}\gamma_{k-1}^2 D^2\phi(x_{k-2})\nabla f(x_{k-2})^{\otimes 2} + 2u(\gamma_{k-1} - 2\gamma_{k-1}^2)D^2\phi(x_{k-2}){U_{k-1}^{(2)}}^{\otimes 2} \\
& + 2u\gamma_{k-1}^2\langle D^2\phi(x_{k-2}); \nabla f(x_{k-2}), v_{k-2}\rangle - 4\sqrt{u}\gamma_{k-1}^{\frac{3}{2}}\langle D^2\phi(x_{k-2}); v_{k-2}, U_{k-1}^{(2)}\rangle \\
& - 2u^{\frac{3}{2}}\gamma_{k-1}^{\frac{3}{2}}\langle D^2\phi(x_{k-2}); \nabla f(x_{k-2}), U_{k-1}^{(2)}\rangle + \frac{1}{4}\gamma_{k-1}^2 D^4\phi(x_{k-2})v_{k-2}^{\otimes 4} \\
& - \gamma_{k-1}^2 D^3\phi(x_{k-2})v_{k-2}^{\otimes 3} - \frac{u}{2}\gamma_{k-1}^2\langle D^3\phi(x_{k-2}); v_{k-2}^{\otimes 2}, \nabla f(x_{k-2})\rangle \\
& + \frac{1}{2}\langle D^3\phi(x_{k-2}); v_{k-2}, \sqrt{u}\sigma_{k-1}^{(1)}U_{k-1}^{(1)}, 2\sqrt{u}\sigma_{k-1}^{(2)}U_{k-1}^{(12}\rangle \\
& + \frac{2u}{3}\gamma_{k-1}^2\langle D^3\phi(x_{k-1}); v_{k-2}; {U_{k-1}^{(2)}}^{\otimes 2}\rangle + o(\gamma_k^2).
\end{aligned}$$

Take the limits and we obtain:

$$\begin{aligned}
\lim_n \frac{\sum_{k=2}^n \gamma_k^2 \nabla f(x_{k-1}) \cdot \nabla\phi(x_{k-1})}{\Gamma_n^{(4)}} = {} & \lim_n \frac{1}{\Gamma_n^{(4)}}\Bigg[\sum_{k=2}^n \gamma_{k-1}^2 \nabla f(x_{k-2}) \cdot \nabla\phi(x_{k-2}) \\
& + \sum_{k=2}^n \gamma_{k-1}^3\langle D^2 f(x_{k-2}); \nabla\phi(x_{k-2}), v_{k-2}\rangle \\
& + \sum_{k=2}^n \gamma_{k-1}^3\langle D^2\phi(x_{k-2}); \nabla f(x_{k-2}), v_{k-2}\rangle\Bigg] \\
& - \frac{u}{2}\int_{\mathbb{R}^d}\langle D^2 f(x); \nabla\phi(x), \nabla f(x)\rangle\pi(dx) \\
& - \frac{u}{2}\int_{\mathbb{R}^d} D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx) \\
& + \frac{1}{2}\int_{\mathbb{R}^d}\left(D^3 f(x)\nabla\phi(x) + 2D^2 f(x)D^2\phi(x) + D^3\phi(x)\nabla f(x)\right)v^{\otimes 2}\nu(dx, dv),
\end{aligned}$$

and

$$\begin{aligned}\lim_n \frac{1}{2}\frac{\sum_{k=2}^n \gamma_k^2 D^2\phi(x_{k-1})v_{k-1}^{\otimes 2}}{\Gamma_n^{(4)}} &= \lim_n \Bigg\{ \frac{1}{\Gamma_n^{(4)}}\sum_{k=2}^n \frac{1}{2}\gamma_{k-1}^2 D^2\phi(x_{k-2})v_{k-2}^{\otimes 2} \\ &+ \frac{1}{\Gamma_n^{(4)}}\sum_{k=2}^n \gamma_{k-1}^3\Bigg[\frac{1}{2}D^3\phi(x_{k-2})v_{k-2}^{\otimes 3} - 2D^2\phi(x_{k-2})v_{k-2}^{\otimes 2} \\ &-u\langle D^2\phi(x_{k-2});\nabla f(x_{k-2}),v_{k-2}\rangle + 2uD^2\phi(x_{k-2}){U_{k-1}^{(2)}}^{\otimes 2}\Bigg]\Bigg\} \\ &- \frac{3u}{4}\int_{\mathbb{R}^{2d}}\langle D^3\phi(x);\nabla f(x),v^{\otimes 2}\rangle\nu(dx,dv) + \frac{1}{4}\int_{\mathbb{R}^{2d}} D^4\phi(x)v^{\otimes 4}\pi(dx) \\ &+ \frac{u^2}{2}\int_{\mathbb{R}^d} D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx).\end{aligned}$$

Claim:

$$a)\lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n \gamma_k^2\nabla\phi(x_{k-1})\cdot v_{k-1} = 0.$$

$$b)\lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n \gamma_k^3(\frac{u}{2}\nabla\phi(x_{k-1})\cdot\nabla f(x_{k-1}) - \frac{1}{2}D^2\phi(x_{k-1})v_{k-1}^{\otimes 2}) = 0.$$

$$c)\lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n \gamma_k^3\nabla\phi(x_{k-1})\cdot v_{k-1} = 0.$$

We'll prove the **Claim** at the end of our proof. We can use the **Claim** and our expansion of $Z_n^{(2)}$ to find the following relation:

$$\begin{aligned}\lim_n \frac{1}{\Gamma_n^{(4)}}Z_n^{(2)} &= \lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=2}^n \gamma_k^2\nabla\phi(x_{k-1})\cdot v_{k-1} \\ &+ \lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=2}^n \gamma_k^2\left[\frac{u}{2}\nabla f(x_{k-1})\cdot\nabla\phi(x_{k-1}) - \frac{1}{2}D^2\phi(x_{k-1})v_{k-1}^{\otimes 2}\right] \\ &= \lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=2}^n \gamma_{k-1}^2\left[\frac{u}{2}\nabla f(x_{k-2})\cdot\nabla\phi(x_{k-2}) - \frac{1}{2}D^2\phi(x_{k-2})v_{k-2}^{\otimes 2}\right] \\ &+ \lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=2}^n \gamma_{k-1}^3\Bigg[\frac{u}{2}\langle D^2 f(x_{k-2});\nabla\phi(x_{k-2}),v_{k-2}\rangle + \frac{3u}{2}\langle D^2\phi(x_{k-2});\nabla f(x_{k-2}),v_{k-2}\rangle \\ &-\frac{1}{2}D^3\phi(x_{k-2})v_{k-2}^{\otimes 3} + 2D^2\phi(x_{k-2})v_{k-2}^{\otimes 2} - 2u\Delta\phi(x_{k-2})\Bigg]\end{aligned}$$

$$\begin{aligned}
&-\frac{u^2}{4}\int_{\mathbb{R}^d}\langle D^2 f(x);\nabla\phi(x),\nabla f(x)\rangle\pi(dx)-\frac{u^2}{4}\int_{\mathbb{R}^d}D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx)\\
&+\frac{u}{4}\int_{\mathbb{R}^d}\left(D^3 f(x)\nabla\phi(x)+2D^2 f(x)D^2\phi(x)+D^3\phi(x)\nabla f(x)\right)v^{\otimes 2}\nu(dx,dv)\\
&+\frac{3u}{4}\int_{\mathbb{R}^{2d}}\langle D^3\phi(x);\nabla f(x),v^{\otimes 2}\rangle\nu(dx,dv)-\frac{1}{4}\int_{\mathbb{R}^{2d}}D^4\phi(x)v^{\otimes 4}\pi(dx)\\
&-\frac{u^2}{2}\int_{\mathbb{R}^d}D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx)\\
&=\lim_n\frac{1}{\Gamma_n^{(4)}}[Z_n^{(2)}+3Z_n^{(3)}]-\frac{u^2}{4}\int_{\mathbb{R}^d}\langle D^2 f(x);\nabla\phi(x),\nabla f(x)\rangle\pi(dx)\\
&-\frac{3u^2}{4}\int_{\mathbb{R}^d}D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx)-\frac{1}{4}\int_{\mathbb{R}^{2d}}D^4\phi(x)v^{\otimes 4}\pi(dx)\\
&+\frac{u}{4}\int_{\mathbb{R}^d}\left(D^3 f(x)\nabla\phi(x)+2D^2 f(x)D^2\phi(x)+4D^3\phi(x)\nabla f(x)\right)v^{\otimes 2}\nu(dx,dv).
\end{aligned}$$

The last identity follows from **Claim**-a),b) and the fact that $\lim_n\frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n\gamma_k^3\langle D^2 f(x_{k-1});\nabla\phi(x_{k-1}),v_{k-1}\rangle=0$. To prove $\lim_n\frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n\gamma_k^3\langle D^2 f(x_{k-1});\nabla\phi(x_{k-1}),v_{k-1}\rangle=0$, we can assume ψ is a new test function satisfying $\nabla\psi(x)=D^2 f(x)\nabla\phi(x)$. Then the statement follows from **Claim**-c). This could be done because ψ satisfies the all assumptions on ϕ stated in the theorem. Therefore we obtain

$$\begin{aligned}
\lim_n\frac{1}{\Gamma_n^{(4)}}Z_n^{(3)}&=\frac{u^2}{12}\int_{\mathbb{R}^d}\langle D^2 f(x);\nabla\phi(x),\nabla f(x)\rangle\pi(dx)+\frac{u^2}{4}\int_{\mathbb{R}^d}D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx)\\
&-\frac{u}{12}\int_{\mathbb{R}^d}\left(D^3 f(x)\nabla\phi(x)+2D^2 f(x)D^2\phi(x)+4D^3\phi(x)\nabla f(x)\right)v^{\otimes 2}\nu(dx,dv)\\
&+\frac{1}{12}\int_{\mathbb{R}^{2d}}D^4\phi(x)v^{\otimes 4}\pi(dx).
\end{aligned}$$

Combine with our previous results on $2Z_n^{(2)}+3Z_n^{(3)}$ and we obtain

$$\begin{aligned}
\lim_n\frac{1}{\Gamma_n^{(4)}}[Z_n^{(2)}+Z_n^{(3)}]&=\lim_n\frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n\sqrt{u}\gamma_k^{\frac{3}{2}}\nabla\phi(x_{k-1})\cdot U_k^{(2)}+\frac{u^2}{8}\int_{\mathbb{R}^d}D^2\phi(x)\nabla f(x)^{\otimes 2}\pi(dx)\\
&-\frac{u}{12}\int_{\mathbb{R}^{2d}}\langle D^3\phi(x);\nabla f(x),v^{\otimes 2}\rangle\nu(dx,dv)+\frac{u}{24}\int_{\mathbb{R}^{2d}}\langle D^3 f(x);\nabla\phi(x),v^{\otimes 2}\rangle\nu(dx,dv)\\
&+\frac{u}{12}\int_{\mathbb{R}^{2d}}(D^2 f D^2\phi)(x)v^{\otimes 2}\nu(dx,dv)-\frac{u^2}{24}\int_{\mathbb{R}^d}\langle D^2 f(x);\nabla\phi(x),\nabla f(x)\rangle\pi(dx)\\
&-\frac{1}{24}\int_{\mathbb{R}^{2d}}D^4\phi(x)v^{\otimes 4}\pi(dx).
\end{aligned}$$

Then we plug this result in our original decomposition:

$$\begin{aligned}
\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^n \gamma_k \mathcal{L}\phi(x_{k-1}) &= \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^n \gamma_k^{\frac{3}{2}} \frac{2\sqrt{3}}{3} \nabla\phi(x_{k-1}) \cdot (\sqrt{u} U_k^{(1)}) \\
&+ \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^n \sqrt{u}\gamma_k^{\frac{3}{2}} \nabla\phi(x_{k-1}) \cdot U_k^{(2)} + \frac{u^2}{8} \int_{\mathbb{R}^d} D^2\phi(x) \nabla f(x)^{\otimes 2} \pi(dx) \\
&- \frac{u}{12} \int_{\mathbb{R}^{2d}} \langle D^3\phi(x); \nabla f(x), v^{\otimes 2}\rangle \nu(dx, dv) + \frac{u}{24} \int_{\mathbb{R}^{2d}} \langle D^3 f(x); \nabla\phi(x), v^{\otimes 2}\rangle \nu(dx, dv) \\
&+ \frac{u}{12} \int_{\mathbb{R}^{2d}} (D^2 f D^2\phi)(x) v^{\otimes 2} \nu(dx, dv) - \frac{u^2}{24} \int_{\mathbb{R}^d} \langle D^2 f(x); \nabla\phi(x), \nabla f(x)\rangle \pi(dx) \\
&- \frac{1}{24} \int_{\mathbb{R}^{2d}} D^4\phi(x) v^{\otimes 4} \pi(dx) + \frac{u}{4} \int_{\mathbb{R}^{2d}} \langle D^3\phi(x); \nabla f(x), v^{\otimes 2}\rangle \nu(dx, dv) \\
&- \frac{u^2}{8} \int_{\mathbb{R}^d} D^2\phi(x) \nabla f(x)^{\otimes 2} \pi(dx) - \frac{1}{24} \int_{\mathbb{R}^{2d}} D^4\phi(x) v^{\otimes 4} \nu(dx, dv) \\
&= \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^n \gamma_k^{\frac{3}{2}} \nabla\phi(x_{k-1}) \cdot (\frac{2\sqrt{3}}{3} \sqrt{u} U_k^{(1)} + \frac{1}{2} 2\sqrt{u} U_k^{(2)}) \\
&+ \frac{u}{6} \int_{\mathbb{R}^{2d}} \langle D^3\phi(x); \nabla f(x), v^{\otimes 2}\rangle \nu(dx, dv) + \frac{u}{24} \int_{\mathbb{R}^{2d}} \langle D^3 f(x); \nabla\phi(x), v^{\otimes 2}\rangle \nu(dx, dv) \\
&+ \frac{u}{12} \int_{\mathbb{R}^{2d}} (D^2\phi D^2 f)(x) v^{\otimes 2} \nu(dx, dv) - \frac{1}{12} \int_{\mathbb{R}^{2d}} D^4\phi(x) v^{\otimes 4} \nu(dx, dv) \\
&- \frac{u^2}{24} \int_{\mathbb{R}^d} \langle D^2 f(x); \nabla\phi(x), \nabla f(x)\rangle \pi(dx).
\end{aligned}$$

It remains to determine the normal limit. Since $(U_k^{(1)}, U_k^{(2)})$ is Gaussian in $\mathbb{R}^{2d}$ with mean zero and covariance matrix $\frac{1+e^{-4\gamma_k} - 2e^{-2\gamma_k}}{4\sigma_k^{(1)}\sigma_k^{(2)}} I_d$, we can find the distribution of $U_k := (\frac{2\sqrt{3}}{3}\sqrt{u}U_k^{(1)} + \frac{1}{2}2\sqrt{u}U_k^{(2)})$. $\{U_k\}$ are independent $2d$-Gaussian Random vectors with $U_k \sim \mathcal{N}(0, \Sigma_k)$, where

$$\begin{aligned}
\Sigma_k &= \mathbb{E}[(\frac{2\sqrt{3}}{3}\sqrt{u}U_k^{(1)} + \frac{1}{2}2\sqrt{u}U_k^{(2)})^T (\frac{2\sqrt{3}}{3}\sqrt{u}U_k^{(1)} + \frac{1}{2}2\sqrt{u}U_k^{(2)})] \\
&= \frac{4u}{3} I_d + \frac{4u\sqrt{3}}{3} \frac{1 + e^{-4\gamma_k} - 2e^{-2\gamma_k}}{4\sigma_k^{(1)}\sigma_k^{(2)}} I_d + u I_d \\
&\sim \frac{10}{3} u I_d + O(\gamma_k) I_d.
\end{aligned}$$

Apply our weak convergence result and CLT for arrays of square-integrable martingale increments, we have that when $0 < \hat{\gamma} < +\infty$:

$$\frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^{n} \gamma_k^{\frac{3}{2}} \nabla\phi(x_{k-1}) \cdot U_k \implies \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \lim_n \frac{1}{{\Gamma_n^{(4)}}^2} \sum_{k=1}^{n} \gamma_k^3 |\nabla\phi(x_{k-1})|^2 (\frac{10}{3}u + O(\gamma_k)) = \frac{10}{3} u \hat{\gamma}^{-2} \int_{\mathbb{R}^d} |\nabla\phi(x)|^2 \pi(dx).$$

In conclusion, when $\hat{\gamma} \in (0, +\infty)$:

$$\frac{\Gamma_n}{\Gamma_n^{(4)}} \nu_n^{\gamma}(\mathcal{L}\phi) \implies \mathcal{N}(\rho, \frac{10}{3} u \hat{\gamma}^{-2} \int_{\mathbb{R}^d} |\nabla\phi(x)|^2 \pi(dx))$$

where

$$\begin{aligned}
\rho = & \frac{u}{6} \int_{\mathbb{R}^{2d}} \langle D^3\phi(x); \nabla f(x), v^{\otimes 2} \rangle \nu(dx, dv) + \frac{u}{24} \int_{\mathbb{R}^{2d}} \langle D^3 f(x); \nabla\phi(x), v^{\otimes 2} \rangle \nu(dx, dv) \\
& + \frac{u}{12} \int_{\mathbb{R}^{2d}} (D^2\phi D^2 f)(x) v^{\otimes 2} \nu(dx, dv) - \frac{1}{12} \int_{\mathbb{R}^{2d}} D^4\phi(x) v^{\otimes 4} \nu(dx, dv) \\
& - \frac{u^2}{24} \int_{\mathbb{R}^d} \langle D^2 f(x); \nabla\phi(x), \nabla f(x) \rangle \pi(dx).
\end{aligned}$$

When $\hat{\gamma} = 0$,

$$\frac{\Gamma_n}{\sqrt{\Gamma_n^{(3)}}} \nu_n^{\gamma}(\mathcal{L}\phi) \implies \mathcal{N}(0, \frac{10}{3} u \int_{\mathbb{R}^d} |\nabla\phi(x)|^2 \pi(dx)).$$

When $\hat{\gamma} = +\infty$,

$$\frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^{n} \gamma_k^{\frac{3}{2}} \nabla\phi(x_{k-1}) \cdot (\frac{2\sqrt{3}}{3} \sqrt{u} U_k^{(1)} + \frac{1}{2} 2\sqrt{u} U_k^{(2)}) \to 0 \qquad \text{in probability.}$$

Therefore when $\hat{\gamma} = +\infty$,

$$\frac{\Gamma_n}{\Gamma_n^{(4)}} \nu_n^{\gamma}(\mathcal{L}\phi) \to \rho \qquad\qquad \text{in probability.}$$

Proof of the claim: First we'll show that $\frac{1}{\Gamma_n^{(3)}} \sum_{k=1}^n \gamma_k^2 \mathcal{L}\phi(x_{k-1}) \to 0$. We can use our decomposition of $\mathcal{L}\phi(x_{k-1})$ and obtain:

$$\begin{aligned}\sum_{k=1}^n \gamma_k^2 \mathcal{L}\phi(x_{k-1}) = \sum_{k=1}^n \Big\{ &\gamma_k \left(\phi(x_k) - \phi(x_{k-1})\right) + \gamma_k^3 v_{k-1} \cdot \nabla\phi(x_{k-1}) \\ &+ \frac{u}{2}\gamma_k^3 \nabla f(x_{k-1}) \cdot \nabla\phi(x_{k-1}) - \frac{1}{2}\gamma_k^3 D^2\phi(x_{k-1}) v_{k-1}^{\otimes 2} \Big\}.\end{aligned}$$

Since $\gamma_{k-1} - \gamma_k \sim o(\gamma_k^4)$ and $\{\phi(x_n)\}$ is tight, $\frac{1}{\Gamma_n^{(3)}} \sum_{k=1}^n \gamma_k \left(\phi(x_k) - \phi(x_{k-1})\right) \to 0$. Then we can apply theorem 9 with $p = 6, s = 1$ and obtain

$$\begin{aligned}\frac{1}{\Gamma_n^{(3)}} \sum_{k=1}^n \gamma_k^2 \mathcal{L}\phi(x_{k-1}) \to &\int_{\mathbb{R}^{2d}} v \cdot \nabla\phi(x) \nu(dx, dv) + \frac{u}{2} \int_{\mathbb{R}^d} \nabla\phi(x) \cdot \nabla f(x) \pi(dx) \\ &- \frac{1}{2} \int_{\mathbb{R}^{2d}} D^2\phi(x) v^{\otimes 2} \nu(dx, dv) \\ &= 0.\end{aligned}$$

The last identity follows from integration by parts and Fubini theorem. In the same way, we can also prove $\frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^n \gamma_k^3 \mathcal{L}\phi(x_{k-1}) \to 0$. Next, we'll show $\lim_n \frac{1}{\Gamma_n^{(3)}} \sum_{k=1}^n \gamma_k^2 (\frac{u}{2} \nabla\phi(x_{k-1}) \cdot \nabla f(x_{k-1}) - \frac{1}{2} D^2\phi(x_{k-1}) v_{k-1}^{\otimes 2}) = 0$, we'll use the same trick as we did in the proof of theorem 5. We Taylor expand $\mathcal{L}\phi(x_{k-1})$ at (x_{k-2}, v_{k-2}):

$$\begin{aligned}\gamma_k^2 \mathcal{L}\phi(x_{k-1}) = &\gamma_k^2 \mathcal{L}\phi(x_{k-2}) + \gamma_k^2 (\gamma_{k-1} - \gamma_{k-1}^2) D^2\phi(x_{k-2}) v_{k-2}^{\otimes 2} \\ &- \frac{u}{2} \gamma_k^2 \gamma_{k-1}^2 \langle D^2\phi(x_{k-2}); v_{k-2}, \nabla f(x_{k-2}) \rangle \\ &- \gamma_k^2 (2\gamma_{k-1} - 2\gamma_{k-1}^2) v_{k-2} \cdot \nabla\phi(x_{k-2}) \\ &- u\gamma_k^2 (\gamma_{k-1} - \gamma_{k-1}^2) \nabla f(x_{k-2}) \cdot \nabla\phi(x_{k-2}) \\ &+ \frac{1}{2} \gamma_k^2 \gamma_{k-1}^2 D^3\phi(x_{k-2}) v_{k-2}^{\otimes 3} - 2\gamma_k^2 \gamma_{k-2}^2 D^2\phi(x_{k-2}) v_{k-2}^{\otimes 2} \\ &- u\gamma_k^2 \gamma_{k-1}^2 \langle D^2\phi(x_{k-2}); v_{k-2}, \nabla f(x_{k-2}) \rangle \\ &+ \gamma_k^2 \langle D^2\phi(x_{k-2}); \sqrt{u}\sigma_{k-1}^{(1)} U_{k-1}^{(1)}, 2\sqrt{u}\sigma_{k-1}^{(2)} U_{k-1}^{(2)} \rangle + o(\gamma_{k-1}^3).\end{aligned}$$

Since $\gamma_{k-1} - \gamma_k = o(\gamma_k^4)$, we can change γ_k on the left hand side to γ_{k-1} when we take limits with scale $\Gamma_n^{(4)}$. Apply theorem 9 with $p = 8, s = 1$ to terms with order $o(\gamma_k^3)$-coefficients.

$$\begin{aligned}\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_k^2 \mathcal{L}\phi(x_{k-1}) = &\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^2 \mathcal{L}\phi(x_{k-2}) - 2\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^3 \mathcal{L}\phi(x_{k-2}) \\ &- 2\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^3 (\frac{u}{2}\nabla\phi(x_{k-2}) \cdot \nabla f(x_{k-2}) - \frac{1}{2} D^2\phi(x_{k-2}) v_{k-2}^{\otimes 2}) \\ &- 3\int_{\mathbb{R}^{2d}} D^2\phi(x) v^{\otimes 2} \nu(dx, dv) + u\int_{\mathbb{R}^d} \nabla\phi(x) \cdot \nabla f(x) \pi(dx) \\ &+ \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^2 \langle D^2\phi(x_{k-2}); \sqrt{u}\sigma_{k-1}^{(1)} U_{k-1}^{(1)}, 2\sqrt{u}\sigma_{k-1}^{(2)} U_{k-1}^{(2)} \rangle.\end{aligned}$$

Since we proved $\frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^n \gamma_k^3 \mathcal{L}\phi(x_{k-1}) \to 0$ and from Theorem 5, we've shown that

$$\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^2 \langle D^2\phi(x_{k-2}); \sqrt{u}\sigma_{k-1}^{(1)} U_{k-1}^{(1)}, 2\sqrt{u}\sigma_{k-1}^{(2)} U_{k-1}^{(2)} \rangle = 2u \int_{\mathbb{R}^d} \Delta\phi(x) \pi(dx).$$

We obtain

$$\begin{aligned}&\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^2 (\frac{u}{2}\nabla\phi(x_{k-2}) \cdot \nabla f(x_{k-2}) - \frac{1}{2} D^2\phi(x_{k-2}) v_{k-2}^{\otimes 2}) \\ &= \frac{1}{2}\left[\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_k^2 \mathcal{L}\phi(x_{k-1}) - \lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=2}^n \gamma_{k-1}^2 \mathcal{L}\phi(x_{k-2})\right] \\ &= 0.\end{aligned}$$

Therefore, $\lim_n \frac{1}{\Gamma_n^{(4)}} \sum_{k=1}^n \gamma_k^3 (\frac{u}{2}\nabla\phi(x_{k-1}) \cdot \nabla f(x_{k-1}) - \frac{1}{2} D^2\phi(x_{k-1}) v_{k-1}^{\otimes 2}) = 0$.

To prove the **Claim**, we need to use the decomposition again:

$$\begin{aligned}\sum_{k=1}^n \gamma_k^2 \mathcal{L}\phi(x_{k-1}) = \sum_{k=1}^n \Bigg\{ &\gamma_k[\phi(x_k) - \phi(x_{k-1})] + (\gamma_k^3 - \frac{2}{3}\gamma_k^4) v_{k-1} \cdot \nabla\phi(x_{k-1}) \\ &+ \frac{u}{2}(\gamma_k^3 - \frac{2}{3}\gamma_k^4) \nabla f(x_{k-1}) \cdot \nabla\phi(x_{k-1}) - \frac{1}{2}(\gamma_k^3 - 2\gamma_k^4) D^2\phi(x_{k-1}) v_{k-1}^{\otimes 2} \\ &- \frac{2u}{3}\gamma_k^4 D^2\phi(x_{k-1}) {U_k^{(1)}}^{\otimes 2} + \frac{u}{2}\gamma_k^4 \langle D^2\phi(x_{k-1}); v_{k-1}, \nabla f(x_{k-1}) \rangle \\ &- \frac{1}{6}\gamma_k^4 D^3\phi(x_{k-1}) v_{k-1}^{\otimes 3} + o(\gamma_k^4) \Bigg\}.\end{aligned}$$

Since $\{\phi(x_n)\}$ is tight and $\gamma_{k-1}-\gamma_k = o(\gamma_k^4)$, we have $\frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n \gamma_k(\phi(x_k)-\phi(x_{k-1})) \to 0$. For the terms with coefficients of order γ_k^3, we can apply theorem 9 with $p=8, s=1$. Then we obtain:

$$\begin{aligned}
\lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n \gamma_k^2 \mathcal{L}\phi(x_{k-1}) &= \lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n \gamma_k^3(\mathcal{L}\phi(x_{k-1}) + \frac{u}{2}\nabla\phi(x_{k-1})\cdot\nabla f(x_{k-1}) - \frac{1}{2}D^2\phi(x_{k-1})v_{k-1}^{\otimes 2}) \\
&- \frac{u}{3}\int_{\mathbb{R}^d}\nabla\phi(x)\cdot\nabla f(x)\pi(dx) + \int_{\mathbb{R}^{2d}} D^2\phi(x)v^{\otimes 2}\nu(dx,dv) \\
&- \frac{2u}{3}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} D^2\phi(x)z^{\otimes 2}\mu(dz)\pi(dx) \\
&= \lim_n \frac{1}{\Gamma_n^{(4)}}\sum_{k=1}^n \gamma_k^3(\mathcal{L}\phi(x_{k-1}) + \frac{u}{2}\nabla\phi(x_{k-1})\cdot\nabla f(x_{k-1}) - \frac{1}{2}D^2\phi(x_{k-1})v_{k-1}^{\otimes 2}) \\
&= 0.
\end{aligned}$$

The second identity follows from integration by parts and Fubini theorem. The last identity follows from the two statements we just proved. ■

CHAPTER 3

Heavy-tailed Sampling

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 [RC99, ADFDJ03, HLW06, BGJM11, MT12, LM16, DMPS18]; 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,

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dB_t, \tag{3.1}$$

where B_t is a d-dimensional Brownian motion. For example, the Unadjusted Langevin Algorithm [RDF78], the Metropolis Adjusted Langevin Algorithm [RT96, RR98] and the proximal sampler [TP18, LST21, VPD22] arise as different discretizations of (3.1). Under light-tailed assumptions, i.e. when the density π has exponentially fast decaying tails, the diffusion X_t in (3.1) converges exponentially fast to π as its stationary density, which motivates the use of discretizations of (3.1) 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 [KN04, JR07, Kam18], multiple comparison procedures [GBH04, GB09], Bayesian statistics [GJPS08, GLM18], and statistical machine learning [BZ17, NŞR19, ŞZTG20, DKTZ20], in this chapter, 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 (3.1) does not converge exponentially to its stationary density in various metrics of interest. In the following sections, we first give an overview of related work. Then we introduce and analyze two new methods, the transformed Langevin Monte Carlo and the Itô discretization, to sample from heavy-tail target densities.

3.1. Related Work

Non-asymptotic iteration complexity of different discretizations of (3.1) 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 [Dal17b, DM17, DK19, DMM19, LST20, SL19, HBE20, CDWY20, DMM19, DKRD19, LE20, CDWY20, CEL$^+$21] and references therein. In particular, [VW19, EH21, CEL$^+$21] analyzed the performance of ULA under various functional inequalities suited to light-tailed densities. Furthermore, the recent work of [BCE$^+$22] 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, [DCWY19, CLA$^+$21, WSC22], analyzed the Metropolis-Adjusted Langevin Algorithm (MALA) in light-tailed settings. The proximal sampler algorithm was analyzed under various light-tailed assumptions in [LST21, CCSW22]. The iteration complexity of the widely used Hamiltonian Monte Carlo algorithm and discretizations of underdamped Langevin diffusions were analyzed, for example, in [DRD20b, BREZ20, CDWY20, MCC$^+$21, Mon21, CLW21, WW22, CV22]. We also refer interested readers to [LW22, DL21] for non-asymptotic analyses of other MCMC algorithms used in practice in light-tailed settings.

In the context of heavy-tailed sampling, [Kam18] considered the scaling limits of appropriately modified Metropolis random walk in an asymptotic setting. [JG12] 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 [YŁR22]. It is also worth highlighting that [DBCD19, DGM20] and [BRZ19] 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 [ADW21] and [ALPW21] 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 [ŞZTG20, HMW21] and [ZZ22] 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. [CDV09] considered the iteration complexity of Metropolis random walk algorithm for sampling from s-concave distributions. [LWME19] analyzed a class of discretizations of general Itô diffusions that admit heavy-tailed equilibrium densities. A detailed comparison to [LWME19] is provided in Section 3.3.7. Recently, [MHFH$^+$23] established non-asymptotic convergence guarantees for LMC(AKa ULA) and the Langevin diffusion (3.1) in Rényi divergence for sufficiently smooth targets that satisfy a weak Poincaré inequality, including heavy-tail targets.

The recent works by [HKRC18, ZPFP20, CLGL$^+$20b, AC21, Jia21, LTVW22] 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 [LTVW22] is based on mean-squared analysis.

3.2. Transformed Langevin Monte Carlo

3.2.1. Motivations. Given a potential function $f : \mathbb{R}^d \to \mathbb{R}$, we consider the problem of sampling from the density

$$\pi(x) \coloneqq Z^{-1} e^{-f(x)}, \tag{3.2}$$

where $Z \coloneqq \int e^{-f(x)} dx$ is an (unknown) normalization constant. When the target density π is heavy-tailed, the solution to (3.1) is not exponentially ergodic, that is, the solution does not converge to the stationary density rapidly. Indeed [RT96, Theorem 2.4] shows that if $|\nabla f(x)| \to 0$ when $|x| \to \infty$, then the solution to (3.1) is *not* exponentially ergodic. In the other direction, standard results in the literature, for example [Wan06, BGL14] show that the solution to (3.1) being exponentially ergodic is equivalent to the density π satisfying the Poincaré inequality, which requires π to have exponentially decaying tails. Furthermore, [Wan06, Chapter 4] shows that when π has polynomially decaying tails, the convergence is only sub-exponential or polynomial.

Turning to time-discretizations of (3.1), the Euler discretization or the Unadjusted Langevin Algorithm (ULA) is given by

$$x_{n+1} = x_n - \gamma \nabla f(x_n) + \sqrt{2\gamma} u_{n+1}, \tag{3.3}$$

where (u_n) is a sequence of independent and identically distributed d-dimensional standard Gaussian vectors and $\gamma > 0$ is a user-defined step size parameter. Over the past decade, non-asymptotic oracle complexity analysis of ULA (and other related discretizations) have been studied intensively. We refer to [Dal17b, DM17, DK19, DMM19, LST20, DCWY19, SL19, HBE20, CDWY20, CLA$^+$21, WSC21] for the case when the potential f is strongly convex, [DMM19, DKRD19, CDWY20, Leh21] when it is convex, and [CCAY$^+$18, MCJ$^+$19, MMS20] when it is non-convex. We also highlight the works of [VW19], [EH20], [Ngu21] and [CEL$^+$21] which analyzed ULA when π satisfies certain functional inequalities. Specifically, [VW19] showed that when π satisfies a Logarithmic Sobolev Inequality (LSI) and has Lipschitz-smooth gradients, ULA with a number of iterations of order $\tilde{O}(1/\epsilon)$ generates a sample which is ϵ-close to π with respect to KL-divergence. A necessary condition for π to satisfy the LSI condition is that it should have sub-Gaussian tails. Furthermore, [EH20] considered densities that satisfy a modified LSI (m-LSI) inequality and showed that the number of iterations becomes of order $\tilde{O}(1/\epsilon^c)$, for some $c \geq 1$ (which depends on certain smoothness conditions). A typical example of a density that satisfies a m-LSI condition but not the LSI condition is $\pi(x) \propto \exp(-|x|)$. Thus, the result in [EH20] could also be viewed as an oracle complexity result for ULA when sampling from sub-exponential densities. Recently [Ngu21] relaxed the conditions required in [EH20] and provided similar results under the assumption that the target density satisfies a Poincaré inequality and dissipativity at the same time. Furthermore, [CEL$^+$21] also presented an analysis of ULA under the so-called Latała-Oleszkiewicz [LO00] inequality, that interpolates between the LSI and Poincaré inequality for the stronger Rényi metric and removes the dissipativity assumptions required in [EH20, Ngu21]. It is worth pointing out here that the proofs of [EH20] and [CEL$^+$21] are based on certain transformations of the target densities.

The above results, however, are not applicable to sampling from polynomially decaying heavy-tailed densities like the multivariate t-distribution, whose density is of the form $\pi(x) \propto (1+|x|^2)^{-\frac{d+\kappa}{2}}$, where $\kappa > 0$ is the degrees-of-freedom parameter. Recently, some attempts have been made to sample from such heavy-tailed densities by considering stable-driven SDEs of the form

$$dX_t = b(X_t)dt + \sqrt{2}dZ_t \tag{3.4}$$

where b is the drift term defined based on the Riesz potential, and Z_t is an ϑ-stable process with $\vartheta \in (1,2)$ [ŞZTG20, Şim17, HMW21]. Specifically [HMW21] established exponential ergodicity of the solution of (3.4), under conditions that allow for much heavier tails than Brownian-driven SDEs. The eventual hope is that discretizations of (3.4) might lead to algorithms with provable non-asymptotic oracle complexity rates. However, it appears to be non-trivial to analyze discretization of (3.4), especially if we are interested in tight non-asymptotic results, due to the difficulties in dealing with the non-smoothness of drift term b.

In this section, we take an alternate approach for heavy-tailed sampling using ULA on a transformed version of the target density. Such an approach was used by [JG12] in the context of Metropolis Random Walk algorithm, which serves as our motivation. The key idea in this approach is to construct smooth invertible maps (also called diffeomorphisms) $h : \mathbb{R}^d \to \mathbb{R}^d$ that transform the heavy-tailed density π to an appropriately light-tailed density π_h. Given such a map, one could first sample from the light-tailed density π_h and subsequently obtain samples from the heavy-tailed density π using the inverse map of h. It is also worth highlighting that [DBCD19, DGM20] and [BRZ19] used the transformation approach for proving asymptotic exponential ergodicity of bouncy particle and zig-zag samplers respectively, in the heavy-tailed setting.

There are several issues to overcome when using the above strategy in the context of ULA. First, note that the constructed map h has to convert the heavy-tailed density π to a light-tailed density π_h. In this process, however, the bulk of the density π_h might become non-smooth, if the map is not constructed carefully. This non-smoothness could subsequently hinder the usage of ULA algorithm to sample from π_h. Second, the constants involved (for example, the LSI or m-LSI constant) in the light-tailed density π_h might start to depend exponentially on the dimension after transformation. This again hinders the efficiency of the ULA when sampling from π_h. Furthermore, the transformation map needs to be efficiently computable. In this work, we propose a family of carefully constructed transformations that overcome the above issues and present non-asymptotic results for sampling from a class of heavy-tailed densities.

3.2.2. Organizations. The rest of the section is organized as follows. In Section 3.2.3 we introduce the notation and preliminary background material used in the rest of the section. In Section 3.2.4, we introduce our transformation map, highlight key properties and present the Transformed Unadjusted Langevin Algorithm (TULA) algorithm. We also discuss a warm-up example regarding exponentially tailed densities, and provide an interpretation of the transformed diffusion as a special case of Itô diffusions. In Section 3.2.5, we present non-asymptotic oracle complexity results for TULA under various assumptions on the potential function that characterize the level of heavy-tails allowed. In Section 3.2.6, we discuss the relationship between our assumptions on the heavy-tails used in Section 3.2.5 and non-local Dirichlet form based functional inequalities (arising in the equilibrium analysis of stable-driven diffusions). Illustrative examples are provided in Section 3.2.7.

3.2.3. Notations and Preliminaries. For a vector $a \in \mathbb{R}^d$, we represent the Euclidean norm by $|a|$. For a mapping $h : \mathbb{R}^d \to \mathbb{R}^d$, we denote the Jacobian matrix by $\nabla h \in \mathbb{R}^{d\times d}$. In the case when $h : \mathbb{R}^d \to \mathbb{R}$, $\nabla h \in \mathbb{R}^d$ denotes the gradient vector and $\triangle h = \nabla \cdot \nabla h$ denotes the Laplacian. For a function $h : \mathbb{R} \to \mathbb{R}$, we simply denote its first, second and third order derivatives by h', h'' and h''' respectively. For a matrix A, we denote its determinant and operator norm by $\det(A)$ and $\|A\|$ respectively. For two symmetric matrices A, B, the relation $A \preceq B$ refers to the fact that $B - A$ is positive semi-definite. The class of function $\mathcal{C}^k(\Omega)$ refers to those functions that have k-times continuously differentiable derivatives on the domain Ω. For a function ϕ, $\|\phi\|_\infty$ refers to the sup-norm.

We also require the following definitions used in the rest of the section. Let ν and μ be two probability densities with full support on $\mathbb{R}^d$. Then, for a convex function $\Phi : \mathbb{R} \to \mathbb{R}$ such that

$\Phi(1) = 0$, the Φ-*divergence* of ν from μ is defined as

$$D_\Phi(\nu|\mu) \coloneqq \int_{\mathbb{R}^d} \Phi\left(\frac{\nu(x)}{\mu(x)}\right) \mu(x)dx.$$

When the function is given by $\Phi(t) = t\log(t)$, we obtain the *Kullback-Leibler (KL) divergence* of ν with respect to μ, given by

$$H_\mu(\nu) \coloneqq \int_{\mathbb{R}^d} \log\frac{\nu(x)}{\mu(x)}\nu(x)dx.$$

Our complexity results later will be provided in terms of KL-divergence. The *Relative Fisher Information* of ν with respect to μ is given by

$$I_\mu(\nu) \coloneqq \int_{\mathbb{R}^d} \left|\nabla\log\frac{\nu(x)}{\mu(x)}\right|^2 \nu(x)dx.$$

The Rényi divergence of order $q > 1$ is defined as

$$R_q(\nu|\mu) = \frac{1}{q-1}\log\left(\int_{\mathbb{R}^d}\left(\frac{\nu(x)}{\mu(x)}\right)^q \mu(x)dx\right).$$

Note that when $q \to 1_+$, we have $R_q(\nu|\mu)$ approaching $H_\mu(\nu)$.

We now introduce additional technical details required for discussing functional inequalities; rigorous expositions could be found in [Wan06, BGL14]. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and let $\mathcal{L}$ denote a linear operator (infinitesimal generator) that is self-adjoint with domain $D(\mathcal{L})$ which generates a Markov semi-group P_t on $L^2(\mu)$. The carré de champ operator associated to the infinitesimal generator $\mathcal{L}$ is given by the bilinear map $\Gamma(\phi_1, \phi_2) = 1/2\left[\mathcal{L}(\phi_1\phi_2) - \phi_1\mathcal{L}\phi_2 - \phi_2\mathcal{L}\phi_1\right]$, for all ϕ_1, ϕ_2 defined in a subspace of $D(\mathcal{L})$ which is an algebra. We call the collection of the measure μ on a state space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and a carré de champ operator Γ a Markov triple, denoted as $(\mathbb{R}^d, \mu, \Gamma)$. It is well-known that the Dirichlet form associated with a Markov semi-group P_t is then given by $\mathcal{E}(\phi_1, \phi_2) = \int \Gamma(\phi_1, \phi_2)d\mu$. By a standard integration-by-parts argument, we also have that $\mathcal{E}(\phi_1, \phi_2) = -\int \phi_1\mathcal{L}\phi_2$. We use the convention $\mathcal{E}(\phi)$ to denote $\mathcal{E}(\phi, \phi)$. The Dirichlet domain $\mathcal{D}(\mathcal{E})$ is defined as $\mathcal{D}(\mathcal{E}) \coloneqq \left\{\phi \in L^2(\mu) : \mathcal{E}(\phi) < \infty\right\}$.

In the case of Brownian driven diffusions as in (3.1), the generator $\mathcal{L}$ is defined based on the Laplacian operator $\triangle$, which is a local operator. Correspondingly, the Dirichlet form is given by

$\mathcal{E}(\phi) = \int |\nabla \phi(x)|^2 \mu(x) dx$, for all $\phi \in \mathcal{D}(\mathcal{E})$. Based on this, we will introduce functional inequalities below. A probability density μ is said to satisfy *Poincaré Inequality (PI)* with constant C_{P}, denoted as $\mu \sim P(C_{\mathrm{P}})$ if for all functions $\phi \in \mathcal{D}(\mathcal{E})$, we have

$$\text{(PI)} \quad \mathrm{Var}_\mu(\phi) := \int_{\mathbb{R}^d} \left(\phi^2(x) - \left(\int_{\mathbb{R}^d} \phi(x)\mu(x)dx \right) \right)^2 \mu(x)dx \leq C_{\mathrm{P}} \int_{\mathbb{R}^d} |\nabla \phi(x)|^2 \mu(x) dx = C_{\mathrm{P}} \mathcal{E}(\phi).$$

Similarly, a probability density μ satisfies a *Logarithmic Sobolev inequality (LSI)* with constant C_{LSI} denoted as $\mu \sim LS(C_{\mathrm{LSI}})$ if for all functions $\phi \in \mathcal{D}(\mathcal{E})$, we have

(LSI)

$$\mathrm{Ent}_\mu(f) := \int_{\mathbb{R}^d} f^2(x) \log \left(\frac{f^2(x)}{\int_{\mathbb{R}^d} f^2(x)\mu(x)dx} \right) \mu(x)dx \leq 2C_{\mathrm{LSI}} \int_{\mathbb{R}^d} |\nabla f(x)|^2 \mu(x)dx = 2C_{\mathrm{LSI}} \mathcal{E}(\phi).$$

An equivalent form of LSI is that for all probability densities $\rho(x)$, we have

$$H_\mu(\rho) \leq \frac{C_{\mathrm{LSI}}}{2} I_\mu(\rho).$$

We refer the reader to [BGL14, Chapter 5] for the derivation of the equivalence. A probability density $\mu(x)$ satisfies a *modified Log-Sobolev Inequality (m-LSI)* if for all probability measure $\rho(x)$ and all $s \geq 2$, there is $\delta \in [0, 1/2)$ (depending on s) such that

$$\text{(m-LSI)} \qquad H_\mu(\rho) \leq C_{\text{M-LSI}} I_\mu(\rho)^{1-\delta} M_s(\rho + \mu)^\delta.$$

where $M_s(\rho) = \int_{\mathbb{R}^d} (1 + |x|^2)^{s/2} \rho(x) dx$. This version of m-LSI was introduced by [EH20] (also see [CEL+21]), motivated by a related definition from [TV00]. It is important to notice that the above version of m-LSI does not contain the Poincaré inequality as a special case, i.e., there exists densities that satisfy the above m-LSI inequality but not Poincaré inequality and vice versa. There exists other modifications to the LSI including the Beckner or Nash inequality [BGL14, Chapter 7] and the Latała-Oleszkiewicz [LO00] refinement to it, that interpolate between the LSI and Poincaré inequalities.

The above discussion is focused on Brownian driven SDEs. It turns out that the above class of functional inequalities are suitable for characterizing light-tailed densities (i.e., tails that decay exponentially fast). In the case of ϑ-stable driven diffusions as in (3.4), the generator is defined based on the non-local fractional Laplacian operator $(-\triangle)^{-\frac{\vartheta}{2}}$; see, for example, [Kwa17]. Correspondingly, in Section 3.2.6, we present more general functional inequalities based on non-local Dirichlet forms that are suitable for characterizing heavy-tailed densities and discuss the connection between our assumptions and such functional inequalities.

3.2.4. The Transformed Unadjusted Langevin Algorithm.

3.2.4.1. *Transformation Map.* We start this section by stating the following important property satisfied by smooth invertible transformation maps $h : \mathbb{R}^d \to \mathbb{R}^d$.

Definition 1 (Transformed density functions). For a probability density $\mu(x)$ with full support in $\mathbb{R}^d$, its transformed density function under a smooth invertible transformation map (or a diffeomorphism) h is given by $\mu_h(x) = \mu(h(x))\det(\nabla h(x))$ for all $x \in \mathbb{R}^d$.

If a random vector X has density μ, then we denote the density of the random vector $Y = h^{-1}(X)$, denoted as μ_h, as the transformed density of μ. Note that in particular if X admits density π of the form in (3.2), then $Y = h^{-1}(X)$ is distributed with density

$$\pi_h(y) = Z^{-1}e^{-f_h(y)} \quad \text{with} \quad f_h(y) = f(h(y)) - \log\det(\nabla h(y)), \tag{3.5}$$

being referred to as the transformed potential. In what follows, we assume that the potential function is isotropic. We emphasize that this assumptions is made for the sake of technical convenience – it is possible to relax this assumption to certain mild regularity conditions on the density, at the expense of having a more cluttered exposition.

Assumption A0. The initial potential function f is isotropic, i.e $f(x) = f(|x|)$ and $f : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable.

Since f is isotropic under Assumption A0, we may consider f to be a function defined on $\mathbb{R}_+$ as well. In the later context, we use $f(|x|)$ when we consider f defined on $\mathbb{R}_+$ and we use $f(x)$ when it is defined on $\mathbb{R}^d$. Similarly, when we use $f'(|x|), f''(|x|)$ and so on, to represent the derivatives, we consider f to be a function defined on $\mathbb{R}_+$.

We now describe the construction of our specific transformation map. Our proposal is motivated by the work of [JG12], who constructed similar maps to show exponential ergodicity of the Metropolis Random Walk (MRW) algorithm. It turns out that a direct application of their construction to analyze Langevin diffusions and their discretization, leads to worse dimensionality dependencies in the non-asymptotic oracle complexities. Indeed, this is expected as [JG12] predominantly focused on establishing asymptotic results. In order to proceed, we first define functions $g : \mathbb{R} \to \mathbb{R}$ which correspond to the first part of the transformation map construction. Specifically, g is defined based on initial function g_{in} as

$$g(r) = \begin{cases} g_{in}(r), & r < b^{-\frac{1}{\beta}}, \\ e^{br^{\beta}} & r \geq b^{-\frac{1}{\beta}}. \end{cases} \tag{3.6}$$

where $\beta \in (1, 2]$. The initial function $g_{in} : [0, b^{-\frac{1}{\beta}}) \to [0, e)$ satisfies the following assumption.

Assumption G1. The initial function $g_{in} : [0, b^{-\frac{1}{\beta}}) \to [0, e)$ is onto, monotone increasing and twice continuously differentiable. Furthermore, it satisfies,

$$\begin{aligned} g_{in}(0) &= 0 \\ \lim_{r \to b_{-}^{-\frac{1}{\beta}}} g_{in}(r) &= e, \\ \lim_{r \to b_{-}^{-\frac{1}{\beta}}} g'_{in}(r) &= \beta b^{\frac{1}{\beta}} e, \\ \lim_{r \to b_{-}^{-\frac{1}{\beta}}} g''_{in}(r) &= (2\beta^2 - \beta) b^{\frac{2}{\beta}} e, \\ \lim_{r \to b_{-}^{-\frac{1}{\beta}}} g'''_{in}(r) &= (5\beta^3 - 6\beta^2 + 2\beta) b^{\frac{3}{\beta}} e, \end{aligned}$$

$$\begin{aligned} \lim_{r \to 0_+} \left| \frac{f'(g_{in}(r)) g'_{in}(r)}{r} \right| &< \infty, \\ \lim_{r \to 0_+} \left| \frac{\frac{d}{dr} \log g'_{in}(r)}{r} \right| &< \infty, \\ \lim_{r \to 0_+} \left| \frac{\frac{d}{dr} \log \frac{g_{in}(r)}{r}}{r} \right| &< \infty, \\ \lim_{r \to 0_+} \left| \frac{d^2}{dr^2} \log \frac{g_{in}(r)}{r} \right| &< \infty, \\ \lim_{r \to 0_+} \left| \frac{d^2}{dr^2} \log g'_{in}(r) \right| &< \infty. \end{aligned}$$

We now show that if g_{in} satisfies Assumption G1, then g is three times continuously differentiable and invertible on $\mathbb{R}$.

Lemma 3.2.1. *For the function g defined in (3.6), if g_{in} satisfies Assumption G1, then we have*

(1) $g \in \mathcal{C}^3((0, \infty))$,

(2) *g is onto, strictly monotonically increasing, and hence invertible.*

The proof of Lemma 3.2.1 is provided in Section 3.2.8.1. We now show that under Assumption G1, the Φ-divergence is preserved after transformation. This property is important to eventually provide our convergence results for sampling.

Proposition 6. Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be a transformation map satisfying Assumption G1. For any two probability densities ν and μ with full support on $\mathbb{R}^d$, let ν_h and μ_h be the two transformed densities under the map h. Then the Φ-divergence is preserved after transformation, i.e., we have

$$D_\Phi(\nu|\mu) = D_\Phi(\nu_h|\mu_h). \tag{3.7}$$

Proof: We start from the right side of (3.7):

$$\begin{aligned} D_\Phi(\nu_h|\mu_h) &= \int_{\mathbb{R}^d} \Phi\left(\frac{\nu_h(y)}{\mu_h(y)}\right)\mu_h(y)dy = \int_{\mathbb{R}^d} \Phi\left(\frac{\nu(h(y))\det(\nabla h(y))}{\mu(h(y))\det(\nabla h(y))}\right)\mu(h(y))\det(\nabla h(y))dy \\ &= \int_{\mathbb{R}^d} \Phi\left(\frac{\nu(x)}{\mu(x)}\right)\mu(x)dx = D_\Phi(\nu|\mu). \end{aligned}$$

The second identity follows by the change of variable $x = h(y)$ and noting $\det(\nabla h(y)) > 0$ under Assumption G1. ■

With the properties of g introduced in Lemma 3.2.1, we can then further define the isotropic transformations $h : \mathbb{R}^d \to \mathbb{R}^d$:

$$h(x) = \begin{cases} \dfrac{g(|x|)x}{|x|} & x \neq 0, \\ 0 & x = 0. \end{cases} \tag{3.8}$$

We call the map $y \mapsto x = h(y)$ to be the transformation map, which is isotropic. Furthermore, h is also three times continuously differentiable and invertible on $\mathbb{R}^d$ and its inverse is

$$h^{-1}(x) = \begin{cases} g^{-1}(|x|)\dfrac{x}{|x|} & x \neq 0, \\ 0 & x = 0. \end{cases}$$

Therefore, we can define the inverse transformation map $x \mapsto y = h^{-1}(x)$.

Algorithm 1: Transformed Unadjusted Langevin Algorithm (TULA)

```
Input  : Step size $\gamma$ and a sample $y_0$ from a starting density $\rho_0$
Output: Sequence $x_1, x_2, \cdots$
for $n = 0, 1, \cdots$ do
    $x_n = h(y_n)$                                    ▷ apply the inverse transformation ;
    $y_{n+1} \sim \mathcal{N}(y_n - \gamma\nabla f_h(y_n), 2\gamma I_d)$          ▷ generate samples ;
end
```

3.2.4.2. *Transformed Langevin Diffusion and its discretization.* With the transformed density defined in (3.5), the transformed overdamped Langevin diffusion is given by

$$dY_t = -\nabla f_h(Y_t)dt + \sqrt{2}dW_t. \tag{3.9}$$

We denote the density of Y_t by ρ_t for all $t \geq 0$. The stationary density function for the diffusion given by (3.9) is π_h as defined in (3.5). We can apply Euler discretization to the transformed overdamped Langevin diffusion in (3.9) and generate a Markov chain $(y_n)_{n\geq 1}$ via the recursion,

$$y_{n+1} = y_n - \gamma_{n+1}\nabla f_h(y_n) + \sqrt{2\gamma_{n+1}}u_{n+1} \tag{3.10}$$

where (u_n) is a sequence of independent and identically distributed d-dimensional standard Gaussian vectors and $\gamma > 0$ is the fixed step size. The Transformed Unadjusted Langevin algorithm (TULA) in order to generate samples from a heavy-tailed density π is given in Algorithm 1.

We use ν_n to denote the density of the nth iterate x_n and π_γ to denote the stationary density of $(x_n)_{n\geq 1}$. Since the step-size γ in Algorithm 1 is a constant, there is a bias between π_γ and π. For arbitrary accuracy $\epsilon > 0$, by choosing small enough step-size γ and large enough number of iterations n, we can bound the distance between ν_n and π by ϵ in terms of KL or Rényi divergence.

A Warm-up Example: Although our main motivation is to sample from densities that have polynomially decaying tails, in this subsection, we provide a warm-up example on sampling from a density that has exponentially decaying tails (see (3.12) for the definition of the potential function) and does not satisfy LSI, by transforming it to satisfy LSI. Towards that goal, we consider the transformation map in (3.8) with the function g defined as

$$g(r) = \begin{cases} dr^2, & r \geq R, \\ g_{in}(r), & 0 \leq r \leq R, \end{cases} \tag{3.11}$$

where $R > 0$ is a constant, with

$$g_{in}(r) = dRr \exp\left(-\frac{5}{6} + \frac{3}{2}\frac{r^2}{R^2} - \frac{2}{3}\frac{r^3}{R^3} \right).$$

The above form for g is motivated by [JG12, Equation 15], where they constructed transformation maps to transform densities that are sub-exponential to sub-Gaussian. We also point out that we consider the form of g in (3.11) only for this section, and it should not be confused with the general form (3.6) considered in the rest of the section. By an argument similar to the proof of Lemma 3.2.1, it could be shown that the transformation map defined with g as in (3.11) is a diffeomorphism.

Now, consider the potential function defined in a piece-wise manner as

(3.12)
$$f(x) = \begin{cases} (1+|x|^2)^{\frac{1}{2}} + \frac{1}{2} d \log |x|, & |x| \geq R, \\ (1+d^2 g_{in}^{-1}(|x|)^4)^{\frac{1}{2}} + (d-1)\ln \frac{|x|}{g_{in}^{-1}(|x|)} + \log g_{in}'(g_{in}^{-1}(|x|)) - \frac{d}{2}\log d - \ln 2, & |x| \in [0, R]. \end{cases}$$

The corresponding probability density induced by the potential f above has a lighter tail than the one with potential $|x|$. But it has a heavier tail than densities with potentials $|x|^\varrho$ for any $\varrho > 1$. For the above potential f, the transformed potential is given by

$$f_h(x) = (1 + d^2|x|^4)^{\frac{1}{2}} - \frac{d}{2}\log d - \log 2.$$

The LSI constant of the density induced by f_h can be studied via the Holley-Stroock Theorem (see Theorem 15). We can write

$$\begin{aligned} f_h(x) &= d|x|^2 + \frac{1}{d|x|^2 + (1+d^2|x|^4)^{\frac{1}{2}}} - \frac{d}{2}\log d - \log 2 \\ &:= d|x|^2 - \frac{d}{2}\log d - \log 2 + \mathrm{Osc}(x), \end{aligned}$$

where $\mathrm{Osc}(x) := \frac{1}{d|x|^2+(1+d^2|x|^4)^{\frac{1}{2}}}$ and is uniformly bounded by 1. Meanwhile the density corresponding to the potential function $e^{-d|x|^2} + \frac{d}{2}\log d + \log 2$ satisfies LSI with constant $1/2d$. Therefore e^{-f_h} satisfies LSI with constant $C_{h,\mathrm{LSI}} = e/2d$. On the other hand, $f_h(x)$ also has Lipschitz gradients with constant $L_h = O(d)$. Hence, according to [VW19] and Proposition 6, the iteration complexity of TULA for sampling from a density with potential f as in (3.12) is of order $\tilde{O}(d/\epsilon)$ where ϵ is the error tolerance in KL-divergence. This is to be contrasted with [CEL$^+$21, Examples 9 and 11] on using ULA to sample from densities with potentials of the from $|x|^{\varrho}$ for $\rho \in [1, 2]$. Specifically, we note that TULA has better oracle complexity as long as $\rho \in (1, 2]$.

3.2.4.3. *Transformed Langevin Diffusions as Itô diffusions.* It is worth noting that the transformed diffusion process in (3.9) could also be interpreted in terms of an Itô diffusion. Specifically, by a direct calculation, the stochastic process $X_t = h(Y_t)$ has the form

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \tag{3.13}$$

with $\sigma(x) := \sqrt{2}(\nabla h)(h^{-1}(x))$ and

$$\begin{aligned} b(x) := &-(\nabla h^T)(h^{-1}(x))(\nabla h)(h^{-1}(x))\nabla f(x) + (\nabla h^T)(h^{-1}(x))(\nabla \log\det \nabla h)(h^{-1}(x)) \\ &+ (\Delta \cdot h)(h^{-1}(x)), \end{aligned}$$

where $\triangle \cdot h(\cdot) \in \mathbb{R}^d$ and is defined co-ordinate wise as $(\triangle \cdot h(x))_i = \triangle h_i(x)$ for all $i \in \{1, \cdots, d\}$ and $x \in \mathbb{R}^d$. Furthermore, we can actually show that

$$b(x) = \frac{1}{2\pi(x)} \langle \nabla, \pi(x)\sigma(x)^T\sigma(x)\rangle,$$

where $\langle \nabla, \cdot\rangle$ is the divergence operator for matrix-valued function, i.e $\langle \nabla, \omega(x)\rangle_i = \sum_{j=1}^d \frac{\partial \omega_{i,j}(x)}{\partial x_j}$ for $\omega : \mathbb{R}^d \to \mathbb{R}^{d\times d}$.

The above form of $b(x)$ follows by noting that from the form of $\pi(x)$ in (3.2), we have

$$\begin{aligned} \frac{1}{2\pi(x)}\langle \nabla, \pi(x)\sigma(x)^T\sigma(x)\rangle &= \frac{1}{2}\langle \nabla, \sigma^T(x)\sigma(x)\rangle - \frac{1}{2}\sigma^T(x)\sigma(x)\nabla f(x) \\ &= -(\nabla h^T)(h^{-1}(x))(\nabla h)(h^{-1}(x))\nabla f(x) + \frac{1}{2}\langle \nabla, \sigma^T(x)\sigma(x)\rangle. \end{aligned}$$

Meanwhile from (3.8), based on elementary algebraic manipulations, we obtain that

$$\frac{1}{2}\langle \nabla, \sigma^T(x)\sigma(x)\rangle = \left[2g''(g^{-1}(|x|)) + (d-1)\frac{g'(g^{-1}(|x|))^2}{|x|} - (d-1)\frac{|x|}{g^{-1}(|x|)^2}\right]\frac{x}{|x|},$$

$$(\Delta \cdot h)(h^{-1}(x)) = \left[g''(g^{-1}(|x|)) + (d-1)\frac{g'(g^{-1}(|x|))}{g^{-1}(|x|)} - (d-1)\frac{|x|}{g^{-1}(|x|)^2}\right]\frac{x}{|x|},$$

and

$$\begin{aligned}&(\nabla h^T)(h^{-1}(x))(\nabla \log \det \nabla h)(h^{-1}(x))\\ &= \left[g''(g^{-1}(|x|)) + (d-1)\frac{g'(g^{-1}(|x|))^2}{|x|} - (d-1)\frac{g'(g^{-1}(|x|))}{g^{-1}(|x|)}\right]\frac{x}{|x|}.\end{aligned}$$

This highlights the fact that transformations provide a way of constructing the drift and diffusion terms in the Itô diffusion that take into account the heavy-tailed nature of the target density. However, it turns out that the results on the analysis of discretizations of Itô diffusion from [EMS18, LWME19], which are in the Wasserstein metric, are not applicable to the class of Itô diffusion of the form above; indeed the stronger Wasserstein contraction conditions made in those works are not satisfied by the above class of Itô diffusions. We leave a detailed investigation of analysis of discretizations of Itô diffusion above, in stronger KL or Rényi metrics, as future work.

3.2.5. Convergence Results. In this section, we will impose assumptions on the potential function f under which we show exponential ergodicity of the transformed Langevin diffusion and convergence results for Algorithm 1.

3.2.5.1. *Convergence along the transformed Langevin diffusions.* We first state convergence results for the continuous time Langevin diffusion under various curvature-related assumptions on the potential function.

ASSUMPTION A1. (Dissipativity) There exists $A, B, N_1 > 0, \alpha \in [1,2]$ such that for all $|x| > N_1$:

$$f'(\psi(|x|))\psi'(|x|)|x| - b\beta d|x|^\beta + (d-\beta) > A|x|^\alpha - B,$$

where $\psi(r) = e^{br^\beta}$ for all $r \geq b^{-\frac{1}{\beta}}$.

Assumption A1 is imposed to guarantee that the transformed potential function satisfies the dissipativity condition. We next recall the dissipativity condition for completeness.

Assumption B1. (α_h-dissipativity) We say that the transformed potential function $f_h : \mathbb{R}^d \to \mathbb{R}$ satisfies the α_h-dissipativity condition with $\alpha_h \in [1,2]$ if there exists $A_h, B_h > 0$ such that for all $x \in \mathbb{R}^d$:

$$\langle \nabla f_h(x), x \rangle > A_h |x|^{\alpha_h} - B_h.$$

If the transformed potential function satisfies the α_h-dissipativity condition with $\alpha_h = 1$, then the corresponding transformed density π_h satisfies a Poincaré inequality with certain constant $C_{h,\mathrm{P}}$ depending on the potential function. Then, similar to [VW19], we obtain the following result.

Theorem 10. Assume the initial potential function f satisfies Assumption A0 and Assumption A1 with $\alpha = 1$. Then, the transformed density π_h with $\beta = 1$ and $b \geq \frac{r}{8(d-1)}$ satisfies a Poincaré inequality with a constant $C_{h,\mathrm{P}}$ depending on f. Therefore along the transformed Langevin diffusion (3.9), we have for $q \geq 2$ that

$$R_q(\rho_t|\pi_h) \leq \begin{cases} R_q(\rho_0|\pi_h) - \dfrac{2C_{h,\mathrm{P}}t}{q} & \text{if } R_q(\rho_0|\pi_h) \geq 1 \text{ as long as } R_q(\rho_t|\pi_h) \geq 1, \\ e^{-\frac{2C_{h,\mathrm{P}}t}{q}} R_q(\rho_0|\pi_h) & \text{if } R_q(\rho_0|\pi_h) \leq 1. \end{cases}$$

Assumption A2. (Degenerate convexity) There exists $\mu, N_2 > 0, \theta \geq 0$ such that for all $|x| > N_2$:

$$f'(\psi(|x|))\psi'(|x|)|x|^{-1} - b\beta d|x|^{\beta-2} + (d-\beta)|x|^{-2} > \frac{\mu}{(1+\frac{1}{4}|x|^2)^{\theta/2}},$$
$$f''(\psi(|x|))\psi'(|x|)^2 + f'(\psi(|x|))\psi''(|x|) - b\beta(\beta-1)d|x|^{\beta-2} - (d-\beta)|x|^{-2} > \frac{\mu}{(1+\frac{1}{4}|x|^2)^{\theta/2}}.$$

where $\psi(r) = e^{br^\beta}$ for all $r \geq b^{-\frac{1}{\beta}}$.

Assumption A2 is imposed to guarantee the transformed potential function is degenerately convex at infinity. We now recall the definition of degenerate convexity at infinity from [EH20].

Assumption B2. (Degenerate convexity at infinity) We say that the transformed potential function $f_h : \mathbb{R}^d \to \mathbb{R}$ is degenerately convex at infinity if there exist a function $\tilde{\phi} : \mathbb{R}^d \to \mathbb{R}$ such that for a constant $\xi_h \geq 0$

$$\left\| f_h - \tilde{\phi} \right\|_\infty \leq \xi_h,$$

where $\tilde{f}$ satisfies,

$$\nabla^2 \tilde{f}(x) \succeq \frac{\mu_h}{(1 + \frac{1}{4}|x|^2)^{\theta_h/2}} I_d,$$

for some $\mu_h > 0$ and $\theta_h \geq 0$.

The degenerate convexity at infinity condition is weaker than the strong convexity at infinity. If a potential function satisfies degenerate convexity at infinity, then the corresponding probability measure satisfies m-LSI. Similar to [TV00], we obtain the following result.

Theorem 11. Assume the initial potential function f satisfies Assumption A0 and Assumption A2. Then the transformed density π_h satisfies a modified Logarithmic Sobolev Inequality with a uniform constant δ (see (m-LSI)) and constant $C_{h,\text{M-LSI}}$ depending on f. Therefore along the transformed Langevin diffusion (3.9), we have

$$H_{\pi_h}(\rho_t) \leq \frac{C}{t^\ell},$$

where the constant C depends on the potential f and the transformation h and $\ell = (1 - 2\delta)/\delta$.

Remark 11. Note that the above rate is faster than any polynomial but not truly exponential. While the above rate could be made exponential with additional assumptions on the tail and/or assumptions on the initial distribution, we do not present such modifications here.

Assumption A3. (Strong convexity at infinity) There exists $N_3, \rho > 0$ such that for all $|x| > N_3$:

$$f'(\psi(|x|))\psi'(|x|)|x|^{-1} - b\beta d|x|^{\beta-2} + (d - \beta)|x|^{-2} > \rho,$$

$$f''(\psi(|x|))\psi'(|x|)^2 + f'(\psi(|x|))\psi''(|x|) - b\beta(\beta - 1)d|x|^{\beta-2} - (d - \beta)|x|^{-2} > \rho,$$

where $\psi(r) = e^{br^\beta}$ for all $r \geq b^{-\frac{1}{\beta}}$.

Assumption A3 is imposed to guarantee that the transformed potential function is strongly convex with parameter ρ_h at infinity. The property that a potential function is strongly convex at infinity implies that the corresponding probability measure satisfies a LSI with a certain parameter depending on the potential function and the transformation map.

Theorem 12. Assume the initial potential function f satisfies Assumption A0 and Assumption A3, then the transformed density π_h satisfies a logarithmic Sobolev inequality with a constant $C_{h,\mathrm{LSI}}$ depending on f. Therefore along the transformed Langevin diffusion (3.9), we have

$$H_{\pi_h}(\rho_t) \leq e^{-2tC_{h,\mathrm{LSI}}} H_{\pi_h}(\rho_0).$$

3.2.5.2. *Convergence along TULA.* In this section, we state two types of convergence results for Algorithm 1, based on Proposition 6 and [VW19, CEL+21]. While the works of [VW19, CEL+21] provide results only for exponentially decaying densities, our results below are applicable for polynomially-decaying densities based on the constructed transformation maps. To proceed, we first list smoothness conditions on the potential function f.

Assumption A4. (Gradient Lipschitz) There exists $N_4, L > 0$ such that for all $|x| > N_4$:

$$\begin{aligned} &f'(\psi(|x|))\psi'(|x|)|x|^{-1} - b\beta d|x|^{\beta-2} + (d-\beta)|x|^{-2} < L, \\ &f''(\psi(|x|))\psi'(|x|)^2 + f'(\psi(|x|))\psi''(|x|) - b\beta(\beta-1)d|x|^{\beta-2} - (d-\beta)|x|^{-2} < L \end{aligned}$$

where $\psi(r) = e^{br^\beta}$ for all $r \geq b^{-\frac{1}{\beta}}$.

Assumption A4 is imposed to guarantee that the transformed potential function has Lipschitz gradients with parameter L_h. Such smoothness conditions on the potential function are required to study the discrete Markov chain generated in the unadjusted Langevin algorithm. We also remark that it is possible to relax the Lipschitz gradient assumption to certain weak-smooth conditions on the gradient; we do not pursue such extensions in this work. While Theorem 11 holds under m-LSI, to get the corresponding result for Algorithm 1, we also require the following additional tail-conditions.

Assumption A5. (Tail assumption) For some $m \geq 0$, $\alpha_1 \in [0,1]$ and $N_5 > 0$, there exists a positive constant C^*_{TAIL} such that for all $\lambda \geq N_5$,

$$\pi\left\{|\cdot| \geq m + \lambda\right\} \leq 2\exp\left(-\left(\frac{\psi^{-1}(\lambda)}{C^*_{\text{TAIL}}}\right)^{\alpha_1}\right),$$

where $\psi(r) = e^{br^\beta}$ for all $r \geq b^{-\frac{1}{\beta}}$.

Assumption B5. For some $m_h \geq 0$ and $\alpha_{h,1} \in [0,1]$, there exists a positive constant $C_{h,\text{TAIL}}$ such that for all $\lambda \geq 0$,

$$\pi_h\left\{|\cdot| \geq m_h + \lambda\right\} \leq 2\exp\left(-\left(\frac{\lambda}{C_{h,\text{TAIL}}}\right)^{\alpha_{h,1}}\right).$$

Theorem 13. In addition to the assumptions in Theorem 11, assume that the initial potential f is such that $\nabla f_h(0) = 0$, and it satisfies Assumption A4 and Assumption A5. Furthermore, let $\epsilon^{-1}, m_h, C_{h,\text{M-LSI}}, C_{h,\text{TAIL}}, L_h, R_2(\rho_0|\hat{\pi}_h) \geq 1$ ($\hat{\pi}_h$ is as defined in (3.52) with $\hat{R} = 2\int_{\mathbb{R}^d}|x|\pi_h(x)dx$ and $\hat{\gamma} = (3072n\gamma)^{-1}$), and $m_h, C_{h,\text{TAIL}}, R_2(\rho_0|\pi) \leq d^{\tilde{O}(1)}$. Then, Algorithm 1 with an step size

$$\gamma = \tilde{\Theta}\left(\frac{\epsilon}{dC^2_{h,\text{M-LSI}}C^\theta_{h,\text{TAIL}}L_h^2R_{2q}(\nu_0|\pi)^{\theta/\alpha_{h,1}}} \times \min\left\{1, \frac{1}{q\epsilon}, \frac{d}{m_h}, \frac{d}{R_2(\rho_0|\hat{\pi}_h)}, \left(\frac{R_{2q}(\nu_0|\pi)^{1/\alpha_{h,1}}}{m_h}\right)^\theta\right\}\right),$$

satisfies $R_q(\nu_n|\pi) \leq \epsilon$, for all $q \geq 2$ after

$$n = \tilde{\Theta}\left(\frac{dR_{2q}(\nu_0|\pi)^{2\theta/\alpha_{h,1}}C^4_{h,\text{M-LSI}}C^{2\theta}_{h,\text{TAIL}}L_h^2}{\epsilon}\max\left\{1, \epsilon, \frac{m_h}{d}, \frac{R_2(\rho_0|\hat{\pi}_h)^{1/2}}{d}, \left(\frac{m_h}{R_{2q}(\nu_0|\pi)^{1/\alpha_{h,1}}}\right)^\theta\right\}\right)$$

iterations, for some $\theta \in [0,1]$ (depending on the parameter δ in (m-LSI)). Explicit form of $C_{h,\text{M-LSI}}$ is the constant λ in (3.42) and $m_h, C_{h,\text{TAIL}}, L_h$ are given in (3.51),(3.50), (3.36) respectively. The $\tilde{\Theta}(\cdot)$ notation hides polylogarithmic factors as well as constants depending on θ, q.

Remark 12. In order to obtain a direct quantitative bound, it is important to obtain a control of $R_2(\rho_0|\hat{\pi}_h)$ and $R_{2q}(\nu_0|\pi)$. We refer to [CEL$^+$21, Section A] for a proof that the conditions required on $R_2(\rho_0|\hat{\pi}_h)$ is satisfied, and for obtaining a control on the term $R_{2q}(\nu_0|\pi)$.

Theorem 14. In addition to the assumptions in Theorem 12, assume that f satisfies Assumption A4. Then Algorithm 1, for any $y_0 \sim \rho_0$ with $H_{\pi_h}(\rho_0) < \infty$, and with step size

$$0 < \gamma \leq \frac{1}{2L_h^2 C_{h,\mathrm{LSI}}} \min\left\{1, \frac{\epsilon}{4d}\right\},$$

satisfies $H_\pi(\nu_n) < \epsilon$, for any $\epsilon > 0$ after

$$n = \tilde{\Theta}\left(\frac{C_{h,\mathrm{LSI}}}{2\gamma} \log \frac{2H_{\pi_h}(\rho_0)}{\epsilon}\right)$$

iterations. Explicit forms of $C_{h,\mathrm{LSI}}$ and L_h are given in (3.35) and (3.36).

Remark 13. As argued in [VW19], if we let ρ_0 to be a Gaussian distribution with mean being any stationary point of f_h and covariance matrix being $(1/L_h)I_d$, then $H_{\pi_h}(\rho_0) = \tilde{O}(d)$. Furthermore, we also remark that similar convergence results in the stronger Rényi metric, for all $q \geq 4$ holds via [CEL$^+$21, Theorem 4].

Remark 14. We leave a detailed study of obtaining convergence results for the underdamped Langevin dynamics and its discretization as future work.

3.2.6. Relation with Poincaré Inequalities based on Non-local Dirichlet Forms. We now discuss the relationship between our assumptions on the potential function and functional inequalities like super and weak Poincaré inequalities that arise in characterizing the heavy-tailed stationary distributions of certain ϑ-stable driven diffusions [RW01, RW03, CGGR10, Wan14, WW15, HMW21]. Recall from Section 3.2.3 that the Dirichlet form associated with Langevin diffusion in (3.1) is of the form $\mathcal{E}(\phi) = \int |\nabla\phi(x)|^2\mu(x)dx$. However, in the case of ϑ-stable driven diffusions the corresponding non-local Dirichlet form is given by

$$\mathcal{E}(\phi) := \iint_{\{x\neq y\}} \frac{(\phi(x)-\phi(y))^2}{|x-y|^{(d+\vartheta)}} dx\mu(dy), \tag{3.14}$$

for all functions in the Dirichlet domain $\mathcal{D}(\mathcal{E})$; see for example [Wan14]. We now introduce similar functional inequalities that are associated with stable-driven diffusions.

Definition 2 (Poincaré-type Inequalities). A Markov triple $(\mathbb{R}^d, \mu, \Gamma)$ (with μ a probability measure), with the Dirichlet form as in (3.14) is said to satisfy

- a *Poincaré inequality* if there exists a consatnt $C > 0$ such that for any function $\phi : \mathbb{R}^d \to \mathbb{R}$ in the Dirichlet domain $\mathcal{D}(\mathcal{E})$,
$$\mathrm{Var}_\mu(\phi) \leq C\mathcal{E}(\phi),$$
- a *weak Poincaré inequality* if there exists a function $\alpha : (0, \infty) \to \mathbb{R}_+$ such that for any function $\phi : \mathbb{R}^d \to \mathbb{R}$ in the Dirichlet domain $\mathcal{D}(\mathcal{E})$ and $r > 0$,
$$\mathrm{Var}_\mu(\phi) \leq \alpha(r)\mathcal{E}(\phi) + r\left\|\phi\right\|_\infty^2,$$
- a *super Poincaré inequality* if there exists non-increasing function $\beta : (0, \infty) \to \mathbb{R}_+$ such that for any function $\phi : \mathbb{R}^d \to \mathbb{R}$ in the Dirichlet domain $\mathcal{D}(\mathcal{E})$ and $r > 0$,
$$\mu(\phi^2) \leq r\mathcal{E}(\phi) + \beta(r)\mu(|\phi|)^2,$$
where $\mu(\varphi) = \int_{\mathbb{R}^d} \varphi(x)\mu(dx)$ for all $\varphi \in L^1(\mu)$.

In the following, we will discuss the relation between Assumption A1, Assumption A3, and Assumption A2, and the Poincaré-type inequalities above. In what follows, the terms $\alpha(r)$ and $\beta(r)$ are defined as

$$\alpha(r) = \inf_{s>0}\left\{\frac{1}{\inf_{0<|x-y|\leq s}[(e^{f(x)} + e^{f(y)})|x-y|^{-(d+\vartheta)}]} : \int\int_{|x-y|>s} e^{-f(x)}e^{-f(y)}dxdy \leq r/2\right\}, \tag{3.15}$$

$$\beta(r) = \inf_{t,s>0}\left\{\frac{2\mu(\omega)}{\inf_{|x|\geq t}\omega(x)} + \beta_t(t\wedge s) : \frac{2}{\inf_{|x|\geq t}\omega(x)} + s \leq r\right\}, \tag{3.16}$$

where for any $t > 0$, we have

$$\beta_t(s) = \inf_{u>0} \left\{ \frac{(\sup_{|z|\le 2t} e^{f(z)})^2}{u^d(\inf_{|z|\le t} e^{f(z)})} : \frac{u^{\vartheta}(\sup_{|z|\le 2t} e^{f(z)})}{(\inf_{|z|\le t} e^{f(z)})} \le s \right\}.$$

The function ω will depend on the properties of the potential f.

Proposition 7. If the original potential satisfies Assumption A1 with parameters α, A, B, then

(1) If $\alpha > \beta$ or $\alpha = \beta, \vartheta < A\beta^{-1}b^{-1}$, the original density function satisfies the super Poincaré inequality with

$$\omega(x) = \frac{C}{2^{d+\vartheta}} |x|^{A\alpha^{-1}b^{-\frac{\alpha}{\beta}} \log^{\frac{\alpha}{\beta}-1}(|x|) - \vartheta} \log^{-\frac{B}{\beta}}(|x|),$$

for some positive constant C.

(2) If $\alpha = \beta, \vartheta \ge A\beta^{-1}b^{-1}$, the original density function satisfies the weak Poincaré inequality.

Proposition 8. If the original potential satisfies Assumption A2 with parameters μ, θ, then

(1) If $\theta < 2 - \beta$ or $\theta = 2 - \beta, \vartheta < \mu\beta^{-1}b^{-1}$, the original density function satisfies the super Poincaré inequality with $\omega(x)$ defined as

$$\omega(x) = \begin{cases} \frac{C}{2^{d+\vartheta}} |x|^{(1-\theta)^{-1}(2-\theta)^{-1}\mu b^{-\frac{2-\theta}{\beta}} \log^{\frac{2-\theta}{\beta}-1}(|x|) + 1 - (d+\vartheta)} \log^{-\frac{d-\beta}{\beta}}(|x|), & \theta < 2-\beta, \\ \frac{C}{2^{d+\vartheta}} |x|^{b^{-\frac{2-\theta}{\beta}} \log^{\frac{2-\theta}{\beta}-1}(|x|) - \vartheta} \log^{-\frac{d-\beta}{\beta}}(|x|), & \theta = 2-\beta, \vartheta < \mu\beta^{-1}b^{-1}. \end{cases}$$

where C is some positive constant.

