Spectral Gap of Replica Exchange Langevin Diffusion on Mixture Distributions

Jing Dong[a,1], Xin T. Tong[b,2,*]

[a] *Columbia University, 3022 Broadway, New York, NY 10027.*
[b] *National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076*

Abstract

Langevin diffusion (LD) is one of the main workhorses for sampling problems. However, its convergence rate can be significantly reduced if the target distribution is a mixture of multiple densities, especially when each component density concentrates around a different mode. Replica exchange Langevin diffusion (ReLD) is a sampling method that can circumvent this issue. This approach can be further extended to multiple replica exchange Langevin diffusion (mReLD). While ReLD and mReLD have been used extensively in statistics, molecular dynamics, and other applications, there is limited existing analysis on its convergence rate and choices of the temperatures. This paper addresses these problems assuming the target distribution is a mixture of log-concave densities. We show ReLD can obtain constant or better convergence rates. We also show mReLD with K additional LDs can achieve the same results while the exchange frequency only needs to be $(1/K)$-th power of the one in ReLD.

Keywords: Diffusion process, spectral gap, Markov Chain Monte Carlo
Poincaré inequality, mixture model

[*] Corresponding author

Email addresses: `jing.dong@gsb.columbia.edu` (Jing Dong), `xin.t.tong@nus.edu.sg` (Xin T. Tong)

[1] Support from NSF grant DMS-1720433 is gratefully acknowledged by J. Dong.

[2] Singapore MOE grants R-146-000-292-114 and MOE2018-T2-1-156 are gratefully acknowledged by X.T. Tong

1. Introduction

Given a d-dimensional distribution $\pi(x) \propto \exp(-H(x))$, a standard way to generate samples from π is simulating the overdamped Langevin diffusion (LD):

$$dX(t) = \nabla \log \pi(X(t))dt + \sqrt{2}dB(t), \tag{1}$$

where $B(t)$ is a d-dimensional Brownian motion, for a long enough time horizon. The main justification of this approach is that, under mild conditions, the invariant measure of $X(t)$ is the target distribution π. This approach can be quite efficient when the potential function/Hamiltonian $H(x)$ is strongly-convex. However, if $H(x)$ has multiple local minima and each of them is located inside a deep potential well, LD can be very inefficient. In such cases, LD will spend a large amount time circulating inside one potential well before it can reach another potential well (see, for example, [1]). Such behavior significantly slows down its convergence rate to stationarity.

Replica exchange Monte Carlo, also known as parallel tempering, is a method that has been used extensively in molecular dynamic (MD) and statistics to improve the convergence rate of the sampling process when the target distribution is multimodal [2]. When combined with LD, it considers simulating additional LDs (beyond $X(t)$), where each of them is targeting a higher tempered version of π. In general, a high temperature flattens the potential wells so that it is easier for the corresponding LD to move between different potential wells. Periodically, the replicas exchange their locations through a Metropolis Hasting mechanism. Such exchanges can help switch $X(t)$ out of its current potential well while keeping π as its invariant measure. The exact formulation of Replica exchange Langevin diffusion (ReLD) can be found in Section 2.3. Numerical versions of ReLD (through appropriate discretization) have been applied to various applications and achieved significant efficiency gain over LD (see, e.g., [3, 4, 5]).

Despite the elegant intuition and empirical success of ReLD, there is limited theoretical analysis of why and when it performs well. This is partly because most analytical framework does not handle non-convex potential functions well.

ReLD as a Markov process is also more complicated to analyze than LD due to the exchange dynamic. In this paper, we invetigate the performance of ReLD by providing an explicit quantification of its spectral gap. This spectral gap characterization provides guidance on how to choose the temperatures and swapping intensities in ReLD.

1.1. Spectral gap and slow LD convergence on mixture distributions

The convergence rate of a continuous time Markov process Z_t can be characterized by a quantity called the *spectral gap*. To formally define the spectral gap, we first define the generator of Z_t as

$$\mathcal{L}(f)(z) := \lim_{t\to 0} \frac{1}{t}\mathbb{E}[f(Z_t) - f(z)|Z_0 = z],$$

for $f \in \mathcal{D}(\mathcal{L})$ where $\mathcal{D}(\mathcal{L})$ is a subset of $\mathbb{C}_c^2(\mathbb{R}^d)$ such that the above limit exists and $\mathbb{C}_c^2(\mathbb{R}^d)$ is the space of twice continuously differentiable functions with compact support. For most of the Markov processes discussed below, one can simply use $\mathcal{D}(\mathcal{L}) = \mathbb{C}_c^2(\mathbb{R}^d)$. We define the associated carré du champ as

$$\Gamma(f) = \frac{1}{2}(\mathcal{L}(f^2) - 2f\mathcal{L}(f)),$$

and the Dirichlet form as $\mathcal{E}(f) = \int \Gamma(f)\pi^Z(dx)$, where π^Z is the invariant distribution of Z_t. The inverse spectral gap of Z_t can then be defined as

$$\kappa = \sup\left\{\frac{\mathrm{var}_{\pi^Z}(f)}{\mathcal{E}(f)}; f \in \mathbb{C}_c^2(\mathbb{R}^d), \mathcal{E}(f) \neq 0\right\}, \tag{2}$$

where we use $\mathrm{var}_\mu(f)$ to denote the variance of f under μ. As a remark, the domain of the operator defined above usually can be further extended, so that $\mathbb{C}_c^2(\mathbb{R}^d)$ is a core of it. The definition of κ can also be extended (see [6] section 1.4 and 1.13). We restricted our discussion to $\mathbb{C}_c^2(\mathbb{R}^d)$ for simplicity.

The reason why κ controls the speed of convergence of Z_t towards π^Z can be found in Theorem 4.2.5 of [6]. In particular, for any test function $f \in L^2(\pi^Z)$, i.e., square-integrable functions, there is a constant C_0 such that

$$\int (\mathbb{E}[f(Z_t)|Z_0 = z] - \mathbb{E}_{\pi^Z} f(Z))^2 \pi^Z(z)dz \leq C_0 e^{-2t/\kappa}.$$

In other words, $\mathbb{E}[f(Z_t)|Z_0 = z]$ converges to the target expectation exponentially fast with π^Z-a.s. initial conditions, and the convergence rate is $1/\kappa$, i.e., a smaller κ leads to a faster convergence rate.

Using the inverse spectral gap, we can show that LD converges very quickly for a singular Gaussian distribution, but very slowly for a mixture of two singular Gaussians. Let ϕ denote the density function of a d-dimensional standard Gaussian random vector, i.e., $\phi(x) = (2\pi)^{-d/2}\exp(-\|x\|^2/2)$.

Proposition 1 *The inverse spectral gap κ for LD satisfies the following bounds:*

1. *If $\pi(x) \propto \phi(x/\epsilon)$, then $\kappa \leq \epsilon^2$.*
2. *If $\pi(x) \propto \frac{1}{2}\phi(x/\epsilon)+\frac{1}{2}\phi((x-m)/\epsilon)$ and $\epsilon \leq \frac{\|m\|}{16\sqrt{d}}$, then $\kappa \geq \frac{\epsilon^4}{80\|m\|^2}\exp\left(\frac{\|m\|^2}{64\epsilon^2}\right)$.*

The proof of Proposition 1 can be found in Appendix A (The first scenario is well known). Proposition 1 indicates that one of the most challenging types of densities for LD to sample is mixtures of "well-separated" singular densities, even if each of them is Gaussian. When sampling a single Gaussian using LD, the spectral gap is lower bounded by ϵ^{-2}. In this case, a smaller value of ϵ leads to a faster convergence rate. However, when sampling a mixture of two such Gaussians with well-separated modes, the convergence can be very slow for small values of ϵ. In particular, the spectral gap of LD is upper bounded by $80\|m\|^2\epsilon^{-4}\exp(-\|m\|^2/(64\epsilon^2))$, which is extremely small when ϵ is small.

1.2. Replica exchange Langevin diffusion

We next introduce the replica exchange method by considering the scenario where there are two replicas. The first one $X(t)$ is defined in (1) and the second one $Y(t)$ has a stronger stochastic force:

$$dY(t) = \nabla \log \pi(Y(t))dt - \tau Y(t)/M^2 dt + \sqrt{2\tau}dW(t), \tag{3}$$

where $W(t)$ is a d-dimensional Brownian motion, independent of $B(t)$ in (1), and τ is a parameter known as the temperature. M is a large number so that the local minima of $H(x)$ satisfy $\max_{1\leq i\leq I}\|m_i\| \leq M$. The stationary distribution of $Y(t)$ takes the form

$$\pi^Y(y) \propto \exp\left(-\frac{1}{\tau}H(y) - \frac{\|y\|^2}{2M^2}\right).$$

When τ is selected to be a large number, the effective Hamiltonian of $Y(t)$ is approximately $\tau^{-1}H(y)$, which has the same local minima as $H(x)$, but the height of the potential wells are only $1/\tau$ of the latter. Thus, it is easier for $Y(t)$ to climb out of potential wells and visit other local minima.

Even though $Y(t)$ is not sampling the target density π, it can be used to help $X(t)$ sample π more efficiently. To do so, let $\rho > 0$ denote a swapping intensity, so that sequential swapping events take place according to an independent exponential clock with rate ρ. At a swapping event time t, $X(t)$ and $Y(t)$ swap their positions (values) with probability $s(X(t), Y(t))$, where

$$s(x, y) = 1 \wedge \frac{\pi(y)\pi^Y(x)}{\pi(x)\pi^Y(y)}. \tag{4}$$

We refer to the joint process $(X(t), Y(t))$ as ReLD. It can be verified that $\pi \otimes \pi^Y$ is the invariant distribution of ReLD under mild ergodicity conditions [7].

Exchanging $X(t)$ with $Y(t)$ can improve the convergence rate of $X(t)$. We demonstrate the basic idea through Figure 1. As mentioned above, the main reason why sampling directly from LD can be slow for multimodal π is that $X(t)$ can be trapped in a potential well for a long time. In Figure 1, suppose $X(t)$ is currently in $B(m_1, r)$, which is a ball of radius r centered at the mode m_1. In order for $X(t)$ to visit a different mode m_2, it needs to visit the boundary of the potential well, i.e., the origin, and this can take a long time. On the other hand, it is much easier for $Y(t)$ to cross the potential wells. In particular, $Y(t)$ can move "freely" in a larger region demonstrated as $B(0, R)$ in Figure 1, which includes all the local minima. The exchange mechanism (4) swaps $X(t)$ and $Y(t)$ with a decent chance if $Y(t)$ is in a different "high-probability" area for $X(t)$, say $B(m_2, r)$. This helps $X(t)$ visit the other potential well, which effectively improves the convergence rate of $X(t)$. Our main objective in this paper is to translate these intuitions into mathematically rigorous statements.

One major issue with ReLD introduced above is that the exchanges may not happen often enough. To see this, note that in Figure 1, when $X(t)$ is near the first mode m_1, the exchange probability (4) can be very small unless $Y(t)$ is in "high-probability" areas $B(m_1, r)$ or $B(m_2, r)$ as well. But since $Y(t)$ is circling inside a large area $B(0, R)$, the chance that it is in $B(m_1, r)$ or $B(m_2, r)$ can be small if $r \ll R$. To amend this issue, we can simulate multiple parallel LDs with an increasing sequence of temperatures. Then, we exchange the positions of neighboring replicas. The above sampling scheme is referred to as mReLD.

Adding intermediate temperatures improves the small exchange probability issue mentioned earlier. We illustrate the basic idea through Figure 2, where we run three parallel LDs. The "high-probability" areas for $X_0(t), X_1(t)$, and $X_2(t)$ are $B(m_1, r_0) \cup B(m_2, r_0)$, $B(m_1, r_1) \cup B(m_2, r_1)$, and $B(0, r_2)$ respectively. We note that $r_0 < r_1 < r_2$. The exchange between $X_0(t)$ and $X_2(t)$ may not happen often, since $X_2(t)$ has only a small chance of being inside $B(m_1, r_0) \cup B(m_2, r_0)$. On the other hand, $X_1(t)$ stays mostly inside $B(m_1, r_1) \cup B(m_2, r_1)$, and thus has a better chance of being inside $B(m_1, r_0) \cup B(m_2, r_0)$ than $X_2(t)$. Hence $X_1(t)$ can exchange with $X_0(t)$ more often. From this discussion, we see that adding additional replicas, for which the neighboring replicas share similar potential functions, improves the chance of successful exchanges. Meanwhile, exchanges between non-adjacent replica are unlikely to happen, so we decide to exclude them in our design of mReLD. In particular, we consider $K + 1$ LDs

$$dX_k(t) = \tau_k \nabla \log \pi_k(X_k(t))dt + \sqrt{2\tau_k} dW_k(t), \quad k = 0, \dots, K$$

with $1 = \tau_0 \leq \tau_1 \leq \cdots \leq \tau_K$ and $\pi_0 = \pi$. Exchange between two adjacent levels takes place according to independent exponential clocks with rate ρ. At a swapping epoch t for the pair $(k, k+1)$, $X_k(t)$ and $X_{k+1}(t)$ exchange their positions (values) with probability $s_k(X_k(t), X_{k+1}(t))$, where

$$s_k(x_k, x_{k+1}) = 1 \wedge \frac{\pi_k(x_{k+1})\pi_{k+1}(x_k)}{\pi_k(x_k)\pi_{k+1}(x_{k+1})}. \tag{5}$$

We next show that by properly choosing the temperature and the swapping intensity in ReLD and mReLD, we can substantially improve the convergence

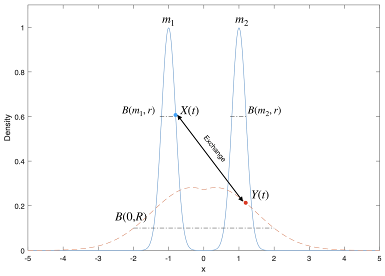

Figure 1: ReLD: The (blue) solid line plots the density function of a bi-modal density π. The (red) dashed line plots the tempered density function π^Y.

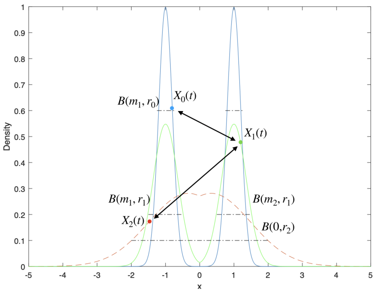

Figure 2: mReLD: The tall (blue) solid, short (green) solid, and (red) dashed lines plot the density functions of π_0, π_1, π_2 respectively.

rate for Gaussian mixtures (including scenario 2 in Proposition 1). To highlight the challenge in sampling efficiency, we focus on the dependence of the inverse spectral gap on the parameter ϵ (i.e., the depth of the potential well), while keeping all other model parameters fixed.

Theorem 1 *Suppose the target density is a mixture of isotropic Gaussian distributions:* $\pi(x) \propto \sum_{i=1}^{I} p_i \phi\left(\frac{x-m_i}{\epsilon}\right)$*, where* $\max_{1\leq i\leq I} \|m_i\| \leq M$ *for some constant* $M < \infty$*. For ReLD with* $\tau, \rho \propto \epsilon^{-d}$*, the inverse spectral gap,* $\kappa = O(1)$*, i.e., is independent of* ϵ*. For mReLD, there exists a sequence of* π_k*'s such that for* $\tau_k = \epsilon^{-\frac{d}{K}}$*,* $k = 1, \ldots, K$*, and* $\rho = \epsilon^{-d/K}$*, the inverse spectral gap,* $\kappa = O(1)$*.*

In this section, we choose the Gaussian mixture due to its simplicity for demonstration. In Section 2, we study the convergence rate of ReLD and mReLD for mixtures of more general distributions. In particular, Theorem 1 is a special case of Theorems 2 and 3 (see Corollaries 1 and 2 for more details).

1.3. Literature review and our contribution

Most standard Markov Chain Monte Carlo (MCMC) methods suffer from a slow convergence rate when the target distribution has multiple isolated modes, i.e., multimodal. Replica exchange Monte Carlo (ReMC), which is also known as parallel tempering, has been proposed to speed up the convergence and has seen promising performance in molecular dynamics and statistical mechanics (see, for example, [3, 4, 2, 8, 9, 5]). In recent year, it has also been applied to machine learning, such as training restricted Boltzmann machines [10, 11, 12]; and solving non-convex optimization problems [13, 14]. There is also a growing interest in designing new ReMC algorithms for improved performance [15, 16, 17, 18], but there are very few existing works analyzing the convergence rate of ReMC. The closest to our work is [19], which establishes an upper bound for the inverse spectral gap of replica exchange samplers. While their bound can be applied to more general samplers than ours, such flexibility comes at a cost of tractability. In particular, to calculate their bound, one needs to design an appropriate partition of the state space and samplers that converge fast on the partition, which can be highly nontrivial. In contrast, ReLD focuses on LDs, which can be seen as concrete samplers. Our results also provide more explicit bounds and there is no need to design the partition to implement the algorithm. Efficiency of ReLD or similar versions of it is also studied in [13, 14]. The work [13] analyzes the spectral gap but does not provide an explicit quantification of the "speed-up" due to swapping. In this work, we are able to quantify the speed-up effect by developing a novel bound for the mean-difference estimates. Focusing on solving non-convex optimization problems, [14] considers a different performance metric than the spectral gap. However, our refined spectral gap bounds can be applied to their setting to quantify the benefit of adding extra replicas and guide related parameter tuning.