(2) If $\theta = 2 - \beta, \vartheta \ge \mu\beta^{-1}b^{-1}$ or $\theta > 2 - \beta$, the original density function satisfies the weak Poincaré inequality.

$$\alpha(r) = \inf \left\{ \frac{1}{\inf_{0<|x-y|\le s}[(e^{f(x)} + e^{f(y)})|x-y|^{-(d+\vartheta)}]} : \int\int_{|x-y|>s} e^{-f(x)} e^{-f(y)} dx dy \le r/2 \right\}.$$

Proposition 9. If the original potential function satisfies Assumption A3 with parameter ρ, then

(1) If $\beta \in (1,2)$ or $\beta = 2, \vartheta < \frac{1}{2}\rho b^{-1}$, the original density function satisfies the super Poincaré inequality with

$$\omega(x) = \frac{C}{2^{d+\vartheta}} |x|^{\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|x|) - \vartheta} \log^{-\frac{d-\beta}{\beta}}(|x|),$$

for some positive constant C.

(2) If $\beta = 2, \vartheta = \frac{1}{2}\rho b^{-1}, d = 1, 2$, the original density function satisfies the Poincaré inequality.

(3) If $\beta = 2, \vartheta = \frac{1}{2}\rho b^{-1}, d \geq 3$ or $\beta = 2, \vartheta > \frac{1}{2}\rho b^{-1}$, the original density function satisfies the weak Poincaré inequality.

Remark 15. For the example of multivariate t-distribution, it is shown later in Lemma 3.2.3, that it satisfies Assumption A3 with $\beta = 2$, $\alpha = 2$, $A = 2b\kappa$, $B \geq 0$ and arbitrary $\mu \in (0, 2b\kappa)$. Therefore when $\kappa > \vartheta$, it falls into the class of densities described by the super Poincaré inequality. When $0 < \kappa \leq \vartheta$, it falls into the class of densities described by the weak Poincaré inequality. This classification of the multivariate t-distributions with different degrees of freedom coincides with [WW15, Corollary 1.2].

Remark 16. For the multivariate t-distribution with degrees of freedom κ, we show later in Lemma 3.2.3 that it satisfies Assumption A2 with arbitrary $\mu \in (0, b\kappa\beta(\beta-1))$ and $\theta = 2 - \beta$. When $\vartheta < \kappa(\beta - 1)$, we can show that multivariate t-distribution with κ degrees of freedom satisfies the super Poincaré inequality which agrees with the results in [Wan14] and our Remark 15 above.

3.2.7. Illustrative Examples. In this section, we introduce a specific transformation map h defined by (3.6) and (3.8) with $\beta = 2$ and g_{in} defined by the following equation. For all $r \leq b^{-\frac{1}{2}}$,

$$g_{in}(r) = rb^{\frac{1}{2}} \exp\left(br^2 - \frac{10}{3}b^{\frac{3}{2}}r^3 + \frac{15}{4}b^2r^4 - \frac{6}{5}b^{\frac{5}{2}}r^5 + \frac{47}{60}\right). \tag{3.17}$$

Using the above transformation map, we analyze the oracle complexity of TULA for sampling from the multivariate t-distribution and related densities.

3.2.7.1. *Example 1.* The density and potential function of the multivariate t-distribution are respectively given by

$$\pi(x) \propto (1+|x|^2)^{-\frac{d+\kappa}{2}}, \quad f(x) = \frac{d+\kappa}{2}\log(1+|x|^2), \tag{3.18}$$

where κ is the degrees of freedom parameter. We first show that the above g_{in} satisfies Assumption G1 and hence the corresponding h is a diffeomorphism.

Lemma 3.2.2. *With g_{in} defined in (3.17), $\beta = 2$ and $f(x) = \frac{d+\kappa}{2}\log(1+|x|^2)$, g defined in (3.6) satisfies Assumption G1.*

Next, we show that the potential function of the multivariate t-distribution satisfies the assumptions we introduced in Section 3.2.5.

Lemma 3.2.3. *We have for the following for the potential function $f(x)$ in (3.18):*

1. *$f(x)$ is isotropic and $f \in \mathcal{C}^2(\mathbb{R}^d)$;*
2. *f satisfies Assumption A4 with some $N_4 > 0$ and $L = 2\kappa b^{\frac{2}{\beta}}\beta$;*
3. *f satisfies Assumption A1 with $\alpha = \beta$, $A = \kappa b\beta$ and some $B \geq 0$,$N_1 > 0$.*
4. *f satisfies Assumption A2 with arbitrary $\mu \in (0, \kappa b\beta(\beta-1))$, $\theta = 2-\beta$ and some $N_2 > 0$.*

Hence, we can apply Theorem 14 with $n = \tilde{O}(L_h^2 C_{h,\mathrm{LSI}}^2 d/\epsilon)$, where $C_{h,\mathrm{LSI}}$ and L_h are two constants that depend on f as introduced in (3.35) and (3.36). However, the dependence of $C_{h,\mathrm{LSI}}$ and L_h on f would affect the order of n significantly, especially in terms of the dimension parameter. Specifically, after explicitly calculating the constants $C_{h,\mathrm{LSI}}$ and L_h, the mixing time of TULA in KL-divergence with error tolerance ϵ is of order $n = \tilde{O}(\exp(2d)d^{d+1}\epsilon^{-1})$. A detailed proof of Lemma 3.2.3 and the calculation for order estimation of the mixing time n are given in Sections 3.2.8.6 and 3.2.8.7 respectively.

Despite the above result for the multivariate t-distribution, we next demonstrate through several examples that as long as the tail becomes slightly lighter, we get linear dependency on both the dimension parameter and inverse of the target accuracy parameter. In the next several examples, we use the following result form [CW97] to calculate the LSI constant. Furthermore, following a similar argument in the proof of Lemma 3.2.3, one can show that the potentials satisfy the assumptions required by Theorem 14. However, for simplicity, we directly calculate the LSI constants of the transformed potential and use the result from [VW19].

Corollary 1. [Simplified version of [CW97, Corollary 1.4]] For the Langevin diffusion process with generator $\mathcal{L} = -\nabla f \cdot \nabla + \triangle$, let $\lambda_f(x)$ be the largest eigenvalue of the matrix $\nabla^2 f(x)$ and let $\bar{\beta}(r) = \inf_{|x| \geq r}\{-\lambda_f(x)\}$. If $\sup_{r \geq 0} \bar{\beta}(r) > 0$, then the stationary measure to this Langevin diffusion satisfies LSI with constant $2/\alpha(\mathcal{L})$ such that

$$\alpha(\mathcal{L}) \geq \frac{2}{a_0^2} \exp\left(1 - \int_0^{a_0} r\bar{\beta}(r)dr\right) > 0,$$

where $a_0 > 0$ is the unique solution to the equation $\int_0^a \bar{\beta}(r)dr = 2/a$.

3.2.7.2. *Example 2.* The next potential function f we consider is given by

$$f(x) = \begin{cases} \frac{d+\kappa}{2}\log(1+|x|^2) - \frac{d+\kappa}{2}\log(1+|x|^{-2}) + (\upsilon_f d + 1)\log\log|x| \\ \quad + \left(\upsilon_f + \frac{1}{2}\right) d\log(1 + 2b(\log|x|)^{-1}) & |x| \geq e, \\ (d-1)\log|x| + \frac{d}{2}\log g_{in}^{-1}(|x|)^2 + \left(\frac{1}{2}+\upsilon_f\right) d\log\left(1 + \frac{1}{2} g_{in}^{-1}(|x|)^2\right) \\ \quad - (d-1)\log g_{in}^{-1}(|x|) + \log g_{in}'(g_{in}^{-1})(|x|) + \upsilon_f d\log b \\ \quad + \left[\left(\frac{1}{2}+\upsilon_f\right)d - 1\right]\log 2, & 0 \leq |x| < e. \end{cases}$$

where $\upsilon_f \in (-\frac{3}{2}, \frac{15}{2})$. With the transformation h defined by (3.6), (3.8) and (3.17) and $b = \frac{d}{2\kappa}$, the transformed potential is

$$f_h(x) = \frac{d}{2}|x|^2 + \left(\frac{1}{2}+\upsilon_f\right) d\log\left(1 + \frac{1}{2}|x|^2\right) + \upsilon_f d\log b + \left[\left(\frac{1}{2}+\upsilon_f\right)d - 1\right]\log 2, \quad \forall x \in \mathbb{R}^d.$$

We can find the LSI constant of the transformed density $\pi_h \propto e^{-f_h(x)}$ by [CW97, Corollary 1.4]. First, note that the two eigenvalues of $\nabla^2 f_h(x)$ are

$$\lambda_1(x) = d\left[1 + \left(\frac{1}{2} + \upsilon_f\right)\frac{1}{1+\frac{1}{2}|x|^2}\right], \quad \text{and} \quad \lambda_2(x) = d\left[1 + \left(\frac{1}{2} + \upsilon_f\right)\frac{1-\frac{1}{2}|x|^2}{(1+\frac{1}{2}|x|^2)^2}\right].$$

We now consider the following cases.

(a) When $\upsilon_f = -\frac{1}{2}$: $\lambda_1(x) = \lambda_2(x) = d$. The LSI constant $C_{h,\mathrm{LSI}} = \frac{2}{d}$.

(b) When $\upsilon_f \in (-\frac{1}{2}, \frac{15}{2})$, $\lambda_2(x) < \lambda_1(x)$ for all $x \in \mathbb{R}^d$. Therefore

$$\overline{\beta}(r) = \begin{cases} \left[1 - \frac{1}{8}\left(\upsilon_f + \frac{1}{2}\right)\right]d, & 0 \le r \le \sqrt{6}, \\ \left[1 + \frac{1-\frac{1}{2}r^2}{(1+\frac{1}{2}r^2)^2}\left(\frac{1}{2} + \upsilon_f\right)\right]d, & r > \sqrt{6}. \end{cases}$$

and

$$\int_0^{a_0} \overline{\beta}(r)dr = \frac{2}{a_0} \quad \Longrightarrow \quad a_0 = \left(\frac{2}{1-\frac{1}{8}(\upsilon_f + \frac{1}{2})}d^{-1}\right)^{\frac{1}{2}}.$$

The LSI constant is hence given by

$$C_{h,\mathrm{LSI}} = a_0^2 \exp\left(\int_0^{a_0} r\overline{\beta}(r)dr - 1\right) = \frac{2}{1-\frac{1}{8}(\upsilon_f+\frac{1}{2})}d^{-1}.$$

(c) When $\upsilon_f \in (-\frac{3}{2}, -\frac{1}{2})$, $\lambda_1(x) < \lambda_2(x)$ for all $x \in \mathbb{R}^d$. Therefore

$$\overline{\beta}(r) = \inf_{|x|>r} \lambda_1(x) = \lambda_1(0) = (\frac{3}{2} + \upsilon_f)d.$$

and

$$\int_0^{a_0} \overline{\beta}(r)dr = \frac{2}{a_0} \quad \Longrightarrow \quad a_0 = \left(\frac{2}{\frac{3}{2}+\upsilon_f}d^{-1}\right)^{\frac{1}{2}}.$$

The LSI constant is

$$C_{h,\mathrm{LSI}} = a_0^2 \exp(\int_0^{a_0} r\overline{\beta}(r)dr - 1) = \frac{2}{\frac{3}{2}+\upsilon_f}d^{-1}.$$

Hence, we have that $C_{h,\mathrm{LSI}} = O(d^{-1})$. Combined with the fact that the gradient Lipschitz constant of f_h is $L_h = O(d)$, according to [VW19], the iteration complexity to achieve ϵ error tolerance in KL-divergence is of order $\tilde{O}(d/\epsilon)$, where $\tilde{O}$ hides numerical constants and poly-logarithmic factors.

3.2.7.3. *Example 3.* The next potential function is given by

$$f(x) = \begin{cases} d(1+\dfrac{1}{2b})\log|x| + (\dfrac{d}{2}+1)\log\log|x| + d\log(1+2b(\log|x|)^{-1}) & \\ -(d-1)\log 2 - \dfrac{d}{2}\log b, & |x| > e, \\ (d-1)\log|x| - (d-1)\log g_{in}^{-1}(|x|) + \dfrac{d}{2}g_{in}^{-1}(|x|)^2 & \\ \quad + d\log(1+\dfrac{1}{2}g_{in}^{-1}(|x|)^2) + \log g_{in}'(g_{in}^{-1})(|x|), & 0 \le |x| \le e. \end{cases}$$

As a point of reference, we compare the potential above to the potential function $\tilde{f}(x) = d(1+\frac{1}{2b})\log(1+|x|) + (\frac{d}{2}+1)\log\log(e+|x|)$. According to [Wan14], if $b = d/2\vartheta$, $\tilde{f}$ satisfies the weak Poincaré inequality with ϑ being the degree of freedom. The corresponding transformed potential is then given by

$$f_h(x) = \frac{d}{2}|x|^2 + d\log\left(1+\frac{1}{2}|x|^2\right).$$

The density function induced by this potential function satisfies the LSI and the log-concavity assumption. This follows from by calculating the two eigenvalues of the Hessian matrix $\nabla^2 f_h(x)$, that are given by

$$\lambda_1 = d\left[1+\frac{1}{1+\frac{1}{2}|x|^2}\right], \text{ and } \quad \lambda_2 = d\left[1+\frac{1-\frac{1}{2}|x|^2}{(1+\frac{1}{2}|x|^2)^2}\right].$$

For all $x \in \mathbb{R}^d$, we have that $0 < \lambda_i \le 2d$ for $i = 1, 2$. Therefore the transformed potential f_h is gradient Lipschitz with parameter $2d$. To find the LSI parameter we use [CW97, Corollary 1.4]. For all $x \in \mathbb{R}^d$: $\lambda_2 \le \lambda_1$.

$$\bar{\beta}(r) = \inf_{|x|>r} \lambda_2 = \begin{cases} \dfrac{7}{8}d, & r \in (0, \sqrt{6}], \\ (1+\dfrac{1-\frac{1}{2}r^2}{(1+\frac{1}{2}r^2)^2})d, & r \in (\sqrt{6}, \infty). \end{cases}$$

The solution to the equation $\int_0^a \bar{\beta}(r)dr = 2/a$ is given by $a_0 = \sqrt{16/7d}$. The LSI constant hence satisfies

$$C_{h,\text{LSI}} \leq a_0^2 \exp\left(\int_0^{a_0} r\bar{\beta}(r)dr - 1\right) = \frac{16}{7d}.$$

According to [VW19], the iteration complexity is of order $\tilde{O}(d/\epsilon)$, where $\tilde{O}$ hides only numerical constants and poly-logarithmic factors.

3.2.7.4. *Example 4.* Our next potential function is given by

$$f(x) = \begin{cases} d(1+\dfrac{1}{2b})\log|x| + \log\log|x| + \dfrac{d}{2}\log(1+2b(\log|x|)^{-1}) & \\ -(\dfrac{d}{2}-1)\log 2, & |x| > e, \\ (d-1)\log|x| - (d-1)\log g_{in}^{-1}(|x|) + \dfrac{d}{2}g_{in}^{-1}(|x|)^2 & \\ +\dfrac{d}{2}\log(1+\dfrac{1}{2}g_{in}^{-1}(|x|)^2) + \log g_{in}'(g_{in}^{-1})(|x|), & 0 \leq |x| \leq e. \end{cases}$$

To study the tail-behavior of the original potential function f, we compare it to another potential function $\tilde{f}(x) = d(1+\frac{1}{2b})\log(1+|x|) + \log\log(e+|x|)$. According to [Wan14], if $b = \frac{d}{2\vartheta}$, $\tilde{f}$ satisfies the weak Poincaré inequality with ϑ being the degree of freedom. But compare to the previous example, it has a heavier tail because $1 < \frac{d}{2}+1$.

The transformed potential in this case is given by

$$f_h(x) = \frac{d}{2}|x|^2 + \frac{d}{2}\log\left(1+\frac{1}{2}|x|^2\right).$$

Similar to the previous example, the corresponding density function satisfies LSI and log-concavity assumption. The two eigenvalues of the Hessian matrix are:

$$\lambda_1 = d\left[1+\frac{1}{2}\frac{1}{1+\frac{1}{2}|x|^2}\right], \qquad \text{and} \quad \lambda_2 = d\left[1+\frac{1}{2}\frac{1-\frac{1}{2}|x|^2}{(1+\frac{1}{2}|x|^2)^2}\right].$$

For all $x \in \mathbb{R}^d$, $0 < \lambda_i \leq \frac{3}{2}d$ for $i = 1,2$. Therefore the transformed potential f_h is gradient Lipschitz with parameter $\frac{3}{2}d$. To find the LSI parameter we use [CW97, Cororllary 1.4]. For all

$x \in \mathbb{R}^d$: $\lambda_2 \leq \lambda_1$. Furthermore, we have

$$\bar{\beta}(r) = \inf_{|x|>r} \lambda_2 = \begin{cases} \dfrac{15}{16} d, & r \in (0, \sqrt{6}], \\ \left(1 + \dfrac{1}{2} \dfrac{1 - \frac{1}{2}r^2}{(1 + \frac{1}{2}r^2)^2}\right) d, & r \in (\sqrt{6}, \infty). \end{cases}$$

The solution to the equation $\int_0^a \bar{\beta}(r)dr = 2/a$ is then $a_0 = \sqrt{\frac{32}{15d}}$. The LSI constant $C_{h,\mathrm{LSI}}$ satisfies

$$C_{h,\mathrm{LSI}} \leq a_0^2 \exp\left(\int_0^{a_0} r\bar{\beta}(r)dr - 1\right) = \frac{32}{15d}.$$

According to [VW19], the iteration complexity is of order $\tilde{O}(d/\epsilon)$, where $\tilde{O}$ hides only numerical constants and poly-logarithmic factors.

3.2.7.5. *Example 5.* We next consider the following potential function given by

$$f(x) = \begin{cases} d(1 + \dfrac{1}{2b}) \log|x| - (\dfrac{d}{4} - 1) \log\log|x| + \dfrac{d}{4} \log(1 + 2b(\log|x|)^{-1}) & \\ -(\dfrac{d}{4} - 1) \log 2 + \dfrac{d}{4} \log b & |x| > e \\ (d-1)\log|x| - (d-1)\log g_{in}^{-1}(|x|) + \dfrac{d}{2} g_{in}^{-1}(|x|)^2 + \log g_{in}'(g_{in}^{-1})(|x|) & 0 \leq |x| \leq e \end{cases}$$

To study the tail-behavior of the original potential function f, we compare it to another potential function $\tilde{f}(x) = d(1 + \frac{1}{2b})\log(1 + |x|) - (\frac{d}{4} - 1)\log\log(e + |x|)$. According to [Wan14], with $b = \frac{d}{2\vartheta}$, if $d < 4$, $\tilde{f}$ satisfies the weak Poincaré inequality with ϑ being the degree of freedom. If $d = 4$, $\tilde{f}$ satisfies Poincaré inequality with ϑ being the degree of freedom. If $d > 4$, $\tilde{f}$ satisfies the super Poincaré inequality with ϑ-degree of freedom.

The transformed potential is given by

$$f_h(x) = \frac{d}{2}|x|^2 + \frac{d}{4}\log\left(1 + \frac{1}{2}|x|^2\right).$$

The corresponding density function satisfies LSI and log-concavity assumption. The two eigenvalues of the Hessian matrix are:

$$\lambda_1 = d\left[1 + \frac{1}{4}\frac{1}{1 + \frac{1}{2}|x|^2}\right], \quad \text{and} \quad \lambda_2 = d\left[1 + \frac{1}{4}\frac{1 - \frac{1}{2}|x|^2}{(1 + \frac{1}{2}|x|^2)^2}\right].$$

For all $x \in \mathbb{R}^d$, $0 < \lambda_i \leq \frac{5}{4}d$ for $i = 1, 2$. Therefore the transformed potential f_h is gradient Lipschitz with parameter $\frac{3}{2}d$. To find the LSI parameter we use [CW97, Cororllary 1.4]. For all $x \in \mathbb{R}^d$: $\lambda_2 \leq \lambda_1$.

$$\bar{\beta}(r) = \inf_{|x|>r} \lambda_2 = \begin{cases} \dfrac{31}{32}d, & r \in (0, \sqrt{6}], \\ \left(1 + \dfrac{1}{2}\dfrac{1 - \frac{1}{2}r^2}{(1+\frac{1}{2}r^2)^2}\right)d, & r \in (\sqrt{6}, \infty). \end{cases}$$

The solution to the equation $\int_0^a \bar{\beta}(r)dr = 2/a$ is then $a_0 = \sqrt{64/31d}$. The LSI constant $C_{h,\text{LSI}}$ satisfies

$$C_{h,\text{LSI}} \leq a_0^2 \exp\left(\int_0^{a_0} r\bar{\beta}(r)dr - 1\right) = \frac{64}{31d}.$$

According to [VW19], the iteration complexity is of order $\tilde{O}(d/\epsilon)$, where $\tilde{O}$ hides only numerical constants and poly-logarithmic factors.

3.2.7.6. *Example 6.* As the limiting example of the previous three examples, we consider the potential function

$$f(x) = \begin{cases} d(1+\frac{1}{2b})\log|x| - (\frac{d}{2}-1)\log\log|x| + \log 2 + \frac{d}{2}\log b & \text{for } |x| > e \\ (d-1)\log|x| - (d-1)\log g_{in}^{-1}(|x|) + \frac{d}{2}g_{in}^{-1}(|x|)^2 + \log g_{in}'(g_{in}^{-1})(|x|), & \text{for } 0 \leq |x| \leq e \end{cases}$$

We introduce $\tilde{f}(x) = d(1+\frac{1}{2b})\log(1+|x|) - (\frac{d}{2}-1)\log\log(e+|x|)$ which has similar tail-behavior as the potential f above. According to [Wan14], with $b = \frac{d}{2\vartheta}$, if $d = 1$, $\tilde{f}$ satisfies the weak Poincaré inequality with ϑ being the degree of freedom. If $d = 2$, $\tilde{f}$ satisfies Poincaré inequality with ϑ being the degree of freedom. If $d > 2$, $\tilde{f}$ satisfies the super Poincaré inequality with ϑ-degree of freedom and it induces a density function which has heavier tail than the multivariate t-distribution with ϑ-degree of freedom.

The transformed potential is $f_h(x) = \frac{d}{2}|x|^2$. The Hessian matrix is $\nabla^2 f_h(x) = dI_d$. Therefore f_h is log-concave with parameter d and the corresponding density satisfies LSI with parameter $C_{h,\text{LSI}} \leq 2/d$. According to [VW19], the iteration complexity is of order $\tilde{O}(d/\epsilon)$, where $\tilde{O}$ hides only numerical constants and poly-logarithmic factors.

3.2.8. Proofs. In this section, we will prove the theorems stated in Sections 3.2.5-3.2.7.

3.2.8.1. *Analysis of the transformation maps.* In this section we first analyze the transformation map induced by g defined in (3.6).

Lemma 3.2.4. *If the potential function f satisfies Assumption A0, then we have*

$$\nabla f_h(x) = \begin{cases} \left[f'(g_{in}(|x|))g'_{in}(|x|) - \frac{g''_{in}(|x|)}{g'_{in}(|x|)} - (d-1)\frac{g'_{in}(|x|)}{g_{in}(|x|)} + \frac{d-1}{|x|}\right]\frac{x}{|x|} & |x| < b^{-\frac{1}{\beta}}, \\ \left[f'(e^{b|x|^\beta})b\beta|x|^{\beta-1}e^{b|x|^\beta} - \beta bd|x|^{\beta-1} + \frac{d-\beta}{|x|}\right]\frac{x}{|x|} & |x| \geq b^{-\frac{1}{\beta}}. \end{cases} \tag{3.19}$$

and $\nabla^2 f_h(x)$ has two eigenvalues $\lambda_1 = \lambda_1(|x|)$ and $\lambda_2 = \lambda_2(|x|)$ with λ_1, λ_2 defined as

(1) When $|x| < b^{-\frac{1}{\beta}}$:

$$\begin{aligned}\lambda_1(|x|) = f''(g_{in}(|x|))(g'_{in}(|x|))^2 + f'(g_{in}(|x|))g''_{in}(|x|) - \frac{g^{(3)}_{in}(|x|)}{g'_{in}(|x|)} + (\frac{g''_{in}(|x|)}{g'_{in}(|x|)})^2 \\ - (d-1)\frac{g''_{in}(|x|)}{g_{in}(|x|)} + (d-1)(\frac{g'_{in}(|x|)}{g_{in}(|x|)})^2 - (d-1)|x|^{-2},\end{aligned} \tag{3.20}$$

$$\lambda_2(|x|) = f'(g_{in}(|x|))g'_{in}(|x|)|x|^{-1} - \frac{g''_{in}(|x|)}{|x|g'_{in}(|x|)} - (d-1)\frac{g'_{in}(|x|)}{|x|g_{in}(|x|)} + (d-1)|x|^{-2}. \tag{3.21}$$

(2) When $|x| \geq b^{-\frac{1}{\beta}}$:

$$\begin{aligned}\lambda_1(|x|) = f''(e^{b|x|^\beta})b^2\beta^2|x|^{2(\beta-1)}e^{2b|x|^\beta} + f'(e^{b|x|^\beta})(\beta(\beta-1)b|x|^{\beta-2} + \beta^2b^2|x|^{2(\beta-1)})e^{b|x|^\beta}, \\ - b\beta(\beta-1)d|x|^{\beta-2} - (d-\beta)|x|^{-2}\end{aligned} \tag{3.22}$$

$$\lambda_2(|x|) = f'(e^{b|x|^\beta})b\beta|x|^{\beta-1}e^{b|x|^\beta}|x|^{-1} - b\beta d|x|^{\beta-2} + (d-\beta)|x|^{-2}. \tag{3.23}$$

Proof of Lemma 3.2.4. For a general transformation map induced by h, the transformed potential f_h can be represented as

$$\begin{aligned}f_h(x) &= f(g(|x|)) - \log\det(\nabla h(x)) \\ &= f(g(|x|)) - \log g'(|x|) - (d-1)\log g(|x|) + (d-1)\log|x|.\end{aligned}$$

The gradient of the transformed potential f_h is

$$\nabla f_h(x) = \left[f'(g(|x|))g'(|x|) - \frac{g''(|x|)}{g'(|x|)} - (d-1)\frac{g'(|x|)}{g(|x|)} + \frac{d-1}{|x|} \right] \frac{x}{|x|}. \tag{3.24}$$

Now, (3.19) follows immediately as a consequence of (3.6) and (3.24). The Hessian matrix of f_h can be represented as

$$\nabla^2 f_h(x) = F_1(|x|)\frac{xx^T}{|x|^2} + F_2(|x|)I_d$$

with

$$\begin{aligned} F_1(|x|) &= f''(g(|x|))g'(|x|)^2 + f'(g(|x|))g''(|x|) - f'(g(|x|))\frac{g'(|x|)}{|x|} - \frac{g'''(|x|)}{g'(|x|)} \\ &+ (\frac{g''(|x|)}{g'(|x|)})^2 - (d-1)\frac{g''(|x|)}{g(|x|)} + (d-1)(\frac{g'(|x|)}{g(|x|)})^2 \\ &+ (d-1)\frac{g'(|x|)}{g(|x|)}\frac{1}{|x|} - \frac{2(d-1)}{|x|^2} + \frac{g''(|x|)}{|x|g'(|x|)}, \\ F_2(|x|) &= \left(f'(g(|x|))g'(|x|) - \frac{g''(|x|)}{g'(|x|)} - (d-1)\frac{g'(|x|)}{g(|x|)} + \frac{d-1}{|x|} \right) |x|^{-1}. \end{aligned}$$

Therefore the two eigenvalues of $\nabla^2 f_h(x)$, λ_1 and λ_2 can be written as

$$\begin{aligned} \lambda_1 &= f''(g(|x|))g'(|x|)^2 + f'(g(|x|))g''(|x|) - \frac{g'''(|x|)}{g'(|x|)} + (\frac{g''(|x|)}{g'(|x|)})^2 \\ &- (d-1)\frac{g''(|x|)}{g(|x|)} + (d-1)(\frac{g'(|x|)}{g(|x|)})^2 - \frac{(d-1)}{|x|^2}, \end{aligned} \tag{3.25}$$

$$\lambda_2 = \left(f'(g(|x|))g'(|x|) - \frac{g''(|x|)}{g'(|x|)} - (d-1)\frac{g'(|x|)}{g(|x|)} + \frac{d-1}{|x|} \right) |x|^{-1}. \tag{3.26}$$

The conclusions in (3.20),(3.21),(3.22),(3.23) can be calculated directly from (3.6), (3.25) and (3.26).

■

With the above result on the transformation map h, we can prove Lemma 3.2.1.

Proof of Lemma 3.2.1: We first show that for all $\beta \in [1,2]$, $g \in \mathcal{C}^3((0,\infty))$. It suffices to show that g is three times continuously differentiable at $r = b^{-1/\beta}$. Based on (3.6), we have

$$\begin{aligned}
g_{in}(b^{-\frac{1}{\beta}}{}_{-}) &= e = e^{br^\beta}|_{r=b^{-\frac{1}{\beta}}},\\
g'_{in}(b^{-\frac{1}{\beta}}{}_{-}) &= \beta b^{\frac{1}{\beta}} e = (e^{br^\beta})'|_{r=b^{-\frac{1}{\beta}}},\\
g''_{in}(b^{-\frac{1}{\beta}}{}_{-}) &= \beta(2\beta-1) b^{\frac{2}{\beta}} e = (e^{br^\beta})''|_{r=b^{-\frac{1}{\beta}}},\\
g'''_{in}(b^{-\frac{1}{\beta}}{}_{-}) &= \beta(5\beta^2-6\beta+2) b^{\frac{3}{\beta}} e = (e^{br^\beta})^{(3)}|_{r=b^{-\frac{1}{\beta}}}.
\end{aligned}$$

Next we show g is monotone increasing. From Assumption G1, we know that g_{in} is increasing on the interval $(0, b^{-\frac{1}{\beta}})$. For $r \in [b^{-\frac{1}{\beta}}, \infty)$, $g'(r) = b\beta r^{\beta-1} e^{br^\beta} > 0$. Combined with the fact that $g \in \mathcal{C}^3((0,\infty))$, we obtain that g is monotone increasing on the interval $(0,\infty)$. Furthermore, $g(0) = g_{in}(0) = 0$ and $\lim_{r\to+\infty} g(r) = \lim_{r\to+\infty} e^{br^\beta} = +\infty$. Therefore g is also onto and invertible. ■

Corollary 2. With function g defined in (3.6) and g_{in} satisfying Assumption G1, if the potential function f satisfies Assumption A0, then the transformed potential function f_h defined by (3.5) is twice continuously differentiable, i.e $f_h \in \mathcal{C}^2(\mathbb{R}^d)$.

Proof of Corollary 2: Under the assumptions in Corollary 2, with the results in Lemma 3.2.1 and (3.25),(3.26), we have $f_h \in \mathcal{C}^2(\mathbb{R}^d \backslash \{0\})$. Therefore it remains to show that $\lim_{|x|\to 0^+} \lambda_i(|x|)$ are well-defined for $i = 1, 2$. According to (3.5), we can represent the two eigenvalues of $\nabla^2 f_h(x)$ as for all $r < b^{-\frac{1}{\beta}}$:

$$\begin{aligned}
\lambda_1(r) &= f''_h(r) = f''(g_{in}(r)) g'_{in}(r)^2 + f'(g_{in}(r)) g''_{in}(r) - (d-1)(\frac{d^2}{dr^2} \log \frac{g_{in}(r)}{r}) + \frac{d^2}{dr^2} \log g'_{in}(r)\\
\lambda_2(r) &= \frac{f'_h(r)}{r} = f'(g_{in}(r)) \frac{g'_{in}(r)}{r} - \frac{\frac{d}{dr} \log g'_{in}(r)}{r} - (d-1) \frac{\frac{d}{dr} \log \frac{g_{in}(r)}{r}}{r}
\end{aligned}$$

Since $f \in \mathcal{C}^2(\mathbb{R})$ and g satisfies Assumption G1, $\lim_{r\to 0_+} |\lambda_i(r)| < \infty$ for $i = 1, 2$, which implies that f_h is twice continuously differentiable at the origin. Therefore $f_h \in \mathcal{C}^2(\mathbb{R}^d)$. ■

3.2.8.2. *Proof of Theorem 10.* We first recall a few definition below. Our proof is based on connections between Lyapunov-based techniques and functional inequality-based techniques for proving ergodicity of diffusion process [BCG08].

Definition 3 (Dissipativity condition). The Langevin diffusion with drift function $b(x)$ is said to satisfy the dissipativity condition if there exists constants $r, M > 0$ such that for all $|x| > M$: $\langle b(x), x\rangle \leq -r|x|$.

Definition 4 (Lyapunov condition). A function $V \in \mathcal{D}(\mathcal{L})$ with $V \geq 1$ is a Lyapunov function if there exist constants $\lambda, c > 0$ and a measurable set $K \subset \mathbb{R}^d$ such that $\mathcal{L}V \leq \lambda V(-1 + c1_K)$. Equivalently, we say the $\mathcal{L}$ satisfies the Lyapunov condition.

Lemma 3.2.5. *Consider the dynamics in* (3.9)*. If the drift function* $-\nabla f_h$ *satisfies the dissipativity condition with* $r > 0, M = 8(d-1)/r$*, then the infinitesimal generator* $\mathcal{L}_h$ *of* (3.9) *satisfies Lyapunov condition.*

Proof of Lemma 3.2.5. We first construct a Lyapunov function V with respect to the generator $\mathcal{L}_h$ as

$$V(x) = \begin{cases} 1 & |x| \leq \dfrac{M}{2}, \\ P(|x|) & \dfrac{M}{2} < |x| < M, \\ e^{a|x|} & |x| \geq M, \end{cases}$$

where $P : [\frac{M}{2}, M] \to [1, e^{aM}]$ is a monotone increasing function such that $V \in \mathcal{C}^2(\mathbb{R}^d)$ and $V \geq 1$ for all $x \in \mathbb{R}^d$. When $|x| \geq M$, we have that

$$\begin{aligned}\mathcal{L}_h V(x) &= -\nabla f_h(x) \cdot \nabla(e^{a|x|}) + \triangle(e^{a|x|}) \\ &= -\nabla f_h(x) \cdot (ae^{a|x|}\frac{x}{|x|}) + ae^{a|x|}\left(\frac{d}{|x|} + a - \frac{1}{|x|}\right) \\ &\leq ae^{a|x|}\left(-r + a + \frac{d-1}{|x|}\right).\end{aligned}$$

Picking $a = r/2$, we obtain that

$$\mathcal{L}_h V(x) \leq \frac{r}{2}V(x)(-\frac{r}{2} + \frac{d-1}{|x|}), \qquad \forall |x| \geq M.$$

Since $M = 8(d-1)/r > 4(d-1)/r$, we obtain that $\mathcal{L}_h V(x) \leq -(r^2/8)V(x)$ for all $|x| \geq M$. When $0 \leq |x| < M$, by the fact that $V \in \mathcal{C}^2(\mathbb{R}^d)$ and $V \geq 1$, there exists $A_{r,d}$ such that

$$\frac{\mathcal{L}_h V(x)}{V(x)} \leq A_{r,d} \qquad\qquad \forall\ 0 \leq |x| < M,$$

where $A_{r,d} = \max_{4(d-1)/r \leq |x| \leq 8(d-1)/r} \left(-rP'(|x|) + \triangle(P(|x|))\right) \vee 0$.

Therefore if $-\nabla f_h$ satisfies dissipativity condition with constant $r > 0$, the corresponding generator $\mathcal{L}_h$ satisfies Lyapunov condition with $\lambda = \lambda_r = r^2/8$, $c = c_{r,d} = A_{r,d}/\lambda_r$ an

$$K = K_{r,d} = \{x \in \mathbb{R}^d : 0 \leq |x| \leq 8(d-1)/r\}. \tag{3.27}$$

■

We further recall additional definitions to proceed.

Definition 5 (Local Poincaré inequality). The Markov triple $(\mathbb{R}^d, \mu, \Gamma)$ satisfies a local Poincaré inequality on a measurable set $K \subset \mathbb{R}^d$ with $\mu(K) \in (0, \infty)$ if for some constant C_K and every function $\phi \in \mathcal{D}(\mathcal{E})$:

$$\int_K (\phi - m_K)^2 d\mu \leq C_K \int_K \Gamma(\phi) d\mu$$

where $m_K = \int_K \phi \; d\mu/\mu(K)$.

Lemma 3.2.6. *If the original density satisfies Assumption A0, then the Markov triple* $(\mathbb{R}^d, \pi_h, \Gamma_h)$ *satisfies a local Poincaré inequality on* $K_{r,d}$ *defined in* (3.27).

Proof of Lemma 3.2.6. According to the classical Poincaré inequality with respect to Lebesgue measure, there is a universal constant $C > 0$ such that for all $u \in W^{1,2}(\mathbb{R}^d) \subset W^{1,2}(K_{r,d})$:

$$\int_{K_{r,d}} (u(x) - u_{K_{r,d}})^2 dx \leq C \frac{d-1}{r} \int_{K_{r,d}} |\nabla u|^2 dx$$

where $u_{K_{r,d}} = \fint_{K_{r,d}} u(x)\ dx$. To prove the local Poincaré inequality for $(\mathbb{R}^d, \pi_h, \Gamma_h)$, without loss of generality, we assume that $\phi \in \mathcal{D}(\mathcal{E})$ and $\phi_{K_{r,d}} = \fint_{K_{r,d}} \phi(x)\ dx = 0$. Then

$$\begin{aligned}
\int_{K_{r,d}} (\phi - m_{K_{r,d}})^2 e^{-f_h(x)} dx &= \int_{K_{r,d}} \phi(x)^2 e^{-f_h(x)} dx - (\int_{K_{r,d}} \phi(x) e^{-f_h(x)} dx)^2 / |K_{r,d}| \\
&\leq \left(\sup_{x \in K_{r,d}} e^{-f_h(x)} \right) \int_{K_{r,d}} \phi(x)^2 dx \\
&= \left(\sup_{x \in K_{r,d}} e^{-f_h(x)} \right) \int_{K_{r,d}} (\phi - \phi_{K_{r,d}})^2 dx \\
&\leq C \frac{d-1}{r} \left(\sup_{x \in K_{r,d}} e^{-f_h(x)} \right) \int_{K_{r,d}} |\nabla \phi(x)|^2 dx \\
&\leq C \frac{d-1}{r} \left(\sup_{x \in K_{r,d}} e^{-f_h(x)} \right) \left(\sup_{x \in K_{r,d}} e^{f_h(x)} \right) \int_{K_{r,d}} \Gamma_h(\phi) e^{-f_h(x)} dx.
\end{aligned}$$

Therefore the Markov triple $(\mathbb{R}^d, \pi_h, \Gamma_h)$ satisfies a local Poincaré inequality on $K_{r,d}$ with constant $C_{r,d} = (C(d-1)/r) \left(\sup_{x \in K_{r,d}} e^{-f_h(x)} \right) \left(\sup_{x \in K_{r,d}} e^{f_h(x)} \right)$. ■

Lemma 3.2.7. *If the infinitesimal generator $\mathcal{L}_h$ satisfies the dissipativity condition with constant $r > 0$ and the original density f satisfies Assumption A0, then the Markov triple $(\mathbb{R}^d, \pi_h, \Gamma_h)$ also satisfies a Poincaré inequality.*

Proof of Lemma 3.2.7. According to Lemma 3.2.5, $\mathcal{L}_h$ satisfies Lyapunov condition with $\lambda = \lambda_r = r^2/8$, K as in (3.27) and $c = c_{r,d} = A_{r,d}/\lambda_r$. Therefore for all $m \in \mathbb{R}$ and $\phi \in \mathcal{D}(\mathcal{E}_h)$:

$$\begin{aligned}
\int_{\mathbb{R}^d} (\phi - m)^2 e^{-f_h(x)} dx &\leq \int_{\mathbb{R}^d} (-\frac{8\mathcal{L}_h V}{r^2 V} + \frac{8A_{r,d}}{r^2} 1_{K_{r,d}})(\phi - m)^2 e^{-f_h(x)} dx \\
&= -\frac{8}{r^2} \int_{\mathbb{R}^d} \frac{\mathcal{L}_h V}{V} (\phi - m)^2 e^{-f_h(x)} dx + \frac{8A_{r,d}}{r^2} \int_{K_{r,d}} (\phi - m)^2 e^{-f_h(x)} dx. \quad (3.28)
\end{aligned}$$

Choosing $m = \int_{K_{r,d}} \phi\ e^{-f_h(x)} dx / \int_{K_{r,d}} e^{-f_h(x)} dx$, the second term in (3.28) can be bounded as a result of Lemma 3.2.6:

$$\frac{8A_{r,d}}{r^2} \int_{K_{r,d}} (\phi - m)^2 e^{-f_h(x)} dx \leq \frac{8A_{r,d}}{r^2} C_{r,d} \int_{K_{r,d}} \Gamma_h(\phi) e^{-f_h(x)} dx. \quad (3.29)$$

The first term in (3.28) can be bounded by integration by parts, and the diffusion property of Γ_h as

$$
\begin{aligned}
-\frac{8}{r^2}\int_{\mathbb{R}^d}\frac{\mathcal{L}_h V}{V}(\phi-m)^2 e^{-f_h(x)}dx &= \frac{8}{r^2}\int_{\mathbb{R}^d}\Gamma_h(\frac{(\phi-m)^2}{V},V)e^{-f_h(x)}dx \\
&= \frac{8}{r^2}\int_{\mathbb{R}^d}\left(\frac{2(\phi-m)}{V}\Gamma_h(\phi-m,V)-\frac{(\phi-m)^2}{V^2}\Gamma_h(V)\right)e^{-f_h(x)}dx \\
&\le \frac{8}{r^2}\int_{\mathbb{R}^d}\Gamma_h(\phi-m)e^{-f_h(x)}dx \\
&= \frac{8}{r^2}\int_{\mathbb{R}^d}\Gamma_h(\phi)e^{-f_h(x)}dx. \qquad (3.30)
\end{aligned}
$$

Combining the results in (3.29) and (3.30), we prove the Markov triple $(\mathbb{R}^d,\pi_h,\Gamma_h)$ satisfies a Poincaré inequality with constant C defined as

$$
\begin{aligned}
C &= \frac{8}{r^2}(1+A_{r,d}C_{r,d}) \qquad \text{with} \\
A_{r,d} &= \max_{4(d-1)/r\le|x|\le 8(d-1)/r}\big(-rP'(|x|)+\triangle(P(|x|))\big)\vee 0, \\
C_{r,d} &= \frac{C(d-1)}{r}\left(\sup_{x\in K_{r,d}} e^{-f_h(x)}\right)\left(\sup_{x\in K_{r,d}} e^{f_h(x)}\right).
\end{aligned}
$$

■

Proof of Theorem 10. Applying (3.19) in Lemma 3.2.4, when $|x|\ge 1/b$, we have

$$
\begin{aligned}
\langle\nabla f_h(x),x\rangle &= [f'(e^{b|x|})be^{b|x|}-b-(d-1)\frac{be^{b|x|}}{e^{b|x|}}+\frac{d-1}{|x|}]|x| \\
&= [f'(e^{b|x|})be^{b|x|}-bd]|x|+(d-1) \\
&\ge A|x|-B, \qquad (3.31)
\end{aligned}
$$

where the last inequality follows from Assumption A1 with $\alpha=1$ and $\beta=1$. We now make the following claim.

Claim: The infinitesimal generator $\mathcal{L}_h$ satisfies the dissipativity condition with the constants

$$
\begin{aligned}
r &= \frac{-8(d-1) + \sqrt{64(d-1)^2 + 32AB(d-1)}}{2B} \in (0, A), \\
M &= \frac{8(d-1)}{r}.
\end{aligned}
\tag{3.32}
$$

If the above **Claim** holds, Theorem 10 follows from Lemma 3.2.7 and Theorem 3 in [VW19]. Furthermore, C_h in the statement is given by

$$
\begin{aligned}
C_h &= \frac{r^2}{8}(1 + A_{r,d}C_{r,d})^{-1} \qquad \text{with} \\
A_{r,d} &= \max_{4(d-1)/r \leq |x| \leq 8(d-1)/r} \big(-rP'(|x|) + \triangle(P(|x|))\big) \vee 0, \\
C_{r,d} &= \frac{C(d-1)}{r^2} \left(\sup_{x \in K_{r,d}} e^{-f_h(x)}\right) \left(\sup_{x \in K_{r,d}} e^{f_h(x)}\right).
\end{aligned}
$$

with r defined in (3.32) and C is a universal constant. We now prove the claim

Proof of the Claim: To prove the **Claim**, it suffices to show for all $|x| \geq \max\{N_1, 8(d-1)/r\}$, we have $\langle \nabla f_h(x), x \rangle \geq r|x|$. Based on (3.31), it further suffices to guarantee

$$
A|x| - B \geq r|x| \qquad \text{and} \qquad \frac{8(d-1)}{r} \geq \frac{1}{b}.
$$

When $r = A/2, 8(d-1)/r$, with $N_1 > 3B/A$, the above conditions are easily satisfied, therby completing the proof. ■

3.2.8.3. *Proof of Theorem 12 and Theorem 14.* Theorem 12 and Theorem 14 are both built on the intermediate result that the transformed measure π_h satisfies a LSI. The proof would rely on the following Holley-Stroock theorem.

Theorem 15 (Holley-Stroock Theorem [HS87]). Let $\mu \sim LS(C_\mu)$ and let $\mu_F = Z_F^{-1} e^{-F} \mu$. If F is bounded, then $\mu_F \sim LS(C_{\mu_F})$ and $C_{\mu_F} \leq e^{OscF} C_\mu$ where $OscF := \sup_{x \in \mathbb{R}^d} F(x) - \inf_{x \in \mathbb{R}^d} F(x)$.

As an immediate corollary of Theorem 15, we have $C_{\mu_F} \geq e^{-OscF} C_\mu$.

Lemma 3.2.8. *If the true target density π satisfies Assumption A0 and Assumption A3, then the transformed density π_h satisfies a LSI.*

Proof of Lemma 3.2.8. Based on (3.22) and (3.23) in Section 3.2.8.1, when $|x| \geq b^{-\frac{1}{\beta}}$, we have

$$\begin{aligned}
\lambda_1 &= f''(e^{b|x|^\beta})b^2\beta^2|x|^{2(\beta-1)}e^{2b|x|^\beta} + f'(e^{b|x|^\beta})(\beta(\beta-1)b|x|^{\beta-2} + \beta^2 b^2|x|^{2(\beta-1)})e^{b|x|^\beta} \\
&\quad - \beta(\beta-1)b|x|^{\beta-2} - (d-\beta)|x|^{-2} \\
&= f''(\psi(|x|))\psi'(|x|)^2 + f'(\psi(|x|))\psi''(|x|) - \beta(\beta-1)b|x|^{\beta-2} - (d-\beta)|x|^{-2} \qquad (3.33) \\
\lambda_2 &= f'(e^{b|x|^\beta} + C_\beta)b\beta|x|^{\beta-1}e^{b|x|^\beta}|x|^{-1} - b\beta d|x|^{\beta-2} + (d-\beta)|x|^{-2} \\
&= f'(\psi(|x|))\psi'(|x|)|x|^{-1} - b\beta d|x|^{\beta-2} + (d-\beta)|x|^{-2}. \qquad (3.34)
\end{aligned}$$

where $\psi(r) = e^{br^\beta}$ for all $r \geq b^{-\frac{1}{\beta}}$. If f satisfies Assumption A3, then for all $|x| \geq \tilde{N}_1 := \max\{N_3, b^{-\frac{1}{\beta}}\}$: $\lambda_1(|x|) \geq \rho$ for $i = 1, 2$. We can then construct two potentials:

$$\tilde{f}_h(x) = \begin{cases} f_h(x) & |x| > \tilde{N}_1, \\ g_h(x) & |x| \leq \tilde{N}_1, \end{cases} \qquad \overline{f}_h(x) = \begin{cases} 0 & |x| > \tilde{N}_1, \\ f_h(x) - g_h(x) & |x| \leq \tilde{N}_1. \end{cases}$$

where $g_h : \{|x| \leq \tilde{N}_1\} \subset \mathbb{R}^d \to \mathbb{R}$ is chosen such that $\tilde{f}_h \in \mathcal{C}^2(\mathbb{R}^d)$ and $\nabla^2 g_h(x) \succeq \rho I_d$ for all $|x| \leq \tilde{N}_1$. Therefore, $\nabla^2 \tilde{f}_h(x) \succeq \rho I_d$ for all $x \in \mathbb{R}^d$ i.e $\tilde{f}_h$ is ρ-strongly convex which implies that the measure $\exp(-\tilde{f}_h(x))dx \sim LS(2/\rho)$(see [BÉ85]). Meanwhile, $\overline{f}_h$ is compactly supported on $\{|x| \leq \tilde{N}_1\}$ and $f_h, g_h \in \mathcal{C}^2(\mathbb{R}^d)$, which implies that $\overline{f}_h$ is bounded, i.e $Osc\overline{f}_h < \infty$. Last according to the Holley-Stroock theorem and the fact that $\pi_h \propto \exp(-f_h) = \exp(-\tilde{f}_h)\exp(-\overline{f}_h)$,

$$\pi_h \sim LS(C_{h,\mathrm{LSI}}) \quad \text{with} \quad C_{h,\mathrm{LSI}} = 2e^{Osc\overline{f}_h}/\rho. \qquad (3.35)$$

■

Proof of Theorem 12. The two inequalities in Theorem 12 follows from Lemma 3.2.8 and Theorem 4 in [VW19]. The constant $C_{h,\mathrm{LSI}}$ in Theorem 12 is the same $C_{h,\mathrm{LSI}}$ in (3.35). ■

Lemma 3.2.9. *If the potential function f satisfies Assumption A4, then the transformed potential $f_h(x)$ satisfies the gradient Lipschitz condition, i.e. there exists $L_h > 0$ such that for all $x, y \in \mathbb{R}^d$, we have $|\nabla f_h(x) - \nabla f_h(y)| \leq L_h|x - y|$.*

Proof of Lemma 3.2.9: It suffices to prove that there is a constant L_h such that $\nabla^2 f_h(x) \preceq L_h I_d$ for all $x \in \mathbb{R}^d$, i.e $\lambda_1(|x|), \lambda_2(|x|) \leq L_h$ for all $x \in \mathbb{R}^d$. Based on (3.33),(3.34) in the proof of Lemma 3.2.8, and the fact that f satisfies Assumption A4, we have when $|x| \geq \tilde{N}_2 := \max\{N_4, b^{-\frac{1}{\beta}}\}$: $\lambda_i(|x|) \leq L$ for $i = 1, 2$. When $|x| \leq \tilde{N}_2$, since $f_h \in \mathcal{C}^2(\mathbb{R}^d)$,

$$\max_{|x|\leq \tilde{N}_2} \left\|\nabla^2 f_h(x)\right\| < \infty.$$

Therefore the transformed density f_h is gradient Lipschitz with parameter L_h defined by

$$L_h = \max\{L, \max_{|x|\leq \tilde{N}_2} \left\|\nabla^2 f_h(x)\right\|\}. \tag{3.36}$$

■

Proof of Theorem 14. From Lemma 3.2.9 we have that the transformed potential f_h has Lipschitz gradients with parameter $L_h = \max\{L, \max_{|x|\leq \tilde{N}_2} \left\|\nabla^2 f_h(x)\right\|\}$. Furthermore, as shown in equation (3.35) in the proof of Lemma 3.2.8, $\pi_h \sim LS(C_{h,\mathrm{LSI}})$ with $C_{h,\mathrm{LSI}} = 2e^{Osc\hat{f}_h}/\rho$. Hence, we can apply [VW19, Theorem 1] to obtain that when $0 < \gamma < \frac{1}{2C_{h,\mathrm{LSI}} L_h^2}$,

$$H_{\pi_h}(\rho_n) \leq e^{-\frac{2\gamma n}{C_{h,\mathrm{LSI}}}} H_{\pi_h}(\rho_0) + 4C_{h,\mathrm{LSI}} L_h^2 \gamma d. \tag{3.37}$$

Now, applying Proposition 6 with $\Phi(x) = x \log x$ to (3.37), we get

$$H_{\pi}(\nu_n) \leq e^{-\frac{2\gamma n}{C_{h,\mathrm{LSI}}}} H_{\pi}(\nu_0) + 4C_{h,\mathrm{LSI}} L_h^2 \gamma d. \tag{3.38}$$

The mixing time estimate in the theorem instantly follows from equation (3.38). ■

3.2.8.4. *Proof of Theorem 11 and Theorem 13.* In this section we will prove Theorem 11 and Theorem 13. First we introduce a result which explains the relation between Assumption A2 and Assumption B2.

Lemma 3.2.10. *If a potential function f satisfies Assumption A2, then the transformed potential f_h satisfies Assumption B2.*

Proof of Lemma 3.2.10: If a potential function f satisfies Assumption A2 with parameters μ, N_2 and θ, then when $|x| \geq b^{-\frac{1}{\beta}}$, the eigenvalues of $\nabla^2 f_h(x)$ are studied in (3.22) and (3.23). Applying $\psi(|x|) = e^{b|x|^\beta}$, we have the following estimates on the eigenvalues: for all $|x| \geq \tilde{N}_5 := \max\{b^{-\frac{1}{\beta}}, N_2\}$ we have

$$\begin{aligned}
\lambda_1 &= f''(\psi(|x|))\psi'(|x|)^2 + f'(\psi(|x|))\psi''(|x|) - b\beta(\beta-1)d|x|^{\beta-2} - (d-\beta)|x|^{-2} \\
&\geq \frac{\mu}{(1+\frac{1}{4}|x|^2)^{\frac{\theta}{2}}}, \\
\lambda_2 &= f'(\psi(|x|))\psi'(|x|)|x|^{-1} - b\beta d|x|^{\beta-2} + (d-\beta)|x|^{-2} \\
&\geq \frac{\mu}{(1+\frac{1}{4}|x|^2)^{\frac{\theta}{2}}}.
\end{aligned}$$

where the inequality follows from Assumption A2. Therefore for all $|x| \geq \tilde{N}_5$, we have that

$$\nabla^2 f_h(x) \succeq \frac{\mu}{(1+\frac{1}{4}|x|^2)^{\frac{\theta}{2}}} I_d.$$

Meanwhile since $f_h \in C^2(\mathbb{R}^d)$, we can construct $\tilde{f}_h \in C^2(\mathbb{R}^d)$ such that $\tilde{f}_h(x) = f_h(x)$ for all $|x| \geq \tilde{N}_5$, $\nabla^2 \tilde{f}_h(x) \succeq \frac{\mu}{(1+\frac{1}{4}|x|^2)^{\frac{\theta}{2}}} I_d$ for all $x \in \mathbb{R}^d$. Furthermore, since both f_h and $\tilde{f}_h$ are continuous,

$$\xi := \left\| f_h - \tilde{f}_h \right\|_\infty = \max_{|x| \leq \tilde{N}_5} |f_h(x) - \tilde{f}_h(x)| < \infty \tag{3.39}$$

Therefore f_h satisfies Assumption B2 with parameters $\xi_h = \xi$, $\mu_h = \mu$ and $\theta_h = \theta$. ■

Next we introduce a result which explains the relation between Assumption A1 and Assumption B1.

Lemma 3.2.11. *If a potential function f satisfies Assumption A1 then the transformed potential f_h satisfies Assumption B1.*

Proof of Lemma 3.2.11: If a potential function f satisfies Assumption A1 with parameters α, A, B, as we have shown in (3.19), for all $|x| \geq b^{-\frac{1}{\beta}}$ we have that

$$\begin{aligned}\langle \nabla f_h(x), x\rangle &= f'(e^{b|x|^\beta}) b\beta |x|^\beta e^{b|x|^\beta} - \beta b d |x|^\beta + (d-\beta)\\ &\geq A|x|^\alpha - B.\end{aligned}$$

where the first inequality follows from Assumption A1. Since $f_h \in \mathcal{C}^2(\mathbb{R}^d)$, we have

$$\min_{|x|\leq b^{-\frac{1}{\beta}}} \langle \nabla f_h(x), x\rangle > -\infty$$

Therefore f_h satisfies Assumption B1 with parameters

$$\alpha_h = \alpha, \quad A_h = A, \quad B_h = \max\{0, B, -\min_{|x|\leq b^{-\frac{1}{\beta}}} \langle \nabla f_h(x), x\rangle\} \in [0,\infty). \tag{3.40}$$

■

With the above two results, we are ready to prove Theorem 11.

Proof of Theorem 11: Taking the derivative of the KL-divergence from ρ_t to π_h, we have

$$\begin{aligned}\frac{d}{dt} H_{\pi_h}(\rho_t) &= \frac{d}{dt}\int_{\mathbb{R}^d} \log\left(\frac{\rho_t(x)}{\pi_h(x)}\right)\rho_t(x)dx\\ &= \int_{\mathbb{R}^d} \frac{\partial \rho_t(x)}{\partial t}\log\left(\frac{\rho_t(x)}{\pi_h(x)}\right)dx + \int_{\mathbb{R}^d} \rho_t(x)\frac{\pi_h(x)}{\rho_t(x)}\frac{1}{\pi_h(x)}\frac{\partial \rho_t(x)}{\partial t}dx\\ &= \int_{\mathbb{R}^d} \nabla\cdot\left(\rho_t(x)\nabla\log\frac{\rho_t(x)}{\pi_h(x)}\right)\log\frac{\rho_t(x)}{\pi_h(x)}dx + 0\\ &= -\int_{\mathbb{R}^d}\rho_t(x)\left|\nabla\log\frac{\rho_t(x)}{\pi_h(x)}\right|^2 dx = -I_{\pi_h}(\rho_t), \qquad (3.41)\end{aligned}$$

where third identity follows from the Fokker-Planck equation

$$\frac{\partial \rho_t}{\partial t} = \nabla\cdot(\rho_t \nabla f_h) + \Delta \rho_t = \nabla\cdot\left(\rho_t\nabla\log\frac{\rho_t}{\pi_h}\right),$$

and the fact that $\int \frac{\partial \rho_t}{\partial t}dx = \frac{d}{dt}\int \rho_t dx = 0$. According to Lemma 3.2.10, we have that π_h satisfies Assumption B2. Hence, according to [EH20, Theorem 1], π_h satisfies a modified LSI, i.e. for all probability densities ρ:

$$H_{\pi_h}(\rho) \leq C_{h,\text{M-LSI}} I_{\pi_h}(\rho)^{1-\delta} M_s(\rho + \pi_h)^{\delta},$$

where $M_s(\rho) = \int_{\mathbb{R}^d} (1 + |x|^2)^{s/2} \rho(x) dx$ is the s-th moment of any function ρ and with ξ defined in (3.39), δ and λ are defined as

$$\delta := \frac{\theta}{s - 2 + 2\theta} \in [0, \frac{1}{2}), \qquad C_{h,\text{M-LSI}} = 4e^{2\xi} \mu^{-\frac{s-2}{s-2+2\theta}}. \tag{3.42}$$

Hence (3.41) can be further written as

$$\frac{d}{dt} H_{\pi_h}(\rho_t) \leq -\lambda^{-\frac{1}{1-\delta}} H_{\pi_h}(\rho_t)^{\frac{1}{1-\delta}} M_s(\rho_t + \pi_h)^{-\frac{\delta}{1-\delta}}. \tag{3.43}$$

According to Lemma 3.2.11, the transformed potential f_h satisfies Assumption B1 with parameters α_h, A_h, B_h. Hence, according to [TV00, Proposition 2], under the α_h-dissipativity of f_h, for all $s \geq 2$:

$$M_s(\rho_t + \pi_h) \leq M_s(\rho_0 + \pi_h) + C_s t,$$

where

$$C_s = \sup_{x \geq 0} \left((ds + s(s-2) - sA_h + sB_h) x^{\frac{s-2}{s-2+\alpha}} - A_h x \right) < \infty. \tag{3.44}$$

Therefore the upper bound in (3.43) can improved as

$$\frac{d}{dt} H_{\pi_h}(\rho_t) \leq -\lambda^{-\frac{1}{1-\delta}} H_{\pi_h}(\rho_t)^{\frac{1}{1-\delta}} \left(M_s(\rho_0 + \pi_h) + C_s t \right)^{-\frac{\delta}{1-\delta}}. \tag{3.45}$$

Rewriting (3.45) as

$$-H_{\pi_h}(\rho_t)^{-\frac{1}{1-\delta}} \frac{d}{dt} H_{\pi_h}(\rho_t) \geq (\lambda C_s^{\delta})^{-\frac{1}{1-\delta}} (M_s(\rho_0 + \pi_h) C_s^{-1} + t)^{-\frac{\delta}{1-\delta}}, \tag{3.46}$$

and applying Gronwall's inequality, we obtain

$$H_{\pi_h}(\rho_t) \leq (\frac{1-2\delta}{\delta})^{\frac{1-\delta}{\delta}} (\lambda C_s^{\delta})^{\frac{1}{\delta}} (M_s(\rho_0 + \pi_h) C_s^{-1} + t)^{-\frac{1-2\delta}{\delta}} \leq \frac{C}{t^l}. \tag{3.47}$$

with $C = (\frac{1-2\delta}{\delta})^{\frac{1-\delta}{\delta}} (\lambda C_s^{\delta})^{\frac{1}{\delta}}$ and $l = (1-2\delta)/\delta$. ■

To prove Theorem 13, we require the following result on the relationship between Assumption A5 and Assumption B5.

Lemma 3.2.12. *If the density π satisfies Assumption A5, then π satisfies Assumption B5.*

Proof of Lemma 3.2.12. Without loss of generality, we can assume that $N_5 \geq e$. When $\lambda \geq N_5 \geq e$,

$$\pi\left\{|\cdot| \geq m+\lambda\right\} \leq 2\exp\left(-\left(\frac{g^{-1}(\lambda)}{C_{\text{TAIL}}}\right)^{\alpha_1}\right)$$

with $C_{\text{TAIL}} = C^*_{\text{TAIL}}$. When $\lambda \in [0, N_5]$, we have

$$\begin{aligned}\pi\left\{|\cdot| \geq m+\lambda\right\} &\leq \pi\left\{|\cdot| \geq m\right\} \\ &\leq 2\exp\left(-\left(\frac{g^{-1}(\lambda)}{C_{\text{TAIL}}}\right)^{\alpha_1}\right)\end{aligned}$$

with $C_{\text{TAIL}} = g^{-1}(N_5)\left(\log\frac{2}{\pi\{|\cdot|\geq m\}}\right)^{\frac{1}{\alpha_1}}$. Therefore for all $\lambda \geq 0$,

$$\pi\left\{|\cdot| \geq m+\lambda\right\} \leq 2\exp\left(-\left(\frac{g^{-1}(\lambda)}{C_{\text{TAIL}}}\right)^{\alpha_1}\right), \tag{3.48}$$

with

$$C_{\text{TAIL}} = \max\left\{C^*_{\text{TAIL}}, g^{-1}(N_5)\left(\log\frac{2}{\pi\left\{|\cdot| \geq m\right\}}\right)^{\frac{1}{\alpha_1}}\right\}.$$

From (3.48), let $X \in \mathbb{R}^d$ be a random variable with density π and $Y := h^{-1}(X)$. Then $Y \in \mathbb{R}^d$ is a random variable with density π_h. We get

$$\begin{aligned}\pi_h\left\{|\cdot| \geq m_h+\lambda\right\} &= \mathbb{P}\left(|Y| \geq m_h+\lambda\right) \\ &= \mathbb{P}\left(g^{-1}(|X|) \geq m_h+\lambda\right) \\ &= \mathbb{P}\left(|X| \geq g(m_h+\lambda)\right).\end{aligned}$$

For any fixed $\lambda \geq 0$, we can choose $m_h(\lambda) = g^{-1}(m + g(\lambda)) - \lambda$ and we get

$$\begin{aligned}
\pi_h \{ |\cdot| \geq m_h(\lambda) + \lambda \} &\leq \mathbb{P}\left(|X| \geq m + g(\lambda) \right) \\
&= \pi \{ |\cdot| \geq m + g(\lambda) \} \\
&\leq 2 \exp\left(- \left(\frac{\psi^{-1}(g(\lambda))}{C_{\text{TAIL}}} \right)^{\alpha_1} \right) \\
&= 2 \exp\left(- \left(\frac{\lambda}{C_{\text{TAIL}}} \right)^{\alpha_1} \right).
\end{aligned}$$

We next claim that there exists a constant m_h such that $m_h(\lambda) \leq m_h$ for all $\lambda \geq 0$. To prove the claim, we apply Taylor expansion in the definition of $m_h(\lambda)$ and we get for any $\lambda \geq 0$, there exists a constant $\theta(\lambda) \in [0, m]$ such that

$$\begin{aligned}
m_h(\lambda) &= g^{-1}\left(g(\lambda) \right) + (g^{-1})'(g(\lambda)) m - \lambda \\
&\leq m \sup_{r \in [0,\infty)} \left(g^{-1})'(r) \right).
\end{aligned}$$

According to our construction of g, we have that $\sup_{r \in [0,\infty)} \left(g^{-1})'(r) \right) < \infty$. Therefore we can pick $m_h = m \left(\sup_{r \in [0,\infty)} \left(g^{-1})'(r) \right) \right)$ which is a constant independent of λ, which proves the claim. Hence, we get for all $\lambda \geq 0$,

$$\pi_h \{ |\cdot| \geq m_h + \lambda \} \leq 2 \exp\left(- \left(\frac{\psi^{-1}(g(\lambda))}{C_{\text{TAIL}}} \right)^{\alpha_1} \right).$$

That is, the transformed density π_h satisfies Assumption B5 with

$$\alpha_h = \alpha_1, \tag{3.49}$$

$$C_{h,\text{TAIL}} = \max \left\{ C^*_{\text{TAIL}}, g^{-1}(N_6) \left(\log \frac{2}{\pi \{ |\cdot| \geq m \}} \right)^{\frac{1}{\alpha_1}} \right\}, \tag{3.50}$$

$$m_h = m \left(\sup_{r \in [0,\infty)} \left(g^{-1})'(r) \right) \right). \tag{3.51}$$

■

Proof of Theorem 13. Let $\hat{\pi}_h$ be a modified density to π_h. It's defined as, for $\hat{\gamma}, \hat{R} > 0$,

$$\hat{\pi}_h \propto \exp(-\hat{f}_h), \qquad \hat{f}_h(x) := f_h(x) + \frac{\hat{\gamma}}{2}(|x| - \hat{R})_+^2. \tag{3.52}$$

Here $(|x| - \hat{R})_+^2$ is interpreted as $\max\left\{|x| - \hat{R}, 0\right\}^2$. Furthermore, $\hat{R}$ is chosen so that $\pi_h(B(0, \hat{R})) \geq \frac{1}{2}$, where $B(0, \hat{R})$ is an Euclidean ball of radius $\hat{R}$ centered at zero. With this definition, the proof follows immediately from Lemma 3.2.10, [EH20, Theorem 1] and [CEL$^+$21, Theorem 8]. ■

3.2.8.5. *Proofs for Section 3.2.6.*

Proof of Proposition 9: When $|x| \geq g(b^{-\frac{1}{\beta}}) = e$, the inverse of g can be represented as

$$g^{-1}(|x|) = b^{-\frac{1}{\beta}} \log^{\frac{1}{\beta}} |x|.$$

Therefore, Assumption A3 can be reformulated as for all $|x| \geq g^{-1}(N_3 \vee b^{-\frac{1}{\beta}}) := g^{-1}(\tilde{N}_1)$:

$$b^{\frac{2}{\beta}}[f'(|x|)\beta \log^{1-\frac{2}{\beta}}(|x|)|x| - \beta d \log^{1-\frac{2}{\beta}}(|x|) + (d-\beta)\log^{-\frac{2}{\beta}}(|x|)] > \rho, \tag{3.53}$$

$$\begin{aligned} &b^{\frac{2}{\beta}}[f''(|x|)\beta^2 \log^{2-\frac{2}{\beta}}(|x|)|x|^2 + f'(|x|)\beta(\beta-1)\log^{1-\frac{2}{\beta}}(|x|)|x| \\ &\quad + f'(|x|)\beta^2 \log^{2-\frac{2}{\beta}}(|x|)|x| - \beta(\beta-1)\log^{1-\frac{2}{\beta}}(|x|) - (d-\beta)\log^{-\frac{2}{\beta}}(|x|)] > \rho. \end{aligned} \tag{3.54}$$

Now, (3.53) gives a lower bound on $f'(|x|)$ of the form:

$$f'(|x|) \geq \rho b^{-\frac{2}{\beta}} \beta^{-1} \log^{-(1-\frac{2}{\beta})}(|x|)|x|^{-1} + d|x|^{-1} - \frac{d-\beta}{\beta}\log^{-1}(|x|)|x|^{-1}.$$

Defining $N := g^{-1}(\tilde{N}_1)$ and integrating from N to a larger value with respect to $|x|$, we obtain

$$\begin{aligned} f(|x|) &\geq f(N) + \frac{1}{2}\rho b^{-\frac{2}{\beta}}(\log^{\frac{2}{\beta}}(|x|) - \log^{\frac{2}{\beta}} N) + d(\log|x| - \log N) - \frac{d-\beta}{\beta}(\log\log|x| - \log\log N) \\ &:= C_{N,d} + \frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}}(|x|) + d\log|x| - \frac{d-\beta}{\beta}\log\log|x|. \end{aligned}$$

Therefore we have for all $|x| > N$:

$$e^{f(x)} \geq C_{N,d}|x|^{\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|x|)+d} \log^{-\frac{d-\beta}{\beta}}(|x|). \tag{3.55}$$

To prove that π satisfies the Poincaré-type inequalities, we leverage the results in [Wan14] and [WW15]. We consider the following quantity with $\vartheta \in (0,2)$ and $x \neq y$:

$$\frac{e^{f(x)} + e^{f(y)}}{|x-y|^{d+\vartheta}}$$

Since $|x-y|^{d+\vartheta} \leq 2^{d+\vartheta-1}(|x|^{d+\vartheta} + |y|^{d+\vartheta})$, we have for all $x \neq y$ and $|x|, |y| > N$, we have

$$\frac{e^{f(x)} + e^{f(y)}}{|x-y|^{d+\vartheta}} \geq \frac{C_{N,d}}{2^{d+\vartheta-1}} \frac{|x|^{\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|x|)+d} \log^{-\frac{d-\beta}{\beta}}(|x|) + |y|^{\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|y|)+d} \log^{-\frac{d-\beta}{\beta}}(|y|)}{|x|^{d+\vartheta} + |y|^{d+\vartheta}}. \tag{3.56}$$

Then, (3.54) gives the lower bound

$$\begin{aligned} &f''(|x|) + f'(|x|)\left(\frac{\beta-1}{\beta}\log^{-1}(|x|)|x|^{-1} + |x|^{-1}\right) \\ &\quad \geq \rho b^{-\frac{2}{\beta}}\beta^{-2}\log^{-(2-\frac{2}{\beta})}(|x|)|x|^{-2} + \frac{\beta-1}{\beta}\log^{-1}(|x|)|x|^{-2} + \frac{d-\beta}{\beta^2}\log^{-2}(|x|)|x|^{-2}. \end{aligned}$$

By multiplying $\log^{1-\frac{1}{\beta}}(|x|)|x|$ on both sides, for all $|x| > N$ we obtain

$$\begin{aligned} \frac{d}{d|x|}\left(f'(|x|)\log^{1-\frac{1}{\beta}}(|x|)|x|\right) \geq & \rho b^{-\frac{2}{\beta}}\beta^{-2}\log^{-(1-\frac{1}{\beta})}(|x|)|x|^{-1} + \frac{\beta-1}{\beta}\log^{-\frac{1}{\beta}}(|x|)|x|^{-1} \\ & + \frac{d-\beta}{\beta}\log^{-1-\frac{1}{\beta}}(|x|)|x|^{-1}, \end{aligned}$$

which implies that

$$f'(|x|) \geq C_{N,d,1}\log^{-(1-\frac{1}{\beta})}(|x|)|x|^{-1} + \rho b^{-\frac{2}{\beta}}\beta^{-1}\log^{-(1-\frac{2}{\beta})}(|x|)|x|^{-1} + |x|^{-1} - (d-\beta)\log^{-1}(|x|)|x|^{-1}.$$

Further integration implies that for all $|x| > N$, we have

$$e^{f(|x|)} \geq C_{N,d,2}|x|^{1+\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|x|)+C_{N,d,1}\beta\log^{\frac{1}{\beta}-1}(|x|)} \log^{-(d-\beta)}|x|. \tag{3.57}$$

Since $d \geq 1$, (3.55) is stronger than (3.57), when we apply results in [Wan14], it's enough for us to consider only (3.55). Therefore we have the following results:

(1) When $\beta \in (1,2)$ or $\beta = 2, \vartheta < \frac{1}{2}\rho b^{-1}$, we can see that for all $|x| > N$:

$$\frac{1}{2}\rho b^{-\frac{2}{\beta}}\log^{\frac{2}{\beta}-1}(|x|) + d - (d+\vartheta) > 0$$

Therefore, with (3.56), conditions in [Wan14, Theorem 1.1-(3)] is satisfied with

$$\omega(x) = \frac{C_{N,d}}{2^{d+\vartheta}} |x|^{\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|x|) - \vartheta} \log^{-\frac{d-\beta}{\beta}}(|x|),$$

such that $\lim_{|x|\to\infty} \omega(x) = \infty$. Hence, $\pi \propto \exp(-f)$ satisfies the super-Poincaré inequality.

(2) When $\beta = 2, \vartheta = \frac{1}{2}\rho b^{-1}, d = 1, 2$, we can see that for all $|x| > N$, we have

$$\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|x|) + d - (d + \vartheta) = 0$$

Therefore with (3.56), since $d = 1, 2$ we obtain

$$\frac{e^{f(x)} + e^{f(y)}}{|x - y|^{d+\vartheta}} \geq \frac{C_{N,d}}{2^{d+\vartheta-1}}.$$

Hence, according to in [Wan14, Theorem 1.1-(1)], $\pi \propto \exp(-f)$ satisfies the Poincaré inequality.

(3) When $\beta = 2, \vartheta = \frac{1}{2}\rho b^{-1}, d \geq 3$, for all $|x| > N$, we have that

$$\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|x|) + d - (d + \vartheta) = 0$$

However, the lower bound in (3.56) goes to zero as $|x|, |y| \to \infty$. Neither Poincaré inequality nor super Poincaré inequality is guaranteed. However, according to [Wan14, Theorem 1.1-(2)], $\pi \propto \exp(-f)$ satisfies the weak Poincaré inequality with $\alpha(r)$ as defined in (3.15). When $\beta = 2, \vartheta > \frac{1}{2}\rho b^{-1}$, we can see that for all $|x| > N$:

$$\frac{1}{2}\rho b^{-\frac{2}{\beta}} \log^{\frac{2}{\beta}-1}(|x|) + d - (d + \vartheta) < 0$$

Neither Poincaré inequality nor super Poincaré inequality is guaranteed. However, according to [Wan14, Theorem 1.1-(2)], $\pi \propto \exp(-f)$ satisfies the weak Poincaré inequality with $\alpha(r)$ as in (3.15).

■

Proof of Proposition 7: Similar to Assumption A3, Assumption A1 are sufficient conditions for Poincaré type inequalities as well. First note that Assumption A1 is equivalent to the following inequality: for all $|x| > N := g^{-1}(N_1 \vee b^{-\frac{1}{\beta}})$ we have

$$f'(|x|) \geq A\beta^{-1}b^{-\frac{\alpha}{\beta}} \log^{\frac{\alpha}{\beta}-1}(|x|)|x|^{-1} + d|x|^{-1} - \frac{B}{\beta}\log^{-1}(|x|)|x|^{-1}.$$

Integrating with respect to $|x|$, we obtain

$$f(|x|) \geq C_{N,d} + Ab^{-\frac{\alpha}{\beta}}\alpha^{-1}\log^{\frac{\alpha}{\beta}}(|x|) + d\log|x| - \frac{B}{\beta}\log\log|x|.$$

For all $|x| \geq N$, we then have

$$e^{f(|x|)} \geq C_{N,d}|x|^{A\alpha^{-1}b^{-\frac{\alpha}{\beta}}\log^{\frac{\alpha}{\beta}-1}(|x|)+d}\log^{-\frac{B}{\beta}}|x|.$$

and

(3.58)
$$\frac{e^{f(x)}+e^{f(y)}}{|x-y|^{d+\vartheta}} \geq \frac{C_{N,d}}{2^{d+\vartheta-1}}\frac{|x|^{A\alpha^{-1}b^{-\frac{\alpha}{\beta}}\log^{\frac{\alpha}{\beta}-1}(|x|)+d}\log^{-\frac{B}{\beta}}|x| + |y|^{A\alpha^{-1}b^{-\frac{\alpha}{\beta}}\log^{\frac{\alpha}{\beta}-1}(|y|)+d}\log^{-\frac{B}{\beta}}|y|}{|x|^{d+\vartheta}+|y|^{d+\vartheta}}.$$

We now consider different cases.

(1) When $\alpha > \beta$ or $\alpha = \beta, \vartheta < A\beta^{-1}b^{-1}$, we can see that for all $|x| > N$ we have that

$$A\alpha^{-1}b^{-\frac{\alpha}{\beta}}\log^{\frac{\alpha}{\beta}-1}(|x|) + d - (d+\vartheta) > 0.$$

Therefore, with (3.58), the conditions in [Wan14, Theorem 1.1-(3)] are satisfied with

$$\omega(x) = \frac{C_{N,d}}{2^{d+\vartheta}}|x|^{A\alpha^{-1}b^{-\frac{\alpha}{\beta}}\log^{\frac{\alpha}{\beta}-1}(|x|)-\vartheta}\log^{-\frac{B}{\beta}}(|x|),$$

such that $\lim_{|x|\to\infty}\omega(x) = \infty$. Hence, $\pi \propto \exp(-f)$ satisfies the super Poincaré inequality.

(2) When $\alpha = \beta, \vartheta = A\beta^{-1}b^{-1}$, we can see that for all $|x| > N$ we have

$$A\alpha^{-1}b^{-\frac{\alpha}{\beta}}\log^{\frac{\alpha}{\beta}-1}(|x|) + d - (d+\vartheta) = 0.$$

However, the lower bound in (3.58) goes to zero as $|x|, |y| \to \infty$. Hence, Neither the Poincaré inequality nor the super Poincaré inequality is satisfied. However, according to

[Wan14, Theorem 1.1-(2)], the density $\pi \propto \exp(-f)$ satisfies the weak Poincaré inequality with $\alpha(r)$ as in (3.15).

(3) When $\alpha = \beta, \vartheta > A\beta^{-1}b^{-1}$, we can see that for all $|x| > N$ we have

$$A\alpha^{-1}b^{-\frac{\alpha}{\beta}}\log^{\frac{\alpha}{\beta}-1}(|x|) + d - (d+\vartheta) < 0.$$

Hence, neither the Poincaré inequality nor super Poincaré inequality is guaranteed. However, according to [Wan14, Theorem 1.1-(2)], $\pi \propto \exp(-f)$ satisfies the weak Poincaré inequality with $\alpha(r)$ as in (3.15).

■

Proof of Proposition 8: Similar as in the proof of Proposition 9, Assumption A2 is equivalent to the following two inequalities: for all $|x| \geq N$,

(3.59)

$$f'(|x|) \geq \mu b^{-\frac{2}{\beta}}\beta^{-1}\log^{-(1-\frac{2}{\beta})}(|x|)(1+\frac{1}{4}b^{-\frac{2}{\beta}}\log^{\frac{2}{\beta}}(|x|))^{-\frac{\theta}{2}}|x|^{-1} + d|x|^{-1} - \frac{d-\beta}{\beta}\log^{-1}(|x|)|x|^{-1},$$

$$f''(|x|) + f'(|x|)\left(\frac{\beta-1}{\beta}\log^{-1}(|x|)|x|^{-1} + |x|^{-1}\right)$$

(3.60)

$$\geq \mu b^{-\frac{2}{\beta}}\beta^{-2}\log^{-(2-\frac{2}{\beta})}(|x|)(1+\frac{1}{4}b^{-\frac{2}{\beta}}\log^{\frac{2}{\beta}}(|x|))^{-\frac{\theta}{2}}|x|^{-2} + \frac{\beta-1}{\beta}\log^{-1}(|x|)|x|^{-2} + \frac{d-\beta}{\beta^2}\log^{-2}(|x|)|x|^{-2}.$$

Choosing $N' > N$ such that for $|x| > N'$, $b^{-\frac{2}{\beta}}\log^{\frac{2}{\beta}}(|x|) > 4/3$, it then implies from (3.59) that for all $|x| > N'$ we have

$$e^{f(|x|)} > C_{N,d}|x|^{(2-\theta)^{-1}\mu b^{-\frac{2-\theta}{\beta}}\log^{\frac{2-\theta}{\beta}-1}(|x|)+d}\log^{-\frac{d-\beta}{\beta}}(|x|). \tag{3.61}$$

Furthermore, (3.60) implies that for all $|x| > N'$,

$$e^{f(|x|)} > C_{N,d}|x|^{(1-\theta)^{-1}(2-\theta)^{-1}\mu b^{-\frac{2-\theta}{\beta}}\log^{\frac{2-\theta}{\beta}-1}(|x|)+C_N\log^{\frac{1}{\beta}-1}(|x|)+1}\log^{-(d-\beta)}(|x|). \tag{3.62}$$

We now consider the different cases as before.

(1) When $\theta < 2 - \beta$, (3.62) is stronger than (3.61). We have that for large $|x|$,

$$(1-\theta)^{-1}(2-\theta)^{-1}\mu b^{-\frac{2-\theta}{\beta}} \log^{\frac{2-\theta}{\beta}-1}(|x|) + C_N \log^{\frac{1}{\beta}-1}(|x|) + 1 - (d+\vartheta) > 0.$$

Therefore, when $\theta < 2 - \beta$, the conditions in [Wan14, Theorem 1.1-(3)] are satisfied with

$$\omega(x) = \frac{C_{N,d}}{2^{d+\vartheta}} |x|^{(1-\theta)^{-1}(2-\theta)^{-1}\mu b^{-\frac{2-\theta}{\beta}} \log^{\frac{2-\theta}{\beta}-1}(|x|)+1-(d+\vartheta)} \log^{-\frac{d-\beta}{\beta}}(|x|),$$

with $\lim_{|x|\to\infty} \omega(x) = \infty$. Hence, $\pi \propto \exp(-f)$ satisfies the super Poincaré inequality.

(2) When $\theta = 2-\beta, \mu\beta^{-1}b^{-1} > \vartheta$, (3.61) is stronger than (3.62). We have that for all $|x| > N'$,

$$(2-\theta)^{-1}\mu b^{-\frac{2-\theta}{\beta}} \log^{\frac{2-\theta}{\beta}-1}(|x|) + d - (d+\vartheta) > 0.$$

Therefore, when $\theta = 2 - \beta, \mu\beta^{-1}b^{-1} > \vartheta$, the conditions in [Wan14, Theorem 1.1-(3)] are satisfied with

$$\omega(x) = \frac{C_{N,d}}{2^{d+\vartheta}} |x|^{b^{-\frac{2-\theta}{\beta}} \log^{\frac{2-\theta}{\beta}-1}(|x|)-\vartheta} \log^{-\frac{d-\beta}{\beta}}(|x|),$$

with $\lim_{|x|\to\infty} \omega(x) = \infty$. Hence, $\pi \propto \exp(-f)$ satisfies the super Poincaré inequality.

(3) When $\theta = 2 - \beta, \mu\beta^{-1}b^{-1} \leq \vartheta$, (3.61) is stronger than (3.62). We have that for all $|x|$ large enough,

$$(2-\theta)^{-1}\mu b^{-\frac{2-\theta}{\beta}} \log^{\frac{2-\theta}{\beta}-1}(|x|) + d - (d+\vartheta) \leq 0.$$

Neither Poincaré inequality nor super Poincaré inequality is guaranteed. According to [Wan14, Theorem 1.1-(2)], $\pi \propto \exp(-f)$ satisfies the weak Poincaré inequality with $\alpha(r)$ in (3.15).

(4) When $\theta > 2 - \beta$, (3.59) is stronger than (3.62). We have that for all $|x|$ large enough,

$$(2-\theta)^{-1}\mu b^{-\frac{2-\theta}{\beta}} \log^{\frac{2-\theta}{\beta}-1}(|x|) + d - (d+\vartheta) < 0.$$

Hencce, neither the Poincaré inequality nor the super Poincaré inequality is guaranteed. However, according to [Wan14, Theorem 1.1-(2)], $\pi \propto \exp(-f)$ satisfies the weak Poincaré inequality with $\alpha(r)$ in (3.15).

■

3.2.8.6. *Proofs for Section 3.2.7.*

Proof of Lemma 3.2.2: First, it's easy to check that $g_{in}(0) = 0$ and $g_{in}(b^{-\frac{1}{2}}) = e$, which implies that $g \in \mathcal{C}([0,\infty))$. Next note that we have

$$\log \frac{g_{in}(r)}{r} = \log(b^{\frac{1}{2}}) + br^2 - \frac{10}{3}b^{\frac{3}{2}}r^3 + \frac{15}{4}b^2r^4 - \frac{6}{5}b^{\frac{5}{2}}r^5 + \frac{47}{60}.$$

Hence, we can then check that

$$\lim_{r\to 0_+} \left| \frac{\frac{d}{dr}\log\frac{g_{in}(r)}{r}}{r} \right| < \infty \text{ and } \lim_{r\to 0_+} \left| \frac{d}{dr^2} \log \frac{g_{in}(r)}{r} \right| < \infty.$$

Note that the first derivative of g_{in} is given by

$$g_{in}'(r) = b^{\frac{1}{2}}\left(1 + 2br^2 - 10b^{\frac{3}{2}}r^3 + 15b^2r^4 - 6b^{\frac{5}{2}}r^5\right)\exp\left(br^2 - \frac{10}{3}b^{\frac{3}{2}}r^3 + \frac{15}{4}b^2r^4 - \frac{6}{5}b^{\frac{5}{2}}r^5 + \frac{47}{60}\right).$$

Hence, we have that

$$\lim_{r\to 0_+} |f'(g_{in}(r))g_{in}'(r)| = (d+\varepsilon) \lim_{r\to 0_+} \left| \frac{g_{in}'(r)}{1+g_{in}(r)^2} \right| \frac{g_{in}(r)}{r} < \infty.$$

Similarly, as $g_{in}'(b^{-\frac{1}{2}}) = 2b^{\frac{1}{2}}e$ and

$$\begin{aligned}\log g_{in}'(r) =& \log(b^{\frac{1}{2}}) + \log(1 + 2br^2 - 10b^{\frac{3}{2}}r^3 + 15b^2r^4 - 6b^{\frac{5}{2}}r^5) \\ &+ br^2 - \frac{10}{3}b^{\frac{3}{2}}r^3 + \frac{15}{4}b^2r^4 - \frac{6}{5}b^{\frac{5}{2}}r^5 + \frac{47}{60},\end{aligned}$$

we can also check that

$$\lim_{r\to 0_+} \left| \frac{d^2}{dr^2} \log g_{in}'(r) \right| < \infty \text{ and } \lim_{r\to 0_+} \left| \frac{\frac{d}{dr}\log g_{in}'(r)}{r} \right| < \infty.$$

Similarly, by taking additional higher order derivatives it is easy to check that $g_{in}^{''}(b^{-\frac{1}{2}}) = 6be$ and $g_{in}^{'''}(b^{-\frac{1}{2}}) = 20b^{\frac{3}{2}}e$. We omit the tedious but elementary calculations here. ■

Proof of Lemma 3.2.3. It's obvious that $f \in \mathcal{C}^2(\mathbb{R}^d)$ and it's isotropic. Note that

$$\begin{aligned}\frac{d}{d|x|}f(|x|) &= (d+\kappa)\frac{|x|}{1+|x|^2},\\ \frac{d^2}{d|x|^2}f(|x|) &= (d+\kappa)\frac{1-|x|^2}{(1+|x|^2)^2}.\end{aligned}$$

With $\psi(r) = e^{br^\beta}$ for all $r \geq b^{-\frac{1}{\beta}}$, based on (3.22) and (3.23), we have that for all $|x| \gg b^{-\frac{1}{\beta}}$ and $k \in \mathbb{Z}^+$,

$$f'(\psi(|x|))\psi'(|x|)|x|^{-1} - b\beta d|x|^{\beta-2} + (d-\beta)|x|^{-2} = \kappa b\beta|x|^{\beta-2} + (d-\beta)|x|^{-2} + o(|x|^{-k}),$$

and

$$\begin{aligned}&f''(\psi(|x|))\psi'(|x|)^2 + f'(\psi(|x|))\psi''(|x|) - b\beta(\beta-1)|x|^{\beta-2} - (d-\beta)|x|^{-2}\\ &= \kappa b\beta(\beta-1)|x|^{\beta-2} - (d-\beta)|x|^{-2} + o(|x|^{-k}).\end{aligned}$$

Note that for all $|x| \geq b^{-\frac{1}{\beta}}$, we have

$$\begin{aligned}&\kappa b\beta|x|^{\beta-2} \leq \kappa\beta b^{\frac{2}{\beta}},\\ &\kappa b\beta(\beta-1)|x|^{\beta-2} \leq \kappa\beta(\beta-1)b^{\frac{2}{\beta}} \leq \kappa\beta b^{\frac{2}{\beta}}.\end{aligned}$$

The last inequality holds since $\beta \in (1, 2]$. Therefore f satisfies Assumption A4 with some $N_4 > 0$ and $L = 2\kappa\beta b^{\frac{2}{\beta}}$.

To check Assumption A1, notice that for all $|x| \gg b^{-\frac{1}{\beta}}$ and $k \in \mathbb{Z}^+$, we have

$$f'(\psi(|x|))\psi'(|x|)|x| - b\beta d|x|^\beta + (d-\beta) = \kappa b\beta|x|^\beta + (d-\beta) + o(|x|^{-k})$$

Therefore Assumption A1 is satisfied with $A = \kappa b\beta$, $\alpha = \beta$ and some $B \geq 0$,$N_1 > 0$.

Lastly, to check Assumption A2, similar to the calculation in checking Assumption A4, for all $|x| \gg b^{-\frac{1}{\beta}}$ and $k \in \mathbb{Z}^+$, we have

$$f'(\psi(|x|))\psi'(|x|)|x|^{-1} - b\beta d|x|^{\beta-2} + (d-\beta)|x|^{-2} = \kappa b\beta|x|^{\beta-2} + (d-\beta)|x|^{-2} + o(|x|^{-k}),$$

and

$$\begin{aligned} &f''(\psi(|x|))\psi'(|x|)^2 + f'(\psi(|x|))\psi''(|x|) - b\beta(\beta-1)|x|^{\beta-2} - (d-\beta)|x|^{-2} \\ &= \kappa b\beta(\beta-1)|x|^{\beta-2} - (d-\beta)|x|^{-2} + o(|x|^{-k}). \end{aligned}$$

Therefore Assumption A2 is satisfied with arbitrary $\mu \in (0, \kappa b\beta(\beta-1))$, $\theta = 2-\beta \geq 0$ and some $N_2 > 0$. ■

3.2.8.7. *Order estimation of mixing time when $\beta = 2$.* When $f(x) = \frac{d+\kappa}{2}\log(1+|x|^2)$, for all $|x| > b^{-\frac{1}{2}}$, the two eigenvalues of $\nabla^2 f_h(x)$ can be studied via (3.22) and (3.23). We obtain

$$\lambda_1 = 2b\kappa + (d-2)|x|^{-2} - 2b(d+\kappa)\frac{1}{1+e^{2b|x|^2}}, \tag{3.63}$$

$$\lambda_2 = 2b\kappa - (d-2)|x|^{-2} + 2b(d+\kappa)\frac{(4b|x|^2-1)e^{2b|x|^2}+1}{(1+e^{2b|x|^2})^2}. \tag{3.64}$$

Therefore, for all $|x| > b^{-\frac{1}{2}}$: we can estimate λ_1:

$$2b\kappa - 2b\kappa\frac{1}{1+e^2} - 2b < \lambda_1 < 2b\kappa + (d-2)b,$$

which can be simplified as

$$2b(\frac{e^2}{1+e^2}\kappa - 1) < \lambda_1 < 2b(\kappa + \frac{d}{2} - 1), \tag{3.65}$$

for all $|x| > b^{-\frac{1}{2}}$. Similarly, we can obtain the following estimate on λ_2:

$$2b\kappa - bd\frac{e^4-3e^2-1}{(1+e^2)^2} < \lambda_2 < 2b\kappa + 2b + 2b\kappa\frac{3e^2+1}{(1+e^2)^2}.$$

The above estimation can be further simplified as

$$2b(\kappa - 0.2d) < \lambda_2 < 2b(1.5\kappa + 1). \tag{3.66}$$

According to (3.65) and (3.66), we instantly have the locally Lipschitz constant, denoted as $L_{h,loc}$, for f_h in the region $\{|x| > b^{-\frac{1}{2}}\}$ being characterized as

$$L_{h,loc} = 2b \max\left\{\kappa + \frac{d}{2} - 1, 1.5\kappa + 1\right\}.$$

Next, for $|x| \leq b^{-\frac{1}{2}}$, we can check that for any fixed d, we have

$$\lim_{|x|\to 0} |\lambda_i(|x|)| < \infty \quad i = 1, 2$$

Therefore we can check that for any fixed $|x| \leq b^{-\frac{1}{2}}$, we have $|\lambda_i(x)| = O(d)$ for $i = 1, 2$ when $d \gg 1$. Thus we can conclude the global Lipschitz constant of f_h, $L_h = O(d)$ for $d \gg 1$.

On the other hand side, from (3.65), we can see for all $\kappa > \frac{1+e^2}{e^2}$, $\lambda_1 > b(\frac{e^2}{1+e^2}\kappa - 1)$. While from (3.66), the lower bound would be negative if $d \gg \kappa$. Therefore to ensure both eigenvalues are lower bounded by $b\kappa$, we need to restrict the region $\{|x| > b^{-\frac{1}{2}}\}$ to set of points with larger magnitudes. For all $|x| > (\frac{d}{b\kappa})^{\frac{1}{2}}$, we have when $d \geq \kappa$ and $d \geq 3$ that

$$\begin{aligned} \lambda_1 &> b\kappa\left(2 - \frac{2}{1 + e^{2d/\kappa}} - 2/d\right) > b\kappa, \\ \lambda_2 &> 2b\kappa - (d-2)(b\kappa/d) = b(\kappa + \kappa/d) > b\kappa. \end{aligned}$$

To determine the LSI constant, we first construct a function G_h such that $\nabla^2 G_h(x) \succeq b\kappa I_d$ for all $x \in \mathbb{R}^d$. Letting $\varpi := \sqrt{(d/b\kappa)}$, the function G_h is defined piecewise as

$$G_h = \begin{cases} f_h & |x| > \varpi \\ \frac{1}{3}A\left(|x| - \varpi\right)^3 + \frac{1}{2}f_h''\left(\varpi\right)\left(|x| - \varpi\right)^2 & \\ + f_h'(\varpi)(|x| - \varpi) + f_h(\varpi) & |x| \leq \varpi, \end{cases}$$

where

$$f_h(\varpi) = \frac{d}{2}\log(1 + e^{-2d/\kappa}) + \frac{\kappa}{2}\log(1 + e^{2d/\kappa}) + (d-2)\log(\varpi) - \log 2.$$

Note that we also have

$$\begin{aligned} f_h'(\varpi) &= 2b\kappa\varpi + (d-2)\left(\frac{d}{b\kappa}\right)^{-\frac{1}{2}} - 2bd(1+\kappa/d)\frac{\varpi}{1+e^{2d/\kappa}}, \\ f_h''(\varpi) &= b\kappa\left(1+\frac{2}{d}\right) + 2bd(1+\kappa/d)\frac{\left(4\frac{d}{\kappa}-1\right)e^{2d/\kappa}+1}{(1+e^{2d/\kappa})^2}, \\ A &= -\frac{b\kappa}{d}\left(-2\varpi - 4bd(1+\kappa/d)\varpi\frac{1-2\frac{d}{\kappa}e^{2d/\kappa}}{(1+e^{2d/\kappa})^2}\right) < 0. \end{aligned}$$

With the above coefficients, we can check $G_h \in \mathcal{C}^2(\mathbb{R}^d)$ and $\nabla G_h(x) \succeq b\kappa I_d$ for all $x \in \mathbb{R}^d$. We now consider different cases.

(1) When $d \gg \kappa$ for all $k \in \mathbb{Z}$:

$$\begin{aligned} f_h(\varpi) &= d + \frac{1}{2}(d-1)\log d + O(1), \\ f'(\varpi) &= 3d\left(\frac{d}{b\kappa}\right)^{-\frac{1}{2}} - 2\left(\frac{d}{b\kappa}\right)^{-\frac{1}{2}} + o(d^{-k}), \\ f''(\varpi) &= b\kappa + 2\left(\frac{d}{b\kappa}\right)^{-1} + o(d^{-k}), \\ A &= -2d\left(\frac{d}{b\kappa}\right)^{-\frac{3}{2}} + 4\left(\frac{d}{b\kappa}\right)^{-\frac{3}{2}} + o(d^{-k}). \end{aligned}$$

Therefore the oscillation between f_h and G_h can be written as

$$Osc(f_h - G_h) = \max_{0\le|x|\le\varpi} |f_h(x) - G_h(x)|.$$

Since both G_h and f_h are monotone increasing with respect to $|x|$, we then have

$$Osc(f_h - G_h) \le G_h(\varpi) + f_h(\varpi) = 2d + (d-1)\log d + O(1)$$

On the other hand,

$$Osc(f_h - G_h) \ge G_h(0) - f_h(0) = \frac{1}{2}(d-1)\log d - \frac{5}{6}d + O(1)$$

Hence, apply Holley-Strook lemma, we can calculate the LSI constant $C_{h,\mathrm{LSI}}$ as

$$C_{h,\mathrm{LSI}} \le 2(b\kappa)^{-1}\exp(Osc(f_h - G_h)) \le C(b\kappa)^{-1}d^{d-1}\exp(2d).$$

Furthermore, because of the lower bound on $Osc(f_h, G_h)$, the factor d^{d-1} can be improved. Hence, according to Theorem 14, to reach ϵ-accuracy in KL-divergence, the mixing time n satisfies:

$$n \sim \tilde{O}(L_h C_h d\epsilon^{-1}) \leq \tilde{O}(\exp(2d) d^{d+1}\epsilon^{-1}).$$

(2) When $d/\kappa = O(1)$, or equivalently when $d/\kappa \to C'$, we have

$$\begin{aligned}
f_h(\varpi) &= \frac{d}{2}[\log(1+e^{-2C'}) + C'^{-1}\log(1+e^{2C'}) + \log(\frac{C'}{b})] + O(1), \\
&:= dC'_1 + O(1), \\
f'(\varpi) &= 3d(\frac{C'}{b})^{-\frac{1}{2}} - 2(\frac{C'}{b})^{-\frac{1}{2}} + o(d^{-k}) \\
&:= b^{\frac{1}{2}} dC'_2 + O(1) \\
f''_h(\varpi) &= bdC'^{-1} + 2(\frac{C'}{b})^{-1} + o(d^{-k}) \\
&:= bdC'_3 + O(1), \\
A &= -2d(\frac{C'}{b})^{-\frac{3}{2}} + 4(\frac{C'}{b})^{-\frac{3}{2}} + o(d^{-k}) \\
&:= b^{\frac{3}{2}} C'_4 d + O(1).
\end{aligned}$$

Therefore for all $|x| \leq (C'/b)^{\frac{1}{2}}$, we have

$$\begin{aligned}
G_h(x) = d\Bigg\{&\frac{1}{3}b^{\frac{3}{2}}C'_4|x|^3 + b(\frac{1}{2}C'_3 - C'^{\frac{1}{2}}C'_4)|x|^2 + b^{\frac{1}{2}}(C'_2 - C'^{\frac{1}{2}}C'_3 + C'C'_4)|x| \\
&+(C'_1 - C'^{\frac{1}{2}}C'_2 + \frac{1}{2}C'C'_3 - \frac{1}{3}C'^{\frac{3}{2}}C'_4)\Bigg\} + O(1)
\end{aligned}$$

Similar to the previous argument, the oscillation can be upper bounded as

$$\begin{aligned}
Osc(f_h - G_h) &= G_h((\frac{C'}{b})^{\frac{1}{2}}) + f_h((\frac{C'}{b})^{\frac{1}{2}}) \\
&= C'_h d + O(1),
\end{aligned}$$

where

$$
\begin{aligned}
C'_h = &\frac{1}{3}C'_4 C'^{\frac{3}{2}} + (\frac{1}{2}C'_3 - C'^{\frac{1}{2}}C'_4)C' + (C'_2 - C'^{\frac{1}{2}}C'_3 + C'C'_4)C'^{\frac{1}{2}} \\
&+ (2C'_1 - C'^{\frac{1}{2}}C'_2 + \frac{1}{2}C'C'_3 - \frac{1}{3}C'^{\frac{3}{2}}C'_4).
\end{aligned}
$$

Hence, applying Holley-Strook Theorem, the LSI constant can be bounded by

$$C_{h,\mathrm{LSI}} \leq 2(b\kappa)^{-1}\exp(Osc(f_h - G_h)) \leq C(bd/C')^{-1}(\exp(C'_h))^d.$$

Hence, according to [VW19], to reach ϵ-accuracy in KL-divergence, the mixing time n satisfies

$$n \sim \tilde{O}(L_h C_{h,\mathrm{LSI}} d\epsilon^{-1}) \leq \tilde{O}((\exp(C'_h))^d d^{-1}\epsilon^{-1}).$$

3.2.9. A Summary of Constants. For the sake of convenience, we provide a list of constants in Table 3.1.

Constant	Description	Equation
ϵ	Accuracy parameter	NA
γ	Step-size parameter	(3.10)
C_{P}	Poincaré constant	(PI)
C_{LSI}	LSI constant	(LSI)
$C_{\text{M-LSI}}$, δ	m-LSI related constants	(m-LSI)
$C_{h,\mathrm{P}}$	Poincaré constant after Transformation	NA
$C_{h,\mathrm{LSI}}$	LSI constant after Transformation	NA
$C_{h,\text{M-LSI}}$	m-LSI constant after Transformation	NA
r, b, β	Parameters related to transformation map	(3.6)
A, B, N_1, α	Parameters related to dissipativity	Assumption A1
μ, N_2, θ	Parameters related to degenerate convexity	Assumption A2
N_3, ρ	Parameters related to convexity	Assumption A3
N_4, L	Parameters related to Lipschitz-gradients	Assumption A4
$N_5, m, \alpha_1, C^*_{\mathrm{TAIL}}$	Parameters related to tail condition	Assumption A5
α_h, A_h, B_h	Dissipativity parameters after transformation	Assumption B1
ξ_h, μ_h, θ_h	Degenerate Convexity at infinity after transformation	Assumption B2
ρ_h	Strong-convexity parameter after transformation	NA
L_h	Lipschitz-gradient parameters after transformation	NA
$m_h, \alpha_{h,1}, C_{h,\mathrm{TAIL}}$	Tail condition parameters after transformation	Assumption B5
κ	Degrees-of-freedom of t distribution	NA
ϑ	Parameter related to super and weak Poincaré inequalities	NA

Table 3.1. A list of all the constants used and their description.

3.3. Itô Discretization

3.3.1. Motivations. Indeed, Theorem 2.4 by [RT96] shows that if $|\nabla \log \pi(x)| \to 0$ as $|x| \to \infty$, then the solution to (3.1) is *not* exponentially ergodic. In the other direction, standard results in the literature, for example [Wan06, BGL14] show that the solution to (3.1) 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 to have exponentially decaying tails. Furthermore, when π has polynomially decaying tails, the convergence is only sub-exponential or polynomial [Wan06, Chapter 4]. Consequently, the algorithms obtained as discretizations of the Langevin diffusion in (3.1) 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 [BBD$^+$09, BL09]. 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.

Theorem 16 (Weighted Poincaré Inequality; Theorem 2.3 in [BL09]). 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 $G = Vg$,

$$(\beta+1)\mathrm{Var}_{\pi_\beta}(g) \leq \int_{\mathbb{R}^d} \frac{\langle (\nabla^2 V)^{-1}\nabla G, \nabla G\rangle}{V} d\pi_\beta + \frac{d}{\beta - d}\left(\int_{\mathbb{R}^d} g d\pi_\beta\right). \tag{3.67}$$

A canonical example of a heavy-tailed density that satisfies the conditions in Theorem 16, and hence (3.67), is the multivariate t-distribution. In particular, we consider the following Itô diffusion process

$$dX_t = -(\beta - 1)\nabla V(X_t)dt + \sqrt{2V(X_t)}dB_t, \tag{3.68}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. The Itô diffusion in (3.68) 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 10 for details. Hence, we study the oracle complexity of the Euler-Maruyama discretization of (3.68), 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, [MT04] for an overview. Our results in this section 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.

In this section, we make the following contributions:

- In Theorem 17, we provide upper bounds on the number of iterations required by the Euler-Maruyama discretization of (3.68) 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 17 assumes access to the exact gradient of the unnormalized target density function (referred to as the first-order setting), in Theorem 18, 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 3.3.7 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.

As mentioned previously, our approach leverages the literature on weighted functional inequalities, that are satisfied by heavy-tailed densities. The weighted Poincare inequality was introduced in [BBD+09] and [BL09], 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 [CGGR10, CGW11, BJM16, CEG17, CGMZ19] for various extensions and improvements of the works of [BBD+09] and [BL09].

3.3.2. Notations. We use the following notation throughout the rest of the section.

- $\langle \cdot, \cdot \rangle$ 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. $\mathcal{C}^2_c(\mathbb{R}^d)$ 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 on $\mathbb{R}^d$, μ and ν 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:

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

- 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.

3.3.3. Organizations. In Section 3.3.4, we first establish the exponential ergodicity of the Itô diffusion in (3.68) 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 (3.68). In Section 3.3.5, we provide moment computations in the heavy-tailed setting that are required to obtain explicit rates from the results in Section 3.3.4. In Section 3.3.6, we provide an extension of our results to the zeroth-order setting. In Section 3.3.7 we provide several illustrative examples. We discuss further implications of our assumptions in Section 3.3.8. The proofs are provided in Section 3.3.9 and in Appendices 3.3.10.1, 3.3.10.2 and 3.3.10.3.

3.3.4. Itô Discretizations and Weighted Poincare Inequalities. In this section, our goal is to analyze the Itô diffusion in (3.68) 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 (3.68) is given by

$$\mathcal{L}\psi = -(\beta - 1)\langle \nabla V, \nabla \psi \rangle + V \Delta \psi. \tag{3.69}$$

Hence, the Fokker-Planck equation corresponding to (3.68) is

$$\partial_t \rho_t = \nabla \cdot (\beta \rho_t \nabla V + V \nabla \rho_t) = \nabla \cdot \left(\rho_t V \nabla \log \frac{\rho_t}{\pi_\beta} \right). \tag{3.70}$$

It follows that, under the conditions in Theorem 16, $\pi_\beta \propto V^{-\beta}$ is the unique stationary density of (3.68). We next examine the convergence properties of (3.68) to its stationary density. To do so, we introduce the following assumption.

Assumption 11. 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 11 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.

We next provide the following corollary to Theorem 16, motivated by the discussion in Section 2 of [BL09].

Corollary 3. Consider the setting of Theorem 16 and suppose further that Assumption 11 holds with $C_V \in (0, \beta+1)$, then for any smooth, π_β-integrable function, ϕ on $\mathbb{R}^d$,

$$Var_{\pi_\beta}(\phi) \leq \left(\sqrt{\beta+1} - \sqrt{C_V}\right)^{-2}\int_{\mathbb{R}^d}\langle V(x)(\nabla^2 V)^{-1}(x)\nabla\phi(x), \nabla\phi(x)\rangle\pi_\beta(x)dx. \tag{3.71}$$

Proof. We start from (3.67), assume that $\int_{\mathbb{R}^d} g d\pi_\beta = 0$. Then (3.67) could be rewritten as

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

Now, note that we have the following elementary bound

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

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

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

Invoking the condition in Assumption 11, we further obtain

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

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 $r := \sqrt{\frac{\beta+1}{C_V}} > 1$, 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 11 holds under the following sufficient condition.

Assumption 12. 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 12.

Lemma 3.3.1. *Let $\beta > d$. If Assumption 12 holds with $C_V\in(0,\beta+1)$, then for any smooth, π_β integrable function ϕ on $\mathbb{R}^d$, we have*

$$Var_{\pi_\beta}(\phi) \le \alpha^{-1}\left(\sqrt{\beta+1}-\sqrt{C_V}\right)^{-2}\int_{\mathbb{R}^d} V(x)|\nabla\phi(x)|^2\pi_\beta(x)dx. \tag{3.72}$$

With (3.72), we can show the exponential decay in χ^2-divergence along (3.68). The proof of the following proposition is standard and we include it here for completeness.

PROPOSITION 10. Under the conditions in Lemma 3.3.1, for (X_t) following diffusion (3.68) with ρ_t being the distribution of X_t, we have

$$\chi^2(\rho_t|\pi_\beta) \le \exp\left(-2\alpha\left(\sqrt{\beta+1}-\sqrt{C_V}\right)^2 t\right)\chi^2(\rho_0|\pi_\beta). \tag{3.73}$$

PROOF OF PROPOSITION 3.3.1. First we can calculate the derivative of $\chi^2(\rho_t|\pi)$ via (3.70),

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

According to (3.72), we get

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

Finally, (3.73) follows from Gronwall's inequality. ■

The above result shows that for the class of π_β satisfying Assumption 12, the Itô diffusion in (3.68), converges exponentially fast to its stationary density. Hence, time-discretizations of (3.68) provide a practical way of sampling from that class of densities. The Euler-Maruyama discretization to (3.68) is given by

$$x_{k+1} = x_k - h(\beta-1)\nabla V(x_k) + \sqrt{2hV(x_k)}\xi_{k+1}, \tag{3.74}$$

where $h>0$ is the step size and $\{\xi\}_{k=1}^{\infty}$ 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 (3.74) 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 10 requires that condition that $\beta > d$, in addition to Assumption 12, whereas Theorem 17 below, does not. In Section 3.3.8, we revisit these conditions and provide additional insights. Obtaining convergence results in the stronger χ^2-divergence is left as future work.

Theorem 17. Let V be gradient-Lipschitz with parameter $L > 0$, and satisfying Assumption 12 with

$$\delta := \frac{\beta - 1 - \frac{1}{4}C_V d}{\frac{1}{4}C_V d} > 0. \tag{3.75}$$

Let $(x_k)_{k=0}^{\infty}$ be generated from (3.74) 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 $k \geq 1$,

$$W_2(\nu_k, \pi_\beta) \leq (1-A)^k W_2(\nu_0, \pi_\beta) + \frac{C}{A} + \frac{B}{\sqrt{A(2-A)}}. \tag{3.76}$$

with A, B and C given respectively in (3.114), (3.115) and (3.116).

Remark 17 (Constant δ). We now motivate the definition and the condition on the constant δ based on exponential contractivity arguments.

Definition 6 (Exponential contractivity). 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 [GDVM19, Theorem 10]. The uniform dissipativity condition for (3.68) can be represented as

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

or equivalently as

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

When V satisfies Assumption 12, a sufficient condition for the above uniform dissipativity condition is given by

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

or equivalently,

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

The sufficient condition coincides with the condition that $\delta > 0$ in Theorem 17, which also motivates the assumption in Theorem 17.

Remark 18 (Iteration complexity). With Theorem 17, we can calculate the order of the iteration complexity to reach an ϵ-accuracy in Wasserstein-2 distance. With (3.114),(3.115),(3.116), we have

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

The above display implies that

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

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

$$h < \min\left\{ \frac{\delta^2 \mathbb{E}_{\pi_\beta}\left[V(X)\right]^{-1}\epsilon^2}{81d(\delta+3)^2(1+\frac{L}{\alpha})^2}, \frac{\delta \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{-\frac{1}{2}}\epsilon}{81(\beta-1)(\delta+3)(1+\frac{L}{\alpha})} \right\}. \tag{3.77}$$

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

$$\begin{aligned} K &= \frac{3(1+\delta)}{\alpha(\beta-1)\delta h^*}K_\epsilon \\ &\leq 273 \max\left\{ \frac{(\delta+3)^3(1+\frac{L}{\alpha})^2 d\mathbb{E}_{\pi_\beta}\left[V(X)\right]}{\alpha\delta^3(\beta-1)\epsilon^2}, \frac{(\delta+3)^2(1+\frac{L}{\alpha})\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}}{\alpha\delta^2\epsilon} \right\} K_\epsilon. \end{aligned} \tag{3.78}$$

Recall the definition of δ in (3.75). 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 18, it is required to compute bounds on the following two quantities: $\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]$ and $\mathbb{E}_{\pi_\beta}\left[V(X)\right]$.

3.3.5. Moment Bounds. In this section, we compute moment bounds under the conditions in Theorem 17.

3.3.5.1. *An Example: Multivariate t-distribution.* We first start with the isotropic case.

PROPOSITION 11. 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

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] = \frac{\beta-1}{\beta-1-\frac{d}{2}} \quad \text{and} \quad \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] = \frac{2d}{\beta-1-\frac{d}{2}}. \tag{3.79}$$

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 $\beta > \frac{d}{2}$,

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

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

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

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

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

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

REMARK 19. If π_β is the class of isotropic multivariate t-distributions, with the results in Proposition 11, 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 [Rot12], 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

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

The above identity implies

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

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

Remark 20. 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

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

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

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

3.3.5.2. *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:

$$\exists\, \varepsilon, R > 0 \text{ such that } \Delta V(x) - (\beta - 1)\frac{|\nabla V(x)|^2}{V(x)} \leq -\varepsilon \qquad \forall\ |x| \geq R. \tag{3.80}$$

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 (3.80).

Proposition 12. If $V \in \mathcal{C}^2(\mathbb{R}^d)$ is positive, L-gradient Lipschitz and satisfies (3.80), then we have

$$\mathbb{E}_{\pi_\beta}\left[V(x)\right] \leq (dL + \varepsilon) \max_{|x| \leq R} V(x), \quad \text{and} \quad \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] \leq \frac{dL\,(dL + \varepsilon)}{(\beta - 1)} \max_{|x| \leq R} V(X). \tag{3.81}$$

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 (3.68) 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 (3.68), then

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

With (3.69), we have

$$\begin{aligned}\mathcal{L}V(x) &= V(x)\left[\Delta V(x) - (\beta-1)\frac{|\nabla V(x)|^2}{V(x)}\right]\\ &\leq V(x)\left(-\varepsilon 1_{|x|\geq R} + dL1_{|x|<R}\right)\\ &\leq -\varepsilon V(x) + (dL+\varepsilon)\max_{|x|\leq R} V(x),\end{aligned}$$

where the first inequality follows from (3.80) and the fact that $\Delta V \leq d\left\|\nabla^2 V\right\|_2$. Therefore we obtain

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

and it follows from Gronwall's inequality that

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

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

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

■

3.3.5.3. *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 13. 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 $r \in (0, \beta - \frac{d}{2} - 1)$,

$$\mathbb{E}_{\pi_\beta}\left[V(X)\right] \leq \left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta - \frac{d}{2} - r}} V(0) \left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})}\right)^{\frac{1}{\beta-\frac{d}{2}-r}}, \tag{3.82}$$

$$\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] \leq \frac{dL}{\beta-1} \left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta - \frac{d}{2} - r}} V(0) \left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})}\right)^{\frac{1}{\beta-\frac{d}{2}-r}}. \tag{3.83}$$

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 3.3.3. 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

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

where the last inequality also follows from Lemma 3.3.3. ■

Remark 21. A ratio between Gamma functions appears in (3.82) and (3.83). The ratio can be written explicitly via properties of Gamma functions.

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

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

- 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 [IM94].

Remark 22. With Theorem 13 and the upper bounds in Remark 21, 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 $r \in (0, \beta - \frac{d}{2} - 1)$,

$$
\mathbb{E}_{\pi_\beta}\left[V(X)\right] \leq V(0)\left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}}\left(\frac{1+r}{r}\right)^{\frac{1}{2(\beta-\frac{d}{2}-r)}}\left(\frac{\beta-\frac{d}{2}}{r}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}}, \tag{3.84}
$$

$$
\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right] \leq \frac{V(0)dL}{\beta-1}\left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}}\left(\frac{1+r}{r}\right)^{\frac{1}{2(\beta-\frac{d}{2}-r)}}\left(\frac{\beta-\frac{d}{2}}{r}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}}. \tag{3.85}
$$

3.3.6. 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 [DCWY19, LST21, RSBG22]. 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 [NS17, RSBG22], for any $x \in \mathbb{R}^d$, we define the zeroth order gradient estimator $g_{\sigma,,m}(x)$ as

$$
g_{\sigma,m}(x) := \frac{1}{m}\sum_{i=1}^{m}\frac{V(x+\sigma u_i)-V(x)}{\sigma}u_i \tag{3.86}
$$

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

$$x_{k+1} = x_k - h(\beta - 1) g_{\sigma,m}(x_k) + \sqrt{2V(x_k)}\xi_{k+1} \tag{3.87}$$

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 [BG22] and [RSBG22], we recall the following property of $g_{\sigma,m}$.

Proposition 14. [RSBG22, 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 (3.86) and $\{x_k\}_{k=0}^{\infty}$ generated by (3.87). We have for any $k \geq 0$,

$$\mathbb{E}\left[\left|\mathbb{E}\left[\zeta_k | x_k\right]\right|^2\right] \leq L^2 \sigma^2 d, \tag{3.88}$$

and

$$\mathbb{E}\left[\left|\zeta_k - \mathbb{E}\left[\zeta_k | x_k\right]\right|^2\right] \leq \frac{\sigma^2}{2m} L^2 (d+3)^3 + \frac{2(d+5)}{m} \mathbb{E}\left[|\nabla V(x_k)|^2\right]. \tag{3.89}$$

Theorem 18. Suppose V is gradient-Lipschitz with parameter $L > 0$ and satisfies Assumption 12 with δ in (3.75). Let $g_{\sigma,m}$ be as defined in (3.86) and $(x_k)_{k=0}^{\infty}$ be generated from (3.87) with $x_k \sim \nu_k$ for all $k \geq 0$. Then with the time step size

$$h < \min\left\{\frac{2\delta}{3(1+\delta)\alpha(\beta-1)}, \frac{\alpha m \delta}{24(1+\delta)(\beta-1)(d+5)L^2}, \frac{1}{4(\beta-1)L}\right\}, \tag{3.90}$$

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

$$W_2(\nu_k, \pi_\beta) \leq (1 - A')^k W_2(\nu_0, \pi_\beta) + \frac{C'}{A'} + \frac{B'}{\sqrt{A'(2-A')}}. \tag{3.91}$$

with A', B' and C' given respectively in (3.123), (3.124) and (3.125).

Remark 23. With Theorem 18, 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 (3.90), we have

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

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

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

$$K = \tilde{O}\left(\max\left\{\frac{1}{\varepsilon^2\delta^3}\mathbb{E}_{\pi_\beta}\left[V(X)\right], \frac{1}{\varepsilon\delta^2}\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}, \frac{d}{\varepsilon^2\delta^2 m}\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]\right\}\right). \tag{3.92}$$

The number of function evaluations is hence mK.

3.3.7. Illustrative Examples. We now provide illustrative examples to highlight the implications of our results.

3.3.7.1. *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 12 with $\alpha = 2$, $C_V = 2$, and V is L-Lipschitz gradient with $L = 2$. The constant δ in Theorem 17 becomes $\delta = 1$. Furthermore, according to proposition 11, $\mathbb{E}_{\pi_\beta}[V(X)] = 2$ and $\mathbb{E}_{\pi_\beta}[|\nabla V(X)|^2] = 4$.

first order algorithm: According to Theorem 17 and (3.78), 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 [LWME19]. The diffusion (3.68) is α'-uniformly dissipative with $\alpha' = d$ and the Euler discretization given in (3.74) 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 3.3.10.2. Hence, by Theorem 1 in [LWME19], 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 [LWME19], we obtain a dimension-free iteration complexity to ensure an ϵ-accuracy in Wasserstein-2 distance.

zeroth order algorithm: According to Theorem 18 and (3.92), 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.

3.3.7.2. *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 12 with $\alpha = 2$, $C_V = 2$ and V is L-Lipschitz gradient with $L = 2$. Hence, the constant δ in Theorem 17 is given by $\delta = 1/d$. According to Proposition 11, $\mathbb{E}_{\pi_\beta}[V(X)] = d+1$ and $\mathbb{E}_{\pi_\beta}[|\nabla V(X)|^2] = 4d$.

first order algorithm: According to Theorem 17 and (3.78), 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 [LWME19]. The diffusion (3.68) is α'-uniformly dissipative with $\alpha' = 1$ and the Euler discretization given in (3.74) 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 3.3.10.2. Hence, according to Theorem 1 in [LWME19], 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 [LWME19], our analysis helps to decrease the dimension exponent by a factor of 2.

zeroth order algorithm: According to Theorem 18 and (3.92), 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 24. The example discussed above highlights the following important observation: Choosing a large batch size does not improve the iteration complexity. To explain this, we understand both (3.74) and (3.87) as approximation to the continuous dynamics (3.68). 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.

3.3.8. Further Results and Additional Insights on Assumptions.

In Section 3.3.4, 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 3.3.4 by introducing the following assumptions.

Assumption 13. 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

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

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

Lemma 3.3.2. *Under Assumption 13, for any smooth function $\phi \in L^2(\pi_\beta)$,*

$$\mathit{Var}_{\pi_\beta}(\phi) \leq C_{WPI} \int_{\mathbb{R}^d} |\nabla\phi(x)|^2 V(x)\pi_\beta(x)dx, \quad \mathit{with} \quad C_{WPI} = C_V(\gamma)\left(\frac{\beta}{\gamma} - 1\right)^{-1}. \tag{3.93}$$

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

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

where the inequality follows from the fact that $\gamma \in \left(0, \frac{\beta}{d+2}\right]$. Therefore we can apply Theorem 16 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 $G = V_\gamma g$,

$$(a+1)\int_{\mathbb{R}^d} g(x)^2 \pi_{\beta'}(x)dx \leq \int_{\mathbb{R}^d} \frac{\langle(\nabla^2 V_\gamma)^{-1}(x)\nabla G(x), \nabla G(x)\rangle}{V_\gamma(x)} \pi_{\beta'}(x)dx. \tag{3.94}$$

Since $\beta' = \beta - 2\gamma$, (3.94) is equivalent to

$$(a+1)\int_{\mathbb{R}^d} \frac{|G(x)|^2}{V(x)} V(x)^{-(\beta-1)}dx \leq \int_{\mathbb{R}^d} \langle(\nabla^2 V_\gamma)^{-1}(x)\nabla G(x), \nabla G(x)\rangle V(x)^{-(\beta'+\gamma)}dx. \tag{3.95}$$

Under Assumption 13, we have

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

where the last identity follows from the fact that $\beta' = \beta - 2\gamma$. Along with (3.95), we get

$$(a+1)\int_{\mathbb{R}^d} \frac{|G(x)|^2}{V(x)} V(x)^{-(\beta-1)}dx \le C_V(\gamma)\int_{\mathbb{R}^d} |\nabla G(x)|^2 V(x)^{-(\beta-1)}dx. \tag{3.96}$$

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

$$\int_{\mathbb{R}^d} |\phi(x) - \mathbb{E}_{\pi_{\beta-\gamma}}[\phi(X)]|^2 \pi_\beta(x)dx \le \frac{C_V(\gamma)}{a+1}\int_{\mathbb{R}^d} |\nabla\phi(x)|^2 V(x)\pi_\beta(x)dx. \tag{3.97}$$

Therefore for any smooth, π_β-square integrable ϕ,

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

which is equivalent to (3.93) with $C_{\mathrm{WPI}} = \frac{C_V(\gamma)}{a+1} = C_V(\gamma)\left(\frac{\beta}{\gamma} - 1\right)^{-1}$. ■

Remark 25. Lemma 3.3.2 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 13 holds with

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

Hence, Lemma 3.3.2 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

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

The detailed calculation for deriving the above mentioned constants is provided in Appendix 3.3.10.3.

As an immediate consequence of Lemma 3.3.2, we have the following χ^2 convergence result for (3.68).

Proposition 15. Under Assumption 13, with (X_t) satisfying (3.68) with ρ_t being the distribution of X_t, we have

$$\chi^2(\rho_t|\pi_\beta) \le \exp\left(-C_V(\gamma)^{-1}\left(\frac{\beta}{\gamma} - 1\right)t\right)\chi^2(\rho_0|\pi_\beta). \tag{3.98}$$

For the case of multivariate t-distributions, Proposition 15 allows us to show exponential convergence of (3.68) in the χ^2 divergence with smaller degrees of freedom (and hence heavier tails) compared to Proposition 10.

3.3.8.1. *Relationship between Lemma 3.3.1 and Lemma 3.3.2.* The result in Lemma 3.3.2 complements that in Lemma 3.3.1. 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 12 with $C_V \in (0, \frac{d+2}{d+2-\beta})$, then V satisfies Assumption 13. Therefore π_β satisfies the WPI. In Proposition 16, this relation is proved formally.

Proposition 16. When $\beta \leq d$, if Assumption 12 holds with $C_V \in (0, \frac{d+2}{d+2-\beta})$, then Assumption 13 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 3.3.10.3, 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 12. 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 $x \in \mathbb{R}^d$. ■

3.3.8.2. *Relationship between Theorem 17 and Proposition 15.* Proposition 15 studies the convergence of the continuous dynamics (3.68) while Theorem 17 studies the convergence of the discretization (3.74). The conditions in Theorem 17 can be shown to imply conditions in proposition 15. In Proposition 15 we only assume Assumption 13. In Theorem 17, we assume (i) Assumption 12, (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 13.

Proposition 17. If Assumption 12 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 13 holds.

Proof of Proposition 15. Under Assumption 12 and L-gradient Lipschitzness assumption, we have that V is 'essential quadratic'. That is, assuming V attains its global minimum at x^*, for all $x \in \mathbb{R}^d$,

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

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

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

which implies that Assumption 12-(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 $d \ge 1$, 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 13-(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 13-(1) is automatically satisfied under Assumption 12. ∎

3.3.9. Proofs of the Main Results.

3.3.9.1. *Proofs of Theorem 17 and Theorem 18.* In this section, we provide the proof of Theorem 17 and Theorem 18 via mean square analysis. We first start with the following intermediate result.

PROPOSITION 18. Let $(X_t)_{t\geq 0}$ follow (3.68) with $X_t \sim \rho_t$ for all $t \geq 0$. If V is gradient Lipschitz with parameter L, then we have

$$\begin{aligned} \mathbb{E}\left[|X_t - X_0|^2\right] &\leq 4\left[(\beta-1)^2 t^2 \mathbb{E}\left[|\nabla V(X_0)|^2\right] + td\mathbb{E}\left[V(X_0)\right]\right] \\ &\qquad \exp\left(4(\beta-1)^2L^2t^2 + d(\beta-1)L^2t^2 + 2dLt\right). \end{aligned} \tag{3.99}$$

PROOF OF PROPOSITION 18. According to (3.68),

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

where

$$\begin{aligned}\mathbb{E}\left[\left|\int_0^t \nabla V(X_s)ds\right|^2\right] &\leq 2\mathbb{E}\left[\left(\int_0^t |\nabla V(X_s)-\nabla V(X_0)|ds\right)^2\right] + 2\mathbb{E}\left[\left(\int_0^t |\nabla V(X_0)|ds\right)^2\right] \\ &\leq 2t\mathbb{E}\left[\int_0^t |\nabla V(X_s)-\nabla V(X_0)|^2 ds\right] + 2t\mathbb{E}\left[\int_0^t |\nabla V(X_0)|^2 ds\right] \\ &\leq 2L^2 t\int_0^t \mathbb{E}\left[|X_s-X_0|^2\right]ds + 2t^2\mathbb{E}\left[|\nabla V(X_0)|^2\right], \end{aligned} \tag{3.100}$$

and

$$\begin{aligned}&\mathbb{E}\left[\int_0^t V(X_s)ds\right] \\ &\leq \mathbb{E}\left[\int_0^t V(X_0) + \langle \nabla V(X_0), X_s - X_0\rangle + \frac{L}{2}|X_s-X_0|^2 ds\right] \\ &= t\mathbb{E}\left[V(X_0)\right] + \frac{L}{2}\mathbb{E}\left[\int_0^t |X_s-X_0|^2 ds\right] - (\beta-1)\mathbb{E}\left[\int_0^t\int_0^s \langle \nabla V(X_0), \nabla V(X_u)\rangle duds\right] \\ &\leq t\mathbb{E}\left[V(X_0)\right] + \frac{L}{2}\mathbb{E}\left[\int_0^t |X_s-X_0|^2 ds\right] - \frac{(\beta-1)t^2}{2}\mathbb{E}\left[|\nabla V(X_0)|^2\right] \\ &\quad - (\beta-1)\mathbb{E}\left[\int_0^t\int_0^s \langle \nabla V(X_0), \nabla V(X_u)-\nabla V(X_0)\rangle duds\right] \\ &\leq t\mathbb{E}\left[V(X_0)\right] + \frac{L}{2}\mathbb{E}\left[\int_0^t |X_s-X_0|^2 ds\right] - \frac{(\beta-1)t^2}{2}\mathbb{E}\left[|\nabla V(X_0)|^2\right] \\ &\quad + \frac{(\beta-1)t^2}{2}\mathbb{E}\left[|\nabla V(X_0)|^2\right] + \frac{\beta-1}{4}\mathbb{E}\left[\int_0^t\int_0^s |\nabla V(X_u)-\nabla V(X_0)|^2 duds\right] \\ &\leq t\mathbb{E}\left[V(X_0)\right] + \frac{L}{2}\mathbb{E}\left[\int_0^t |X_s-X_0|^2 ds\right] + \frac{(\beta-1)L^2}{4}\mathbb{E}\left[\int_0^t\int_0^s |X_u-X_0|^2 duds\right] \\ &\leq t\mathbb{E}\left[V(X_0)\right] + \left(\frac{L}{2}+\frac{(\beta-1)L^2 t}{4}\right)\mathbb{E}\left[\int_0^t |X_s-X_0|^2 ds\right]. \end{aligned} \tag{3.101}$$

With (3.100) and (3.101), we get

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

By Gronwall's inequality, we hence have

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

■

Based on the above proposition, we now prove Theorem 17 below.

Proof of theorem 17. We perform mean square analysis to (3.74). Let $(X_t)_{t\geq 0}$ follow (3.68) with $X_0 \sim \pi_\beta$. Since π_β is the unique stationary distribution to (3.68), $X_t \sim \pi_\beta$ for all $t \geq 0$. With (3.74), we can calculate the difference between X_h and x_1,

$$\begin{aligned} X_h - x_1 &= X_0 - \int_0^h (\beta-1)\nabla V(X_t)dt + \int_0^t \sqrt{2V(X_t)}dB_t - \left(x_0 - (\beta-1)hy_0 + \sqrt{2hV(x_0)}\xi_1\right) \\ &= (X_0 - x_0) - (\beta-1)h\left(\nabla V(X_0) - \nabla V(x_0)\right) - \int_0^h (\beta-1)\left(\nabla V(X_t) - \nabla V(X_0)\right)dt \\ &\int_0^h \left(\sqrt{2V(X_t)} - \sqrt{2V(x_0)}\right)dB_t \\ &:= U_1 + U_2 + U_3, \end{aligned}$$

where

$$U_1 := (X_0 - x_0) - (\beta-1)h\left(\nabla V(X_0) - \nabla V(x_0)\right), \tag{3.102}$$

$$U_2 := -\int_0^h (\beta-1)\left(\nabla V(X_t) - \nabla V(X_0)\right)dt, \tag{3.103}$$

$$U_3 := \int_0^h \left(\sqrt{2V(X_t)} - \sqrt{2V(x_0)}\right)dB_t. \tag{3.104}$$

Therefore according to triangle inequality,

$$\mathbb{E}[|X_h - x_1|^2|]^{\frac{1}{2}} \leq \mathbb{E}[|U_1 + U_3|^2]^{\frac{1}{2}} + \mathbb{E}[|U_2|^2]^{\frac{1}{2}}.$$

Since U_1 is adapted to $\mathcal{F}_0$ and $\mathbb{E}[U_3|\mathcal{F}_0] = 0$, we get

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

Since V is α-strongly convex and L-gradient Lipschitz, it satisfies

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

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

$$\begin{aligned}
&|(X_0 - x_0) - (\beta - 1)h\left(\nabla V(X_0) - \nabla V(x_0)\right)|^2 \\
=&|X_0 - x_0|^2 - 2(\beta - 1)h\langle X_0 - x_0, \nabla V(X_0) - \nabla V(x_0)\rangle + (\beta - 1)^2h^2|\nabla V(X_0) - \nabla V(x_0)|^2 \\
\leq&\left(1 - \frac{2(\beta - 1)\alpha L h}{\alpha + L}\right)|X_0 - x_0|^2 + (\beta - 1)h\left((\beta - 1)h - \frac{2}{\alpha + L}\right)|\nabla V(X_0) - \nabla V(x_0)|^2 \\
\leq&\left(1 - (\beta - 1)\alpha h\right)^2|X_0 - x_0|^2. \qquad (3.105)
\end{aligned}$$

Meanwhile, for arbitrary $r > 0$, we have

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

Notice that under Assumption 12, 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

$$(\sqrt{2V(X_0)} - \sqrt{2V(x_0)})^2 \leq \frac{\alpha C_V}{2}|X_0 - x_0|^2, \tag{3.106}$$

and

$$\int_0^h |\sqrt{2V(X_t)} - \sqrt{2V(X_0)}|^2 dt \leq \frac{\alpha C_V}{2} \int_0^h |X_t - X_0|^2 dt. \tag{3.107}$$

With (3.106) and (3.107), we get

$$\begin{aligned} \mathbb{E}[\int_0^h \left\| \sqrt{2V(X_t)} - \sqrt{2V(x_0)} \right\|_F^2 dt] \leq & \frac{\alpha C_V dh(1+r)}{2} \mathbb{E}[|X_0 - x_0|^2] \\ & + \frac{\alpha C_V d(1+r^{-1})}{2} \int_0^h \mathbb{E}[|X_t - X_0|^2] dt. \end{aligned} \tag{3.108}$$

Next we apply Proposition 18 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

$$\begin{aligned} \mathbb{E}[|X_t - X_0|^2] &\leq \left(4dt\mathbb{E}\left[V(X_0)\right] + 4(\beta-1)^2 t^2 \mathbb{E}\left[|\nabla V(X_0)|^2\right]\right) \exp(1) \\ &\leq 12dt\mathbb{E}\left[V(X_0)\right] + 12(\beta-1)^2 t^2 \mathbb{E}\left[|\nabla V(X_0)|^2\right]. \end{aligned} \tag{3.109}$$

Combining (3.108) and (3.109), when $h < \frac{1}{4(\beta-1)L}$, we have that

$$\begin{aligned} & \mathbb{E}[\int_0^h \left\| \sqrt{2V(X_t)} - \sqrt{2V(x_0)} \right\|_F^2 dt] \\ \leq & \frac{1}{2}\alpha C_V d(1+r)h\mathbb{E}[|X_0 - x_0|^2] \\ & \quad + 6\alpha C_V d(1+r^{-1}) \int_0^h \left(dt\mathbb{E}\left[V(X_0)\right] + (\beta-1)^2 t^2 \mathbb{E}\left[|\nabla V(X_0)|^2\right]\right) dt \\ = & \frac{1}{2}\alpha C_V d(1+r)h\mathbb{E}[|X_0 - x_0|^2] \\ & \quad + 3\alpha C_V d^2(1+r^{-1})h^2\mathbb{E}\left[V(X_0)\right] + 2\alpha C_V d(\beta-1)^2(1+r^{-1})h^3\mathbb{E}\left[|\nabla V(X_0)|^2\right]. \end{aligned} \tag{3.110}$$

With (3.105) and (3.110), we get

$$\begin{aligned}
&\mathbb{E}[|U_1+U_3|^2]\\
&\leq \left(1-2(\beta-1)\alpha h+(\beta-1)^2\alpha^2h^2+\frac{1}{2}\alpha C_V d(1+r)h\right)\mathbb{E}\left[|X_0-x_0|^2\right]\\
&\quad +3\alpha C_V d^2(1+r^{-1})h^2\mathbb{E}\left[V(X_0)\right]+2\alpha C_V d(\beta-1)^2(1+r^{-1})h^3\mathbb{E}\left[|\nabla V(X_0)|^2\right]\\
&\leq \left(1-2(\beta-1)\alpha h+(\beta-1)^2\alpha^2h^2+\frac{1}{2}\alpha C_V d(1+r)h\right)\mathbb{E}\left[|X_0-x_0|^2\right]\\
&\quad +2\alpha C_V d(1+r^{-1})h^2\left(3d\mathbb{E}\left[V(X_0)\right]+2(\beta-1)^2h\mathbb{E}\left[|\nabla V(X_0)|^2\right]\right). \qquad (3.111)
\end{aligned}$$

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

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

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 $r=\delta/3$, (3.111) could be rewritten as

$$\begin{aligned}
\mathbb{E}[|U_1+U_3|^2]&\leq\left(1-\frac{\alpha(\beta-1)\delta}{3(1+\delta)}h\right)^2\mathbb{E}[|X_0-x_0|^2]\\
&\quad+\frac{8\alpha(\beta-1)(3+\delta)h^2}{(1+\delta)\delta}\left(3d\mathbb{E}\left[V(X_0)\right]+2(\beta-1)^2h\mathbb{E}\left[|\nabla V(X_0)|^2\right]\right). \qquad (3.112)
\end{aligned}$$

Next, with the bound in (3.109), we get when $h < \frac{1}{4(\beta-1)L}$,

$$\begin{aligned}
\mathbb{E}[|U_2|^2] &\leq (\beta-1)^2 L^2 \mathbb{E}\left[\left(\int_0^h |X_t - X_0| dt\right)^2\right] \\
&\leq (\beta-1)^2 L^2 h \int_0^h \mathbb{E}\left[|X_t - X_0|^2\right] dt \\
&\leq 6d(\beta-1)^2 L^2 h^3 \mathbb{E}\left[V(X_0)\right] + 4(\beta-1)^4 L^2 h^4 \mathbb{E}\left[|\nabla V(X_0)|^2\right]. \qquad (3.113)
\end{aligned}$$

With (3.112) and (3.113), 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

$$A = \frac{\alpha(\beta-1)\delta}{3(1+\delta)} h, \qquad (3.114)$$

$$B = \frac{4\alpha^{\frac{1}{2}}(\beta-1)^{\frac{1}{2}}(3+\delta)^{\frac{1}{2}} h}{(1+\delta)^{\frac{1}{2}}\delta^{\frac{1}{2}}}\left(d^{\frac{1}{2}}\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + (\beta-1)h^{\frac{1}{2}}\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}\right), \qquad (3.115)$$

$$C = 3d^{\frac{1}{2}}(\beta-1)Lh^{\frac{3}{2}}\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + 2(\beta-1)^2 Lh^2 \mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}}. \qquad (3.116)$$

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 [DK19, Lemma 9], with A, B, C given in (3.114),(3.115),(3.116), 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 (3.76). ■

We now prove Theorem 18.

Proof of Theorem 18. Following the same strategy and notation in the proof of Theorem 17, we have

$$X_h - x_1 = U_1 + U_2 + U_3 + (\beta-1)h\mathbb{E}[\zeta_0|x_0] + (\beta-1)h\left(\zeta_0 - \mathbb{E}[\zeta_0|x_0]\right), \tag{3.117}$$

where U_1, U_2, U_3 are defined in (3.102),(3.103),(3.104) respectively and $\zeta_0 = g_{\sigma,m}(x_0) - \nabla V(x_0)$. Therefore we have

$$\begin{aligned}\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} &\le \mathbb{E}\left[|U_1 + U_3 + (\beta-1)h\left(\zeta_0 - \mathbb{E}[\zeta_0|x_0]\right)|^2\right]^{\frac{1}{2}} \\ &\quad + \mathbb{E}\left[|U_2|^2\right]^{\frac{1}{2}} + (\beta-1)h\mathbb{E}\left[|\mathbb{E}[\zeta_0|x_0]|^2\right]^{\frac{1}{2}} \\ &= \left\{\mathbb{E}\left[|U_1+U_3|^2\right] + (\beta-1)^2h^2\mathbb{E}\left[|\zeta_0 - \mathbb{E}[\zeta_0|x_0]|^2\right]\right\}^{\frac{1}{2}} \\ &\quad + \mathbb{E}\left[|U_2|^2\right]^{\frac{1}{2}} + (\beta-1)h\mathbb{E}\left[|\mathbb{E}[\zeta_0|x_0]|^2\right]^{\frac{1}{2}}.\end{aligned} \tag{3.118}$$

From the proof of Theorem 17 and Proposition 14, when

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

we have that

$$\begin{aligned}\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} \le \Bigg\{&(1-A)^2\mathbb{E}\left[|X_0 - x_0|^2\right] + B^2 + \frac{\sigma^2}{2m}L^2(\beta-1)^2(d+3)^3h^2 \\ &+\frac{2(d+5)(\beta-1)^2h^2}{m}\mathbb{E}\left[|\nabla V(x_0)|^2\right]\Bigg\}^{\frac{1}{2}} + C + L\sigma(\beta-1)d^{\frac{1}{2}}h,\end{aligned} \tag{3.119}$$

where A, B, C are defined in (3.114),(3.115),(3.116). Using the fact that V is gradient Lipshcitz, we have

$$\begin{aligned}\mathbb{E}\left[|\nabla V(x_0)|^2\right] &\le \mathbb{E}\left[\left(|\nabla V(X_0)| + L|X_0 - x_0|\right)^2\right] \\ &\le 2\mathbb{E}\left[|\nabla V(X_0)|^2\right] + 2L^2\mathbb{E}\left[|X_0 - x_0|^2\right].\end{aligned} \tag{3.120}$$

Plugging (3.120) in (3.119), we get

$$\begin{aligned}\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} \leq &\bigg\{(1-A)^2\mathbb{E}\left[|X_0 - x_0|^2\right] + \frac{4(d+5)(\beta-1)^2L^2h^2}{m}\mathbb{E}\left[|X_0 - x_0|^2\right] + B^2 \\ &+\frac{\sigma^2}{2m}L^2(\beta-1)^2(d+3)^3h^2 + \frac{4(d+5)(\beta-1)^2h^2}{m}\mathbb{E}\left[|\nabla V(X_0)|^2\right]\bigg\}^{\frac{1}{2}} \\ &+ C + L\sigma(\beta-1)d^{\frac{1}{2}}h. \end{aligned} \tag{3.121}$$

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

$$\mathbb{E}\left[|X_h - x_1|^2\right]^{\frac{1}{2}} \leq \left\{(1-A')^2\mathbb{E}\left[|X_0 - x_0|^2\right] + B'^2\right\}^{\frac{1}{2}} + C', \tag{3.122}$$

where

$$A' = \frac{\alpha(\beta-1)\delta}{6(1+\delta)}h, \tag{3.123}$$

$$\begin{aligned} B' = &\left(\frac{4\alpha^{\frac{1}{2}}(\beta-1)^{\frac{3}{2}}(3+\delta)^{\frac{1}{2}}h^{\frac{3}{2}}}{(1+\delta)^{\frac{1}{2}}\delta^{\frac{1}{2}}} + \frac{2(\beta-1)(d+5)^{\frac{1}{2}}h}{m^{\frac{1}{2}}}\right)\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}} \\ &+ \frac{4\alpha^{\frac{1}{2}}(\beta-1)^{\frac{1}{2}}d^{\frac{1}{2}}(3+\delta)^{\frac{1}{2}}h}{(1+\delta)^{\frac{1}{2}}\delta^{\frac{1}{2}}}\,\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + \frac{\sigma L(\beta-1)(d+3)^{\frac{3}{2}}}{m^{\frac{1}{2}}}h, \end{aligned} \tag{3.124}$$

$$C' = 3L(\beta-1)d^{\frac{1}{2}}h^{\frac{3}{2}}\mathbb{E}_{\pi_\beta}\left[V(X)\right]^{\frac{1}{2}} + 2L(\beta-1)^2h^2\mathbb{E}_{\pi_\beta}\left[|\nabla V(X)|^2\right]^{\frac{1}{2}} + \sigma L(\beta-1)d^{\frac{1}{2}}h. \tag{3.125}$$

The rest of the proof is the same as the proof of Theorem 17, and hence we get (3.91). ■

3.3.10. Appendix.

3.3.10.1. *Computations for Section 3.3.5.3.*

Lemma 3.3.3. *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 $r \in (0, \beta - \frac{d}{2} - 1)$,*

$$\frac{Z_{\beta-1}}{Z_\beta} \leq \left(\frac{L}{\alpha}\right)^{\frac{\frac{d}{2}}{\beta-\frac{d}{2}-r}} V(0) \left(\frac{\Gamma(\beta)\Gamma(r)}{\Gamma(\frac{d}{2}+r)\Gamma(\beta-\frac{d}{2})}\right)^{\frac{1}{\beta-\frac{d}{2}-r}} . \tag{3.126}$$

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,

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

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

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

Therefore, we can further get

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

■

3.3.10.2. *Computations for Sections 3.3.7.1 and 3.3.7.2.* 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 $L = 2$. (3.68) reduces to

$$dX_t = b(x)dt + \sigma(X_t)dB_t, \tag{3.127}$$

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:

$$\begin{aligned}
&\langle b(x) - b(y), x - y\rangle + \frac{1}{2}\left\|(1 + |x|^2)^{\frac{1}{2}} I_d - (1 + |y|^2)^{\frac{1}{2}} I_d\right\|_F^2 \\
&= -2(\beta - 1)|x - y|^2 + d|(1 + |x|^2)^{\frac{1}{2}} - (1 + |y|^2)^{\frac{1}{2}}|^2 \\
&\le -2(\beta - 1 - \frac{d}{2})|x - y|^2, \qquad (3.128)
\end{aligned}$$

where the inequality follows from the fact that $x \mapsto (1 + |x|^2)^{\frac{1}{2}}$ is 1-Lipschitz. Therefore diffusion (3.127) 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 (3.127). We use the same notations in [LWME19]. According to [LWME19, 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 [LWME19, 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

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

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

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

3.3.10.3. *Computations for Remark 25.* 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

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

and

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

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 13 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 3.3.2, π_β satisfies the weighted Poincaré inequality with weight $1+|x|^2$ with weighted Poincaré constant

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

CHAPTER 4

Regularized Stein Variational Gradient Flow

Given a potential function $V : \mathbb{R}^d \to \mathbb{R}$, the sampling problem involves generating samples from the density

$$\pi(x) := Z^{-1}e^{-V(x)}, \qquad \text{with} \qquad Z := \int e^{-V(x)}dx \tag{4.1}$$

being the normalization constant, which is typically assumed to be unknown or hard to compute. The task of sampling arises in several fields of applied mathematics including Bayesian statistics and machine learning in the context of numerical integration. There are two widely-used approaches for sampling: (i) diffusion-based *randomized* algorithms, which are based on discretizations of certain diffusion processes, and (ii) particle-based *deterministic* algorithms, which are discretizations of certain *approximate* gradient flows. A central idea connecting the two approaches is the seminal work by [JKO98] which provided a variational interpretation of the Langevin diffusion as the Wasserstein Gradient Flow (WGF),

$$\partial_t\mu_t = \nabla \cdot (\mu_t \; \nabla_{W_2}F(\mu_t)) = \nabla \cdot \left(\mu_t \; \nabla \log \frac{\mu_t}{\pi}\right) \tag{4.2}$$

where the term $\nabla_{W_2}F(\mu_t) = \nabla \log \frac{\mu_t}{\pi}$ could be interpreted as the Wasserstein gradient[1] of the relative entropy functional (also called as the Kullback–Leibler divergence), defined by s

$$F(\mu_t) = \mathrm{KL}(\mu_t|\pi) := \int_{\mathbb{R}^d} \log \frac{\mu_t(x)}{\pi(x)} \mu_t(x)dx,$$

evaluated at μ_t. This leads to the idea that sampling could be viewed as *optimization on the space of measures*, a viewpoint that has provided a deeper understanding of the sampling problem [Wib18, TSA20].

[1] See, for example, [AGS05, San17] for the exact definition.

There are several merits and disadvantages to both the randomized and deterministic discretization of the (approximate) WGF. First, note that obtaining exact space-time discretization of the WGF in (4.2) is not possible. Indeed, due to the presence of the diffusion term, when initialized with an N-particle based empirical measure, the particles do not remain as particles for any time $t > 0$. Hence, on the one hand, randomized discretizations like the Langevin Monte Carlo algorithm, are used as implementable space-time discretizations of the WGF. On the other hand, motivated by applications where the randomness in the discretization is undesirable, in the applied mathematics literature, other discretizations of *approximate* WGF were developed. Such methods are predominantly based on using mollifiers and we refer the reader to [Rav85, Rus90, DM90, CB16, CCP19] for a partial list and to [Che17], for a comprehensive overview.

Recently, in the machine learning community, the Stein Variational Gradient Descent [LW16, Liu17] was proposed as another deterministic discretization of approximate WGF, and has gathered significant attention due to applications to reinforcement learning [LRLP17], graphical modeling [WZL18], measure quantization [XKS22], and other fields of machine learning and applied mathematics [WTBL19, CLGL$^+$20a, CLGL$^+$20a, KSA$^+$20]. Due to the use of the reproducing kernels, the Stein Variational Gradient Descent (SVGD) algorithm provides a space-time discretization of the following *approximate* Wasserstein Gradient Flow (which we refer to as the Stein Variational Gradient Flow (SVGF) for simplicity)

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \ \mathcal{T}_{k,\mu_t} \nabla \log \frac{\mu_t}{\pi} \right), \tag{4.3}$$

where $\mathcal{T}_{k,\mu} : L_2^d(\mu) \to L_2^d(\mu)$ is the integral operator defined as $\mathcal{T}_{k,\mu} f(x) = \int k(x,y) f(y) d\mu(y)$ for a function $f \in L_2^d(\mu)$, and for a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$; see, for example [LLN19]. Hence, SVGD (which is based on the SVGF), in this context, while being deterministic only provides a discretization of a *constant-order approximation* to the Wasserstein Gradient Flow due to the presence of the kernel integral operator. Indeed, if $\text{supp}(\mu_t) = \mathbb{R}^d$ and k is bounded continuous translation invariant *characteristic* kernel [SGF$^+$10] on $\mathbb{R}^d$ (e.g., Gaussian, Laplacian kernels), then

$$\begin{aligned}\|\mathcal{T}_{k,\mu_t} - I\|_{\text{op}} &= \sup\{\|\mathcal{T}_{k,\mu_t} f - f\|_{L_2^d(\mu_t)} : \|f\|_{L_2^d(\mu_t)} = 1\} \geq \|\mathcal{T}_{k,\mu_t}\mathbf{1} - \mathbf{1}\|_{L_2^d(\mu_t)} \\ &\geq \|1 - \textstyle\int k(\cdot, x)\mu_t(x)\, dx\|_{L_2(\mu_t)} > 0,\end{aligned}$$

where $\mathbf{1} = (1, \overset{d}{\ldots}, 1)^\top$. This shows that the order of the error is crucially dependent on the choice of the kernel k.

To overcome the above issue with the SVGF, in this work, we propose the Regularized Stein Variational Gradient Flow (R-SVGF). To motivate the proposed flow, we first note that the Wasserstein gradient $\nabla \log(\mu_t/\pi)$ lives in $L_2^d(\mu_t)$, while the kernelized Wasserstein gradient $\mathcal{T}_{k,\mu_t}\nabla \log(\mu_t/\pi)$ *morally* lives in $\mathcal{H}_k^d \subset L_2^d(\mu_t)$. If $\nabla \log(\mu_t/\pi) \in \overline{\text{Ran}(\mathcal{T}_{k,\mu_t})}$, then it is easy to verify that

$$\|((1-\nu)\mathcal{T}_{k,\mu_t} + \nu I)^{-1}\mathcal{T}_{k,\mu_t}\nabla \log(\mu_t/\pi) - \nabla \log(\mu_t/\pi)\|_{L^d(\mu_t)} \to 0, \quad \text{as} \quad \nu \to 0.$$

Additionally, if $\nabla \log(\mu_t/\pi)$ is sufficiently smooth, i.e., there exists $\gamma \in \left(0, \frac{1}{2}\right]$ such that $\nabla \log(\mu_t/\pi) = \mathcal{T}_{k,\mu_t}^{\gamma} h$, for some $h \in L_2^d(\mu_t)$ (see, for example, [CZ07]), then

$$\|((1-\nu)\mathcal{T}_{k,\mu_t} + \nu I)^{-1}\mathcal{T}_{k,\mu_t}\nabla \log(\mu_t/\pi) - \nabla \log(\mu_t/\pi)\|_{L^d(\mu_t)} = O(\nu^{2\gamma}), \quad \text{as} \quad \nu \to 0.$$

In other words, $\|((1-\nu)\mathcal{T}_{k,\mu_t} + \nu I)^{-1}\mathcal{T}_{k,\mu_t}\nabla \log(\mu_t/\pi)$ is a good approximation to $\nabla \log(\mu_t/\pi)$ for small ν. With this motivation, we propose the following R-SVGF given by

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \ \ ((1-\nu)\mathcal{T}_{k,\mu_t} + \nu I)^{-1}\, \mathcal{T}_{k,\mu_t}\left(\nabla \log \frac{\mu_t}{\pi}\right)\right), \tag{4.4}$$

for some regularization parameter $\nu \in (0, 1]$, where R-SVGF arbitrarily approximates the WGF as $\nu \to 0$. It is important to note that in the case of $\gamma = 1/2$, we have $\nabla \log(\mu_t/\pi) \in \mathcal{H}_k^d$, yet, (4.3) suffers from the drawback of providing only a constant-order approximation to (4.2).

Summary of Contributions. Our contributions in this work are as follows:

(1) We propose the Regularized SVGF (R-SVGF) that interpolates between the Wasserstein Gradient Flow and the SVGF. The advantage of the proposed flow is that one could obtain an implementable space-time discretization as long as the regularization parameter

is bounded away from zero. The main intuition behind the proposed flow is to pick an appropriately small choice of regularization parameter so that we could arbitrarily approximate the WGF (Theorems 19 and 21).

(2) For the R-SVGF, we provide rates of convergence to the equilibrium density in two cases: (i) in the Fisher Information metric under no assumptions on the target (Theorem 20) and (ii) in the KL-divergence metric under an LSI assumption on the target (Theorem 22). We also establish similar results for the time-discretized R-SVGF (Theorems 23 and 24).

(3) We characterize the existence and uniqueness (Theorem 25), and stability (Theorem 26) of the solutions to the R-SVGF in the mean-field limit.

(4) We provide preliminary numerical experiments demonstrating the advantage of the space-time discretization of the R-SVGF, which we call as the the Regularized Stein Variational Gradient Descent (R-SVGD) algorithm, over the standard SVGD algorithm.

4.1. Organizations

The rest of the chapter is organized as follows. In Section 4.2, we introduce the notations used in the rest of the paper. In Section 4.3, we provide the preliminaries on reproducing kernel Hilbert spaces required for our work. In Section 4.4, we introduce the R-SVGF, along with the notion of regularized Stein-Fisher information, required for our analysis. Due to the technical nature of the proofs, we postpone the results on existence and uniqueness of the R-SVGF, and related stability results respectively to Sections 4.6 and 4.7. In Section 4.5, we provide convergence results on the R-SVGF flow and its time-discretized version. We conclude in Section 4.8 with a space-time discretization which provides a practically implementable algorithm, and provide preliminary empirical results.

4.2. Notations

We use the following notations throughout this work:

- For a matrix, $\|\cdot\|_2$ denotes the matrix 2-norm (spectral norm) and $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm which is defined as $\|A\|_{HS}^2 = \sum_{i,j=1}^d |a_{ij}|^2$ for any matrix $A = (a_{ij})_{i,j\in[d]}$.
- The term id denotes the $d \times d$ identity matrix. I_d corresponds to the identity operator in the RKHS. I corresponds to the identity operator in $L_2(\mu)$.
- $\mathcal{P}(\mathbb{R}^d)$ denotes the space of all probability measures on $\mathbb{R}^d$, and $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of all probability measures on $\mathbb{R}^d$ with finite second moments.
- $\Big(L_\infty(\mathbb{R}^d), \|\cdot\|_{L_\infty(\mathbb{R}^d)}\Big)$ denotes the space of all essentially bounded measurable functions on $\mathbb{R}^d$ with $\|f\|_{L_\infty(\mathbb{R}^d)} := \inf\{C : |f(x)| \le C \text{ for almost every } x \in \mathbb{R}^d\}$ for any $f \in L_\infty(\mathbb{R}^d)$.
- For any $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\Big(L_2(\mu), \|\cdot\|_{L_2(\mu)}\Big)$ is the space of all μ-square integrable measurable function on $\mathbb{R}^d$ with $\|f\|_{L_2(\mu)}^2 := \int_{\mathbb{R}^d} |f(x)|^2 \mu(dx)$.
- Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $\big(\mathcal{G}, \|\cdot\|_{\mathcal{G}}\big)$ denote two function spaces. For an operator $A : \mathcal{H} \to \mathcal{G}$, we denote the adjoint operator of A by A^*. We denote the operator norm by $\|A\|_{\mathcal{H}\to\mathcal{G}}$, which is defined as $\|A\|_{\mathcal{H}\to\mathcal{G}} := \sup_{\|u\|_{\mathcal{H}}\le 1} \|Au\|_{\mathcal{G}}$. When we don't emphasize the spaces, we denote the operator norm of A by $\|A\|_{\mathrm{op}}$ for simplicity.
- Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ denote two Hilbert spaces. For an operator $A : \mathcal{H} \to \mathcal{G}$, we denote the Hilbert-Schmidt norm by $\|A\|_{HS}$ which is defined as $\|A\|_{HS}^2 := \sum_{i\in I} \|Ae_i\|_{\mathcal{G}}^2$ where $\{e_i\}_{i\in I}$ is an orthonormal basis of $\mathcal{H}$. We denote the nuclear norm by $\|A\|_{nuc}$ which is defined as $\|A\|_{nuc} := \sum_{i\in I} \langle (A^*A)^{\frac{1}{2}} e_i, e_i\rangle_{\mathcal{H}}$ where $\{e_i\}_{i\in I}$ is an orthonormal basis of $\mathcal{H}$.
- For a smooth function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\nabla_1 f$ denotes the gradient of f in the first variable and $\nabla_2 f$ denotes the gradient of f in the second variable.
- For a map $\phi : \mathbb{R}^d \to \mathbb{R}^d$, ϕ_i denotes the i-th component of the function value and $\nabla\phi$ denotes the Jacobian, i.e., $(\nabla\phi)_{ij} = \partial_j \phi_i$.
- $T_\# \rho$ represents the push-forward of the density ρ under a map T.
- $\langle\cdot,\cdot\rangle_H$ denotes inner-product in the Hilbert space H. $\langle\cdot,\cdot\rangle$ denotes inner-product in the Euclidean space $\mathbb{R}^d$.

- $\mathcal{C}(\mathbb{R}^d;\mathbb{R}^d)$ is the space of all $\mathbb{R}^d$-valued continuous functions on $\mathbb{R}^d$.
- For any function space $\mathcal{H}$ on $\mathbb{R}^d$, $\mathcal{C}([0,T];\mathcal{H})$ is the space of functions f such that for any fixed $t \in [0,T]$, $f(t,\cdot) \in \mathcal{H}$ and for any fixed $x \in \mathbb{R}^d$, $f(\cdot,x)$ is a continuous function on $[0,T]$. $\mathcal{C}^1([0,T];\mathcal{H})$ is the space of functions f such that for any fixed $t \in [0,T]$, $f(t,\cdot) \in \mathcal{H}$ and for any fixed $x \in \mathbb{R}^d$, $f(\cdot,x)$ is a continuous function with continuous first order derivative on $[0,T]$.
- $\mathcal{C}_0^\infty([0,\infty)\times\mathbb{R}^d)$ is the space of all measurable functions on $[0,\infty)\times\mathbb{R}^d$ that vanish at infinity, i.e., for any $f \in \mathcal{C}_0^\infty([0,\infty)\times\mathbb{R}^d)$, $f(t,x) \to 0$ as $t \to \infty$ and $f(t,x) \to 0$ as $x \to \infty$.
- Suppose $f : \mathbb{R}^d \to \mathbb{R}^d$ is a vector-valued function. For a function space $\mathcal{H}$, we say $f \in \mathcal{H}^d$ if $f = [f_1, \cdots, f_d]$ such that $f_i \in \mathcal{H}$ for all $i \in [d]$.

4.3. Preliminaries on Reproducing Kernel Hilbert Space

In this section, we introduce some properties of RKHS which would be used later in the formulation and analysis of R-SVGF. We refer the reader to [SC08, BTA11, PR16] for the basics of RKHS. We let $\mathcal{H}_k$ to be a separable RKHS over $\mathbb{R}^d$ with the reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{>0}$ and with $\|\cdot\|_{\mathcal{H}_k}$ denoting the associated RKHS norm. We make the following assumption on the kernel function k throughout the chapter.

Assumption A1. The kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is strictly positive definite, continuous and bounded.

The following results are essentially based on [SC08, Lemma 4.23, and Theorems 4.26 and 4.27].

Proposition 19 ([SC08]). Under Assumption A1, the following holds.

(i) The kernel function k is bounded if and only if every $f \in \mathcal{H}_k$ is bounded. Moreover, the inclusion $i_d : \mathcal{H}_k \to L_\infty(\mathbb{R}^d)$ is continuous and $\|i_d\|_{\mathcal{H}_k \to L_\infty(\mathbb{R}^d)} = \|k\|_\infty$, where $\|k\|_\infty := \sup_{x\in\mathbb{R}^d} \sqrt{k(x,x)}$.