A key question in implementation of ReLD or ReMC is how to set/tune the temperature and the swapping rate. Most previous investigations rely on extensive simulation experiments and heuristic arguments [20, 21, 22]. The work [7] uses the large deviation theory to define a rate of convergence for the empirical measure of ReLD. It shows that the rate increases with the swapping intensity ρ. Thus, an infinite swapping algorithm (ISA) is proposed. A more detailed large deviation analysis of ISA is provided in [23]. Similar to our work, [24] studies the ergodicity properties of ISA and derive bounds for the Poincaré inequality constant. Recently, a series of works provide rigorous analysis on how to tune the temperatures to achieve an asymptotically optimal exchange probability in the high dimensional limit [15, 16, 18]. Similar to [19], these analyses assume the existence of some exact samplers of the target distributions and focus mostly on the equilibrium behavior. In contrast, our spectral gap analysis focuses on concrete diffusion processes and characterizes the non-equilibrium behavior.

A similar but slightly different sampling idea to ReLD is simulated tempering, which considers dynamically changing the temperature of LD [25]. Several tempering-based MCMC methods have been studied in the literature, including annealing MCMC [26], tempered transition method [27], etc. Like ReLD, there are very few theoretical results about its efficiency. The work [28] develops lower bounds for the spectral gap of general simulated tempering chains, but the bounds are too loose to provide concrete guidance on how to choose the hyperparameters. Recently, [29] establishes a tighter bound for simulated tempering LD. Their analysis specifics how to set the temperatures in the setting where the target distribution is a mixture of log-concave densities with different modes but the same shape. In contrast, our results allow the mixture components to be of different shapes. One main challenge in implementing simulated tempering is that one needs to estimate the normalizing constants of the target distributions. In contrast, replica exchange avoids the need to deal with these normalizing constants, as they are cancelled out in the exchange probabilities.

1.4. Organization and notation

The rest of the paper is organized as follow. In Section 2, we present the main results, Theorems 2 and 3, which provide estimates on the inverse spectral gap for ReLD and mReLD respectively. In Section 3, we demonstrate how to apply our results to mixtures of log-concave densities and the connection between mixture models and the Morse function assumption in [1]. The proof of the main results (Theorems 2 and 3) are provided in Section 4. To keep the discussion concise, all the technical results are proved in the appendices.

Given two vectors $u, v \in \mathbb{R}^d$, we use $\|v\|$ to denote the l_2 norm of v, and $\langle v, u\rangle := u^T v$. Given a matrix A, we use $\|A\|$ to denote its l_2-operator norm. For any $f \in \mathbb{C}^2(\mathbb{R}^d)$, i.e., twice continuously differentiable functions, we use $\nabla f \in \mathbb{R}^d$ to denote its gradient, $\nabla^2 f \in \mathbb{R}^{d\times d}$ to denote its Hessian, and $\Delta f := \mathrm{tr}(\nabla^2 f)$. We also denote $B(x_0, R)$ as a ball with center x_0 and radius R.

When a distribution π is given, we use $\mathbb{E}_\pi f$ and $\mathrm{var}_\pi(f)$ to denote the mean and variance of f under π. For two distributions π and ν on $\mathbb{R}^d$, we write their product measure on $\mathbb{R}^{2d}$ as $\pi \otimes \nu$. Since we consider mostly diffusion-type of stochastic processes, it is reasonable to assume the associate distributions are absolutely continuous with respect to the Lebesgue measure. When we refer to a distribution π, we assume it has a probability density function $\pi(x)$. Then, we can use $\pi(x)/\nu(x)$ to denote the Radon-Nikodym derivative between π and ν. We define $0/0 \equiv 0$.

Our goal is to develop a proper upper bound for the inverse spectral gap κ, which can be translated to a lower bound for the spectral gap. As the underlying distribution/process may involve several parameters, e.g., ϵ, d, M in the Gaussian mixture example and τ, ρ for ReLD, the exact characterization of the upper bound can get quite involved. Therefore, we adopt the O notation. For a nonnegative function f and a sequence of non-negative quantities A_ϵ indexed by ϵ, we write $A_\epsilon = O(f(\epsilon))$ if there is a constant $C > 0$ independent of ϵ, such that $A_\epsilon \leq Cf(\epsilon)$. $A_\epsilon = O(1)$ means $A_\epsilon \leq C$. We also write $A_\epsilon = \Omega(f(\epsilon))$ if there is a constant $C > 0$ independent of ϵ, such that $A_\epsilon \geq Cf(\epsilon)$. Our goal is to quantify the dependence of κ on ϵ, τ, and ρ.

2. General problem setup and results

We introduce the general setup of ReLD and study its performance when the target distribution is of certain mixture type in this section. Our development relies on applications of Poincaré inequality (PI). Therefore, we start by introducing some basic properties of PI. We then discuss general assumptions for the type of density mixtures that our framework can handle. The main results are presented in Theorems 2 and 3 in Sections 2.3 and 2.4 respectively.

2.1. Poincaré inequality and Lyapunov function

Recall that the basic LD is given by

$$dX(t) = \nabla \log \pi(X(t))dt + \sqrt{2}dB(t).$$

We denote $\mathcal{L}_\pi$ as it generator, which takes the following form:

$$\mathcal{L}_\pi(f) = \langle \nabla f, \nabla \log \pi \rangle + \Delta f,$$

for $f \in \mathbb{C}_c^2(\mathbb{R}^d)$. Then, the associated carré du champ takes the form $\Gamma(f) = \|\nabla f\|^2$. The inverse spectral gap κ in (2) can also be viewed as the coefficient in the PI, which is often refer to as the *PI constant*.

Definition 1 *A density π follows κ-PI if the following holds*

$$var_\pi(f) \leq \kappa \int \|\nabla f(x)\|^2 \pi(x)dx, \quad \forall f \in \mathbb{C}_c^2(\mathbb{R}^d).$$

We next review some existing results of PI.

Proposition 2 (Holley–Stroock perturbation principle) *Suppose for some operator Γ and density π, $var_\pi(f) \leq \kappa \int \Gamma(f)(x)\pi(x)dx$ for all $f \in \mathbb{C}_c^2(\mathbb{R}^d)$. Moreover, suppose there exists a constant $C \in (0, \infty)$ such that $C^{-1} \leq \pi(x)/\mu(x) \leq C$ for all x. Then, $var_\mu(f) \leq C^2\kappa \int \Gamma(f)(x)\mu(x)dx$ for all $f \in \mathbb{C}_c^2(\mathbb{R}^d)$.*

Proposition 2 indicates that if a density π follows a κ-PI, then a mild perturbation of π also follows a PI. Note that Γ here can be the carré du champ of LD, but it can also be the carré du champ of ReLD.

The next result connects the Lyapunov function to the PI constant. The connection was first established in [30]. Here, we present a slightly different version of it.

Definition 2 *A $\mathbb{C}^2$ function $V(x) : \mathbb{R}^d \to [1, \infty)$ is a (λ, h, B, C)-Lyapunov function for a density $\nu(x)$ if the following holds*

$$\mathcal{L}_\nu V(x) \leq -\lambda V(x) + h1_B(x), \quad \frac{\sup_{x\in B} \nu(x)}{\inf_{x\in B} \nu(x)} \leq C,$$

where $\lambda, h, C \in (0, \infty)$ are positive constants and $B \subset \mathbb{R}^d$ is a bounded domain.

Proposition 3 *Suppose ν has a $(\lambda, h, B(x_0, R), C)$-Lyapunov function. Then,*

$$var_\nu(f) \leq \frac{1 + hR^2C^2}{\lambda}\mathbb{E}_\nu[\|\nabla f(X)\|^2].$$

Proposition 3 provides a convenient way to compute (upper bound) the PI constant for a given density ν. Based on Proposition 3, we define the following notion of a density:

Definition 3 *We say ν is an $\mathbf{Ly}(R, q, a)$-density with the center x_0, if it has a $(\lambda, h, B(x_0, R), C)$-Lyapunov function, with*

$$\frac{1 + hR^2C^2}{\lambda} \leq q \quad \textit{and} \quad \sup_{x \in B(x_0,R)} \frac{u_{B(x_0,R)}(x)}{\nu(x)} \leq a,$$

where $u_{B(x_0,R)}$ denotes the uniform distribution on $B(x_0, R)$.

Remark 1 *In our main theoretical development, we will consider replacing ν with $u_{B(x_0,R)}$, since the latter is easier to handle. The constant a in the $\mathbf{Ly}(q, R, a)$-density roughly measures how well the uniform approximation is.*

2.2. Mixture Density

As discussed in Section 1, we are interested in understanding how replica exchange improves the convergence of LD on a multimodal target density. Multimodal densities often arise from mixture models:

$$\pi(x) = \sum_{i=1}^{I} p_i \nu_i(x), \tag{6}$$

where $p_i \geq 0$ with $\sum_{i=1}^{I} p_i = 1$, and each ν_i has a single mode m_i.

We next discuss what kind of mixture model would allow a replica exchange process $(X(t), Y(t))$ to sample efficiently. First, each ν_i should be “easy” for an LD of the form (1) to sample directly, since the exchange mechanism can only help $X(t)$ visiting different modes but not sampling an individual ν_i faster. This requirement can be formulated through the existence of an appropriate Lyapunov function for ν_i based on Proposition 3:

Assumption 1 *There are positive constants r_i, q, a such that for $i = 1, \ldots, I$, ν_i is an $\mathbf{Ly}(r_i, q, a)$-density with the center m_i.*

We will show in Propostion 4 that log-concave densities satisfy Assumption 1.

Second, $Y(t)$ should be able to visit different m_i's "easily". Otherwise, it cannot help $X(t)$ reach certain modes. This requirement can be formulated as requiring that m_i's are not too far from each other. Since our problem is shift invariant, this is equivalent to assuming that there exists a constant $M < \infty$ such that $\max_{1\leq i\leq I} \|m_i\| \leq M$. In particular, M does not depend on d or q.

Remark 2 *It is worth mentioning that [1] imposes different assumptions on the Hamiltonian $H(x)$. In particular, it assumes $H(x)$ is a Morse function and there is an admissible partition so that a proper Lyapunov function exists within each partition. Admittedly, this might be a more general assumption, since not all densities can be written as a mixture (6). However, this set of assumptions requires more technical definitions and verification. Moreover, it can be shown that under mild conditions, the setting in [1] can be converted to a mixture. We will provide more details of the connection in Section 3.3.*

2.3. Spectral Gap for ReLD

We next formulate a general ReLD. First, pick a density π^Y and consider the following two LDs driven by independent d-dimensional Browian motions $W^x(t)$ and $W^y(t)$:

$$\begin{aligned} dX(t) &= \nabla \log \pi(X(t))dt + \sqrt{2}dW^x(t), \\ dY(t) &= \tau\nabla \log \pi^Y(Y(t))dt + \sqrt{2\tau}dW^y(t). \end{aligned} \tag{7}$$

Swapping epochs are generated by an independent exponential clock with rate ρ. At a swapping epoch t, we swap the positions of $X(t)$ and $Y(t)$ with probability $s(X(t), Y(t))$, where s is defined in (4). It is easy to see that the ReLD discussed in Section 1 is a special case of (7) with $\pi^Y(y) = \exp(-\frac{1}{\tau}H(y) - \frac{\|y\|^2}{2M^2})$.

We consider a general π^Y here for two reasons. First, as we will discuss in Section 2.3, the temperature τ is often "required" to be a large number. Then, direct simulation of $Y(t)$ with the Euler-Maruyama scheme would require a very small stepsize. If π^Y is a simple density, for example, a Gaussian density, we can have direct access to the transition kernel of $Y(t)$ and avoid using any discretization scheme. Second, it is easier to impose requirements on π^Y for the replica exchange process to achieve good convergence rate. For a mixture-type target distribution as in (6), we impose the following assumption on π^Y:

Assumption 2 *There are constants (R_i, Q, A) so that for each mode m_i, π^Y is an $\mathbf{Ly}(R_i, Q, A)$-density with center m_i, $i = 1, \dots, I$.*

We will show in Proposition 6 that many forms of π^Y satisfy Assumption 2.

The generator of ReLD, denoted by $\mathcal{L}_R$, is then given by

$$\begin{aligned}\mathcal{L}_R f(x,y) &= \lim_{t\to 0} \frac{1}{t}\mathbb{E}[f(X_t, Y_t) - f(x,y) | X_0 = x, Y_0 = y]\\ &= \mathcal{L}_x f(x,y) + \tau\mathcal{L}_y f(x,y) + \rho s(x,y)(f(y,x) - f(x,y)),\end{aligned}$$

for $f \in \mathbb{C}_c^2(\mathbb{R}^{2d})$, where $\mathcal{L}_x f(x,y) := \langle \nabla_x f(x,y), \nabla_x \log \pi(x)\rangle + \Delta_x f(x,y)$ and $\mathcal{L}_y f(x,y) := \langle \nabla_y f(x,y), \nabla_y \log \pi^Y(y)\rangle + \Delta_y f(x,y)$. It is easy to verify that $\pi \otimes \pi^Y$ is an invariant measure for ReLD. In particular, $\mathbb{E}_{\pi\otimes\pi^Y}\mathcal{L}_R f = 0$. The associated carré du champ for ReLD is given by

$$\begin{aligned}\Gamma_R f(x,y) &= \frac{1}{2}(\mathcal{L}_R(f^2) - 2f\mathcal{L}_R(f))\\ &= \|\nabla_x f(x,y)\|^2 + \tau\|\nabla_y f(x,y)\|^2 + \frac{1}{2}\rho s(x,y)(f(y,x) - f(x,y))^2\end{aligned} \tag{8}$$

for $f \in \mathbb{C}_c^2(\mathbb{R}^{2d})$. Note that if we simply simulate $X(t)$ and $Y(t)$ according to (7) without the exchange, the carre du champ will be $\|\nabla_x f(x,y)\|^2 + \tau\|\nabla_y f(x,y)\|^2$. The exchange mechanism contributes to the additional positive term $\frac{1}{2}\rho s(x,y)(f(y,x) - f(x,y))^2$ in Γ_R. While this helps lowering the inverse spectral gap κ in (2), the extent of improvement is far from obvious.

We next quantify the effect of the exchange mechanism on the spectral gap. In addition to Assumptions 1 and 2, we also impose the following assumption:

Assumption 3 *There are $r > 0$ and $R > 0$, such that the constants R_i, $i = 1, \dots, I$, and r_i, $i = 1 \dots, I$, from Assumptions 1 and 2 satisfy $R_i \le R$ and $\frac{R_i}{r_i} \le \frac{R}{r}$ for all $1 \le i \le I$.*

Theorem 2 *For ReLD defined in (7), under Assumptions 1, 2, and 3,*

$$var_{\pi\otimes\pi^Y}(f(X,Y)) \leq \kappa \mathbb{E}_{\pi\otimes\pi^Y}[\Gamma_R(f(X,Y))],$$

for all $f \in \mathbb{C}_c^2(\mathbb{R}^{2d})$, where

$$\kappa = \max\left\{3(56A+1)q,\ \frac{3}{\tau}\left(57Q + 14aA\left(\frac{R^{d+1}}{r^{d-1}}\right)\left(\log\left(\frac{R}{r}\right)\right)^{1_{d=1}}\right), \frac{7aA}{\rho}\left(\frac{R}{r}\right)^d\right\}.$$

In particular, if R, A, Q, a are $O(1)$ constants, then $\kappa = O\left(q + \left(\frac{1}{\tau} + \frac{1}{\rho}\right)\frac{1}{r^d}\right)$. When $q < 1$, if we set $\tau, \rho \geq Uq^{-\alpha}r^{-d}$ for any $\alpha \leq 1$ and $U > 0$, then $\kappa = O(U^{-1}q^{\alpha})$.

For mixture of singular densities with isolated modes, r and q can be very small. For example, as we will explain in more details in Section 3.1, $r^2, q = \Theta(\epsilon^2)$ for the Gaussian mixture model in Proposition 1. If we choose $\tau, \rho \geq r^{-d}$, then $\kappa = O(1)$, i.e., it does not depend on r or q. If we choose $\tau, \rho \geq q^{-1}r^{-d}$, then $\kappa = O(q)$. In this case, the spectral gap is of the same order as the smallest spectral gap of the component densities in the mixture.