(ii) Let μ be a σ-finite measure on $\mathbb{R}^d$. Assume that

$$\|k\|_{L_2(\mu)} := \left(\int_{\mathbb{R}^d} k(x,x)d\mu(x)\right)^{\frac{1}{2}} < \infty.$$

Then $\mathcal{H}_k$ consists of 2-integrable functions and the inclusion $\iota_{k,\mu} : \mathcal{H}_k \to L_2(\mu)$ is continuous with $\|\iota_{k,\mu}\|_{\mathcal{H}_k \to L_2(\mu)} \leq \|k\|_{L_2(\mu)}$. Moreover, the adjoint of this inclusion is the operator $\iota_{k,\mu}^* : L_2(\mu) \to \mathcal{H}_k$ defined by

$$\iota_{k,\mu}^* g(x) := \int_{\mathbb{R}^d} k(x,y)g(y)\mu(y)dy, \qquad g \in L_2(\mu),\ x \in \mathbb{R}^d.$$

(iii) $\mathcal{H}_k$ is dense in $L_2(\mu)$ if and only if $\iota_{k,\mu}^* : L_2(\mu) \to \mathcal{H}_k$ is injective. Alternatively, $\iota_{k,\mu}^* : L_2(\mu) \to \mathcal{H}_k$ has a dense image if and only if $\iota_{k,\mu} : \mathcal{H}_k \to L_2(\mu)$ is injective.

(iv) $\iota_{k,\mu} : \mathcal{H}_k \to L_2(\mu)$ is a Hilbert-Schmidt operator with $\|\iota_{k,\mu}\|_{HS} = \|k\|_{L_2(\mu)}$. Moreover, the integral operator $\mathcal{T}_{k,\mu} = \iota_{k,\mu}\iota_{k,\mu}^* : L_2(\mu) \to L_2(\mu)$ is compact, positive, self-adjoint, and nuclear with $\|\mathcal{T}_{k,\mu}\|_{nuc} = \|\iota_{k,\mu}\|_{HS} = \|k\|_{L_2(\mu)}$.

The RKHS norm of $f \in \mathcal{H}_k^d$ is given by $\|f\|_{\mathcal{H}_k^d}^2 := \sum_{i=1}^d \|f_i\|_{\mathcal{H}_k}^2$. The $L_2^d(\mu)$ norm of $f \in L_2^d(\mu)$ is given by $\|f\|_{L_2^d(\mu)}^2 := \sum_{i=1}^d \|f_i\|_{L_2(\mu)}^2$. When $f \in \mathcal{H}_k^d$ with $f = [f_1, \cdots, f_d]$ and $g \in \mathcal{H}_k$, we define $\langle f, g\rangle_{\mathcal{H}_k}$ as a vector in $\mathbb{R}^d$ and $(\langle f, g\rangle_{\mathcal{H}_k})_i = \langle f_i, g\rangle_{\mathcal{H}_k}$ for all $i \in [d]$. When $f \in L_2^d(\mu)$ with $f = [f_1, \cdots, f_d]$ and $g \in L_2(\mu)$, we define $\langle f, g\rangle_{L_2(\mu)}$ as a vector in $\mathbb{R}^d$ and $\left(\langle f, g\rangle_{L_2(\mu)}\right)_i = \langle f_i, g\rangle_{L_2(\mu)}$ for all $i \in [d]$. Note also that $\mathrm{Ran}((\iota_{k,\mu}\iota_{k,\mu}^*)^{1/2}) = \mathcal{H}_k^d \subset L_2^d(\mu)$. We refer the interested reader to [CZ07] for more details.

Finally, we remark that by letting $(\lambda_i, e_i)_{i=1}^\infty$ to be the set of eigenvalues and eigenfunctions of the operator $\iota_{k,\mu}\iota_{k,\mu}^*$ where $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ and $(e_i)_{i=1}^\infty$ form an orthonormal system in $\mathrm{Ran}(\iota_{k,\mu}\iota_{k,\mu}^*)$, we have the following spectral representation that, for all $f \in \mathrm{Ran}(\iota_{k,\mu}\iota_{k,\mu}^*)$,

$$\iota_{k,\mu}\iota_{k,\mu}^* f = \sum_{i=1}^\infty \lambda_i \langle f, e_i\rangle_{L_2(\mu)} e_i. \tag{4.5}$$

Computing the spectral representation, in general for any given μ and kernel k is a non-trivial task. Results are only known on a case-by-case basis; see, for example, [MNY06, AM14, CX20, SH21].

However, we use the decomposition only in our analysis. For the purely practical algorithm that we describe eventually in Section 4.8, we do not need to know the decomposition explicitly.

4.4. Regularized SVGF

We now introduce the formulation of the Regularized-SVGF and discuss its connection with SVGF and the WGF. Recall that in the mean-field limit, the SVGF in (4.3) only provides a constant order approximation to the WGF in (4.2), due to the presence of the operator $\mathcal{T}_{k,\mu}$. As the operator $\mathcal{T}_{k,\mu}$ is not invertible, we seek to obtain a regularized inverse so that we end up with the following Regularized-SVGF, as in (4.4), for some regularization parameter $\nu \in (0,1]$. Note in particular that as $\nu \to 0$, the Regularized-SVGF gets arbitrarily close to the WGF. Our goal in this section is to derive the above mentioned R-SVGF from first principles.

The central operator required in our formulation is the following *Stein operator*, which is defined for all $p \in \mathcal{P}(\mathbb{R}^d)$, and for all smooth maps $\phi : \mathbb{R}^d \to \mathbb{R}^d$, as

$$\mathcal{A}_p\phi(x) = \phi(x) \otimes \nabla \log p(x) + \nabla\phi(x),$$

where $\otimes$ denotes the outer-product. Now, the Wasserstein Gradient Flow in (4.2) could be thought of as follows. Consider moving a particle $x \sim \rho$ (for some $\rho \in \mathcal{P}(\mathbb{R}^d)$) based on the mapping $x \mapsto T(x) \coloneqq x + h\phi(x)$, where $h > 0$ is a step-size parameter, and ϕ is a vector-field chosen so that the KL-divergence between the pushforward of ρ according to T, denoted as $T_{\#}\rho$, and the target density π in minimal. Liu and Wang [LW16, Theorem 3.1], showed that

$$\nabla_h \mathrm{KL}(T_{\#}\rho|\pi)|_{h=0} = -\mathbb{E}_{x\sim\rho}[\mathrm{trace}(\mathcal{A}_\pi\phi(x))].$$

We also refer to [JKO98] for an earlier version of the same result. Based on this observation, if we try to find the vector-field ϕ in the unit-ball of $L_2^d(\rho)$ that maximizes the quantity $[\mathbb{E}_{x\sim\rho}[\mathrm{trace}(\mathcal{A}_\pi\phi(x))]]^2$, a straight-forward calculation based on integration-by-parts, results in the optimal ϕ being the Wasserstein gradient $\nabla \log \frac{\rho}{\pi}$. To have a practical implementation, [LW16] considered maximizing $[\mathbb{E}_{x\sim\rho}[\mathrm{trace}(\mathcal{A}_\pi\phi(x))]]^2$ over the unit-ball in the RKHS $\mathcal{H}_k^d$, which results in the optimal vector-field being equal to $\mathcal{T}_{k,\rho}\nabla \log \frac{\rho}{\pi}$, and correspondingly results in the SVGF in (4.3).

In this work, we propose to find the vector field ϕ that maximizes $[\mathbb{E}_{x\sim\rho}[\mathrm{trace}(\mathcal{A}_\pi\phi(x))]]^2$ over the unit-ball with respect to an interpolated norm between $L_2^d(\rho)$ and $\mathcal{H}_k^d$. Specifically, the interpolation norm that we consider is of the form $\nu\left\|\cdot\right\|^2_{\mathcal{H}_k^d} + (1-\nu)\left\|\cdot\right\|^2_{L_2^d(\rho)}$, for some regularization parameter $\nu \in (0,1]$, which trades-off between $\|\cdot\|^2_{\mathcal{H}_k^d}$ and $\|\cdot\|^2_{L_2^d(\rho)}$. We also remark here that a similar idea has been leveraged in the context of RKHS-based statistical hypothesis testing [BLY21]. Formally, for $\rho, \pi \in \mathcal{P}(\mathbb{R}^d)$, we consider the following optimization problem.

$$S(\rho,\pi) := \max_{\phi\in\mathcal{H}_k^d}\left\{[\mathbb{E}_{x\sim q}[\mathrm{trace}(\mathcal{A}_\pi\phi(x))]]^2 \qquad \text{such that} \quad \nu\left\|\phi\right\|^2_{\mathcal{H}_k^d} + (1-\nu)\left\|\phi\right\|^2_{L_2^d(\rho)} \leq 1\right\}.$$

For any $\rho \in \mathcal{P}(\mathbb{R}^d)$, the optimal vector field, ϕ that minimizes $\mathrm{KL}(T_\#\rho|\pi)$ can be described via the following result.

Proposition 20. Let $T(x) = x + h\phi(x)$ and $T_\#\rho(z)$ be the density of $z = T(x)$ when $x \sim \rho$, for some density $\rho \in \mathcal{P}(\mathbb{R}^d)$. For $\nu \in (0,1]$, define

$$\mathcal{B} := \{\phi \in \mathcal{H}_k^d : \nu\left\|\phi\right\|^2_{\mathcal{H}_k^d} + (1-\nu)\left\|\phi\right\|^2_{L_2^d(\rho)} \leq 1\}.$$

Then the direction of steepest descent in $\mathcal{B}$ that maximizes $-\nabla_h KL(T_\#\rho|\pi)|_{h=0}$ is given by

$$\phi^*_{\rho,\pi}(\cdot) \propto \left((1-\nu)\iota^*_{k,\rho}\iota_{k,\rho} + \nu I_d\right)^{-1}\mathbb{E}_{x\sim\rho}[-\nabla V(x)k(x,\cdot) + \nabla k(x,\cdot)],$$

where $\iota_{k,\rho} : \mathcal{H}_k^d \to L_2^d(\rho)$ is the inclusion operator and $\iota^*_{k,\rho}$ is its adjoint as in Proposition 19. Furthermore, under the optimal vector field $\phi^*_{\rho,\pi}$, we have $-\nabla_h KL(T_\#\rho|\pi)|_{h=0} = S(\rho,\pi)$.

Proof. First note that according to [LW16, Theorem 3.1], we have

$$\nabla_h \mathrm{KL}(T_\#\rho|\pi)|_{h=0} = -\mathbb{E}_{x\sim q}[\mathrm{trace}(\mathcal{A}_\pi\phi(x))].$$

Therefore, we have

$$\phi^*_{\rho,\pi} = \arg\max_{\phi\in\mathcal{H}_k^d}\left\{[\mathbb{E}_{x\sim\rho}[\mathrm{trace}(\mathcal{A}_\pi\phi(x))]]^2 \qquad \text{such that} \quad \nu\left\|\phi\right\|^2_{\mathcal{H}_k^d} + (1-\nu)\left\|\phi\right\|^2_{L_2^d(\rho)} \leq 1\right\}.$$

Next, observe that we have

$$\begin{aligned}\mathbb{E}_{x\sim\rho}[\mathrm{trace}(\mathcal{A}_\pi\phi(x))] &= \sum_{i=1}^d \mathbb{E}_{x\sim\rho}[-\partial_i V(x)\phi_i(x) + \partial_i\phi_i(x)] \\ &= \sum_{i=1}^d \mathbb{E}_{x\sim\rho}[-\partial_i V(x)\langle\phi_i, k(x,\cdot)\rangle_{\mathcal{H}_k} + \langle\phi_i, \partial_i k(x,\cdot)\rangle_{\mathcal{H}_k}] \\ &= \langle\phi, \mathbb{E}_{x\sim\rho}[-\nabla V(x)k(x,\cdot) + \nabla k(x,\cdot)]\rangle_{\mathcal{H}_k^d}.\end{aligned}$$

Meanwhile, the constraint can be written as

$$\begin{aligned}\nu\left\|\phi\right\|_{\mathcal{H}_k^d}^2 + (1-\nu)\left\|\phi\right\|_{L_2^d(\rho)}^2 &= \nu\langle\phi,\phi\rangle_{\mathcal{H}_k^d} + (1-\nu)\langle\iota_{k,\rho}\phi, \iota_{k,\rho}\phi\rangle_{L_2^d(\rho)} \\ &= \langle\left(\nu I_d + (1-\nu)\iota_{k,\rho}^*\iota_{k,\rho}\right)\phi, \phi\rangle_{\mathcal{H}_k^d} \\ &= \left\|\left((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\right)^{\frac{1}{2}}\phi\right\|_{\mathcal{H}_k^d}^2,\end{aligned}$$

where $I_d : \mathcal{H}_k \to \mathcal{H}_k$ is the identity operator. Now, note that $\left((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\right)^{\frac{1}{2}}$ is well-defined since $\iota_{k,\rho}^*\iota_{k,\rho} : \mathcal{H}_k^d \to \mathcal{H}_k^d$ is positive, compact and self-adjoint. Therefore based on the above display, the constraint $\{\phi\in\mathcal{H}_k^d : \nu\left\|\phi\right\|_{\mathcal{H}_k^d}^2 + (1-\nu)\left\|\phi\right\|_{L_2^d(\rho)}^2 \leq 1\}$ is equivalent to

$$\{\phi\in\mathcal{H}_k^d : \psi = \left((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\right)^{\frac{1}{2}}\phi \quad and \quad \|\psi\|_{\mathcal{H}_k^d} \leq 1\}.$$

Since the spectrum of $\iota_{k,\rho}^*\iota_{k,\rho}$ is positive and $\nu\in(0,1]$, $(1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d$ is invertible. For all $\phi\in\mathcal{H}_k^d$, there exists a unique $\psi\in\mathcal{H}_k^d$ such that $\left((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\right)^{-\frac{1}{2}}\psi = \phi$. Applying this fact along with the equivalent form of the constraint, we have

$$\begin{aligned}\mathbb{E}_{x\sim\rho}(\mathrm{trace}(\mathcal{A}_\pi\phi(x))) &= \left\langle\left((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\right)^{-\frac{1}{2}}\psi, \mathbb{E}_{x\sim\rho}[-\nabla V(x)k(x,\cdot) + \nabla k(x,\cdot)]\right\rangle_{\mathcal{H}_k^d} \\ &= \left\langle\psi, \left((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\right)^{-\frac{1}{2}}\mathbb{E}_{x\sim\rho}[-\nabla V(x)k(x,\cdot) + \nabla k(x,\cdot)]\right\rangle_{\mathcal{H}_k^d} \\ &\leq \left\|\left((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\right)^{-\frac{1}{2}}\mathbb{E}_{x\sim\rho}[-\nabla V(x)k(x,\cdot) + \nabla k(x,\cdot)]\right\|_{\mathcal{H}_k^d}\end{aligned}$$

where the second identity follows from the fact that $\Big((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\Big)^{-\frac{1}{2}}$ is self-adjoint and the upper bound in the last inequality is achieved when

$$\psi^* \propto \big((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\big)^{-\frac{1}{2}} \mathbb{E}_{x\sim\rho}[-\nabla V(x)k(x,\cdot) + \nabla k(x,\cdot)],$$

and the result hence follows. ■

With the optimal-vector field as derived above, we consider the following mean-field partial differential equation (PDE) as the R-SVGF:

$$\partial_t\rho_t = \nabla\cdot\left(\rho_t\ \iota_{k,\rho_t}\big((1-\nu)\iota_{k,\rho_t}^*\iota_{k,\rho_t} + \nu I_d\big)^{-1}\iota_{k,\rho_t}^*\left(\nabla\log\frac{\rho_t}{\pi}\right)\right). \tag{4.6}$$

It is important to notice that the R-SVGF interpolates between SVGF and WGF. However, the regime of interest for us is when $\nu \to 0$, as we get arbitrarily close to the WGF. We quantify this statement precisely in the later sections. On the other hand, when $\nu \to 1$ R-SVGF becomes the SVGF.

Remark 26. We now make the following remarks about the above result.

(i) We can alternatively write $\phi_{\rho,\pi}^*$ from Proposition 20 as

$$\phi_{\rho,\pi}^* \propto -\big((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\big)^{-1}\iota_{k,\rho}^*\left(\nabla\log\frac{\rho}{\pi}\right),$$

since we have

$$\begin{aligned}\mathbb{E}_{x\sim\rho}[-\nabla V(x)k(x,\cdot) + \nabla k(x,\cdot)] &= \int_{\mathbb{R}^d} k(\cdot,x)\left(-\nabla V(x) - \frac{\nabla\rho(x)}{\rho(x)}\right)\rho(x)dx\\ &= -\,\iota_{k,\rho}^*\left(\nabla\log\frac{\rho}{\pi}\right).\end{aligned}$$

(ii) The operator in (4.6) has an equivalent expression as we discuss below. First, we claim that

$$\iota_{k,\rho}\big((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\big)^{-1}\iota_{k,\rho}^* = \big((1-\nu)\iota_{k,\rho}\iota_{k,\rho}^* + \nu I\big)^{-1}\iota_{k,\rho}\iota_{k,\rho}^*.$$

To see that, we start with the trivial identity in the first line below and proceed as

$$\begin{aligned}
&\left((1-\nu)\iota_{k,\rho}\iota_{k,\rho}^{*}+\nu I\right)\iota_{k,\rho}=\iota_{k,\rho}\left((1-\nu)\iota_{k,\rho}^{*}\iota_{k,\rho}+\nu I_d\right),\\
\Longrightarrow\ &\iota_{k,\rho}=\left((1-\nu)\iota_{k,\rho}\iota_{k,\rho}^{*}+\nu I\right)^{-1}\iota_{k,\rho}\left((1-\nu)\iota_{k,\rho}^{*}\iota_{k,\rho}+\nu I_d\right)\\
\Longrightarrow\ &\iota_{k,\rho}\left((1-\nu)\iota_{k,\rho}^{*}\iota_{k,\rho}+\nu I_d\right)^{-1}=\left((1-\nu)\iota_{k,\rho}\iota_{k,\rho}^{*}+\nu I\right)^{-1}\iota_{k,\rho}\\
\Longrightarrow\ &\iota_{k,\rho}\left((1-\nu)\iota_{k,\rho}^{*}\iota_{k,\rho}+\nu I_d\right)^{-1}\iota_{k,\rho}^{*}=\left((1-\nu)\iota_{k,\rho}\iota_{k,\rho}^{*}+\nu I\right)^{-1}\iota_{k,\rho}\iota_{k,\rho}^{*}.
\end{aligned}$$

According to this observation, (4.6) can also be written in the following form

$$\begin{aligned}
\partial_t\rho_t&=\nabla\cdot\left(\rho_t\ \left((1-\nu)\iota_{k,\rho_t}\iota_{k,\rho_t}^{*}+\nu I\right)^{-1}\iota_{k,\rho_t}\iota_{k,\rho_t}^{*}\left(\nabla\log\frac{\rho_t}{\pi}\right)\right)\\
&=\nabla\cdot\left(\rho_t\ \left((1-\nu)\mathcal{T}_{k,\rho_t}+\nu I\right)^{-1}\mathcal{T}_{k,\rho_t}\left(\nabla\log\frac{\rho_t}{\pi}\right)\right),
\end{aligned}$$

thereby providing the R-SVGF introduced in (4.4).

(iii) **Particle-based spatial discretization.** We now describe the spatial discretization of the R-SVGF. Based on the results in Proposition 20 and Remark 26, we obtain the following ODE system:

$$\begin{cases}
\dfrac{dx_i(t)}{dt}=-\Bigg((1-\nu)\iota_{k,\rho_t^N}^{*}\iota_{k,\rho_t^N}+\nu I_d\Bigg)^{-1}\Bigg(\dfrac{1}{N}\displaystyle\sum_{j=1}^{N}-\nabla_2 k\left(x_i(t),x_j(t)\right)+k\left(x_i(t),x_j(t)\right)\nabla V(x_j(t))\\
bigg)\\
x_i(0)=x_i^0\in\mathbb{R}^d,\quad i=1,2,\ldots,N
\end{cases},$$

where $\{x_i(t)\}_{i=1}^{N}$ is the set of N particles. $\rho_t^N=\frac{1}{N}\sum_{j=1}^{N}\delta_{x_j(t)}$ is the empirical distribution at time t, provides a N-particle spatial discretization of the R-SVGF.

(iv) **Time discretization.** We also have the following time-discretization of the R-SVGF. Let $\{h_n\}_{n=1}^{\infty}$ be the sequence of time step-size. We denote the density at the n-th iterate by ρ^n for all integers $n\geq 1$. Then the time discretization of the R-SVGF can be written as

$$\rho^{n+1}=\left(id-h_{n+1}\mathcal{D}_{\nu_{n+1},\rho^n}\nabla\log\frac{\rho^n}{\pi}\right)_{\#}\rho^n, \tag{4.7}$$

where $\mathcal{D}_{\nu_n,\rho^n}=\Big((1-\nu_n)\iota_{k,\rho^n}\iota_{k,\rho^n}^{*}+\nu_n I_d\Big)^{-1}\iota_{k,\rho^n}\iota_{k,\rho^n}^{*}$.

(v) The parameter ν can also be made to be dependent on t or n; in fact, in our analysis we pick a time-varying regularization parameter.

4.5. Convergence Results in Continuous and Discrete Time

Our goal in this section is derive convergence guarantees for the R-SVGF. Before we proceed, we introduce the notion of Regularized Stein-Fisher information (or Regularized Kernel Stein Discrepancy).

4.5.1. Regularized Stein-Fisher Information and its Properties. Note that several works, for example [KSA+20, DNS19, SSR22], used the notion of Stein-Fisher Information to understand the convergence properties of the SVGD algorithm. The Stein-Fisher information was introduced in [CSG16, LLJ16, GM17] under the name Kernel Stein Discrepancy. However, a drawback of the Stein-Fisher information is that it is a weaker metric, for example in comparison to the Fisher information metric; see [GM17, GDVM19, SGBSM20]. Below, we introduce a regularized version of the Stein-Fisher information and show that as the regularization parameter tends to zero, it converges to the standard Fisher information.

Let $\rho, \pi \in \mathcal{P}(\mathbb{R}^d)$, then, the Fisher information corresponds to

$$I(\rho|\pi) = \left\|\nabla \log \frac{\rho}{\pi}\right\|_{L_2(\rho)}^2 = \sum_{i=1}^{\infty} \left|\left\langle \nabla \log \frac{\rho}{\pi}, e_i \right\rangle_{L_2(\rho)}\right|^2,$$

with $(e_i)_{i=1}^{\infty}$ being an orthonormal basis to $L_2(\rho)$. Correspondingly, the Stein-Fisher information is defined as

$$I_{Stein}(\rho|\pi) := \left\|\iota_{k,\rho}^* \nabla \log \frac{\rho}{\pi}\right\|_{\mathcal{H}_k^d}^2 = \left\langle \nabla \log \frac{\rho}{\pi}, \iota_{k,\rho}\iota_{k,\rho}^* \nabla \log \frac{\rho}{\pi} \right\rangle_{L_2^d(\rho)} = \sum_{i=1}^{\infty} \lambda_i \left|\left\langle \nabla \log \frac{\rho}{\pi}, e_i \right\rangle_{L_2(\rho)}\right|^2,$$

where $(\lambda_i, e_i)_{i=1}^{\infty}$ are the set of eigenvalues and eigenvectors of the operator $\iota_{k,\rho}\iota_{k,\rho}^*$, with $\lambda_1 \geq \lambda_2 \geq \cdots > 0$.

Remark 27. Strictly speaking, the above notation implicitly assumes that the operator $\mathcal{T}_{k,\rho}$ has a trivial null space, in which case the $\overline{\operatorname{Ran}(\mathcal{T}_{k,\rho})} \equiv L_2(\rho)$ and hence the eigenfunctions $(e_i)_{i=1}^{\infty}$ form an orthonormal basis to $L_2(\rho)$. However, our analysis does not require this condition on $\mathcal{T}_{k,\rho}$. In particular, if $\mathcal{T}_{k,\rho}$ has a non-trivial null-space, then $\overline{\mathrm{Ran}(\mathcal{T}_{k,\rho})} \subset L_2(\rho)$. In this case, our analysis still holds true. For example, with a slight abuse of notation, if we let e_i, for certain values of i, to also denote the basis of the null-space of $\mathcal{T}_{k,\rho}$, conclusions similar to our results hold.

With this representation for the Fisher information and the Stein-Fisher information, it is immediately clear that the Stein-Fisher information is severely restrictive, in particular when the eigenvalues of the chosen RKHS decay faster. To counter this effect, we introduce the following regularized Stein-Fisher information and show that when the regularization parameter is chosen appropriately, the regularized Stein-Fisher information upper and lower bounds Fisher information.

DEFINITION 7 (Regularized Stein-Fisher Information). For any probability measure ρ, the regularized Stein Fisher information from ρ to π, denoted as $I_{\nu,Stein}(\rho|\pi)$, is defined as

$$I_{\nu,Stein}(\rho|\pi) := \left\langle \iota_{k,\rho}^* \nabla \log \frac{\rho}{\pi}, \left((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\right)^{-1} \iota_{k,\rho}^* \nabla \log \frac{\rho}{\pi} \right\rangle_{\mathcal{H}_k^d}. \tag{4.8}$$

The regularized Stein Fisher information in (4.8) is well-defined because the operator

$$(1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d : \mathcal{H}_k^d \to \mathcal{H}_k^d$$

is positive and for any $f \in \mathcal{H}_k^d$, $\Big((1-\nu)\iota_{k,\rho}^*\iota_{k,\rho} + \nu I_d\Big) f = 0$ if and only if $f = 0$.

REMARK 28. The regularized Stein Fisher information has the following alternative representation:

$$I_{\nu,Stein}(\rho|\pi) = \sum_{i=1}^{\infty} \frac{\lambda_i}{(1-\nu)\lambda_i + \nu} \left| \left\langle \nabla \log \frac{\rho}{\pi}, e_i \right\rangle_{L_2(\rho)} \right|^2. \tag{4.9}$$

For $\nu > 0$, with the fact that λ_i decreases to zero as $i \to \infty$, the regularized Stein Fisher information and the Stein Fisher information both encode the spectral decay information of $\iota_{k,\rho}\iota_{k,\rho}^*$. However, note that the regularized Stein Fisher information tends to the Fisher information as $\nu \to 0$. Hypothetically speaking, if ν is set to zero, then the regularized Stein Fisher information actually becomes the Fisher information. In our analysis, we will take advantage of the relation between the regularized Stein Fisher information and the Fisher information, while studying the convergence properties of R-SVGF under Log-Sobolev inequality assumptions on the target π. A precise relation between the regularized Stein Fisher information and the Fisher information is stated in the following result. Before stating the result, we introduce the following notation for convenience. For $\gamma \in (0, \frac{1}{2}]$, we denote the pre-image of $\nabla \log \frac{\rho}{\pi} \in L_2^d(\rho)$ under $(\iota_{k,\rho}\iota_{k,\rho}^*)^\gamma$ as

$$\mathfrak{I}(\rho, \gamma) := (\iota_{k,\rho}\iota_{k,\rho}^*)^{-\gamma} \nabla \log \frac{\rho}{\pi}.$$

Note that $\|\mathfrak{I}(\rho,\gamma)\|_{L_2^d(\rho)}$ is finite if and only if $\nabla \log \frac{\rho}{\pi} \in \mathrm{Ran}((\iota_{k,\rho}\iota_{k,\rho}^*)^\gamma)$.

Proposition 21 (Equivalence relation between $I(\rho|\pi)$ and $I_{\nu,Stein}(\rho|\pi)$). Let ρ be a probability measure in $\mathbb{R}^d$ such that $I(\rho|\pi)$ and $I_{\nu,Stein}(\rho|\pi)$ are well-defined. Suppose there exists $\gamma \in (0, \frac{1}{2}]$ such that $\|\mathfrak{I}(\rho,\gamma)\|_{L_2^d(\rho)} < \infty$. If the regularization parameter is chosen to satisfy the following condition,

$$\frac{\nu}{1-\nu} \leq \left(\frac{I(\rho|\pi)}{2\left\|\mathfrak{I}(\rho,\gamma)\right\|_{L_2^d(\rho)}^2} \right)^{\frac{1}{2\gamma}}, \tag{4.10}$$

then we have that

$$\frac{1}{2}(1-\nu)^{-1} I(\rho|\pi) \leq I_{\nu,Stein}(\rho|\pi) \leq (1-\nu)^{-1} I(\rho|\pi).$$

Proof of Proposition 21. According to (4.9), we have

$$\begin{aligned} I_{\nu,Stein}(\rho|\pi) &= \sum_{i=1}^{\infty} \frac{\lambda_i}{(1-\nu)\lambda_i + \nu} \left| \left\langle \nabla \log \frac{\rho}{\pi}, e_i \right\rangle_{L_2(\rho)} \right|^2 \\ &\leq (1-\nu)^{-1} \sum_{i=1}^{\infty} \left| \left\langle \nabla \log \frac{\rho}{\pi}, e_i \right\rangle_{L_2(\rho)} \right|^2 \leq (1-\nu)^{-1} I(\mu|\pi). \end{aligned}$$

On the other hand, since $\|\mathfrak{I}(\rho,\gamma)\|_{L_2^d(\rho)} < \infty$ for some $\gamma \in (0, \frac{1}{2}]$, there exists $h = \mathfrak{I}(\rho,\gamma) \in L_2^d(\rho)$ such that

$$\nabla \log \frac{\rho}{\pi} = (\iota_{k,\rho}\iota_{k,\rho}^*)^\gamma h.$$

Therefore

$$(4.11)\qquad (1-\nu)^{-1}I(\rho|\pi) - I_{\nu,Stein}(\rho|\pi) = \sum_{i=1}^{\infty} \frac{(1-\nu)^{-1}\nu}{(1-\nu)\lambda_i+\nu}\left|\langle(\iota_{k,\rho}\iota_{k,\rho}^{*})^{\gamma}h, e_i\rangle_{L_2(\rho)}\right|^2$$
$$= \sum_{i=1}^{\infty} \frac{(1-\nu)^{-1}\nu\lambda_i^{2\gamma}}{(1-\nu)\lambda_i+\nu}\left|\langle h, e_i\rangle_{L_2(\rho)}\right|^2$$
$$\leq (1-\nu)^{-1-2\gamma}\nu^{2\gamma}\left\|\mathfrak{I}(\rho,\gamma)\right\|_{L_2^d(\rho)}^2$$
$$\leq \frac{1}{2}(1-\nu)^{-1}I(\rho|\pi),$$

where the second to last inequality follows from the fact that

$$\sup_i\left(\frac{(1-\nu)^{-1}\nu\lambda_i^{2\gamma}}{(1-\nu)\lambda_i+\nu}\right) = (1-\nu)^{-1-2\gamma}\nu^{2\gamma}\sup_i\left(\frac{(1-\nu)\lambda_i}{(1-\nu)\lambda_i+\nu}\right)^{2\gamma}\left(\frac{\nu}{(1-\nu)\lambda_i+\nu}\right)^{1-2\gamma}$$
$$\leq (1-\nu)^{-1}\nu^{2\gamma},$$

and the last inequality follows from the condition in (4.10). ■

4.5.2. Convergence results for R-SVGF.

Relationship between R-SVGF and WGF. We now provide the relationship between the R-SVGF and the WGF in various metrics. We first start with the relationship in the Fisher information metric, without any stringent assumptions on the target distribution (thereby allowing for multi-modal and complex densities that arise in practice). Note that the Fisher information metric corresponds to the first-order stationarity metric for the WGF obtained by minimizing the KL divergence. This metric has been recently proposed as a meaningful metric to consider in the case of sampling from general non-log-concave densities in [BCE$^+$22]. Note in particular under mild conditions on q (e.g., connected support) that having the Fisher information $I(p|q) = 0$ implies $p \equiv q$. However, even when $I(p|q) \leq \epsilon$, for some $\epsilon > 0$, we have that the modes of the two densities are well-aligned, as argued in [BCE$^+$22].

Theorem 19 (Relation to the WGF in Relative Fisher Information). Let (ρ_t) be the solution to (4.6) and (μ_t) be the solution to the WGF, i.e.,

$$\begin{cases} \partial_t \mu = \nabla \cdot \left(\mu \nabla \log \frac{\mu}{\pi} \right), \\ \mu(0, \cdot) = \mu_0(\cdot). \end{cases} \tag{4.12}$$

For any $t > 0$, suppose there exists $\gamma_t \in (0, \frac{1}{2}]$ such that $\|\mathfrak{I}(\rho_t, \gamma_t)\|_{L_2^d(\rho_t)} < \infty$. Then, for any initial distribution $\mu_0 \in \mathcal{P}(\mathbb{R}_2^d)$, and for any $T \in (0, \infty)$, we have

$$\int_0^T I(\rho_t|\mu_t)dt \le \frac{4}{3}\mathrm{KL}(\rho_0|\mu_0) + \frac{8}{3}\|k\|_\infty^2 \int_0^T \nu^{2\gamma_t} \|\mathfrak{I}(\rho_t, \gamma_t)\|_{L_2^d(\rho_t)}^2 dt. \tag{4.13}$$

Proof of Theorem 19. First note that we have the following upper bound on $\frac{d}{dt}\mathrm{KL}(\rho_t|\mu_t)$:

$$\begin{aligned}
&\frac{d}{dt}\mathrm{KL}(\rho_t|\mu_t) \\
&= \frac{d}{dt}\int_{\mathbb{R}^d} \log \frac{\rho_t(x)}{\mu_t(x)} \rho_t(x) dx \\
&= \int_{\mathbb{R}^d} \partial_t \rho_t(x) \log \frac{\rho_t(x)}{\mu_t(x)} dx + \int_{\mathbb{R}^d} \left(\frac{\partial_t \rho_t(x)}{\mu_t(x)} + \rho_t(x) \partial_t \left(\frac{1}{\mu_t(x)} \right) \right) \frac{\mu_t(x)}{\rho_t(x)} \rho_t(x) dx \\
&= \int_{\mathbb{R}^d} \partial_t \rho_t(x) \log \frac{\rho_t(x)}{\mu_t(x)} dx + \int_{\mathbb{R}^d} \partial_t \rho_t(x) dx - \int_{\mathbb{R}^d} \partial_t \mu_t(x) \frac{\rho_t(x)}{\mu_t(x)} dx \\
&= -\int_{\mathbb{R}^d} \left\langle \iota_{k,\rho_t} \left((1-\nu)\iota_{k,\rho_t}^* \iota_{k,\rho_t} + \nu I_d \right)^{-1} \iota_{k,\rho_t}^* \nabla \log \frac{\rho_t(x)}{\pi(x)}, \nabla \log \frac{\rho_t(x)}{\mu_t(x)} \right\rangle \rho_t(x) dx \\
&\quad + 0 + \int_{\mathbb{R}^d} \left\langle \mu_t(x) \nabla \log \frac{\mu_t(x)}{\pi(x)}, \nabla \left(\frac{\rho_t(x)}{\mu_t(x)} \right) \right\rangle dx
\end{aligned}$$

$$
\begin{aligned}
&= -\int_{\mathbb{R}^d} \left\langle \iota_{k,\rho_t} \left((1-\nu)\iota_{k,\rho_t}^* \iota_{k,\rho_t} + \nu I_d\right)^{-1} \iota_{k,\rho_t}^* \nabla \log \frac{\rho_t(x)}{\pi(x)}, \nabla \log \frac{\rho_t(x)}{\mu_t(x)} \right\rangle \rho_t(x) dx \\
&\quad + \int_{\mathbb{R}^d} \left\langle \nabla \log \frac{\rho_t(x)}{\mu_t(x)}, \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle \rho_t(x) dx \\
&= -\int_{\mathbb{R}^d} \left\langle \nabla \log \frac{\rho_t(x)}{\mu_t(x)}, \nabla \log \frac{\rho_t(x)}{\pi(x)} - \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle \rho_t(x) dx \\
&\quad - \int_{\mathbb{R}^d} \left\langle \left(\iota_{k,\rho_t} \left((1-\nu)\iota_{k,\rho_t}^* \iota_{k,\rho_t} + \nu I_d\right)^{-1} \iota_{k,\rho_t}^* - I\right) \nabla \log \frac{\rho_t(x)}{\pi(x)}, \nabla \log \frac{\rho_t(x)}{\mu_t(x)} \right\rangle \rho_t(x) dx \\
&\leq -\int_{\mathbb{R}^d} \left| \nabla \log \frac{\rho_t(x)}{\mu_t(x)} \right|^2 \rho_t(x) dx \\
&\quad + \frac{1}{4} \int_{\mathbb{R}^d} \left| \nabla \log \frac{\rho_t(x)}{\mu_t(x)} \right|^2 \rho_t(x) dx + 2 \left\| \left(\iota_{k,\rho_t} \left((1-\nu)\iota_{k,\rho_t}^* \iota_{k,\rho_t} + \nu I_d\right)^{-1} \iota_{k,\rho_t}^* - I\right) \nabla \log \frac{\rho_t}{\pi} \right\|_{L_2^d(\rho_t)}^2 \\
&= -\frac{3}{4} I(\rho_t|\mu_t) + \sum_{i=1}^{\infty} \frac{2(1-\lambda_i)^2 \nu^2}{((1-\nu)\lambda_i + \nu)^2} \left| \left\langle \nabla \log \frac{\rho_t}{\pi}, e_i \right\rangle_{L_2(\rho_t)} \right|^2 .
\end{aligned}
$$

In the above calculation, the fourth equality follows by integration-by-parts, the inequality follows by Young's inequality for the inner product (i.e., $\langle p, q \rangle \leq \frac{1}{2} c|p|^2 + \frac{1}{2c}|q|^2$ for any $p, q \in \mathbb{R}^d$) and the last equality follows from the proof of Proposition 21. Since $\nabla \log \frac{\rho_t}{\pi} = (\iota_{k,\rho_t} \iota_{k,\rho_t}^*)^{\gamma_t} h_t$ for some $\gamma_t \in (0, 1/2]$ with $h_t := \mathfrak{I}(\rho_t, \gamma_t) \in L_2^d(\rho_t)$, we obtain

$$
\begin{aligned}
\frac{d}{dt} \mathrm{KL}(\rho_t|\mu_t) &\leq -\frac{3}{4} I(\rho_t|\mu_t) + 2 \left(\max_i \lambda_i \wedge 1 \right)^2 \left(\max_i \frac{\lambda_i^{\gamma_t} \nu}{\lambda_i + \nu} \right)^2 \|h_t\|_{L_2^d(\rho_t)}^2 \\
&\leq -\frac{3}{4} I(\rho_t|\mu_t) + 2 \|k\|_\infty^2 \nu^{2\gamma_t} \|\mathfrak{I}(\rho_t, \gamma_t)\|_{L_2^d(\rho_t)}^2 ,
\end{aligned}
$$

where the last inequality follows from the fact that $\max_i \lambda_i = \left\| \iota_{k,\rho_t} \iota_{k,\rho_t}^* \right\|_{L_2(\rho_t) \to L_2(\rho_t)} \leq \|k\|_\infty$. Integrating from $t = 0$ to $t = T$, we get

$$
\mathrm{KL}(\rho_T|\mu_T) - \mathrm{KL}(\rho_0|\mu_0) \leq -\frac{3}{4} \int_0^T I(\rho_t|\mu_t) dt + 2 \|k\|_\infty^2 \int_0^T \nu^{2\gamma_t} \|\mathfrak{I}(\rho_t, \gamma_t)\|_{L_2^d(\rho_t)}^2 dt.
$$

Since KL-divergence is non-negative, (4.13) is proved. ■

REMARK 29. The above result shows that as long as $\rho_0 = \mu_0$, i.e., both the WGF and R-SVGF are initialized with the same density, and ν is chosen such that $T^{-1} \int_0^T \nu^{2\gamma_t} \|\mathfrak{I}(\rho_t, \gamma_t)\|_{L_2^d(\rho_t)}^2 dt \to 0$, the *averaged* Fisher information along the path tends to zero. This shows the benefit of regularizing the

SVGF – it enables one to closely approximate the WGF with appropriate choice of the regularization parameters.

Convergence to Equilibrium along the Fisher Information. We now provide results on the convergence to equilibrium along the Fisher information for the R-SVGF. We re-emphasize here that our result below is provided without any assumptions on the target π.

THEOREM 20 (**Convergence of Fisher information**). Let (ρ_t) be the solution to (4.6). For any $t > 0$, supppose there exists $\gamma_t \in (0, \frac{1}{2}]$ such that $\|\mathfrak{I}(\rho_t, \gamma_t)\|_{L_2^d(\rho_t)} < \infty$. Then

$$\int_0^\infty I(\rho_t|\pi)dt \leq (1-\nu)\mathrm{KL}(\rho_0|\pi) + \int_0^\infty \nu^{2\gamma_t}(1-\nu)^{-2\gamma_t} \|\mathfrak{I}(\rho_t, \gamma_t)\|^2_{L_2^d(\rho_t)} \, dt.$$

Furthermore, if $\int_0^\infty \nu^{2\gamma_t}(1-\nu)^{-2\gamma_t} \|\mathfrak{I}(\rho_t, \gamma_t)\|^2_{L_2^d(\rho_t)} \, dt < \infty$, then we get $I(\rho_t|\pi) \to 0$ as $t \to \infty$.

Before proving the above theorem, we introduce a few intermediate results.

PROPOSITION 22 (Decay of the KL-divergence). For the solution $(\rho_t)_{t\geq 0}$ to the PDE (4.6), it holds that

$$\frac{d}{dt}\mathrm{KL}(\rho_t|\pi) \leq 0. \tag{4.14}$$

and consequently

$$\frac{d}{dt}\mathrm{KL}(\rho_t|\pi) = -I_{\nu,Stein}(\rho_t|\pi). \tag{4.15}$$

PROOF OF PROPOSITION 22. Note that

$$\begin{aligned}
\frac{d}{dt}\mathrm{KL}(\rho_t|\pi) &= \frac{d}{dt}\int_{\mathbb{R}^d}\rho_t \log\frac{\rho_t}{\pi}dx \\
&= \int_{\mathbb{R}^d}\partial_t\rho_t\log\frac{\rho_t}{\pi}dx + \int_{\mathbb{R}^d}\partial_t\rho_t dx \\
&= -\int_{\mathbb{R}^d}\left\langle \nabla\log\frac{\rho_t}{\pi}(x), \iota_{k,\rho_t}\left((1-\nu)\iota_{k,\rho_t}^*\iota_{k,\rho_t} + \nu I_d\right)^{-1}\iota_{k,\rho_t}^*\left(\nabla\log\frac{\rho_t}{\pi}\right)(x)\right\rangle \rho_t(x)dx + 0 \\
&= -\left\langle \nabla\log\frac{\rho_t}{\pi}, \iota_{k,\rho_t}\left((1-\nu)\iota_{k,\rho_t}^*\iota_{k,\rho_t} + \nu I_d\right)^{-1}\iota_{k,\rho_t}^*\left(\nabla\log\frac{\rho_t}{\pi}\right)\right\rangle_{L_2^d(\rho_t)} \\
&= -\left\langle \iota_{k,\rho_t}^*\nabla\log\frac{\rho_t}{\pi}, \left((1-\nu)\iota_{k,\rho_t}^*\iota_{k,\rho_t} + \nu I_d\right)^{-1}\iota_{k,\rho_t}^*\left(\nabla\log\frac{\rho_t}{\pi}\right)\right\rangle_{\mathcal{H}_k^d}.
\end{aligned}$$

It suffices to show that for all $\nu > 0$, $\left((1-\nu)\iota_{k,\rho_t}^*\iota_{k,\rho_t} + \nu I_d\right)^{-1}$ is a positive operator from $\mathcal{H}_k^d$ to $\mathcal{H}_k^d$. By the definition of ι_{k,ρ_t}, for any $f \in \mathcal{H}_k^d$ with $\|f\|_{\mathcal{H}_k^d} = 1$,

$$\begin{aligned}
\langle f, \left((1-\nu)\iota_{k,\rho_t}^*\iota_{k,\rho_t} + \nu I_d\right) f\rangle_{\mathcal{H}_k^d} &= (1-\nu)\langle \iota_{k,\rho_t} f, \iota_{k,\rho_t} f\rangle_{L_2^d(\rho_t)} + \nu\|f\|_{\mathcal{H}_k^d}^2 \\
&= (1-\nu)\|\iota_{k,\rho_t} f\|_{L_2^d(\rho_t)}^2 + \nu > 0
\end{aligned}$$

for all $\nu > 0$. Therefore $(1-\nu)\iota_{k,\rho_t}^*\iota_{k,\rho_t} + \nu I_d$ is a positive operator from $\mathcal{H}_k^d$ to $\mathcal{H}_k^d$. So is the operator $\left((1-\nu)\iota_{k,\rho_t}^*\iota_{k,\rho_t} + \nu I_d\right)^{-1}$. Hence, we have (4.14). The claim in (4.15) follows directly from (4.14), (4.5) and Definition 7. ■

PROOF OF THEOREM 20. From Proposition 22 and (4.11), we know that

$$\frac{d}{dt}\mathrm{KL}(\rho_t|\pi) = -(1-\nu)^{-1}I(\rho_t|\pi) + \sum_{i=1}^{\infty}\frac{(1-\nu)^{-1}\nu}{(1-\nu)\lambda_i + \nu}\left|\left\langle\nabla\log\frac{\rho_t}{\pi}, e_i\right\rangle_{L_2(\rho_t)}\right|^2,$$

where $\nabla \log \frac{\rho_t}{\pi} = (\iota_{k,\rho_t}\iota_{k,\rho_t}^*)^{\gamma_t} h_t$ for some $\gamma_t \in (0, \frac{1}{2}]$ with $h_t := \mathfrak{I}(\rho_t, \gamma_t) \in L_2^d(\rho_t)$. Therefore we have,

$$\begin{aligned}
&\frac{d}{dt}\mathrm{KL}(\rho_t|\pi) \\
=& -(1-\nu)^{-1} I(\rho_t|\pi) + \sum_{i=1}^{\infty} \frac{(1-\nu)^{-1}\nu}{(1-\nu)\lambda_i + \nu} \left| \left\langle (\iota_{k,\rho_t}\iota_{k,\rho_t}^*)^{\gamma_t} h_t, e_i \right\rangle_{L_2(\rho_t)} \right|^2 \\
=& -(1-\nu)^{-1} I(\rho_t|\pi) + \sum_{i=1}^{\infty} \frac{(1-\nu)^{-1}\nu\lambda_i^{2\gamma_t}}{(1-\nu)\lambda_i + \nu} \left| \langle h_t, e_i \rangle_{L_2(\rho_t)} \right|^2 \\
=& \sum_{i=1}^{\infty} (1-\nu)^{-1-2\gamma_t} \nu^{2\gamma_t} \left(\frac{(1-\nu)\lambda_i}{(1-\nu)\lambda_i + \nu} \right)^{2\gamma_t} \left(\frac{\nu}{(1-\nu)\lambda_i + \nu} \right)^{1-2\gamma_t} \left| \langle h_t, e_i \rangle_{L_2(\rho_t)} \right|^2 \\
&\qquad - (1-\nu)^{-1} I(\rho_t|\pi) \\
\leq& -(1-\nu)^{-1} I(\rho_t|\pi) + (1-\nu)^{-1-2\gamma_t} \nu^{2\gamma_t} \left\| \mathfrak{I}(\rho_t, \gamma_t) \right\|_{L_2^d(\rho_t)}^2 .
\end{aligned}$$

The result follows by integrating over t and noting that the KL-divergence is non-negative. Now, with ρ_t denoting the solution to (9), we have that $I(\rho_t|\pi)$ is non-negative and continuous in t. The claim of convergence holds because for a continuous function h, if we have that $\int_0^\infty h(t)dt < \infty$, then we have $h(t) \to 0$ as $t \to \infty$. ■

Convergence in KL-divergence under LSI. While the previous result was provided without any further assumptions on the target density $\pi \in \mathcal{P}(\mathbb{R}_2^d)$, in this section, we provide improved convergence results of the R-SVGF under the assumption that the π satisfies the Log-Sobolev Inequality. Recall that we say that $\pi \in \mathcal{P}(\mathbb{R}^d)$ satisfies the Log-Sobolev inequality with constant $\lambda > 0$ if for all $\mu \in \mathcal{P}(\mathbb{R}^d)$:

$$\mathrm{KL}(\mu|\pi) \leq \frac{1}{2\lambda} I(\mu|\pi).$$

Our first result below is a stronger version of the result in Theorem 19, under the assumption that the target π satisfies LSI and Assumption 4.5.1 on the initialization of the WGF.

ASSUMPTION 4.5.1. *The initial density μ_0 is chosen so that the solution (μ_t) to* (4.12) *also satisfies LSI with parameter λ, for all $t > 0$.*

Under the stronger assumption that the target density π is strongly log-concave, following the arguments in [VW19, Theorem 8], it is easy to show that Assumption 4.5.1 is satisfied as long as μ_0 is chosen such that it satisfies LSI. We conjecture that the same holds true even when the target density satisfies LSI and additional mild smoothnes assumptions (i.e., LSI is preserved along the trajectory as long as the initial density μ_0 satisfies LSI, presumably with additional milder assumptions). However, a proof of this conjecture has eluded us thus far.

Theorem 21 (Relation to the WGF under LSI). Assume π satisfies the log-Sobolev inequality with parameter λ. Let (ρ_t) be the solution to (4.6). Let (μ_t) be the solution to WGF, defined in (4.12), with μ_0 satisfying Assumption 4.5.1. For any $t > 0$, suppose there exists $\gamma_t \in (0, \frac{1}{2}]$ such that $\|\mathfrak{I}(\rho_t, \gamma_t)\|_{L_2^d(\rho_t)} < \infty$. Then, for any $T \in (0, \infty)$, we have

$$\mathrm{KL}(\rho_T|\mu_T) \leq e^{-3\lambda T/2}\mathrm{KL}(\rho_0|\mu_0) + 2\left\|k\right\|_\infty^2 \int_0^T \nu^{2\gamma_t} e^{-3\lambda(T-t)/2} \left\|\mathfrak{I}(\rho_t, \gamma_t)\right\|_{L_2^d(\rho_t)}^2 dt. \tag{4.16}$$

Proof of Theorem 21. Following the same arguments as in the proof of Theorem 19, we obtain that for any $t > 0$,

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(\rho_t|\mu_t) &\leq -\frac{3}{4}I(\rho_t|\mu_t) + \sum_{i=1}^{\infty} \frac{2(1-\lambda_i)^2\nu^2}{((1-\nu)\lambda_i + \nu)^2} \left|\left\langle \nabla \log \frac{\rho_t}{\pi}, e_i\right\rangle_{L_2(\rho_t)}\right|^2 \\ &\leq -\frac{3}{4}I(\rho_t|\mu_t) + 2\left\|k\right\|_\infty^2 \nu^{2\gamma_t} \left\|\mathfrak{I}(\rho_t, \gamma_t)\right\|_{L_2^d(\rho_t)}^2.\end{aligned}$$

Hence, under Assumption 4.5.1 we obtain

$$\frac{d}{dt}\mathrm{KL}(\rho_t|\mu_t) \leq -\frac{3\lambda}{2}\mathrm{KL}(\rho_t|\mu_t) + 2\left\|k\right\|_\infty^2 \nu^{2\gamma_t} \left\|\mathfrak{I}(\rho_t, \gamma_t)\right\|_{L_2^d(\rho_t)}^2.$$

Finally, (4.16) follows from the Gronwall's inequality. ■

Our second result is a stronger version of the result in Theorem 20, under the assumption that the target distribution π satisfies LSI. We remark that convergence to equilibrium of the related WGF under various functional inequalities is a well-studied topic. We refer the interested reader to [BGL14] for a detailed overview.

Theorem 22 (Decay of KL-divergence under LSI). Assume that π satisfies the log-Sobolev inequality with $\lambda > 0$. Let (ρ_t) be the solution to (4.6). For any $t > 0$, suppose there exists $\gamma_t \in (0, \frac{1}{2}]$ such that $\|\mathfrak{I}(\rho_t, \gamma_t)\|_{L_2^d(\rho_t)} < \infty$. Then, for any $T \in (0, \infty)$, we have

$$\mathrm{KL}(\rho_T|\pi) \leq e^{-2(1-\nu)^{-1}\lambda T}\mathrm{KL}(\rho_0|\pi) + \int_0^T \nu^{2\gamma_t}(1-\nu)^{-2\gamma_t - 1} \|\mathfrak{I}(\rho_t, \gamma_t)\|^2_{L_2^d(\rho_t)} e^{2(1-\nu)^{-1}\lambda(t-T)} dt.$$

Proof of Theorem 22. From the proof of Theorem 20, we have

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(\rho_t|\pi) &\leq -(1-\nu)^{-1} I(\rho_t|\pi) + (1-\nu)^{-1-2\gamma_t}\nu^{2\gamma_t} \|\mathfrak{I}(\rho_t, \gamma_t)\|^2_{L_2^d(\rho_t)} \\ &\leq -2(1-\nu)^{-1}\lambda \mathrm{KL}(\rho_t|\pi) + (1-\nu)^{-1-2\gamma_t}\nu^{2\gamma_t} \|\mathfrak{I}(\rho_t, \gamma_t)\|^2_{L_2^d(\rho_t)},\end{aligned}$$

where the last inequality follows the log-Sobolev inequality. The final statement follows from Gronwall's inequality. ■

Remark 30 (Exponential Decay of KL-divergence). Yet another way to state the above result is via the introducing the following regularized Stein-LSI, similar to the introduction of Stein-LSI in [DNS19]. However, the introduction of Stein-LSI is quite restrictive in the sense that it couples assumptions on the target and the chosen RKHS. This makes verifying the conditions more delicate. To counter this effect, we now introduce the notion of Regularized Stein-LSI. We say that $\pi \in \mathcal{P}(\mathbb{R}^d)$ satisfies the regularized Stein log-Sobolev inequality with constant $\lambda > 0$ if for all $\mu \in \mathcal{P}(\mathbb{R}^d)$:

$$\mathrm{KL}(\mu|\pi) \leq \frac{1}{2\lambda} I_{\nu,Stein}(\mu|\pi). \tag{4.17}$$

An advantage of the above condition is that, as $\nu \to 0$ the regularized Stein-LSI inequality becomes equivalent to the standard LSI inequality. Under the condition that the target density π satisfies (4.17), and letting $(\rho_t)_{t\geq 0}$ be the solution to (4.6), it holds that

$$\mathrm{KL}(\rho_t|\pi) \leq e^{-2\lambda t}\mathrm{KL}(\rho_0|\pi). \tag{4.18}$$

The proof of (4.18) follows immediately from Proposition 22 and (4.17).

4.5.3. Convergence results for Time-discretized R-SVGF. In this section we analyze the convergence properties of the time-discretized R-SVGF in (4.7). To do so, we require the following additional assumptions.

Assumption A2. The following conditions hold:

(1) There exists a constant $B > 0$ such that $\|\nabla_1 k(x,\cdot)\|_{\mathcal{H}_k^d} \le B$ for all $x \in \mathbb{R}^d$.

(2) The potential function $V : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable and gradient Lipschitz with parameter L.

(3) Along the population limit (4.7), $I(\rho^n|\pi) < \infty$ for all fixed $n \ge 0$.

The smoothness assumptions in points (1) and (2) of Assumption A2 are commonly required in analyzing any discrete-time algorithms, albeit deterministic [KSA$^+$20, SSR22] or randomized [VW19, CEL$^+$21, BCE$^+$22]. While it could be relaxed (see, for example, [SR22]), in general it is impossible to completely avoid them as in the case of analyzing the corresponding flows. Before stating our results, we also introduce some convenient notations. We let

$$\mathfrak{S}_n := \left(\sup_i \frac{{\lambda_i^{(n)}}^{1+2\gamma_n}}{\left((1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}\right)^2} \right) \quad \text{and} \quad R_n := \|\mathfrak{I}(\rho^n, \gamma_n)\|_{L_2^d(\rho^n)},$$

where the sequences $\{\lambda_i^{(n)}\}$ corresponds to the positive eigenvalues of the operator $\iota_{k,\rho^n}\iota_{k,\rho^n}^*$ in the order of decreasing values.

Theorem 23 (Convergence in Fisher Divergence). Suppose Assumption A2 holds. Let (ρ^n) be the time discretization of the R-SVGF described in (4.7) with initial condition $\rho^0 = \rho_0$ such that $\mathrm{KL}(\rho_0|\pi) \le R$. For each n, suppose that ν_{n+1} and the step-size h_{n+1} are chosen such that,

$$h_{n+1} < \min\left\{ \frac{1-\nu_{n+1}}{L}, \frac{\alpha - 1}{\alpha B R_n \sqrt{\mathfrak{S}_n}} \right\}, \tag{4.19}$$

where $\alpha \in (1,2)$ is some constant, and suppose that there exists $\gamma_n \in (0, \frac{1}{2}]$, such that $\mathfrak{I}(\rho^n, \gamma_n) \in L_2^d(\rho^n)$. Then,

$$\text{(4.20)} \quad \sum_{n=0}^{\infty} \frac{h_{n+1}}{2(1-\nu_{n+1})} I(\rho^n|\pi) \leq \sum_{n=0}^{\infty} \nu_{n+1}^{2\gamma_n} (1-\nu_{n+1})^{-2\gamma_n - 1} h_{n+1} \left(1 + \frac{1}{2}\nu_{n+1}^{-1} \alpha^2 B^2 h_{n+1}\right) R_n^2 + R.$$

Before proving Theorem 23, we first prove the following intermediate result.

LEMMA 4.5.1. *For each $n \geq 1$, define $g = \mathcal{D}_{\nu_{n+1},\rho^n} \nabla \log \frac{\rho^n}{\pi}$. Under the conditions in Theorem 23, we have that, for any $x \in \mathbb{R}^d$ and $t \in [0, h_{n+1}]$,*

$$\|\nabla g(x)\|_{HS}^2 \leq B^2 R_n^2 \mathfrak{S}_n \text{ and } \left\|(id - t\nabla g(x))^{-1}\right\|_2 \leq \alpha.$$

PROOF OF LEMMA 4.5.1. Since for each n, there exists $\gamma_n \in (0, 1/2]$ and a function $h = \mathfrak{I}(\rho^n, \gamma_n) \in L_2^d(\rho^n)$ such that $(\iota_{k,\rho^n} \iota_{k,\rho^n}^*)^{2\gamma_n} h_j = \partial_j \log \frac{\rho^n}{\pi}$, where h_j is the j-th component of the function value of h, we have

$$\begin{aligned}
\|\nabla g(x)\|_{HS}^2 &= \sum_{j,l=1}^{d} \left|\frac{\partial g_j(x)}{\partial x_l}\right|^2 \\
&= \sum_{j,l=1}^{d} \left(\sum_{i=1}^{\infty} \frac{\lambda_i^{(n)}}{(1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}} \langle \partial_j \log \frac{\rho^n}{\pi}, e_i \rangle_{L_2(\rho^n)} \partial_l e_i(x) \right)^2 \\
&= \sum_{j,l=1}^{d} \left(\sum_{i=1}^{\infty} \frac{{\lambda_i^{(n)}}^{1+\gamma_n}}{(1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}} \langle h_j, e_i \rangle_{L_2(\rho^n)} \partial_l e_i(x) \right)^2 \\
&\leq \sum_{j,l=1}^{d} \left(\sum_{i=1}^{\infty} \langle h_j, e_i \rangle_{L_2(\rho^n)}^2 \right) \left(\sum_{i=1}^{\infty} \frac{{\lambda_i^{(n)}}^{2+2\gamma_n}}{\left((1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}\right)^2} |\partial_l e_i(x)|^2 \right) \\
&= \left(\sum_{i=1}^{\infty} \left| \langle h, e_i \rangle_{L_2(\rho^n)} \right|^2 \right) \left(\sum_{i=1}^{\infty} \frac{{\lambda_i^{(n)}}^{2+2\gamma_n}}{\left((1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}\right)^2} |\nabla e_i(x)|^2 \right) \\
&\leq \sup_i \left(\frac{{\lambda_i^{(n)}}^{1+2\gamma_n}}{\left((1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}\right)^2} \right) \|\nabla_1 k(x, \cdot)\|_{\mathcal{H}_k^d}^2 R_n^2 \\
&\leq B^2 R_n^2 \sup_i \left(\frac{{\lambda_i^{(n)}}^{1+2\gamma_n}}{\left((1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}\right)^2} \right).
\end{aligned}$$

In the above, the first inequality follows from Cauchy-Schwartz inequality, the second inequality follows from the fact that

$$\sum_{i=1}^{\infty} |\nabla e_i(x)|^2 = \sum_{i=1}^{\infty} \langle \nabla_1 k(x,\cdot), e_i \rangle_{\mathcal{H}_k^d}^2 = \|\nabla_1 k(x,\cdot)\|_{\mathcal{H}_k}^2 ,$$

and the last inequality follows from Assumption A2. Meanwhile, since $\|\nabla g(x)\|_2 \leq \|\nabla g(x)\|_{HS}$ for all $x \in \mathbb{R}^d$, we have for every $t \in [0, h_{n+1}]$,

$$\begin{aligned}
\left\|(id - t\nabla g(x))^{-1}\right\|_2 &\leq \sum_{m=0}^{\infty} \|t\nabla g(x)\|_2^m \leq \sum_{m=0}^{\infty} \|t\nabla g(x)\|_{HS}^m \\
&\leq \sum_{m=0}^{\infty} \left(h_{n+1} B R_n \sup_i \left(\frac{\lambda_i^{(n)\,1+2\gamma_n}}{\left((1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}\right)^2} \right)^{\frac{1}{2}} \right)^m \\
&\leq \sum_{m=0}^{\infty} \left(\frac{\alpha - 1}{\alpha} \right)^m = \alpha.
\end{aligned}$$

where the last inequality follows from (4.19). ■

PROOF OF THEOREM 23. We start from studying the single step along (4.7). In the following analysis, for each $n \geq 1$, we denote $g = \mathcal{D}_{\nu_{n+1},\rho^n} \nabla \log \frac{\rho^n}{\pi}$, $\phi_t(x) = x - tg(x)$ for all $x \in \mathbb{R}^d$, $t \in [0, h_{n+1}]$ and $\tilde{\rho}_t = (\phi_t)_{\#}\rho^n$. Therefore we have

$$\rho^n = \tilde{\rho}_0 \quad \text{and} \quad \rho^{n+1} = (\phi_{h_{n+1}})_{\#}\rho^n = \tilde{\rho}_{h_{n+1}}.$$

The following analysis is motivated by [SSR22, Proposition 3.1]. According to [Vil21, Theorem 5.34], the velocity field ruling the evolution of $\tilde{\rho}_t$ is $\omega_t \in L_2^d(\tilde{\rho}_t)$ and $\omega_t(x) = -g(\phi_t^{-1}(x))$. Define $\psi(t) = \mathrm{KL}(\tilde{\rho}_t|\pi)$, according to the chain rule in [Vil21, section 8.2],

$$\begin{aligned}
\psi'(t) &= \langle \nabla_{W_2} \mathrm{KL}(\tilde{\rho}_t|\pi), \omega_t \rangle_{L_2^d(\tilde{\rho}_t)} , \\
\psi''(t) &= \left\langle \omega_t, \mathrm{Hess}_{\mathrm{KL}(\cdot|\pi)}(\tilde{\rho}_t)\omega_t \right\rangle_{L_2^d(\tilde{\rho}_t)} .
\end{aligned}$$

where $\mathrm{Hess}_{\mathrm{KL}(\cdot|\pi)}(\tilde{\rho}_t)$ is the Wasserstein Hessian of $\mathrm{KL}(\cdot|\pi)$ at $\tilde{\rho}_t$. For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and any v in the Wasserstein tangent space at μ, the Wasserstein Hessian is given by

$$\left\langle v, \mathrm{Hess}_{\mathrm{KL}(\cdot|\pi)}(\mu) v\right\rangle_{L_2^d(\mu)} = \left\langle v, \nabla^2 V v\right\rangle_{L_2^d(\mu)} + \mathbb{E}_\mu[\|\nabla v(X)\|_{HS}^2].$$

Therefore we can expand the difference in KL-divergence between the two consecutive iterations as

$$\begin{aligned}
&\psi(h_{n+1}) - \psi(0) \\
&= \psi'(0)h_{n+1} + \int_0^{h_{n+1}} (h_{n+1} - t)\psi''(t)dt \\
&= -h_{n+1} \left\langle \nabla_{W_2} \mathrm{KL}(\rho^n|\pi), g\right\rangle_{L_2^d(\rho^n)} + \int_0^{h_{n+1}} (h_{n+1} - t)\langle \omega_t, \mathrm{Hess}_{\mathrm{KL}(\cdot|\pi)}(\tilde{\rho}_t)\omega_t\rangle_{L_2^d(\tilde{\rho}_t)} dt. \qquad (4.21)
\end{aligned}$$

The first term on the right hand side of (4.21) can be studied via the spectrum of the operator $\iota_{k,\rho^n}\iota_{k,\rho^n}^*$.

$$\begin{aligned}
&- h_{n+1} \left\langle \nabla_{W_2} \mathrm{KL}(\rho^n|\pi), g\right\rangle_{L_2^d(\rho^n)} \\
&= -h_{n+1} \left\langle \nabla \log \frac{\rho^n}{\pi}, \left((1-\nu_{n+1})\iota_{k,\rho^n}^*\iota_{k,\rho^n} + \nu_{n+1} I_d\right)^{-1} \iota_{k,\rho^n}^* \nabla \log \frac{\rho^n}{\pi}\right\rangle_{L_2^d(\rho^n)} \\
&= -h_{n+1} \sum_{i=1}^{\infty} \frac{\lambda_i^{(n)}}{(1-\nu_{n+1})\lambda_i^{(n)} + \nu_{n+1}} \left| \left\langle \nabla \log \frac{\rho^n}{\pi}, e_i^{(n)} \right\rangle_{L_2(\rho^n)} \right|^2 \\
&= -h_{n+1} I_{\nu_{n+1}, Stein}(\rho^n|\pi).
\end{aligned}$$

Since $\tilde{\rho}_t = (\phi_t)_\# \rho^n$, for any function h we have $\mathbb{E}_{X\sim\tilde{\rho}_t}\left[h(X)\right] = \mathbb{E}_{Y\sim\rho^n}\left[h(\phi_t(Y))\right]$. Hence, for the second term on the right side of (4.21), we obtain

$$\begin{aligned}
\left\langle \omega_t, \mathrm{Hess}_{\mathrm{KL}(\cdot|\pi)}(\tilde{\rho}_t)\omega_t\right\rangle_{L_2^d(\tilde{\rho}_t)} &= \left\langle \omega_t, \nabla^2 V \omega_t\right\rangle_{L_2^d(\tilde{\rho}_t)} + \mathbb{E}_{\tilde{\rho}_t}[\|\nabla \omega_t(x)\|_{HS}^2] \\
&= \left\langle g(\phi_t^{-1}), \nabla^2 V g(\phi_t^{-1})\right\rangle_{L_2^d(\tilde{\rho}_t)} + \mathbb{E}_{\rho^n}[\|\nabla \omega_t \circ \phi_t(x)\|_{HS}^2] \\
&= \mathbb{E}_{\rho^n}\left[g(x)^T \nabla V^2(\phi_t(x)) g(x)\right] + \mathbb{E}_{\rho^n}\left[\left\|\nabla g(x)(\nabla\phi_t(x))^{-1}\right\|_{HS}^2\right] \\
&\leq L\left\|g\right\|_{L_2^d(\rho^n)}^2 + \mathbb{E}_{\rho^n}\left[\left\|\nabla g(x)(\nabla\phi_t(x))^{-1}\right\|_{HS}^2\right],
\end{aligned}$$

where the last inequality follows from Assumption A2-(2). Therefore we obtain

$$\begin{aligned}\mathrm{KL}(\rho^{n+1}|\pi) - \mathrm{KL}(\rho^n|\pi) \leq -h_{n+1} I_{\nu_{n+1},Stein}(\rho^n|\pi) + \frac{Lh_{n+1}^2}{2}\|g\|^2_{L_2^d(\rho^n)} \\ + \frac{h_{n+1}^2}{2}\max_{t\in[0,h_{n+1}]}\mathbb{E}_{\rho^n}\left[\left\|\nabla g(x)(\nabla\phi_t(x))^{-1}\right\|^2_{HS}\right],\end{aligned}$$

where

$$\begin{aligned}\|g\|^2_{L_2^d(\rho^n)} &= \left\|\mathcal{D}_{\nu_{n+1},\rho^n}\nabla\log\frac{\rho^n}{\pi}\right\|^2_{L_2^d(\rho^n)} \\ &= \left\|\left((1-\nu_{n+1})\iota_{k,\rho^n}\iota^*_{k,\rho^n} + \nu_{n+1}I_d\right)^{-1}\iota_{k,\rho^n}\iota^*_{k,\rho^n}\nabla\log\frac{\rho^n}{\pi}\right\|^2_{L_2^d(\rho^n)} \\ &= \sum_{i=1}^d\left(\frac{\lambda_i^{(n)}}{(1-\nu_{n+1})\lambda_i^{(n)}+\nu_{n+1}}\right)^2\left|\left\langle\nabla\log\frac{\rho^n}{\pi}, e_i^{(n)}\right\rangle_{L_2(\rho^n)}\right|^2 \\ &\leq (1-\nu_{n+1})^{-2} I(\rho_n|\pi),\end{aligned}$$

with $(\lambda_i^{(n)}, e_i^{(n)})_{i=1}^\infty$ being the sequence of eigenvalues and eigenvectors to the operator $\iota_{k,\rho^n}\iota^*_{k,\rho^n}$ such that $\lambda_1^{(n)} \geq \cdots \geq \lambda_i^{(n)} \geq \cdots > 0$ and $(e_i^{(n)})_{i=1}^\infty$ is an orthonormal basis of $L_2(\rho^n)$. According to Lemma 4.5.1 and Assumption A2,

$$\begin{aligned}&\|\nabla g(x)\|^2_{HS} \\ \leq &\sup_i\left(\frac{{\lambda_i^{(n)}}^{1+2\gamma_n}}{\left((1-\nu_{n+1})\lambda_i^{(n)}+\nu_{n+1}\right)^2}\right)B^2R_n^2 \\ \leq &\sup_i\left(\frac{\nu_{n+1}^{2\gamma_n-1}}{(1-\nu_{n+1})^{2\gamma_n+1}}\left(\frac{(1-\nu_{n+1})\lambda_i^{(n)}}{(1-\nu_{n+1})\lambda_i^{(n)}+\nu_{n+1}}\right)^{1+2\gamma_n}\left(\frac{\nu_{n+1}}{(1-\nu_{n+1})\lambda_i^{(n)}+\nu_{n+1}}\right)^{1-2\gamma_n}\right)B^2R_n^2 \\ \leq &\nu_{n+1}^{2\gamma_n-1}(1-\nu_{n+1})^{-2\gamma_n-1}B^2R_n^2,\end{aligned}$$

and furthermore according to Lemma 4.5.1, $\left\|(id - t\nabla g(x))^{-1}\right\|_2^2 \leq \alpha^2$. Therefore we get

$$\begin{aligned}
\mathrm{KL}(\rho^{n+1}|\pi) - \mathrm{KL}(\rho^n|\pi) &\leq -h_{n+1} I_{\nu_{n+1},Stein}(\rho^n|\pi) + \frac{Lh_{n+1}^2(1-\nu_{n+1})^{-2}}{2} I(\rho^n|\pi) \\
&\quad + \frac{1}{2}\alpha^2 B^2 \nu_{n+1}^{2\gamma_n - 1}(1-\nu_{n+1})^{-2\gamma_n - 1} R_n^2 h_{n+1}^2 \\
&\leq -h_{n+1}(1-\nu_{n+1})^{-1}\left(1 - \frac{Lh_{n+1}(1-\nu_{n+1})^{-1}}{2}\right) I(\rho^n|\pi) \\
&\quad + h_{n+1}\nu_{n+1}^{2\gamma_n}(1-\nu_{n+1})^{-2\gamma_n - 1} R_n^2 \left(1 + \frac{1}{2}h_{n+1}\nu_{n+1}^{-1}\alpha^2 B^2\right) \\
&\leq -\frac{1}{2}h_{n+1}(1-\nu_{n+1})^{-1} I(\rho^n|\pi) \\
&\quad + h_{n+1}\nu_{n+1}^{2\gamma_n}(1-\nu_{n+1})^{-2\gamma_n - 1} R_n^2 \left(1 + \frac{1}{2}h_{n+1}\nu_{n+1}^{-1}\alpha^2 B^2\right).
\end{aligned}$$

where the last inequality follows from (4.19) and the second inequality follows from the fact that

$$I(\rho^n|\pi) - I_{\nu_{n+1},Stein}(\rho^n|\pi) \leq \nu_{n+1}^{2\gamma_n}(1-\nu_{n+1})^{-2\gamma_n - 1} R_n^2,$$

which is proved in Proposition 21. Lastly, summing over n and we obtain

$$\begin{aligned}
\sum_{n=0}^{\infty} \frac{h_{n+1}}{2(1-\nu_{n+1})} I(\rho^n|\pi) &\leq \sum_{n=0}^{\infty}\left(\mathrm{KL}(\rho^n|\pi) - \mathrm{KL}(\rho^{n+1})\right) \\
&\quad + \sum_{n=0}^{\infty} h_{n+1}\nu_{n+1}^{2\gamma_n}(1-\nu_{n+1})^{-2\gamma_n - 1} R_n^2 \left(1 + \frac{1}{2}h_{n+1}\nu_{n+1}^{-1}\alpha^2 B^2\right) \\
&\leq \mathrm{KL}(\rho^0|\pi) + \sum_{n=0}^{\infty} h_{n+1}\nu_{n+1}^{2\gamma_n}(1-\nu_{n+1})^{-2\gamma_n - 1} R_n^2 \left(1 + \frac{1}{2}h_{n+1}\nu_{n+1}^{-1}\alpha^2 B^2\right),
\end{aligned}$$

where the last inequality follows from the fact that KL divergence is non-negative. Therefore (4.20) is proved. ■

Remark 31. We emphasize that the above result does not make any assumptions on the target density π, except for the Lipschitz gradient assumption. In particular, it holds for multi-modal densities. However, the metric of convergence is the weaker Fisher information metric.

We now provide a stronger result under LSI assumptions.

Theorem 24. Suppose Assumption A2 holds and π satisfies the log-Sobolev inequality with parameter λ. Let (ρ^n) be as described in (4.7) with initial condition $\rho^0 = \rho_0$ such that $\mathrm{KL}(\rho_0|\pi) \le R$. Assume the regularization parameter and the step-size parameters are chosen such that for all $n \ge 0$, they satisfy

$$h_{n+1} < \min\left\{\frac{1-\nu_{n+1}}{L}, \frac{\alpha-1}{\alpha B R_n \sqrt{\mathfrak{S}_n}}, \frac{2(1-\nu_{n+1})}{\lambda}\right\},\ \frac{\nu_{n+1}}{1-\nu_{n+1}} \le \left(\frac{I(\rho^n|\pi)}{R_n^2}\right)^{\frac{1}{2\gamma_n}}, \tag{4.22}$$

where $\alpha \in (1,2)$ is a constant, $\gamma_n \in (0, \frac{1}{2}]$, and $\mathfrak{I}(\rho^n, \gamma_n) \in L_2^d(\rho^n)$. Then, for all $n \ge 1$,

$$\mathrm{KL}(\rho^n|\pi) \le R \prod_{i=1}^{n} \left(1 - \frac{1}{2}\lambda(1-\nu_i)^{-1} h_i\right). \tag{4.23}$$

Proof of Theorem 24. From the proof of Theorem 23, we can bound the difference in KL-divergence between two consecutive iterations by

$$\begin{aligned}
&\mathrm{KL}(\rho^{n+1}|\pi) - \mathrm{KL}(\rho^n|\pi)\\
&\le -\frac{1}{2}h_{n+1}(1-\nu_{n+1})^{-1} I(\rho^n|\pi) + h_{n+1}\nu_{n+1}^{2\gamma_n}(1-\nu_{n+1})^{-2\gamma_n-1} R_n^2 \left(1 + \frac{1}{2}h_{n+1}\nu_{n+1}^{-1}\alpha^2 B^2\right)\\
&= -\frac{1}{4}h_{n+1}(1-\nu_{n+1})^{-1} I(\rho^n|\pi)\left(2 - \frac{\nu_{n+1}^{2\gamma_n}}{(1-\nu_{n+1})^{2\gamma_n}} \frac{R_n^2\left(1+\frac{1}{2}h_{n+1}\nu_{n+1}^{-1}\alpha^2 B^2\right)}{I(\rho^n|\pi)}\right)\\
&\le -\frac{1}{4}h_{n+1}(1-\nu_{n+1})^{-1} I(\rho^n|\pi)\left(2 - \frac{\nu_{n+1}^{2\gamma_n}}{(1-\nu_{n+1})^{2\gamma_n}} \frac{R_n^2}{I(\rho^n|\pi)}\right)\\
&\le -\frac{1}{4}h_{n+1}(1-\nu_{n+1})^{-1} I(\rho^n|\pi),
\end{aligned}$$

where the last inequality follows from (4.22). Last, since π satisfies the log-Sobolev inequality with parameter λ, we get

$$\mathrm{KL}(\rho^{n+1}|\pi) \le \left(1 - \frac{1}{2}\lambda(1-\nu_{n+1})^{-1}h_{n+1}\right)\mathrm{KL}(\rho^n|\pi),$$

and (4.23) follows from the above recursive inequality. ■

Remark 32. We make the following remarks about the above result.

(i) Prior results on the analysis of time-discretization of the SVGF under functional inequality assumptions are established only in the weaker Stein-Fisher information metric [KSA+20, SSR22]. Our results above are established for the KL-divergence and is more in line with similar results established for other randomized Monte Carlo algorithms [VW19, CEL$^+$21, BCE$^+$22].

(ii) According to (4.23), to reach an ϵ-accuracy in KL-divergence, we need the number of iterations to be at least n_ϵ such that $\prod_{i=1}^{n_\epsilon}\left(1-\frac{1}{2}\lambda(1-\nu_i)^{-1}h_i\right)R \le \epsilon$. With the fact that $\log(1-x) < -x$ for all $x \in (0,1)$, we get n_ϵ satisfies

$$\sum_{i=1}^{n_\epsilon}(1-\nu_i)^{-1}h_i \ge \frac{2}{\lambda}\log\left(\frac{R}{\epsilon}\right).$$

Under (4.22), if we can choose the time step sizes $(h_i)_{i=1}^{\infty}$ to be constant $h > 0$, then we have $n_\epsilon = O(\log(R/\epsilon))$. For comparison, in the following Table, we provide the iteration complexity results for different methods, to obtain $\mathrm{KL}(\rho_n|\pi) \le \epsilon$, under the assumption that the target π satisfies LSI.

Algorithm	Source	Type	Iterations
SVGD	NA	Deterministic	unknown
LMC	[VW19, CEL$^+$21]	Randomized	$\mathcal{O}(\frac{1}{\epsilon})$
MALA	NA	Randomized	unknown
Proximal sampler	[CCSW22]	Randomized	$\mathcal{O}\left(\log^{\lambda}(\frac{1}{\epsilon})\right)$
Regularized SVGF	Theorem 24	Deterministic	$\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$

Table 4.1. The results from [VW19, CEL$^+$21, CCSW22] are presented in a simplified manner to convey the dependency on the accuracy parameter ϵ. The result for the proximal sampler holds only in expectation. Currently it is not clear how to obtain a high-probability result in KL-divergence; see [CCSW22] for details.

4.6. Existence and Uniqueness

The existence and uniqueness of the SVGF was studied in [LLN19]. Motivated by their approach, in this section we study the existence and uniqueness of solutions to (4.6) under appropriate assumptions. Our main difficulty is in handling the non-linear operator $\left((1-\nu)\mathcal{T}_{k,\mu}+\nu I\right)^{-1}\mathcal{T}_{k,\mu}$ in the R-SVGF.

We first introduce the definition of weak solutions to (4.6). We restrict the initial conditions in the probability measure space $\mathcal{P}_V$ which defined as

$$\mathcal{P}_V := \left\{ \rho \in \mathcal{P} \ : \ \|\rho\|_{\mathcal{P}_V} := \int_{\mathbb{R}^d} (1 + V(x))\rho(x)dx < \infty \right\},$$

where $\mathcal{P}$ denotes the set of all probability measures on $\mathbb{R}^d$. We say that a measure-value function $\rho \in \mathcal{C}([0,\infty), \mathcal{P})$ is a weak solution to (4.6) with initial condition $\rho_0 \in \mathcal{P}_V$ if

$$\sup_{t \in [0,T]} \|\rho_t\|_{\mathcal{P}_V} < \infty, \ \forall\, T > 0,$$

and

$$\int_0^\infty \int_{\mathbb{R}^d} (\partial_t \phi(t,x) + \nabla \phi(t,x) \cdot U[\rho_t](x))\rho_t(x)dxdt + \int_{\mathbb{R}^d} \phi(0,x)\rho_0(x)dx = 0,$$

for all $\phi \in \mathcal{C}_0^\infty([0,\infty) \times \mathbb{R}^d)$ and $U[\rho] := -\Big((1-\nu)\iota_{k,\rho}\iota_{k,\rho}^* + \nu I\Big)^{-1} \iota_{k,\rho}\iota_{k,\rho}^*(\nabla \log \frac{\rho}{\pi})$.