2.4. Spectral Gap for mReLD

Considering $K+1$ LDs

$$dX_i(t) = \tau_i \nabla \log \pi_i(X_i(t))dt + \sqrt{2\tau_i}dW_i(t), \quad i = 0, \ldots, K \tag{9}$$

with $1 = \tau_0 \leq \tau_1 \leq \cdots \leq \tau_K$ and $\pi_0 = \pi$. Exchange between two adjacent levels takes place according to independent exponential clocks with rate ρ. At a swapping epoch t for the pair $(k, k+1)$, $k = 0, \ldots, K-1$, $X_k(t)$ and $X_{k+1}(t)$ exchange their positions with probability $s_k(X_k(t), X_{k+1}(t))$, which is defined in (5). Let $\mathbf{x}_{k:l} = (x_k, \ldots, x_l)$ and $\pi_{k:l} = \pi_k \otimes \cdots \otimes \pi_l$. Note that each $x_k \in \mathbb{R}^d$ for $k = 0, 1, \ldots, K$. The generator of mReLD takes the form: for $f \in \mathbb{C}_c^2(\mathbb{R}^{d(K+1)})$,

$$\begin{aligned}\mathcal{L}_R^K(f(\mathbf{x}_{0:k})) = &\sum_{k=0}^{K}(\tau_k\langle\nabla_{x_k}f(\mathbf{x}_{0:k}), \nabla\log\pi_k(\mathbf{x}_{0:k}))\rangle + \tau_k\Delta_{x_k}f(\mathbf{x}_{0:k}))\\ &+ \sum_{k=0}^{K}\rho s_k(x_k, x_{k+1})(f(\mathbf{x}_{0:K}) - f(\mathbf{x}_{0:(k-1)}, x_{k+1}, x_k, \mathbf{x}_{(k+2):K})).\end{aligned}$$

The corresponding carré du champ and Dirichlet from are

$$\begin{aligned}\Gamma_R^K(f(\mathbf{x}_{0:K})) :=& \sum_{k=0}^{K} \tau_k \|\nabla_{x_k} f(\mathbf{x}_{0:K})\|^2 \\ &+ \sum_{k=0}^{K} \rho s_k(x_k, x_{k+1})(f(\mathbf{x}_{0:K}) - f(\mathbf{x}_{0:(k-1)}, x_{k+1}, x_k, \mathbf{x}_{(k+2):K}))^2\end{aligned}$$

and $\mathcal{E}_R^K(f) = \int \Gamma_R^K(f)\pi_{0:K}(d\mathbf{x}_{0:K})$ respectively for $f \in \mathbb{C}_c^2(\mathbb{R}^{d(K+1)})$.

We make the following assumptions about π_k's.

Assumption 4 *There are positive constants $q_k, r_{k,i}, a_k$ for $k = 0, \dots, K$, $i = 1, \dots, I$, such that: 1) $\pi_k(x) = \sum_{i=1}^{I} p_i \nu_{k,i}(x)$, where $\nu_{k,i}$ is an $\mathbf{Ly}(r_{k,i}, q_k, a_k)$-density with center m_i; 2) For each m_i, π_K is an $\mathbf{Ly}(r_{K,i}, q_K, a_K)$-density with center m_i.*

Assumption 5 *There is an increasing sequence $0 < r_0 < r_1 < \cdots < r_K$, such that the constants $r_{k,i}$, $k = 0, \dots, K$, $i = 1, \dots, I$, from Assumption 4 satisfy $r_{k+1,i}/r_{k,i} \le r_{k+1}/r_k$, for all $0 \le k \le K-1$ and $1 \le i \le I$, and $r_{K,i} \le r_K$, for all $1 \le i \le I$.*

Theorem 3 *For mReLD defined in (9), suppose Assumptions 4 and 5 hold, and K, q_k, a_k, r_k, $k = 1, \dots, K$, are all $O(1)$ constants. Then,*

$$var_{\pi_{0:K}}(f(\mathbf{X}_{0:K})) \le \kappa \mathbb{E}_{\pi_{0:K}}[\Gamma_R^K(f(\mathbf{X}_{0:K}))], \tag{10}$$

for all $f \in \mathbb{C}_c^2(\mathbb{R}^{d(K+1)})$, where

$$\kappa = O\left(\max\left\{\frac{q_0}{\tau_0}, \left(\frac{1}{\tau_k} + \frac{1}{\rho}\right)\left(\frac{r_k}{r_{k-1}}\right)^d, 1 \le k \le K\right\}\right). \tag{11}$$

In particular, when $q_0 < 1$, for any $\alpha \le 1, U > 0$, if we choose $\tau_k \ge U\left(\frac{1}{q_0}\right)^\alpha \left(\frac{r_k}{r_{k-1}}\right)^d$ and $\rho \ge U \max_{1 \le k \le K} \left(\frac{1}{q_0}\right)^\alpha \left(\frac{r_k}{r_{k-1}}\right)^d$, then $\kappa = O(U^{-1} q_0^\alpha)$.

The exact estimate of κ is quite complicated. We provide the explicit expression in Theorem 4. In Theorem 3, we assume K, q_k, a_k, r_k, $1 \le k \le K$, are all fixed $O(1)$ constants to simplify the estimate. For mixture models with small values of r_0 and q_0, if we can construct π_k's such that $r_k/r_{k-1} = \Theta\big((r_K/r_0)^{1/K} \big)$ for $k = 1, \ldots, K$, we can set $\tau_k, \rho \geq (r_K/r_0)^{d/K}$ to achieve $\kappa = O(1)$. If we further enlarge $\tau_k, \rho \geq q_0^{-1} (r_K/r_0)^{d/K}$, then $\kappa = O(q_0)$. In this case, the spectral gap matches the smallest spectral gap of the component densities in the mixture.

3. Applying replica-exchange to different densities

In this section, we investigate ReLD for some specific examples. We first present some general properties of log-concave densities. In particular, we show they are $\mathbf{Ly}(R, q, a)$-densities (Definition 3).

Definition 4 *A density ν is a (c, L)-log-concave density if $H = -\log \nu$ is $\mathbb{C}^2$ and $\langle \nabla H(x) - \nabla H(y), x - y\rangle \geq c\|x-y\|^2$, $\|\nabla^2 H(x)\| \leq L, \forall x, y$.*

Proposition 4 *If ν is (c, L)-log-concave and m is its mode, then $V(x) = \frac{c}{d}\|x-m\|^2 + 1$ is a (λ, h, B, C)-Lyapunov function of ν with $\lambda = c$, $h = 3c$, $B = B\left(m, \sqrt{\frac{3d}{c}}\right)$,and $C = \exp\left(\frac{3dL}{2c}\right)$. This implies that ν is a $\mathbf{Ly}(r, q, a)$-density with $q = c^{-1} + \frac{9d}{c}\exp\left(\frac{3dL}{c}\right)$, $r = \sqrt{\frac{3d}{c}}$, and $a = \frac{1}{V_d}\exp\left(\frac{3Ld}{2c}\right)\left(\frac{4\pi}{3d}\right)^{d/2}$, where V_d denotes the volume of a d-dimensional ball with unit radius.*

We also provide a bound for a in the $\mathbf{Ly}(R, q, a)$-density, based on a specific form of the $(\lambda, h, B(x_0, R), C)$-Lyapunov function.

Proposition 5 *Suppose ν has a $(\lambda, h, B(x_0, R), C)$-Lyapunov function of form $V(x) = \gamma\|x-x_0\|^2+1$. Then ν is a $\mathbf{Ly}(R, q, a)$-density with $a = \frac{C}{V_d}\exp\left(\frac{1}{4}\lambda R^2\right)\left(\frac{4\pi}{\lambda R^2}\right)^{d/2}$.*

We next provide some specific forms of π^Y that satisfies Assumption 2.

Proposition 6 *Assume $\max_{1\leq i\leq I}\|m_i\| \leq M$ for some $M < \infty$.*

1. *If $\pi^Y(x) \propto \phi(x/M)$, then Assumption 2 holds with $R^2 = O(M^2 d)$, $Q = O(M^2 d\exp(12d))$, and $A = O(\exp(6d))$.*
2. *If $\pi^Y(x) \propto \phi(x/M)\pi(x)^\beta$, $\nu_i(x)$'s are (c, L)-log concave densities, and $\beta \leq (dM^2c + dM^2L^2/c)^{-1}$, then Assumption 2 holds with $R^2 = O(M^2 d)$, $Q = O(M^2 d\exp(20d))$, and $A = O(\exp(12d))$.*

Proposition 6 indicates that Assumption 2 is similar to requiring all modes, m_i's, being bounded by a constant that does not depend on d.

3.1. ReLD for mixture of log-concave densities

A general Gaussian mixture model can be written as $\pi(x) = \sum_{i=1}^{I} p_i \nu_i(x)$, where $\nu_i(x) = \frac{1}{\sqrt{\det(2\pi\Sigma_i)}} \exp(-\frac{1}{2}(x-m_i)^T \Sigma_i^{-1}(x-m_i))$. Suppose $\mathcal{C}^{-1} l_i^2 \preceq \Sigma_i \preceq l_i^2$, $l_m \leq l_i \leq l_M \leq 1$, where $\mathcal{C}$ is known as the condition number. Then, ν_i is $(l_i^{-2}, \mathcal{C} l_i^{-2})$-log-concave. For general $(l_i^{-2}, \mathcal{C} l_i^{-2})$-log-concave densities, we have the following result:

Corollary 1 *Suppose $\pi = \sum_{i=1}^{I} p_i \nu_i$ where ν_i's are $(l_i^{-2}, l_i^{-2}\mathcal{C})$-log concave densities with modes m_i's, and $\|m_i\| \leq M$. Let $l_m = \min_i l_i$, $l_M = \max_i l_i$, and $\tau \geq dM^2 l_M^{-2} + dM^2 l_M^2 l_m^{-4} \mathcal{C}^2$. Then, $var_{\pi \otimes \pi^Y} f \leq \kappa \mathbb{E}_{\pi \otimes \pi^Y} \Gamma_R(f)$ holds with*

$$\kappa = O\left(\exp(\mathcal{C}Dd) \max\left\{ dl_M^2, \frac{1}{\tau} dM l_m \left(\frac{M}{l_m}\right)^d, \frac{1}{\rho}\left(\frac{M}{l_m}\right)^d \right\}\right),$$

where D is a fixed constant.

We next provide some interpretations of Corollary 1. As κ is the inverse spectral gap, we refer to $1/\kappa$ as the convergence rate. First, consider the Gaussian mixture model in scenario 2 of Proposition 1 where $l_m^2 = l_M^2 = \epsilon^2$ and $\mathcal{C} = 1$. By choosing $\tau, \rho = \Omega(\epsilon^{-d-2})$, $\beta = \tau^{-1} \leq \epsilon^2$ and $\kappa = O(\epsilon^2)$. This matches the convergence rate of LD when $\pi(x) \propto \phi(x/\epsilon)$. We can also set $\tau, \rho = \Theta(\epsilon^{-d})$, which leads to $\kappa = O(1)$. In addition, our result allows the Gaussian components to be of different scales. For example, $l_1^2 = l_m^2 = \epsilon^2$ and $l_2^2 = l_M^2 = \epsilon$. In this case, if $\tau = \Omega(\max\{\epsilon^{-d}, \epsilon^{-3}\})$ and $\rho = \Omega(\epsilon^{-d-1})$, $\beta = \tau^{-1} \leq \epsilon^{-3}$ and $\kappa = O(l_M^2) = O(\epsilon)$. This matches the convergence rate of LD for $\pi = \nu_2$.

In general, for fixed values of d and $\mathcal{C}$, τ and ρ need to scale as $(M/l_m)^d$ for the convergence rate to be of a constant order. To see the intuition behind this, note that with a high temperature, $Y(t)$ can be seen as a random search in the set $\{\|x\| \leq M\}$ with speed τ. At any time t, the chance that it is in a radius-l_m neighborhood of a mode m_i is $(l_m/M)^d$. Thus, to have a constant convergence rate, it is necessary for Y_t to run at a speed $\tau = \Theta((M/l_m)^d)$. Meanwhile, ρ

is rate of checking whether the exchange takes place, and it needs to be of the same scale as τ.

In implementations, when applying discretization schemes like Euler-Maruyama for ReLD, the step size often needs to scale as $\min\{1/\tau, 1/\rho\}$. If $M = O(1)$ and $l_m = O(\epsilon)$, the computational cost of ReLD is roughly $O(\epsilon^{-d})$. While this can be quite high, it is much better than the computational cost of using LD alone, which is roughly $O(\exp(D\epsilon^{-2}))$ as shown in Proposition 1. When taking computational cost into account, it is of practical interest to further reduce τ and ρ, which can be achieved by mReLD.

3.2. mReLD for mixture of log-concave densities

In this section, we demonstrate how the mReLD result applies to the mixture models discussed in Section 3.1. Following the practical choice in MD simulation, we assume the invariant measure for $X_k(t)$ takes the form

$$\pi_k(x) \propto (\pi(x))^{\beta_k}, \quad k = 0, 1, \ldots, K-1$$

for some inverse temperature $\beta_k \in [0, 1]$. Note that this choice makes the drift term of $X_k(t)$ being a multiple of $\nabla \log(\pi(X_k(t)))$, which is generally accessible.

When the target distribution is a mixture of log concave densities, our characterization of the spectral gap depends on whether we need to synchronize τ_k with β_k. In particular, if the speed of simulation for $X_k(t)$, which is described by τ_k, does not need to match the temperature $\frac{1}{\beta_k}$, then $(\beta_k)_{1\le k\le K}$ can be chosen as a geometric sequence for efficient simulation. If τ_k needs to be $\frac{1}{\beta_k}$, $(\beta_k)_{1\le k\le K}$ can be a geometric sequence for $d = 1, 2$. But for $d \geq 3$, our analysis requires β_k to be log geometric.

Corollary 2 *Suppose $\pi_0 = \pi = \sum_{i=1}^{I} p_i \nu_i$, where ν_i is $(l_i^{-2}, l_i^{-2}\mathcal{C})$-log concave densities with modes m_i and $\|m_i\| \leq M$ for $i = 1, \ldots, I$. Let $l_m = \min_i l_i$, $l_M = \max_i l_i$. Consider running mReLD with*

$$\pi_k(x) \propto (\pi(x))^{\beta_k}, \quad k = 1, \ldots, K-1, \quad \pi_K(x) \propto (\pi(x))^{\beta_K} \phi(x/M).$$

With $K, d, \mathcal{C}, M$ all being O(1) constants,

1. *if* $\beta_k = l_m^{\frac{2k}{K}}, \tau_0 = 1$, *and* $\tau_k, \rho \geq l_m^{-\alpha-\frac{d}{K}}$ *for* $0 \leq \alpha \leq 1$, $k = 1, \dots, K$, *then* (10) *holds with* $\kappa = O(l_M^{2\alpha})$;
2. *if* $d \leq 2$, $\tau_k = \beta_k^{-1} = l_m^{-\frac{2k}{K}}$ *for* $k = 0, 1, \dots, K$, *and* $\rho \geq l_m^{-d/K}$, *then* (10) *holds with* $\kappa = O(1)$;
3. *if* $d \geq 3$, $\tau_0 = \beta_0 = 1$, $\tau_k = \beta_k^{-1} = l_m^{-2(\frac{d-2}{d})^{K-k}}$ *for* $k = 1, \dots, K$, *and* $\rho \geq l_m^{-2}$, *then* (10) *holds with* $\kappa = O\big(l_m^{-d(\frac{d-2}{d})^{K-1}}\big)$.

Consider the mixture of Gaussian densities in scenario 2 of Proposition 1 where $l_m^2 = l_M^2 = \epsilon^2$ and $\mathcal{C} = 1$. By choosing $\tau_k, \rho = \Omega(\epsilon^{-\frac{d}{K}-2})$, $\beta_k = \epsilon^{\frac{2k}{K}}$, for $k = 1, \dots, K$, we have $\kappa = O(\epsilon^{-2})$. This matches the LD convergence rate when $\pi(x) \propto \phi(x/\epsilon)$, i.e., a single Gaussian. We can also set $\tau_k, \rho = \Omega(\epsilon^{-d/K})$ and $\beta_k = \epsilon^{\frac{2k}{K}}$, which leads to $\kappa = O(1)$. Comparing the discussion following Corollary 1, we note that the parameters τ_k, ρ are reduced from ϵ^{-d} to $\epsilon^{-d/K}$. This in practice can be computationally more desirable. Lastly, Corollary 1 combined with Corollary 2 (scenario 1 with $\alpha = 0$) proves Theorem 1.

3.3. Morse Hamiltonian functions

The paper [1] considers a general density model based on the Morse function:

$$\pi(x) \propto \exp(-H(x)/\epsilon),$$

where $H(x)$ is a nonnegative Morse function. Due to Proposition 2, we say $\pi_\epsilon(x) \propto \exp(-H_\epsilon(x)/\epsilon)$ (or $H_\epsilon(x)$) is an ϵ perturbation of $\pi(x)$ (or $H(x)$) if

$$|H(x) - H_\epsilon(x)| \leq D\epsilon, \quad \forall x \in \mathbb{R}^d \text{ for some constant } D \in (0, \infty).$$

The paper [1] further assumes that $H(x)$ has a finite set of local minima $\{m_1, \dots, m_I\}$, a partition $\{\Omega_i\}_{1\leq i\leq I}$ of $\mathbb{R}^d$, and a ϵ-perturbation of $H(x)$, $H_\epsilon(x)$ so that

$$\frac{1}{2\epsilon}\Delta H_\epsilon(x) - \frac{1}{4\epsilon^2}\|\nabla H_\epsilon(x)\|^2 \leq -\frac{\lambda_0}{\epsilon}, \quad \forall x \notin \cup B(m_i, a\sqrt{\epsilon}), \tag{12}$$

where $B(m_i, a\sqrt{\epsilon}) \subset \Omega_i$. Moreover, Ω_i is the attraction basin of m_i for gradient flows driven by ∇H_ϵ, i.e., $\Omega_i := \{x \in \mathbb{R}^d : \lim_{t\to\infty} x_t = m_i, \dot{x}_t = -\nabla H_\epsilon(x_t), x_0 = x\}$.