In order to study the existence of weak solutions, we consider the characteristic flow (see, for example [MRZ16] and [LLN19, Definition 3.1]) induced by (4.6), which is written as

$$\begin{cases} \dfrac{d}{dt}\Phi(t,x,\rho_0) = -\mathcal{D}_{\nu,\rho_t}\nabla \log \dfrac{\rho_t}{\pi}(\Phi(t,x,\rho_0)), \\ \rho_t = (\Phi(t,\cdot,\rho_0))_{\#}\rho_0, \\ \Phi(0,x,\rho_0) = x, \end{cases} \tag{4.24}$$

where $\mathcal{D}_{\nu,\rho_t} = \Big((1-\nu)\iota_{k,\rho_t}\iota_{k,\rho_t}^* + \nu I\Big)^{-1} \iota_{k,\rho_t}\iota_{k,\rho_t}^*$ for all $t > 0$. Here, the expression $\rho_t = \Phi(t,\cdot,\rho_0)_{\#}\rho_0$ means that the measure ρ_t is the push-forward measure of ρ_0 under the map $x \to \Phi(t,x,\rho_0)$. We think of $\{X(t,\cdot,\rho_0)\}_{t\geq 0,\rho_0}$ as a family of maps from $\mathbb{R}^d$ to $\mathbb{R}^d$ parameterized by t and ρ_0. The existence and uniqueness to the weak solutions of (4.6) is equivalent to the existence and uniqueness of solutions to (4.24). In Theorem 25, we first prove that the mean field characteristic flow in (4.24) is well-defined. To do so, we also require the following additional assumptions on the kernel and the potential functions.

Assumption K1. The kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is symmetric, positive definite and fourth continuously differentiable in both variables with bounded derivatives up to fourth order. More explicitly, we assume

(1) $\|k\|_\infty := \sup_{x\in\mathbb{R}^d} \sqrt{k(x,x)} < \infty$.

(2) $\|\nabla k\|_\infty := \sup_{x,y\in\mathbb{R}^d} |\nabla_1 k(x,y)| = \sup_{x,y\in\mathbb{R}^d} |\nabla_2 k(x,y)| < \infty$.

(3) $\|\nabla_1 \cdot \nabla_2 k\|_\infty := \sup_{x,y\in\mathbb{R}^d} |\nabla_x \cdot \nabla_y k(x,y)| < \infty$.

(4) $\left\|\nabla^2 k\right\|_\infty := \sup_{x,y\in\mathbb{R}^d} \left\|\nabla_x^2 k(x,y)\right\|_2 < \infty$.

(5) $\|\nabla_1 \nabla_2 k(x,y)\|_\infty := \sup_{x,y\in\mathbb{R}^d} \|\nabla_x \nabla_y k(x,y)\|_2 < \infty$.

(6) $\left\|\nabla^2(\nabla_1 \cdot \nabla_2 k)\right\|_\infty := \sup_{x,y\in\mathbb{R}^d} \left\|\nabla_x^2(\nabla_x \cdot \nabla_y k(x,y))\right\|_2 < \infty$.

(7) $\|\nabla_1\nabla_2(\nabla_1 \cdot \nabla_2 k)\|_\infty := \sup_{x,y\in\mathbb{R}^d} \|\nabla_x\nabla_y(\nabla_x \cdot \nabla_y k(x,y))\|_2 < \infty$.

(8) $\left\|\nabla_1^2 \cdot \nabla_2^2 k\right\|_\infty := \sup_{x,y\in\mathbb{R}^d} \sum_{i,j=1}^d |\partial_{x_i}\partial_{x_j}\partial_{y_i}\partial_{y_j} k(x,y)| < \infty$.

We emphasize here that [LLN19] required that the kernel is radial for their analysis. However, our analysis does not require this assumption. A classical example of a kernel satisfying the above conditions is the Gaussian kernel.

Assumption V1. The potential function $V : \mathbb{R}^d \to \mathbb{R}$ satisfies

(1) $V \in \mathcal{C}^2(\mathbb{R}^d)$, $V \geq 0$ and $V(x) \to +\infty$ as $|x| \to +\infty$.

(2) For any $\alpha, \beta > 0$, there exists a constant $C_{\alpha,\beta} > 0$ such that if $|y| \leq \alpha|x| + \beta$, then

$$(1 + |x|)(|\nabla V(y)| + \left\|\nabla^2 V(y)\right\|_2) \leq C_{\alpha,\beta}(1 + V(x)).$$

(3) V is gradient Lipschitz with parameter L_V, i.e., for all $x \in \mathbb{R}^d$, $\left\|\nabla^2 V(x)\right\|_2 \leq L_V$.

To present our result, we define the set of functions

$$Y := \left\{ u \in \mathcal{C}(\mathbb{R}^d;\mathbb{R}^d) |\sup_{x\in\mathbb{R}^d} |u(x) - x| < \infty \right\},$$

which is a complete metric space with the uniform metric $d_Y(u, v) = \sup_{x\in\mathbb{R}^d} |u(x) - v(x)|$.

THEOREM 25. Let k satisfy Assumption K1, V satisfy Assumption V1 and $\rho_0 \in \mathcal{P}_V$.

(i) For any $T > 0$, there exists a unique solution $\Phi(\cdot, \cdot, \rho_0) \in \mathcal{C}^1([0, T]; Y)$ to (4.24). Moreover, the measure $\rho_t = \Phi(t, \cdot, \rho_0)_{\#}\rho_0$ satisfies

$$\|\rho_t\|_{\mathcal{P}_V} \leq \|\rho_0\|_{\mathcal{P}_V} \exp(C_{1,0}\nu^{-1/2} \|k\|_\infty \mathrm{KL}(\rho_0|\pi)^{1/2}t^{1/2}).$$

(ii) For any $\rho_0 \in \mathcal{P}_V$, there is a unique $\rho \in \mathcal{C}([0, \infty); \mathcal{P}_V)$ which is a weak solution to (4.6). Moreover, for all $t \geq 0$,

$$\|\rho_t\|_{\mathcal{P}_V} \leq \|\rho_0\|_{\mathcal{P}_V} \exp(C_{1,0}\nu^{-1/2} \|k\|_\infty \mathrm{KL}(\rho_0|\pi)^{1/2}t^{1/2}).$$

REMARK 33. In Theorem 25, we introduce an upper bound to the $\mathcal{P}_V$-norm of the solution to (4.6) for any $\nu \in (0, 1]$. A similar result is established for the case of SVGF, i.e., when $\nu = 1$ in [LLN19, Theorem 2.4]. In comparison to [LLN19, Theorem 2.4], our result requires that the initial KL-divergence to the target is bounded. Furthermore, if we set $\nu = 1$ in our result, we do not end up recovering their result. When $\nu = 1$, there is an explicit integral formula to $\mathcal{D}_{1,\rho_t}\nabla \log \frac{\rho_t}{\pi}$ which is leveraged in [LLN19] for their proof. For $\nu \in (0, 1)$, due to the absence of an explicit representation, we get the result in Theorem 25 by carefully analyzing the quantity $\mathcal{D}_{\nu,\rho_t}(\nabla \log \frac{\rho_t}{\pi})$ along with the decay of KL-divergence property introduced in Proposition 22.

PROOF OF THEOREM 25. Our proof leverages the approach of [LLN19, Theorem 3.2] for the case of SVGF. In comparison to [LLN19], we handle various difficulties arising with the non-linear operator in R-SVGF. We first prove claim (i) based on the following two steps. Claim (ii) follows directly from claim (i) and [Vil21, Theorem 5.34].

Step 1 (Local well-posedness): Fix $r > 0$ and define

$$Y_r := \left\{ u \in Y | \sup_{x\in\mathbb{R}^d} |u(x) - x| \leq r \right\}. \tag{4.25}$$

We will prove that there exists $T_0 > 0$ such that (4.24) has a unique solution $\Phi(t, x, \rho_0)$ in the set $S_r = \mathcal{C}([0, T_0]; Y_r)$ which is a complete metric space with metric

$$d_S(u, v) = \sup_{t \in [0, T_0]} d_Y(u(t, \cdot), v(t, \cdot)).$$

The integral formulation of (4.24) is

$$\Phi(t, x, \rho_0) = x - \int_0^t \mathcal{D}_{\nu,\rho_s} \nabla \log \frac{\rho_s}{\pi}(\Phi(s, x, \rho_0)) ds. \tag{4.26}$$

Let us define the operator $\mathcal{F} : u(t, \cdot) \mapsto \mathcal{F}(u)(t, \cdot)$ by

$$\mathcal{F}(u)(t, x) = x - \int_0^t \mathcal{D}_{\nu,\rho_s} \nabla \log \frac{\rho_s}{\pi}(u(s, x)) ds,$$

where $\rho_t = (u(t, \cdot))_{\#} \rho_0$. We will show that $\mathcal{F}$ is a contraction from S_r to S_r and thus has a unique fixed point. First we show that $\mathcal{F}$ maps S_r into S_r for some $T_0 > 0$. For some $u \in S_r$, checking that $(t, x) \mapsto \mathcal{F}(u)(t, x)$ is continuous is straightforward. We need to show that $\sup_{x \in \mathbb{R}^d} |\mathcal{F}(u)(t, x) - x| \leq r$ for any $u \in S_r$.

$$\begin{aligned}
\mathcal{F}(u)(t, x) - x &= -\int_0^t \mathcal{D}_{\nu,\rho_s} \nabla \log \frac{\rho_s}{\pi}(u(s, x)) ds \\
&= -\int_0^t \left((1-\nu)\iota_{k,\rho_s}\iota_{k,\rho_s}^* + \nu I\right)^{-1} \iota_{k,\rho_s}\iota_{k,\rho_s}^* \nabla \log \frac{\rho_s}{\pi}(\Phi(s, x)) ds \\
&= -\int_0^t \left((1-\nu)\iota_{k,\rho_s}\iota_{k,\rho_s}^* + \nu I\right)^{-1} \mathbb{E}_{y \sim \rho_s}[k(\cdot, y)\nabla V(y) - \nabla k(y, \cdot)](u(s, x)) ds.
\end{aligned}$$

Note that there is an equivalent representation for $\mathcal{D}_{\nu,\rho_s}$:

$$\mathcal{D}_{\nu,\rho_s} = \iota_{k,\rho_s} \left((1-\nu)\iota_{k,\rho_s}^* \iota_{k,\rho_s} + \nu I_d\right)^{-1} \iota_{k,\rho_s}^*.$$

We then analyze the operators ι_{k,ρ_s} and $\left((1-\nu)\iota_{k,\rho_s}^{*}\iota_{k,\rho_s}+\nu I_d\right)^{-1}\iota_{k,\rho_s}^{*}$ respectively. Since $\|k\|_{\infty}<\infty$, according to Proposition 19, ι_{k,ρ_s} is the inclusion operator from $\mathcal{H}_k^d$ to $L_{\infty}^d(\mathbb{R}^d)$. The corresponding operator norm, denoted as $\|\iota_{k,\rho_s}\|_{\mathcal{H}_k^d\to L_\infty^d}$ can be bounded in the following way:

$$\begin{aligned}\|\iota_{k,\rho_s}\|_{\mathcal{H}_k^d\to L_\infty^d} &:= \sup_{\|f\|_{\mathcal{H}_k^d}=1}\sup_{x\in\mathbb{R}^d}|f(x)|\\ &= \sup_{\|f\|_{\mathcal{H}_k^d}=1}\sup_{x\in\mathbb{R}^d}|\langle k(x,\cdot),f\rangle_{\mathcal{H}_k}|\\ &\le \sup_{x\in\mathbb{R}^d}\sqrt{k(x,x)}:=\|k\|_{\infty}\,. \end{aligned}\tag{4.27}$$

Meanwhile, let $(\lambda_i,e_i)_{i=1}^{\infty}$ be the spectrum of $\iota_{k,\rho_s}\iota_{k,\rho_s}^{*}$ with $(e_i)_{i=1}^{\infty}$ being an orthonormal basis of $L_2^d(\rho_s)\equiv\overline{\mathrm{Ran}(\iota_{k,\rho_s}\iota_{k,\rho_s}^{*})}$ according to Proposition 19, $(\sqrt{\lambda_i}e_i)_{i=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}_k^d$; see also Remark 27. Hence, we have

$$\begin{aligned}&\left\|\left((1-\nu)\iota_{k,\rho_s}^{*}\iota_{k,\rho_s}+\nu I_d\right)^{-1}\iota_{k,\rho_s}^{*}\nabla\log\frac{\rho_s}{\pi}\right\|_{\mathcal{H}_k^d}^2\\ &\le\left\|\left((1-\nu)\iota_{k,\rho_s}^{*}\iota_{k,\rho_s}+\nu I_d\right)^{-\frac{1}{2}}\right\|_{\mathcal{H}_k^d\to\mathcal{H}_k^d}^2\left\|\left((1-\nu)\iota_{k,\rho_s}^{*}\iota_{k,\rho_s}+\nu I_d\right)^{-\frac{1}{2}}\iota_{k,\rho_s}^{*}\nabla\log\frac{\rho_s}{\pi}\right\|_{\mathcal{H}_k^d}^2\\ &\le\nu^{-1}I_{\nu,Stein}(\rho_s|\pi).\end{aligned}\tag{4.28}$$

where the last inequality follows from (4.8) and the fact that $(1-\nu)\iota_{k,\rho_s}^{*}\iota_{k,\rho_s}$ is positive. With (4.27) and (4.28), we get the following uniform bound on $|\mathcal{D}_{\nu,\rho_s}\log\frac{\rho_s}{\pi}(x)|$ for all $x\in\mathbb{R}^d$,

$$\begin{aligned}\left|\mathcal{D}_{\nu,\rho_s}\nabla\log\frac{\rho_s}{\pi}(x)\right| &\le\left\|\mathcal{D}_{\nu,\rho_s}\nabla\log\frac{\rho_s}{\pi}\right\|_{L_\infty^d}\\ &\le\|\iota_{k,\rho_s}\|_{\mathcal{H}_k^d\to L_\infty^d}\left\|\left((1-\nu)\iota_{k,\rho_s}^{*}\iota_{k,\rho_s}+\nu I_d\right)^{-1}\iota_{k,\rho_s}^{*}\nabla\log\frac{\rho_s}{\pi}\right\|_{\mathcal{H}_k^d}\\ &\le\nu^{-\frac{1}{2}}\|k\|_{\infty}I_{\nu,Stein}(\rho_s|\pi)^{\frac{1}{2}}.\end{aligned}$$

Therefore, for all $t\in[0,T]$ and all $x\in\mathbb{R}^d$:

$$|\mathcal{F}(u)(t,x)-x|\le\nu^{-\frac{1}{2}}\|k\|_{\infty}\int_0^T I_{\nu,Stein}(\rho_s|\pi)^{\frac{1}{2}}ds.\tag{4.29}$$

According to Lemma 4.6.1, there exists $T_0 > 0$ such that for all $u \in S_r$,

$$\int_0^{T_0} I_{\nu,Stein}(\rho_t|\pi)^{\frac{1}{2}} dt < \nu^{1/2} \left\|k\right\|_\infty^{-1} r,$$

which along with (4.29) implies $|\mathcal{F}(u)(t,x) - x| \leq r$ for all $u \in S_r$.

Next we show that $\mathcal{F}$ is a contraction on S_r. Our goal is to show that there exists $T_0 > 0$ and $C \in (0,1)$ such that for any $u, v \in S_r$,

$$\sup_{t\in[0,T_0]} \sup_{x\in\mathbb{R}^d} |\mathcal{F}(u)(t,x) - \mathcal{F}(v)(t,x)| < C \sup_{t\in[0,T_0]} \sup_{x\in\mathbb{R}^d} |u(t,x) - v(t,x)|.$$

Observe that

$$\begin{aligned}
|\mathcal{F}(u)(t,x) - \mathcal{F}(v)(t,x)| &= \left| \int_0^t \mathcal{D}_{\nu,\rho_{1,s}} \nabla \log \frac{\rho_{1,s}}{\pi} \left(u(s,x)\right) - \mathcal{D}_{\nu,\rho_{2,s}} \nabla \log \frac{\rho_{2,s}}{\pi} \left(v(s,x)\right) ds \right| \\
&\leq \left| \int_0^t \mathcal{D}_{\nu,\rho_{1,s}} \nabla \log \frac{\rho_{1,s}}{\pi} (u(s,x)) - \mathcal{D}_{\nu,\rho_{2,s}} \nabla \log \frac{\rho_{2,s}}{\pi} (u(s,x)) ds \right| \\
&\quad + \left| \int_0^t \mathcal{D}_{\nu,\rho_{2,s}} \nabla \log \frac{\rho_{2,s}}{\pi} (u(s,x)) - \mathcal{D}_{\nu,\rho_{2,s}} \nabla \log \frac{\rho_{2,s}}{\pi} (v(s,x)) ds \right| \\
&\leq d_S(u,v) \int_0^{T_0} C_1(t) dt + \int_0^{T_0} L(t) dt \sup_{t\in[0,T_0]} \sup_{x\in\mathbb{R}^d} |u(t,x) - v(t,x)| \\
&= d_S(u,v) \int_0^{T_0} C_1(t) + L(t) \ dt,
\end{aligned}$$

where the second inequality follows from Lemma 4.6.2 and Lemma 4.6.3. Furthermore, according to (4.33) and (4.34), there exists $T_0 > 0$ such that

$$\int_0^{T_0} C_1(t) + L(t) \ dt < 1.$$

Therefore we have proved there exists $T_0 > 0$ such that $\mathcal{F}$ is a contraction from S_r into S_r. According to the contraction theorem, $\mathcal{F}$ has a unique fixed point $\Phi(\cdot, \cdot, \rho_0) \in S_r$ which solves (4.24). Defining $\rho_t = (\Phi(t, \cdot, \rho_0))_{\#}\rho_0$, one sees that $\Phi(t, x, \rho_0)$ solves (4.24) in the time interval $[0, T_0]$.

Step 2 (Extension of local solution): According to (4.32) and (4.34), we can extend the local solution beyond time T_0 as long as the quantity

$$\|\rho_t\|_{\mathcal{P}_V} = \int_{\mathbb{R}^d} (1 + V(\Phi(t,x,\rho_0)))\rho_0(dx)$$

remains finite. Next we establish a bound on this quantity showing that the local solution can be extended for any $t > 0$.

$$\begin{aligned}
\partial_t \int_{\mathbb{R}^d} (1 + V(\Phi(t,x,\rho_0)))\,\rho_0(dx) &= -\int_{\mathbb{R}^d} \left\langle \nabla V(\Phi(t,x,\rho_0)), \mathcal{D}_{\nu,\rho_t} \nabla \log \frac{\rho_t}{\pi}(\Phi(t,x,\rho_0)) \right\rangle \rho_0(dx) \\
&= -\left\langle \nabla V, \iota_{k,\rho_t} \left((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\right)^{-1} \iota^*_{k,\rho_t} \nabla \log \frac{\rho_t}{\pi} \right\rangle_{L_2^d(\rho_t)} \\
&= -\left\langle \iota^*_{k,\rho_t}\nabla V, \left((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\right)^{-1} \iota^*_{k,\rho_t} \nabla \log \frac{\rho_t}{\pi} \right\rangle_{\mathcal{H}_k^d} \\
&\le \left\|\iota^*_{k,\rho_t}\nabla V\right\|_{\mathcal{H}_k^d} \left\| \left((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\right)^{-1} \iota^*_{k,\rho_t} \nabla \log \frac{\rho_t}{\pi} \right\|_{\mathcal{H}_k^d},
\end{aligned}$$

where

$$\begin{aligned}
\left\|\iota^*_{k,\rho_t}\nabla V\right\|^2_{\mathcal{H}_k^d} &= \left\langle \nabla V, \iota_{k,\rho}\iota^*_{k,\rho_t}\nabla V \right\rangle_{L_2^d(\rho_t)} \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(y,z)\left\langle \nabla V(y), \nabla V(z)\right\rangle \rho_t(y)\rho_t(z)dydz \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(\Phi(t,y,\rho_0),\Phi(t,z,\rho_0))\left\langle \nabla V(\Phi(t,y,\rho_0)), \nabla V(\Phi(t,z,\rho_0))\right\rangle \rho_0(y)\rho_0(z)dydz \\
&\le \|k\|^2_\infty \left(\int_{\mathbb{R}^d} |\nabla V(\Phi(t,y),\rho_0)|\rho_0(y)dy\right)^2 \\
&\le \|k\|^2_\infty\, C^2_{1,0}\, \|\rho_t\|^2_{\mathcal{P}_V},
\end{aligned}$$

where the last inequality follows from Assumption V1. Therefore

$$\begin{aligned}
\partial_t \|\rho_t\|_{\mathcal{P}_V} &\le C_{1,0}\,\|k\|_\infty\,\|\rho_t\|_{\mathcal{P}_V} \left\| ((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d)^{-1}\iota^*_{k,\rho_t}\nabla \log \frac{\rho_t}{\pi} \right\|_{\mathcal{H}_k^d} \\
&\le C_{1,0}\,\|k\|_\infty\,\nu^{-\frac{1}{2}} I_{\nu,Stein}(\rho_t|\pi)^{\frac{1}{2}}\,\|\rho_t\|_{\mathcal{P}_V},
\end{aligned}$$

where the last inequality follows from (4.28). It follows from Gronwall's inequality that

$$
\begin{aligned}
\|\rho_t\|_{\mathcal{P}_V} &\leq \|\rho_0\|_{\mathcal{P}_V} \exp\left(C_{1,0}\nu^{-\frac{1}{2}} \|k\|_\infty \int_0^t I_{\nu,Stein}(\rho_s|\pi)^{\frac{1}{2}} ds\right) \\
&\leq \|\rho_0\|_{\mathcal{P}_V} \exp\left(C_{1,0}\nu^{-\frac{1}{2}} \|k\|_\infty \sqrt{t\int_0^t I_{\nu,Stein}(\rho_s|\pi) ds}\right) \\
&\leq \|\rho_0\|_{\mathcal{P}_V} \exp\left(C_{1,0}\nu^{-\frac{1}{2}} \|k\|_\infty \sqrt{t\mathrm{KL}(\rho_0|\pi)}\right), \qquad (4.30)
\end{aligned}
$$

where the second inequality follows from Jensen's inequality and the last inequality follows from (4.15). With this bound we can iterate the argument to extend the local solution defined on $[0, T_0] \times \mathbb{R}^d$ to all of $[0, \infty) \times \mathbb{R}^d$, so that (4.30) holds for all $t > 0$. Finally $\Phi(\cdot, x, \rho_0))$ has continuous first order derivative due to the integral formulation in (4.26), Assumption K1 and Assumption V1. The proof is thus complete. ■

Lemma 4.6.1. *Let $\rho_0 \in \mathcal{P}_V$ and suppose the assumptions of Theorem 25 hold. Then, for any $\epsilon > 0$, there exists a constant $T > 0$ such that for all $u \in S_r$ and $t \in [0, T]$, with $\rho_t = u(t, \cdot)_\# \rho_0$, we have*

$$
\int_0^T I_{\nu,Stein}(\rho_t|\pi)^{\frac{1}{2}} dt < \epsilon. \qquad (4.31)
$$

Proof of Lemma 4.6.1. According to Lemma 20, the regularized kernelized Stein discrepancy can be written as

$$
\begin{aligned}
S(\rho_t|\pi)^2 &= \left(\mathbb{E}_{x\sim\rho_t}\left[\mathrm{trace}(\mathcal{A}_\pi \phi^*_{\rho_t,\pi}(x))\right]\right)^2 \\
&= \left\|\left((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\right)^{-\frac{1}{2}} \iota^*_{k,\rho_t} \nabla \log \frac{\rho_t}{\pi}\right\|^2_{\mathcal{H}_k^d} \\
&= \left\langle \iota^*_{k,\rho_t} \nabla \log \frac{\rho_t}{\pi}, \left((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\right)^{-1} \iota^*_{k,\rho_t} \nabla \log \frac{\rho_t}{\pi}\right\rangle_{\mathcal{H}_k^d} = I_{\nu,Stein}(\rho_t, \pi).
\end{aligned}
$$

Meanwhile, since $\rho_t = u(t, \cdot)_\# \rho_0$ with $u \in S_r$, for any $\mathbb{R}^d$-valued random vector X, $u(t, X) \sim \rho_t$ and $|u(t, X) - X| \leq r$ almost surely. Therefore

$$
W_1(\rho_t, \pi) \leq W_1(\rho_0, \rho_t) + W_1(\rho_0, \pi) = \inf_{X\sim\rho_0, Y\sim\rho_t} \mathbb{E}\left[|X - Y|\right] + W_1(\rho_0, \pi) \leq r + W_1(\rho_0, \pi).
$$

Next we upper bound the regularized kernelized Stein discrepancy by the Wasserstein-2 distance. According to [GM17, Lemma 18], for any general vector field $\phi \in \mathcal{H}_k^d$, we have

$$\begin{aligned} \left|\mathbb{E}_{x\sim\rho_t}\left[\operatorname{trace}(\mathcal{A}_\pi\phi(x))\right]\right| \leq{}& \left(M_0(\phi)M_1(\nabla V) + M_2(\phi)d\right)W_1(\rho_t,\pi) \\ &+ \sqrt{2M_0(\phi)M_1(\phi)\mathbb{E}_{x\sim\pi}\left[|\nabla V(x)|^2\right]W_1(\rho_t,\pi)}, \end{aligned}$$

where for any $g : \mathbb{R}^d \to \mathbb{R}^d$ and $g \in C^1(\mathbb{R}^d)$,

$$M_0(g) := \sup_{x\in\mathbb{R}^d}|g(x)|, \quad M_1(g) := \sup_{x\neq y}\frac{|g(x)-g(y)|}{|x-y|}, \quad M_2(g) := \sup_{x\neq y}\frac{\|\nabla g(x)-\nabla g(y)\|_2}{|x-y|}.$$

For any $\phi \in \mathcal{H}_k^d$ and $\phi = [\phi_1, \cdots, \phi_d]^T$, according to [SC08, Lemma 4.34],

$$\sup_{x\in\mathbb{R}^d}|D^\alpha\phi_i(x)| = \sup_{x\in\mathbb{R}^d}\left|D^\alpha\left\langle\phi_i, k(x,\cdot)\right\rangle_{\mathcal{H}_k}\right| \leq \|\phi_i\|_{\mathcal{H}_k^d}\sup_{x\in\mathbb{R}^d}|D_1^\alpha D_2^\alpha k(x,x)|^{\frac{1}{2}}.$$

Therefore,

$$M_0(\phi) = \sup_{x\in\mathbb{R}^d}\sqrt{\sum_{i=1}^d\phi_i(x)^2} \leq \sqrt{\sum_{i=1}\|\phi_i\|_{\mathcal{H}_k}^2\sup_{x\in\mathbb{R}^d}k(x,x)} = \|k\|_\infty\|\phi\|_{\mathcal{H}_k^d},$$

$$\begin{aligned} M_1(\phi) &= \sup_{x\neq y}\frac{\sqrt{\sum_{i=1}^d\left(\phi_i(x)-\phi_i(y)\right)^2}}{|x-y|} \leq \sqrt{\sum_{i=1}^d\sup_{x\in\mathbb{R}^d}|\nabla\phi_i(x)|^2} \\ &\leq \sqrt{\sum_{i=1}^d\sum_{j=1}^d\|\phi_i\|_{\mathcal{H}_k}^2\sup_{x\in\mathbb{R}^d}D_1^{e_j}D_2^{e_j}k(x,x)} = \left(\sup_{x\in\mathbb{R}^d}\operatorname{trace}\left(\nabla_1\nabla_2k(x,x)\right)\right)^{\frac{1}{2}}\|\phi\|_{\mathcal{H}_k^d} \\ &\leq \|\nabla_1\cdot\nabla_2k\|_\infty^{\frac{1}{2}}\|\phi\|_{\mathcal{H}_k^d}, \end{aligned}$$

$$\begin{aligned} M_2(\phi) &= \sup_{x\neq y}\frac{\|\nabla\phi(x)-\nabla\phi(y)\|_2}{|x-y|} \leq \sup_{x\neq y}\frac{\|\nabla\phi(x)-\nabla\phi(y)\|_F}{|x-y|} \leq \sqrt{\sum_{i,j=1}^d\sup_{x\neq y}\frac{|\partial_j\phi_i(x)-\partial_j\phi_i(y)|^2}{|x-y|^2}} \\ &\leq \sqrt{\sum_{i,j,l=1}^d\sup_{x\in\mathbb{R}^d}D_1^{e_j+e_l}D_2^{e_j+e_l}k(x,x)\|\phi_i\|_{\mathcal{H}_k}^2} \leq \left\|\nabla_1^2\cdot\nabla_2^2k\right\|_\infty^{\frac{1}{2}}\|\phi\|_{\mathcal{H}_k^d}. \end{aligned}$$

According to Assumption V1, $M_1(\nabla V) = L_V$ and $\mathbb{E}_{x\sim\pi}\left[|\nabla V(x)|^2\right] \leq C_{1,0}\left\|\pi\right\|_{\mathcal{P}_V}$. Therefore

$$\begin{aligned}|\mathbb{E}_{x\sim\rho_t}\left[\mathrm{trace}(\mathcal{A}_\pi\phi(x))\right]| \leq{}& \left(\left\|k\right\|_\infty L_V + \left\|\nabla_1^2\cdot\nabla_2^2 k\right\|_\infty^{\frac{1}{2}} d\right)\left\|\phi\right\|_{\mathcal{H}_k^d}(W_1(\rho_0,\pi)+r)\\ &+\sqrt{2\left\|k\right\|_\infty\left\|\nabla_1\cdot\nabla_2 k\right\|_\infty^{\frac{1}{2}} C_{1,0}\left\|\pi\right\|_{\mathcal{P}_V}(W_1(\rho_0,\pi)+r)}\left\|\phi\right\|_{\mathcal{H}_k^d}.\end{aligned}$$

Note that ϕ^*_{k,ρ_t} satisfies that $\nu\left\|\phi^*_{k,\rho_t}\right\|^2_{\mathcal{H}_k^d} + (1-\nu)\left\|\phi^*_{k,\rho_t}\right\|_{L_2^d(\rho_t)^d} \leq 1$. Therefore

$$\left\|\phi^*_{k,\rho_t}\right\|_{\mathcal{H}_k^d} \leq \nu^{-1/2},$$

and

$$\begin{aligned}I_{\nu,Stein}(\rho_t|\pi)^{\frac{1}{2}} = S(\rho_t,\pi) &= \left|\mathbb{E}_{x\sim\rho_t}\left[\mathrm{trace}(\mathcal{A}_\pi\phi^*_{k,\rho_t}(x))\right]\right|\\ &\leq \nu^{-\frac{1}{2}}\left(\left\|k\right\|_\infty L_V + \left\|\nabla_1^2\cdot\nabla_2^2 k\right\|_\infty^{\frac{1}{2}} d\right)(W_1(\rho_0,\pi)+r)\\ &\quad+\nu^{-\frac{1}{2}}\sqrt{2\left\|k\right\|_\infty\left\|\nabla_1\cdot\nabla_2 k\right\|_\infty^{\frac{1}{2}} C_{1,0}\left\|\pi\right\|_{\mathcal{P}_V}(W_1(\rho_0,\pi)+r)}.\end{aligned}$$

Since the upper bound is independent of the choice of $u(t,\cdot)\in S_r$ and the time variable t, for any $\epsilon > 0$, we can choose a T small enough such that (4.31) holds. ■

Lemma 4.6.2. *Under the assumptions of Theorem 25, let $S_r = \mathcal{C}([0,T];Y_r)$ with Y_r defined in (4.25). Then for any $t\in[0,T]$ there exists $L(t) > 0$ such that for any $u\in S_r$, for all $x,y\in\mathbb{R}^d$ and $t\in[0,T]$,*

$$\left|\mathcal{D}_{\nu,\rho_t}\nabla\log\frac{\rho_t}{\pi}(x) - \mathcal{D}_{\nu,\rho_t}\nabla\log\frac{\rho_t}{\pi}(y)\right| \leq L(t)|x-y|, \tag{4.32}$$

where for all $t\in[0,T]$, $\rho_t = (u(t,\cdot))_{\#}\rho_0$ and

$$L(t) = \nu^{-\frac{1}{2}}\left(2\left\|\nabla_1\nabla_2 k\right\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} I_{\nu,Stein}(\rho_t|\pi)^{\frac{1}{2}}. \tag{4.33}$$

Proof of Lemma 4.6.2. Since $\mathcal{D}_{\nu,\rho_t} = \iota_{k,\rho_t}\Big((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\Big)^{-1}\iota^*_{k,\rho_t}$ and ι_{k,ρ_t} is the inclusion operator,

$$\begin{aligned}
&\left|\mathcal{D}_{\nu,\rho_t}\nabla\log\frac{\rho_t}{\pi}(x) - \mathcal{D}_{\nu,\rho_t}\nabla\log\frac{\rho_t}{\pi}(y)\right| \\
&= \left|\big((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\big)^{-1}\iota^*_{k,\rho_t}\nabla\log\frac{\rho_t}{\pi}(x) - \big((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\big)^{-1}\iota^*_{k,\rho_t}\nabla\log\frac{\rho_t}{\pi}(y)\right| \\
&= \left|\Big\langle k(x,\cdot) - k(y,\cdot), \big((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\big)^{-1}\iota^*_{k,\rho_t}\nabla\log\frac{\rho_t}{\pi}\Big\rangle_{\mathcal{H}_k}\right| \\
&\leq \|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}_k}\left\|\big((1-\nu)\iota^*_{k,\rho_t}\iota_{k,\rho_t} + \nu I_d\big)^{-1}\iota^*_{k,\rho_t}\nabla\log\frac{\rho_t}{\pi}\right\|_{\mathcal{H}_k^d} \\
&\leq \nu^{-\frac{1}{2}} I_{\nu,Stein}(\rho_t|\pi)^{\frac{1}{2}}\,\|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}_k^d},
\end{aligned}$$

where the second identity follows from the reproducing property and the last inequality follows from (4.28). Furthermore, we can write

$$\begin{aligned}
\|k(x,\cdot) - k(y,\cdot)\|^2_{\mathcal{H}_k^d} &= k(x,x) - 2k(x,y) + k(y,y) \\
&\leq \left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)|x-y|^2,
\end{aligned}$$

where the first identity follows from the RKHS property and the second identity follows from Taylor expansion and Assumption K1. Therefore (4.32) holds with $L(t)$ defined in (4.33). ■

Lemma 4.6.3. *Under the assumptions in Theorem 25, let $S_r = \mathcal{C}([0,T];Y_r)$ with Y_r defined in (4.25). Then for any $t \in [0,T]$, there exists $C_1(t) > 0$ such that for any $u, v \in S_r$,*

$$\sup_{x\in\mathbb{R}^d}\left|\mathcal{D}_{\nu,\rho_{1,t}}\nabla\log\frac{\rho_{1,t}}{\pi}(x) - \mathcal{D}_{\nu,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\right| \leq C_1(t)d_S(u,v), \tag{4.34}$$

where for all $t \in [0,T]$, $\rho_{1,t} = (u(t,\cdot))_\#\rho_0$, $\rho_{2,t} = (v(t,\cdot))_\#\rho_0$ and

(4.35)

$$C_1(t) = 2\nu^{-\frac{3}{2}}(1-\nu)\|k\|^2_\infty\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}}\|\nabla k\|_\infty I_{\nu,Stein}(\rho_{1,t}|\pi)^{\frac{1}{2}} + \nu^{-1}L_r\|k\|_\infty,$$

with L_r being defined in (4.39).

Proof of Lemma 4.6.3. With the facts that $\mathcal{D}_{\nu,\mu} = \iota_{k,\mu}(\iota_{k,\mu}^{*}\iota_{k,\mu} + \nu I)^{-1}\iota_{k,\mu}^{*}$ and $\iota_{k,\mu}$ is the inclusion operator we get,

$$\begin{aligned}
&\left|\mathcal{D}_{\nu,\rho_{1,t}}\nabla\log\frac{\rho_{1,t}}{\pi}(x) - \mathcal{D}_{\nu,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\right| \\
&= \left|\Big((1-\nu)\iota_{k,\rho_{1,t}}^{*}\iota_{k,\rho_{1.t}} + \nu I\Big)^{-1}\iota_{k,\rho_{1.t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}(x)\right. \\
&\qquad\left. - \Big((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}} + \nu I\Big)^{-1}\iota_{k,\rho_{2,t}}^{*}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\right| \\
&\leq \left|\left(\Big((1-\nu)\iota_{k,\rho_{1,t}}^{*}\iota_{k,\rho_{1.t}} + \nu I\Big)^{-1} - \Big((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}} + \nu I\Big)^{-1}\right)\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}(x)\right| \\
&\qquad + \left|\Big((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}} + \nu I\Big)^{-1}\left(\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}(x) - \iota_{k,\rho_{2,t}}^{*}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\right)\right|.
\end{aligned}$$

We then turn to study the two terms in the upper bound separately.

First term: Note that, we have

$$\begin{aligned}
&\left|\left(\Big((1-\nu)\iota_{k,\rho_{1,t}}^{*}\iota_{k,\rho_{1.t}} + \nu I\Big)^{-1} - \Big((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}} + \nu I\Big)^{-1}\right)\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}(x)\right| \\
&= \left|\iota_{k,\rho_{1,t}}\Big((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}} + \nu I\Big)^{-1}\Big((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}} - (1-\nu)\iota_{k,\rho_{1,t}}^{*}\iota_{k,\rho_{1,t}}\Big)\right. \\
&\qquad\left.\times\Big((1-\nu)\iota_{k,\rho_{1,t}}^{*}\iota_{k,\rho_{1,t}} + \nu I\Big)^{-1}\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}(x)\right| \\
&\leq \left\|\iota_{k,\rho_{1,t}}\right\|_{\mathcal{H}_k^d\to L_\infty^d}\left\|\Big((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}} + \nu I\Big)^{-1}\right\|_{\mathcal{H}_k^d\to\mathcal{H}_k^d}(1-\nu)\left\|\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}} - \iota_{k,\rho_{1,t}}^{*}\iota_{k,\rho_{1,t}}\right\|_{\mathcal{H}_k^d\to\mathcal{H}_k^d} \\
&\qquad\times\left\|\Big((1-\nu)\iota_{k,\rho_{1,t}}^{*}\iota_{k,\rho_{1,t}} + \nu I\Big)^{-1}\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}\right\|_{\mathcal{H}_k^d} \\
&\leq \|k\|_\infty\,\nu^{-1}(1-\nu)\nu^{-\frac{1}{2}}I_{\nu,Stein}(\rho_{1,t}|\pi)^{\frac{1}{2}} \\
&\qquad\times\sup_{\|\phi\|_{\mathcal{H}_k^d}=1}\left\langle\int_{\mathbb{R}^d}k(\cdot,x)\phi(x)(d\rho_{1,t}(x) - d\rho_{2,t}(x)), \int_{\mathbb{R}^d}k(\cdot,y)\phi(y)(d\rho_{1,t}(y) - d\rho_{2,t}(y))\right\rangle_{\mathcal{H}_k^d}^{\frac{1}{2}}
\end{aligned}$$

(4.36)

$$\leq 2\nu^{-\frac{3}{2}}(1-\nu)\|k\|_\infty^2\left(3\left\|\nabla^2 k\right\|_\infty + 2\|\nabla_1\nabla_2 k\|_\infty\right)^{\frac{1}{2}}I_{\nu,Stein}(\rho_{1,t}|\pi)^{\frac{1}{2}}d_S(u,v).$$

As we are bounding the function value by its L_∞^d norm, the second step allows the function to be in the space of L_∞^d, without which we think of the function as belonging to the RKHS. The second inequality follows from (4.27) and (4.28). The last inequality follows from the fact that

$$
\begin{aligned}
&\left\langle \int_{\mathbb{R}^d} k(\cdot,x)\phi(x)(d\rho_{1,t}(x)-d\rho_{2,t}(x)), \int_{\mathbb{R}^d} k(\cdot,y)\phi(y)(d\rho_{1,t}(y)-d\rho_{2,t}(y)) \right\rangle_{\mathcal{H}_k^d}^{\frac{1}{2}} \\
&= \Bigg(\sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Big\langle k\left(u(t,x),\cdot\right)\phi\left(u(t,x)\right) - k\left(v(t,x),\cdot\right)\phi\left(v(t,x)\right), \\
&\qquad\qquad k\left(u(t,y),\cdot\right)\phi\left(u(t,y)\right) - k\left(v(t,y),\cdot\right)\phi\left(v(t,y)\right) \Big\rangle_{\mathcal{H}_k^d} d\rho_0(x) d\rho_0(y) \Bigg)^{\frac{1}{2}} \\
&\leq \sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d} \left\| k\left(u(t,x),\cdot\right)\phi\left(u(t,x)\right) - k\left(v(t,x),\cdot\right)\phi\left(v(t,x)\right) \right\|_{\mathcal{H}_k^d} \rho_0(x)dx \\
&\leq \sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d} \left\| \left(k\left(u(t,x),\cdot\right) - k\left(v(t,x),\cdot\right)\right)\phi\left(u(t,x)\right) \right\|_{\mathcal{H}_k^d} \rho_0(x)dx \\
&\quad + \sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d} \left\| k\left(v(t,x),\cdot\right)\left(\phi\left(u(t,x)\right) - \phi\left(v(t,x)\right)\right) \right\|_{\mathcal{H}_k^d} \rho_0(x)dx \\
&= \sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d} \left\| \left(k\left(u(t,x),\cdot\right) - k\left(v(t,x),\cdot\right)\right) \right\|_{\mathcal{H}_k} \left| \langle \phi(\cdot), k\left(u\left(t,x\right),\cdot\right)\rangle_{\mathcal{H}_k} \right| \rho_0(x)dx \\
&\quad + \sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d} \left\| k\left(v(t,x),\cdot\right) \right\|_{\mathcal{H}_k} \left| \langle \phi(\cdot), k\left(u(t,x),\cdot\right) - k\left(v(t,x),\cdot\right)\rangle_{\mathcal{H}_k} \right| \rho_0(x)dx \\
&\leq \sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d} \left\| \left(k\left(u(t,x),\cdot\right) - k\left(v(t,x),\cdot\right)\right) \right\|_{\mathcal{H}_k} \|\phi\|_{\mathcal{H}_k^d} \left\| k\left(u\left(t,x\right),\cdot\right) \right\|_{\mathcal{H}_k} \rho_0(x)dx \\
&\quad + \sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d} \left\| k\left(v(t,x),\cdot\right) \right\|_{\mathcal{H}_k} \|\phi\|_{\mathcal{H}_k^d} \left\| k\left(u(t,x),\cdot\right) - k\left(v(t,x),\cdot\right) \right\|_{\mathcal{H}_k} \rho_0(x)dx \\
&= \sup_{x\in\mathbb{R}^d} \Bigg(\sqrt{k\left(u(t,x),u(t,x)\right) + k\left(v(t,x),v(t,x)\right) - 2k\left(u(t,x),v(t,x)\right)} \\
&\qquad \times \left(\sqrt{k\left(u(t,x),u(t,x)\right)} + \sqrt{k\left(v(t,x),v(t,x)\right)} \right) \Bigg) \\
&\leq 2\|k\|_\infty \left(3\left\|\nabla^2 k\right\|_\infty + 2\|\nabla_1\nabla_2 k\|_\infty\right)^{\frac{1}{2}} d_S(u,v),
\end{aligned}
$$

where the first identity follows from the definition of $\rho_{i,t}$ for $i = 1,2$ and change of variable, the second inequality holds due to the symmetry in x and y, the first identity follows from the reproducing property of the RKHS, the last identity follows from the fact that $\|k(x,\cdot)\|_{\mathcal{H}_k} = \sqrt{k(x,x)}$ for all x and the last inequality follows from Assumption K1 and Taylor expansion on both variables in k up to second order.

Second term: Note that we have

$$\begin{aligned}
&\left|\left((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}}+\nu I\right)^{-1}\left(\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}(x)-\iota_{k,\rho_{2,t}}^{*}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\right)\right|\\
&=\left|\iota_{k,\rho_{2,t}}\left((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}}+\nu I\right)^{-1}\left(\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}(x)-\iota_{k,\rho_{2,t}}^{*}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\right)\right|\\
&\leq\left\|\iota_{k,\rho_{2,t}}\right\|_{\mathcal{H}_k^d\to L_\infty^d}\left\|\left((1-\nu)\iota_{k,\rho_{2,t}}^{*}\iota_{k,\rho_{2,t}}+\nu I\right)^{-1}\right\|_{\mathcal{H}_k^d\to\mathcal{H}_k^d}\left\|\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}-\iota_{k,\rho_{2,t}}^{*}\nabla\log\frac{\rho_{2,t}}{\pi}\right\|_{\mathcal{H}_k^d}\\
&\leq\|k\|_\infty\,\nu^{-1}\left\|\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}-\iota_{k,\rho_{2,t}}^{*}\nabla\log\frac{\rho_{2,t}}{\pi}\right\|_{\mathcal{H}_k^d},
\end{aligned}$$

where the last inequality follows from (4.27) and for all $x\in\mathbb{R}^d$,

$$\begin{aligned}
&\iota_{k,\rho_{1,t}}^{*}\nabla\log\frac{\rho_{1,t}}{\pi}(x)-\iota_{k,\rho_{2,t}}^{*}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\\
&=\int_{\mathbb{R}^d}k(x,y)\nabla\log\frac{\rho_{1,t}}{\pi}(y)d\rho_{1,t}(y)-\int_{\mathbb{R}^d}k(x,y)\nabla\log\frac{\rho_{2,t}}{\pi}d\rho_{2,t}(y)\\
&=\int_{\mathbb{R}^d}\left(k(x,y)\nabla V(y)-\nabla_2k(x,y)\right)d\rho_{1,t}(y)-\int_{\mathbb{R}^d}\left(k(x,y)\nabla V(y)-\nabla_2k(x,y)\right)d\rho_{2,t}(y)\\
&=\int_{\mathbb{R}^d}\left(k(x,u(t,y))\nabla V(u(t,y))-k(x,v(t,y))\nabla V(v(t,y))\right)d\rho_0(y)\\
&\qquad-\int_{\mathbb{R}^d}\left(\nabla_2k(x,u(t,y))-\nabla_2k(x,v(t,y))\right)d\rho_0(y).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\left\|\iota_{k,\rho_t}^{*}\nabla\log\frac{\rho_t}{\pi}-\iota_{k,\mu_t}^{*}\nabla\log\frac{\mu_t}{\pi}\right\|_{\mathcal{H}_k^d}\\
&\leq\int_{\mathbb{R}^d}\Bigg(\|k(\cdot,u(t,y))\nabla V(u(t,y))-k(\cdot,v(t,y))\nabla V(v(t,y))\|_{\mathcal{H}_k^d}\\
&\qquad+\|\nabla_2k(\cdot,u(t,y))-\nabla_2k(\cdot,v(t,y))\|_{\mathcal{H}_k^d}\Bigg)d\rho_0(y).
\end{aligned}$$

For simplicity, in the following calculations, we denote $u(t,y)$ and $v(t,y)$ as u and v respectively. We will bound $\|k(\cdot,u)\nabla V(u)-k(\cdot,v)\nabla V(v)\|_{\mathcal{H}_k^d}$ and $\|\nabla_2k(\cdot,u)-\nabla_2k(\cdot,v)\|_{\mathcal{H}_k^d}$ respectively. Note that we have

$$
\begin{aligned}
&\|k(\cdot,u)\nabla V(u)-k(\cdot,v)\nabla V(v)\|_{\mathcal{H}_k^d}^2\\
&=\langle k(\cdot,u)\nabla V(u)-k(\cdot,v)\nabla V(v),k(\cdot,u)\nabla V(u)-k(\cdot,v)\nabla V(v)\rangle_{\mathcal{H}_k^d}\\
&=|\nabla V(u)|^2k(u,u)-2\left\langle\nabla V(u),\nabla V(v)\right\rangle k(u,v)+|\nabla V(v)|^2k(v,v)\\
&\le|\langle\nabla V(u)-\nabla V(v),\nabla V(u)k(u,u)-\nabla V(v)k(v,v)\rangle|\\
&\quad+|\langle\nabla V(u),\nabla V(v)\rangle\left(k(u,u)+k(v,v)-2k(u,v)\right)|,
\end{aligned}
$$

where

$$
\begin{aligned}
&|\langle\nabla V(u)-\nabla V(v),\nabla V(u)k(u,u)-\nabla V(v)k(v,v)\rangle|\\
&\le|\nabla V(u)-\nabla V(v)|^2k(u,u)+|\nabla V(u)-\nabla V(v)|\,|\nabla V(v)|\,|k(u,u)-k(v,v)|\\
&\le C_{1,r}^2(1+V(y))^2d_{S_T}(u,v)^2k(u,u)+C_{1,r}^2(1+V(y))^2d_S(u,v)\,|k(u,u)-k(v,v)|\\
&\le C_{1,r}^2(1+V(y))^2d_S(u,v)^2\left\|k\right\|_\infty^2+2C_{1,r}^2\left\|\nabla k\right\|_\infty(1+V(y))^2d_S(u,v)^2.
\end{aligned}
$$

The second inequality follows from Assumption V1 and the last inequality follows from Assumption K1 and Taylor expansion on both variables in k up to first order. And, we also have

$$
\begin{aligned}
&|\langle\nabla V(u),\nabla V(v)\rangle\left(k(u,u)+k(v,v)-2k(u,v)\right)|\\
&\le C_{1,r}^2(1+V(y))^2\left|k(u,u)+k(v,v)-2k(u,v)\right|\\
&\le C_{1,r}^2(1+V(y))^2\left(3\left\|\nabla^2k\right\|_\infty+2\left\|\nabla_1\nabla_2k\right\|_\infty\right)d_S(u,v)^2,
\end{aligned}
$$

where the first inequality follows from Assumption V1 and the last inequality follows from Assumption K1 and Taylor expansion on both variables in k up to second order. With the above two inequalities, we have

$$
\begin{aligned}
&\|k(\cdot,u(t,y))\nabla V(u(t,y))-k(\cdot,v(t,y))\nabla V(v(t,y))\|_{\mathcal{H}_k^d}\\
&\le C_{1,r}\left(\|k\|_\infty+2\|\nabla k\|_\infty{}^{\frac{1}{2}}+3\left\|\nabla^2k\right\|_\infty{}^{\frac{1}{2}}+2\|\nabla_1\nabla_2k\|_\infty{}^{\frac{1}{2}}\right)d_S(u,v)(1+V(y)).
\end{aligned}
\tag{4.37}
$$

Observe that for all $x, y \in \mathbb{R}^d$,

$$\begin{aligned}\langle \nabla_2 k(\cdot, x), \nabla_2 k(\cdot, y)\rangle_{\mathcal{H}_k^d} &= \nabla_1 \cdot \nabla_2 \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}_k^d} \\ &= \nabla_1 \cdot \nabla_2 k(x, y),\end{aligned}$$

If we denote the function $\nabla_1 \cdot \nabla_2 k = D_{1,2}k$ where $D_{1,2}k$ is symmetric since k is symmetric, we get

$$\begin{aligned}\|\nabla_2 k(\cdot, u) - \nabla_2 k(\cdot, v)\|^2_{\mathcal{H}_k^d} &= D_{1,2}k(u, u) + D_{1,2}k(v, v) - 2D_{1,2}k(u, v) \\ &\leq \left(2\left\|\nabla^2(D_{1,2}k)\right\|_\infty + \|\nabla_1\nabla_2(D_{1,2}k)\|_\infty\right) d_S(u, v)^2.\end{aligned}$$

where the inequality follows from Taylor expansion on both variables of $D_{1,2}k$. Therefore

$$\|\nabla_2 k(\cdot, u) - \nabla_2 k(\cdot, v)\|_{\mathcal{H}_k^d} \leq \left(2\left\|\nabla^2(D_{1,2}k)\right\|_\infty^{\frac{1}{2}} + \|\nabla_1\nabla_2(D_{1,2}k)\|_\infty^{\frac{1}{2}}\right) d_S(u, v). \tag{4.38}$$

According to (4.37) and (4.38), we get

$$\left\|\iota^*_{k,\rho_t}\nabla \log \frac{\rho_t}{\pi} - \iota^*_{k,\mu_t}\nabla \log \frac{\mu_t}{\pi}\right\|_{\mathcal{H}_k^d} \leq L_r d_s(u, v)$$

with

$$\begin{aligned}L_r = C_{1,r}\|\rho_0\|_{\mathcal{P}_V}&\left(\|k\|_\infty + 2\|\nabla k\|_\infty{}^{\frac{1}{2}} + 3\left\|\nabla^2 k\right\|_\infty{}^{\frac{1}{2}} + 2\|\nabla_1\nabla_2 k\|_\infty{}^{\frac{1}{2}}\right) \\ &+ 2\left\|\nabla^2(D_{1,2}k)\right\|_\infty{}^{\frac{1}{2}} + \|\nabla_1\nabla_2(D_{1,2}k)\|_\infty{}^{\frac{1}{2}}.\end{aligned} \tag{4.39}$$

Therefore, the second term is bounded as

$$\left|\left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}} + \nu I\right)^{-1}\left(\iota^*_{k,\rho_{1,t}}\nabla \log \frac{\rho_{1,t}}{\pi}(x) - \iota^*_{k,\rho_{2,t}}\nabla \log \frac{\rho_{2,t}}{\pi}(x)\right)\right| \leq \nu^{-1}L_r\|k\|_\infty d_S(u, v). \tag{4.40}$$

With (4.36) and (4.40), we obtain

$$\sup_{x\in\mathbb{R}^d}\left|\mathcal{D}_{\nu,\rho_{1,t}}\nabla \log \frac{\rho_{1,t}}{\pi}(x) - \mathcal{D}_{\nu,\rho_{2,t}}\nabla \log \frac{\rho_{2,t}}{\pi}(x)\right| \leq C_1(t)d_S(u, v)$$

with $C_1(t)$ being defined in (4.35). ■

4.7. Stability

In this section we prove a stability estimate for the weak solutions to (4.6). To do this, we introduce a space of probability measures on $\mathbb{R}^d$ and assumptions on V as follows,

$$\mathcal{P}_p := \left\{ \rho \in \mathcal{P} : \|\rho\|_{\mathcal{P}_p} := \int_{\mathbb{R}^d} |x|^p \rho(x) dx < \infty \right\},$$

where $\mathcal{P}$ denotes the set of all probability measures on $\mathbb{R}^d$.

Assumption V2. In addition to Assumption V1, there exists a constant $C_V > 0$ and $q > 1$ such that $|\nabla V(x)|^q \leq C_V(1 + V(x))$ for all $x \in \mathbb{R}^d$ and $\sup_{\theta \in [0,1]} \left\| \nabla^2 V(\theta x + (1-\theta)y) \right\|_2^q \leq C_V(1 + V(x) + V(y))$.

Theorem 26. Let V satisfy Assumption V2 with $q \in (1, \infty)$ and k satisfies Assumption K1. Let p be the conjugate of q, i.e., $p^{-1} + q^{-1} = 1$. Let $\rho_1, \rho_2 \in \mathcal{P}_p$ be two initial probability measures satisfying $\|\rho_i\|_{\mathcal{P}_p} \leq R$ for some constant $R > 0$ and $i = 1, 2$. Let $\rho_{1,t}$ and $\rho_{2,t}$ be the associated weak solution to (4.6). Then given any $T > 0$, there exists a constant $C > 0$ depending on $k, V, q, \nu, \rho_1, \rho_2$ such that

$$\sup_{t \in [0,T]} \mathcal{W}_p(\rho_{1,t}, \rho_{2,t}) \leq C \mathcal{W}_p(\rho_1, \rho_2).$$

More explicitly, the constant C is given by

$$C = \exp\Bigg(\nu^{-1} \|k\|_\infty C(T, k, V, \nu, \rho_1, \rho_2, q) T + + \nu^{-\frac{1}{2}} \left(2 \|\nabla_1 \nabla_2 k\|_\infty + 3 \left\|\nabla^2 k\right\|_\infty \right)^{\frac{1}{2}} \sqrt{\mathrm{KL}(\rho_2|\pi)} T^{\frac{1}{2}}$$

(4.41)

$$+ 2\nu^{-\frac{3}{2}}(1-\nu) \|k\|_\infty^2 \left(2 \left\| \nabla_1 \nabla_2 k + 3 \left\|\nabla^2 k\right\|_\infty \right\|_\infty \right)^{\frac{1}{2}} \sqrt{\mathrm{KL}(\rho_1|\pi)} T^{\frac{1}{2}} \Bigg),$$

where $C(T, k, V, \nu, \rho_1, \rho_2, q)$ is defined in (4.46).

Remark 34. If we focus on the dependency on ν and T in (4.41), we have

$$C \leq C' \exp\left(\nu^{-1} T \exp\left(C' \nu^{-\frac{1}{2}} T^{\frac{1}{2}} \right) + \nu^{-\frac{3}{2}}(1-\nu) T^{\frac{1}{2}} + \nu^{-\frac{1}{2}} T^{\frac{1}{2}} \right)$$

where C' is a constant independent of ν and T.

The proof is inspired by that of [LLN19, Theorem 2.7] which in turn is motivated by the Dobrushin's coupling argument (see, for example, [Dob79] and [MRZ16, Theorem 1.4.1]). In the following proof we mainly highlight the parts of our proof that are different from the proof of [LLN19, Theorem 2.7].

Proof of Theorem 26. First, under Assumption V2, there exists a constant $C_0 > 0$ such that $V(x) \leq C_0(1+|x|^p)$ for all $x \in \mathbb{R}^d$. Therefore, $\mathcal{P}_p \subset \mathcal{P}_V$ and $\|\rho_i\|_{\mathcal{P}_V} \leq C(R) < \infty$ for $i = 1, 2$. By Theorem 25, the weak solutions take the form

$$\rho_{i,t} = (\Phi(t, \cdot, \rho_i))_{\rho_i}, \quad i = 1, 2$$

where $\Phi(\cdot, \cdot, \rho_i)$ solves (4.24) with $\rho_0 = \rho_i$. Let π^0 be a coupling measure between ρ_1 and ρ_2. For $\delta > 0$, define $\phi_\delta(x) = \frac{1}{p}(|x|^2 + \delta)^{p/2}$ which satisfies

$$\lim_{\delta \to 0^+} \phi_\delta(x) = \frac{1}{p}|x|^p \text{ and } |\nabla \phi_\delta(x)| \leq |x|^{p-1}, \quad \text{for all} \quad x \in \mathbb{R}^d.$$

We start from estimating the derivative of ϕ_δ in the time variable, for which we have

$$\begin{aligned} &\partial_t \phi_\delta(\Phi(t, x_1, \rho_1) - \Phi(t, x_2, \rho_2)) \\ =& -\nabla \phi_\delta(\Phi(t, x_1, \rho_1) - \Phi(t, x_2, \rho_2)) \left(\mathcal{D}_{\nu, \rho_{1,t}} \nabla \log \frac{\rho_{1,t}}{\pi}(\Phi(t, x_1, \rho_1)) - \mathcal{D}_{\nu, \rho_{2,t}} \nabla \log \frac{\rho_{2,t}}{\pi}(\Phi(t, x_2, \rho_2)) \right). \end{aligned}$$

The next step is to estimate

$$\left| \mathcal{D}_{\nu, \rho_{1,t}} \nabla \log \frac{\rho_{1,t}}{\pi}(\Phi(t, x_1, \rho_1)) - \mathcal{D}_{\nu, \rho_{2,t}} \nabla \log \frac{\rho_{2,t}}{\pi}(\Phi(t, x_2, \rho_2)) \right|.$$

Note that

$$\mathcal{D}_{\nu, \rho_{1,t}} \nabla \log \frac{\rho_{1,t}}{\pi}(\Phi(t, x_1, \rho_1)) - \mathcal{D}_{\nu, \rho_{2,t}} \nabla \log \frac{\rho_{2,t}}{\pi}(\Phi(t, x_2, \rho_2)) := I_1 + I_2,$$

where

$$\begin{aligned} I_1 &:= \mathcal{D}_{\nu, \rho_{1,t}} \nabla \log \frac{\rho_{1,t}}{\pi}(\Phi(t, x_1, \rho_1)) - \mathcal{D}_{\nu, \rho_{2,t}} \nabla \log \frac{\rho_{2,t}}{\pi}(\Phi(t, x_1, \rho_1)), \\ I_2 &:= \mathcal{D}_{\nu, \rho_{2,t}} \nabla \log \frac{\rho_{2,t}}{\pi}(\Phi(t, x_1, \rho_1)) - \mathcal{D}_{\nu, \rho_{2,t}} \nabla \log \frac{\rho_{2,t}}{\pi}(\Phi(t, x_2, \rho_2)). \end{aligned}$$

According to Proposition 23, we have

$$
\begin{aligned}
&|-\nabla\phi_\delta(\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2))\cdot I_1|\\
&\le 2\nu^{-\frac{3}{2}}(1-\nu)\left\|k\right\|_\infty^2\left(2\left\|\nabla_1\nabla_2 k\right\|_\infty+3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} I_{\nu,Stein}(\rho_{1,t}|\pi)^{\frac{1}{2}}|\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2)|^{p-1}\\
&\qquad\qquad\times\left(\int_{\mathbb{R}^d}|\Phi(t,y_1,\rho_1)-\Phi(t,y_2,\rho_2)|\pi^0(dy_1,dy_2)\right)\\
&\quad+\nu^{-1}\left\|k\right\|_\infty C(t,k,V,\nu,\rho_1,\rho_2,q)|\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2)|^{p-1}\\
&\qquad\qquad\times\left(\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi(t,y_1,\rho_1)-\Phi(t,y_2,\rho_2)|^p\pi^0(dy_1,dy_2)\right)^{1/p}
\end{aligned}
$$

and

$$
\begin{aligned}
&|-\nabla\phi_\delta(\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2))\cdot I_2|\\
&\quad\le\nu^{-\frac{1}{2}}\left(2\left\|\nabla_1\nabla_2 k\right\|_\infty+3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} I_{\nu,Stein}(\rho_{2,t}|\pi)^{\frac{1}{2}}|\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2)|^p.
\end{aligned}
$$

Now, defining

$$
D_p(\pi)(s):=\left(\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi(s,y_1,\rho_1)-\Phi(s,y_2,\rho_2)|^p\pi^0(dy_1,dy_2)\right)^{1/p},
$$

we have, for any $t\in[0,T]$ that

$$
\begin{aligned}
&\phi_\delta(\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2))\\
&\le\nu^{-\frac{1}{2}}\left(2\left\|\nabla_1\nabla_2 k\right\|_\infty+3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}}\int_0^t I_{\nu,Stein}(\rho_{2,s}|\pi)^{\frac{1}{2}}|\Phi(s,x_1,\rho_1)-\Phi(s,x_2,\rho_2)|^p ds\\
&+\phi_\delta(x_1-x_2)+\nu^{-1}\left\|k\right\|_\infty C(T,k,V,\nu,\rho_1,\rho_2,q)\int_0^t|\Phi(s,x_1,\rho_1)-\Phi(s,x_2,\rho_2)|^{p-1}D_p(\pi^0)(s)ds\\
&+2\nu^{-\frac{3}{2}}(1-\nu)\left\|k\right\|_\infty^2\left(2\left\|\nabla_1\nabla_2 k\right\|_\infty+3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}}\\
&\qquad\qquad\times\int_0^t I_{\nu,Stein}(\rho_{1,s}|\pi)^{\frac{1}{2}}|\Phi(s,x_1,\rho_1)-\Phi(s,x_2,\rho_2)|^{p-1}D_p(\pi^0)(s)ds.
\end{aligned}
$$

Integrating the above inequality w.r.t. the coupling π^0, and using the fact that

$$
\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2)|^{p-1}\pi^0(dx_1,dx_2)\le D_p(\pi^0)(s)^{p-1},
$$

and letting $\delta \to 0$, we get

$$D_p(\pi^0)(t)^p \le D_p(\pi^0)(0)^p + \nu^{-1} \|k\|_\infty C(T,k,V,\nu,\rho_1,\rho_2,q) \int_0^t D_p(\pi^0)(s)^p ds$$
$$+ 2\nu^{-\frac{3}{2}}(1-\nu) \|k\|_\infty^2 \left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \int_0^t I_{\nu,Stein}(\rho_{1,s}|\pi)^{\frac{1}{2}} D_p(\pi^0)(s)^p ds$$
$$+ \left(\nu^{-\frac{1}{2}}\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}}\right) \int_0^t I_{\nu,Stein}(\rho_{2,s}|\pi)^{\frac{1}{2}} D_p(\pi^0)(s)^p ds.$$

By using Gronwall's inequality, we further obtain

$$D_p(\pi^0)(t)^p$$
$$\le D_p(\pi^0)(0)^p \exp\left(\nu^{-1}\|k\|_\infty C(T,k,V,\nu,\rho_1,\rho_2,q)t\right)$$
$$\times \exp\left(2\nu^{-\frac{3}{2}}(1-\nu)\|k\|_\infty^2 \left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \int_0^t I_{\nu,Stein}(\rho_{1,s}|\pi)^{\frac{1}{2}} ds\right)$$
$$\times \exp\left(\nu^{-\frac{1}{2}}\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \int_0^t I_{\nu,Stein}(\rho_{2,s}|\pi)^{\frac{1}{2}} ds\right)$$
$$\le D_p(\pi^0)(0)^p \exp\Bigg(\nu^{-\frac{1}{2}}\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \sqrt{\mathrm{KL}(\rho_2|\pi)} t^{\frac{1}{2}}$$
$$+ \nu^{-1}\|k\| C(T,k,V,\nu,\rho_1,\rho_2,q)t + 2\nu^{-\frac{3}{2}}(1-\nu)\|k\|_\infty^2\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \sqrt{\mathrm{KL}(\rho_1|\pi)} t^{\frac{1}{2}}\Bigg).$$

Hence, we obtain

$$\mathcal{W}_p^p(\rho_{1,t},\rho_{2,t})$$
$$= \inf_{\pi\in\Gamma(\rho_{1,t},\rho_{2,t})} \int_{\mathbb{R}^d\times\mathbb{R}^d} |x_1-x_2|^p \pi(dx_1,dx_2) \le \inf_{\pi^0\in\Gamma(\rho_1,\rho_2)} D_p(\pi^0)(t)^p$$
$$\le \exp\Bigg(\nu^{-1}\|k\|_\infty C(T,k,V,\nu,\rho_1,\rho_2,q)t + \nu^{-\frac{1}{2}}\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \sqrt{\mathrm{KL}(\rho_2|\pi)} t^{\frac{1}{2}}$$
$$+ 2\nu^{-\frac{3}{2}}(1-\nu)\|k\|_\infty^2\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \sqrt{\mathrm{KL}(\rho_1|\pi)} t^{\frac{1}{2}}\Bigg) \inf_{\pi^0\in\Gamma(\rho_1,\rho_2)} D_p(\pi^0)(0)^p$$
$$= \exp\Bigg(\nu^{-1}\|k\|_\infty C(T,k,V,\nu,\rho_1,\rho_2,q)t + \nu^{-\frac{1}{2}}\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \sqrt{\mathrm{KL}(\rho_2|\pi)} t^{\frac{1}{2}}$$
$$+ 2\nu^{-\frac{3}{2}}(1-\nu)\|k\|_\infty^2\left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \sqrt{\mathrm{KL}(\rho_1|\pi)} t^{\frac{1}{2}}\Bigg) \mathcal{W}_p^p(\rho_1,\rho_2),$$

yielding the result. ■

Proposition 23. Under the assumption of Theorem 26, let $\Phi(\cdot,\cdot,\rho_i)$ be the solution to (4.24) with $\rho_0=\rho_i$ for $i=1,2$. Let π^0 be a probability measure on $\mathbb{R}^{2d}$ with marginals ρ_1 and ρ_2. Then we have

$$\begin{aligned}
&\left|\mathcal{D}_{\nu,\rho_{1,t}}\nabla\log\frac{\rho_{1,t}}{\pi}(\Phi(t,x_1,\rho_1))-\mathcal{D}_{\nu,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(\Phi(t,x_1,\rho_1))\right|\\
&\le 2\nu^{-\frac{3}{2}}(1-\nu)\left\|k\right\|_\infty^2\left(2\left\|\nabla_1\nabla_2k\right\|_\infty+3\left\|\nabla^2k\right\|_\infty\right)^{\frac{1}{2}}I_{\nu,Stein}(\rho_{1,t}|\pi)^{\frac{1}{2}}\\
&\qquad\times\int_{\mathbb{R}^d}|\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2)|\pi^0(dx_1,dx_2)\\
&\quad+\nu^{-1}\left\|k\right\|_\infty C(t,k,V,\nu,\rho_1,\rho_2,q)\left(\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi(t,y_1,\rho_1)-\Phi(t,y_2,\rho_2)|^p\,\pi^0(dy_1,dy_2)\right)^{1/p},
\end{aligned}\tag{4.42}$$

and

$$\begin{aligned}
&\left|\mathcal{D}_{\nu,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(\Phi(t,x_1,\rho_1))-\mathcal{D}_{\nu,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(\Phi(t,x_2,\rho_2))\right|\\
&\le\nu^{-\frac{1}{2}}\left(2\left\|\nabla_1\nabla_2k\right\|_\infty+3\left\|\nabla^2k\right\|_\infty\right)^{\frac{1}{2}}I_{\nu,Stein}(\rho_{2,t}|\pi)^{\frac{1}{2}}\left|\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2)\right|,
\end{aligned}\tag{4.43}$$

where $C(t,k,V,\nu,\rho_1,\rho_2,q)$ is given in (4.46).

Proof of Proposition 23. First we prove (4.43). For any $x\in\mathbb{R}^d$,

$$\begin{aligned}
&\left|\mathcal{D}_{\nu,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(x)-\mathcal{D}_{\nu,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(y)\right|\\
&=\left|\left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}}+\nu I_d\right)^{-1}\iota^*_{k,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\right.\\
&\qquad\left.-\left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}}+\nu I_d\right)^{-1}\iota^*_{k,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(y)\right|\\
&=\left|\left\langle k(x,\cdot)-k(y,\cdot),\left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}}+\nu I_d\right)^{-1}\iota^*_{k,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(\cdot)\right\rangle_{\mathcal{H}_k}\right|\\
&\le\left\|k(x,\cdot)-k(y,\cdot)\right\|_{\mathcal{H}_k}\left\|\left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}}+\nu I_d\right)^{-1}\iota^*_{k,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}\right\|_{\mathcal{H}_k^d}\\
&\le\nu^{-\frac{1}{2}}I_{\nu,Stein}(\rho_{2,t}|\pi)^{\frac{1}{2}}\left\|k(x,\cdot)-k(y,\cdot)\right\|_{\mathcal{H}_k^d}\\
&\le\nu^{-\frac{1}{2}}I_{\nu,Stein}(\rho_{2,t}|\pi)^{\frac{1}{2}}\left(2\left\|\nabla_1\nabla_2k\right\|_\infty+3\left\|\nabla^2k\right\|_\infty\right)^{\frac{1}{2}}|x-y|,
\end{aligned}$$

where the second inequality follows from (4.28) and the last inequality follows from the reproducing property and Taylor expansion. The claim in (4.43) then follows from the above inequality.

To prove (4.42), first according to the proof of Lemma 4.6.3, for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}
&\left|\mathcal{D}_{\nu,\rho_{1,t}} \nabla \log \frac{\rho_{1,t}}{\pi}(x) - \mathcal{D}_{\nu,\rho_{2,t}} \nabla \log \frac{\rho_{2,t}}{\pi}(x)\right| \\
&\leq \left|\left(\left((1-\nu)\iota^*_{k,\rho_{1,t}}\iota_{k,\rho_{1,t}} + \nu I_d\right)^{-1} - \left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}} + \nu I_d\right)^{-1}\right)\iota^*_{k,\rho_{1,t}} \nabla \log \frac{\rho_{1,t}}{\pi}(x)\right| \\
&+ \left|\left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}} + \nu I_d\right)^{-1}\left(\iota^*_{k,\rho_{1,t}} \nabla \log \frac{\rho_{1,t}}{\pi}(x) - \iota^*_{k,\rho_{2,t}} \nabla \log \frac{\rho_{2,t}}{\pi}(x)\right)\right|
\end{aligned}$$

and first term on the right hand side is bounded as

$$\begin{aligned}
&\left|\left(\left((1-\nu)\iota^*_{k,\rho_{1,t}}\iota_{k,\rho_{1,t}} + \nu I_d\right)^{-1} - \left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}} + \nu I_d\right)^{-1}\right)\iota^*_{k,\rho_{1,t}} \nabla \log \frac{\rho_{1,t}}{\pi}(x)\right| \\
&\leq \|k\|_\infty \nu^{-\frac{3}{2}}(1-\nu) I_{\nu,Stein}(\rho_{1,t}|\pi)^{\frac{1}{2}} \\
&\quad \sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \left\langle \int_{\mathbb{R}^d} k(\cdot,x)\phi(x)(d\rho_{1,t}(x) - d\rho_{2,t}(x)), \int_{\mathbb{R}^d} k(\cdot,y)\phi(y)(d\rho_{1,t}(y) - d\rho_{2,t}(y)) \right\rangle_{\mathcal{H}_k^d}^{\frac{1}{2}}
\end{aligned}$$

By Theorem 25, the weak solutions to (4.6) take the form

$$\rho_{i,t} = (\Phi(t,\cdot,\rho_i))_{\rho_i}, \quad i = 1, 2.$$

Similar to the proof in Lemma 4.6.3, we have

$$\begin{aligned}
&\sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \left\langle \int_{\mathbb{R}^d} k(\cdot,x)\phi(x)(d\rho_{1,t}(x) - d\rho_{2,t}(x)), \int_{\mathbb{R}^d} k(\cdot,y)\phi(y)(d\rho_{1,t}(y) - d\rho_{2,t}(y)) \right\rangle_{\mathcal{H}_k^d}^{\frac{1}{2}} \\
&= \Bigg(\sup_{\|\phi\|_{\mathcal{H}_k^d}=1} \int_{\mathbb{R}^d\times\mathbb{R}^d} \int_{\mathbb{R}^d\times\mathbb{R}^d} \Big\langle k\left(\Phi(t,x_1,\rho_1),\cdot\right)\phi\left(\Phi(t,x_1,\rho_1)\right) - k\left(\Phi(t,x_2,\rho_2),\cdot\right)\phi\left(\Phi(t,x_2,\rho_2)\right), \\
&k\left(\Phi(t,y_1,\rho_1),\cdot\right)\phi\left(\Phi(t,y_1,\rho_1)\right) - k\left(\Phi(t,y_2,\rho_2),\cdot\right)\phi\left(\Phi(t,y_2,\rho_2)\right) \Big\rangle_{\mathcal{H}_k^d} \pi^0(dx_1,dx_2)\pi^0(dy_1,y_2) \Bigg)^{\frac{1}{2}} \\
&\leq \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \left\|k\left(\Phi(t,x_1,\rho_1),\cdot\right)\phi\left(\Phi(t,x_1,\rho_1)\right) - k\left(\Phi(t,x_2,\rho_2),\cdot\right)\phi\left(\Phi(t,x_2,\rho_2)\right)\right\|_{\mathcal{H}_k^d} \pi^0(dx_1,dx_2) \\
&\leq 2\|k\|_\infty \left(2\|\nabla_1\nabla_2 k\|_\infty + 3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} \int_{\mathbb{R}^d\times\mathbb{R}^d} |\Phi(t,x_1,\rho_1) - \Phi(t,x_2,\rho_2)|\pi^0(dx_1,dx_2).
\end{aligned}$$

Therefore

$$\begin{aligned}
&\left|\left((\iota^*_{k,\rho_{1,t}}\iota_{k,\rho_{1,t}}+\nu I)^{-1}-(\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}}+\nu I)^{-1}\right)\iota^*_{k,\rho_{1,t}}\nabla\log\frac{\rho_{1,t}}{\pi}(x)\right|\\
&\quad\le 2\nu^{-\frac{3}{2}}(1-\nu)\left\|k\right\|_\infty^2\left(2\left\|\nabla_1\nabla_2 k\right\|_\infty+3\left\|\nabla^2 k\right\|_\infty\right)^{\frac{1}{2}} I_{\nu,Stein}(\rho_{1,t}|\pi)^{\frac{1}{2}}\\
&\qquad\times\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi(t,x_1,\rho_1)-\Phi(t,x_2,\rho_2)|\pi^0(dx_1,dx_2).
\end{aligned}$$

According to the proof of Lemma 4.6.3, the second term is bounded as

$$\begin{aligned}
&\left|\left((1-\nu)\iota^*_{k,\rho_{2,t}}\iota_{k,\rho_{2,t}}+\nu I\right)^{-1}\left(\iota^*_{k,\rho_{1,t}}\nabla\log\frac{\rho_{1,t}}{\pi}(x)-\iota^*_{k,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\right)\right|\\
&\quad\le\nu^{-1}\left\|k\right\|_\infty\left\|\iota^*_{k,\rho_{1,t}}\nabla\log\frac{\rho_{1,t}}{\pi}-\iota^*_{k,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}\right\|_{\mathcal{H}_k^d}.
\end{aligned}$$

For the factor $\left\|\iota^*_{k,\rho_{1,t}}\nabla\log\frac{\rho_{1,t}}{\pi}-\iota^*_{k,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}\right\|_{\mathcal{H}_k^d}$, notice that

$$\begin{aligned}
&\iota^*_{k,\rho_{1,t}}\nabla\log\frac{\rho_{1,t}}{\pi}(x)-\iota^*_{k,\rho_{2,t}}\nabla\log\frac{\rho_{2,t}}{\pi}(x)\\
&=\int_{\mathbb{R}^d}k(x,y)\nabla\log\frac{\rho_{1,t}}{\pi}(y)d\rho_{1,t}(y)-\int_{\mathbb{R}^d}k(x,y)\nabla\log\frac{\rho_{2,t}}{\pi}d\rho_{2,t}(y)\\
&=\int_{\mathbb{R}^d}\left(k(x,y)\nabla V(y)-\nabla_2 k(x,y)\right)d\rho_{1,t}(y)-\int_{\mathbb{R}^d}\left(k(x,y)\nabla V(y)-\nabla_2 k(x,y)\right)d\rho_{2,t}(y)\\
&=\int_{\mathbb{R}^d\times\mathbb{R}^d}\left(k(x,\Phi(t,y_1,\rho_1))\nabla V(\Phi(t,y_1,\rho_1))-k(x,\Phi(t,y_2,\rho_2))\nabla V(\Phi(t,y_2,\rho_2))\right)d\pi^0(dy_1,dy_2)\\
&\quad-\int_{\mathbb{R}^d\times\mathbb{R}^d}\left(\nabla_2 k(x,\Phi(t,y_1,\rho_1))-\nabla_2 k(x,\Phi(t,y_2,\rho_2))\right)d\pi^0(dy_1,dy_2)
\end{aligned}$$

and we get

$$\begin{aligned}
&\left\|\iota^*_{k,\rho_t}\nabla\log\frac{\rho_t}{\pi}-\iota^*_{k,\mu_t}\nabla\log\frac{\mu_t}{\pi}\right\|_{\mathcal{H}_k^d}\\
&\le\int_{\mathbb{R}^d\times\mathbb{R}^d}\left\|k(\cdot,\Phi(t,y_1,\rho_1))\nabla V(\Phi(t,y_1,\rho_1))-k(\cdot,\Phi(t,y_2,\rho_2))\nabla V(\Phi(t,y_2,\rho_2))\right\|_{\mathcal{H}_k^d}d\pi^0(dy_1,dy_2)\\
&\quad+\int_{\mathbb{R}^d\times\mathbb{R}^d}\left\|\nabla_2 k(\cdot,\Phi(t,y_1,\rho_1))-\nabla_2 k(\cdot,\Phi(t,y_2,\rho_2))\right\|_{\mathcal{H}_k^d}d\pi^0(dy_1,dy_2)
\end{aligned}$$

For simplicity, we denote $\Phi(t,y_1,\rho_1),\Phi(t,y_2,\rho_2)$ as Φ_1 and Φ_2 respectively in the following calculations. We will bound the two integrals separately.