We next consider a transformation of the partition framework in [1] into a mixture model. Define

$$d_i(x) = \min\{\|x-y\| | y \in \Omega_i\} \ \text{ and } \ \Omega_i' = \left\{x : d_i^2(x) \leq \frac{1}{n}\right\}.$$

We assume $d_i^2(x)$ is $\mathbb{C}^2$ on Ω_i' for sufficiently large n with bounded derivatives.

Proposition 7 *Suppose $\pi(x) \propto \exp(-\frac{1}{\epsilon}H(x))$, $\Omega_i' = \{x : 0 < d_i(x) < \frac{1}{\sqrt{n}}\}$, and the following conditions hold:*

1. *There is an ϵ perturbation $H_\epsilon(x)$ such that (12) holds.*
2. *The boundary of Ω_i is regular enough so that $d_i^2(x)$ is $\mathbb{C}^2$ on Ω_i', and for any $x_n \to x \in \partial\Omega_i$, $\nabla d_i(x_n) \to v_\perp(x)$, where $v_\perp(x)$ is the outward direction orthogonal to $\partial\Omega_i$.*
3. *There exists $D_\epsilon \in (0,\infty)$ such that $\Delta d_i(x) \leq D_\epsilon$, $\|\nabla H_\epsilon(x)\| \leq D_\epsilon$, and $\Delta H_\epsilon(x) \leq D_\epsilon$.*

Then, for ϵ sufficiently small, there exists a density π_ϵ, which is an ϵ perturbation of π and $\pi_\epsilon(x) \propto \sum_{i=1}^I p_i \nu_i(x)$, where ν_i has a $(\lambda_0/\epsilon, h_0/\epsilon, B(m_i, a\sqrt{\epsilon}), C)$-Lyapunov function for certain fixed constants h_0 and C.

We next provide a simple concrete example to demonstrate how mixtures of singular densities arise in practice, and how to implement the Morse function framework discussed above. Suppose we want to obtain the posterior density $p(x|y_1,\ldots,y_n)$ where the prior is $\mathcal{N}(0,2)$ and the observation is $y_i = x^2+\xi_i, \xi_i \sim \mathcal{N}(0,1)$. The posterior density is given by

$$p(x|y_1,\ldots,y_n) \propto \exp\left(-\frac{1}{2}\left(2x^2 + \sum_{i=1}^n (x^2-y_i)^2\right)\right) \propto \exp\left(-\frac{n}{2}(x^2-m_n)^2\right).$$

where $m_n = \frac{1}{n}\sum_{i=1}^n y_i - \frac{1}{n}$. It is easy to see that when $m_n > 0$, $p(x|y_1,\ldots,y_n)$ has two modes: $\pm\sqrt{m_n}$. For $m_n = 1$, this density is also known as the double-well potential. Following Proposition 7, we can decompose it into a mixture:

Corollary 3 *For $\pi(x) \propto \exp(-\frac{1}{2}n(x^2-a^2)^2)$ with $a > 0$, $\pi(x) \propto \nu_+(x)+\nu_-(x)$ where $\nu_+(x) = \exp(-\frac{1}{2}n(x^2-a^2)^2)1\{x \geq 0\}$ and $\nu_-(x) = \exp(-\frac{1}{2}n(x^2 -$*

$a^2)^2)1\{x < 0\}$. Moreover, for ϵ sufficiently small, there is a density π_ϵ, which is an ϵ perturbation of π and $\pi_\epsilon(x) \propto \nu_1(x)+\nu_2(x)$ where ν_1 has a $(na^2, nh, B(a, \sqrt{n}r), C)$-Lyapunov function and ν_2 has a $(na^2, nh, B(-a, \sqrt{n}r), C)$-Lyapunov function for certain fixed constants h, C.

4. Proof techniques

In this section, we provide detailed analysis on how the replica-exchange mechanism speeds up the convergence. To make the presentation concise, we allocate most of the technical verification to the appendix.

4.1. Analysis of ReLD

We first explain how to prove Theorem 2. Our proof utilizes the PI. The key is to match (bound) the variance, $\mathrm{var}_{\pi\otimes\pi^Y}(f(X,Y))$, with the carré du champ of ReLD, i.e., Γ_R in (8).

Let $\bar\theta = \mathbb{E}_{\pi\otimes\pi^Y}[f(X,Y)]$, $\eta_i(y) = \int f(x,y)\nu_i(x)dx$ and $\theta_i = \int \eta_i(y)\pi^Y(y)dy$, for $i = 1, 2, \ldots, I$. First, based on the form of π, the variance of $f(X,Y)$ can be decomposed as

$$\mathrm{var}_{\pi\otimes\pi^Y}(f(X,Y)) = \sum_{i=1}^{I} p_i \int (f(x,y) - \bar\theta)^2 \nu_i(x)\pi^Y(y)dxdy.$$

Then, because $f(x,y)-\bar\theta = (f(x,y)-\eta_i(y))+(\eta_i(y)-\theta_i)+(\theta_i-\bar\theta)$, by Cauchy-Schwarz inequality, we can further decompose the variance as

$$\begin{aligned}\mathrm{var}_{\pi\otimes\pi^Y}(f(X,Y)) \le{}& 3\sum_{i=1}^{I} p_i \underbrace{\int (f(x,y) - \eta_i(y))^2\nu_i(x)\pi^Y(y)dxdy}_{\text{(A)}}\\ &+3\sum_{i=1}^{I} p_i \underbrace{\int (\eta_i(y)-\theta_i)^2\pi^Y(y)dy}_{\text{(B)}} +3\sum_{i,j=1}^{I} p_i p_j \underbrace{(\theta_i-\theta_j)^2}_{\text{(C)}}.\end{aligned} \tag{13}$$

To see part (C), note that as $\bar\theta = \sum_{i=1}^I p_i\theta_i$,

$$\sum_{i=1}^{I} p_i(\theta_i-\bar\theta)^2 = \sum_{i=1}^{I} p_i \left(\sum_{j=1}^{I} p_j(\theta_i-\theta_i)\right)^2 \le \sum_{i,j} p_i p_j (\theta_i-\theta_j)^2.$$

In the decomposition (13), part (A) is the variance of f under ν_i with y being fixed. Part (B) is the variance of η_i under π^Y. Since ν_i and π^Y satisfy the Lyapunov condition, parts (A) and (B) can be controlled using Proposition 3. Thus, the key is to develop an upper bound for the mean difference square in part (C): $(\theta_i - \theta_j)^2 = \left(\mathbb{E}_{\nu_i \otimes \pi^Y}[f(X,Y)] - \mathbb{E}_{\nu_j \otimes \pi^Y}[f(X,Y)]\right)^2$.

When running LD alone, [1] provides an estimate of the difference between $\mathbb{E}_{\nu_i}[f(X)]$ and $\mathbb{E}_{\nu_j}[f(X)]$ (see Theorem 2.12 in [1]). The estimate depends on the saddle height, and when $\nu_i \propto \phi((x-m_i)/\epsilon)$, it grows exponentially in $1/\epsilon$. One of the main technical contribution of this paper is to find an upper bound for the mean difference in the ReLD setting. In particular, we establish that the ratio between the mean difference square and the carré du champ of ReLD stays invariant when ϵ goes to zero. To achieve a better PI constant, we need to exploit the additional exchange term that arises in the carré du champ for ReLD:

$$\begin{aligned}
&\mathbb{E}_{\pi \otimes \pi^Y}\left[\rho s(X,Y)(f(Y,X) - f(X,Y))^2\right] \\
=&\sum_{i,j} p_i p_j \rho \int (f(y,x) - f(x,y))^2 (\nu_i(x)\pi^Y(y)) \wedge (\nu_j(y)\pi^Y(x)) dx dy \\
\le&\rho \int (f(y,x) - f(x,y))^2 (\pi(x)\pi^Y(y)) \wedge (\pi(y)\pi^Y(x)) dx dy.
\end{aligned}$$

In the following, we refer to $(\nu_i(x)\pi^Y(y)) \wedge (\nu_j(y)\pi^Y(x))$ as a "maximal coupling density" as its formulation is similar to the L_1-maximal coupling between $\nu_i(x)\pi^Y(y)$ and $\nu_j(y)\pi^Y(x)$ [31]. However, this "maximal coupling density" is still difficult to deal with. To resolve the challenge, we replace ν_i by $u_{B(m_i,r_i)}$, which is the uniform distribution on $B(m_i, r_i)$, and π^Y by $u_{B(m_j,R_j)}$ using appropriate bounding arguments. The "maximal coupling density" with uniform distributions is much easier to handle, and we can build proper bound for the transformed mean difference square under uniform distributions. Following this idea, we establish the following bound for the mean difference square.

Proposition 8 *Consider four densities $\nu_1^X, \nu_2^X, \nu_1^Y, \nu_2^Y$. Suppose ν_i^X is a $\mathbf{Ly}(r_i, q, a)$-density with center m_i for $i = 1, 2$. Similarly, suppose ν_i^Y is a $\mathbf{Ly}(R_i, Q, A)$-density with center m_i for $i = 1, 2$. Moreover, for $i = 1, 2$, there are constants*

R, r, a, A such that $R_i \le R$ and $R_i/r_i \le R/r$. Then

$$\begin{aligned}
&\left(\mathbb{E}_{\nu_1^X \otimes \nu_2^Y}[f(X,Y)] - \mathbb{E}_{\nu_2^X \otimes \nu_1^Y}[f(X,Y)]\right)^2 \\
\le &\Xi_x \int \|\nabla_x f(x,y)\|^2 (\nu_1^X(x)\nu_2^Y(y) + \nu_2^X(x)\nu_1^Y(y)) dxdy \\
&+ \Xi_y \int \|\nabla_y f(x,y)\|^2 (\nu_1^X(x)\nu_2^Y(y) + \nu_2^X(x)\nu_1^Y(y)) dxdy \\
&+ \Xi_e \int (f(x,y) - f(y,x))^2 \left(\nu_1^X(x)\nu_2^Y(y) \wedge \nu_2^X(y)\nu_1^Y(x)\right) dxdy,
\end{aligned}$$

where $\Xi_x = 14(q + r^2a^2)A$, $\Xi_y = 14(Q + R^2A^2) + 7aA\left(\frac{R^{d+1}}{r^{d-1}}\right)\left(\log\left(\frac{R}{r}\right)\right)^{1_{d=1}}$, and $\Xi_e = 7\left(\frac{R}{r}\right)^d aA$.

The proof Proposition 8 is in Appendix D.

Proof [Proof of Theorem 2] Recall the decomposition in (13).

For part (A). By Assumption 1 and Proposition 3, we have

$$\int (f(x,y) - \eta_i(y))^2 \nu_i(x)\pi^Y(y) dxdy \le q \int \|\nabla_x f(x,y)\|^2 \nu_i(x)\pi^Y(y) dxdy.$$

For part (B). By Assumption 2 and Proposition 3, we have

$$\begin{aligned}
\int (\eta_i(y) - \theta_i)^2 \pi^Y(y) dy &\le Q \int \|\nabla \eta_i(y)\|^2 \pi^Y(y) dy \\
&\le \frac{Q}{\tau}\tau \int \|\nabla_y f(x,y)\|^2 \nu_i(x)\pi^Y(y) dxdy,
\end{aligned}$$

where the second inequality follows from Jensen's inequality since $\nabla\eta_i(y) = \int \nabla_y f(x,y)\nu_i(x)dx$.

For part (C). Under Assumptions 2 and 3, for each center m_i, $i = 1, \ldots, I$, π^Y is a $\mathbf{Ly}(R_i, Q, A)$-density with $R_i \le R, R_i/r_i \le R/r$. Thus, by setting $\nu_i^X = \nu_i$ and $\nu_i^Y = \pi^Y$ in Proposition 8, we have

$$\begin{aligned}
&\left(\int f(x,y)\nu_i(x)\pi^Y(y)dxdy - \int f(x,y)\nu_j(x)\pi^Y(y)dxdy\right)^2 \\
\le &\Xi_x \int \|\nabla_x f(x,y)\|^2 (\nu_i(x)\pi^Y(y) + \nu_j(x)\pi^Y(y)) dxdy \\
&+ \frac{\Xi_y}{\tau}\tau \int \|\nabla_y f(x,y)\|^2 (\nu_i(x)\pi^Y(y) + \nu_j(x)\pi^Y(y)) dxdy \\
&+ \frac{\Xi_e}{\rho}\rho \int (f(x,y) - f(y,x))^2 \left(\nu_i(x)\pi^Y(y) \wedge \nu_j(y)\pi^Y(x)\right) dxdy.
\end{aligned}$$

Putting the bounds for (A) – (C) together, because $\sum_{i,j} p_i p_j(\nu_i(x)\pi^Y(y) + \nu_j(x)\pi^Y(y)) = 2\pi(x)\pi^Y(y)$ and

$$\sum_{i,j} p_i p_j(\nu_i(x)\pi^Y(y) \wedge \nu_j(y)\pi^Y(x)) \leq (\pi(x)\pi^Y(y)) \wedge (\pi(y)\pi^Y(x)),$$

$$\begin{aligned}\mathrm{var}_{\pi\otimes\pi^Y}(f(X,Y)) \leq& 3\left(q + 2\Xi_x\right)\int \|\nabla_x f(x,y)\|^2\pi(x)\pi^Y(y)dxdy\\ &+ 3\left(\frac{Q}{\tau} + 2\frac{\Xi_y}{\tau}\right)\int \tau\|\nabla_y f(x,y)\|^2\pi(x)\pi^Y(y)dxdy\\ &+ 3\frac{\Xi_e}{\rho}\rho\int (f(x,y) - f(y,x))^2\left(\pi(x)\pi^Y(y) \wedge \pi(y)\pi^Y(x)\right)dxdy\\ \leq& \kappa\mathbb{E}_{\pi\otimes\pi^Y}[\Gamma_R(f(X,Y))],\end{aligned}$$

where

$$\kappa = \max\left\{3(56A+1)q,\ \frac{3}{\tau}\left(57Q + 14aA\left(\frac{R^{d+1}}{r^{d-1}}\right)\left(\log\left(\frac{R}{r}\right)\right)^{1_{d=1}}\right), \frac{7aA}{\rho}\left(\frac{R}{r}\right)^d\right\}.$$

□

4.2. Analysis of Multiple ReLD

We first rephrase Theorem 3 into a more detailed version as follows:

Theorem 4 *For mReLD defined in (9), under Assumptions 4 and 5,*

$$var_{\pi_{0:K}}(f(\mathbf{X}_{0:K})) \leq \kappa\mathbb{E}_{\pi_{0:K}}[\Gamma_R^K(f(\mathbf{X}_{0:K}))],$$

where

$$\begin{aligned}\kappa = \max_{0\leq k\leq K-1}\max\Bigg\{&\sum_{h=2}^{k-2}\frac{3(4\alpha)^{k-h+1}}{\tau_k}\left(8\alpha\gamma\Xi_{x_k} + 2\gamma\Xi_{y_{k-1}}\right)\\ &+\frac{3}{\tau_k}\left((8\alpha\gamma+2\gamma)\Xi_{x_k} + 2\gamma\Xi_{y_{k-1}} + 2q_k\right), \sum_{h=0}^{k}\frac{3(4\alpha)^{k-h+2}}{\rho}\gamma\Xi_{e_k}\Bigg\},\end{aligned}$$

for any $\alpha,\gamma > 1$ with $1/\alpha + 1/\gamma = 1$, and

$$\Xi_{x_k} = 28q_k a_{k+1},\quad \Xi_{y_k} = 28q_{k+1} + 7\frac{(r_{k+1})^{d+1}}{(r_k)^{d-1}}a_k a_{k+1}\left(\log\left(\frac{r_{k+1}}{r_k}\right)\right)^{1_{d=1}},$$

$$\Xi_{y_{(-1)}} = 0,\quad \Xi_{e_k} = 7\left(\frac{r_{k+1}}{r_k}\right)^d a_k a_{k+1}.$$

The proof of Theorem 4 builds on the analysis of ReLD and induction arguments. We provide a roadmap of our proving strategy in this section.

Denote

$$\mathbb{E}_{r:h}[f(\mathbf{X}_{0:K})] := \int f(\mathbf{X}_{0:r-1}, \mathbf{y}_{r:h}, \mathbf{X}_{h+1:K})\pi_{r:h}(\mathbf{y}_{r:h})d\mathbf{y}_{r:h},$$
$$\mathbb{E}_{k}[f(\mathbf{X}_{0:K})] := \int f(\mathbf{X}_{0:k-1}, y, \mathbf{X}_{k+1:K})\pi_{k}(y)dx,$$

and we write $\mathbb{E}_{(K+1):K}[f(\mathbf{X}_{0:K})] = f(\mathbf{X}_{0:K})$ for convenience. We also write

$$\mathrm{var}_{0:K}(f(\mathbf{X}_{0:K})) := \mathbb{E}_{0:K}\left[\left(f(\mathbf{X}_{0:K}) - \mathbb{E}_{0:K}f(\mathbf{X}_{0:K})\right)^2\right]$$

We first note that

$$f(\mathbf{X}_{0:K}) - \mathbb{E}[f(\mathbf{X}_{0:K})] = \sum_{k=0}^{K}\left(\mathbb{E}_{(k+1):K}[f(\mathbf{X}_{0:K})] - \mathbb{E}_{k:K}[f(\mathbf{X}_{0:K})]\right).$$

For $j < k$,

$$\begin{aligned}&\mathbb{E}_{0:K}\left[\left(\mathbb{E}_{(j+1):K}f(\mathbf{X}_{0:K}) - \mathbb{E}_{j:K}f(\mathbf{X}_{0:K})\right)\left(\mathbb{E}_{(k+1):K}f(\mathbf{X}_{0:K}) - \mathbb{E}_{k:K}f(\mathbf{X}_{0:K})\right)\right]\\ =&\mathbb{E}_{0:K}\left[\left(\mathbb{E}_{(k+1):K}f(\mathbf{X}_{0:K}) - \mathbb{E}_{k:K}f(\mathbf{X}_{0:K})\right)\mathbb{E}_{k:K}\left[\left(\mathbb{E}_{(j+1):K}f(\mathbf{X}_{0:K}) - \mathbb{E}_{j:K}f(\mathbf{X}_{0:K})\right)\right]\right] = 0.\end{aligned}$$

Thus, we have the following variance decomposition

$$\mathrm{var}_{0:K}(f(\mathbf{X}_{0:K})) = \sum_{k=0}^{K}\mathbb{E}_{0:K}\left[\mathbb{E}_k\left[\left(\mathbb{E}_{(k+1):K}f(\mathbf{X}_{0:K}) - \mathbb{E}_{k:K}f(\mathbf{X}_{0:K})\right)^2\right]\right].$$

The above decomposition allows us to focus on

$$\mathbb{E}_k\left[\left(\mathbb{E}_{(k+1):K}f(\mathbf{X}_{0:K}) - \mathbb{E}_{k:K}f(\mathbf{X}_{0:K})\right)^2\right]$$

individually. Let $\mathbf{W}_k = \mathbf{X}_{0:(k-1)}$, $Y_k = X_{k+1}$, $\mathbf{Z}_k = \mathbf{X}_{(k+2):K}$. We also write $\pi_k^{\mathbf{Z}} = \pi_{(k+2):K}$. For a fixed $\mathbf{W}_k = \mathbf{w}_k$, we define

$$g_k(\mathbf{w}_k, x_k, y_k) = \int f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k)\pi_k^{\mathbf{Z}}(\mathbf{z}_k)d\mathbf{z}_k,$$
$$\eta_{k,i}(\mathbf{w}_k, y_k) = \int g_k(\mathbf{w}_k, x_k, y_k)\nu_{k,i}(x_k)dx_k,$$
$$\theta_{k,i}(\mathbf{w}_k) = \int \eta_{k,i}(\mathbf{w}_k, y_k)\pi_{k+1}(y_k)dy_k, \text{ and } \bar{\theta}_k(\mathbf{w}_k) = \sum_{i=1}^{I}p_i\theta_{k,i}(\mathbf{w}_k).$$

Note that with these notations,

$$\mathbb{E}_{k+1:K}[f(\mathbf{X}_{0:K})] = \int g_k(\mathbf{W}_k, X_k, y_k)\pi_{k+1}(y_k)dy_k, \quad \mathbb{E}_{k:K}[f(\mathbf{X}_{0:K})] = \bar{\theta}_k(\mathbf{W}_k).$$

Following similar lines of argument as (13), we have

$$\begin{aligned}
&\mathbb{E}_k\left[\left(\mathbb{E}_{(k+1):K}f(\mathbf{X}_{0:K}) - \mathbb{E}_{k:K}f(\mathbf{X}_{0:K})\right)^2\right] \\
=&\int\left(\int g_k(\mathbf{w}_k, x_k, y_k)\pi_{k+1}(y_k)dy_k - \bar{\theta}_k(\mathbf{w}_k)\right)^2\pi_k(x_k)dx_k \\
\leq&3\sum_{i=1}^{I} p_i \underbrace{\int\left(g_k(\mathbf{w}_k, x_k, y_k) - \eta_{k,i}(\mathbf{w}_k, y_k)\right)^2 \nu_{k,i}(x_k)dx_k\pi_{k+1}(y_k)dy_k}_{\text{(A)}} \\
&+3\sum_{i=1}^{I} p_i \underbrace{\int(\eta_{k,i}(\mathbf{w}_k, y_k) - \theta_{k,i}(\mathbf{w}_k))^2\pi_{k+1}(y_k)dy_k}_{\text{(B)}} + 3\sum_{i,j} p_i p_j \underbrace{(\theta_{k,i}(\mathbf{w}_k) - \theta_{k,j}(\mathbf{w}_k))^2}_{\text{(C)}}.
\end{aligned} \tag{14}$$

We note that part (A) and (B) are variances of functions with respect to individual mixture component. Thus, they are easy to bound using Proposition 3. For part (C), we utilize Proposition 8 and an induction argument on k to develop a proper upper bound for the mean difference square. The details can be found in Proposition 10 in Appendix E. The proof of Theorem 4 is also provided in Appendix E.

5. Conclusion and future directions

LD is a popular sampling method, but its convergence rate can be significantly reduced if the target distribution is a mixture of singular densities. ReLD is a method that can circumvent this issue. It employs an additional LD process sampling a high temperature version of the target distribution, and swaps the values of the two processes according to a Metropolis-Hasting mechanism. More generally, mReLD employs K additional LD processes sampling with different temperature coefficients. In this work, we formulate a framework to quantify the spectral gap of ReLD and mReLD. Our analysis shows that the spectral gap of ReLD does not degenerate when the mixture component becomes singular, as long as the simulation parameters of ReLD scale properly with the singularity parameter ϵ. While using mReLD can achieve the same convergence rate, the simulation parameters have a weaker dependence on the singularity parameter.

While our results close some theoretical gaps for ReLD and mReLD, there are several questions left unanswered. First, ReLD and mReLD are stochastic processes, but not executable sampling algorithms. How to derive efficient MCMC algorithms from them is an interesting research question. Notably, direct simulation methods like Euler-Maruyama will incur sampling bias. While using Metropolis adjusted Langevin algorithm (MALA) can remove such bias, whether the spectral gap of mReLD can be inherited by its MALA implementation requires further analysis.

Second, high dimensionality is another major challenge for sampling problems besides multi-modality. Replica exchange alone may not be a good tool to handle high dimensionality. Implementing ReLD on high dimensional distributions also have additional computational challenges, which often require novel techniques to handle [16, 17, 18]. Our estimate for the spectral gap has an exponential dependence on the dimension. This is mainly because our assumptions on the target distribution are quite general. Better scaling on the dimension can be obtained if we assume the existence of a lower effective dimension [32, 33] or a sparse conditional structure [34, 35]. We also note that convergence metrics other than the spectral gap may have better dependence on the dimension. Examples include MCMC variance [15] and round trip rates [17].

Lastly, we remark that ReLD and mReLD does not require any prior knowledge of the locations of the mixture components. However, knowing the locations and other information can potentially lead to more efficient algorithms, examples of which can be found in [36, 37].

References

[1] G. Menz, A. Schlichting, Poincaré and logarithmic sobolev inequalities by decomposition of the energy landscape, The Annals of Probability 42 (5) (2014) 1809–1884.

[2] D. J. Earl, M. W. Deem, Parallel tempering: Theory, applications, and new perspectives, Physical Chemistry Chemical Physics 7 (23) (2005) 3910–3916.

[3] R. H. Swendsen, J.-S. Wang, Replica Monte Carlo simulation of spin-glasses, Physical review letters 57 (21) (1986) 2607.

[4] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chemical physics letters 314 (1-2) (1999) 141–151.

[5] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX 1 (2015) 19–25.

[6] D. Bakry, I. Gentil, M. Ledoux, Analysis and geometry of Markov diffusion operators, Vol. 348, Springer Science & Business Media, 2013.

[7] P. Dupuis, Y. Liu, N. Plattner, J. Doll, On the infinite swapping limit for parallel tempering, Multiscale Model Simulation 10 (3) (2012) 986–1022.

[8] J. D. Chodera, M. R. Shirts, Replica exchange and expanded ensemble simulations as gibbs sampling: Simple improvements for enhanced mixing, The Journal of chemical physics 135 (19) (2011) 194110.

[9] M. Ebbers, H. Knöpfel, M. Löwe, F. Vermet, Mixing times for the swapping algorithm on the blume-emery-griffiths model, Random Structures & Algorithms 45 (1) (2014) 38–77.

[10] K. Cho, T. Raiko, A. Ilin, Parallel tempering is efficient for learning restricted boltzmann machines, in: The 2010 international joint conference on neural networks (ijcnn), IEEE, 2010, pp. 1–8.

[11] G. Desjardins, A. Courville, Y. Bengio, P. Vincent, O. Delalleau, et al., Parallel tempering for training of restricted boltzmann machines, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, MIT Press Cambridge, MA, 2010, pp. 145–152.

[12] H. Hult, P. Nyquist, C. Ringqvist, Infinite swapping algorithm for training restricted boltzmann machines, in: International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Springer, 2018, pp. 285–307.

[13] Y. Chen, J. Chen, J. Dong, J. Peng, Z. Wang, Accelerating nonconvex learning via replica exchange Langevin diffusion, in: International Conference on Learning Representations, 2019.

[14] J. Dong, X. T. Tong, Replica exchange for non-convex optimization, Journal of Machine Learning Research 22 (173) (2021) 1–59.

[15] G. O. Roberts, J. S. Rosenthal, Minimising MCMC variance via diffusion limits, with an application to simulated temperingmcmc variance via diffusion limits, with an application to simulated tempering, The Annals of Applied Probability 24 (1) (2014) 131–149.

[16] N. G. Tawn, G. O. Roberts, Accelerating parallel tempering: Quantile tempering algorithm (quanta), Advances in Applied Probability 51 (3) (2019) 802–834.

[17] S. Syed, A. Bouchard-Côté, G. Deligiannidis, A. Doucet, Non-reversible parallel tempering: A scalable highly parallel mcmc scheme, arXiv preprint arXiv:1905.02939.

[18] N. G. Tawn, G. O. Roberts, J. S. Rosenthal, Weight-preserving simulated tempering, Statistics and Computing 30 (1) (2020) 27–41.

[19] D. B. Woodard, S. C. Schmidler, M. Huber, et al., Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions, The Annals of Applied Probability 19 (2) (2009) 617–640.

[20] D. J. Sindhikara, D. J. Emerson, A. E. Roitberg, Exchange often and properly in replica exchange molecular dynamics, Journal of Chemical Theory and Computation 6 (9) (2010) 2804–2808.

[21] D. A. Kofke, On the acceptance probability of replica-exchange Monte Carlo trials, The Journal of chemical physics 117 (15) (2002) 6911–6914.

[22] M. J. Abraham, J. E. Gready, Ensuring mixing efficiency of replica-exchange molecular dynamics simulations, Journal of Chemical Theory and Computation 4 (7) (2008) 1119–1128.

[23] J. Doll, P. Dupuis, P. Nyquist, A large deviations analysis of certain qualitative properties of parallel tempering and infinite swapping algorithms, Applied Mathematics & Optimization 78 (1) (2018) 103–144.

[24] G. Menz, A. Schlichting, W. Tang, T. Wu, Ergodicity of the infinite swapping algorithm at low temperature, arXiv preprint arXiv:1811.10174.

[25] E. Marinari, G. Parisi, Simulated tempering: a new Monte Carlo scheme, EPL (Europhysics Letters) 19 (6) (1992) 451.

[26] C. J. Geyer, E. A. Thompson, Annealing Markov chain Monte Carlo with applications to ancestral inference, Journal of the American Statistical Association 90 (431) (1995) 909–920.

[27] R. M. Neal, Sampling from multimodal distributions using tempered transitions, Statistics and computing 6 (4) (1996) 353–366.

[28] N. Madras, D. Randall, Markov chain decomposition for convergence rate analysis, Annals of Applied Probability (2002) 581–606.

[29] R. Ge, H. Lee, A. Risteski, Simulated tempering Langevin Monte Carlo ii: An improved proof using soft markov chain decomposition, arXiv preprint arXiv:1812.00793.

[30] D. Bakry, F. Barthe, P. Cattiaux, A. Guillin, A simple proof of the Poincaré inequality for a large class of probability measures, Electronic Communications in Probability 13 (2008) 60–66.

[31] T. Lindvall, Lectures on the coupling method, Courier Corporation, 2002.

[32] M. Hairer, A. Stuart, S. Vollmer, Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions, Ann. Appl. Probab. 24 (6) (2014) 2455–2490.

[33] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, A. Spantini, Likelihood-informed dimension reduction for nonlinear inverse problems, Inverse Problems 30 (11) (2014) 114015.

[34] M. Morzfeld, X. T. Tong, Y. M. Marzouk, Localization for mcmc: sampling high-dimensional posterior distributions with local structure, Journal of Computational Physics 380 (2019) 1–28.

[35] X. T. Tong, M. Morzfeld, Y. M. Marzouk, Mala-within-gibbs samplers for high-dimensional distributions with sparse conditional structure, SIAM Journal on Scientific Computing 42 (3) (2020) A1765–A1788.

[36] D. Aristoff, T. Lelievre, Mathematical analysis of temperature accelerated dynamics, Multiscale Modeling & Simulation 12 (1) (2014) 290–317.

[37] T. Lelievre, G. Stoltz, Partial differential equations and stochastic methods in molecular dynamics, Acta Numerica 25 (2016) 681–880.

[38] M. Bebendorf, A note on the poincaré inequality for convex domains, Zeitschrift für Analysis und ihre Anwendungen 22 (4) (2003) 751–756.

Appendix A. Proof of Proposition 1

Proof **Claim 1)** This is classical result one can find in [1, 6].

Claim 2) Let $\nu(x) \propto \phi(x/\epsilon)$, i.e., $\nu(x) = \frac{1}{\epsilon^d}\phi(x/\epsilon)$.

Because $\nabla\phi(x/\epsilon) = -\frac{1}{\epsilon^2}\phi(x/\epsilon)x$,

$$
\begin{aligned}
&\mathcal{E}\left(\frac{\nu}{\pi}\right) = 4\int \left\|\nabla \frac{\phi(x/\epsilon)}{\phi(x/\epsilon)+\phi((x-m)/\epsilon)}\right\|^2 \pi(x)dx \\
=&\frac{4}{\epsilon^4}\int \left\|\frac{\phi(x/\epsilon)x(\phi(x/\epsilon)+\phi((x-m)/\epsilon)) - \phi(x/\epsilon)(\phi(x/\epsilon)x + \phi((x-m)/\epsilon)(x-m))}{(\phi(x/\epsilon)+\phi((x-m)/\epsilon))^2}\right\|^2 \pi(x)dx \\
=&\frac{4\|m\|^2}{\epsilon^4}\int \left|\frac{\phi(x/\epsilon)\phi((x-m)/\epsilon)}{(\phi(x/\epsilon)+\phi((x-m)/\epsilon))^2}\right|^2 \pi(x)dx \\
=&\frac{4\|m\|^2}{\epsilon^4}\int \left|r(x)+1/r(x)+2\right|^{-2}\pi(x)dx \quad \text{ where } r(x) = \phi((x-m)/\epsilon)/\phi(x/\epsilon) \\
=&\frac{4\|m\|^2}{\epsilon^4}\left(\int_{A\bigcup B} \left|r(x)+1/r(x)+2\right|^{-2}\pi(x)dx + \int_{A^c\bigcap B^c}\left|r(x)+1/r(x)+2\right|^{-2}\pi(x)dx\right),
\end{aligned}
$$

where $A = \{\|x\|^2 \le \|m\|^2/16\}, B = \{\|x-m\|^2 \le \|m\|^2/16\}$. When $x \in A$,

$$
\begin{aligned}
r(x) = \exp\left(-\frac{\|x-m\|^2 - \|x\|^2}{2\epsilon^2}\right) &\le \exp\left(\frac{2\|x\|\|m\| - \|m\|^2}{2\epsilon^2}\right) \\
&\le \exp\left(\frac{\frac{2}{4}\|m\|^2 - \|m\|^2}{2\epsilon^2}\right) = \exp\left(-\frac{\|m\|^2}{4\epsilon^2}\right).
\end{aligned} \tag{A.1}
$$

Likewise, we can show that when $x \in B$, $\frac{1}{r(x)} \le \exp\left(-\frac{\|m\|^2}{4\epsilon^2}\right)$. Thus,

$$
\int_{A\bigcup B}\left|r(x)+1/r(x)+2\right|^{-2}\pi(x)dx \le \exp\left(-\frac{\|m\|^2}{4\epsilon^2}\right). \tag{A.2}
$$

Next, we note that $\left|r(x)+1/r(x)+2\right|^{-2} \le \frac{1}{16}$ always hold. Therefore,

$$
\begin{aligned}
\int_{A^c\bigcap B^c}&\left|r(x)+1/r(x)+2\right|^{-2}\pi(x)dx \le \frac{1}{16}\int_{A^c\bigcap B^c}\pi(x)dx \\
\le&\frac{1}{16}\left(\int_{\left\{\|z\|^2 > \frac{\|m\|^2}{16\epsilon^2}\right\}}\phi(z)dz + \int_{\left\{\|z-m\|^2 > \frac{\|m\|^2}{16\epsilon^2}\right\}}\phi(z-m)dz\right) \\
\le&\frac{1}{16}2\exp\left(-\frac{d}{2}\left(\frac{\|m\|^2}{16d\epsilon^2} - \frac{1}{2} - \log\left(\frac{\|m\|^2}{8d\epsilon^2}\right)\right)\right) \quad \text{by Cramer's bound} \\
\le&\frac{1}{8}\exp\left(-\frac{1}{64}\frac{\|m\|^2}{\epsilon^2}\right) \quad \text{for } \epsilon \le \tfrac{\|m\|}{16\sqrt{d}}.
\end{aligned} \tag{A.3}
$$