First integral: Similar to the proof in Lemma 4.6.3, we have

$$\begin{aligned}
&\|k(\cdot,\Phi_1)\nabla V(\Phi_1) - k(\cdot,\Phi_2)\nabla V(\Phi_2)\|^2_{\mathcal{H}_k^d} \\
&= \langle \nabla V(\Phi_1) - \nabla V(\Phi_2), \nabla V(\Phi_1)k(\Phi_1,\Phi_1) - \nabla V(\Phi_2)k(\Phi_2,\Phi_2)\rangle \\
&\quad + \langle \nabla V(\Phi_1), \nabla V(\Phi_2)\rangle \left(k(\Phi_1,\Phi_1) + k(\Phi_2,\Phi_2) - 2k(\Phi_1,\Phi_2)\right)
\end{aligned}$$

where

$$\begin{aligned}
&|\langle \nabla V(\Phi_1) - \nabla V(\Phi_2), \nabla V(\Phi_1)k(\Phi_1,\Phi_1) - \nabla V(\Phi_2)k(\Phi_2,\Phi_2)\rangle| \\
&\le |\nabla V(\Phi_1) - \nabla V(\Phi_2)|^2 k(\Phi_1,\Phi_1) + |\nabla V(\Phi_1) - \nabla V(\Phi_2)||\nabla V(\Phi_2)||k(\Phi_1,\Phi_1) - k(\Phi_2,\Phi_2)| \\
&\le \sup_{\theta\in[0,1]} \left\|\nabla^2(\theta\Phi_1 + (1-\theta)\Phi_2)\right\|_2^2 |\Phi_1 - \Phi_2|^2 \|k\|_\infty^2 \\
&\quad + \sup_{\theta\in[0,1]} \left\|\nabla^2(\theta\Phi_1 + (1-\theta)\Phi_2)\right\|_2 |\Phi_1 - \Phi_2| C_V^{q^{-1}} (1 + V(\Phi_2))^{1/q} |k(\Phi_1,\Phi_1) - k(\Phi_2,\Phi_2)| \\
&\le C_V^{2/q}(1 + V(\Phi_1) + V(\Phi_2))^{2/q} \|k\|_\infty^2 |\Phi_1 - \Phi_2|^2 \\
&\quad + C_V^{1/q}(1 + V(\Phi_1) + V(\Phi_2))^{1/q} C_V^{q^{-1}} (1 + V(\Phi_2))^{1/q} \|\nabla k\|_\infty |\Phi_1 - \Phi_2|^2 \\
&\le C_V^{2/q}\left(\|k\|_\infty^2 + \|\nabla k\|_\infty\right)(1 + V(\Phi_1) + V(\Phi_2))^{2/q} |\Phi_1 - \Phi_2|^2
\end{aligned}$$

The third inequality follows from Assumption V2 and the last inequality follows from Assumption K1 and Taylor expansion on both variables in k up to first order. Furthermore, we have

$$\begin{aligned}
&|\langle \nabla V(\Phi_1), \nabla V(\Phi_2)\rangle \left(k(\Phi_1,\Phi_1) + k(\Phi_2,\Phi_2) - 2k(\Phi_1,\Phi_2)\right)| \\
&\le C_V^{2/q}(1 + V(\Phi_1))^{1/q}(1 + V(\Phi_2))^{1/q} |k(\Phi_1,\Phi_1) + k(\Phi_2,\Phi_2) - 2k(\Phi_1,\Phi_2)| \\
&\le C_V^{2/q}(1 + V(\Phi_1))^{1/q}(1 + V(\Phi_2))^{1/q}\left(3\left\|\nabla^2 k\right\|_\infty + 2\left\|\nabla_1\nabla_2 k\right\|_\infty\right)|\Phi_1 - \Phi_2|^2,
\end{aligned}$$

where the first inequality follows from Assumption V2 and the last inequality follows from Assumption K1 and Taylor expansion on both variables in k up to second order.

With the above two inequalities, we get

$$\begin{aligned}
&\int_{\mathbb{R}^d\times\mathbb{R}^d} \|k(\cdot,\Phi(t,y_1,\rho_1))\nabla V(\Phi(t,y_1,\rho_1)) - k(\cdot,\Phi(t,y_2,\rho_2))\nabla V(\Phi(t,y_2,\rho_2))\|_{\mathcal{H}_k^d}\, d\pi^0(dy_1,dy_2)\\
&\le C_V^{1/q}\left(\|k\|_\infty + \|\nabla k\|_\infty{}^{\frac{1}{2}} + 3\left\|\nabla^2 k\right\|_\infty{}^{\frac{1}{2}} + 2\|\nabla_1\nabla_2 k\|_\infty{}^{\frac{1}{2}}\right)\\
&\qquad \times \int_{\mathbb{R}^d\times\mathbb{R}^d} |\Phi_1(t,y_1,\rho_1) - \Phi(t,y_2,\rho_2)|(1+V(\Phi(t,y_1,\rho_1)) + V(\Phi(t,y_2,\rho_2)))^{1/q}\pi^0(dy_1,dy_2)\\
&\le C_V^{1/q}\left(\|k\|_\infty + \|\nabla k\|_\infty{}^{\frac{1}{2}} + 3\left\|\nabla^2 k\right\|_\infty{}^{\frac{1}{2}} + 2\|\nabla_1\nabla_2 k\|_\infty{}^{\frac{1}{2}}\right)\\
&\qquad \times \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |\Phi_1(t,y_1,\rho_1) - \Phi(t,y_2,\rho_2)|^p\pi^0(dy_1,dy_2)\right)^{1/p}\\
&\qquad\quad \times \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} 1+V(\Phi(t,y_1,\rho_1)) + V(\Phi(t,y_2,\rho_2))\pi^0(dy_1,dy_2)\right)^{1/q}\\
&\le \Big(\|\rho_1\|_{\mathcal{P}_V}\exp(C_{1,0}\nu^{-1/2}q^{-1}\,\|k\|_\infty\sqrt{tKL(\rho_1|\pi)})\\
&\qquad\qquad + \|\rho_2\|_{\mathcal{P}_V}\exp(C_{1,0}\nu^{-1/2}q^{-1}\,\|k\|_\infty\sqrt{tKL(\rho_2|\pi))}\Big)^{1/q}\\
&\qquad \times C_V^{1/q}\left(\|k\|_\infty + \|\nabla k\|_\infty{}^{\frac{1}{2}} + 3\left\|\nabla^2 k\right\|_\infty{}^{\frac{1}{2}} + 2\|\nabla_1\nabla_2 k\|_\infty{}^{\frac{1}{2}}\right)\\
&\qquad\quad \times \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |\Phi_1(t,y_1,\rho_1) - \Phi(t,y_2,\rho_2)|^p\pi^0(dy_1,dy_2)\right)^{1/p}\\
&\le 3C_V^{1/q}\left(\|k\|_\infty + \|\nabla k\|_\infty{}^{\frac{1}{2}} + \left\|\nabla^2 k\right\|_\infty{}^{\frac{1}{2}} + \|\nabla_1\nabla_2 k\|_\infty{}^{\frac{1}{2}}\right)\left(\|\rho_1\|_{\mathcal{P}_V} + \|\rho_2\|_{\mathcal{P}_V}\right)^{1/q}\\
&\qquad \times \exp(C_{1,0}\nu^{-1/2}q^{-1}\,\|k\|_\infty\sqrt{t(\mathrm{KL}(\rho_1|\pi) + \mathrm{KL}(\rho_2|\pi))})\\
&\qquad\quad \times \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |\Phi_1(t,y_1,\rho_1) - \Phi(t,y_2,\rho_2)|^p\pi^0(dy_1,dy_2)\right)^{1/p}\\
&:= C_1(k,V)\left(\|\rho_1\|_{\mathcal{P}_V} + \|\rho_2\|_{\mathcal{P}_V}\right)^{\frac{1}{q}}\exp\left(D_1(k,\nu,q)\,(\mathrm{KL}(\rho_1|\pi) + \mathrm{KL}(\rho_2|\pi))^{\frac{1}{2}}\,t^{\frac{1}{2}}\right)
\end{aligned}$$

(4.44)

$$\times \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |\Phi_1(t,y_1,\rho_1) - \Phi(t,y_2,\rho_2)|^p\pi^0(dy_1,dy_2)\right)^{1/p}.$$

Second integral: Denoting the function $\nabla_1\cdot\nabla_2 k = D_{1,2}k$, we first note that $D_{1,2}k$ is symmetric since k is symmetric. According to the above identity, we get

$$\|\nabla_2 k(\cdot,\Phi_1) - \nabla_2 k(\cdot,\Phi_2)\|^2_{\mathcal{H}_k^d} = D_{1,2}k(\Phi_1,\Phi_1) + D_{1,2}k(\Phi_2,\Phi_2) - 2D_{1,2}k(\Phi_1,\Phi_2)$$

Applying Taylor's series expansion on both variables of $D_{1,2}k$, we get

$$\|\nabla_2 k(\cdot,\Phi_1) - \nabla_2 k(\cdot,\Phi_2)\|^2_{\mathcal{H}_k^d} \le \left(2\left\|\nabla^2(D_{1,2}k)\right\|_\infty + \|\nabla_1\nabla_2(D_{1,2}k)\|_\infty\right)|\Phi_1-\Phi_2|^2.$$

With the above inequality, we obtain

$$\tag{4.45}\begin{aligned}&\int_{\mathbb{R}^d} \|\nabla_2 k(\cdot,\Phi(t,y_1,\rho_1)) - \nabla_2 k(\cdot,\Phi(t,y_2,\rho_2))\|_{\mathcal{H}_k^d}\, d\pi^0(dy_1,dy_2)\\ &\le \left(2\left\|\nabla^2(D_{1,2}k)\right\|_\infty^{\frac{1}{2}} + \|\nabla_1\nabla_2(D_{1,2}k)\|_\infty^{\frac{1}{2}}\right)\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi(t,y_1,\rho_1)-\Phi(t,y_2,\rho_2)|\pi^0(dy_1,dy_2).\end{aligned}$$

Based on (4.44),(4.45), we then get

$$\begin{aligned}&\left\|\iota^*_{k,\rho_t}\nabla\log\frac{\rho_t}{\pi} - \iota^*_{k,\mu_t}\nabla\log\frac{\mu_t}{\pi}\right\|_{\mathcal{H}_k^d}\\ &\le C_1(k,V)\left(\|\rho_1\|_{\mathcal{P}_V}+\|\rho_2\|_{\mathcal{P}_V}\right)^{\frac{1}{q}}\exp\left(D_1(k,\nu,q)\left(\mathrm{KL}(\rho_1|\pi)+\mathrm{KL}(\rho_2|\pi)\right)^{\frac{1}{2}}t^{\frac{1}{2}}\right)\\ &\qquad\left(\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi_1(t,y_1,\rho_1)-\Phi(t,y_2,\rho_2)|^p\pi^0(dy_1,dy_2)\right)^{1/p}\\ &\quad+\left(2\left\|\nabla^2(D_{1,2}k)\right\|_\infty^{\frac{1}{2}} + \|\nabla_1\nabla_2(D_{1,2}k)\|_\infty^{\frac{1}{2}}\right)\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi(t,y_1,\rho_1)-\Phi(t,y_2,\rho_2)|\pi^0(dy_1,dy_2)\\ &\le C(t,k,V,\nu,\rho_1,\rho_2,q)\left(\int_{\mathbb{R}^d\times\mathbb{R}^d}|\Phi_1(t,y_1,\rho_1)-\Phi(t,y_2,\rho_2)|^p\pi^0(dy_1,dy_2)\right)^{1/p}.\end{aligned}$$

where

$$\tag{4.46}\begin{aligned}&C(t,k,V,\nu,\rho_1,\rho_2,q)\\ &=C_1(k,V)\left(\|\rho_1\|_{\mathcal{P}_V}+\|\rho_2\|_{\mathcal{P}_V}\right)^{\frac{1}{q}}\exp\left(D_1(k,\nu,q)\left(\mathrm{KL}(\rho_1|\pi)+\mathrm{KL}(\rho_2|\pi)\right)^{\frac{1}{2}}t^{\frac{1}{2}}\right)+C_2(k)\end{aligned}$$

with

$$\begin{aligned}C_1(k,V)&=3C_V^{1/q}\left(\|k\|_\infty+\|\nabla k\|_\infty^{\frac{1}{2}}+\left\|\nabla^2 k\right\|_\infty^{\frac{1}{2}}+\|\nabla_1\nabla_2 k\|_\infty^{\frac{1}{2}}\right),\\ D_1(k,\nu,q)&=C_{1,0}\nu^{-1/2}q^{-1}\|k\|_\infty,\\ C_2(k)&=2\left\|\nabla^2(\nabla_1\cdot\nabla_2 k)\right\|_\infty^{\frac{1}{2}}+\|\nabla_1\nabla_2(\nabla_1\cdot\nabla_2 k)\|_\infty^{\frac{1}{2}}.\end{aligned}$$

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

$$\left| \Big((1-\nu)\iota_{k,\rho_{2,t}}^* \iota_{k,\rho_{2,t}} + \nu I \Big)^{-1} \Big(\iota_{k,\rho_{1,t}}^* \nabla \log \frac{\rho_{1,t}}{\pi}(x) - \iota_{k,\rho_{2,t}}^* \nabla \log \frac{\rho_{2,t}}{\pi}(x) \Big) \right|$$

$$\leq \nu^{-1} \left\| k \right\|_\infty C(t,k,V,\nu,\rho_1,\rho_2,q) \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |\Phi_1(t,y_1,\rho_1) - \Phi(t,y_2,\rho_2)|^p \pi^0(dy_1,dy_2) \right)^{1/p}.$$

Therefore, we obtain the desired result. ■

4.8. Space-time Discretization: A Practical Algorithm

In this section we introduce a practical space-time discretization to the R-SVGF described in (4.24). In the algorithm we let positive integers N and n to denote the number of particles and (discrete) iterations. We note by $(X_n^i)_{i=1}^N$ the position of the N particles at the n-th step. We let $\bar{X}_n := [X_n^1, \ldots, X_n^N]^T$. For all functions $f : \mathbb{R}^d \to \mathbb{R}^d$, we define the operator L_n as

$$L_n f := [f(X_n^1), \cdots, f(X_n^N)]^T.$$

The positions of the particles are then updated as

(4.47)

$$\bar{X}_{n+1} = \bar{X}_n - h_{n+1} \left(\frac{(1-\nu_{n+1})}{N} K_n + \nu_{n+1} I_N \right)^{-1} \left(\frac{1}{N} K_n (L_n \nabla V) - \frac{1}{N} \sum_{j=1}^N L_n \nabla k(X_n^j, \cdot) \right),$$

where $(h_n)_{n=1}^\infty$ is the sequence of step-sizes, $I_{N\times N}$ is the $N \times N$ identity matrix and $K_n \in \mathbb{R}^{N\times N}$ is the gram matrix defined as $(K_n)_{ij} = k(X_n^i, X_n^j)$ for all $i, j \in [N]$. We call the above algorithm as the Regularized SVGD algorithm. The iterates in (4.47) follow from Lemma 20 and the finite-sample representations for the operators $\iota_{k,\hat{\rho}^n} \iota_{k,\hat{\rho}^n}$ where $\hat{\rho}^n$ is the empirical measure of the particles at the n-th step, i.e., $\hat{\rho}^n = \sum_{i=1}^N \delta_{X_n^i}$.

While the convergence analysis of space-time discretization of the SVGF (i.e., the SVGD algorithm) and the R-SVGD (i.e., the regularized SVGD algorithm) is an interesting and challenging open question, in this section we demonstrate the improved performance of the regularized SVGD algorithm over the SVGD algorithm in some simulation examples. Specifically, we consider the simulation setup in [LW16]: We let the target $\pi := (1/3)\pi_1 + (2/3)\pi_2$, where $\pi_1 \equiv \text{Normal}(-2, 1)$

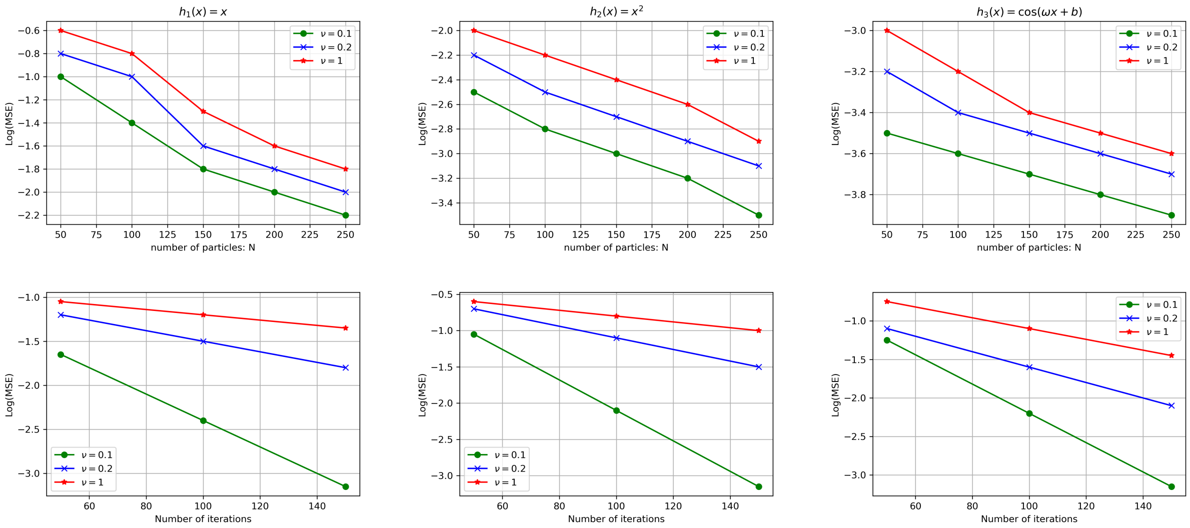

FIGURE 4.1. R-SVGD for various values of the regularization parameter ν. The case of $\nu = 1$ corresponds to SVGD. Left, Middle and Right columns correspond respectively to $h_1(x) := x$, $h_2(x) := x^2$ and $h_3(x) := \cos(\omega x + b)$. Top and bottom rows correspond respectively to log(MSE) versus number of particles and number of iterations.

and $\pi_2 \equiv \text{Normal}(+2, 1)$, and we let the initial distribution to be Normal$(-10, 1)$. We now focus on numerically computing the expectations of the form $\mathbb{E}_{x\sim\pi}[h_i(x)]$, for three cases, $h_1(x) := x$, $h_2(x) := x^2$ and $h_3(x) := \cos(\omega x + b)$, where $\omega \sim \text{Normal}(0, 1)$ and $b \sim \text{Uniform}([0, 2\pi])$.

In Figure 4.1, we plot the mean-squared error in estimating the above expectations with the regularized and unregularized SVGD algorithm. Here, the expectation is over the intialization (and over ω and b for h_3). In the top row, we report the logarithm of the mean-squared error versus the number of particles N for a fixed number of iterations (set to 100). In the bottom row, we report the logarithm of the mean-squared error versus the number of iterations for a fixed number of particles (set to 200). For both algorithms, we use the Gaussian kernel $k(u, v) = \exp\left(-\frac{1}{\gamma}\|u - v\|_2^2\right)$, where the bandwidth parameter γ is set using the median heuristic [LW16]. We use the Adagrad step-size choice for both cases, following [LW16]. For the choice of the regularization parameter, we report results for various choices of ν. The case of $\nu = 1$ corresponds exactly to the SVGD algorithm. We notice that for small values of ν the regularized SVGD algorithm performs better than the SVGD algorithm.

In terms of computational complexity, in comparison to the SVGD algorithm, each iteration of regularized SVGD algorithm requires inverting an $N \times N$ matrix. It is possible to speed-up the regularized SVGD algorithm by reinterpreting the iterations as solving a system of linear equations, and using fast implementations of linear systems solvers. Other techniques for speeding-up include using Random Fourier Features and Nyström method. We leave a detailed study of speeding up the regularized SVGD algorithm as future work.

CHAPTER 5

Online Least-squares Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is an algorithm that was initially developed by [RM51] for root-finding. Today, SGD and its variants are the most commonly used algorithms for training machine learning models ranging from large-scale linear models to deep neural networks. One of the main challenges in understanding SGD is comprehending its convergence properties. In the case of fixed-dimensional problems, the learning theory and optimization communities have focused on providing non-asymptotic bounds, either in expectation or with high-probability, over the past two decades. However, such bounds often tend to be overly conservative in predicting the actual behaviour of the SGD algorithm on large-scale statistical problems occurring in practice that are invariably based on specific data generating models. To address this, recent research has concentrated on characterizing the exact dynamics of SGD in large-scale high-dimensional problems. Specifically, the focus is on obtaining the precise asymptotic behavior of SGD and its fluctuations when the number of iterations or observations and the data dimension tend to infinity under appropriate scalings. The main idea behind this approach is to demonstrate that, under the considered scaling, the noise effects in SGD average out, so the exact asymptotic behavior and fluctuations are determined by a particular set of dynamical system equations.

Our goal in this chapter is to consider the SGD algorithm on a specific statistical problem, namely the linear regression problem, and provide a fine-grained analysis of its behaviour under high-dimensional scalings. Specifically, we consider the linear regression model, $Y = X^\top \theta^* + \mathcal{E}$, where $\theta^* \in \mathbb{R}^d$ is the true regression coefficient, $X \in \mathbb{R}^d$ is the zero-mean input random vector with covariance matrix $\mathbb{E}[XX^\top] = \Sigma_d \in \mathbb{R}^{d\times d}$, $Y \in \mathbb{R}$ is the output or response random variable, and $\mathcal{E} \in \mathbb{R}$ is a zero-mean noise with at least finite-variance. For this statistical model, we consider minimizing the following population least-squares stochastic optimization problem

$$\min_{\theta\in\mathbb{R}^d} \mathbb{E}[(Y - \langle X, \theta\rangle)^2],$$

using the online SGD with an initial guess $\theta^0 \in \mathbb{R}^d$, given by the following iterations

$$\theta^{t+1} = \theta^t + \eta \left(y^t - \langle x^t, \theta^t\rangle\right) x^t, \tag{5.1}$$

where $\theta^t \in \mathbb{R}^d$ is the output at time t and $\eta > 0$ is the step-size parameter and plays a crucial in obtaining our scaling limits and fluctuations. Specific choices for η will be detailed shortly in Section 5.3. The sequences $\{x^t\}_{t\geq 0}$ and $\{\varepsilon^t\}_{t\geq 0}$ are assumed to be independent and identical copies of the random vector X and noise $\mathcal{E}$ respectively. Note in particular that for the online SGD in (5.1), the number of iterations is equal to the number of observations used.

In our analysis, we view the least-squares online SGD in (5.1), as a *discrete* space-time interacting particle system, where the space-axis corresponds to the coordinates of the vector θ^t and the time-axis corresponds to the evolution of the algorithm. Specifically, note that the online SGD updates in (5.1) can be viewed in the following coordinate-wise form. For any $1 \leq i \leq d$,

$$\begin{aligned}
\theta_i^{t+1} &= \theta_i^t + \eta x_i^t y^t - \eta \sum_{j=1}^d x_i^t x_j^t \theta_j^t \\
&= \theta_i^t + \eta x_i^t \Big(\sum_{j=1}^d x_j^t \theta_j^* + \varepsilon^t - \sum_{j=1}^d x_j^t \theta_j^t \Big) \\
&= \theta_i^t - \eta \sum_{j=1}^d x_i^t x_j^t (\theta_j^t - \theta_j^*) + \eta x_i^t \varepsilon^t.
\end{aligned}$$

Now, defining the centralized iterates as $\Delta\theta^t \coloneqq \theta^t - \theta^*$ and letting $\Delta\theta_i^t$ denote its i-th coordinate for all $1 \leq i \leq d$ and $t \geq 0$, the least-squares online SGD can be then alternatively be represented as the following interacting particle system:

$$\Delta\theta_i^{t+1} = \Delta\theta_i^t - \underbrace{\eta \sum_{j=1}^d x_i^t x_j^t \Delta\theta_j^t}_{\textit{random}\text{ interaction}} + \eta x_i^t \varepsilon^t, \qquad 1 \leq i \leq d, \tag{5.2}$$

where the particles $\{\Delta\theta_i^t\}_{1\leq i\leq d}$ are interacting and evolve over a discrete-time scale. In particular, the interaction among the particles $\{\Delta\theta_i^t\}_{1\leq i\leq d}$ is random for any t, and the *expected interaction* is captured by the covariance matrix Σ_d of the input vector X. Therefore, to analyze the high-dimensional asymptotic properties of the least-squares online SGD, we analyze the scaling limit and fluctuations of the interacting particle system given by (5.2). In particular, our limits are derived in the form of infinite-dimensional Ordinary Differential Equations (ODEs) and Stochastic Differential Equations (SDEs); for some background we refer to [Arn92, KX95, DPZ14].

Our approach is also motivated by the larger literature available on analyzing interacting particle systems. See, for example, [KL98, Lig99, Zei04, Szn04, DN08, Spo12]. The interacting particle system in (5.2) can either exhibit long-range or short-range interactions depending on the structure of the covariance matrix Σ_d. The case when the covariance matrix Σ_d is "smooth" in an appropriate sense, thereby prohibiting abrupt changes in the entries of the covariance matrix, corresponds to the regime of long-range interactions. Examples of such covariance matrices include bandable and circulant covariance matrices. In this work, we work in the long-range interaction regime. Alternatively, the case when covariance matrix is "rough", which allows for certain degree of abrupt changes between the entries (allowing, for example, various patterns of structured sparsity in Σ_d) corresponds to the regime of short-range interactions. For last few decades, a large number of works on interacting particle system focused on the connection between the scaling limit of the fluctuation in interacting particle systems and the Kardar-Parisi-Zhang (KPZ) equation. See, for example, [BG97, BS10, Qua11, Cor12, CST18, Gho17, CGST20]. Many of those works demonstrated that the fluctuation of the height function of the associated particle system converges to the KPZ equation under weak noise scaling. A handful set of those works including some recent breakthroughs [MQR21, DOV22, QS23] showed convergence towards the KPZ fixed point under the so called KPZ scaling of space, time and fluctuation. Our fluctuation results in the current work for the long-range interaction case does not reproduce the KPZ equation or the KPZ fixed point in the limit. However, we do believe that in the short-range regime, the scaling limits would lead to those unprecedented limiting behavior even for the seemingly simple problem of solving least-squares with online SGD. In a forthcoming work, we investigate the case of short-range interactions in detail.

5.1. Preliminaries

Before we proceed, we list the notations we make in this work. For a positive integer a, we let $[a] \coloneqq \{1, \ldots, a\}$. For vectors, superscripts denote time-index and subscripts denote coordinates. The space of square-integrable functions on S, a subset of Euclidean space, is denoted as $L^2(S)$ with the squared-norm $\|g\|^2_{L^2(S)} \coloneqq \int_S g(x)^2 dx$ for any $g \in L^2(S)$. The space of continuous functions in $[0,1]$, is denoted by $C([0,1])$, and is equipped with the topology of uniform convergence over $[0,1]$. The space of continuous functions with continuous derivatives up to k^{th}-order in $[0,1]$, is denoted by $C^k([0,1])$, and is equipped with the topology of uniform convergence in function value and derivatives up to k^{th}-order over $[0,1]$ (we mainly use $k = 1, 2$). For any topological space $\mathcal{H}$ and $\tau > 0$, $C([0,\tau]; \mathcal{H})$ represents the space of $\mathcal{H}$-valued functions with continuous trajectories.

We also require the definition of a Gaussian random field or Gaussian random field process, that arise in characterizing the limiting behavior of the SGD iterates. We refer to [AT07] for additional background.

Definition 5.1.1. A *Gaussian random field* is defined as a random field g on a parameter set $[0,1]$ for which the finite-dimensional distributions of $(g(x_1), \cdots, g(x_K))$ are multivariate Gaussian for each $1 \le K < \infty$ and each $(x_1, \cdots, x_K) \in [0,1]^K$. A *Gaussian random field process* is defined as a time-indexed random field g on a parameter set $[0,\infty) \times [0,1]$ for which $\{(g(t,x_1), \cdots, g(t,x_K))\}_{t \ge 0}$ are multivariate Gaussian processes for each $1 \le K < \infty$ and each $(x_1, \cdots, x_K) \in [0,1]^K$.

Space-time Interpolation. Our approach starts by constructing a space-time stochastic process by performing a piecewise linear interpolation of the discrete particles in (5.2). This process, denoted by $\{\overline{\Theta}^{d,T}(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ is continuous both in time and in space. The spatial coordinate at the macroscopic scale is indexed by the set $[0,1]$ and is denoted by the spatial variable x. The spatial resolution at the microscopic scale is of order $1/d$. Let $T \in \mathbb{N}_+$ be a positive integer parameter. The parameters $\tau \in (0,\infty)$ and $1/T$ corresponds to the time-scale till which we would like to observe the trajectory and the resolution of the time-axis respectively. For any τ, we consider the first $\lfloor \tau T \rfloor$ least-squares online SGD iterates. Specifically, $\lfloor \tau T \rfloor$ corresponds to the overall number of iterations, which also corresponds to the number of observations used.

We now describe how to construct the interpolation of the discrete particles in (5.2). Consider the function $\Theta^{d,T}(\cdot,\cdot) : [0,\tau] \times [0,1] \to \mathbb{R}$ that satisfies the following conditions:

(a) it is evaluated as the positions of SGD particles on the grid points with grid width T^{-1} in the time variable and d^{-1} in the space variable, i.e. for any $i \in [d]$ and any $0 \le t \le \lfloor \tau T \rfloor$,

$$\Theta^{d,T}\Big(\frac{t}{T}, \frac{i}{d}\Big) = \Delta\theta_i^t.$$

We can artificially define $\Delta\theta_0^t = \Delta\theta_1^t$.

(b) it is piecewise-constant in the time variable, i.e. for any $s \in [0,\tau]$,

$$\Theta^{d,T}(s,\cdot) = \Theta^{d,T}\Big(\frac{\lfloor sT \rfloor}{T}, \cdot\Big).$$

(c) it is piecewise linear in the space variable, i.e. for any $x \in [0,1]$,

$$\Theta^{d,T}(\cdot,x) = (\lfloor dx \rfloor + 1 - dx)\Theta^{d,T}\Big(\cdot, \frac{\lfloor dx \rfloor}{d}\Big) + (dx - \lfloor dx \rfloor)\Theta\Big(\cdot, \frac{\lfloor dx \rfloor + 1}{d}\Big).$$

From the above construction, we can give an explicit form of $\Theta^{d,T}$ based on $\{\Delta\theta_i^t\}_{i\in[d],0\le t\le\lfloor\tau T\rfloor}$: for any $s \in [0,\tau]$ and any $x \in [0,1]$,

$$\begin{aligned}\Theta^{d,T}(s,x) &= (\lfloor dx \rfloor + 1 - dx)\Theta^{d,T}\Big(\frac{\lfloor sT \rfloor}{T}, \frac{\lfloor dx \rfloor}{d}\Big) + (dx - \lfloor dx \rfloor)\Theta\Big(\frac{\lfloor sT \rfloor}{T}, \frac{\lfloor dx \rfloor + 1}{d}\Big)\\ &= (\lfloor dx \rfloor + 1 - dx)\Delta\theta_{\lfloor dx \rfloor}^{\lfloor sT \rfloor} + (dx - \lfloor dx \rfloor)\Delta\theta_{\lfloor dx \rfloor + 1}^{\lfloor sT \rfloor}.\end{aligned}$$

Condition (b) in particular implies that $\Theta^{d,T}$ is càdlaġ in the time variable. From condition (c), it is easy to observe that $\Theta^{d,T}$ is continuous with well-defined weak derivative in space. Actually the first order space derivative to $\Theta^{d,T}$ is well-defined for all x except the grid points $\{0, \frac{1}{d}, \cdots, \frac{d-1}{d}, 1\}$. Next we construct $\overline{\Theta}^{d,T}$ which is piecewise linear both in time and in space. For any $s \in [0,\tau]$ and $x \in [0,1]$,

$$\overline{\Theta}^{d,T}(s,x) := (\lfloor sT \rfloor + 1 - sT)\,\Theta^{d,T}\Big(\frac{\lfloor sT \rfloor}{T}, x\Big) + (sT - \lfloor sT \rfloor)\,\Theta^{d,T}\Big(\frac{\lfloor sT \rfloor + 1}{T}, x\Big). \tag{5.3}$$

From this construction we can immediately see that $\overline{\Theta}^{d,T} \in C([0,\tau]; C[0,1])$. Our main objective in this work, is to characterize the limiting behavior of the space-time stochastic process

$\{\overline{\Theta}^{d,T}(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ when both T and d go to infinity, under appropriate scalings for appropriate choices of the step-size parameter η, and under various assumptions on the noise $\mathcal{E}$.

5.2. Assumptions

We start with the following definition of a space-time stochastic process, which characterizes the data generating process for X (or alternatively the sequence $\{x^t\}_{t\geq 0}$).

Definition 5.2.1 (Data Generating Process). Let A, B, and E be real-valued symmetric functions defined on $[0,1]^2$, $[0,1]^4$ and $[0,1]^8$ respectively. $\{W(s,x)\}_{s\geq 0, x\in[0,1]}$ is a stochastic process which is white in the time variable, such that for all $s\geq 0$ and $x, x_1, \cdots, x_8 \in [0,1]$:

$$\mathbb{E}\left[W(s,x)\right]=0, \qquad \mathbb{E}\left[W(s,x_1)W(s,x_2)\right]=A(x_1,x_2),$$

$$\mathbb{E}\big[\prod_{i=1}^{4}W(s,x_i)\big]=B(x_1,x_2,x_3,x_4), \qquad \mathbb{E}\big[\prod_{i=1}^{8}W(s,x_i)\big]=E(x_1,\cdots,x_8).$$

We now introduce the assumptions on the samples $\{x_i^t\}_{t\geq 0, i\in[d]}$ and random noises $\{\varepsilon^t\}_{t\geq 0}$.

Assumption 5.2.1. *The samples $\{x_i^t\}_{1\leq i\leq d, t\geq 0}$ and the noises $\{\varepsilon^t\}_{t\geq 0}$ are independent and satisfy*

(a) *There exists a stochastic process $\{W(s,x)\}_{s\geq 0, x\in[0,1]}$, defined in Definition 5.2.1 such that*

$$x_i^t = W\Big(\frac{t}{T},\frac{i}{d}\Big).$$

(b) *There exists a universal constant $C_1>0$ such that for any $t\geq 0$,*

$$\mathbb{E}\big[\varepsilon^t\big]=0, \quad \mathbb{E}\big[\left|\varepsilon^t\right|^2\big]\leq \sigma_d^2 \quad \textit{and} \quad \mathbb{E}\big[\left|\varepsilon^t\right|^4\big]\leq C_1\sigma_d^4.$$

The noise-variance σ_d^2 plays a crucial role in our scaling limits. In particular, it is allowed to grow with d, with the growth rate determining the precise scaling limit of the SGD iterates. In the following, we drop the subscript d for convenience. We now introduce the main assumption we make regarding smoothness of the covariance and higher-moments of the process W.

Assumption 5.2.2 (Smoothness Conditions). *The stochastic process* $\{W(s,x)\}_{s\geq 0, x\in[0,1]}$ *satisfies the following smoothness conditions.*

1. *The function* $A : [0,1]^2 \to \mathbb{R}$ *is such that* $A(x,\cdot) \in C^1([0,1])$ *for any* $x \in [0,1]$*. As a result, there exist constants* C_2, C_3 *such that for any* $x, y, z \in [0,1]$*, we have*

$$|A(x,y)| \leq C_2 \qquad \textit{and} \qquad |A(x,z) - A(y,z)| \leq C_3|x-y|.$$

We further assume that there exists a constant C_4 *such that*

$$|A(x,x) + A(y,y) - 2A(x,y)| \leq C_4^2|x-y|^2.$$

2. *For the function* $B : [0,1]^4 \to \mathbb{R}$*, there exist constants* $C_5, C_6, C_7 > 0$ *such that for any* $x_1, x_2, x_3, x_4 \in [0,1]$*,* $|B(x_1,x_2,x_3,x_4)| \leq C_5$ *and*

$$\begin{aligned}|B(x_1,x_3,x_1,x_3) + B(x_2,x_3,x_2,x_3) - 2B(x_1,x_3,x_2,x_3)| &\leq C_6^2|x_1-x_2|^2,\\ \big|B(x_1,x_1,x_1,x_1) + B(x_2,x_2,x_2,x_2) + 6B(x_1,x_1,x_2,x_2) &- 4B(x_1,x_1,x_1,x_2)\\ - 4B(x_1,x_2,x_2,x_2)\big| &\leq C_7^4\big|x_1-x_2\big|^4.\end{aligned}$$

3. *For the function* $E : [0,1]^8 \to \mathbb{R}$*, there exist constants* C_8, C_9 *such that for all* $x_1, \cdots, x_8 \in [0,1]$*,* $E(x_1,\cdots,x_8) \leq C_8$ *and*

$$\begin{aligned}&\big|E(x_1,x_1,x_1,x_1,x_3,\cdots,x_6) + E(x_2,x_2,x_2,x_2,x_3,\cdots,x_6) + 6E(x_1,x_1,x_2,x_2,x_3,\cdots,x_6)\\ &- 4E(x_1,x_1,x_1,x_2,x_3,\cdots,x_6) - 4E(x_1,x_2,x_2,x_2,x_3,\cdots,x_6)\big| \leq C_9^4\big|x_1-x_2\big|^4.\end{aligned}$$

Assumptions 5.2.1 and 5.2.2 are made to ensure tightness of the interpolated process in (5.3). In particular, the assumptions on the second and fourth-order moments are required to derive the tightness conditions needed to establish the scaling limits, with the second-order moment information actually showing up in the limit. Additionally, assumptions on the eighth-order moments are required to derive the tightness conditions needed to establish the fluctuations. In this case, the second-order moments information appears in the drift terms, and both the second and fourth-order moment information show up in the diffusion term.

Our assumptions above allow for a relative general class of distributions for the data and the noise sequence; see Section 5.5.2 for additional insights and examples. Importantly, we emphasize here that we do not make any isotropy or Gaussianity (or sub-Gaussianity) type assumptions made in several recent works (see Section 5.4). Instead, without assuming the independence of different coordinates in each sample, we take into account the structure of the k^{th}-order moments of the sample distribution for $k = 2, 4, 8$.

Finally, we remark that our analysis is general enough to allow for certain degree of dependence within the data sequence $\{x^t\}_{t\geq 0}$ and within the error sequence $\{\varepsilon^t\}_{t\geq 0}$. Such assumptions are typically made in several applications including reinforcement learning [Mey22] and sequential or online decision making [CLS21, KDL$^+$21]. Furthermore, we could also allow for some level of dependency between the data and error sequences. We do not discuss these extensions in detail to keep our exposition simpler.

5.3. Main Results

5.3.1. Scaling Limits. We now state our results on the scaling limits of the least-squares online SGD under different orders of noise variances.

THEOREM 5.3.1 (Scaling Limits). *Let the initial conditions satisfy: For any $0 \leq i \leq d$, there exist constants R, L such that the initial condition satisfies*

$$\mathbb{E}\left[|\Delta\theta_i^0|^4\right] \leq R^4 \qquad \text{and} \qquad \mathbb{E}\left[|\Delta\theta_i^0 - \Delta\theta_{i-1}^0|^4\right] \leq L^4 d^{-4}.$$

Also, let Assumption 5.2.1 and Assumption 5.2.2 hold and let $\{\xi_1(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ denote a Gaussian random field process with covariance given by (5.29). Further, let there exist a uniform constant $C_{s,1}$ such that $\max(\eta dT, \sigma\eta T^{\frac{1}{2}}) \leq C_{s,1}$. Then, we have the following scaling limits.

(1) ***Low-noise, i.e., $\sigma^2/d^2T \to 0$:*** *Assume further that there exists a uniform positive constant α such that $\lim_{d,T\to\infty} \sigma^2/(d^2T) = 0$ and $\lim_{d,T\to\infty} \eta dT = \alpha$. Then for any $\tau \in (0,\infty)$, $\{\overline{\Theta}^{d,T}\}_{d\geq 1,T>0}$ converges weakly to a function Θ as $d, T \to \infty$. Furthermore,*

the limit $\Theta \in C^1([0,\tau]; C^1([0,1]))$ *is the unique continuous solution to the following ODE:*

$$\partial_s \Theta(s,x) = -\alpha \int_0^1 A(x,y)\Theta(s,y)\mathrm{d}y. \tag{5.4}$$

(2) ***Moderate-noise, i.e.,*** $\sigma^2/d^2T \to (0,\infty)$***:*** *Assume further that there exist uniform positive constants* α, β *such that* $\lim_{d,T\to\infty} \sigma^2/(d^2T) = \beta^2$ *and* $\lim_{d,T\to\infty} \eta dT = \alpha$*. Then for any* $\tau \in (0,\infty)$*,* $\{\overline{\Theta}^{d,T}\}_{d\geq 1, T>0}$ *converges weakly to a process* $\{\Theta(s,\cdot)\}_{s\in[0,\tau]}$ *as* $d,T\to\infty$*. Furthermore, the limit* Θ *is the unique solution in* $C([0,\tau]; C([0,1]))$ *to the following SDE:*

$$\mathrm{d}\Theta(s,x) = -\alpha \int_0^1 A(x,y)\Theta(s,y)\mathrm{d}y\mathrm{d}s + \alpha\beta\mathrm{d}\xi_1(s,x). \tag{5.5}$$

(3) ***High-noise, i.e.,*** $\sigma^2/d^2T \to \infty$***:*** *Assume further that there exists a uniform positive constant* α *such that* $\lim_{d,T\to\infty} \eta\sigma T^{\frac{1}{2}} = \alpha$*. Then for any* $\tau \in (0,\infty)$*,* $\{\overline{\Theta}^{d,T}\}_{d\geq 1, T>0}$ *converges weakly to a process* $\{\Theta(s,\cdot)\}_{s\in[0,\tau]}$ *as* $d,T\to\infty$*. Furthermore, the limit* Θ *is the unique solution in* $C([0,\tau]; C([0,1]))$ *to the following SDE:*

$$\mathrm{d}\Theta(s,x) = \alpha\mathrm{d}\xi_1(s,x). \tag{5.6}$$

Remark 5.3.2. Theorem 5.3.1 shows that under high-dimensional scalings, the limiting behavior of the SGD trajectory exhibits a three-step phase transition: It goes from being ballistic (i.e., characterized by an infinite-dimensional ODE) in the low-noise setting, to diffusive (i.e., characterized by an infinite-dimensional SDE) in the moderate-noise setting, to purely random in the high-noise setting. The boundaries of this three-step phase transition are precisely characterized, and explicit dependencies on the order of dimension, iterations and step-size choices are identified. In the moderate and high-noise setting, the covariance of the diffusion term ξ_1 is determined by the second moment function A from Assumption 5.2.2. We also remark that our initial conditions are made coordinate-wise and are rather mild.

Remark 5.3.3. In the low-noise setting, according to (5.4), we see that $\Theta(s,\cdot)$ has the same order of smoothness as $A(\cdot,y)$ for any $s \in [0,\tau]$ and $y \in [0,1]$. This phenomenon is more general, i.e., if we further assume that $A(\cdot, y) \in C^k([0,1])$, for some finite positive integer k and for any $y \in [0,1]$, then it can be shown that $\Theta \in C^1([0,\tau]; C^k([0,1]))$.

Remark 5.3.4. Assumption 5.2.2 is required to show tightness results in Proposition 5.7.1 and Proposition 5.7.2, and hence to prove existence of the weak limit. It is worth mentioning that in the low-noise setup when the noise variance satisfies $\sigma^2 = O(d^2)$, Assumption 5.2.2 can be relaxed by dropping the last condition on B and the condition on E. The tightness results under the relaxed assumptions can be proved in the same way by considering second moments bounds rather than the fourth moments bounds in Proposition 5.7.1 and Proposition 5.7.2. For the sake of simplicity and consistency of our analysis, the proof of Theorem 5.3.1 is based on Assumption 5.2.2.

5.3.2. Fluctuations. We now study the fluctuation of $\{\Delta\theta_i^t\}_{i\in[d],0\le t\le N}$ in the low-noise setting. In order to do so, we look at a re-scaled difference between $\{\Delta\theta_i^t\}_{i\in[d],0\le t\le N}$ and its scaling limit. Specifically, for any $s \in [0,\tau]$, $x \in [0,1]$, define

$$U^{d,T}(s,x) := \gamma\big(\overline{\Theta}^{d,T}(s,x) - \Theta(s,x)\big) \tag{5.7}$$

where γ is the scaling parameter and we expect $\gamma \to \infty$ as $d, T \to \infty$. We now state our fluctuation results.

Theorem 5.3.5 (Fluctuations). *Let the initial conditions follow: For any $0 \le i \le d$, there exist constants D, M such that the initial condition satisfies*

$$\mathbb{E}\left[|U^{d,T}(0,\frac{i}{d})|^4\right] \le D^4 \qquad \textit{and} \qquad \mathbb{E}\left[|U^{d,T}(0,\frac{i}{d}) - U^{d,T}(0,\frac{i-1}{d})|^4\right] \le M^4 d^{-4}.$$

Furthermore, let the assumptions made in Theorem 5.3.1 in the low-noise setup, i.e. $\sigma^2/dT^2 \to 0$, hold. Let $\eta = \frac{\alpha}{dT}$ and assume that there exists a uniform constant $C_{s,2}$ such that for all $d \ge 1, T > 0$, we have $\max(\gamma T^{-\frac{1}{2}}, \gamma d^{-1}, \gamma\sigma d^{-1} T^{-\frac{1}{2}}) \le C_{s,2}$. Then we have the following fluctuation results.

(1) ***Particle interaction dominates:*** *Assume further that, as $d, T \to \infty$, we have*

$$T = o(d^2), \quad \sigma = O(d), \quad \gamma T^{-\frac{1}{2}} \to \zeta \in (0,\infty), \quad \textit{and} \quad \gamma\sigma d^{-1} T^{-\frac{1}{2}} \to \beta \in [0,\infty),$$

for some uniform constants ζ, β. Then for any $\tau \in (0, \infty)$, the fluctuation of SGD particles, $\{U^{d,T}\}_{d\geq 1, T>0}$ converges in distribution to a function $U \in C([0,\tau]; C([0,1]))$ as $d, T \to \infty$. Furthermore, for any $s \in [0,\tau]$, $x \in [0,1]$, the limit U is the unique solution in $C([0,\tau]; C([0,1]))$ to the SDE

$$\mathrm{d}U(s,x) = -\alpha \int_0^1 A(x,y)U(s,y)\mathrm{d}y\mathrm{d}s + \alpha\beta\mathrm{d}\xi_2(s,x) + \alpha\zeta\mathrm{d}\xi_3(s,x), \tag{5.8}$$

where $\{\xi_2(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ and $\{\xi_3(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ are two independent Gaussian random field processes with covariances given by (5.39) *and* (5.41) *respectively.*

(2) ***Noise dominates:*** *Assume further that, as $d, T \to \infty$, we have*

$$\max(d, T^{\frac{1}{2}}) \ll \sigma \ll dT^{\frac{1}{2}}, \qquad \textit{and} \qquad \gamma\sigma d^{-1}T^{-\frac{1}{2}} \to \beta \in (0,\infty),$$

for a uniform constant β. Then for any $\tau \in (0, \infty)$, the fluctuation of SGD particles, $\{U^{d,T}\}_{d\geq 1, T>0}$ converges in distribution to a function $U \in C([0,\tau]; C([0,1]))$ as $d, T \to \infty$. Furthermore, for any $s \in [0,\tau]$, $x \in [0,1]$, the limit U is the unique solution in $C([0,\tau]; C([0,1]))$ to the SDE

$$\mathrm{d}U(s,x) = -\alpha \int_0^1 A(x,y)U(s,y)\mathrm{d}y\mathrm{d}s + \alpha\beta\mathrm{d}\xi_2(s,x), \tag{5.9}$$

where $\{\xi_2(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ is a Gaussian random field process with covariance given by (5.39)*.*

(3) ***Interpolation error dominates:*** *Assume further that, as $d, T \to \infty$, we have*

$$d = O(T^{\frac{1}{2}}), \qquad \sigma = O(T^{\frac{1}{2}}) \qquad \textit{and} \qquad \gamma d^{-1} \to 0.$$

Then for any $\tau \in (0, \infty)$, the fluctuation of SGD particles, $\{U^{d,T}\}_{d\geq 1, T>0}$ converges in distribution to a function $U \in C([0,\tau]; C([0,1]))$ as $d, T \to \infty$. Furthermore, for any $s \in [0,\tau]$, $x \in [0,1]$, the limit U is the unique continuous solution to the following ODE (with random initial conditions)

$$\mathrm{d}U(s,x) = -\alpha \int_0^1 A(x,y)U(s,y)\mathrm{d}y\mathrm{d}s. \tag{5.10}$$

Remark 5.3.6. The setting in Theorem 5.3.5 further splits the low-noise regime of Theorem 5.3.5 into three sub-regimes. The first two regimes correspond to the case when the dominating terms leading to the fluctuations are due to the particle interaction (whose expectation is characterized by the covariance function A), and the noise variance respectively. In particular, when the particle interaction dominates, we have two independent diffusion terms; while the covariance of the process ξ_2 is determined only by the second-moment function A in Assumption 5.2.2, the covariance of the process ξ_3 appearing in the second diffusion term is determined by both A and the fourth-moments function B. The third sub-regime comes from having to deal with the approximation of the integral on the right hand side of (5.4) with its Riemann sum, and we refer to this regime as the interpolation error regime. In this regime, we choose a small order of γ to ensure the interpolation error vanishes. Doing so, fluctuations from the particle interaction and noises are suppressed in the limit. Therefore we obtain the degenerate convergence to the ODE as in (5.10) for the fluctuations. The entire limit identification result is provided in Theorem 5.6.4.

Remark 5.3.7. Scaling limits and fluctuations developed in Theorems 5.3.1 and 5.3.5 respectively, also hold (with slight modifications) for the online multiplier bootstrap version of SGD developed in [FXY18, Equations (7) and (8)]; such results may be leveraged for practical high-dimensional statistical inference.

5.4. Related Work

High-dimensional scaling limits of SGD. [BAGJ22] studied the scaling limits of online SGD in the high-dimensional setting for a class of non-convex problems. Their scaling limits are derived for finite-dimensional summary statistics of the online SGD, and no fluctuation results are provided. Furthermore, the precise scaling relationship between the dimension and the number of iterations, and the impact of the data generating process, in particular the covariance structure, is left unexplored. [WML17] analyzed the online SGD algorithms for least-squares regression and principal component analysis, and derived the scaling limit of the empirical densities as solutions to PDEs. However, their analysis was restricted to the special case of isotropic covariance matrices and no fluctuation results are provided. See also [WL19, VSL$^+$22, PP21, PPAP22a, PPAP22b] for related works on specific models for online and mini-batch SGD. However, such works do not identify any phase transition phenomenon (from ballistic to diffusive behaviour), are applicable only for specific statistics of the SGD iterates, and do not characterize the fluctuations.

The works by [CCM21] and [GTM$^+$22] also characterized the asymptotic behaviors of variants of SGD for a class of optimization problems in the high-dimensional setting. Their approach was based on the so-called *dynamical mean-field theory* from statistical physics. Different from our work, which considers online SGD on expected objective functions, both [CCM21] and [GTM$^+$22] considered mini-batch SGD on finite-sum objective functions (or empirical risk minimization). [CCM21] require isotropic sub-Gaussian inputs for their analysis. While [GTM$^+$22] allows for non-isotrpic covariance, they required Gaussianity assumptions on the inputs. Finally, they only track a real-valued functional of the trajectory as the dimension grows and no fluctuation results are provided.

To our knowledge, our work provides the first result on characterizing the entire infinite-dimensional trajectorial limit and related fluctuations of the online SGD, for the specific problem of least-squares regression with smooth covariance structures.

Other high-dimensional analysis of SGD. Random matrix theory is also used to analyze full and mini-batch gradient-based iterative algorithms for specific high-dimensional models; see, for example, [DT19, PLPP21, BES$^+$22, DT22, PvMPP22].

[CLP22, CPT23] studied mini-batch SGD for certain high-dimensional non-convex problems using Gaussian process techniques. Their work relies heavily on the isotropy and Gaussianity assumptions. State-space approaches for high-dimensional analysis of online SGD was carried out in [TV23] under isotropic Gaussianity assumptions. The work of [BAGJ21] also used a similar approach to establish high-probability bounds in a signal-recovery setup. Recently, high-dimensional normal approximation results and tail bounds are also established in [ABG23] and [DMN$^+$21, DMNS22] respectively for online least-squares SGD.

Mean-field analysis of SGD for overparametrized neural networks is also explored intensely in the recent past. While assuming growing parameter dimension, such works, however, assume the data-dimension is fixed. Due to the flurry of recent works in this direction, it is impossible to list them all here. We refer to [CB18b, MMN18, SS20b, SS20a, PN21, RVE22, DCLW22, SS22, AAM22, GGK22] for a sampling of such works.

Fixed-dimensional analyses of SGD. The study of diffusion limits of SGD in the fixed-dimensional setting is a classical topic. We refer to [KGY03, KC12, BMP12, LPW12] for a textbook treatment of this topic. The main idea behind such works is to show that appropriately time-interpolated SGD iterates converge to a multi-dimensional Ornstein-Uhlenbeck process, under specific scalings. Recently, [LWME19] developed a framework to approximate the dynamics of a relative general class of stochastic gradient algorithms by stochastic differential equations. See also, [KB17, LWLZ18, GCPT20, FDBD21] for a partial list of other recent related works.

Almost sure convergence and central limit theorems for SGD in the fixed-dimensional setting is also well-studied. See [MHKC20, SGD21, LY22] and references therein for almost-sure convergence results. With regards to CLT, we refer to [PJ92, Rup88, DR20, TA17, AD19, YBVE21, DDB20, BBHS21, DDJ23] and references therein for a partial list of related works. We also highlight the works of [ABE19] and [SZ22], where non-asymptotic normal approximation for SGD is established. For a survey of expectation and high-probability bounds for SGD and its variants, see [BCN18], and [Lan20].

Scaling limits of MCMC algorithms. Finally, we remark that high-dimensional scaling limits of iterative sampling algorithms like the Random-Walk Metropolis (RWM) algorithm, Unadjusted Langevin Algorithm (ULA), and Metropolis Adjusted Langevin Algorithm (MALA) is well-studied. For example, [PST12] and [MPS12] characterize the scaling limits in the form of infinite-dimensional SDE (or equivalently as stochastic PDEs); see also the references therein for other related works in this direction. While being morally related to our approach, the sampling algorithms studied in those works correspond to a different setup than us, as the interactions are not characterized by any data generating process.

5.5. Applications and Examples

As an application of our main results in Theorem 5.3.1 and 5.3.5, in this section, we show how they can be leveraged to study certain specific properties of the least-squares online SGD, like the Mean Square Error (MSE) and Predictive Error (PE).

Mean Squared Error (MSE) and Predictive Error (PE): For a given d and T, the time-interpolated mean square error and prediction error of the least-squares online SGD estimator at time $s \in [0, \tau]$ are defined respectively as

$$\mathsf{MSE}^{d,T}(s) := \frac{1}{d}\sum_{i=1}^{d} \overline{\Theta}^{d,T}(\frac{\lfloor sT \rfloor}{T}, \frac{i}{d})^2,$$

$$\mathsf{PE}^{d,T}(s) := \frac{1}{d^2}\sum_{i,j=1}^{d} A(\frac{i}{d}, \frac{j}{d})\overline{\Theta}^{d,T}(\frac{\lfloor sT \rfloor}{T}, \frac{i}{d})\overline{\Theta}^{d,T}(\frac{\lfloor sT \rfloor}{T}, \frac{j}{d}).$$

Specifically, we first calculate the high-dimensional scaling limits of MSE and PE in terms of the solutions to (5.4),(5.5) and (5.6), and show the decay properties of the limiting MSE and limiting PE in the low-noise setup. We next calculate the fluctuations of the MSE and PE in terms of the solutions to (5.4), and (5.8),(5.9) and (5.10).

5.5.1. Scaling Limits and Fluctuations of MSE and PE.

Leveraging Theorem 5.3.1, we have the following result on the scaling limits of MSE and PE of the least-squares online SGD. In particular, it exhibits the three-step phase-transition depending on the noise level. To proceed, we defined the limiting MSE and PE as follows:

$$\mathsf{MSE}(s) := \int_0^1 \Theta(s,x)^2 dx, \qquad \mathsf{PE}(s) := \int_0^1\int_0^1 \Theta(s,x)A(x,y)\Theta(s,y)dxdy. \tag{5.11}$$

Proposition 5.5.1. *Assume that Assumption 5.2.1, Assumption 5.2.2 and the initial conditions in Theorem 5.3.1 hold. Then, we have the following convergences, (i)* $\mathsf{MSE}^{d,T}(s) \to \mathsf{MSE}(s)$ *and (ii)* $\mathsf{PE}^{d,T}(s) \to \mathsf{PE}(s)$ *(in probability in the low-noise setting and in distribution in the other two settings), provided one of the following scaling condition is satisfied:*

(1) ***Low-noise:*** $\sigma d^{-1}T^{-\frac{1}{2}} \to 0$, $\eta dT \to \alpha \in (0,\infty)$ *as* $d, T \to \infty$ *and* Θ *is the solution to* (5.4).

(2) ***Moderate-noise:*** $\sigma d^{-1}T^{-\frac{1}{2}} \to \beta \in (0,\infty)$, $\eta dT \to \alpha \in (0,\infty)$ *as* $d, T \to \infty$ *and* Θ *is the solution to* (5.5).

(3) ***High-noise:*** $\sigma d^{-1}T^{-\frac{1}{2}} \to \infty$, $\eta\sigma T^{\frac{1}{2}} \to \alpha \in (0,\infty)$ *as* $d, T \to \infty$ *and* Θ *is the solution to* (5.6).

We next characterize the fluctuations of $\mathsf{MSE}^{d,T}(s)$ and $\mathsf{PE}^{d,T}(s)$, i.e $\gamma\left(\mathsf{MSE}^{d,T}(s) - \mathsf{MSE}(s)\right)$ and $\gamma\left(\mathsf{PE}^{d,T}(s) - \mathsf{PE}(s)\right)$, respectively.

Proposition 5.5.2. *Assume that Assumption 5.2.1, Assumption 5.2.2 and the initial conditions in Theorem 5.3.1 and Theorem 5.3.5 hold. If* $\eta = \frac{\alpha}{dT}$ *then*

(i) $\gamma\left(\mathsf{MSE}^{d,T}(s) - \mathsf{MSE}(s)\right) \to 2\int_0^1 \Theta(s,x)U(s,x)\mathrm{d}x$ *in distribution,*

(ii) $\gamma\left(\mathsf{PE}^{d,T}(s) - \mathsf{PE}(s)\right) \to 2\int_0^1\int_0^1 \Theta(s,x)A(x,y)U(s,y)\mathrm{d}x\mathrm{d}y$ *in distribution,*

provided one of the following scaling is satisfied:

(1) $\gamma = \zeta T^{\frac{1}{2}}$ *for some* $\zeta \in (0,\infty)$. $T = o(d^2), \sigma = O(d)$ *as* $d, T \to \infty$ *and* Θ, U *are the solutions to* (5.4) *and* (5.8) *respectively.*

(2) $\gamma = \beta\sigma^{-1}dT^{\frac{1}{2}}$ *for some* $\beta \in (0,\infty)$. $\max(d, T^{\frac{1}{2}}) \ll \sigma \ll dT^{\frac{1}{2}}$ *as* $d, T \to \infty$ *and* Θ, U *are the solutions to* (5.4) *and* (5.9) *respectively.*

(3) $1 \ll \gamma \ll d$, $d = O(T^{\frac{1}{2}})$, $\sigma = O(T^{\frac{1}{2}})$ *as* $d, T \to \infty$ *and* Θ, U *are the solutions to* (5.4) *and* (5.10) *respectively.*

5.5.2. Additional Insights and Specific Example. In this section, we show that under Assumption 5.2.1 (part (a)) and Assumption 5.2.2, the functions A, B, E in Definition 5.2.1 are the limiting descriptions of the corresponding moments of finite-dimensional data. We first show that a piecewise-constant function induced by the finite-dimensional covariance matrix Σ_d of the data $\{x_i^t\}_{1\le i\le d}$ converges uniformly to the function A defined in Definition 5.2.1. Before we state the convergence result, we introduce some necessary definitions and notations.

Definition 5.5.3. Given a symmetric $d \times d$ matrix Σ with real entries, define a piecewise-constant function on $[0,1]^2$ by dividing $[0,1]^2$ into d^2 smaller squares each of length $1/d$ and set

$$W_\Sigma(x,y) := \Sigma(i,j) \text{ if } \lceil dx \rceil = i, \lceil py \rceil = j.$$

Recalling that Σ_d denotes the covariance matrix of data $\{x_i^t\}_{1 \le i \le d}$, it is easy to see that as $d \to \infty$, under Assumption 5.2.1 and Assumption 5.2.2, we have

$$\|W_{\Sigma_d} - A\|_\infty := \sup_{x,y \in [0,1]} |W_{\Sigma_d}(x,y) - A(x,y)| \to 0.$$

Indeed, from the Definition 5.5.3, we have

$$W_{\Sigma_d}(x,y) - A(x,y) = \begin{cases} 0, & \lceil dx \rceil, \lceil dy \rceil \in \{0,1,\cdots,d\} \\ A\Big(\frac{\lceil dx \rceil}{d}, \frac{\lceil dy \rceil}{d}\Big) - A(x,y), & \text{otherwise.} \end{cases}$$

Due to Assumption 5.2.2, we have

$$\Big|A\Big(\frac{\lceil dx \rceil}{d}, \frac{\lceil dy \rceil}{d}\Big) - A(x,y)\Big| \le \Big|A\Big(\frac{\lceil dx \rceil}{d}, \frac{\lceil dy \rceil}{d}\Big) - A\Big(x, \frac{\lceil dy \rceil}{d}\Big)\Big| + \big|A\Big(x, \frac{\lceil dy \rceil}{d}\Big) - A(x,y)\big| \le \frac{2C_3}{d}.$$

Therefore $|W_{\Sigma_d}(x,y) - A(x,y)| \le 2C_3 d^{-1}$ for all $x, y \in [0,1]$. The claimed convergence result hence follows from definition of $\|\cdot\|_\infty$. Similarly, we can extend Definition 5.5.3 to any m-tensor and study the uniform convergence for functions defined $[0,1]^m$ with any $m \in \mathbb{N}$. In that way we can interpret B and E in Definition 5.2.1 as the $C([0,1]^4)$-limit and the $C([0,1]^8)$-limit of the functions corresponding to the fourth moment tensor and the eighth moment tensor of x^t respectively.

We now provide a concrete example of a model that satisfies our Assumption 5.2.2. Recall that we consider the data $\{x_i^t\}_{t \ge 1, i \in [d]}$ as being generated as the discretizations of the process $\{W(s,x)\}_{s \in [0,\tau], x \in [0,1]}$ defined in Definition 5.2.1.

Sinusoidal Covariance. For any $s \ge 0$, let $\{W(s,x)\}_{x \in [0,1]}$ be a centered stochastic process with the covariance function A given by

$$A(x,y) = a_0 + \sum_{k=1}^{\infty} b_k \cos\left(2\pi k\,(x-y)\right), \qquad \forall\, x,y \in [0,1]. \tag{5.12}$$

such that there exists $\varepsilon > 0$ such that

$$\sum_{k=1}^{\infty} k^{5+\varepsilon} b_k^2 < \infty. \tag{5.13}$$

Define the finite dimensional data covariance matrix as

$$\Sigma_d(i,j) = a_0 + \sum_{k=1}^{\infty} b_k \cos\left(2\pi k(i-j)/d\right), \qquad \forall\, i,j \in [d]. \tag{5.14}$$

Let A be as defined in (5.12) and W_{Σ_d} be as defined in Definition 5.5.3. When (5.13) holds, it is easy to see that $\|A - W_{\Sigma_d}\|_\infty \to 0$ as $d \to \infty$. Condition (5.13) guarantees that A defined in (5.12) satisfies Assumption 5.2.2.

If we further assume that $W(s,\cdot)$ is a Gaussian process, we have relatively easy expressions for the fourth and eighth moments, thanks to Isserlis' Theorem. That is, for any $k \in \mathbb{N}$, we have

$$\mathbb{E}\big[\prod_{i=1}^{2k} W(s,x_i)\big] = \sum_{p \in P_{2k}^2} \prod_{(i,j)\in p} \mathbb{E}\big[W(s,x_i)W(s,x_j)\big], \tag{5.15}$$

where P_{2k}^2 is the set of all pairings of $\{1,\cdots,2k\}$. Based on this, we have the following result.

Lemma 5.5.4. *The function A defined by* (5.12) *and* (5.13) *satisfies Assumption 5.2.2. Furthermore, if W is Gaussian, its fourth moment B and eighth moment E satisfy Assumption 5.2.2.*

We emphasize here that the Gaussian assumption in Lemma 5.5.4 is purely made for the sake of simplicity. Based on generalizations of Isserlis' theorem, computations similar to those required to prove Lemma 5.5.4 could be carried out in the elliptical setting [ZYB21], mixture of Gaussian setting [Vig12] and beyond.

Next, we provide more explicit decay properties of the scaling limits that govern the limiting behavior of the finite-dimensional and finite-iteration MSE and PE in the low-noise setup.

Proposition 5.5.5. *Invoke the the assumptions in Theorem 5.3.1, under the low-noise set up.*

(a) *For an orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ of $L^2([0,1])$, if $A : [0,1]^2 \to \mathbb{R}$ admits the following decomposition $A(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)$, where the sum converges in $L^2([0,1]^2)$ and*

$\lambda_1 \geq \lambda_2 \geq \cdots > 0$, with $\lambda := \inf_i \lambda_i > 0$, we have that $\mathsf{MSE}(\tau) \leq \mathsf{MSE}(0)\exp(-2\alpha\lambda\tau)$, for any $\tau \in (0,\infty)$.

(b) *For any $\phi \in L^2([0,1])$, if A satisfies $\int_0^1\int_0^1 \phi(x)A(x,y)\phi(y)\mathrm{d}x\mathrm{d}y \geq 0$, we have that $\mathsf{PE}(\tau) \leq \mathsf{MSE}(0)/\alpha\tau$, for any $\tau \in (0,\infty)$.*

In the following, we provide two specific examples illustrating the above propositions.

(1) Let covariance matrix be defined as

$$\Sigma_d(i,j) = 1 + \cos\left(2\pi(i-j)/d\right) \quad \text{for all} \quad i,j \in [d].$$

Then, the function A is given by $A(x,y) = 1 + \cos(2\pi(x-y))$ for all $x, y \in [0,1]$. Note that, A could also be represented as

$$A(x,y) = 1 + \frac{1}{2}\big(\sqrt{2}\cos(2\pi x)\big)\big(\sqrt{2}\cos(2\pi y)\big) + \frac{1}{2}\big(\sqrt{2}\sin(2\pi x)\big)\big(\sqrt{2}\sin(2\pi y)\big).$$

Therefore A satisfies the conditions in part (a) of Proposition 5.5.5 with $\lambda_1 = 1, \lambda_2 = \lambda_3 = 1/2$. Hence, we look at the MSE as defined in (5.11). According to Proposition 5.5.5, the scaling limit of the MSE satisfies $\mathsf{MSE}(\tau) \leq \mathsf{MSE}(0)\exp(-\alpha\tau)$ for any $\tau \in (0,\infty)$.

(2) Now, let the covariance matrix be given by

$$\Sigma_d(i,j) = \Big(\frac{|i-j|}{d} - \frac{1}{2}\Big)^2 - 2\Big(\frac{|i-j|}{d} - \frac{1}{2}\Big)^4 \quad \text{for all} \quad i,j \in [d].$$

Then the function A is given by $A(x,y) = \left(|x-y| - \frac{1}{2}\right)^2 - 2\left(|x-y| - \frac{1}{2}\right)^4$ for all $x, y \in [0,1]$. By a Fourier series expansion, we also have that

$$A(x,y) = a_0 + \sum_{k=1}^{\infty}\frac{b_k}{2}\big(\sqrt{2}\cos(2\pi x)\big)\big(\sqrt{2}\cos(2\pi y)\big) + \sum_{k=1}^{\infty}\frac{b_k}{2}\big(\sqrt{2}\sin(2\pi x)\big)\big(\sqrt{2}\sin(2\pi y)\big),$$

where $a_0 = 7/120$ and $b_k = 6/\pi^2 k^4$ for all $k \geq 1$. It can also be checked that (5.13) is satisfied with $\varepsilon = 1$. As there is no positive uniform lower bound of the eigenvalues, we look at the PE as defined in (5.11). According to part (b) of Proposition 5.5.5, we have for any $\tau \in (0,\infty)$, $\mathsf{PE}(s) \leq \mathsf{MSE}(0)/\alpha\tau$.

More generally, to compute the scaling limits and the fluctuations, in particular in the moderate and high-noise setups, one invariably has to resort to numerical procedures. We refer, for example, to [LPS14] regarding the details, and leave a thorough investigation of this as future work.

5.6. Proof Overview

The high level idea behind our proofs is to study the interpolated SGD iterates in (5.3) as a sequence of random variables in the space $C([0,\tau];C([0,1]))$. Under appropriate assumptions, we first prove that the sequence of random variables is tight in $C([0,\tau];C([0,1]))$; this forms a major portion of our analysis. As a consequence, we have that the sequential weak limit exists. Next, we identify the sequential limits by deriving an ODE/SDE that the limits satisfy. Last, we prove that all the sequential weak limits are the same and therefore the sequence of random variables converges weakly to a unique limit that solves the ODE/SDEs.

5.6.1. For Theorem 5.3.1.

Our first result shows the tightness of $\{\overline{\Theta}^{d,T}(\cdot,\cdot)\}_{d\geq 1,T>0}$ in the space $C\left([0,\tau];C([0,1])\right)$. It is based on two technical results (see Proposition 5.7.1 and Proposition 5.7.2) which forms the major part of our analysis.

Theorem 5.6.1 (Tightness). *Let Assumptions 5.2.1 and 5.2.2 hold, and the initial conditions in Theorem 5.3.1 are satisfied. Further suppose that there is a uniform constant $C_{s,1}>0$ such that $\max(\eta dT,\eta\sigma T^{\frac{1}{2}})\leq C_{s,1}$. Then $\{\overline{\Theta}^{d,T}\}_{d\geq 1,T>0}$ defined in (5.3) is tight in $C\left([0,\tau];C([0,1])\right)$. Hence any subsequence of $\{\overline{\Theta}^{d,T}\}_{d\geq 1,T>0}$ has a further weakly convergent subsequence with its limit in $C\left([0,\tau];C([0,1])\right)$.*

Theorem 5.6.2 (Limit Identification). *Invoke the conditions in Theorem 5.6.1. Then for any $\tau>0$ and any subsequence $\{\Theta^{d_k,T_k}\}_{k\geq 1}$ of $\{\Theta^{d,T}\}_{d\geq 1,T>0}$, there further exists a subsequence converging weakly to a function $\Theta\in C([0,\tau];C([0,1]))$ as $d_k,T_k\to\infty$. Furthermore, for any bounded smooth function $f:[0,\tau]\to\mathbb{R}$, for any $s\in[0,\tau]$, $x\in[0,1]$, we have*

$$
\begin{aligned}
&-\int_0^s \Theta(u,x)f'(u)\mathrm{d}u+f(s)\Theta(s,x)-f(0)\Theta(0,x)\\
&=-\eta dT\big(\int_0^s\int_0^1 f(u)A(x,y)\Theta(u,y)\mathrm{d}y\mathrm{d}u+o(1)\big)+\sigma^2\eta^2T\big(\int_0^s f(u)\mathrm{d}\xi_1(u,x)+o(1)\big).
\end{aligned}
$$

where $\{\xi_1(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ *is a Gaussian field process with covariance given by* (5.29).

The above theorem is proved in Section 5.7.2. As an immediately consequence, we have that:

(a) When $\eta dT \to \alpha$ and $\sigma^2/d^2T \to 0$ as $d,T\to\infty$, we get

$$-\int_0^s \Theta(u,x)f'(u)\mathrm{d}u + f(s)\Theta(s,x) - f(0)\Theta(0,x) = -\alpha\int_0^s\int_0^1 f(u)A(x,y)\Theta(u,y)\mathrm{d}y\mathrm{d}u. \tag{5.16}$$

(b) When $\eta dT \to \alpha$ and $\sigma^2/d^2T \to \beta^2$ as $d,T\to\infty$, we get

$$\begin{aligned}&-\int_0^s \Theta(u,x)f'(u)\mathrm{d}u + f(s)\Theta(s,x) - f(0)\Theta(0,x)\\ &= -\alpha\int_0^s\int_0^1 f(u)A(x,y)\Theta(u,y)\mathrm{d}y\mathrm{d}u + \alpha^2\beta^2\int_0^s f(u)\mathrm{d}\xi_1(u,x).\end{aligned}$$

(c) When $\eta\sigma T^{\frac{1}{2}} \to \alpha$ and $\sigma^2/d^2T \to \infty$ as $d,T\to\infty$, we get

$$-\int_0^s \Theta(u,x)f'(u)\mathrm{d}u + f(s)\Theta(s,x) - f(0)\Theta(0,x) = \alpha^2\int_0^s f(u)\mathrm{d}\xi_1(u,x).$$

Our main result in Theorem 5.3.1 follows from the above results; proof is provided in Section 5.7.2.

5.6.2. For Theorem 5.3.5. We next turn to proving the fluctuation results. From the definition of our interpolation process in Section 5.1, we immediately have that $U^{d,T} \in C([0,\tau];C([0,1]))$. According to the definition of $\Theta^{d,T}$, under the low-noise setup in Theorem 5.3.1, for any $0 \le t \le$

$\lfloor \tau T \rfloor - 1 := N-1$, $i \in [d]$, we have

$$\begin{aligned}
&U^{d,T}(\frac{t+1}{T},\frac{i}{d}) - U^{d,T}(\frac{t}{T},\frac{i}{d}) \\
=&-\eta\sum_{j=1}^{d} W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})U^{d,T}(\frac{t}{T},\frac{j}{d}) \\
&-\eta\gamma\sum_{j=1}^{d}\big(W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}\big[W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})\big]\big)\Theta(\frac{t}{T},\frac{j}{d}) \\
&-\gamma(\Theta(\frac{t+1}{T},\frac{i}{d}) - \Theta(\frac{t}{T},\frac{i}{d}) + \eta\sum_{j=1}^{d} A(\frac{i}{d},\frac{j}{d})\Theta(\frac{t}{T},\frac{j}{d})) \\
&+\eta\gamma W(\frac{t}{T},\frac{i}{d})\varepsilon^{t}.
\end{aligned} \tag{5.17}$$

The idea of deriving the limit of $\{U^{d,T}\}_{d\geq 1, T>0}$ is now similar to the idea of deriving the limit of $\{\overline{\Theta}^{d,T}\}_{d\geq 1, T>0}$. Specifically, we first have the following tightness result.

Theorem 5.6.3 (Tightness). *Let Assumptions 5.2.1 and 5.2.2 hold and further suppose the initial conditions in Theorem 5.3.1 and Theorem 5.3.5 are satisfied. If $\eta = \frac{\alpha}{dT}$, $\sigma = o(dT^{\frac{1}{2}})$ and there exists a uniform constant $C_{s,2}$ such that $\max(\gamma T^{-\frac{1}{2}}, \gamma d^{-1}, \gamma\sigma d^{-1}T^{-\frac{1}{2}}) \leq C_{s,2}$ for all $d \geq 1, T > 0$, then $\{U^{d,T}(\cdot,\cdot)\}_{d\geq 1,T>0}$ as defined in (5.7) is tight in $C\left([0,\tau]; C([0,1])\right)$. Hence any subsequence of $\{U^{d,T}(\cdot,\cdot)\}_{d\geq 1,T>0}$ has a further weakly convergent subsequence with limit in $C\left([0,\tau]; C([0,1])\right)$.*

Theorem 5.6.4 (Limit Identification). *Invoke the conditions in Theorem 5.6.3. For any $\tau > 0$ and any subsequence $\{U^{d_k,T_k}\}_{k\geq 1}$ of $\{U^{d,T}\}_{d\geq 1,T>0}$, there exists a further subsequence converging weakly to a stochastic function $U \in C([0,\tau]; C([0,1]))$ as $d_k, T_k \to \infty$. Furthermore, for any bounded smooth function $f : [0,\tau] \to \mathbb{R}$, U satisfies that for any $s \in [0,\tau], x \in [0,1]$,*

$$\begin{aligned}
&-\int_0^s U(u,x)f'(u)\mathrm{d}u + f(s)U(s,x) - f(0)U(0,x) \\
=&-\alpha\int_0^s\int_0^1 f(u)A(x,y)U(u,y)\mathrm{d}y\mathrm{d}u + \alpha\beta\int_0^s f(u)\mathrm{d}\xi_2(u,x)1_{\{\gamma\sigma d^{-1}T^{-\frac{1}{2}}\to\beta\}} \\
&+\alpha\zeta\int_0^s f(u)\mathrm{d}\xi_3(u,x)1_{\{\gamma T^{-\frac{1}{2}}\to\zeta\}} + O(\gamma d^{-1} + \gamma T^{-1}) + o(1).
\end{aligned}$$

The above theorem is proved in Section 5.8.2. As an immediately consequence, we have that:

(a) When $T = o(d^2)$, $\sigma = O(d)$ and $\gamma T^{-\frac{1}{2}} \to \zeta \in (0, \infty)$, we can show that $O(\gamma d^{-1} + \gamma T^{-1}) = o(1)$. Therefore, we get

$$\begin{aligned} &- \alpha^{-1}\big(\int_0^s U(u,x) f'(u) \mathrm{d}u + f(s)U(s,x) - f(0)U(0,x)\big) \\ = &- \int_0^s \int_0^1 f(u) A(x,y) U(u,y) \mathrm{d}y \mathrm{d}u + \beta \int_0^s f(u) \mathrm{d}\xi_2(u,x) 1_{\{\gamma\sigma d^{-1} T^{-\frac{1}{2}} \to \beta\}} + \zeta \int_0^s f(u) \mathrm{d}\xi_3(u,x). \end{aligned}$$

(b) When $\max(d, T^{\frac{1}{2}}) \ll \sigma \ll dT^{\frac{1}{2}}$ and $\gamma\sigma d^{-1} T^{-\frac{1}{2}} \to \beta \in (0,\infty)$, we can show that $\gamma T^{-\frac{1}{2}} \to 0$ and $O(\gamma d^{-1} + \gamma T^{-1}) = o(1)$. Therefore we get

$$\begin{aligned} &- \alpha^{-1}\big(\int_0^s U(u,x) f'(u) \mathrm{d}u + f(s)U(s,x) - f(0)U(0,x)\big) \\ = &- \int_0^s \int_0^1 f(u) A(x,y) U(u,y) \mathrm{d}y \mathrm{d}u + \beta \int_0^s f(u) \mathrm{d}\xi_2(u,x). \end{aligned}$$

(c) When $d = O(T^{\frac{1}{2}})$, $\sigma = O(T^{\frac{1}{2}})$ and $\gamma d^{-1} \to 0$, we can show that $\gamma\sigma d^{-1} T^{-\frac{1}{2}} \to 0$, $\gamma T^{-\frac{1}{2}} \to 0$ and $O(\gamma d^{-1} + \gamma T^{-1}) = o(1)$. Therefore we get

$$- \int_0^s U(u,x) f'(u) \mathrm{d}u + f(s)U(s,x) - f(0)U(0,x) = -\alpha \int_0^s \int_0^1 f(u) A(x,y) U(u,y) \mathrm{d}y \mathrm{d}u.$$

This proves Theorem 5.3.5.

5.6.3. Existence and Uniqueness of the Solutions. To fully complete the proofs of our main results in Theorems 5.3.1 and 5.3.5, we also need to provide the following existence and uniqueness results on the solutions to the corresponding SDEs. The ODE case follows by Picard-Lindelöf theorem (see, for example, [Arn92]). Furthermore, since (5.5), (5.6), (5.9) and (5.10) can be considered as certain degenerate forms of (5.8), we specifically focus on the existence and uniqueness of solution to (5.8). Results similar to Proposition 5.6.5 below also hold for (5.5), (5.6), (5.9) and (5.10).

Theorem 5.6.5. *Let Assumption 5.2.2 hold.*

(1) If the initial condition of (5.8) *satisfies* $U(0,\cdot) \in C([0,1])$, *then there exists a unique solution* $\{U(s,x)\}_{s\in[0,\tau], x\in[0,1]} \in C([0,\tau]; C([0,1]))$ *to the SDE* (5.8).

(2) If the initial condition of (5.8) *satisfies* $U(0,\cdot) \in L^2([0,1])$, *then there exists a unique solution* $\{U(s,x)\}_{s\in[0,\tau],x\in[0,1]} \in C([0,\tau];L^2([0,1]))$ *to the SDE* (5.8). *Furthermore, the solution satisfies the following stability property:*

$$\begin{aligned}\mathbb{E}\big[\sup_{s\in[0,\tau]} \|U(s,\cdot)\|^2_{L^2([0,1])}\big] &\le 4\big(\mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big] + C_2\alpha^2\beta^2\tau \\ &\quad + (C_5 + C_2^2)\alpha^2\zeta^2 \int_0^\tau \|\Theta(s,\cdot)\|^2_{L^2([0,1])}\,\mathrm{d}s\big) \exp\left(4C_2^2\alpha^2\tau^2\right).\end{aligned} \tag{5.18}$$

We remark that the above results not only provide theoretical support but also provide the necessary conditions for provably computing the solutions numerically; see, for example, [LPS14] for details.

5.7. Proofs for the Scaling Limits

5.7.1. Tightness of $\{\overline{\Theta}^{d,T}\}_{d\ge1,T>0}$.

We first start with the following two main propositions.