The last inequality holds because when $\epsilon \leq \frac{\|m\|}{16\sqrt{d}}$,

$$\frac{\|m\|^2}{8d\epsilon^2} \geq 32 \text{ and } \log\left(\frac{\|m\|^2}{8d\epsilon^2}\right) < \frac{\|m\|^2}{32d\epsilon^2}.$$

Putting (A.2) and (A.3) together, we have

$$\mathcal{E}\left(\frac{\nu}{\pi}\right) \leq \frac{4\|m\|^2}{\epsilon^4}\left(\exp\left(-\frac{\|m\|^2}{4\epsilon^2}\right) + \frac{1}{8}\exp\left(-\frac{1}{64}\frac{\|m\|^2}{\epsilon^2}\right)\right) \leq \frac{5\|m\|^2}{\epsilon^4}\exp\left(-\frac{\|m\|^2}{64\epsilon^2}\right).$$

On the other hand, $\chi^2(\nu\|\pi) = \int\left(\frac{\nu(x)}{\pi(x)} - 1\right)^2 \pi(x)dx \geq \left(\int\left|\frac{\nu(x)}{\pi(x)} - 1\right|\pi(x)dx\right)^2$. We also note that $\nu(x) - \pi(x) = \frac{1}{2}\frac{1}{\epsilon^d}\phi(x/\epsilon) - \frac{1}{2}\frac{1}{\epsilon^d}\phi((x-m)/\epsilon)$. Thus,

$$\begin{aligned}
\left(\int\left|\frac{\nu(x)}{\pi(x)} - 1\right|\pi(x)dx\right)^2 &= \frac{1}{4}\frac{1}{\epsilon^{2d}}\left(\int|\phi(x/\epsilon) - \phi((x-m)/\epsilon))|\,dx\right)^2 \\
&= \frac{1}{4}\frac{1}{\epsilon^{2d}}\left(\int|1 - r(x)|\,\phi(x/\epsilon)dx\right)^2 \\
&\geq \frac{1}{4}\frac{1}{\epsilon^{2d}}\left(\int_{\{\|x\|^2\leq\|m\|^2/16\}}\left|1 - \exp\left(-\frac{\|m\|^2}{4\epsilon^2}\right)\right|\phi(x/\epsilon)dx\right)^2 \quad \text{by (A.1)} \\
&\geq \frac{1}{8}\left(\int_{\left\{\|z\|^2\leq\frac{\|m\|^2}{16\epsilon^2}\right\}}\phi(z)dz\right)^2 \quad \text{by replacing } x/\epsilon \text{ with } z \\
&= \frac{1}{8}\left(1 - \int_{\left\{\|z\|^2>\frac{\|m\|^2}{16\epsilon^2}\right\}}\phi(z)dz\right)^2 \\
&\geq \frac{1}{8}\left(1 - \exp\left(-\frac{1}{64}\frac{\|m\|^2}{\epsilon^2}\right)\right)^2 \geq \frac{1}{16},
\end{aligned}$$

where we use Cramer bound again for $\epsilon \leq \frac{\|m\|}{16\sqrt{d}}$. Above all,

$$\kappa = \max_{u:u\ll\pi}\frac{\chi^2(u\|\pi)}{\mathcal{E}(u/\pi)} \geq \frac{\chi^2(\nu\|\pi)}{\mathcal{E}(\nu/\pi)} \geq \frac{\epsilon^4}{80\|m\|^2}\exp\left(\frac{\|m\|^2}{64\epsilon^2}\right).$$

□

Appendix B. Proof of results in Section 2

Appendix B.1. Proof of Proposition 2

Proof Let $\bar{f}_\mu$ and $\bar{f}_\pi$ be the mean of f under μ and π.

$$\begin{aligned}
\mathrm{var}_\mu(f(X)) = \int(f(x) - \bar{f}_\mu)^2\mu(x)dx &\leq \int(f(x) - \bar{f}_\pi)^2\mu(x)dx \\
&\leq C\int(f(x) - \bar{f}_\pi)^2\pi(x)dx \leq C^2\kappa\int\Gamma(f)(x)\mu(x)dx.
\end{aligned}$$

□

Appendix B.2. Proof of Proposition 3

Before we prove Proposition 3, we first introduce a few auxiliary lemmas.

Lemma 1 *Given a ball $B = B(x_0, R) \subset \mathbb{R}^d$, u_B satisfies a R^2-PI:*

$$var_{u_B}(f(X)) \leq R^2 \mathbb{E}_{u_B}[\|\nabla f(X)\|^2]. \tag{B.1}$$

This is a classical result, which can be found in [38].

For a given measure μ, we denote $\mu_D(x) = \frac{\mu(x) 1_D(x)}{\int_D \mu(y) dy}$, i.e., the measure μ conditional on being in the bounded domain D.

Lemma 2 *Given a ball $B = B(x_0, R) \subset \mathbb{R}^d$, suppose* $\max_{x \in B} \mu(x) / \min_{x \in B} \mu(x) \leq C$*. Then $var_{\mu_B}(f(X)) \leq C^2 R^2 \mathbb{E}_{\mu_B}[\|\nabla f(X)\|^2]$.*

Proof Apply Proposition 2 and Lemma 1 we have the result. □

Proof [Proof of Proposition 3] The arguments we use here are similar to the ones used in [30]. The only difference is that we use Lemma 2 to find the bounding constants explicitly. Note that for any constant c,

$$\int (f(x) - c)^2 \nu(x) dx \leq \underbrace{\int \frac{-\mathcal{L}_\nu V(x)}{\lambda V(x)} (f(x) - c)^2 \nu(x) dx}_{\text{(I)}} + \underbrace{\int (f(x) - c)^2 \frac{b}{\lambda V(x)} 1_B(x) \nu(x) dx}_{\text{(II)}}.$$

For part (I), note that

$$\begin{aligned}
&\int \frac{-\mathcal{L}_\nu V(x)}{V(x)} (f(x) - c)^2 \nu(x) dx \\
=&\int \left\langle \nabla \left(\frac{(f(x) - c)^2}{V(x)} \right), \nabla V(x) \right\rangle \nu(x) dx \text{ by equation (1.7.1) in [6]} \\
=&2 \int \frac{f(x) - c}{V(x)} \langle \nabla f(x), \nabla V(x) \rangle \nu(x) dx - \int \frac{(f(x) - c)^2}{V(x)^2} \|\nabla V(x)\|^2 \nu(x) dx \\
=&\int \|\nabla f(x)\|^2 \nu(x) dx - \int \left\| \nabla f(x) - \frac{f(x) - c}{V(x)} \nabla V(x) \right\|^2 \nu(x) dx \leq \int \|\nabla f(x)\|^2 \nu(x) dx.
\end{aligned}$$

For part (II), recall ν_B is ν conditioned on being in B. Set $c = \int f(x)\nu_B(x)dx$.

$$\begin{aligned}\int_B \frac{(f(x)-c)^2}{V(x)}\nu(x)dx &= \mathbb{P}_\nu(X\in B)\int_B \frac{(f(x)-c)^2}{V(x)}\nu_B(x)dx\\ &\leq \mathbb{P}_\nu(X\in B)\int_B (f(x)-c)^2\nu_B(x)dx \text{ as } V(x)\geq 1\\ &\leq \mathbb{P}_\nu(X\in B)C^2R^2\int_B \|\nabla f(x)\|^2\nu_B(x)dx \text{ by Lemma 2}\\ &\leq C^2R^2\int \|\nabla f(x)\|^2\nu(x)dx.\end{aligned}$$

Putting the two parts together, we have $\mathrm{var}_\nu(f) \leq \left(\frac{1}{\lambda} + \frac{bC^2R^2}{\lambda}\right)\mathbb{E}_\nu[\|\nabla f(X)\|^2]$.

□

Appendix C. Proof of Results in Section 3

Appendix C.1. Proof of Proposition 4

Proof Recall that $H(x) = -\log\nu(x)$. Without loss of generality, we assume $m = 0$ and $H(0) = 0$. We first note that $\|\nabla H(x)\|\|x\| \geq \langle\nabla H(x), x\rangle \geq c\|x\|^2$ and $\|\nabla H(x)\|^2 \geq c\langle\nabla H(x), x\rangle$. Then, by convexity of H, we have $H(0) \geq H(x) - \langle\nabla H(x), x\rangle$, which implies that $cH(x) \leq c\langle\nabla H(x), x\rangle \leq \|\nabla H(x)\|^2$. For $V(x) = \frac{c}{d}\|x\|^2 + 1$,

$$\mathcal{L}_\nu V(x) = -\frac{2c}{d}\langle\nabla H(x), x\rangle + 2c \leq -2\frac{c^2}{d}\|x\|^2 + 2c \leq -cV(x) + 3c1_{\|x\|^2\leq\frac{3d}{c}}.$$

In addition, as $\|\nabla^2 H(x)\| \leq L$, for some x' on the line segment between x and 0, $H(x) = H(0) + \frac{1}{2}x^T\nabla^2 H(x')x \leq \frac{1}{2}L\|x\|^2$. Thus, if $\|x\|^2 \leq \frac{3d}{c}$, $\frac{\sup_{x\in B}\nu(x)}{\inf_{x\in B}\nu(x)} \leq \exp\left(\frac{3dL}{2c}\right)$. So Definition 2 is verified. Applying Proposition 3, we get

$$q = \frac{1}{c}\left(1 + 3c\frac{3d}{c}\exp(\frac{3dL}{c})\right).$$

Next, note that because

$$H(x) - H(0) \geq H(x) - H(x/2) \geq \langle\nabla H(x/2), x/2\rangle \geq \frac{c}{4}\|x\|^2,$$

we have $\frac{1}{4}c\|x\|^2 \leq H(x) \leq \frac{1}{2}L\|x\|^2$, which implies that

$$\exp\left(-\frac{1}{2}L\|x\|^2\right) \leq \exp(-H(x)) \leq \exp\left(-\frac{1}{4}c\|x\|^2\right).$$

Therefore, $\int \exp(-H(x))dx \leq \int \exp\left(-\frac{1}{4}c\|x\|^2\right) dx \leq \left(\frac{4\pi}{c}\right)^{d/2}$, and for $\|x\|^2 \leq \frac{3d}{c}$, $\exp(-H(x)) \geq \exp(-\frac{3dL}{2c})$. This leads to our estimate of a:

$$a = \frac{\int_B \exp(-H(x))dx}{\exp(-H(x))V_d(3d/c)^{\frac{d}{2}}} \leq \frac{1}{V_d}\exp\left(\frac{3Ld}{2c}\right)\left(\frac{4\pi}{3d}\right)^{d/2}.$$

□

Appendix C.2. Proof of Proposition 5

Proof Without loss of generality, we assume $x_0 = 0$. Let $\nu_0 = \min_{x\in B(0,R)} \nu(x)$ and $H(x) = -\log \nu(x)$. Note that because $\mathcal{L}_\nu V(x) = -2\gamma\langle\nabla H(x), x\rangle + d\gamma$ and $\mathcal{L}_\nu V(x) \leq -\lambda V(x)$ when $\|x\| \geq R$, $\langle\nabla H(x), x\rangle \geq \frac{1}{2}\lambda\|x\|^2$. Let $y = \frac{R}{\|x\|}x$, then

$$\begin{aligned} H(x) - H(y) &= \int_0^1 \langle\nabla H(y + s(x-y)), x-y\rangle ds \\ &= \int_0^1 \langle\nabla H(y + s(x-y)), y + s(x-y)\rangle \frac{\|x\| - R}{R + s(\|x\| - R)} ds \\ &\geq \frac{1}{2}\lambda \int_0^1 \|y + s(x-y)\|\|x-y\| ds \\ &\geq \frac{1}{2}\lambda \int_0^1 \langle y + s(x-y), x-y\rangle ds = \frac{1}{4}\lambda(\|x\|^2 - R^2). \end{aligned}$$

Next, as $\|y\|^2 = R^2$,

$$\nu(x) \leq \nu(y)\exp\left(-\frac{1}{4}\lambda(\|x\|^2 - R^2)\right) \leq C\nu_0 \exp\left(-\frac{1}{4}\lambda(\|x\|^2 - R^2)\right),$$

Meanwhile, for $\|x\| \leq R$, $\nu(x) \leq C\nu_0 \leq C\nu_0\exp\left(-\frac{1}{4}\lambda(\|x\|^2 - R^2)\right)$. Then, because $\int \nu(x)dx = 1$, $1 \leq C\nu_0\exp\left(\frac{1}{4}\lambda R^2\right)\left(\frac{4\pi}{\lambda}\right)^{d/2}$. This implies

$$a \leq \frac{u_{B(0,R)}(x)}{\nu_0} \leq \frac{C}{V_d R^d}\exp\left(\frac{1}{4}\lambda R^2\right)\left(\frac{4\pi}{\lambda}\right)^{d/2}.$$

□

Appendix C.3. Proof of Proposition 6

Next we prove a more general version of Proposition 6:

Proposition 9 1. *If $\pi^Y(x) \propto \phi(x/M)$, then Assumption 2 holds with*

$$R^2 = 3M^2(2d+1), \quad Q = 2M^2\left(1 + \frac{9}{2}(2d+1)\exp(12d+8)\right),$$
$$A = \frac{1}{V_d}\left(\frac{2\pi}{3(2d+1)}\right)^{d/2}\exp(6d+4).$$

2. *Suppose $\pi(x) = \sum_{i=1}^{I} p_i\nu_i(x)$, where $\nu_i(x)$ are (c,L)-log concave densities with modes $\|m_i\| \le M$. If $\pi^Y(x) \propto \pi(x)^\beta$ with $\beta = d(2M^2c + 2M^2L^2/c)^{-1}$, then Assumption 2 holds with*

$$R^2 = 20M^2\left(1+\frac{L^2}{c^2}\right), \quad Q = M^2\left(1+\frac{L^2}{c^2}\right)\left(\frac{4}{d} + 100\exp\left(44\frac{dL}{c}\right)\right),$$
$$A = \frac{1}{V_d}\left(\frac{4\pi}{5d}\right)^{d/2}\exp\left(22\frac{dL}{c} + \frac{5}{4}d\right).$$

3. *If $\pi^Y(x) \propto \phi(x/M)\pi(x)^\beta$, where $\nu_i(x)$ are (c,L)-log concave densities with modes satisfying $\|m_i\| \le M$ and $\beta \le (dM^2c + dM^2L^2/c)^{-1}$, then Assumption 2 holds with*

$$R^2 = 5M^2(2d+1), \quad Q = 2M^2\left(1+\frac{25}{2}(2d+1)\exp(20d+30)\right),$$
$$A = \frac{1}{V_d}\left(\frac{8\pi}{5(2d+1)}\right)^{d/2}\exp(12d+16).$$

Proof For claim 1), $\pi^Y(x) \propto \phi(x/M)$. Consider $V_i(x) = \gamma\|x-m_i\|^2 + 1$ with $\gamma = \big(M^2(2d+1)\big)^{-1}$. We first note that

$$\begin{aligned}
\mathcal{L}_{\pi^Y}V_i(x) &= -\frac{2\gamma}{M^2}\langle x-m_i, x\rangle + 2d\gamma\\
&= -\frac{\gamma}{M^2}\|x-m_i\|^2 + \frac{\gamma}{M^2}\|m_i\|^2 - \frac{\gamma}{M^2}\|x\|^2 + 2d\gamma\\
&\le -\frac{\gamma}{M^2}\|x-m_i\|^2 + (2d+1)\gamma\\
&\le -\frac{1}{2M^2}V_i(x) + \left(\frac{1}{2M^2} + (2d+1)\gamma\right)1_{\|x-m_i\|^2\le 3M^2(2d+1)}\\
&= -\frac{1}{2M^2}V_i(x) + \frac{3}{2M^2}1_{\|x-m_i\|^2\le 3M^2(2d+1)}.
\end{aligned} \tag{C.1}$$

Then, the bounding constants for the Lyapunov function are

$$\lambda = \frac{1}{2M^2}, \quad h = \frac{3}{2M^2}, \quad R^2 = 3M^2(2d+1),$$

In addition, for $x \in B(m_i, R)$, we have $\|x\|^2 \leq 2R^2 + 2M^2$. Thus, the density ratio can be bounded by $C = \exp\left(\frac{R^2+M^2}{M^2}\right) = \exp(6d+4)$. By Proposition 3,

$$Q = (1 + hR^2C^2)/\lambda = 2M^2 + 9M^2(2d+1)\exp(12d+8).$$

Moreover, $A = \frac{(2\pi M^2)^{d/2}\exp(\frac{R^2+M^2}{M^2})}{V_d R^d} = \frac{1}{V_d}\left(\frac{2\pi}{3(2d+1)}\right)^{d/2} \exp\left(6d+4\right)$.

For claim 2), $\pi^Y(x) \propto \pi(x)^\beta$. Consider $V_i(x) = \gamma\|x - m_i\|^2 + 1$ with

$$\gamma \leq \left(2M^2 + 2M^2\frac{L^2}{c^2} + \frac{2d}{\beta c}\right)^{-1}.$$

Let $H_i(x) = -\log \nu_i(x)$. We first note that

$$\nabla \log \pi^Y(x) = \beta \nabla \log \pi(x) = -\frac{\beta\sum_{i=1}^I p_i \nu_i(x) \nabla H_i(x)}{\sum_{i=1}^I p_i \nu_i(x)}$$

and

$$\begin{aligned}
-\langle \nabla V_j(x), &\nabla H_i(x)\rangle = -2\gamma\langle x - m_j, \nabla H_i(x) - \nabla H_i(m_i)\rangle \\
&\leq -2\gamma\langle x - m_i, \nabla H_i(x) - \nabla H_i(m_i)\rangle + 2\gamma\langle m_j - m_i, \nabla H_i(x) - \nabla H_i(m_i)\rangle \\
&\leq -2c\gamma\|x - m_i\|^2 + 2\gamma L\|m_j - m_i\|\|x - m_i\| \\
&\leq -c\gamma\|x - m_i\|^2 + \frac{\gamma L^2}{c}\|m_j - m_i\|^2 \\
&\leq -\frac{1}{2}c\gamma\|x - m_j\|^2 + c\gamma\|m_j - m_i\|^2 + \frac{\gamma L^2}{c}\|m_j - m_i\|^2 \\
&\leq -\frac{1}{2}cV_j(x) + 2c\gamma M^2 + 2\frac{\gamma L^2}{c}M^2 + \frac{1}{2}c.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathcal{L}_{\pi^Y} V_j(x) &\leq -\frac{1}{2}\beta c V_j(x) + 2\beta c\gamma M^2 + 2\beta\frac{\gamma L^2}{c}M^2 + \frac{1}{2}\beta c + 2\gamma d \\
&= -\frac{1}{4}\beta c V_j(x) - \frac{1}{4}\beta c\gamma\|x - m_j\|^2 + \beta c\gamma\left(2M^2 + 2\frac{L^2M^2}{c^2} + \frac{2d}{\beta c}\right) + \frac{1}{4}\beta c \\
&\leq -\frac{1}{4}\beta c V_j(x) + \frac{5}{4}\beta c 1_{\|x - m_j\|^2 \leq \frac{5}{\gamma}}.
\end{aligned}$$

For $\beta = d\left(M^2c + M^2L^2/c\right)^{-1}$, $R^2 = \frac{5}{\gamma} = 20M^2\left(1 + \frac{L^2}{c^2}\right)$. Next, we note that if $\|x - m_j\|^2 \leq \frac{5}{\gamma}$, $\|x - m_i\|^2 \leq \frac{10}{\gamma} + 4M^2 \leq \frac{11}{5}R^2$. Then, note that for any region B, if we let $\psi(x) := \max_i \frac{\max_{x\in B}\nu_i(x)}{\min_{x\in B}\nu_i(x)}$,

$$\frac{\max_{x\in B}\pi^Y(x)}{\min_{x\in B}\pi^Y(x)} \leq \frac{(\sum_i p_i \max_{x\in B}\nu_i(x))^\beta}{(\sum_i p_i \min_{x\in B}\nu_i(x))^\beta} \leq \frac{(\sum_i p_i \psi \min_{x\in B}\nu_i(x))^\beta}{(\sum_i p_i \min_{x\in B}\nu_i(x))^\beta} = (\psi(x))^\beta.$$

Therefore

$$C = \frac{\max_{B(m_j,R)} \pi^Y(x)}{\min_{B(m_j,R)} \pi^Y(x)} \leq \max_i \left(\frac{\max_{B(m_j,R)} \nu_i(x)}{\min_{B(m_j,R)} \nu_i(x)} \right)^\beta$$

$$\leq \exp\left(\frac{1}{2}L\frac{11}{5}R^2\beta\right) = \exp\left(22\frac{dL}{c}\right).$$

By Proposition 3,

$$Q = \frac{1+\frac{5}{4}\beta c \frac{5}{\gamma}\exp(44dL/c)}{\frac{1}{4}\beta c} = M^2\left(1+\frac{L^2}{c^2}\right)\left(\frac{4}{d}+100\exp\left(44\frac{dL}{c}\right)\right).$$

The estimate of A can be obtained by Lemma 5,

$$A = \frac{C}{V_d}\exp\left(\frac{1}{16}\beta c R^2\right)\left(\frac{16\pi}{\beta c R^2}\right)^{d/2} = \frac{1}{V_d}\exp\left(22\frac{dL}{c}+\frac{5}{4}d\right)\left(\frac{4\pi}{5d}\right)^{d/2}.$$

For claim 3), $\pi^Y(x) \propto \phi(x/M)\pi(x)^\beta$. Consider $V_i(x) = \gamma\|x-m_i\|^2+1$. Combining our analysis in claim 1) and claim 2), we have

$$\begin{aligned}
\mathcal{L}_{\pi^Y}V_j(x) \leq & -\left(\frac{1}{2M^2}+\frac{\beta c}{4}\right)V_j(x) - \left(\frac{1}{2M^2}+\frac{\beta c}{4}\right)\gamma\|x-m_j\|^2 \\
& + \frac{1}{2M^2} + \beta c\gamma\left(2M^2+2\frac{L^2M^2}{c^2}+\frac{2d+1}{\beta c}\right)+\frac{1}{4}\beta c \\
\leq & -\frac{1}{2M^2}V_j(x) - \frac{1}{2M^2}\gamma\|x-m_j\|^2 \\
& + \frac{1}{2M^2} + \beta c\gamma\left(2M^2+2\frac{L^2M^2}{c^2}\right)+\gamma(2d+1).
\end{aligned}$$

For $\beta \leq (dM^2c + dM^2L^2/c)^{-1}$ and $\gamma = (M^2(2d+1))^{-1}$, we can set

$$R^2 = \frac{5}{\gamma} = 5M^2(2d+1) \text{ and } h = \frac{5}{2M^2}.$$

Then $\mathcal{L}_{\pi^Y}V_j(x) \leq -\frac{1}{2M^2}V_j(x) + h1_{\|x-m_j\|^2\leq R^2}$. We next note that as

$$\frac{\max_{x\in B(m_j,R)}\phi(x/M)}{\min_{x\in B(m_j,R)}\phi(x/M)} \leq \exp\left(\frac{R^2+M^2}{M^2}\right) = \exp(10d+6) \text{ and}$$

$$\frac{\max_{x\in B(m_j,R)}\pi(x)^\beta}{\min_{x\in B(m_j,R)}\pi(x)^\beta} \leq \exp\left(\frac{1}{2}L(2R^2+4M^2)\beta\right) \leq \exp\left(\frac{10+7/d}{1+L^2/c^2}\frac{L}{c}\right) \leq \exp(17/2),$$

we have

$$C \leq \frac{\max_{x\in B(m_j,R)}\phi(x/M)}{\min_{x\in B(m_j,R)}\phi(x/M)}\frac{\max_{x\in B(m_j,R)}\pi(x)^\beta}{\min_{x\in B(m_j,R)}\pi(x)^\beta} \leq \exp(10d+15).$$

By Proposition 3, $Q = \frac{1+hR^2C^2}{\frac{1}{2M^2}} = 2M^2\left(1 + \frac{25}{2}(2d+1)\exp\left(20d+30\right)\right)$. Lastly, the constant A can be obtain from Lemma 5:

$$A = \frac{C}{V_d}\exp\left(\frac{R^2}{8M^2}\right)\left(\frac{8M^2\pi}{R^2}\right)^{d/2} \leq \frac{1}{V_d}\exp\left(12d+16\right)\left(\frac{8\pi}{5(2d+1)}\right)^{d/2}.$$

□

Appendix C.4. Proof of Corollary 1

Proof Since ν_i is $(l_M^{-2}, l_m^{-2}\mathcal{C})$-log concave, by Lemma 4, Assumption 1 holds with $q = l_M^2 + 9dl_M^2\exp(3d\mathcal{C})$, $\quad r_i^2 = 3dl_i^2$, $\quad a = \frac{1}{V_d}\left(\frac{4\pi}{3d}\right)^{d/2}\exp\left(\frac{3d\mathcal{C}}{2}\right)$. Choosing $\pi^Y(x) \propto \phi(x/M)(\pi(x))^\beta$ with $\beta = \frac{1}{\tau} \leq (dM^2l_M^{-2} + dM^2l_M^2l_m^{-4}\mathcal{C}^2)^{-1}$, Proposition 6 gives us $R^2 = O(M^2d)$, $Q = O(M^2d\exp(20d))$, and $A = O(\exp(12d))$. Plug these estimates and $r^2 = 3dl_m^2$ into Theorem 2, we have

$$\kappa = \max\left\{3(56A+1)q,\ \frac{3}{\tau}\left(57Q + 14aA\left(\frac{R^{d+1}}{r^{d-1}}\right)\left(\log\left(\frac{R}{r}\right)\right)^{1_{d=1}}\right), \frac{7aA}{\rho}\left(\frac{R}{r}\right)^d\right\}$$
$$= O\left(\exp(\mathcal{C}Dd)\max\left\{dl_M^2, \frac{1}{\tau}dMl_m\left(\frac{M}{l_m}\right)^d, \frac{1}{\rho}\left(\frac{M}{l_m}\right)^d\right\}\right).$$

□

Appendix C.5. Proof of Corollary 2

To prove Corollary 2, we first introduce an auxiliary lemma.

Lemma 3 *For any given $\beta \in (0,1]$, if ν is a $(l^{-2}, l^{-2}\mathcal{C})$-log concave density with mode m, then $\mu(x) \propto (\nu(x))^\beta$ is a $\mathbf{Ly}(R_\beta, q_\beta, a_\beta)$ with $\lambda_\beta = \beta l^{-2}$, and a suitable constant D so that*

$$q_\beta = O\left(\frac{d\exp(Dd\mathcal{C})}{\lambda_\beta}\right), \quad R_\beta^2 = \frac{4d}{\lambda_\beta} = O\left(\frac{d}{\lambda_\beta}\right), \quad A_\beta = O(\exp(Dd\mathcal{C})).$$

Proof We consider using $V(x) = \gamma\|x-m\|^2 + 1$, with $\gamma = \lambda_\beta/(2d)$. Denote $H(x) = -\log\nu(x)$. Then,

$$\begin{aligned}\mathcal{L}_\mu V(x) &= -2\gamma\beta\langle x-m, \nabla H(x)\rangle + 2d\gamma \\ &\leq -2\gamma\beta l^{-2}\|x-m\|^2 + 2d\gamma \\ &= -2\lambda_\beta\gamma\|x-m\|^2 + 2d\gamma \\ &\leq -\lambda_\beta V(x) + \left(-\lambda_\beta\gamma\|x-m\|^2 + \lambda_\beta + 2d\gamma\right) \\ &\leq -\lambda_\beta V(x) + b_\beta 1_{\|x-m\|^2 \leq R_\beta^2}.\end{aligned}$$

where $b_\beta = \lambda_\beta + 2d\gamma = 2\lambda_\beta$ by our choice of γ, and $R_\beta^2 = \frac{b_\beta}{\gamma\lambda_\beta} = \frac{4d}{\lambda_\beta}$. Note that

$$\max(\log\mu(x) - \log\mu(y)) = \beta\max(\log\nu(x) - \log\nu(y)).$$

Because $\beta\max_{x,y\in B(m,R_\beta)}(\log\nu(x) - \log\nu(y)) \leq \frac{1}{2}\beta l^{-2}\mathcal{C}R_\beta^2 \leq 2d\mathcal{C}, C_\beta \leq \exp(2d\mathcal{C})$. Lastly, by Lemma 5, $A_\beta = \frac{C_\beta}{V_d}\exp\left(\frac{1}{2}\lambda_\beta R_\beta^2\right)\left(\frac{4\pi}{\lambda_\beta R_\beta^2}\right)^{d/2} = O(\exp(Dd\mathcal{C}))$, and by Definition 3,$q_\beta = \frac{1+R_\beta^2 C_\beta^2 b_\beta}{\lambda_\beta} = O\left(\frac{d\exp(Dd\mathcal{C})}{\lambda_\beta}\right)$. □

Proof [Proof of Corollary 2] Consider the following density:

$$\pi'_k(x) \propto \sum_{i=1}^{I} p_i(\nu_i(x))^{\beta_k}, k = 1, \ldots, K-1, \quad \pi'_0 = \pi_0, \pi'_K = \pi_K.$$

By Lemma 3, π'_k satisfies Assumptions 4 and 5 with

$$r_{k,i}^2 = \frac{4dl_i^2}{\beta_k}, \quad r_k^2 = \frac{4dl_m^2}{\beta_k}, \quad q_k = O(1), \quad a_k = O(1), \quad k = 0, \ldots, K-1.$$

Moreover, by Lemma 4, $q_0 = l_M^2(1 + 9d\exp(3d\mathcal{C}))$. From Proposition 6, for $\beta_K \leq \frac{1}{dM^2(l_M^{-2} + l_m^{-4}l_M^2\mathcal{C}^2)}$, we have

$$r_{K,i}^2 = O(M^2 d) = O(1),\ q_K = O(M^2 d\exp(20d)) = O(1),\ a_K = O(\exp(10d)) = O(1).$$

Then, by Theorem 3 $\mathrm{var}_{\pi'_{0:K}}(f(\mathbf{X}_{0:K})) \leq \kappa'\mathbb{E}_{\pi'_{0:K}}[\Gamma_R^K(f(\mathbf{X}_{0:K}))]$, with $\kappa' = O(l_M^{2\alpha})$ for some $\alpha \leq 1$, if the parameters τ_k, ρ satisfy

$$\begin{aligned}&\tau_k \geq U l_M^{-2\alpha}\left(\frac{\beta_{k-1}}{\beta_k}\right)^{d/2}, \quad k = 1, \ldots, K-1, \quad \tau_K \geq U l_M^{-2\alpha}\left(\frac{\beta_{K-1}}{l_m^2}\right)^{d/2}, \\ &\rho \geq U l_M^{-2\alpha} max\left\{\left(\frac{\beta_{K-1}}{l_m^2}\right)^{d/2}, \left(\frac{\beta_{k-1}}{\beta_k}\right)^{d/2}\right\}.\end{aligned} \tag{C.2}$$

for some $U > 0$. Note that $\left(\sum_{i=1}^{I} p_i\nu_i(x)\right)^{\beta_k} \geq \sum_{i=1}^{I} p_i\nu_i(x)^{\beta_k}$, since x^{β_k} is concave for $0 \leq \beta_k \leq 1$. On the other hand, for $p_0 = \min_{i\leq I} p_i$,

$$\left(\sum_{i=1}^{I} p_i\nu_i(x)\right)^{\beta_k} \leq \max_i \nu_i(x)^{\beta_k} \leq \frac{1}{p_0}\sum_{i=1}^{I} p_i\nu_i(x)^{\beta_k}.$$

Therefore, $p_0\pi_k(x) \leq \pi'_k(x) \leq \pi_k(x) \leq \frac{1}{p_0}\pi_k(x)$ for $k = 1, \ldots, K-1$. By Proposition 2, $\mathrm{var}_{\pi_{0:K}}(f(\mathbf{X}_{0:K})) \leq \kappa\mathbb{E}_{\pi_{0:K}}[\Gamma_R^K(f(\mathbf{X}_{0:K}))]$, with $\kappa = p_0^{-2K}\kappa'$. We next verify that (C.2) holds. In scenario 1, for $k = 1, \ldots, K$, as $\beta_k = l_m^{2k/K}$,

$$l_M^{-2\alpha}\left(\frac{\beta_{k-1}}{\beta_k}\right)^{d/2} \leq l_m^{-2\alpha-d/K} \leq \tau_k \text{ and } l_M^{-2\alpha}\max_k\left(\frac{\beta_{k-1}}{\beta_k}\right)^{d/2} \leq l_m^{-2\alpha-d/K} \leq \rho.$$

Thus, (C.2) holds. In scenario 2, $\beta_k = l_m^{\frac{2k}{K}} < 1$ and $d \leq 2$,

$$\left(\frac{\beta_{k-1}}{\beta_k}\right)^{d/2} = l_m^{-d/K} \leq l_m^{-2k/K} = \tau_k, \quad \max_k\left(\frac{\beta_{k-1}}{\beta_k}\right)^{d/2} = l_m^{-d/K} \leq \rho.$$

Thus, (C.2) holds with $\alpha = 0$. In scenario 3, note that with our choice of β_k and τ_k, $k = 1, \ldots, K$

$$\left(\frac{\beta_{k-1}}{\beta_k}\right)^{d/2} = l_m^{\frac{d}{2}(2(\frac{d-2}{d})^{K-k+1}-2(\frac{d-2}{d})^{K-k})} = l_m^{-2(\frac{d-2}{d})^{K-k}} = \tau_k.$$

Meanwhile, because $\tau_0 = 1 = l_m^{-d(\frac{d-2}{d})^{K-1}}\left(\frac{\beta_0}{\beta_1}\right)^{d/2}$, (C.2) holds with constant $\alpha = -d(\frac{d-2}{d})^{K-1}$. □

Appendix C.6. Proof of Proposition 7

Before we prove Proposition 7, we first present an auxiliary lemma.