Proposition 5.7.1. *Assume that Assumption 5.2.1 and Assumption 5.2.2 hold and there is a uniform constant* $C_{s,1} > 0$ *such that* $\max(\eta dT, \eta\sigma T^{\frac{1}{2}}) \le C_{s,1}$. *Then there exists a positive constant* C *depending on* τ *and* $C_{s,1}$ *such that for any* $d, T > 0$ *and* $i \in [d]$ *and any* $0 \le t_1 < t_2 \le \lfloor \tau T \rfloor$,

$$\mathbb{E}\big[\,\big|\Delta\theta_i^{t_2} - \Delta\theta_i^{t_1}\big|^4\,\big] \le C\left(\frac{t_2 - t_1}{T}\right)^2. \tag{5.19}$$

Proof of Proposition 5.7.1. For any $0 \le t_1 < t_2 \le \lfloor \tau T \rfloor \coloneqq N$, we have

$$\Delta\theta_i^{t_2} - \Delta\theta_i^{t_1} = -\eta \sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d} \mathbb{E}[x_i^t x_j^t]\Delta\theta_j^t - \eta \sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d} \left(x_i^t x_j^t - \mathbb{E}[x_i^t x_j^t]\right)\Delta\theta_j^t + \eta\sum_{t=t_1}^{t_2-1} x_i^t \varepsilon^t$$

which implies that

$$
\begin{aligned}
\big|\Delta\theta_i^{t_2} - \Delta\theta_i^{t_1}\big|^4 &\leq 27\eta^4 \underbrace{\big(\sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d} \mathbb{E}[x_i^t x_j^t]\Delta\theta_j^t\big)^4}_{S_{i,1}} \\
&+ 27\eta^4 \underbrace{\big(\sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d}\big(x_i^t x_j^t - \mathbb{E}[x_i^t x_j^t]\big)\Delta\theta_j^t\big)^4}_{S_{i,2}} + 27\eta^4 \underbrace{\big(\sum_{t=t_1}^{t_2-1} x_i^t\varepsilon^t\big)^4}_{S_{i,3}}.
\end{aligned}
\tag{5.20}
$$

Sum over the space index i and take expectation and we get

$$
\frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\big[|\Delta\theta_i^{t_2} - \Delta\theta_i^{t_1}|^4\big] \leq \frac{27\eta^4}{d}\sum_{i=1}^{d}\mathbb{E}\big[S_{i,1}\big] + \frac{27\eta^4}{d}\sum_{i=1}^{d}\mathbb{E}\big[S_{i,2}\big] + \frac{27\eta^4}{d}\sum_{i=1}^{d}\mathbb{E}\big[S_{i,3}\big]. \tag{5.21}
$$

Next we will estimate the right hand side of (5.21) term by term. Define $m_t^d \coloneqq \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\left[|\Delta\theta_i^t|^4\right]$. First we have

$$
\begin{aligned}
\mathbb{E}\big[S_{i,1}\big] = \mathbb{E}\big[\sum_{r_1,r_2,r_3,r_4=t_1}^{t_2-1}\ \sum_{j_1,j_2,j_3,j_4=1}^{d}\prod_{k=1}^{4} A(\frac{i}{d},\frac{j_k}{d})\prod_{k=1}^{4}\Delta\theta_{j_k}^{r_k}\big] &\leq C_2^4\mathbb{E}\big[\sum_{r_1,r_2,r_3,r_4}\ \sum_{j_1,j_2,j_3,j_4}\prod_{k=1}^{4}\Delta\theta_{j_k}^{r_k}\big] \\
\leq C_2^4 d^4 (t_2-t_1)^3 \sum_{t=t_1}^{t_2-1} m_t^d,
\end{aligned}
$$

where the first inequality follows from Assumption 5.2.2 and the last inequality follows from $abcd \leq \frac{1}{4}(a^4+b^4+c^4+d^4)$ for any $a,b,c,d\in\mathbb{R}$. We also have that

$$
\mathbb{E}\big[S_{i,2}\big] = \mathbb{E}\big[\sum_{r_1,r_2,r_3,r_4=t_1}^{t_2-1}\ \sum_{j_1,j_2,j_3,j_4=1}^{d}\prod_{k=1}^{4}\big(x_i^{r_k}x_{j_k}^{r_k} - \mathbb{E}[x_i^{r_k}x_{j_k}^{r_k}]\big)\prod_{k=1}^{4}\Delta\theta_{j_k}^{r_k}\big].
$$

Most terms in the above sum are zeros due to the independence between x^t and $x^{t'}$ when $t\neq t'$. There are two types of terms would be preserved. To proceed, we let $\mathcal{F}_k^N$ to be the σ-algebra generated by $\big\{\{x_i^t\}_{1\leq i\leq d, 0\leq t\leq k}, \{\Delta\theta_i^t\}_{1\leq i\leq d, 0\leq t\leq N}\big\}$.

Type 1: When $r_1 = r_2 = r_3 = r_4$, nonzero terms are in the form of

$$
\sum_{j_1,j_2,j_3,j_4=1}^{d}\mathbb{E}\big[\prod_{k=1}^{4}\big(x_i^t x_{j_k}^t - \mathbb{E}\left[x_i^t x_{j_k}^t\right]\big)\prod_{k=1}^{4}\Delta\theta_{j_k}^t\big],
$$

and there are $3(t_2 - t_1)$ such terms. The sum of such terms can be upper bounded as

$$
\begin{aligned}
&3\sum_{t=t_1}^{t_2-1}\sum_{j_1,j_2,j_3,j_4=1}^{d}\mathbb{E}\big[\prod_{k=1}^{4}\left(x_i^t x_{j_k}^t - \mathbb{E}\left[x_i^t x_{j_k}^t\right]\right)\prod_{k=1}^{4}\Delta\theta_{j_k}^t\big] \\
&= 3\sum_{t=t_1}^{t_2-1}\sum_{j_1,j_2,j_3,j_4=1}^{d}\mathbb{E}\Big[\mathbb{E}\big[\prod_{k=1}^{4}\left(x_i^t x_{j_k}^t - \mathbb{E}\left[x_i^t x_{j_k}^t\right]\right)\prod_{k=1}^{4}\Delta\theta_{j_k}^t\big]|\mathcal{F}_0^N\Big] \\
&\le 3C_8'\sum_{t=t_1}^{t_2-1}\sum_{j_1,j_2,j_3,j_4=1}^{d}\mathbb{E}\big[\prod_{k=1}^{4}\Delta\theta_{j_k}^t\big] \le C_8' d^4\sum_{t=t_1}^{t_2-1} m_t^d,
\end{aligned}
$$

where the first inequality follows from Assumption 5.2.2 and $C_8' = 16C_8 + 32C_2^2C_5 + 16C_2^4$. The last inequality follows from $abcd \le \frac{1}{4}(a^4 + b^4 + c^4 + d^4)$ for any $a, b, c, d \in \mathbb{R}$.

Type 2: When r_1, r_2, r_3, r_4 are pairwise equal, nonzero terms are in the form of

$$
\sum_{j_1,j_2,j_3,j_4=1}^{d}\mathbb{E}\big[\prod_{k=1}^{2}\left(x_i^t x_{j_k}^t - \mathbb{E}\left[x_i^t x_{j_k}^t\right]\right)\prod_{k=3}^{4}\left(x_i^{t'} x_{j_k}^{t'} - \mathbb{E}\left[x_i^{t'} x_{j_k}^{t'}\right]\right)\prod_{k=1}^{2}\Delta\theta_{j_k}^t\prod_{k=3}^{4}\Delta\theta_{j_k}^{t'}\big]
$$

with $t \ne t'$ and there are $3(t_2 - t_1)^2 - 3(t_2 - t_1)$ such terms. Sum of such terms can be upper bounded as

$$
\begin{aligned}
&3\sum_{t\ne t'}\sum_{j_1,j_2,j_3,j_4=1}^{d}\mathbb{E}\big[\prod_{k=1}^{2}\left(x_i^t x_{j_k}^t - \mathbb{E}\left[x_i^t x_{j_k}^t\right]\right)\prod_{k=3}^{4}\left(x_i^{t'} x_{j_k}^{t'} - \mathbb{E}\left[x_i^{t'} x_{j_k}^{t'}\right]\right)\prod_{k=1}^{2}\Delta\theta_{j_k}^t\prod_{k=3}^{4}\Delta\theta_{j_k}^{t'}\big] \\
&= 3\sum_{t\ne t'}\mathbb{E}\Big[\sum_{j_1,j_2,j_3,j_4=1}^{d}\mathbb{E}\big[\prod_{k=1}^{2}\left(x_i^t x_{j_k}^t - \mathbb{E}\left[x_i^t x_{j_k}^t\right]\right)\prod_{k=3}^{4}\left(x_i^{t'} x_{j_k}^{t'} - \mathbb{E}\left[x_i^{t'} x_{j_k}^{t'}\right]\right)\prod_{k=1}^{2}\Delta\theta_{j_k}^t\prod_{k=3}^{4}\Delta\theta_{j_k}^{t'}\big]|\mathcal{F}_0^N\Big] \\
&\le 3C_5'^2\sum_{t\ne t'}\sum_{j_1,j_2,j_3,j_4=1}^{d}\mathbb{E}\big[\prod_{k=1}^{2}\Delta\theta_{j_k}^t\prod_{k=3}^{4}\Delta\theta_{j_k}^{t'}\big] \le 3C_5'^2 d^4(t_2 - t_1)\sum_{t=t_1}^{t_2-1} m_t^d,
\end{aligned}
$$

where the first inequality follows from Assumption 5.2.2 and $C_5' = C_5 + C_2^2$. The last inequality follows from $4(abcd) \le (a^4 + b^4 + c^4 + d^4)$ for any $a, b, c, d \in \mathbb{R}$. Combine the two upper bounds and we have

$$
\mathbb{E}\big[S_{i,2}\big] \le 3\big(C_8' + C_5'^2(t_2 - t_1)\big)d^4\sum_{t=t_1}^{t_2-1} m_t^d \le \big(C_8' + C_5'^2\big)(t_2 - t_1)d^4\sum_{t=t_1}^{t_2-1} m_t^d.
$$

Last, due to Assumption 5.2.2, we estimate $\mathbb{E}[S_{i,3}]$ as follows:

$$\begin{aligned}\mathbb{E}[S_{i,3}] &= \sum_{t=t_1}^{t_2-1} B(\frac{i}{d},\frac{i}{d},\frac{i}{d},\frac{i}{d})\mathbb{E}\big[(\varepsilon^t)^4\big] + \sum_{t\neq t'} A(\frac{i}{d},\frac{i}{d})A(\frac{i}{d},\frac{i}{d})\mathbb{E}\left[(\varepsilon^t)^2\right]\mathbb{E}\big[(\varepsilon^{t'})^2\big] \\ &\leq C_5 C_1 \sigma^4 (t_2-t_1) + C_2^2\sigma^4(t_2-t_1)^2 \leq \big(C_5C_1 + C_2^2\big)\sigma^4(t_2-t_1)^2.\end{aligned}$$

Plug our estimations of $\mathbb{E}[S_{i,,1}]$, $\mathbb{E}[S_{i,2}]$, $\mathbb{E}[S_{i,3}]$ into (5.21) and we have

$$\begin{aligned}\frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\big[|\Delta\theta_i^{t_2} - \Delta\theta_i^{t_1}|^4\big] &\leq 27\eta^4 d^4\big[C_2^4(t_2-t_1)^3 + 3\big(C_8' + C_5'^2\big)(t_2-t_1)\big] \\ &\qquad \times \sum_{t=t_1}^{t_2-1} m_t^d + 27\big(C_5C_1 + C_2^2\big)\eta^4\sigma^4(t_2-t_1)^2.\end{aligned}\tag{5.22}$$

Pick $t_1 = 0,\ t_2 = t \leq N$ in (5.22). We have for any $t \leq N$,

$$\begin{aligned} m_t^d &\leq 8m_0^d + 216\eta^4 d^4\big[C_2^4 t^3 + 3\big(C_8' + C_5'^2\big)t\big]\sum_{k=0}^{t-1} m_k^d + 216\big(C_5C_1 + C_2^2\big)\eta^4\sigma^4 t^2 \\ &\leq 8m_0^d + 216\big(C_2^4 + 3C_8' + 3C_5'^2\big)\eta^4 d^4 t^3 \sum_{k=0}^{t-1} m_k^d + 216\big(C_5C_1 + C_2^2\big)\eta^4\sigma^4 t^2.\end{aligned}$$

According to the discrete Gronwall's inequality, we have for any $t \leq N$,

$$\begin{aligned} m_t^d &\leq 216\big(C_5C_1 + C_2^2\big)\eta^4\sigma^4 t^2 \\ &\quad + 216^2\big(C_5C_1 + C_2^2\big)\eta^8 d^4\sigma^4 t^3 \sum_{k=0}^{t-1} k^2 \exp\big(216\big(C_2^4 + 3C_8' + 3C_5'^2\big)\eta^4 d^4 t^3 (t-k-1)\big) \\ &\leq 216\big(C_5C_1 + C_2^2\big)\eta^4\sigma^4 t^2 + 216^2\big(C_5C_1 + C_2^2\big)\eta^8 d^4 \sigma^4 t^6 \exp\big(216\big(C_2^4 + 3C_8' + 3C_5'^2\big)\eta^4 d^4 t^4\big) \\ &\leq 216\big(C_5C_1 + C_2^2\big)C_{s,1}^4\tau^2 + 216^2\big(C_5C_1 + C_2^2\big)C_{s,1}^8\tau^6 \exp\big(216\big(C_2^4 + 3C_8' + 3C_5'^2\big)C_{s,1}^4\tau^4\big),\end{aligned}$$

where the last inequality follows from $\max(\eta d T, \eta\sigma T^{\frac{1}{2}}) \leq C_{s,1}$ and $N \leq \tau T$. We simplify the upper bound of m_t^d as $m_t^d \leq C_\tau$ for some positive constant C_τ independent of d, T, σ. Apply the upper bound of m_t^d to (5.20) along with estimations of $\mathbb{E}[S_{i,,1}]$, $\mathbb{E}[S_{i,2}]$, $\mathbb{E}[S_{i,3}]$ and we get

$$\begin{aligned}
&\mathbb{E}\left[|\Delta\theta_i^{t_2} - \Delta\theta_i^{t_1}|^2\right] \\
&\leq 27\eta^4 d^4 \big(C_2^4(t_2-t_1)^3 + 3(C_8' + C_5'^2)(t_2-t_1)\big)\sum_{t=t_1}^{t_2-1} m_t^d + 27(C_5C_1 + C_2^2)\eta^4\sigma^4(t_2-t_1)^2 \\
&\leq 27C_\tau\big(C_2^4 + 3C_8' + 3C_5'^2\big)\eta^4 d^4 (t_2-t_1)^4 + 27(C_5C_1 + C_2^2)\eta^4\sigma^4(t_2-t_1)^2 \\
&\leq 27C_\tau\big(C_2^4 + 3C_8' + 3C_5'^2\big)C_{s,1}^4\left(\frac{t_2-t_1}{T}\right)^4 + 27(C_5C_1 + C_2^2)C_{s,1}^4\left(\frac{t_2-t_1}{T}\right)^2,
\end{aligned}$$

where the last inequality follows from $\max(\eta dT, \eta\sigma T^{\frac{1}{2}}) \leq C_{s,1}$. Last due to $\left(\frac{t_2-t_1}{T}\right)^2 \leq \tau^2$ for any $0 \leq t_1 < t_2 \leq N$, (5.19) is proved. ■

Proposition 5.7.2. *Assume that Assumption 5.2.1 and Assumption 5.2.2 hold and there exists a uniform constant $C_{s,1} > 0$ such that $\max(\eta dT, \eta\sigma T^{\frac{1}{2}}) \leq C_{s,1}$. Then there exists a positive constant C depending on τ and $C_{s,1}$ such that for any $d, T > 0$, $i, j \in [d]$ and $0 \leq t_1 < t_2 \leq \lfloor \tau T \rfloor$,*

$$\mathbb{E}\big[\big|(\Delta\theta_i^{t_2} - \Delta\theta_j^{t_2}) - (\Delta\theta_i^{t_1} - \Delta\theta_j^{t_1})\big|^4\big] \leq C\left(\frac{i-j}{d}\right)^4\left(\frac{t_2-t_1}{T}\right)^2. \tag{5.23}$$

Furthermore, if initial conditions in Theorem 5.3.1 hold, then for any $i, j \in [d]$ and any $0 \leq t \leq \lfloor \tau T \rfloor$,

$$\mathbb{E}\big[\big|\Delta\theta_i^t - \Delta\theta_j^t\big|^4\big] \leq \big(8L^4 + 8\tau^2 C\big)\left(\frac{i-j}{d}\right)^4. \tag{5.24}$$

Proof of Proposition 5.7.2. According to (5.2), for any $i, j \in [d]$ and any $0 \leq t \leq \lfloor \tau T \rfloor - 1 := N - 1$, we have

$$\begin{aligned}
\Delta\theta_i^{t+1} - \Delta\theta_j^{t+1} = (\Delta\theta_i^t - \Delta\theta_j^t) &- \eta\sum_{l=1}^d \left(\mathbb{E}\left[x_i^t x_l^t\right] - \mathbb{E}\left[x_j^t x_l^t\right]\right)\Delta\theta_l^t \\
&- \eta\sum_{l=1}^d \left(x_i^t x_l^t - \mathbb{E}\left[x_i^t x_l^t\right] - x_j^t x_l^t + \mathbb{E}\left[x_j^t x_l^t\right]\right)\Delta\theta_l^t + \eta(x_i^t - x_j^t)\varepsilon^t.
\end{aligned}$$

Summing over the time index and taking expectation of the absolute value of both sides, we get for any $0 \leq t_1 < t_2 \leq N$,

$$
\begin{aligned}
&\mathbb{E}\big[\big|(\Delta\theta_i^{t_2} - \Delta\theta_j^{t_2}) - (\Delta\theta_i^{t_1} - \Delta\theta_j^{t_1})\big|^4\big] \\
&\leq 27\eta^4\mathbb{E}\big[\underbrace{\big(\sum_{t=t_1}^{t_2-1}\sum_{l=1}^{d}\big(\mathbb{E}\left[x_i^t x_l^t\right] - \mathbb{E}\left[x_j^t x_l^t\right]\big)\,\Delta\theta_l^t\big)^4}_{\mathbb{E}[S_{i,j,1}]}\big] \\
&\quad + 27\eta^4\mathbb{E}\big[\underbrace{\big(\sum_{t=t_1}^{t_2-1}\sum_{l=1}^{d}\big(x_i^t x_l^t - \mathbb{E}\left[x_i^t x_l^t\right] - x_j^t x_l^t + \mathbb{E}\left[x_j^t x_l^t\right]\big)\,\Delta\theta_l^t\big)^4}_{\mathbb{E}[S_{i,j,2}]}\big] \qquad (5.25) \\
&\quad + 27\eta^4\mathbb{E}\big[\underbrace{\big(\sum_{t=t_1}^{t_2-1}(x_i^t - x_j^t)\varepsilon^t\big)^4}_{\mathbb{E}[S_{i,j,3}]}\big].
\end{aligned}
$$

Next we will estimate the right hand side of (5.25) term by term. First we have

$$
\begin{aligned}
\mathbb{E}\big[S_{i,j,1}\big] &= \mathbb{E}\big[\sum_{r_1,r_2,r_3,r_4=t_1}^{t_2-1}\;\sum_{l_1,l_2,l_3,l_4=1}^{d}\prod_{k=1}^{4}\big(A(\frac{i}{d},\frac{j_k}{d}) - A(\frac{j}{d},\frac{j_k}{d})\big)\prod_{k=1}^{4}\Delta\theta_l^t\big] \\
&\leq C_3^4\big(\frac{i-j}{d}\big)^4(t_2-t_1)^3 d^4\sum_{t=t_1}^{t_2-1} m_t^d,
\end{aligned}
$$

where the inequality follows from Assumption 5.2.2 and $4(abcd) \leq a^4+b^4+c^4+d^4$ for all $a, b, c, d \in \mathbb{R}$. Similar to $\mathbb{E}[S_{i,2}]$ in the proof of Proposition 5.7.1, in the estimation of $\mathbb{E}[S_{i,j,2}]$, there are two types of nonzero terms.

Type 1: The first type of nonzero terms are in the form of

$$
\sum_{l_1,l_2,l_3,l_4=1}^{d}\mathbb{E}\big[\prod_{k=1}^{4}\big(x_i^t x_{l_k}^t - \mathbb{E}[x_i^t x_{l_k}^t] - x_j^t x_{l_k}^t + \mathbb{E}[x_j^t x_{l_k}^t]\big)\,\big|\prod_{k=1}^{4}\Delta\theta_{l_k}^t\big|\big],
$$

and there are $3(t_2 - t_1)$ such terms. Sum of such terms can be bounded by

$$\begin{aligned}
&3\sum_{t=t_1}^{t_2-1}\sum_{l_1,l_2,l_3,l_4=1}^{d}\mathbb{E}[\prod_{k=1}^{4}\left(x_i^t x_{l_k}^t - \mathbb{E}[x_i^t x_{l_k}^t] - x_j^t x_{l_k}^t + \mathbb{E}[x_j^t x_{l_k}^t]\right)|\prod_{k=1}^{4}\Delta\theta_{l_k}^t|] \\
&= 3\sum_{t=t_1}^{t_2-1}\sum_{l_1,l_2,l_3,l_4=1}^{d}\mathbb{E}\Big[\mathbb{E}[\prod_{k=1}^{4}\left(x_i^t x_{l_k}^t - \mathbb{E}[x_i^t x_{l_k}^t] - x_j^t x_{l_k}^t + \mathbb{E}[x_j^t x_{l_k}^t]\right)|\prod_{k=1}^{4}\Delta\theta_{l_k}^t|]|\mathcal{F}_0^N\Big] \\
&\leq 3C_9'^4\big(\frac{i-j}{d}\big)^4\sum_{t=t_1}^{t_2-1}\sum_{l_1,l_2,l_3,l_4=1}^{d}\mathbb{E}[|\prod_{k=1}^{4}\Delta\theta_{l_k}^t|] \leq C_9'^4\big(\frac{i-j}{d}\big)^4 d^4\sum_{t=t_1}^{t_2-1}m_t^d,
\end{aligned}$$

where the first inequality follows from Assumption 5.2.2 and $C_9'^4 = C_9^4 + 2C_3^2C_6^2 + C_3^4$. The last inequality follows from $4(abcd) \leq a^4 + b^4 + c^4 + d^4$ for all $a, b, c, d \in \mathbb{R}$.

Type 2: The second type nonzero terms are in the form of

$$\begin{aligned}
\sum_{l_1,l_2,l_3,l_4=1}^{d}\mathbb{E}\Bigg[&\underbrace{\prod_{k=1}^{2}\left(x_i^t x_{l_k}^t - \mathbb{E}[x_i^t x_{l_k}^t] - x_j^t x_{l_k}^t + \mathbb{E}[x_j^t x_{l_k}^t]\right)\prod_{k=3}^{4}\Big(x_i^{t'} x_{l_k}^{t'} - \mathbb{E}[x_i^{t'} x_{l_k}^{t'}] - x_j^{t'} x_{l_k}^{t'} + \mathbb{E}[x_j^{t'} x_{l_k}^{t'}]\Big)}_{P_{i,j,l_1,l_2,l_3,l_4}^{t,t'}} \\
&\times|\prod_{k=1}^{2}\Delta\theta_{l_k}^t\prod_{k=3}^{4}\Delta\theta_{l_k}^{t'}|\Bigg],
\end{aligned}$$

with $t \neq t'$ and there are $3(t_2 - t_1)^2 - 3(t_2 - t_1)$ such terms. Sum of such terms can be upper bounded by

$$\begin{aligned}
&\sum_{t\neq t'}\sum_{l_1,l_2,l_3,l_4=1}^{d}\mathbb{E}\Big[\mathbb{E}[P_{i,j,l_1,l_2,l_3,l_4}^{t,t'}\prod_{k=1}^{2}|\Delta\theta_{l_k}^t|\prod_{k=3}^{4}|\Delta\theta_{l_k}^{t'}|]|\mathcal{F}_0^N\Big] \\
&\leq C_6'^4\big(\frac{i-j}{d}\big)^4\sum_{t\neq t'}\sum_{l_1,l_2,l_3,l_4=1}^{d}\mathbb{E}[\prod_{k=1}^{2}|\Delta\theta_{l_k}^t|\prod_{k=3}^{4}|\Delta\theta_{l_k}^{t'}|] \leq C_6'^4\big(\frac{i-j}{d}\big)^4 d^4(t_2-t_1)\sum_{t=t_1}^{t_2-1}m_t^d,
\end{aligned}$$

where the first inequality follows from Assumption 5.2.2 and $C_6'^4 = C_3^2 + C_6^2$. The last inequality follows from $abcd \leq \frac{a^4+b^4+c^4+d^4}{4}$ for all $a, b, c, d \in \mathbb{R}$. Therefore $\mathbb{E}[S_{i,j,2}]$ can be upper bounded as

$$\mathbb{E}[S_{i,j,2}] \leq \big(C_9'^4 + C_6'^4(t_2-t_1)\big)\big(\frac{i-j}{d}\big)^4 d^4\sum_{t=t_1}^{t_2-1}m_t^d \leq \big(C_9'^4 + C_6'^4\big)(t_2-t_1)\big(\frac{i-j}{d}\big)^4 d^4\sum_{t=t_1}^{t_2-1}m_t^d.$$

Due to Assumption 5.2.2, we can bound $\mathbb{E}[S_{i,j,3}]$ as

$$\mathbb{E}[S_{i,j,3}] = \sum_{t=t_1}^{t_2-1} \mathbb{E}\big[(x_i^t - x_j^t)^4\big]\mathbb{E}\big[\,\big(\varepsilon^t\big)^4\,\big] + \sum_{t\neq t'} \mathbb{E}\big[(x_i^t - x_j^t)^2\big]\mathbb{E}\big[(x_i^{t'} - x_j^{t'})^2\big]\mathbb{E}\big[\,\big(\varepsilon^t\big)^2\,\big]\mathbb{E}\big[\big(\varepsilon^{t'}\big)^2\big]$$

$$\le C_1 C_7^4 (t_2 - t_1)\big(\frac{i-j}{d}\big)^4\sigma^4 + C_4^4 (t_2 - t_1)^2\big(\frac{i-j}{d}\big)^4\sigma^4 \le \big(C_1 C_7^4 + C_4^4\big)(t_2 - t_1)^2\big(\frac{i-j}{d}\big)^4\sigma^4.$$

We have shown $m_t^d \le C_\tau$ for any $0 \le t \le N$ in the proof of Proposition 5.7.1. With (5.25) and the estimations on $\mathbb{E}[S_{i,j,1}]$, $\mathbb{E}[S_{i,j,2}]$, $\mathbb{E}[S_{i,j,3}]$, we have

$$\mathbb{E}\big[\big|(\Delta\theta_i^{t_2} - \Delta\theta_j^{t_2}) - (\Delta\theta_i^{t_1} - \Delta\theta_j^{t_1})\big|^4\big]$$

$$\le 27(C_9'^4 + C_6'^4 + C_3^4)(t_2 - t_1)^3\big(\frac{i-j}{d}\big)^4\eta^4 d^4 \sum_{t=t_1}^{t_2-1} m_t^d + 27\big(C_1 C_7^4 + C_4^4\big)(t_2 - t_1)^2\big(\frac{i-j}{d}\big)^4\eta^4\sigma^4$$

$$\le 27 C_\tau (C_9'^4 + C_6'^4 + C_3^4) C_{s,1}^4 \big(\frac{t_2 - t_1}{T}\big)^4\big(\frac{i-j}{d}\big)^4 + 27 C_{s,1}^4\big(\frac{t_2 - t_1}{T}\big)^2\big(\frac{i-j}{d}\big)^4,$$

where the last inequality follows from $\max(\eta d T, \eta\sigma T^{\frac{1}{2}}) \le C_{s,1}$. Therefore (5.23) is proved. Last (5.24) follows from (5.23) and the initial conditions in Theorem 5.3.1. ■

We are now ready to prove Theorem 5.6.1 based on the above two propositions.

Proof of Theorem 5.6.1. Tightness can be proved by the Kolmogorov tightness criteria [KS12, Chapter 4]. The last statement simply follows from tightness property. To apply the Kolmogorov tightness criteria, we need to verify the following two conditions:

(a) $\{\overline{\Theta}^{d,T}(0,0)\}_{d\ge 1, T>0}$ is tight in the probability space.

(b) There exists a positive constant $C_{\textbf{tight}}$ such that for any $s_1, s_2 \in [0,\tau]$ and $x_1, x_2 \in [0,1]$, we have

$$\sup_{d,T} \mathbb{E}\big[\big|\overline{\Theta}^{d,T}(s_1,x_1) - \overline{\Theta}^{d,T}(s_2,x_2)\big|^4\big] \le C_{\textbf{tight}}\big(\,|s_1 - s_2|^2 + |x_1 - x_2|^4\,\big).$$

To verify (a), it is easy to see that for any $d \ge 1, T > 0$ and $N > 0$,

$$\mathbb{P}\big(\big|\overline{\Theta}^{d,T}(0,0)\big| > N\big) \le N^{-2}\mathbb{E}\big[\big|\Delta\theta_0^0\big|^2\big] \le N^{-2}R^2 \to 0 \qquad \text{as } N \to \infty.$$

To verify (b), without loss of generality we assume that $0 \leq s_1 < s_2 \leq \tau$, $0 \leq x_1 < x_2 \leq 1$ and $\lfloor Ts_1 \rfloor = t_1 \leq \lfloor Ts_2 \rfloor = t_2$, $\lfloor dx_1 \rfloor = i \leq \lfloor dx_2 \rfloor = j$. According to (5.3), $\overline{\Theta}^{d,T}(\cdot,\cdot)$ is linear in both variables. We have

$$\begin{aligned}\mathbb{E}\big[\big|\overline{\Theta}^{d,T}(s_1,x_1) - \overline{\Theta}^{d,T}(s_2,x_2)\big|^4\big] \leq 8\mathbb{E}\big[\big|\overline{\Theta}^{d,T}(s_1,x_1) - \overline{\Theta}^{d,T}(s_2,x_1)\big|^4\big] \\ + 8\mathbb{E}\big[\big|\overline{\Theta}^{d,T}(s_2,x_1) - \overline{\Theta}^{d,T}(s_2,x_2)\big|^4\big].\end{aligned}$$

According to Proposition 5.7.1, we have

$$\mathbb{E}\big[\big|\overline{\Theta}^{d,T}(s_1,x_1) - \overline{\Theta}^{d,T}(s_2,x_1)\big|^4\big] \leq C\left(\frac{t_2+1-t_1}{T}\right)^2 \leq 5C\,(s_2-s_1)^2\,.$$

and according to Proposition 5.7.2, we have

$$\mathbb{E}\big[\big|\overline{\Theta}^{d,T}(s_2,x_1) - \overline{\Theta}^{d,T}(s_2,x_2)\big|^4\big] \leq (8M^4+8\tau^2C)\left(\frac{j+1-i}{d}\right)^4 \leq 40(M^4+\tau^2C)\,(x_2-x_1)^4\,.$$

Therefore (b) holds with $C_{\textbf{tight}} = 40C + 320(M^4+\tau^2C)$. ■

5.7.2. Limit Identification.

Proof of Theorem 5.6.2. According to Theorem 5.6.1, any subsequence

$$\{\overline{\Theta}^{d_k,T_k}(\cdot,\cdot)\}_{k\geq 1} \text{ of } \{\overline{\Theta}^{d,T}(\cdot,\cdot)\}_{d\geq 1,T>0}$$

has a further weakly convergent subsequence with limit $\Theta \in C\left([0,\tau];C([0,1])\right)$ as $d_k, T_k \to \infty$. For the simplicity of notations, we denote the convergent subsequence of $\{\overline{\Theta}^{d_k,T_k}(\cdot,\cdot)\}_{k\geq 1}$ by $\{\overline{\Theta}^{d,T}(\cdot,\cdot)\}_{d\geq 1,T>0}$ in the proof.

To identify the limit, first we rewrite (5.2) in terms of $\overline{\Theta}^{d,T}$. For any $0 \leq t \leq \lfloor \tau T \rfloor - 1$ and any $i \in [d]$:

$$\overline{\Theta}^{d,T}(\frac{t+1}{T},\frac{i}{d}) - \overline{\Theta}^{d,T}(\frac{t}{T},\frac{i}{d}) = -\eta\sum_{j=1}^{d} W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d}) + \eta W(\frac{t}{T},\frac{i}{d})\varepsilon^t. \tag{5.26}$$

Therefore for any bounded smooth function $f : [0,\tau] \to \mathbb{R}$, we have for any $s \in [0,\tau]$, $i \in [d-1]$ and $x \in [\frac{i}{d}, \frac{i+1}{d}]$,

$$\begin{aligned}
&\sum_{t=0}^{\lfloor sT\rfloor -1} f(\frac{t}{T})\big(\overline{\Theta}^{d,T}(\frac{t+1}{T},x) - \overline{\Theta}^{d,T}(\frac{t}{T},x)\big) \\
&= -\eta(i+1-dx)\sum_{t=0}^{\lfloor sT\rfloor -1}\sum_{j=1}^{d} f(\frac{t}{T})W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d}) + \eta(i+1-dx)\sum_{t=0}^{\lfloor sT\rfloor -1} f(\frac{t}{T})W(\frac{t}{T},\frac{i}{d})\varepsilon^t \\
&\quad - \eta(dx-i)\sum_{t=0}^{\lfloor sT\rfloor -1}\sum_{j=1}^{d} f(\frac{t}{T})W(\frac{t}{T},\frac{i+1}{d})W(\frac{t}{T},\frac{j}{d})\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d}) + \eta(dx-i)\sum_{t=0}^{\lfloor sT\rfloor -1} f(\frac{t}{T})W(\frac{t}{T},\frac{i+1}{d})\varepsilon^t.
\end{aligned} \tag{5.27}$$

We can rewrite the left hand side of (5.27) as

$$\text{LHS(5.27)} = \sum_{t=1}^{\lfloor sT\rfloor -1}\overline{\Theta}^{d,T}(\frac{t}{T},x)\big(f(\frac{t-1}{T}) - f(\frac{t}{T})\big) + f(\frac{\lfloor sT\rfloor -1}{T})\overline{\Theta}^{d,T}(\frac{\lfloor sT\rfloor}{T},x) - f(0)\overline{\Theta}^{d,T}(0,x).$$

When $d, T \to \infty$, since f is bounded and smooth, for any $s \in (0,\tau)$, we have

$$f(s \pm T^{-1}) = f(s) + O(T^{-1}), \qquad T\left(f(s) - f(s-T^{-1})\right) = f'(s) + O(T^{-1}).$$

Since $\{\overline{\Theta}^{d,T}\}_{d\geq 1, T>0}$ converges weakly to Θ and f, f' are continuously bounded, we have

$$\frac{1}{T}\sum_{t=1}^{\lfloor sT\rfloor -1}\overline{\Theta}^{d,T}(\frac{t}{T},x)\left(f'(\frac{t}{T}) + O(T^{-1})\right) = \int_0^{\frac{\lfloor sT\rfloor}{T}}\Theta(u,x)f'(u)\mathrm{d}u + o(1).$$

Therefore

$$\text{LHS(5.27)} = -\int_0^{\frac{\lfloor sT\rfloor}{T}}\Theta(u,x)f'(u)\mathrm{d}u + f(s)\Theta(\frac{\lfloor sT\rfloor}{T},x) - f(0)\Theta(0,x) + o(1). \tag{5.28}$$

Next, we look at the right hand side of (5.27). For any $i \in [d]$, RHS(5.27) $= -(i+1-dx)I_i^2 - (dx-i)I_{i+1}^s$,

$$I_i^s = \underbrace{-(\eta dT)\frac{1}{dT}\sum_{t=0}^{\lfloor sT\rfloor-1}\sum_{j=1}^{d} f(\frac{t}{T})A(\frac{i}{d},\frac{j}{d})\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d})}_{N_{i,1}^s} + \underbrace{(\eta dT)\frac{1}{dT}\sum_{t=0}^{\lfloor sT\rfloor-1} f(\frac{t}{T})W(\frac{t}{T},\frac{i}{d})\varepsilon^t}_{N_{i,2}^s}$$
$$\underbrace{-(\eta dT)\frac{1}{dT}\sum_{t=0}^{\lfloor sT\rfloor-1}\sum_{j=1}^{d} f(\frac{t}{T})\big(W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}\big[W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})\big]\big)\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d})}_{N_{i,3}^s}.$$

Due to the facts that f, A are continuously bounded, and $\{\overline{\Theta}^{d,T}\}_{d\geq 1, T>0}$ converges weakly to Θ, we have

$$\begin{aligned} N_{i,1}^s &= -\eta dT\big(\frac{1}{dT}\sum_{t=0}^{\lfloor sT\rfloor-1}\sum_{j=1}^{d} f(\frac{t}{T})A(\frac{i}{d},\frac{j}{d})\Theta(\frac{t}{T},\frac{j}{d}) + o(1)\big) \\ &= -\eta dT\big(\int_0^{\frac{\lfloor sT\rfloor}{T}}\int_0^1 f(u)A(\frac{i}{d},y)\Theta(u,y)\mathrm{d}y\mathrm{d}u + o(1)\big). \end{aligned}$$

For $N_{i,2}^s = (\eta dT)\frac{1}{dT}\sum_{t=0}^{\lfloor sT\rfloor-1} f(\frac{t}{T})W(\frac{t}{T},\frac{i}{d})\varepsilon^t$, note that $\mathbb{E}\left[N_{i,2^s}\right] = 0$. Since $\{W(\frac{t}{T},\frac{i}{d})\varepsilon^t\}_{t=1}^T$ is a sequence of i.i.d. random variables, the limit of $N_{i,2}^s$ can be studied via standard Central Limit Theorems. In particular, we have

$$\mathbb{E}[N_{i,2}^{s_1}N_{j,2}^{s_2}] = \sigma^2\eta^2 TA(\frac{i}{d},\frac{j}{d})\big(\int_0^{s_1\wedge s_2} f(u)^2\mathrm{d}u + o(1)\big).$$

Therefore $N_{i,2}^s = \sigma^2\eta^2 T\big(\int_0^s f(u)\mathrm{d}\xi_1(u,\frac{i}{d}) + o(1)\big)$ where $\{\xi_1(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ is a Gaussian field process such that for any $s_1, s_2 \in [0,\tau]$, $\xi_1(s_1,\cdot) - \xi_1(s_2,\cdot) \sim \mathcal{N}(0, |s_1-s_2|\sigma_1^2)$ with

$$\sigma_1^2(x,y) = A(x,y), \qquad \forall x,y\in[0,1]. \tag{5.29}$$

For $N_{i,3}^s$, we have

$$N_{i,3}^s = -(\eta dT)\sum_{j=1}^{d}\sum_{t=0}^{\lfloor sT\rfloor-1} Z_{t,j},$$

where $Z_{t,j} := \frac{1}{dT} f(\frac{t}{T})\big(W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}\big[W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})\big]\big)\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d})$. Notice that

$$\mathbb{E}\left[Z_{t,j}\right] = \frac{1}{T} f(\frac{t}{T})\mathbb{E}\Big[\mathbb{E}\big[\big(W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}\big[W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})\big]\big)|\mathcal{F}_0^N\big]\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d})\Big] = 0.$$

Furthermore, we can check that for any $t_1 \neq t_2$ and any $j, l \in [d]$,

$$\mathbb{E}\left[Z_{t_1,j}Z_{t_2,l}\right] = \frac{1}{d^2T^2} f(\frac{t_1}{T})f(\frac{t_2}{T})\big(B(\frac{i}{d},\frac{j}{d},\frac{i}{d},\frac{l}{d}) - A(\frac{i}{d},\frac{j}{d})A(\frac{i}{d},\frac{l}{d})\big)\mathbb{E}\big[\overline{\Theta}^{d,T}(\frac{t_1}{T},\frac{j}{d})\overline{\Theta}^{d,T}(\frac{t_2}{T},\frac{l}{d})\big]\delta(t_1 - t_2).$$

Therefore, we have $\mathbb{E}\left[N_{i,3}^s\right] = 0$ and

$$\begin{aligned}
\mathrm{Var}(N_{i,3}) &= \eta^2 \sum_{j,l=1}^{d} \sum_{t=0}^{\lfloor sT \rfloor - 1} f(\frac{t}{T})^2\big(B(\frac{i}{d},\frac{j}{d},\frac{i}{d},\frac{l}{d}) - A(\frac{i}{d},\frac{j}{d})A(\frac{i}{d},\frac{l}{d})\big)\mathbb{E}\big[\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d})\overline{\Theta}^{d,T}(\frac{t}{T},\frac{l}{d})\big] \\
&\leq \left(C_2^2 + C_5\right) \|f\|_\infty^2 s\eta^2 dT \sum_{j=1}^{d} \mathbb{E}\big[|\overline{\Theta}^{d,T}(\frac{t}{T},\frac{j}{d})|^2\big] \\
&\leq C_{s,1}^2 \left(C_2^2 + C_5\right) \|f\|_\infty^2 \frac{s}{T}\big(\frac{1}{d}\sum_{j=1}^{d}\mathbb{E}\big[|\Delta\theta_j^0|^2\big] + \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}\big[|\Delta\theta_j^t - \Delta\theta_j^0|^2\big]\big) \\
&\to 0, \qquad \text{as } d, T \to \infty,
\end{aligned}$$

where the first inequality follows from Assumption 5.2.2. The last limit follows from Proposition 5.7.1 and initial conditions in Theorem 5.3.1. Therefore we have shown that $N_{i,3}^s \to 0$ in probability and we write it as $N_{i,3}^s = o(1)$ in the following calculation. Combine our approximations on $N_{i,1}^s, N_{i,2}^s, N_{i,3}^s$ for any $i \in [d]$ and $s \in [0, \tau]$, we can write the right hand side of (5.27) as

$$\begin{aligned}
\mathrm{RHS}(5.27) &= -(i+1-dx)(N_{i,1}^s + N_{i,2}^s + N_{i,3}^s) - (dx - i)(N_{i+1,1}^s + N_{i+1,2}^s + N_{i+1,3}^s) \\
&= -\eta dT\big(\int_0^{\frac{\lfloor sT \rfloor}{T}} \int_0^1 f(u)\big((i+1-dx)A(\frac{i}{d},y) + (dx-i)A(\frac{i+1}{d},y)\big)\Theta(u,y)\mathrm{d}y\mathrm{d}u + o(1)\big) \\
&\quad + \sigma^2\eta^2 T\big((i+1-dx)\int_0^s f(u)\mathrm{d}\xi_1(u,\frac{i}{d}) + (dx-i)\int_0^s f(u)\mathrm{d}\xi_1(u,\frac{i+1}{d}) + o(1)\big) \\
&= -\eta dT\big(\int_0^{\frac{\lfloor sT \rfloor}{T}} \int_0^1 f(u)A(x,y)\Theta(u,y)\mathrm{d}y\mathrm{d}u + o(1)\big) + \sigma^2\eta^2 T\big(\int_0^s f(u)\mathrm{d}\xi_1(u,x) + o(1)\big). \qquad (5.30)
\end{aligned}$$

Finally, with (5.28) and(5.30), we have for any $s \in [0, \tau]$,

$$
\begin{aligned}
&-\int_0^{\frac{\lfloor sT\rfloor}{T}}\Theta(u,x)f'(u)\mathrm{d}u+f(s)\Theta(\frac{\lfloor sT\rfloor}{T},x)-f(0)\Theta(0,x)+o(1)\\
&=-\eta dT\big(\int_0^{\frac{\lfloor sT\rfloor}{T}}\int_0^1 f(u)A(x,y)\Theta(u,y)\mathrm{d}y\mathrm{d}u+o(1)\big)+\sigma^2\eta^2T\big(\int_0^s f(u)\mathrm{d}\xi_1(u,x)+o(1)\big).
\end{aligned}
$$

Hence, letting $d,T\to\infty$ we obtain that for any bounded smooth test function f and for any $s\in[0,\tau]$, $x\in[0,1]$,

$$
\begin{aligned}
&-\int_0^{s}\Theta(u,x)f'(u)\mathrm{d}u+f(s)\Theta(s,x)-f(0)\Theta(0,x)\\
&=-\eta dT\big(\int_0^{s}\int_0^1 f(u)A(x,y)\Theta(u,y)\mathrm{d}y\mathrm{d}u+o(1)\big)+\sigma^2\eta^2T\big(\int_0^s f(u)\mathrm{d}\xi_1(u,x)+o(1)\big).
\end{aligned}
$$

■

From Theorem 5.6.2 to Theorem 5.3.1. The moderate-noise setup and the high-noise setup simply follow from integration by parts according to the consequences 2 and 3 of Theorem 5.6.2 respectively as we discussed in Section 5.6.1. The uniqueness and existence of solution in $C([0,\tau];C([0,1]))$ follows from part (a) of Theorem 5.6.5.

For the low-noise setup, since f is smooth, we know that

$$
\int_0^{(\cdot)}\Theta(u,x)f'(u)\mathrm{d}u\in C^1([0,\tau])\quad\text{and}\quad\int_0^{(\cdot)}f(u)A(x,y)\Theta(u,y)\mathrm{d}u\mathrm{d}y\in C^1([0,\tau]).
$$

Therefore according to (5.16), $f(\cdot)\Theta(\cdot,x)\in C^1([0,\tau])$ for any $x\in[0,1]$ which implies that $\Theta(\cdot,x)\in C^1([0,\tau])$ for any $x\in[0,1]$. We can then apply integration by parts to the left side of (5.16). Therefore Θ satisfies (5.4).

Since A satisfies Assumption 5.2.2, for any $\Theta_1,\Theta_2\in C([0,\tau];C([0,1]))$, we have

$$
\begin{aligned}
\alpha\sup_{x\in[0,1]}\left|\int_0^1 A(x,y)\left(\Theta_1(s,y)-\Theta_2(s,y)\right)\mathrm{d}y\right|&\le\alpha\sup_{x,y\in[0,1]}|A(x,y)|\sup_{y\in[0,1]}|\Theta_1(s,y)-\Theta_2(s,y)|\\
&\le\alpha C_2\sup_{y\in[0,1]}|\Theta_1(s,y)-\Theta_2(s,y)|.
\end{aligned}
$$

Therefore the right hand side of (5.4) is Lipschitz in $\Theta(s,\cdot)$ for any $s \in [0,\tau]$. According to the Picard-Lindelöf theorem (see, for example, [Arn92]), there exists a unique solution in $C([0,\tau];C([0,1]))$ to (5.4) with any initial condition $\Theta(0,\cdot) = \Theta_0(\cdot) \in C([0,1])$.

With the uniqueness of solution to (5.4) and Theorem 5.6.2, we know that every subsequence of $\{\Theta^{d,T}\}_{d\geq 1,T>0}$ has a further subsequence converging weakly to a unique $\Theta \in C([0,\tau];C([0,1]))$. Therefore $\{\Theta^{d,T}\}_{d\geq 1,T>0}$ converges weakly to Θ as $d,T\to\infty$. ■

5.8. Proofs for the Fluctuations

5.8.1. Tightness of $\{U^{d,T}\}_{d\geq 1,T>0}$.

Proposition 5.8.1. *Under the assumptions in Theorem 5.6.2, if there exists a uniform positive constant $C_{s,2}$ such that $\max(\gamma T^{-\frac{1}{2}}, \gamma d^{-1}, \gamma\sigma d^{-1}T^{-\frac{1}{2}}) \leq C_{s,2}$, then there exists a uniform constant $C>0$ such that for any $d\geq 1, T>0$, any $i\in[d]$ and any $0\leq t_1 < t_2 \leq \lfloor \tau T\rfloor$,*

$$\mathbb{E}\big[\big|U^{d,T}(\frac{t_2}{T},\frac{i}{d}) - U^{d,T}(\frac{t_1}{T},\frac{i}{d})\big|^4\big] \leq C\left(\frac{t_2-t_1}{T}\right)^2. \tag{5.31}$$

Proof of Proposition 5.8.1. From (5.17), we have that for any $0\leq t_1<t_2\leq N := \lfloor\tau T\rfloor$,

$$\begin{aligned}
&U^{d,T}(\frac{t_2}{T},\frac{i}{d}) - U^{d,T}(\frac{t_1}{T},\frac{i}{d})\\
=&\underbrace{-\eta\sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d}A(\frac{i}{d},\frac{j}{d})U^{d,T}(\frac{t}{T},\frac{j}{d})}_{M_{i,1}}\underbrace{-\eta\sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d}\big(W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}[W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})]\big)U^{d,T}(\frac{t}{T},\frac{j}{d})}_{M_{i,2}}\\
&\underbrace{-\eta\gamma\sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d}\big(W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}[W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})]\big)\Theta(\frac{t}{T},\frac{j}{d})}_{M_{i,3}}
\end{aligned}$$

$$\underbrace{-\gamma(\Theta(\frac{t_2}{T},\frac{i}{d}) - \Theta(\frac{t_1}{T},\frac{i}{d}) + \eta\sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d}A(\frac{i}{d},\frac{j}{d})\Theta(\frac{t}{T},\frac{j}{d}))}_{M_{i,4}} + \underbrace{\eta\gamma\sum_{t=t_1}^{t_2-1}W(\frac{t}{T},\frac{i}{d})\varepsilon^t}_{M_{i,5}}. \tag{5.32}$$

Therefore we have

$$\mathbb{E}\big[\big|U^{d,T}(\frac{t_2}{T},\frac{i}{d}) - U^{d,T}(\frac{t_1}{T},\frac{i}{d})\big|^4\big] \leq 125\big(\mathbb{E}[M_{i,1}^4] + \mathbb{E}[M_{i,2}^4] + \mathbb{E}[M_{i,3}^4] + \mathbb{E}[M_{i,4}^4] + \mathbb{E}[M_{i,5}^4]\big). \tag{5.33}$$

Next we will bound the expectation of the right hand side term by term. Many terms can be estimated using the proof of Proposition 5.7.1. Define $n_t^d := \frac{1}{d}\sum_{i=1}^d \mathbb{E}\big[\big|U^{d,T}(\frac{t}{T},\frac{i}{d})\big|^4\big]$. $\mathbb{E}[M_{i,1}^4]$ can be estimated similar to $\mathbb{E}[S_{i,1}]$ in the proof of Proposition 5.7.1 and we have

$$\begin{aligned}\mathbb{E}\big[M_{i,1}^4\big] &= \eta^4\mathbb{E}\big[\sum_{r_1,r_2,r_3,r_4=t_1}^{t_2-1}\sum_{j_1,j_2,j_3,j_4=1}^{d}\prod_{k=1}^4 A(\frac{i}{d},\frac{j_k}{d})\prod_{k=1}^4 U^{d,T}(\frac{r_k}{T},\frac{j_k}{d})\big]\\ &\le C_2^4\eta^4 d^4(t_2-t_1)^3\sum_{t=t_1}^{t_2-1} n_t^d.\end{aligned}$$

$\mathbb{E}[M_{i,2}^4]$ can be estimated similar to $\mathbb{E}[S_{i,2}]$ in the proof of Proposition 5.7.1 and we have

$$\mathbb{E}\big[N_{i,2}^4\big] \le 3\left(C_8' + C_5'^2\right)\eta^4 d^4(t_2-t_1)\sum_{t=t_1}^{t_2-1} n_t^d.$$

Similar to the how we bound $\mathbb{E}[M_{i,2}^4]$, we have

$$\begin{aligned}\mathbb{E}\big[M_{i,3}^4\big] &\le 3\big(C_8' + C_5'^2\big)(t_2-t_1)\eta^4\gamma^4\mathbb{E}\big[\sum_{t=t_1}^{t_2-1}\sum_{j_1,j_2,j_3,j_4=1}^{d}\prod_{k=1}^4\Theta(\frac{t}{T},\frac{j_k}{d})\big]\\ &\le 3\left(C_8' + C_5'^2\right)\|\Theta\|_\infty^4\gamma^4\eta^4 d^4(t_2-t_1)^2\\ &= O(\gamma^4\eta^4 d^4(t_2-t_1)^2),\end{aligned}$$

where the second inequality follows from the fact that $\|\Theta\|_\infty := \sup_{s\in[0,\tau],x\in[0,1]}|\Theta(s,x)| < \infty$. Next due to the fact that A,Θ are C^1 in all variables, we have

$$\begin{aligned}\mathbb{E}\left[M_{i,4}^2\right] &= \gamma^4\eta^4 d^4T^4\big(-\int_{\frac{t_1}{T}}^{\frac{t_2}{T}}\int_0^1 A(\frac{i}{d},y)\Theta(s,y)\mathrm{d}y\mathrm{d}s + \frac{1}{dT}\sum_{t=t_1}^{t_2-1}\sum_{j=1}^{d}A(\frac{i}{d},\frac{j}{d})\Theta(\frac{t}{T},\frac{j}{d})\big)^4\\ &= O(\gamma^4\eta^4(t_2-t_1)^4) + O(\gamma^4\eta^4 d^4(t_2-t_1)^4T^{-4}).\end{aligned}$$

$\mathbb{E}[M_{i,5}^4]$ can be estimated similar to $\mathbb{E}[S_{i,3}]$ in the proof of Proposition 5.7.1 and we have

$$\mathbb{E}\left[M_{i,5}^4\right] \le (C_5C_1 + C_2^2)\sigma^4\eta^4\gamma^4(t_2-t_1)^2 = O(\gamma^4\eta^4\sigma^4(t_2-t_1)^2).$$

Combining all the estimations and pick $t_2 = t, t_1 = 0$ in (5.33),

$$\begin{aligned} n_t^d = \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\big[\big|U^{d,T}(\frac{t}{T},\frac{i}{d})\big|^4\big] &\le 8n_0^d + \frac{8}{d}\sum_{i=1}^{d}\mathbb{E}\big[\big|U^{d,T}(\frac{t}{T},\frac{i}{d}) - U^{d,T}(0,\frac{i}{d})\big|^4\big] \\ &\le 8n_0^d + 2000\big(C_2^4 + 3(C_8 + C_5^2)\big)\eta^4 d^4 t^3 \sum_{k=0}^{t-1} n_k^d + C_{d,T}(t), \end{aligned}$$

where $C_{d,T}(t) \coloneqq O(\gamma^4\eta^4 d^4 t^2 + \gamma^4\eta^4 t^4 + \gamma^4\eta^4 d^4 t^4 T^{-4} + \gamma^4\eta^4\sigma^4 t^2)$. According to the discrete Gronwall's inequality, we have

$$n_t^d \le C_{d,T}(t) + 2000\big(C_2^4 + 3(C_8 + C_5^2)\big)\eta^4 d^4 t^3 \sum_{k=0}^{t-1} C_{d,T}(k) \exp\big(2000\eta^4 d^4\big(C_2^4 t^3 + 3(C_8 + C_5^2)\big)t^3(t-k-1)\big).$$

Since $\eta dT \le C_{s,1}$, $2000\big(C_2^4 t^3 + 3(C_8 + C_5^2)\big)\eta^4 d^4 t^3 (t-k-1) \le C_\tau C_{s,1}^4 \tau^4$ for any $0 \le k < t$, $t \le N$ and $C_{\tau,1}$ is a constant independent of d, T, σ. Since $C_{d,T}(k)$ is increasing with k, there exists a constant $C_{\tau,2}$ independent of d, T, σ such that

(5.34)

$$n_t^d \le C_{d,T}(t) + 2000\big(C_2^4 t^3 + 3(C_8 + C_5^2)\big) \exp\big(C_{\tau,1} C_{s,1}^4 \sigma^{-4}\tau^4\big)\,\eta^4 d^4 t^3 \sum_{k=0}^{t-1} C_{d,T}(k) \le C_{\tau,2} C_{d,T}(t).$$

Plug (5.34) into (5.33) and take expectations. Then, we get

$$\begin{aligned} &\mathbb{E}\big[\big|U^{d,T}(\frac{t_2}{T},\frac{i}{d}) - U^{d,T}(\frac{t_1}{T},\frac{i}{d})\big|^4\big] \\ \le& 125\big(C_2^4 + 3(C_8' + C_5'^2)\big)\eta^4 d^4 (t_2 - t_1)^3 \sum_{t=t_1}^{t_2-1} n_t^d + C_{d,T}(t_2 - t_1) \\ \le& 125 C_{\tau,2}\big(C_2^4 + 3(C_8' + C_5'^2)\big)\eta^4 d^4 (t_2 - t_1)^3 \sum_{t=t_1}^{t_2-1} C_{d,T}(t) + C_{d,T}(t_2 - t_1). \end{aligned}$$

Observe that since $\max(\eta dT, \eta\sigma T^{\frac{1}{2}}) \le C_{s,1}$, we have

$$C_{d,T}(t) = \big(\frac{t}{T}\big)^2 O(\gamma^4 T^{-2} + \gamma^4 d^{-4} + \gamma^4\sigma^4 d^{-4} T^{-2}).$$

Furthermore, note that

$$\max(\gamma T^{-\frac{1}{2}}, \gamma d^{-1}, \gamma\sigma d^{-1} T^{-\frac{1}{2}}) \le C_{s,2} \quad \text{and} \quad C_{d,T}(t) \le C_{\tau,3} C_{s,2}^4 \big(\frac{t}{T}\big)^2$$

for some positive constant $C_{\tau,3}$ independent of d, T, σ, γ. Therefore

$$\begin{aligned}\mathbb{E}\big[\big|U^{d,T}(\frac{t_2}{T},\frac{i}{d})-U^{d,T}(\frac{t_1}{T},\frac{i}{d})\big|^4\big] &\leq 125C_{\tau,2}\big(C_2^4+3(C_8'+C_5'^2)\big)\eta^4 d^4(t_2-t_1)^4 C_{\tau,3}C_{s,2}^4\tau^2+C_{\tau,3}\big(\frac{t_2-t_1}{T}\big)^2\\ &\leq 125C_{\tau,2}\big(C_2^4+3(C_8'+C_5'^2)\big)C_{s,1}^4C_{\tau,3}C_{s,2}^4\tau^2\big(\frac{t_2-t_1}{T}\big)^4+C_{\tau,3}\big(\frac{t_2-t_1}{T}\big)^2.\end{aligned}$$

Last (5.35) is proved since $\big(\frac{t_2-t_1}{T}\big)^2 \leq \tau^2$. ■

Proposition 5.8.2. *Under the assumptions in Proposition 5.8.1, for any $d, T > 0$, then there exists a constant $C > 0$ independent of d, T such that for any $d \geq 1, T > 0$, any $i, j \in [d]$ and any $0 \leq t \leq \lfloor \tau T \rfloor$,*

$$\mathbb{E}\big[\big|U^{d,T}(\frac{t}{T},\frac{i}{d})-U^{d,T}(\frac{t}{T},\frac{j}{d})\big|^4\big] \leq 8\left(C\tau^2+M^4\right)\big(\frac{i-j}{d}\big)^4. \tag{5.35}$$

Proof of Proposition 5.8.2. From (5.7) we have for any $0 \leq t_1 < t_2 \leq N \coloneqq \lfloor \tau T \rfloor$ and any $i, j \in [d]$,

$$\begin{aligned}&\big(U^{d,T}(\frac{t_2}{T},\frac{i}{d})-U^{d,T}(\frac{t_1}{T},\frac{j}{d})\big)-\big(U^{d,T}(\frac{t}{T},\frac{i}{d})-U^{d,T}(\frac{t}{T},\frac{j}{d})\big)\\ &=\underbrace{-\eta\sum_{t=t_1}^{t_2-1}\sum_{l=1}^{d}\big(A(\frac{i}{d},\frac{l}{d})-A(\frac{j}{d},\frac{l}{d})\big)U^{d,T}(\frac{t}{T},\frac{l}{d})}_{M_{i,j,1}}\underbrace{-\eta\sum_{t=t_1}^{t_2-1}\sum_{l=1}^{d}\left(x_i^tx_l^t-\mathbb{E}\left[x_i^tx_l^t\right]-x_j^tx_l^t+\mathbb{E}\left[x_j^tx_l^t\right]\right)U^{d,T}(\frac{t}{T},\frac{l}{d})}_{M_{i,j,2}}\\ &\quad\underbrace{+\alpha\gamma\int_{\frac{t_1}{T}}^{\frac{t_2}{T}}\int_0^1\big(A(\frac{i}{d},y)-A(\frac{j}{d},y)\big)\Theta(s,y)dyds-\eta\gamma\sum_{l=1}^{d}\big(A(\frac{i}{d},\frac{l}{d})-A(\frac{j}{d},\frac{l}{d})\big)\Theta(\frac{t}{T},\frac{l}{d})}_{M_{i,j,3}}\\ &\quad\underbrace{-\eta\gamma\sum_{t=t_1}^{t_2-1}\sum_{l=1}^{d}\left(x_i^tx_l^t-\mathbb{E}\left[x_i^tx_l^t\right]-x_j^tx_l^t+\mathbb{E}\left[x_j^tx_l^t\right]\right)\Theta(\frac{t}{T},\frac{l}{d})}_{M_{i,j,4}}\underbrace{+\eta\gamma\sum_{t=t_1}^{t_2-1}\left(x_i^t-x_j^t\right)\varepsilon^t}_{M_{i,j,5}}.\end{aligned}$$

Therefore if we define $\Delta^{d,T}(t,i,j) \coloneqq U^{d,T}(\frac{t}{T},\frac{i}{d})-U^{d,T}(\frac{t}{T},\frac{j}{d})$, we have

$$\begin{aligned}&\mathbb{E}\big[\big|\Delta^{d,T}(t_2,i,j)-\Delta^{d,T}(t_1,i,j)\big|^4\big]\\ &\leq 125\big(\mathbb{E}\big[M_{i,j,1}\big]+\mathbb{E}\big[M_{i,j,2}\big]+\mathbb{E}\big[M_{i,j,3}\big]+\mathbb{E}\big[M_{i,j,4}\big]+\mathbb{E}\big[M_{i,j,5}\big]\big).\end{aligned}\tag{5.36}$$

Next we estimate the right hand side of (5.36) term by term and most terms are bounded based the proof of Proposition 5.7.2. $\mathbb{E}[M_{i,j,1}]$ can be upper bounded similar to $\mathbb{E}[S_{i,j,1}]$ in the proof of

Proposition 5.7.2 and we have

$$\mathbb{E}\big[M_{i,j,1}\big] \leq C_3^4 \big(\frac{i-j}{d}\big)^4 (t_2-t_1)^3 \eta^4 d^4 \sum_{t=t_1}^{t_2-1} n_t^d = \big(\frac{i-j}{d}\big)^4 \big(\frac{t_2-t_1}{T}\big)^2 O(\eta^4 d^4 T^3 \sum_{t=t_1}^{t_2-1} n_t^d).$$

$\mathbb{E}[M_{i,j,2}]$ can be upper bounded similar to $\mathbb{E}[S_{i,j,2}]$ in the proof of Proposition 5.7.2 and we have

$$\mathbb{E}\big[M_{i,j,2}\big] \leq \big(C_9'^4 + C_6'^4\big) \big(\frac{i-j}{d}\big)^4 (t_2-t_1) \eta^4 d^4 \sum_{t=t_1}^{t_2-1} n_t^d = \big(\frac{i-j}{d}\big)^4 \big(\frac{t_2-t_1}{T}\big)^2 O(\eta^4 d^4 T^2 \sum_{t=t_1}^{t_2-1} n_t^d).$$

Since Θ is C^1 in both variables and $\partial_1 A(x,\cdot) \in C^1([0,1])$ for any $x \in [0,1]$, $\mathbb{E}[M_{i,j,3}]$ can be estimated as

$$\begin{aligned}
\mathbb{E}\big[M_{i,j,3}\big] &= \alpha^4 \gamma^4 \big(\int_{\frac{i}{d}}^{\frac{j}{d}} \int_{\frac{t_1}{T}}^{\frac{t_2}{T}} \int_0^1 \partial_1 A(z,y) \Theta(s,y) \mathrm{d}y \mathrm{d}s \mathrm{d}z - \int_{\frac{i}{d}}^{\frac{j}{d}} \frac{1}{dT} \sum_{l=1}^{d} \partial_1 A(z, \frac{l}{d}) \Theta(\frac{t}{T}, \frac{l}{d}) \mathrm{d}z \big)^4 \\
&= O\bigg(\gamma^4 \big(\frac{i-j}{d}\big)^4 \big(\frac{t_2-t_1}{T}\big)^4 \frac{1}{T^4} \bigg) + O\bigg(\gamma^4 \big(\frac{i-j}{d}\big)^4 \big(\frac{t_2-t_1}{T}\big)^4 \frac{1}{d^4} \bigg) \\
&= \big(\frac{i-j}{d}\big)^4 \big(\frac{t_2-t_1}{T}\big)^4 O\bigg(\gamma^4 T^{-4} + \gamma^4 d^{-4} \bigg).
\end{aligned}$$

Next, $\mathbb{E}[M_{i,j,4}]$ can be estimated similar to $\mathbb{E}[M_{i,j,2}]$:

$$\begin{aligned}
\mathbb{E}\big[M_{i,j,4}\big] \leq\ & C_9'^4 \eta^4 \gamma^4 \big(\frac{i-j}{d}\big)^4 \sum_{t=t_1}^{t_2-1} \sum_{l_1,l_2,l_3,l_4=1}^{d} \mathbb{E}[| \prod_{k=1}^{4} \Theta(\frac{t}{T}, \frac{l_k}{d}) |] \\
& + C_6'^4 \eta^4 \gamma^4 \big(\frac{i-j}{d}\big)^4 \sum_{t',t=t_1}^{t_2-1} \sum_{l_1,l_2,l_3,l_4=1}^{d} \mathbb{E}[| \prod_{k=1}^{2} \Theta(\frac{t}{T}, \frac{l_k}{d}) \prod_{k=3}^{4} \Theta(\frac{t'}{T}, \frac{l_k}{d}) |] \\
\leq\ & \big(C_9'^4 + C_6'^4\big) \left\|\Theta\right\|_\infty^4 \big(\frac{i-j}{d}\big)^4 \eta^4 \gamma^4 d^4 (t_2-t_1)^2 \\
=\ & \big(\frac{i-j}{d}\big)^4 \big(\frac{t_2-t_1}{T}\big)^2 O(\eta^4 \gamma^4 d^4 T^2).
\end{aligned}$$

Last $\mathbb{E}[M_{i,j,5}]$ can be upper bounded similar to $\mathbb{E}[S_{i,j,3}]$ in the proof of Proposition 5.7.2 and we have

$$\mathbb{E}\big[M_{i,j,5}\big] \leq \big(C_1 C_7^4 + C_4^4\big) \big(\frac{i-j}{d}\big)^4 \big(\frac{t_2-t_1}{T}\big)^2 \eta^4 \gamma^4 \sigma^4 T^2.$$

In the proof of Proposition 5.8.1, we have shown that $n_t^d \leq C_{\tau,2}C_{d,T}(t) \leq C_{\tau,2}C_{\tau,3}C_{s,2}^4\big(\frac{t}{T}\big)^2 = O(1)$. Therefore apply this estimation to (5.36), along with the estimations on $\mathbb{E}[M_{i,j,\cdot}]$'s and we have

$$\begin{aligned}
&\mathbb{E}\big[\big|\Delta^{d,T}(t_2,i,j) - \Delta^{d,T}(t_1,i,j)\big|^4\big] \\
&\leq \big(\frac{i-j}{d}\big)^4\big(\frac{t_2-t_1}{T}\big)^2 O\Big(\eta^4 d^4 T^3 \sum_{t=t_1}^{t_2-1} n_t^d + \gamma^4 T^{-4} + \gamma^4 d^{-4} + \eta^4\gamma^4 d^4 T^2 + \eta^4\gamma^4\sigma^4 T^2\Big) \\
&= \big(\frac{i-j}{d}\big)^4\big(\frac{t_2-t_1}{T}\big)^2 O\big(\eta^4 d^4 T^4 + \gamma^4 T^{-4} + \gamma^4 d^{-4} + \eta^4\gamma^4 d^4 T^2 + \eta^4\gamma^4\sigma^4 T^2\big).
\end{aligned}$$

Since $\max(\eta dT, \eta\sigma T^{\frac{1}{2}}) \leq C_{s,1}$ and $\max(\gamma T^{-\frac{1}{2}}, \gamma d^{-1}, \gamma\sigma d^{-1}T^{-\frac{1}{2}}) \leq C_{s,2}$,

$$O\big(\eta^4 d^4 T^4 + \gamma^4 T^{-4} + \gamma^4 d^{-4} + \eta^4\gamma^4 d^4 T^2 + \eta^4\gamma^4\sigma^4 T^2\big) = O(1),$$

and therefore there exists uniform constant C such that

$$\mathbb{E}\big[\big|\Delta^{d,T}(t_2,i,j) - \Delta^{d,T}(t_1,i,j)\big|^4\big] \leq C\big(\frac{i-j}{d}\big)^4\big(\frac{t_2-t_1}{T}\big)^2.$$

Pick $t_2 = t$ and $t_1 = 0$, we get

$$\begin{aligned}
\mathbb{E}\big[\big|U^{d,T}(\frac{t}{T},\frac{i}{d}) - U^{d,T}(\frac{t}{T},\frac{j}{d})\big|^4\big] &\leq 8\mathbb{E}\big[\big|U^{d,T}(0,\frac{i}{d}) - U^{d,T}(0,\frac{j}{d})\big|^4\big] + 8C\big(\frac{i-j}{d}\big)^4\big(\frac{t_2-t_1}{T}\big)^2 \\
&\leq 8\big(C\tau^2 + M^4\big)\big(\frac{i-j}{d}\big)^4.
\end{aligned}$$

■

Proof of Theorem 5.6.3. When $\eta = \frac{\alpha}{dT}$ and $\sigma = o(dT^{\frac{1}{2}})$, there exists a uniform constant $C_{s,1}$ such that $\max(\eta dT, \eta\sigma T^{\frac{1}{2}}) \leq C_{s,1}$. Therefore given Proposition 5.8.1 and Proposition 5.8.2, the proof of Theorem 5.6.3 is exactly the same as the proof of Theorem 5.6.1. Hence we skip the details. ■

5.8.2. Limit Identification.

Proof of Theorem 5.6.4. For any $0 \le t \le \lfloor \tau T \rfloor - 1 \coloneqq N-1$, $i \in [d]$, according to (5.3) and (5.7), we have

$$\begin{aligned}
&U^{d,T}(\frac{t+1}{T}, x) - U^{d,T}(\frac{t}{T}, x)\\
&= -\eta(i+1-dx)\sum_{j=1}^{d} W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})U^{d,T}(\frac{t}{T},\frac{j}{d}) - \eta(dx-i)\sum_{j=1}^{d} W(\frac{t}{T},\frac{i+1}{d})W(\frac{t}{T},\frac{j}{d})U^{d,T}(\frac{t}{T},\frac{j}{d})\\
&\quad - \eta\gamma(i+1-dx)\sum_{j=1}^{d}\big(W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}\big[W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})\big]\big)\Theta(\frac{t}{T},\frac{j}{d})\\
&\quad - \eta\gamma(dx-i)\sum_{j=1}^{d}\big(W(\frac{t}{T},\frac{i+1}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}\big[W(\frac{t}{T},\frac{i+1}{d})W(\frac{t}{T},\frac{j}{d})\big]\big)\Theta(\frac{t}{T},\frac{j}{d})\\
&\quad - \gamma\bigg(\Theta(\frac{t+1}{T},x) - \Theta(\frac{t}{T},x) + \eta\sum_{j=1}^{d}\big((i+1-dx)A(\frac{i}{d},\frac{j}{d}) + (dx-i)A(\frac{i+1}{d},\frac{j}{d})\big)\Theta(\frac{t}{T},\frac{j}{d})\bigg)\\
&\quad + \eta\gamma(i+1-dx)W(\frac{t}{T},\frac{i}{d})\varepsilon^t + \eta\gamma(dx-i)W(\frac{t}{T},\frac{i+1}{d})\varepsilon^t.
\end{aligned}$$

If we apply a bounded smooth test function $f : [0,\tau] \to \mathbb{R}$ on both sides, we get for any $s \in (0,\tau)$:

(5.37)
$$\sum_{t=0}^{\lfloor sT \rfloor - 1} f(\frac{t}{T})\big(U^{d,T}(\frac{t+1}{T},x) - U^{d,T}(\frac{t}{T},x)\big) = -\,(i+1-dx)\sum_{k=1}^{3} P^s_{i,k} - (dx-i)\sum_{k=1}^{3} P^s_{i+1,k} - P^s_4,$$

where for any $i \in [d]$,

$$\begin{aligned}
P^s_{i,1} &= \eta\sum_{t=0}^{\lfloor sT \rfloor - 1}\sum_{j=1}^{d} f(\frac{t}{T})W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})U^{d,T}(\frac{t}{T},\frac{j}{d}),\\
P^s_{i,2} &= -\eta\gamma\sum_{t=0}^{\lfloor sT \rfloor - 1} f(\frac{t}{T})W(\frac{t}{T},\frac{i}{d})\varepsilon^t,\\
P^s_{i,3} &= \eta\gamma\sum_{t=0}^{\lfloor sT \rfloor - 1}\sum_{j=1}^{d} f(\frac{t}{T})\big(W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d}) - \mathbb{E}\big[W(\frac{t}{T},\frac{i}{d})W(\frac{t}{T},\frac{j}{d})\big]\big)\Theta(\frac{t}{T},\frac{j}{d}),\\
P^s_4 &= \gamma\sum_{t=0}^{\lfloor sT \rfloor - 1} f(\frac{t}{T})\big(\Theta(\frac{t+1}{T},x) - \Theta(\frac{t}{T},x) + \eta\sum_{j=1}^{d}\big((i+1-dx)A(\frac{i}{d},\frac{j}{d}) + (dx-i)A(\frac{i+1}{d},\frac{j}{d})\big)\Theta(\frac{t}{T},\frac{j}{d})\big).
\end{aligned}$$

Similar to the way we derive (5.28), we can write the left hand side of (5.37) as

$$\text{LHS}(5.37) = -\int_0^{\frac{\lfloor sT\rfloor}{T}} U(u,x)f'(u)\mathrm{d}u + f(s)U(s,x) - f(0)U(0,x) + o(1). \tag{5.38}$$

Next, we look at the right hand side of (5.37). There are 4 types of terms on the right hand side of (5.37).

(1) For terms related to $P^s_{i,1}$ for some $i \in [d]$, we can deal with them akin to how we deal with $N^s_{i,1}, N^s_{2,i}$ in the proof of Theorem 5.6.2. We get

$$-(i+1-dx)P^s_{i,1} - (dx-i)P^s_{i+1,1} = -\alpha \int_0^{\frac{\lfloor sT\rfloor}{T}} \int_0^1 f(u)A(x,y)U(u,y)\mathrm{d}y\mathrm{d}u + o(1).$$

(2) For terms related to $P^s_{i,2}$ for some $i \in [d]$, we can deal with them akin to how we deal with $N^s_{i,2}$ in the proofs of Theorem 5.6.2.

$$P^s_{i,2} = \begin{cases} o(1) & \text{if } \gamma\sigma d^{-1}T^{-\frac{1}{2}} \to 0, \\ \alpha\beta \displaystyle\int_0^s f(u)\mathrm{d}\xi_2(u,\frac{i}{d}) + o(1) & \text{if } \gamma\sigma d^{-1}T^{-\frac{1}{2}} \to \beta \in (0,\infty). \end{cases}$$

where $\{\xi_2(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ is a Gaussian field process with covariance given by

$$\sigma_2^2(x,y) = A(x,y), \qquad \forall x,y \in [0,1]. \tag{5.39}$$

Therefore

(2.1) When $\gamma\sigma d^{-1}T^{-\frac{1}{2}} \to 0$, $-(i+1-dx)P^s_{i,2} - (dx-i)P^s_{i+1,2} \to 0$ in probability.

(2.2) When $\gamma\sigma d^{-1}T^{-\frac{1}{2}} \to \beta \in (0,\infty)$,

$$-(i+1-dx)P^s_{i,2} - (dx-i)P^s_{i+1,2} = \alpha\beta \int_0^{\frac{\lfloor sT\rfloor}{T}} f(u)\mathrm{d}\xi_2(u,x) + o(1).$$

(3) To study the terms related to $P_{i,3}$ for some $i \in [d]$, we first define

$$Z^{d,T}_\gamma(\frac{t}{T},\frac{j}{d}) := \frac{\gamma}{dT}\sum_{l=1}^d f(\frac{t}{T})\big(W(\frac{t}{T},\frac{j}{d})W(\frac{t}{T},\frac{l}{d}) - \mathbb{E}[W(\frac{t}{T},\frac{j}{d})W(\frac{t}{T},\frac{l}{d})]\big)\Theta(\frac{t}{T},\frac{l}{d}),$$

for any $t \leq \lfloor sT \rfloor$, $j \in [d]$. Then we can write

$$P_{i,3}^{s} = (\eta dT) \sum_{t=0}^{\lfloor sT \rfloor - 1} Z_{\gamma}^{d,T}(\frac{t}{T}, \frac{i}{d}).$$

Notice that $\mathbb{E}\big[Z_{\gamma}^{d,T}(\frac{t}{T}, \frac{i}{d})\big] = 0$ and $\{Z_{\gamma}^{d,T}(\frac{t_1}{T}, \frac{i}{d})\}_{i\in[d]}$ is independent of $\{Z_{\gamma}^{d,T}(\frac{t_2}{T}, \frac{i}{d})\}_{i\in[d]}$ when $t_1 \neq t_2$. For any $0 \leq s_1 < s_2 \leq \tau$, $i, j \in [d]$, we compute

$$\begin{aligned}
\mathbb{E}\big[P_{i,3}^{s_1} P_{j,3}^{s_2}\big] &= (\eta dT)^2\, \mathbb{E}\big[\sum_{t=0}^{\lfloor s_1 T \rfloor - 1} Z_{\gamma}^{d,T}(\frac{t}{T}, \frac{i}{d}) Z_{\gamma}^{d,T}(\frac{t}{T}, \frac{j}{d})\big] \\
&= (\eta dT)^2\, \frac{\gamma^2}{d^2 T^2} \sum_{t=0}^{\lfloor s_1 T \rfloor - 1} \sum_{l_1, l_2 = 1}^{d} f(\frac{t}{T})^2 \Theta(\frac{t}{T}, \frac{l_1}{d}) \Theta(\frac{t}{T}, \frac{l_2}{d}) \tilde{B}(\frac{i}{d}, \frac{l_1}{d}, \frac{j}{d}, \frac{l_2}{d}) \\
&= \alpha^2 (\frac{\gamma^2}{T}) \frac{1}{T} \sum_{t=0}^{\lfloor s_1 T \rfloor - 1} \big(\int_0^1 \int_0^1 f(\frac{t}{T})^2 \Theta(\frac{t}{T}, x) \Theta(\frac{t}{T}, y) \tilde{B}(\frac{i}{d}, x, \frac{j}{d}, y) \mathrm{d}x \mathrm{d}y + o(1) \big),
\end{aligned}$$

where $\tilde{B}(x_1, x_2, x_3, x_4) = B(x_1, x_2, x_3, x_4) - A(x_1, x_3) A(x_2, x_4)$ for any $x_1, x_2, x_3, x_4 \in [0,1]$. The first identity follows from independence and the last identity follows from the fact that f, Θ, B are continuous and bounded. Therefore, we have

(5.40)

$$\mathbb{E}\big[P_{i,3}^{s_1} P_{j,3}^{s_2}\big] = \gamma^2 T^{-1} \alpha^2 \int_0^{\frac{\lfloor s_1 T \rfloor}{T}} \int_0^1 \int_0^1 f(u)^2 \Theta(u, x) \Theta(u, y) \tilde{B}(\frac{i}{d}, x, \frac{j}{d}, y) \mathrm{d}x \mathrm{d}y \mathrm{d}u + o(\gamma^2 T^{-1}).$$

Now, we have the following observations.

(3.1) If $\gamma = o(T^{\frac{1}{2}})$, then $P_{i,3}^{s} \to 0$ in probability and $-(i + 1 - dx) P_{i,2}^{s} - (dx - i) P_{i+1,2}^{s} \to 0$ in probability.

(3.2) If $\gamma T^{-\frac{1}{2}} \to \zeta \in (0,\infty)$, then with (5.40), we can apply the Lindeberg-Feller CLT to $P_{i,3}$. Since

$$\begin{aligned}\mathbb{E}\big[Z_\gamma^{d,T}(\frac{t}{T},\frac{i}{d})^2\big] &= (\frac{\gamma}{dT})^2\sum_{l_1,l_2}^{d} f(\frac{t}{T})^2\Theta(\frac{t}{T},\frac{l_1}{d})\Theta(\frac{t}{T},\frac{l_2}{d})\tilde{B}(\frac{i}{d},\frac{l_1}{d},\frac{j}{d},\frac{l_2}{d})\\ &= \frac{\zeta^2}{T}f(\frac{t}{T})^2\Big(\int_0^1\int_0^1\Theta(\frac{t}{T},z_1)\Theta(\frac{t}{T},z_2)\tilde{B}(\frac{i}{d},z_1,\frac{i}{d},z_2)\mathrm{d}x\mathrm{d}y+o(1)\Big),\\ \mathrm{Var}(P_{i,3}^s) &= \zeta^2\alpha^2\int_0^{\frac{\lfloor sT\rfloor}{T}}\int_0^1\int_0^1 f(u)^2\Theta(u,z_1)\Theta(u,z_2)\tilde{B}(\frac{i}{d},x,\frac{i}{d},y)\mathrm{d}z_1\mathrm{d}z_2\mathrm{d}u+o(1).\end{aligned}$$

As $T\to\infty$, we have

$$\frac{\max_{0\le t\le\lfloor sT\rfloor-1}\mathbb{E}\big[Z_\gamma^{d,T}(\frac{t}{T},\frac{i}{d})^2\big]}{\mathrm{Var}(P_{i,3}^s)}\to 0.$$

Therefore, $P_{i,3}^s \to \alpha\zeta\int_0^s f(u)\mathrm{d}\xi_3(u,\frac{i}{d})$ in distribution with $\{\xi_3(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ being a Gaussian field process such that $\xi_3(s,x)\sim\mathcal{N}(0,\sigma_3(s,x)^2)$ and

$$\sigma_3(s,x)^2 \coloneqq \int_0^1\int_0^1\Theta(s,z_1)\Theta(s,z_2)\tilde{B}(x,z_1,x,z_2)\mathrm{d}z_1\mathrm{d}z_2. \tag{5.41}$$

Furthermore, using the facts that $A\in C([0,1]^2)$ and $B(\cdot,x,\cdot,y)\in C([0,1]^2)$ we have for any $z_1\in[\frac{i}{d},\frac{i+1}{d})$, $z_2\in[\frac{j}{d},\frac{j+1}{d})$,

$$\begin{aligned}&(i+1-dz_1)(j+1-dz_2)\tilde{B}(\frac{i}{d},x,\frac{j}{d},y)+(i+1-dz_1)(dz_2-j)\tilde{B}(\frac{i}{d},x,\frac{j+1}{d},y)\\ &+(dz_1-i)(j+1-dz_2)\tilde{B}(\frac{i+1}{d},x,\frac{j+1}{d},y)+(dz_1-i)(dz_2-j)\tilde{B}(\frac{i+1}{d},x,\frac{j+1}{d},y)\\ &=\tilde{B}(z_1,x,z_2,y)+o(1)\end{aligned}$$

and, that any $x\in[\frac{i}{d},\frac{i+1}{d})$:

$$-(i+1-dx)P_{i,3}^s-(dx-i)P_{i+1,3}^s\xrightarrow{\mathrm{d}}\alpha\zeta\int_0^s f(u)\mathrm{d}\xi_3(u,x). \tag{5.42}$$

Also notice that $\mathbb{E}[P_{i,3}^sP_{j,2}^s]=0$ for any $s\in[0,\tau]$ and $i,j\in[d]$, and so we have $\mathbb{E}[\xi_2(s,x)\xi_3(s,y)]=0$ for any $s\in[0,\tau]$ and $x,y\in[0,1]$. Therefore the Gaussian field process $\{\xi_3(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ is independent of the Gaussian field process $\{\xi_2(s,x)\}_{s\in[0,\tau],x\in[0,1]}$.

4. For the term P_4^s, since $\Theta \in C^1([0,\tau];C^1([0,1]))$ and $\partial_s\Theta(s,x) = -\alpha\int_0^1 A(x,y)\Theta(s,y)\mathrm{d}y$, we have

$$\begin{aligned}
&\Theta(\frac{t+1}{T},x) - \Theta(\frac{t}{T},x)\\
=& -\alpha\int_{\frac{t}{T}}^{\frac{t+1}{T}}\int_0^1 A(x,y)\Theta(s,y)\mathrm{d}y\mathrm{d}s = \frac{\alpha}{dT}\sum_{j=1}^{d} A(x,\frac{j}{d})\Theta(\frac{t}{T},\frac{j}{d}) + O(d^{-1}T^{-1}+T^{-2})\\
=& \eta\sum_{j=1}^{d}\big((i+1-dx)A(\frac{i}{d},\frac{j}{d}) + (dx-i)A(\frac{i+1}{d},\frac{j}{d}) + O(d^{-1})\big)\Theta(\frac{t}{T},\frac{j}{d}) + O(d^{-1}T^{-1}+T^{-2}).
\end{aligned}$$

Since $\eta dT \to \alpha$, we have

$$P_4^s = \gamma\sum_{t=1}^{\lfloor sT\rfloor -1} f(\frac{t}{T})O(\eta + d^{-1}T^{-1}+T^{-2}) = \frac{1}{T}\sum_{t=1}^{\lfloor sT\rfloor -1} f(\frac{t}{T})O(d^{-1}\gamma + T^{-1}\gamma).$$

Therefore, combining our approximations, we get

$$\begin{aligned}
\text{RHS}(5.37) =& -\alpha\int_0^{\frac{\lfloor sT\rfloor}{T}}\int_0^1 f(u)A(x,y)U(u,y)\mathrm{d}y\mathrm{d}u + \alpha\beta\int_0^s f(u)\mathrm{d}\xi_2(u,x)1_{\{\gamma\sigma d^{-1}T^{-\frac{1}{2}}\to\beta\}}\\
&+\alpha\zeta\int_0^s f(u)\mathrm{d}\xi_3(u,x)1_{\{\gamma T^{-\frac{1}{2}}\to\zeta\}} + \int_0^{\frac{\lfloor sT\rfloor}{T}} f(u)\mathrm{d}uO(\gamma d^{-1}+\gamma T^{-1}) + o(1),
\end{aligned} \tag{5.43}$$

where $\{\xi_2(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ and $\{\xi_3(s,x)\}_{s\in[0,\tau],x\in[0,1]}$ are two independent Gaussian field processes adapted to the same filtration $\{\mathcal{F}_s\}_{s\in[0,\tau]}$ with covariances given by (5.39) and (5.41) respectively. $\beta,\zeta \in [0,\infty)$ because $\max(\gamma T^{-\frac{1}{2}},\gamma d^{-1},\gamma\sigma d^{-1}T^{-\frac{1}{2}}) \le C_{s,2}$. Combine (5.38) and (5.43), and we have for any $s\in[0,\tau]$,

$$\begin{aligned}
&-\int_0^s U(u,x)f'(u)\mathrm{d}u + f(s)U(s,x) - f(0)U(0,x)\\
=& -\alpha\int_0^s\int_0^1 f(u)A(x,y)U(u,y)\mathrm{d}y\mathrm{d}u + \alpha\beta\int_0^s f(u)\mathrm{d}\xi_2(u,x)1_{\{\gamma\sigma d^{-1}T^{-\frac{1}{2}}\to\beta\}}\\
&+\alpha\zeta\int_0^s f(u)\mathrm{d}\xi_3(u,x)1_{\{\gamma T^{-\frac{1}{2}}\to\zeta\}} + O(\gamma d^{-1}+\gamma T^{-1}) + o(1),
\end{aligned}$$

giving us the desired result. ■

5.9. Proofs for the Existence and Uniqueness of the SDE

Proof of part (a) of Theorem 5.6.5. **Existence:** We prove existence based on the Picard iteration argument. Let $U_0(s,x) = U(0,x)$ for all $s \in [0,\tau]$. The Picard iteration is given by

$$U_k(s,x) = U(0,x) - \alpha \int_0^s \int_0^1 A(x,y) U_{k-1}(u,y) \mathrm{d}y \mathrm{d}u + \alpha\beta \int_0^s \mathrm{d}\xi_2(u,x) + \alpha\zeta \int_0^s \mathrm{d}\xi_3(u,x). \tag{5.44}$$

According to the definition of ξ_2, ξ_3, if $U(0,\cdot) \in C([0,1])$, then $U_k(s,\cdot) \in C([0,1])$ for all $s \in [0,\tau]$ and $k \geq 0$. Next we show that $U_k \in C([0,\tau]; C([0,1]))$. For any $K > 0$, define the stopping time

$$\tau_{K,N} \coloneqq \min(\tau, \inf\{s \in [0,\tau] : \max_{0 \leq k \leq K} \|U_k(s,\cdot)\|_\infty \geq N\}).$$

It is easy to see that $\tau_{K,N} \to \tau$ almost surely as $N \to \infty$. Define $U_k^N(s,x) \coloneqq U_k(s \wedge \tau_{K,N}, x)$. We have

$$\begin{aligned} U_k^N(s,x) = U(0,x) &- \alpha \int_0^s \int_0^1 A(x,y) U_{k-1}(u \wedge \tau_N, y) 1_{(0,\tau_N)}(u) \mathrm{d}y \mathrm{d}u \\ &+ \alpha\beta \int_0^s 1_{(0,\tau_N)}(u) \mathrm{d}\xi_2(u \wedge \tau_N, x) + \alpha\zeta \int_0^s 1_{(0,\tau_N)}(u) \mathrm{d}\xi_3(u \wedge \tau_N, x). \end{aligned}$$

Therefore under Assumption 5.2.1 we obtain

$$\begin{aligned} \max_{0 \leq k \leq K} \left\|U_k^N(s,\cdot)\right\|_\infty \leq \|U(0,\cdot)\|_\infty &+ C_2\alpha \int_0^s \int_0^1 \max_{0 \leq k \leq K} \left\|U_k^N(u,\cdot)\right\|_\infty \mathrm{d}y \mathrm{d}u \\ &+ \alpha\beta \sup_{x \in [0,1]} \Big| \int_0^s 1_{(0,\tau_N)}(u) \mathrm{d}\xi_2(u,x) \Big| + \alpha\zeta \sup_{x \in [0,1]} \Big| \int_0^s 1_{(0,\tau_N)}(u) \mathrm{d}\xi_3(u,x) \Big|. \end{aligned}$$

Taking the expectation, we get

$$\begin{aligned} \mathbb{E}\big[\max_{0 \leq k \leq K} \sup_{s \in [0,\tau]} \left\|U_k^N(s,\cdot)\right\|_\infty\big] \leq \mathbb{E}\big[\|U(0,\cdot)\|_\infty\big] &+ C_2\alpha \int_0^\tau \mathbb{E}\big[\max_{0 \leq k \leq K} \left\|U_k^N(u,\cdot)\right\|_\infty\big] \mathrm{d}u \\ &+ \alpha\beta \mathbb{E}\big[\sup_{x \in [0,1]} \big(\int_0^\tau 1_{(0,\tau_{K,N})}(u) \mathrm{d}\xi_2(u,x)\big)^2\big]^{\frac{1}{2}} \\ &+ \alpha\zeta \mathbb{E}\big[\sup_{x \in [0,1]} \big(\int_0^\tau 1_{(0,\tau_{K,N})}(u) \mathrm{d}\xi_3(u,x)\big)^2\big]^{\frac{1}{2}}. \end{aligned}$$

According to (5.39) and (5.41), under Assumption 5.2.1, we have

$$\begin{aligned}
&\mathbb{E}\big[\max_{0\le k\le K}\sup_{s\in[0,\tau]}\left\|U_k^N(s,\cdot)\right\|_\infty\big]\\
\le&\mathbb{E}\big[\left\|U(0,\cdot)\right\|_\infty\big]+C_2\alpha\int_0^\tau\mathbb{E}\big[\max_{0\le k\le K}\left\|U_k^N(u,\cdot)\right\|_\infty\big]\mathrm{d}u\\
&+\alpha\beta\mathbb{E}\big[\sup_{x\in[0,1]}\int_0^\tau A(x,x)\mathrm{d}(u\wedge\tau_{K,N})\big]^{\frac{1}{2}}\\
&+\alpha\zeta\mathbb{E}\big[\sup_{x\in[0,1]}\int_0^\tau\int_0^1\int_0^1\Theta(u,z_1)\Theta(u,z_2)\tilde{B}(x,z_1,x,z_2)\mathrm{d}z_1\mathrm{d}z_2\mathrm{d}(u\wedge\tau_{K,N})\big]^{\frac{1}{2}}\\
\le&\mathbb{E}\left[\left\|U(0,\cdot)\right\|_\infty\right]+C_2\alpha\int_0^\tau\mathbb{E}\big[\max_{0\le k\le K}\sup_{u\in[0,s]}\left\|U_k^N(u,\cdot)\right\|_\infty\big]\mathrm{d}s\\
&+C_2^{\frac{1}{2}}\alpha\beta\tau^{\frac{1}{2}}+(C_2^2+C_5)^{\frac{1}{2}}\alpha\zeta\left\|\Theta\right\|_\infty\tau^{\frac{1}{2}},
\end{aligned}$$

where $\tilde{B}(x,z_1,x,z_2)=B(x,z_1,x,z_2)-A(x,z_1)A(x,z_2)$. With Gronwall's inequality, we then get

$$\mathbb{E}\big[\max_{0\le k\le K}\sup_{s\in[0,\tau]}\left\|U_k^N(s,\cdot)\right\|_\infty\big]\le\left(\mathbb{E}\big[\left\|U(0,\cdot)\right\|_\infty\big]+C_2^{\frac{1}{2}}\alpha\beta\tau^{\frac{1}{2}}+(C_5+C_2^2)^{\frac{1}{2}}\alpha\zeta\left\|\Theta\right\|_\infty\tau^{\frac{1}{2}}\right)\exp(C_2\alpha\tau).$$

Since K is arbitrary, we can push $N\to\infty$ and according to Monotone convergence theorem, we prove for any k,

$$\mathbb{E}\big[\sup_{s\in[0,\tau]}\left\|U_k^N(s,\cdot)\right\|_\infty\big]\le\left(\mathbb{E}\big[\left\|U(0,\cdot)\right\|_\infty\big]+C_2^{\frac{1}{2}}\alpha\beta\tau^{\frac{1}{2}}+(C_5+C_2^2)^{\frac{1}{2}}\alpha\zeta\left\|\Theta\right\|_\infty\tau^{\frac{1}{2}}\right)\exp(C_2\alpha\tau).$$

Therefore for any k, $U_k\in C\left([0,\tau];C([0,1])\right)$. Next we show that $\{U_k\}_{k=1}^\infty$ converges in the space $C([0,\tau];C([0,1]))$. From (5.44), we have for any $k\ge 1$ and $s\in[0,\tau]$,

$$\begin{aligned}
\left\|U_{k+1}(s,\cdot)-U_k(s,\cdot)\right\|_\infty&=\alpha\left\|\int_0^s\int_0^1 A(\cdot,y)\left(U_k(u,y)-U_{k-1}(u,y)\right)\mathrm{d}y\mathrm{d}u\right\|_\infty\\
&\le C_2\alpha s\int_0^s\left\|U_k(u,\cdot)-U_{k-1}(u,\cdot)\right\|_\infty\mathrm{d}u,
\end{aligned}$$

and

$$\begin{aligned}
\left\|U_1(s,\cdot)-U_0(s,\cdot)\right\|_\infty\le\alpha&\left\|\int_0^s\int_0^1 A(\cdot,y)U_0(y)\mathrm{d}y\mathrm{d}u\right\|_\infty\\
&+\alpha\beta\sup_{x\in[0,1]}\big|\int_0^s\mathrm{d}\xi_2(u,x)\big|+\alpha\zeta\sup_{x\in[0,1]}\big|\int_0^s\mathrm{d}\xi_3(u,x)\big|.
\end{aligned}$$

Hence, for any $s \in [0,\tau]$,

$$\mathbb{E}\big[\sup_{r\in[0,s]} \|U_1(r,\cdot)-U_0(r,\cdot)\|_\infty\big] \le (1+C_2\alpha s)\mathbb{E}\big[\|U(0,\cdot)\|_\infty\big] + C_2^{\frac{1}{2}}\alpha\beta s^{\frac{1}{2}} + (C_5+C_2^2)^{\frac{1}{2}}\alpha\zeta\,\|\Theta\|_\infty\, s^{\frac{1}{2}}$$

$$:= C(s) \le C(\tau) < \infty,$$

and

$$\mathbb{E}\big[\sup_{r\in[0,s]} \|U_{k+1}(r,\cdot)-U_k(r,\cdot)\|_\infty\big] \le C_2\alpha\int_0^s \mathbb{E}\big[\sup_{u\in[0,r]}\|U_k(u,\cdot)-U_{k-1}(u,\cdot)\|_\infty\big]\mathrm{d}u.$$

By induction we get for any $k \ge 1$,

$$\mathbb{E}\big[\sup_{r\in[0,s]} \|U_{k+1}(r,\cdot)-U_k(r,\cdot)\|_\infty\big] \le \frac{C(\tau)\,(C_2\alpha s)^k}{k!}.$$

Therefore according to Markov's inequality,

$$\mathbb{P}\big(\|U_{k+1}(r,\cdot)-U_k(r,\cdot)\|_\infty > \frac{1}{2^k}\big) \le \frac{C(\tau)\,(2C_2 s)^k}{k!}.$$

Let Ω be the path space on which the Gaussian field processes ξ_2 and ξ_3 are defined. According to Borel-Cantelli Lemma we have for almost every $\omega\in\Omega$, there exists $k(\omega)$ such that $\sup_{r\in[0,s]}\|U_{k+1}(r,\cdot)-U_k(r,\cdot)\|_\infty \le \frac{1}{2^k}$ for any $k\ge k(\omega)$. Therefore with probability 1, $\{U_k(\cdot,\cdot)\}_{k\ge1}$ converges in $C([0,\tau];C([0,1]))$ with limit $U(\cdot,\cdot)\in C([0,\tau];C([0,1]))$. Furthermore we can check that U satisfies SDE (5.8). Therefore existence is proved.

Uniqueness: Suppose that there exist two solutions, $U,\overline{U}$ to SDE (5.8), from (5.45), we have for any $s\in[0,\tau]$, $x\in[0,1]$,

$$\overline{U}(s,x)-U(s,x) = -\alpha\int_0^s\int_0^1 A(x,y)\left(\overline{U}(u,y)-U(u,y)\right)\mathrm{d}y\mathrm{d}u,$$

which implies that

$$\mathbb{E}\big[\sup_{s\in[0,\tau]}\left\|\overline{U}(s,\cdot)-U(s,\cdot)\right\|_\infty\big] \le C_2\alpha\mathbb{E}\big[\int_0^\tau \sup_{r\in[0,s]}\left\|\overline{U}(r,\cdot)-U(r,\cdot)\right\|_\infty \mathrm{d}s\big].$$

By Gronwall's inequality we have

$$\mathbb{E}\big[\sup_{s\in[0,\tau]}\big\|\overline{U}(s,\cdot)-U(s,\cdot)\big\|_\infty\big]\le\mathbb{E}\big[\big\|\overline{U}(0,\cdot)-U(0,\cdot)\big\|_\infty\big]\exp\left(C_2\alpha\tau\right)=0.$$

Therefore there is a unique solution to (5.8) in $C([0,\tau];C([0,1]))$. ■

Proof of part (b) of Theorem 5.6.5. **Stability:** Suppose there exist solutions to the SDE (5.8), then the integral form solution can be written as $\forall\ s\in[0,\tau]$

$$U(s,x)=U(0,x)-\alpha\int_0^s\int_0^1 A(x,y)U(u,y)\mathrm{d}y\mathrm{d}u+\alpha\beta\int_0^s\mathrm{d}\xi_2(u,x)+\alpha\zeta\int_0^s\mathrm{d}\xi_3(u,x). \tag{5.45}$$

Define the stopping time $\tau_N:=\min(\tau,\inf\{s\in[0,\tau]:\|U(s,\cdot)\|_{L^2([0,1])}\ge N\})$. It is easy to see that $\tau_N\to\tau$ almost surely as $N\to\infty$. Define $U^N(s,x):=U(s\wedge\tau_N,x)$. We have

$$\begin{aligned}U^N(s,x)=U(0,x)-\alpha\int_0^s\int_0^1 A(x,y)U(u\wedge\tau_N,y)1_{(0,\tau_N)}(u)\mathrm{d}y\mathrm{d}u\\+\alpha\beta\int_0^s 1_{(0,\tau_N)}(u)\mathrm{d}\xi_2(u\wedge\tau_N,x)+\alpha\zeta\int_0^s 1_{(0,\tau_N)}(u)\mathrm{d}\xi_3(u\wedge\tau_N,x).\end{aligned}$$

Therefore we obtain an estimation of $U^N(s,\cdot)$ in the space of $L^2([0,1])$,

$$\begin{aligned}\int_0^1\big|U^N(s,x)\big|^2\mathrm{d}x\le 4\left\|U(0,\cdot)\right\|^2_{L^2([0,1])}+4\alpha^2 s\int_0^s\int_0^1\big(\int_0^1 A(x,y)U^N(u,y)\mathrm{d}y\big)^2\mathrm{d}x\mathrm{d}u\\+4\alpha^2\beta^2\int_0^1\big|\int_0^s 1_{(0,\tau_N)}(u)\mathrm{d}\xi_2(u,x)\big|^2\mathrm{d}x+4\alpha^2\zeta^2\int_0^1\big|\int_0^s 1_{(0,\tau_N)}(u)\mathrm{d}\xi_3(u,x)\big|^2\mathrm{d}x.\end{aligned}$$

Under Assumption 5.2.2, we have that for the integral operator

$$\mathcal{A}:g\in L^2([0,1])\mapsto\int_0^1 A(x,y)g(y)\mathrm{d}y\in L^2([0,1]),$$

$$\|\mathcal{A}\|_{op}:=\sup_{\|g\|_{L^2([0,1])}=1}\int_0^1\big(\int_0^1 A(x,y)g(y)\mathrm{d}y\big)^2\mathrm{d}x\le C_2.$$

Therefore we get

$$\begin{aligned}\big\|U^N(s,\cdot)\big\|^2_{L^2([0,1])}\le 4\left\|U(0,\cdot)\right\|^2_{L^2([0,1])}+4C_2^2\alpha^2 s\int_0^s\big\|U^N(u,\cdot)\big\|^2_{L^2([0,1])}\mathrm{d}s\\+4\alpha^2\beta^2\int_0^1\Big(\int_0^s 1_{(0,\tau_N)}(u)\mathrm{d}\xi_2(u,x)\Big)^2\mathrm{d}x+4\alpha^2\zeta^2\int_0^1\Big(\int_0^s 1_{(0,\tau_N)}(u)\mathrm{d}\xi_3(u,x)\Big)^2\mathrm{d}x.\end{aligned}$$

Taking the supreme over $[0,\tau]$ and taking the expectation, we get

$$\begin{aligned}\mathbb{E}\big[\sup_{s\in[0,\tau]}\big\|U^N(s,\cdot)\big\|^2_{L^2([0,1])}\big] &\leq 4\mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big]+4C_2^2\alpha^2\tau\int_0^\tau\mathbb{E}\big[\,\big\|U^N(s,\cdot)\big\|^2_{L^2([0,1])}\big]\mathrm{d}s\\ &\quad+4\alpha^2\beta^2\sup_{s\in[0,\tau]}\int_0^1\mathbb{E}\big[\big(\int_0^s 1_{(0,\tau_N)}(u)\mathrm{d}\xi_2(u,x)\big)^2\big]\mathrm{d}x\\ &\quad 4\alpha^2\zeta^2\sup_{s\in[0,\tau]}\int_0^1\mathbb{E}\big[\big(\int_0^s 1_{(0,\tau_N)}(u)\mathrm{d}\xi_3(u,x)\big)^2\big]\mathrm{d}x.\end{aligned}$$

According to (5.39) and (5.41), we have

$$\begin{aligned}&\mathbb{E}\left[\langle\mathrm{d}\xi_2(s,x),\mathrm{d}\xi_2(s,x)\rangle\right]=A(x,x)\mathrm{d}s,\\ &\mathbb{E}\left[\langle\mathrm{d}\xi_3(s,x),\mathrm{d}\xi_3(s,x)\rangle\right]=\int_0^1\int_0^1\Theta(s,z_1)\Theta(s,z_2)\tilde{B}(x,z_1,x,z_2)\mathrm{d}z_1\mathrm{d}z_2\mathrm{d}s,\end{aligned}$$

where $\tilde{B}(x,z_1,x,z_2)=B(x,x_1,x,z_2)-A(x,z_1)A(x,z_2)$. Under Assumption 5.2.2, we get

$$\begin{aligned}\mathbb{E}\big[\sup_{s\in[0,\tau]}\big\|U^N(s,\cdot)\big\|^2_{L^2([0,1])}\big] &\leq 4\mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big]+4C_2^2\alpha^2\tau\int_0^\tau\mathbb{E}\big[\,\big\|U^N(s,\cdot)\big\|^2_{L^2([0,1])}\big]\mathrm{d}s\\ &\quad+4\alpha^2\beta^2\int_0^1\int_0^\tau A(x,x)\mathrm{d}(s\wedge\tau_N)\mathrm{d}x\\ &\quad+4\alpha^2\zeta^2\int_0^1\int_0^\tau\int_0^1\int_0^1\Theta(s,z_1)\Theta(s,z_2)\tilde{B}(x,z_1,x,z_2)\mathrm{d}z_1\mathrm{d}z_2 d(s\wedge\tau_N)\mathrm{d}x\\ &\leq 4\mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big]+4C_2^2\alpha^2\tau\int_0^\tau\mathbb{E}\big[\,\big\|U^N(s,\cdot)\big\|^2_{L^2([0,1])}\big]\mathrm{d}s\\ &\quad+4C_2\alpha^2\beta^2\tau+4\left(C_5+C_2^2\right)\alpha^2\zeta^2\int_0^\tau\|\Theta(s,\cdot)\|^2_{L^2([0,1])}\mathrm{d}s\\ &\leq 4\mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big]+4C_2^2\alpha^2\tau\int_0^\tau\mathbb{E}\big[\sup_{r\in[0,s]}\big\|U^N(r,\cdot)\big\|^2_{L^2([0,1])}\big]\mathrm{d}s\\ &\quad+4C_2\alpha^2\beta^2\tau+4(C_5+C_2^2)\alpha^2\zeta^2\int_0^\tau\|\Theta(s,\cdot)\|^2_{L^2([0,1])}\mathrm{d}s.\end{aligned}$$

With Gronwall's inequality we get

$$\begin{aligned}\mathbb{E}\big[\sup_{s\in[0,\tau]}\big\|U^N(s,\cdot)\big\|^2_{L^2([0,1])}\big] &\leq 4\bigg(\mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big]+C_2\alpha^2\beta^2\tau\\ &\quad+(C_5+C_2^2)\alpha^2\zeta^2\int_0^\tau\|\Theta(s,\cdot)\|^2_{L^2([0,1])}\mathrm{d}s\bigg)\times\exp\left(4C_2^2\alpha^2\tau^2\right).\end{aligned}$$

Last letting $N\to\infty$ and according to Monotone convergence theorem, we prove (5.18).

Existence: As before, we use the Picard iteration argument to show existence. Let $U_0(s,x) = U(0,x)$ for all $s \in [0,\tau]$. The Picard iteration is given by

$$(5.46)\quad U_k(s,x) = U(0,x) - \alpha \int_0^s \int_0^1 A(x,y)U_{k-1}(u,y)\mathrm{d}y\mathrm{d}u + \alpha\beta \int_0^s \mathrm{d}\xi_2(u,x) + \alpha\zeta \int_0^s \mathrm{d}\xi_3(u,x).$$

With similar argument as in the stability part, we get for any $K \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\big[\max_{1\le k\le K} \|U_k(s,\cdot)\|^2_{L^2([0,1])}\big] &\le 4C_2^2\alpha^2 s \int_0^s \max_{1\le k\le K} \mathbb{E}\big[\,\|U_k(u,\cdot)\|^2_{L^2([0,1])}\big]\mathrm{d}u \\ &+ 4\big(\mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big] + C_2\alpha^2\beta^2 s + (C_5 + C_2^2)\alpha^2\zeta^2 \int_0^s \|\Theta(u,\cdot)\|^2_{L^2([0,1])}\,\mathrm{d}u\big).\end{aligned}$$

Gronwall's inequality implies that for any $k \in \mathbb{N}$,

$$\begin{aligned}&\mathbb{E}\big[\,\|U_k(s,\cdot)\|^2_{L^2([0,1])}\big] \\ &\le 4\bigg(\mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big] + C_2\alpha^2\beta^2 s + (C_5 + C_2^2)\alpha^2\zeta^2 \int_0^s \|\Theta(u,\cdot)\|^2_{L^2([0,1])}\,\mathrm{d}u\bigg) \exp\left(4C_2^2\alpha^2 s^2\right).\end{aligned}$$

Therefore for any k and $s \in [0,\tau]$, $U(s,\cdot) \in L^2([0,1])$. Next we show that $\{U_k\}_{k=1}^\infty$ converges in the space $C([0,\tau]; L^2([0,1]))$. From (5.46), we have for any $k \ge 1$ and $s \in [0,\tau]$,

$$\begin{aligned}\|U_{k+1}(s,\cdot) - U_k(s,\cdot)\|^2_{L^2([0,1])} &= \alpha^2 \left\| \int_0^s \int_0^1 A(x,y)\left(U_k(u,y) - U_{k-1}(u,y)\right)\mathrm{d}y\mathrm{d}u \right\|^2_{L^2([0,1])} \\ &\le C_2^2\alpha^2 s \int_0^s \|U_k(u,\cdot) - U_{k-1}(u,\cdot)\|^2_{L^2([0,1])}\,\mathrm{d}u,\end{aligned}$$

and

$$\begin{aligned}\|U_1(s,\cdot) - U_0(s,\cdot)\|^2_{L^2([0,1])} \le 3\alpha^2 \int_0^1 \big(\int_0^s \int_0^1 A(x,y)U_0(y)\mathrm{d}y\mathrm{d}u\big)^2\mathrm{d}x + 3\alpha^2\beta^2 \int_0^1 \big(\int_0^s \mathrm{d}\xi_2(u,x)\big)^2\mathrm{d}x \\ + 3\alpha^2\zeta^2 \int_0^1 \big(\int_0^s \mathrm{d}\xi_3(u,x)\big)^2\mathrm{d}x.\end{aligned}$$

Therefore, for any $s \in [0,\tau]$, we have that

$$\begin{aligned}\mathbb{E}\big[\sup_{r\in[0,s]} \|U_1(r,\cdot) - U_0(r,\cdot)\|^2_{L^2([0,1])}\big] &\le 3C_2^2\alpha^2 s^2 \mathbb{E}\big[\,\|U(0,\cdot)\|^2_{L^2([0,1])}\big] + 3C_2\alpha^2\beta^2 s \\ &\quad + 3\left(C_5 + C_2^2\right)\alpha^2\zeta^2 \int_0^s \|\Theta(u,\cdot)\|^2_{L^2([0,1])}\,\mathrm{d}u \\ &:= C(s) \le C(\tau) < \infty,\end{aligned}$$

and

$$\mathbb{E}\big[\sup_{r\in[0,s]} \|U_{k+1}(r,\cdot) - U_k(r,\cdot)\|^2_{L^2([0,1])}\big] \le C_2^2\alpha^2 s \int_0^s \mathbb{E}\big[\sup_{u\in[0,r]} \|U_k(u,\cdot) - U_{k-1}(u,\cdot)\|^2_{L^2([0,1])}\big]\mathrm{d}u.$$

Hence, by induction, we get that for any $k \ge 1$,

$$\mathbb{E}\big[\sup_{r\in[0,s]} \|U_{k+1}(r,\cdot) - U_k(r,\cdot)\|^2_{L^2([0,1])}\big] \le \frac{C(\tau)\left(C_2^2\alpha^2\tau s\right)^k}{k!}.$$

Therefore according to Markov inequality,

$$\mathbb{P}\big(\,\|U_{k+1}(r,\cdot) - U_k(r,\cdot)\|^2_{L^2([0,1])} > \frac{1}{2^k}\big) \le \frac{C(\tau)\left(2C_2^2\alpha^2\tau s\right)^k}{k!}.$$

Let Ω be the path space on which the Gaussian field processes ξ_2 and ξ_3 are defined. According to Borel-Cantelli Lemma, we have have that for almost every $\omega \in \Omega$, there exists $k(\omega)$ such that $\sup_{r\in[0,s]} \|U_{k+1}(r,\cdot) - U_k(r,\cdot)\|^2_{L^2([0,1])} \le \frac{1}{2^k}$ for any $k \ge k(\omega)$. Therefore with probability 1, $\{U_k(\cdot,\cdot)\}_{k\ge1}$ converges in $C([0,\tau]; L^2([0,1]))$ with limit $U(\cdot,\cdot) \in C([0,\tau]; L^2([0,1]))$. Furthermore we can check that U satisfies SDE (5.8). Therefore the solution to (5.8) exists.

Uniqueness: Suppose that there exist two solutions, $U, \overline{U}$ to SDE (5.8), from (5.45), we have for any $s \in [0,\tau]$, $x \in [0,1]$,

$$\overline{U}(s,x) - U(s,x) = -\alpha \int_0^s \int_0^1 A(x,y)\left(\overline{U}(u,y) - U(u,y)\right)\mathrm{d}y\mathrm{d}u,$$

which implies that

$$\mathbb{E}\big[\sup_{s\in[0,\tau]} \left\|\overline{U}(s,\cdot) - U(s,\cdot)\right\|_{L^2([0,1])}\big] \le C_2^2\alpha^2\tau\mathbb{E}\big[\int_0^\tau \sup_{r\in[0,s]} \left\|\overline{U}(r,\cdot) - U(r,\cdot)\right\|_{L^2([0,1])}\mathrm{d}s\big].$$

By Gronwall's inequality we have

$$\mathbb{E}\big[\sup_{s\in[0,\tau]}\big\|\overline{U}(s,\cdot)-U(s,\cdot)\big\|_{L^2([0,1])}\big]\le\mathbb{E}\big[\big\|\overline{U}(0,\cdot)-U(0,\cdot)\big\|^2_{L^2([0,1])}\big]\exp\big(C_2^2\alpha^2\tau^2\big)=0.$$

Therefore these is a unique solution to (5.8) in $C([0,\tau];L^2([0,1]))$. ■

5.10. Proofs for Applications from Section 5.5

Proof of Proposition 5.5.1. First, according to the definition of $\mathsf{MSE}^{d,T}$, we have

$$\begin{aligned}\mathsf{MSE}^{d,T}(s)&=\frac{1}{d}\sum_{i=1}^{d}\overline{\Theta}^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})^2=\int_0^1\overline{\Theta}^{d,T}(s,y)^2\mathrm{d}y+o(1)\\&\to\mathsf{MSE}(s):=\int_0^1\Theta(s,y)^2\mathrm{d}y\quad\text{as }d,T\to\infty,\end{aligned}$$

where the second identity follows from the fact that $\overline{\Theta}^{d,T}\in C([0,\tau];C([0,1]))$. The last step follows from Theorem 5.3.1. Next for the predictive error, according to the definition of $\mathsf{PE}^{d,T}$,

$$\begin{aligned}\mathsf{PE}^{d,T}(s)&=\frac{1}{d^2}\sum_{i,j=1}^{d}A(\frac{i}{d},\frac{j}{d})\overline{\Theta}^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\overline{\Theta}^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{j}{d})\\&=\int_0^1\int_0^1A(x,y)\overline{\Theta}^{d,T}(s,x)\overline{\Theta}^{d,T}(s,y)\mathrm{d}x\mathrm{d}y+o(1)\\&\to\mathsf{PE}(s)\quad\text{as }d,T\to\infty,\end{aligned}$$

where the second identity follows from the fact that $\overline{\Theta}^{d,T}\in C([0,\tau];C([0,1]))$. The last step follows from Theorem 5.3.1. Last, different equations that characterize Θ follows directly from Theorem 5.3.1. ■

Proof of Proposition 5.5.2. Scaling conditions in (1), (2) and (3) corresponds to the different scalings in Theorem 5.3.5. Therefore for each cases, we have $\overline{\Theta}^{d,T}\xrightarrow{\mathrm{p}}\Theta$, $U^{d,T}\xrightarrow{\mathrm{d}}U$ with Θ being the solution to (5.4). U solves (5.8), (5.9), (5.10) in (1), (2) and (3) respectively. Next we apply these convergence results to prove (i).

$$\begin{aligned}&\gamma\big(\mathsf{MSE}^{d,T}(s)-\mathsf{MSE}(s)\big)\\ =&\underbrace{-\gamma\big(\int_0^1\Theta(s,x)^2\mathrm{d}x-\frac{1}{d}\sum_{i=1}^d\Theta(s,\frac{i}{d})^2\big)}_{N_1^s}\underbrace{-\frac{\gamma}{d}\sum_{i=1}^d\big(\Theta(s,\frac{i}{d})^2-\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})^2\big)}_{N_2^s}\\ &\underbrace{-\frac{2\gamma}{d}\sum_{i=1}^d\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\big(\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})-\overline{\Theta}^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\big)}_{N_3^s}+\underbrace{\frac{\gamma}{d}\sum_{i=1}^d\big(\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})-\overline{\Theta}^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\big)^2}_{N_4^s}.\end{aligned}$$

(a) For N_1^s, since $\Theta(s,\cdot)\in C^1([0,1])\cap L^\infty([0,1])$ for any $s\in[0,\tau]$, we have

$$\big|\int_0^1\Theta(s,x)^2\mathrm{d}x-\frac{1}{d}\sum_{i=1}^d\Theta(s,\frac{i}{d})^2\big|\le 2\left\|\Theta(s,\cdot)\right\|_\infty\left\|\partial_x\Theta(s,\cdot)\right\|_\infty d^{-1}$$

Therefore

- in (1), $N_1^s=O(T^{\frac{1}{2}d^{-1}})=o(1)$ because $T=o(d^2)$.
- in (2), $N_1^s=O(\sigma^{-1}T^{\frac{1}{2}})=o(1)$ because $\max(d,T^{\frac{1}{2}})\ll\sigma$.
- in (3), $N_1^s=O(\gamma d^{-1})=o(1)$ because $\gamma\ll d$.

(b) For N_2^s, since $\Theta(\cdot,x)\in C^1([0,\tau])\cap L^\infty([0,\tau])$ for any $x\in[0,1]$, we have

$$\big|\Theta(s,\frac{i}{d})^2-\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})^2\big|\le 2\left\|\Theta(\cdot,x)\right\|_\infty\left\|\partial_s\Theta(\cdot,x)\right\|_\infty T^{-1}.$$

Therefore

- in (1), $N_2^s=O(\gamma T^{-1})=O(T^{-\frac{1}{2}})=o(1)$.
- in (2), $N_2^s=O(\gamma T^{-1})=O(\sigma^{-1}dT^{-\frac{1}{2}})=o(1)$ because $\max(d,T^{\frac{1}{2}})\ll\sigma$.
- in (3), $N_2^s=O(\gamma T^{-1})=o(dT^{-1})=o(1)$ because $d=O(T^{\frac{1}{2}})$.

(c) For N_3^s, according to the definition of $U^{d,T}$, we have

$$\begin{aligned}
N_3^s &= -\frac{2}{d}\sum_{i=1}^{d}\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})U^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\\
&= -2\int_0^1 \Theta(s,x)U(s,x)\mathrm{d}x + 2\big(\int_0^1 \Theta(s,x)U(s,x)\mathrm{d}x - \frac{1}{d}\sum_{i=1}^{d}\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})U(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\big)\\
&\quad + \frac{2}{d}\sum_{i=1}^{d}\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\big(U(\frac{\lfloor sT\rfloor}{T},\frac{i}{d}) - U^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\big).
\end{aligned}$$

As $U \in C([0,\tau];C([0,1]))$, $U^{d,T} \xrightarrow{\mathrm{d}} U$ and $\Theta \in C^1([0,\tau];C^1([0,1]))$, we have that

$$\begin{aligned}
\int_0^1 \Theta(s,x)U(s,x)\mathrm{d}x - \frac{1}{d}\sum_{i=1}^{d}\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})U(\frac{\lfloor sT\rfloor}{T},\frac{i}{d}) &= o(1)\\
\frac{2}{d}\sum_{i=1}^{d}\Theta(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\big(U(\frac{\lfloor sT\rfloor}{T},\frac{i}{d}) - U^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})\big) &= o(1).
\end{aligned}$$

Therefore, we get

$$N_3^s \xrightarrow{\mathrm{d}} -2\int_0^1 \Theta(s,x)U(s,x)\mathrm{d}x.$$

(d) For N_4^s, according to the definition of $U^{d,T}$, we have

$$\begin{aligned}
N_4^s &= -\frac{1}{\gamma d}\sum_{i=1}^{d}U^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})^2\\
&= -\gamma^{-1}\int_0^1 U(s,x)^2\mathrm{d}x - \gamma^{-1}\big(\int_0^1 U(s,x)^2\mathrm{d}x - \frac{1}{d}\sum_{i=1}^{d}U(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})^2\big)\\
&\quad - \gamma^{-1}d^{-1}\sum_{i=1}^{d}\big(U^{d,T}(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})^2 - U(\frac{\lfloor sT\rfloor}{T},\frac{i}{d})^2\big).
\end{aligned}$$

According to (5.18) and the fact that $\gamma \gg 1$, $\gamma^{-1}\int_0^1 U(s,x)^2 dx = o(1)$. As $U \in C([0,\tau]; C([0,1]))$ and $U^{d,T} \xrightarrow{\mathrm{d}} U$, we have that

$$\int_0^1 U(s,x)^2 \mathrm{d}x - \frac{1}{d}\sum_{i=1}^{d} U(\frac{\lfloor sT \rfloor}{T}, \frac{i}{d})^2 = o(1)$$

$$\frac{1}{d}\sum_{i=1}^{d}\big(U^{d,T}(\frac{\lfloor sT \rfloor}{T}, \frac{i}{d})^2 - U(\frac{\lfloor sT \rfloor}{T}, \frac{i}{d})^2\big) = o(1).$$

Therefore, $N_4^s = o(1)$.

According to points (a), (b), (c), and (d) above, (i) is proved. Statement (ii) can be proved similarly under Assumption 5.2.2. We will leave it to the readers. ■

Proof of Lemma 5.5.4. Let $\bar{A}(x) = a_0 + \sum_{k=1}^{\infty} b_k \cos(2\pi k x)$ for all $x \in [0,1]$. Then $A(x,y) = \bar{A}(|x-y|)$ for all $x, y \in [0,1]$. To prove that A satisfies Assumption 5.2.2, it suffices to show that $\bar{A} \in C^2([0,1])$. Note that $\bar{A}$ is given in the form of Fourier series with orthonormal basis $\{1, \sqrt{2}\cos(2\pi k x), \sqrt{2}\sin(2\pi k x)\}_{k\geq 1}$. Under condition (5.13), we have $\bar{A} \in H^{(5+\varepsilon)/2}([0,1])$, where $H^{(5+\varepsilon)/2}([0,1])$ is the Sobolev space of functions on $[0,1]$ with square-integrable weak derivatives up to order $(5+\varepsilon)/2$; see, for example, [AF03] for details about Sobolev spaces. By the Sobolev embedding theorem [AF03, Chapter 4], we hence have that $\bar{A} \in C^2([0,1])$, thus implying the desired result.

According to (5.15), we have

$$B(x_1, x_2, x_3, x_4) = \sum_{p \in P_4^2} \prod_{(i,j)\in p} A(x_i, x_j) \tag{5.47}$$

From Lemma 5.5.4, we know that A is bounded and twice continuously differentiable. Therefore B is continuous and bounded. Furthermore with (5.47), we have

$$\begin{aligned}&\big|B(x_1,x_3,x_1,x_3) + B(x_2,x_3,x_2,x_3) - 2B(x_1,x_3,x_2,x_3)\big| \\ =&\big|A(x_1,x_3)\left(A(x_1,x_1) + A(x_2,x_2) - 2A(x_1,x_2)\right) - 2\left(A(x_1,x_3) - A(x_2,x_3)\right)^2\big| \leq C|x_1 - x_2|^2,\end{aligned}$$

and

$$
\begin{aligned}
&\big|B(x_1,x_1,x_1,x_1)+B(x_2,x_2,x_2,x_2)+6B(x_1,x_1,x_2,x_2)-4B(x_1,x_1,x_1,x_2)-4B(x_1,x_2,x_2,x_2)\big|\\
=&3\big|A(x_1,x_1)+A(x_2,x_2)-2A(x_1,x_2)\big|^2\leq C|x_1-x_2|^4,
\end{aligned}
$$

where the last inequality follows from Lemma 5.5.4. Therefore the function B satisfies Assumption 5.2.2. For the eighth moments, similarly we have

$$
E(x_1,\cdots,x_8)=\sum_{p\in P_8^2}\prod_{(i,j)\in p}A(x_i,x_j) \tag{5.48}
$$

Therefore according to Lemma 5.5.4, E is continuous and bounded. Furthermore with (5.48), letting $\tilde{x}=(x_3,x_4,x_5,x_6)$ we have that

$$
\begin{aligned}
&\big|E(x_1,x_1,x_1,x_1,\tilde{x})+E(x_2,x_2,x_2,x_2,\tilde{x})+6E(x_1,x_1,x_2,x_2,\tilde{x})\\
&\qquad\qquad\qquad\qquad -4E(x_1,x_1,x_1,x_2,\tilde{x})-4E(x_1,x_2,x_2,x_2,\tilde{x})\big|\\
=&\ \big|3(A(x_1,x_1)+A(x_2,x_2)-2A(x-1,x_2))^2\sum_{p\in P_{3-6}^2}\prod_{(i,j)\in p}A(x_i,x_j)+\sum_{\sigma(\mathbf{i})}\prod_{j=1}^{4}(A(x_1,x_{i_j})-A(x_2,x_{i_j}))\\
&\quad+3(A(x_1,x_1)+A(x_2,x_2)-2A(x_1,x_2))\big(\sum_{\sigma(\mathbf{i})}A(x_{i_3},x_{i_4})\prod_{j=1}^{2}\big(A(x_1,x_{i_j})-A(x_2,x_{i_j})\big)\big)\big|\\
\leq&\ C\big|x_1-x_2\big|^4,
\end{aligned}
$$

where we use P_{3-6}^2 to denote the set of all pairings in $\{3,4,5,6\}$ and $\sum\limits_{\sigma(\mathbf{i})}$ to denote summing the 4-tuple $\mathbf{i}:=(i_1,i_2,i_3,i_4)$ over all permutations of the set $\{3,4,5,6\}$. The last inequality above follows from Lemma 5.5.4. Therefore, the function E satisfies Assumption 5.2.2. ■

Proof of Proposition 5.5.5. We starting by proving part (a). According to (5.4),

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s}\mathsf{MSE}(s)=\frac{\mathrm{d}}{\mathrm{d}s}\int_0^1\Theta(s,x)^2\mathrm{d}x&=-2\alpha\int_0^1\int_0^1\Theta(s,x)A(x,y)\Theta(s,y)\mathrm{d}x\mathrm{d}y\\
&=-2\alpha\sum_{i=1}^{\infty}\lambda_i\langle\Theta(s,\cdot),\phi_i\rangle_{L^2([0,1])}^2\\
&\leq-2\alpha\lambda\mathsf{MSE}(s),
\end{aligned}
$$

where the equality in the second line follows from our assumption on A and the inequality in the last line follows from the assumption that $\lambda \coloneqq \inf_i \lambda_i > 0$. The final claim follows from Gronwall's inequality.

Next, we prove part (b). Define the functional $\mathcal{F} : L^2([0,1]) \to \mathbb{R}$ as

$$\mathcal{F}[\phi] \coloneqq \int_0^1 \int_0^1 A(x,y)\phi(x)\phi(y)\mathrm{d}x\mathrm{d}y, \qquad \forall \phi \in L^2([0,1]).$$

Its functional gradient $\frac{\delta \mathcal{F}[\phi]}{\delta \phi} : [0,1] \to \mathbb{R}$ and its functional Hessian $\frac{\delta^2 \mathcal{F}[\phi]}{\delta \phi^2} : [0,1]^2 \to \mathbb{R}$ are given by

$$\frac{\delta \mathcal{F}[\phi]}{\delta \phi}(x) = 2\int_0^1 A(x,y)\phi(y)\mathrm{d}y, \quad \frac{\delta^2 \mathcal{F}[\phi]}{\delta \phi^2}(x,y) = 2A(x,y)$$

Let Θ be the solution to (5.4). For any $s \in [0,\tau]$, we have

$$\begin{aligned}\mathsf{PE}'(s) = \frac{\mathrm{d}}{\mathrm{d}s}\mathcal{F}[\Theta(s,\cdot)] &= -2\alpha \int_0^1 \int_0^1 \int_0^1 A(x,y)A(y,z)\Theta(s,x)\Theta(s,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\ &= -\frac{\alpha}{2}\int_0^1 \left|\frac{\delta \mathcal{F}[\Theta(s,\cdot)]}{\delta \Theta(s,\cdot)}(y)\right|^2 \mathrm{d}y,\end{aligned}$$

where the second identity follows from (5.4). Integrate from on interval $[0,s]$ and we get

$$\begin{aligned}\mathsf{PE}(s) &= \mathsf{PE}(0) - \frac{\alpha}{2}\int_0^s \int_0^1 \left|\frac{\delta \mathcal{F}[\Theta(r,\cdot)]}{\delta \Theta(r,\cdot)}(y)\right|^2 \mathrm{d}y\mathrm{d}r \\ &= \mathcal{F}[\Theta(0,\cdot)] - \mathcal{F}[0] - \frac{\alpha}{2}\int_0^s \int_0^1 \left|\frac{\delta \mathcal{F}[\Theta(r,\cdot)]}{\delta \Theta(r,\cdot)}(y)\right|^2 \mathrm{d}y\mathrm{d}r, \qquad (5.49)\end{aligned}$$

where the last step follows from the fact that $\mathcal{F}[0] = 0$. Notice that

$$\frac{\delta \mathcal{F}[\Theta(r,\cdot)]}{\delta \Theta(r,\cdot)}(y) = 2\int_0^1 A(x,y)\Theta(r,y)\mathrm{d}y = -\frac{2}{\alpha}\partial_r \Theta(r,x). \qquad (5.50)$$

Therefore, we have

$$\begin{aligned}\mathsf{PE}(0) - \mathcal{F}[0] &= -\left\langle \frac{\delta \mathcal{F}[\Theta(0,\cdot)]}{\delta \Theta(0,\cdot)}, -\Theta(0,\cdot)\right\rangle_{L^2([0,1])} - \int_0^1 \int_0^1 \Theta(0,x)A(x,y)\Theta(0,y)\mathrm{d}x\mathrm{d}y \\ &\leq -\frac{2}{\alpha}\langle \partial_s \Theta(s,x)|_{s=0}, \Theta(0,\cdot)\rangle_{L^2([0,1])} \\ &= -\frac{1}{\alpha}\left(\frac{\mathrm{d}}{\mathrm{d}s}\int_0^1 \Theta(s,y)^2 \mathrm{d}y\right)\Bigg|_{s=0}, \qquad (5.51)\end{aligned}$$

where the first step follows from Taylor expansion and the inequality follows from our assumption on A and (5.50). Meanwhile, according to (5.50),

$$\begin{aligned}
-\frac{\alpha}{2}\int_0^s\int_0^1\left|\frac{\delta\mathcal{F}[\Theta(r,\cdot)]}{\delta\Theta(r,\cdot)}(y)\right|^2\mathrm{d}y\mathrm{d}r &= -\frac{2}{\alpha}\int_0^s\int_0^1\left(\partial_r\Theta(r,y)\right)^2\mathrm{d}y\mathrm{d}r = -\frac{2}{\alpha}\int_0^1\int_0^s\partial_r\Theta(r,y)\Theta(r,y)\mathrm{d}r\mathrm{d}y\\
&= -\frac{2}{\alpha}\int_0^1\Theta(s,y)\partial_r\Theta(r,y)|_{r=s}-\Theta(0,y)\partial_r\Theta(r,y)|_{r=0}\mathrm{d}y\\
&\quad+\frac{2}{\alpha}\int_0^1\int_0^s\Theta(r,y)\partial^2_{rr}\Theta(r,y)\mathrm{d}r\mathrm{d}y\\
&= \frac{2}{\alpha}\left(\frac{\mathrm{d}}{\mathrm{d}r}\int_0^1\Theta(r,x)^2\mathrm{d}x\bigg|_{r=0}-\frac{\mathrm{d}}{\mathrm{d}r}\int_0^1\Theta(r,x)^2\mathrm{d}x\bigg|_{r=s}\right)\\
&\quad+\frac{\alpha}{2}\int_0^s\int_0^1\left|\frac{\delta\mathcal{F}[\Theta(r,\cdot)]}{\delta\Theta(r,\cdot)}(y)\right|^2\mathrm{d}y\mathrm{d}r \qquad (5.52)
\end{aligned}$$

where the last step follows because

$$\begin{aligned}
\frac{2}{\alpha}\int_0^1\int_0^s\Theta(r,y)\partial^2_{rr}\Theta(r,y)\mathrm{d}r\mathrm{d}y &= -2\int_0^1\int_0^s\Theta(r,y)\int_0^1A(y,z)\partial_r\Theta(r,z)\mathrm{d}z\mathrm{d}r\mathrm{d}y\\
&= 2\alpha\int_0^1\int_0^s\Theta(r,y)\int_0^1A(y,z)\int_0^1A(z,x)\Theta(r,x)\mathrm{d}x\mathrm{d}z\mathrm{d}r\mathrm{d}y\\
&= 2\alpha\int_0^s\int_0^1\left(\int_0^1A(y,z)\Theta(r,z)\mathrm{d}z\right)^2\mathrm{d}y\mathrm{d}s\\
&= \frac{\alpha}{2}\int_0^s\int_0^1\left|\frac{\delta\mathcal{F}[\Theta(r,\cdot)]}{\delta\Theta(r,\cdot)}(y)\right|^2\mathrm{d}y\mathrm{d}r.
\end{aligned}$$

Therefore based on combining (5.49), (5.51) and (5.52), we have that

$$\begin{aligned}
\mathsf{PE}(s) &\le -\frac{1}{\alpha}\left(\frac{\mathrm{d}}{\mathrm{d}s}\int_0^1\Theta(s,y)^2\mathrm{d}y\right)\bigg|_{s=0}+\frac{1}{\alpha}\left(\frac{\mathrm{d}}{\mathrm{d}r}\int_0^1\Theta(r,x)^2\mathrm{d}x\bigg|_{r=0}-\frac{\mathrm{d}}{\mathrm{d}r}\int_0^1\Theta(r,x)^2\mathrm{d}x\bigg|_{r=s}\right)\\
&= -\frac{1}{\alpha}\frac{\mathrm{d}}{\mathrm{d}s}\int_0^1\Theta(s,x)^2\mathrm{d}x.
\end{aligned}$$

Integrating over $[0,\tau]$, we have

$$\int_0^\tau\mathsf{PE}(s)\mathrm{d}s\le-\frac{\mathsf{MSE}(\tau)}{\alpha}+\frac{\mathsf{MSE}(0)}{\alpha}\le\frac{\mathsf{MSE}(0)}{\alpha}.$$

From previous calculations we know that $\mathsf{PE}'(s) = -\frac{\alpha}{2}\int_0^1 \left|\frac{\delta\mathcal{F}[\Theta(s,\cdot)]}{\delta\Theta(s,\cdot)}(y)\right|^2 \mathrm{d}y \leq 0$. Therefore

$$\mathsf{PE}(\tau) \leq \frac{\int_0^\tau \mathsf{PE}(s)\mathrm{d}s}{\tau} \leq \frac{\mathsf{MSE}(0)}{\alpha\tau}.$$

■

Bibliography

[AAM22] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[ABE19] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*. PMLR, 2019.

[ABG23] Bhavya Agrawalla, Krishnakumar Balasubramanian, and Promit Ghosal. High-dimensional Central Limit Theorems for Linear Functionals of Online Least-Squares SGD. *arXiv preprint arXiv:2302.09727*, 2023.

[AC21] Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. *Advances in Neural Information Processing Systems*, 34:28405–28418, 2021.

[AD19] Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 2019.

[ADFDJ03] 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.

[ADW21] Christophe Andrieu, Paul Dobson, and Andi Q. Wang. Subgeometric hypocoercivity for piecewise-deterministic Markov process Monte carlo methods. *Electronic Journal of Probability*, 26, 2021.

[AF03] Robert A Adams and John JF Fournier. *Sobolev spaces*. Elsevier, 2003.

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

[ALPW21] Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q Wang. Comparison of Markov chains via weak Poincaré inequalities with application to pseudo-marginal MCMC. *arXiv preprint arXiv:2112.05605*, 2021.

[AM14] Douglas Azevedo and Valdir Antonio Menegatto. Sharp estimates for eigenvalues of integral operators generated by dot product kernels on the sphere. *Journal of Approximation Theory*, 177:57–68, 2014.

[Arn92] Vladimir I Arnold. *Ordinary differential equations.* Springer Science & Business Media, 1992.

[AT07] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*, volume 80. Springer, 2007.

[BAGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

[BAGJ22] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

[BBD$^+$09] 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.

[BBHS21] Anas Barakat, Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance. *Electronic Journal of Statistics*, 15(2):3892–3947, 2021.

[BCE$^+$22] Krishnakumar 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.

[BCG08] Dominique Bakry, Patrick Cattiaux, and Arnaud Guillin. Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *Journal of Functional Analysis*, 254(3):727–759, 2008.

[BCN18] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[BÉ85] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Seminaire de probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

[BES+22] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation. In *Advances in Neural Information Processing Systems*, 2022.

[BG97] Lorenzo Bertini and Giambattista Giacomin. Stochastic Burgers and KPZ equations from particle systems. *Comm. Math. Phys.*, 183(3):571–607, 1997.

[BG22] 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.

[BGJM11] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 103. Springer, 2014.

[BJM16] 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.

[BL09] Sergey Bobkov and Michel Ledoux. Weighted Poincaré-type inequalities for Cauchy and other convex measures. *The Annals of Probability*, 37(2):403–427, 2009.

[BLY21] Krishnakumar Balasubramanian, Tong Li, and Ming Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 22(1), 2021.

[BMP12] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

[BREZ20] 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.

[BRZ19] Joris Bierkens, Gareth Roberts, and Pierre-André Zitt. Ergodicity of the zigzag process. *The Annals of Applied Probability*, 29(4):2266–2301, 2019.

[BS10] Márton Balázs and Timo Seppäläinen. Order of current variance and diffusivity in the asymmetric simple exclusion process. *Ann. of Math. (2)*, 171(2):1237–1265, 2010.

[BTA11] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer Science & Business Media, 2011.

[BZ17] 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.

[CB16] Katy Craig and Andrea Bertozzi. A blob method for the aggregation equation. *Mathematics of Computation*, 85(300):1681–1717, 2016.

[CB18a] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211, 2018.

[CB18b] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[CCAY$^+$18] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

[CCBJ17] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

[CCM21] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

[CCP19] José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):1–53, 2019.

[CCSW22] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 2984–3014, 2022.

[CDV09] 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.

[CDWY20] 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.

[CEG17] Dario Cordero-Erausquin and Nathael Gozlan. Transport proofs of weighted poincaré inequalities for log-concave distributions. *Bernoulli*, 23(1):134–158, 2017.

[CEL$^+$21] Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

[CFG14] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691, 2014.

[CGGR10] 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.

[CGMZ19] 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.

[CGST20] Ivan Corwin, Promit Ghosal, Hao Shen, and Li-Cheng Tsai. Stochastic PDE limit of the six vertex model. *Communications in Mathematical Physics*, 375(3):1945–2038, 2020.

[CGW11] Patrick Cattiaux, Arnaud Guillin, and Li-Ming Wu. Some remarks on Weighted Logarithmic Sobolev Inequality. *Indiana University Mathematics Journal*, pages 1885–1904, 2011.

[Che17] Alina Chertock. A practical guide to deterministic particle methods. In *Handbook of Numerical Analysis*, volume 18, pages 177–202. Elsevier, 2017.

[CLA$^+$21] 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.

[CLGL$^+$20a] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109, 2020.

[CLGL$^+$20b] 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.

[CLP22] Kabir Aladin Chandrasekher, Mengqi Lou, and Ashwin Pananjady. Alternating minimization for generalized rank one matrix sensing: Sharp predictions from a random initialization. *arXiv preprint arXiv:2207.09660*, 2022.

[CLS21] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2021.

[CLW20] Yu Cao, Jianfeng Lu, and Lihan Wang. Complexity of randomized algorithms for underdamped Langevin dynamics. *arXiv preprint arXiv:2003.09906*, 2020.

[CLW21] Yu Cao, Jianfeng Lu, and Lihan Wang. Complexity of randomized algorithms for underdamped Langevin dynamics. *Communications in Mathematical Sciences*, 19(7):1827–1853, 2021.

[Cor12] Ivan Corwin. The Kardar–Parisi–Zhang equation and universality class. *Random matrices: Theory and applications*, 1(01):1130001, 2012.

[CPT23] Kabir Aladin Chandrasekher, Ashwin Pananjady, and Christos Thrampoulidis. Sharp global convergence guarantees for iterative nonconvex optimization with random data. *The Annals of Statistics*, 51(1):179–210, 2023.

[CSG16] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *JMLR: Workshop and Conference Proceedings*, 2016.

[CST18] Ivan Corwin, Hao Shen, and Li-Cheng Tsai. ASEP(q, j) converges to the KPZ equation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 54(2):995–1012, 2018.

[CV22] 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.

[CW97] Mu-Fa Chen and Feng-Yu Wang. Estimates of logarithmic Sobolev constant: An improvement of Bakry–Emery criterion. *Journal of functional analysis*, 144(2):287–300, 1997.

[CX20] Lin Chen and Sheng Xu. Deep neural tangent kernel and Laplace kernel have the same RKHS. In *International Conference on Learning Representations*, 2020.

[CZ07] Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

[Dal17a] Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689. PMLR, 07–10 Jul 2017.

[Dal17b] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[DBCD19] George Deligiannidis, Alexandre Bouchard-Côté, and Arnaud Doucet. Exponential ergodicity of the bouncy particle sampler. *The Annals of Statistics*, 47(3):1268–1287, 2019.

[DCLW22] Zhiyan Ding, Shi Chen, Qin Li, and Stephen J Wright. Overparameterization of deep ResNet: Zero loss and mean-field analysis. *Journal of machine learning research*,

2022.

[DCWY19] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

[DDB20] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3):1348–1382, 2020.

[DDJ23] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *arXiv preprint arXiv:2301.06632*, 2023.

[DGM20] Alain Durmus, Arnaud Guillin, and Pierre Monmarché. Geometric ergodicity of the bouncy particle sampler. *The Annals of Applied Probability*, 30(5):2069–2098, 2020.

[DK19] 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.

[DKRD19] Arnak S Dalalyan, Avetik Karagulyan, and Lionel Riou-Durand. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *arXiv preprint arXiv:1906.08530*, 2019.

[DKTZ20] 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.

[DL21] Zhiyan Ding and Qin Li. Langevin Monte Carlo: Random coordinate descent and variance reduction. *J. Mach. Learn. Res.*, 22:205–1, 2021.

[DM90] Pierre Degond and Francisco-José Mustieles. A deterministic approximation of diffusion equations using particles. *SIAM Journal on Scientific and Statistical Computing*, 11(2):293–310, 1990.

[DM17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the Unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 06 2017.

[DM19] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[DMM19] Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

[DMN$^+$21] Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. *Advances in Neural Information Processing Systems*, 34:30063–30074, 2021.

[DMNS22] Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for Polyak-Ruppert averaged iterates of linear stochastic approximation. *arXiv preprint arXiv:2207.04475*, 2022.

[DMP18] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov Chain Monte Carlo: When Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[DMPS18] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains.* Springer, 2018.

[DMS19] Alain Durmus, Eric Moulines, and Eero Saksman. On the convergence of Hamiltonian Monte Carlo. *The Annals of Statistics (to appear)*, 2019+.

[DN08] RWR Darling and JR Norris. Differential equation approximations for Markov chains. *Probability Surveys*, 5:37–79, 2008.

[DNS19] Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

[Dob79] Roland Dobrushin. Vlasov equations. *Functional Analysis and Its Applications*, 13(2):115–123, 1979.

[DOV22] Duncan Dauvergne, Janosch Ortmann, and Bálint Virág. The directed landscape. *Acta Math.*, 229(2):201–285, 2022.

[DPZ14] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions.* Cambridge university press, 2014.

[DR20] John Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 2020.

[DRD20a] Arnak Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

[DRD20b] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

[DT19] Percy Deift and Thomas Trogdon. Universality in numerical computation with random data: Case studies and analytical results. *Journal of Mathematical Physics*, 60(10):103306, 2019.

[DT22] Xiucai Ding and Thomas Trogdon. The conjugate gradient algorithm on a general class of spiked covariance matrices. *Quarterly of Applied Mathematics*, 80(1):99–155, 2022.

[Ebe16] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3-4):851–886, 2016.

[EGZ19] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

[EH20] Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. *arXiv preprint arXiv:2005.13097*, 2020.

[EH21] 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.

[EMS18] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9671–9680, 2018.

[FDBD21] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for SGD and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

[FXY18] Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 2018.

[GB09] Alan Genz and Frank Bretz. *Computation of multivariate normal and t-probabilities*, volume 195. Springer Science & Business Media, 2009.

[GBH04] 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.

[GCPT20] Abhishek Gupta, Hao Chen, Jianzong Pi, and Gaurav Tendolkar. Some limit properties of Markov chains induced by recursive stochastic algorithms. *SIAM Journal on Mathematics of Data Science*, 2(4):967–1003, 2020.

[GDVM16] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *arXiv preprint arXiv:1611.06972*, 2016.

[GDVM19] 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.

[GGK22] Benjamin Gess, Rishabh S Gvalani, and Vitalii Konarovskyi. Conservative SPDEs as fluctuating mean field limits of stochastic gradient descent. *arXiv preprint arXiv:2207.05705*, 2022.

[Gho17] Promit Ghosal. Hall-Littlewood-PushTASEP and its KPZ limit. *arXiv e-prints*, page arXiv:1701.07308, January 2017.

[GJPS08] 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.

[GLM18] 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.

[GM17] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1292–1301. JMLR. org, 2017.

[GTM$^+$22] Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean field theory for stochastic gradient descent methods. *arXiv preprint arXiv:2210.06591*, 2022.

[HBE20] 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.

[HBSL22] Ye He, Krishnakumar Balasubramanian, Bharath Sriperumbudur, and Jianfeng Lu. Regularized Stein Variational Gradient Flow. *Submitted to Foundations of Computational Mathematics*, 2022.

[HKRC18] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 2878–2887, 2018.

[HLW06] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

[HMW21] Lu-Jing Huang, Mateusz B Majka, and Jian Wang. Approximation of heavy-tailed distributions via stable-driven SDEs. *Bernoulli*, 27(3):2040–2068, 2021.

[HS87] Richard Holley and Daniel Stroock. Logarithmic Sobolev Inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5-6):1159–1194, 1987.

[IM94] 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.

[JG12] 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.

[Jia21] Qijia Jiang. Mirror Langevin Monte Carlo: the Case Under Isoperimetry. *Advances in Neural Information Processing Systems*, 34:715–725, 2021.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17,

1998.

[JR07] Søren Jarner and Gareth Roberts. Convergence of heavy-tailed Monte Carlo Markov Chain algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815, 2007.

[Kam18] 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.

[KB17] Walid Krichene and Peter L Bartlett. Acceleration and averaging in stochastic descent dynamics. *Advances in Neural Information Processing Systems*, 30, 2017.

[KC12] Harold Joseph Kushner and Dean S Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 2012.

[KDL$^+$21] Koulik Khamaru, Yash Deshpande, Tor Lattimore, Lester Mackey, and Martin J Wainwright. Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266v3*, 2021.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[KGY03] Harold J. Kushner and Gang George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[KL98] Claude Kipnis and Claudio Landim. *Scaling limits of interacting particle systems*, volume 320. Springer Science & Business Media, 1998.

[KN04] Samuel Kotz and Saralees Nadarajah. *Multivariate t-distributions and their applications.* Cambridge University Press, 2004.

[KS12] Ioannis Karatzas and Steven Shreve. *Brownian Motion and Stochastic Calculus*, volume 113. Springer Science & Business Media, 2012.

[KSA$^+$20] Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for Stein Variational Gradient Descent. *Advances in Neural Information Processing Systems*, 33, 2020.

[Kwa17] Mateusz Kwaśnicki. Ten equivalent definitions of the fractional laplace operator. *Fractional Calculus and Applied Analysis*, 20(1):7–51, 2017.

[KX95] Gopinath Kallianpur and Jie Xiong. Stochastic differential equations in infinite-dimensional spaces. *Lecture Notes-Monograph Series*, 26:iii–342, 1995.

[Lan20] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

[LBBG19] Samuel Livingstone, Michael Betancourt, Simon Byrne, and Mark Girolami. On the geometric ergodicity of Hamiltonian Monte Carlo. *Bernoulli*, 25(4A):3109–3138, 2019.

[LCCC16] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned Stochastic Gradient Langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[LE20] Mufan Bill Li and Murat A Erdogdu. Riemannian langevin algorithm for solving semidefinite programs. *arXiv preprint arXiv:2010.11176*, 2020.

[Leh21] Joseph Lehec. The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *arXiv preprint arXiv:2101.10695*, 2021.

[Lig99] Thomas Liggett. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*, volume 324. springer science & Business Media, 1999.

[Liu08] Jun S Liu. *Monte Carlo strategies in scientific computing.* Springer Science & Business Media, 2008.

[Liu17] Qiang Liu. Stein Variational Gradient Descent as gradient flow. *Advances in Neural Information Processing Systems*, 30, 2017.

[LLJ16] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284, 2016.

[LLN19] Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the Stein Variational Gradient Descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019.

[LM16] Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods.* Springer, 2016.

[LO00] Rafał Latała and Krzysztof Oleszkiewicz. Between Sobolev and Poincaré. In *Geometric aspects of functional analysis*, pages 147–168. Springer, 2000.

[LP02] Damien Lamberton and Gilles Pages. Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3):367–405, 2002.

[LPS14] Gabriel J Lord, Catherine E Powell, and Tony Shardlow. *An introduction to computational stochastic PDEs*, volume 50. Cambridge University Press, 2014.

[LPW12] Lennart Ljung, Georg Pflug, and Harro Walk. *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser, 2012.

[LRLP17] Yang Liu, Prajit Ramachandran, Qiang Liu, and Jian Peng. Stein variational policy gradient. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017*, 2017.

[LS16] Tony Lelievre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016.

[LST20] 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.

[LST21] 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.

[LT93] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[LTVW22] 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.

[LW16] Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

[LW22] Jianfeng Lu and Lihan Wang. Complexity of zigzag sampling algorithm for strongly log-concave distributions. *Statistics and Computing*, 32(3):1–12, 2022.

[LWLZ18] Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167(1):75–97, 2018.

[LWME19] Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems*, pages 7748–7760, 2019.

[LY22] Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory*, pages 2963–2983. PMLR, 2022.

[MCC$^+$19] Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

[MCC$^+$21] 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.

[MCJ$^+$19] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

[Mey22] Sean Meyn. *Control systems and reinforcement learning*. Cambridge University Press, 2022.

[MHB17] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate Bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

[MHFH$^+$23] Alireza Mousavi-Hosseini, Tyler Farghly, Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. Towards a complete analysis of langevin monte carlo: Beyond poincar\'e inequality. *arXiv preprint arXiv:2303.03589*, 2023.

[MHKC20] Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[MMS20] Mateusz B Majka, Aleksandar Mijatović, and Łukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, 30(4):1534–1581, 2020.

[MMW$^+$19] Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm. *arXiv preprint arXiv:1908.10859*, 2019.

[MNY06] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer, 2006.

[Mon21] Pierre Monmarché. High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electronic Journal of Statistics*, 15(2):4117–4166, 2021.

[MPM$^+$20] Eric Mazumdar, Aldo Pacchiano, Yi-an Ma, Peter L Bartlett, and Michael I Jordan. On thompson sampling with langevin algorithms. *arXiv preprint arXiv:2002.10002*, 2020.

[MPS12] Jonathan C Mattingly, Natesh S Pillai, and Andrew M Stuart. Diffusion limits of the Random Walk Metropolis algorithm in high dimensions. *The Annals of Applied Probability*, 22(3):881–930, 2012.

[MQR21] Konstantin Matetski, Jeremy Quastel, and Daniel Remenik. The KPZ fixed point. *Acta Math.*, 227(1):115–203, 2021.

[MRZ16] Adrian Muntean, Jens Rademacher, and Antonios Zagaris. *Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity.* Springer, 2016.

[MSH02] Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: Locally lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

[MT04] Grigori N Milstein and Michael V Tretyakov. *Stochastic numerics for mathematical physics*, volume 456. Springer, 2004.

[MT12] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

[MT13] Grigori Noah Milstein and Michael V Tretyakov. *Stochastic numerics for mathematical physics.* Springer Science & Business Media, 2013.

[Nea11] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2(11):2, 2011.

[Ngu21] Dao Nguyen. Unadjusted Langevin Algorithm for sampling a mixture of weakly smooth potentials. *arXiv preprint arXiv:2112.09311*, 2021.

[NS17] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[NŞR19] 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.

[PJ92] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[PLPP21] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pages 3548–3626. PMLR, 2021.

[PN21] Huy Tuan Pham and Phan-Minh Nguyen. Limiting fluctuation and trajectorial stability of multilayer neural networks with mean field training. *Advances in Neural Information Processing Systems*, 34:4843–4855, 2021.

[PP21] Courtney Paquette and Elliot Paquette. Dynamics of stochastic momentum methods on large-scale, quadratic models. *Advances in Neural Information Processing Systems*, 34:9229–9240, 2021.

[PPAP22a] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *arXiv preprint arXiv:2205.07069*, 2022.

[PPAP22b] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit Regularization or Implicit Conditioning? Exact Risk Trajectories of SGD in High Dimensions. In *Advances in Neural Information Processing Systems*, 2022.

[PR16] Vern Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152. Cambridge University Press, 2016.

[PST12] Natesh S Pillai, Andrew M Stuart, and Alexandre H Thiéry. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, pages 2320–2356, 2012.

[PvMPP22] Courtney Paquette, Bart van Merriënboer, Elliot Paquette, and Fabian Pedregosa. Halting time is predictable for large models: A universality property and average-case analysis. *Foundations of Computational Mathematics*, pages 1–77, 2022.

[QS23] Jeremy Quastel and Sourav Sarkar. Convergence of exclusion processes and the KPZ equation to the KPZ fixed point. *J. Amer. Math. Soc.*, 36(1):251–289, 2023.

[Qua11] Jeremy Quastel. Introduction to KPZ. *Current developments in mathematics*, 2011(1), 2011.

[Rav85] Pierre-Arnaud Raviart. An analysis of particle methods. In *Numerical Methods in Fluid Dynamics*, pages 243–324. Springer, 1985.

[RC99] Christian Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[RDF78] 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.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Rot12] Michael Roth. *On the multivariate t distribution.* Linköping University Electronic Press, 2012.

[RR98] 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.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1674–1703, 2017.

[RSBG22] 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.

[RT96] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[Rup88] David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[Rus90] Giovanni Russo. Deterministic diffusion of particles. *Communications on Pure and Applied Mathematics*, 43(6):697–733, 1990.

[RVE22] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

[RW01] Michael Röckner and Feng-Yu Wang. Weak Poincaré inequalities and L_2 convergence rates of Markov semigroups. *Journal of Functional Analysis*, 185(2):564–603, 2001.

[RW03] Michael Röckner and Feng-Yu Wang. Harnack and functional inequalities for generalized Mehler semigroups. *Journal of Functional Analysis*, 203(1):237–261, 2003.

[San17] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: An overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[SC08] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[SGBSM20] Carl-Johann Simon-Gabriel, Alessandro Barp, Bernhard Schölkopf, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv preprint arXiv:2006.09268*, 2020.

[SGD21] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on*

Learning Theory, pages 3935–3971. PMLR, 2021.

[SGF$^+$10] Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

[SH21] Meyer Scetbon and Zaid Harchaoui. A spectral analysis of dot-product kernels. In *International conference on Artificial Intelligence and Statistics*, pages 3394–3402. PMLR, 2021.

[Şim17] Umut Şimşekli. Fractional Langevin Monte Carlo: Exploring Lévy driven stochastic differential equations for Markov Chain Monte Carlo. In *International Conference on Machine Learning*, pages 3200–3209, 2017.

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2098–2109, 2019.

[Spo12] Herbert Spohn. *Large scale dynamics of interacting particles*. Springer Science & Business Media, 2012.

[SR22] Lukang Sun and Peter Richtárik. A note on the convergence of mirrored Stein Variational Gradient Descent under (L_0, L_1)-smoothness condition. *arXiv preprint arXiv:2206.09709*, 2022.

[SS20a] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[SS20b] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[SS22] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

[SSR22] Adil Salim, Lukang Sun, and Peter Richtarik. A convergence theory for SVGD in the population limit under Talagrand's inequality T_1. In *International Conference on Machine Learning*, pages 19139–19152. PMLR, 2022.

[SZ19] Sotirios Sabanis and Ying Zhang. Higher order Langevin Monte Carlo algorithm. *Electronic Journal of Statistics*, 13(2):3805–3850, 2019.

[SZ22] Qi-Man Shao and Zhuo-Song Zhang. Berry–Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3):1548–1576, 2022.

[Szn04] Alain-Sol Sznitman. Topics in random walks in random environment. In *School and conference on probability theory: 13-17 May 2002*, volume 17, pages 203–266, 2004.

[ŞZTG20] 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.

[TA17] Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

[TP18] 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.

[TSA20] Nicolas Garcia Trillos and Daniel Sanz-Alonso. The Bayesian update: Variational formulations and gradient flows. *Bayesian Analysis*, 15(1):29–56, 2020.

[TTV16] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for Stochastic Gradient Langevin Dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

[TV00] Guiseppe Toscani and Cédric Villani. On the trend to equilibrium for some dissipative systems with slowly increasing a priori bounds. *Journal of Statistical Physics*, 98(5-6):1279–1309, 2000.

[TV23] Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *The Journal of Machine Learning Research*, 24, 2023.

[Vem10] Santosh S Vempala. Recent progress and open problems in algorithmic convex geometry. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2010)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2010.

[Vig12] C Vignat. A generalized Isserlis theorem for location mixtures of Gaussian random vectors. *Statistics & probability letters*, 82(1):67–71, 2012.

[Vil09] Cédric Villani. The Wasserstein distances. In *Optimal Transport*, pages 93–111. Springer, 2009.

[Vil21] Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Soc., 2021.

[VPD22] 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.

[VSL$^+$22] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[VW19] Santosh Vempala and Andre Wibisono. Rapid convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices. In *Advances in Neural Information Processing Systems*, pages 8092–8104, 2019.

[Wan06] Feng-Yu Wang. *Functional inequalities Markov semigroups and spectral theory.* Elsevier, 2006.

[Wan14] Jian Wang. A simple approach to functional inequalities for non-local Dirichlet forms. *ESAIM: Probability and Statistics*, 18:503–513, 2014.

[Wib18] Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference On Learning Theory*, pages 2093–3027, 2018.

[WL19] Chuang Wang and Yue M Lu. The scaling limit of high-dimensional online independent component analysis. *Journal of Statistical Mechanics*, 2019(12), 2019.

[WML17] Chuang Wang, Jonathan Mattingly, and Yue M Lu. Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and PCA. *arXiv preprint arXiv:1712.04332*, 2017.

[WSC21] Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-Adjusted Langevin Algorithm for log-concave sampling. *arXiv preprint arXiv:2109.13055*, 2021.

[WSC22] 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.

[WT11] Max Welling and Yee W Teh. Bayesian learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[WTBL19] Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein Variational Gradient Descent with matrix-valued kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

[WW15] Feng-Yu Wang and Jian Wang. Functional inequalities for stable-like Dirichlet forms. *Journal of Theoretical Probability*, 28(2):423–448, 2015.

[WW22] Jun-Kun Wang and Andre Wibisono. Accelerating Hamiltonian Monte Carlo via Chebyshev Integration Time. *arXiv preprint arXiv:2207.02189*, 2022.

[WZL18] Dilin Wang, Zhe Zeng, and Qiang Liu. Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning*, pages 5219–5227. PMLR, 2018.

[XKS22] Lantian Xu, Anna Korba, and Dejan Slepčev. Accurate quantization of measures via interacting particle-based optimization. In *International Conference on Machine Learning*, pages 24576–24595. PMLR, 2022.

[YBVE21] Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. *Advances in Neural Information Processing*, 2021.

[YŁR22] Jun Yang, Krzysztof Łatuszyński, and Gareth Roberts. Stereographic Markov Chain Monte Carlo. *arXiv preprint arXiv:2205.12112*, 2022.

[Zei04] Ofer Zeitouni. Random walks in random environment. *Lecture notes in Mathematics*, 1837:190–312, 2004.

[ZPFP20] 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.

[ZYB21] Baishuai Zuo, Chuancun Yin, and Narayanaswamy Balakrishnan. Expressions for joint moments of elliptical distributions. *Journal of Computational and Applied Mathematics*, 391:113418, 2021.

[ZZ22] Xiaolong Zhang and Xicheng Zhang. Ergodicity of supercritical SDEs driven by α-stable processes and heavy-tailed sampling. *arXiv preprint arXiv:2201.10158*, 2022.