Lemma 4 *Suppose $V(x) = \exp(\frac{1}{2}H(x))$ is $\mathbb{C}^2(\mathbb{R}^d)$ with $H(x) = -\infty$ and $V(x) = 0$ for $x \notin \Omega$. Moreover, for a region $B \subset \Omega$, $\frac{1}{2}\Delta H(x) - \frac{1}{4}\|\nabla H(x)\|^2 \leq -\lambda_0$ for $x \in \Omega \setminus B$. Then $V(x)$ is a (λ_0, h, B, C)-Lyapunov function for $\nu \propto \exp(-H(x))$ with*

$$h = \max_{x\in B}\left(-\frac{1}{4}\|\nabla H(x)\|^2 + \frac{1}{2}\Delta H(x) + \lambda_0\right)V(x), \quad C = \frac{\max_{x\in B}\nu(x)}{\min_{x\in B}\nu(x)}.$$

Proof For $x \notin \Omega$, $\mathcal{L}_\nu V(x) = 0$. For $x \in \Omega$,

$$\mathcal{L}_\nu V(x) = \left(-\frac{1}{4}\|\nabla H(x)\|^2 + \frac{1}{2}\Delta H(x)\right) V(x) \leq -\lambda_0 V(x) + h 1_{x\in B}.$$

□

Proof [Proof of Proposition 7] Consider a clamp function $\psi : \mathbb{R} \to \mathbb{R}$ satisfying 1) ψ is $\mathbb{C}^2$; 2) $\dot{\psi} < 0, \ddot{\psi}/(\dot{\psi})^2 \leq C$; 3) $\psi(x) = 1$ for all $x \leq 0$; 4) $\psi(x) = 0$ for all $x \geq 1$. Let $\Psi_i(x) = \exp\left(\frac{1}{\epsilon}\log\psi(\sqrt{n}d_i(x))\right)$. Then, we can construct

$$\pi_\epsilon \propto \sum_{i=1}^{I} \Psi_i(x)\exp(-\tfrac{1}{\epsilon}H_\epsilon(x)) = \sum_{i=1}^{I}\exp\left(-\frac{1}{\epsilon}Q_{\epsilon,i}(x)\right),$$

where $Q_{\epsilon,i}(x) = -\log\psi(\sqrt{n}d_i(x)) + H_\epsilon(x)$. We next verify that

$$\frac{1}{2\epsilon}\Delta Q_{\epsilon,i} - \frac{1}{4\epsilon^2}\|\nabla Q_{\epsilon,i}\|^2 \leq -\frac{\lambda_0}{\epsilon}. \tag{C.3}$$

Note that (C.3) holds for any $x \in \Omega_i$ since $Q_{\epsilon,i}(x) = H_\epsilon(x)$. When $x \in \Omega_i' \setminus \Omega_i$,

$$\nabla Q_{\epsilon,i}(x) = -\sqrt{n}\frac{\dot{\psi}(\sqrt{n}d_i(x))}{\psi(\sqrt{n}d_i(x))}\nabla d_i(x) + \nabla H_\epsilon(x)$$

We first note that because i) $\nabla d_i(x_n) \to v_\perp(x)$ for any $x_n \to x \in \partial\Omega_i$, ii) $-\nabla H_\epsilon(x)$ points toward the inside of Ω_i for $x \in \partial\Omega_i$, and iii) $\nabla^2 d_i$ and $\nabla^2 H_\epsilon$ are bounded, for n large enough, $-\langle\nabla d_i(x), \nabla H_\epsilon(x)\rangle < 0$ for $x \in \Omega_i' \setminus \Omega_i$. Then,

$$\frac{1}{4\epsilon^2}\|\nabla Q_{\epsilon,n}(x)\|^2 \geq \frac{1}{4\epsilon^2}n\|\nabla d_i(x)\|^2\frac{\dot{\psi}(\sqrt{n}d_i(x))^2}{\psi(\sqrt{n}d_i(x))^2} + \frac{1}{4\epsilon^2}\|\nabla H_\epsilon(x)\|^2.$$

We next note that

$$\begin{aligned}\Delta Q_{\epsilon,n}(x) = &- n\frac{\ddot{\psi}(\sqrt{n}d_i(x))}{\psi(\sqrt{n}d_i(x))}\|\nabla d_i(x)\|^2 + n\frac{\dot{\psi}(\sqrt{n}d_i(x))^2}{\psi(\sqrt{n}d_i(x))^2}\|\nabla d_i(x)\|^2 \\ &- \sqrt{n}\frac{\dot{\psi}(\sqrt{n}d_i(x))}{\psi(\sqrt{n}d_i(x))}\Delta d_i(x) + \Delta H_\epsilon(x)\end{aligned}$$

Thus, for ϵ small enough,

$$\begin{aligned}&\frac{1}{2\epsilon}\Delta Q_\epsilon(x) - \frac{1}{4\epsilon^2}\|\nabla Q_{\epsilon,n}(x)\|^2 \\ \leq &- \frac{n}{2\epsilon}\frac{\ddot{\psi}(\sqrt{n}d_i(x))}{\psi(\sqrt{n}d_i(x))}\|\nabla d_i(x)\|^2 + \frac{n}{2\epsilon}\frac{\dot{\psi}(\sqrt{n}d_i(x))^2}{\psi(\sqrt{n}d_i(x))^2}\|\nabla d_i(x)\|^2 - \frac{n}{2\epsilon}\frac{\dot{\psi}(\sqrt{n}d_i(x))}{\psi(\sqrt{n}d_i(x))}\Delta d_i(x) \\ &+ \frac{1}{2\epsilon}\Delta H_\epsilon(x) - \frac{1}{4\epsilon^2}n\frac{\dot{\psi}(\sqrt{n}d_i(x))^2}{\psi(\sqrt{n}d_i(x))^2}\|\nabla d_i(x)\|^2 - \frac{1}{4\epsilon^2}\|\nabla H_\epsilon(x)\|^2 \\ \leq &\frac{1}{2\epsilon}\Delta H_\epsilon(x) - \frac{1}{4\epsilon^2}\|\nabla H_\epsilon(x)\|^2 \leq -\frac{\lambda_0}{\epsilon}\end{aligned}$$

Lastly, we note that

$$\exp\left(-\frac{1}{\epsilon}H_\epsilon(x)\right) \leq \sum_{i=1}^{I} \exp\left(-\frac{1}{\epsilon}Q_{\epsilon,i}(x)\right) \leq I \exp\left(-\frac{1}{\epsilon}H_\epsilon(x)\right).$$

Moreover $q(x) \propto \exp\left(-\frac{1}{\epsilon}H_\epsilon(x)\right)$ is a ϵ perturbation of π. Thus, π_ϵ is a ϵ perturbation of π. □

Appendix C.7. Proof of Corollary 3

Proof Let $H(x) = \frac{1}{2}(x^2 - a^2)^2$ and $\epsilon = 1/n$. We first note that $\nabla H(x) = 2x(x^2 - a^2)$ and $\nabla^2 H(x) = 6x^2 - 2a^2$. Thus,

$$\frac{1}{2\epsilon}\nabla^2 H(x) - \frac{1}{4\epsilon^2}\|\nabla H(x)\|^2 = \frac{3x^2}{\epsilon} - \frac{a^2}{\epsilon} - \frac{1}{\epsilon^2}x^2(x^2 - a^2)^2.$$

When $|x - a|^2 \geq 3\epsilon/a^2$ and $x > 0$,

$$\frac{1}{\epsilon^2}x^2(x^2 - a^2)^2 = \frac{1}{\epsilon^2}x^2(x - a)^2(x + a)^2 \geq \frac{1}{\epsilon^2}x^2\frac{3\epsilon}{a^2}a^2 \geq \frac{3x^2}{\epsilon}.$$

Then, $\frac{1}{2\epsilon}\nabla^2 H(x) - \frac{1}{4\epsilon^2}\|\nabla H(x)\|^2 \leq -\frac{a^2}{\epsilon}$. Similarly, when $|x + a|^2 \geq 3\epsilon/a^2$ and $x < 0$, we also have $\frac{1}{2\epsilon}\nabla^2 H(x) - \frac{1}{4\epsilon^2}\|\nabla H(x)\|^2 \leq -\frac{a^2}{\epsilon}$. In this case, $H_\epsilon = H$ already satisfies (12). (There is no saddle point for this problem.)

Next if we split R into $\Omega_1 = [0, \infty)$ and $\Omega_2 = (-\infty, 0]$. It is easy to see that $d_1(x) = -x$ is $\mathbb{C}^2$ in $(-\infty, 0)$. In addition, $\nabla d_1(x) = -1$, which is the same as the outward direction for Ω_1 at $x = 0$. Similarly, $d_2(x) = x$ is $\mathbb{C}^2$ in $(0, \infty)$ and $\nabla d_2(x) = 1$ is the same as the outward direction for Ω_2 at $x = 0$. Thus, the existence of the π_ϵ follows from Proposition 7.

□

Appendix D. Proof of Proposition 8

Before we prove Proposition 8, we first present some auxiliary lemmas. Our first result shows that we can replace a density having a $(\lambda, b, B(x_0, R), C)$-Lyapunov Lyapunov function with a uniform distribution, while keeping the difference controlled.

Lemma 5 *Suppose ν has a $(\lambda, b, B(x_0, R), C)$-Lyapunov function, then*

$$(\mathbb{E}_{\nu}[f(X)] - \mathbb{E}_{u_B}[f(U)])^2 \leq 2\frac{1+(b+\lambda)R^2C^2}{\lambda}\mathbb{E}_{\nu}[\|\nabla f(X)\|^2].$$

Proof Let $\bar{f}_{\nu} = \mathbb{E}_{\nu}[f(X)]$ and $\bar{f}_{u_B} = \mathbb{E}_{u_B}[f(U)]$. Then

$$\begin{aligned}
\left(\bar{f}_{\nu} - \bar{f}_{u_B}\right)^2 &\leq 2\mathbb{E}_{\nu}[(f(X)-\bar{f}_{\nu})^2] + 2\mathbb{E}_{\nu}[(f(X)-\bar{f}_{u_B})^2]\\
&\leq 2\frac{1+bR^2C^2}{\lambda}\mathbb{E}_{\nu}[\|\nabla f(X)\|^2] + 2C\mathbb{E}_{u_B}[(f(X)-\bar{f}_{u_B})^2] \text{ by Proposition 3}\\
&\leq 2\frac{1+bR^2C^2}{\lambda}\mathbb{E}_{\nu}[\|\nabla f(X)\|^2] + 2CR^2\mathbb{E}_{u_B}[\|\nabla f(X)\|^2] \text{ by Lemma 1}\\
&\leq 2\frac{1+bR^2C^2}{\lambda}\mathbb{E}_{\nu}[\|\nabla f(X)\|^2] + 2C^2R^2\mathbb{E}_{\nu}[\|\nabla f(X)\|^2]\\
&= 2\frac{1+(b+\lambda)R^2C^2}{\lambda}\mathbb{E}_{\nu}[\|\nabla f(X)\|^2].
\end{aligned}$$

□

Our second result bounds the mean difference square when moving from a big Uniform ball to a small Uniform ball with the same center.

Lemma 6 *Consider $B_r = B(x_0, r)$ and $B_R = B(x_0, R)$ with $R \geq r$. Then when $d = 1$, $\left(\mathbb{E}_{u_{B_r}}[f(X)] - \mathbb{E}_{u_{B_R}}[f(X)]\right)^2 \leq R^2 \log(R/r)\mathbb{E}_{u_{B_R}}[\|\nabla f(X)\|^2]$; when $d \geq 2$, $\left(\mathbb{E}_{u_{B_r}}[f(X)] - \mathbb{E}_{u_{B_R}}[f(X)]\right)^2 \leq \frac{R^{d+1}}{(d-1)r^{d-1}}\mathbb{E}_{u_{B_R}}[\|\nabla f(X)\|^2]$.*

Proof Without loss of generality, we assume $x_0 = 0$.

We first consider the case in which $r = 1$ and $d \geq 2$. Let C_V denote the volume of a d-dimensional unit ball. Consider the spherical coordinate of x. In particular, let $t \in [0, R]$ denote the radial coordinate, and $\theta = (\theta_1, \theta_2, \ldots, \theta_{d-1})$ denote the angular coordinate, i.e., it is a $(d-1)$ dimensional vector with $\theta_i \in [0, \pi]$ for $i = 1, \ldots, d-2$ and $\theta_{n-1} \in [0, 2\pi)$. We also write $\xi(\theta)$ be a d-dimensional vector on S^{d-1} with $\xi_1(\theta) = \cos(\theta_1)$, $\xi_i(\theta) = \sin(\theta_1)\cdots\sin(\theta_{i-1})\cos(\theta_i)$. for $1 < i < d$, and $\xi_d(\theta) = \sin(\theta_1)\cdots\sin(\theta_{d-1})$. Then, $x = r\xi(\theta)$. We also write

$$d_{S^{d-1}}\theta = \sin^{d-2}(\theta_1)\sin^{d-3}(\theta_2)\ldots\sin(\theta_{d-1})d\theta$$

and $\Omega = [0, \pi]^{d-2} \times [0, 2\pi)$. Then

$$\mathbb{E}_{u_{B_R}}[f(X)] = \frac{1}{C_V R^d}\int_{\Omega}\int_0^R f(t\xi(\theta))t^{d-1}dtd_{S^{d-1}}\theta = \frac{1}{C_V}\int_{\Omega}\int_0^1 f(Rt\xi(\theta))t^{d-1}dtd_{S^{d-1}}\theta.$$

Using the spherical coordinate representation, we have

$$
\begin{aligned}
&\Big(\mathbb{E}_{u_{B_1}}[f(X)] - \mathbb{E}_{u_{B_R}}[f(X)]\Big)^2 \\
=&\left(\frac{1}{C_V}\int_{\Omega}\int_0^1 f(t\xi(\theta))t^{d-1}dtd_{S^{d-1}}\theta - \frac{1}{C_V}\int_{\Omega}\int_0^1 f(Rt\xi(\theta))t^{d-1}dtd_{S^{d-1}}\theta\right)^2 \\
\leq&\frac{1}{C_V}\int_{\Omega}\int_0^1 (f(Rt\xi(\theta)) - f(t\xi(\theta)))^2 t^{d-1}dtd_{S^{d-1}}\theta \ \text{ by Jensen's inequality} \\
=&\frac{1}{C_V}\int_{\Omega}\int_0^1 \left(\int_1^R \sum_{i=1}^d \nabla_{x_i} f(st\xi(\theta))t\xi_i(\theta)ds\right)^2 t^{d-1}dtd_{S^{d-1}}\theta \\
\leq&\frac{1}{C_V}\int_{\Omega}\int_0^1 R\int_1^R \left(\sum_{i=1}^d \nabla_{x_i} f(st\xi(\theta))t\xi_i(\theta)\right)^2 dst^{d-1}dtd_{S^{d-1}}\theta \ \text{ by Jensen's inequality} \\
\leq&\frac{R}{C_V}\int_{\Omega}\int_0^1\int_1^R \|\nabla f(st\xi(\theta))\|^2 dst^{d+1}dtd_{S^{d-1}}\theta \ \text{ by Cauchy-Schwarz inequality and } \|\xi\| = 1 \\
=&\frac{R}{C_V}\int_1^R\int_{\Omega}\int_0^1 \|\nabla f(st\xi(\theta))\|^2 t^{d+1}dtd_{S^{d-1}}\theta ds \\
=&\frac{R}{C_V}\int_1^R\int_{\Omega}\int_0^s \|\nabla f(r\xi(\theta))\|^2 r^{d+1}drd_{S^{d-1}}\theta\frac{1}{s^{d+2}}ds \text{ by letting } r = st \\
\leq&\frac{R}{C_V}\int_1^R\int_{\Omega}\int_0^s \|\nabla f(r\xi(\theta))\|^2 r^{d-1}drd_{S^{d-1}}\theta\frac{1}{s^{d}}ds \\
\leq&R^{d+1}\left(\frac{1}{C_V R^d}\int_{\Omega}\int_0^R \|\nabla f(r\xi(\theta))\|^2 r^{d-1}drd_{S^{d-1}}\theta\right)\int_1^R\frac{1}{s^d}ds \leq \frac{R^{d+1}}{d-1}\mathbb{E}_{u_{B_R}}[\|\nabla f(X)\|^2].
\end{aligned}
$$

When $d = 1$**,** following similar arguments as above, we can show that

$$\Big(\mathbb{E}_{u_{B_1}}[f(X)] - \mathbb{E}_{u_{B_R}}[f(X)]\Big)^2 = \left(\frac{1}{2}\int_{-1}^1 f(Rt) - f(t)dt\right)^2 \leq \frac{1}{2}\int_{-1}^1 R\int_1^R \|\nabla f(st)\|^2 dst^2 dt$$

We then change variable by $r = st$ and find

$$\int_{-1}^1\int_1^R \|\nabla f(st)\|^2 dst^2 dt \leq \int_1^R\int_{-s}^s \|\nabla f(r)\|^2 dr\frac{s^2}{s^3}ds \leq \log(R){-}1)\int_{-R}^R \|\nabla f(r)\|^2 dr.$$

For general $r > 0$**,** we can simply set $Z = X/r$, $g(X) = f(X/r)$ and $q = R/r$.
Because $\mathbb{E}_{u_{B_1}}[g(Z)] = \mathbb{E}_{u_{B_r}}[f(X)]$ and $\mathbb{E}_{u_{B_q}}[g(Z)] = \mathbb{E}_{u_{B_R}}[f(X)]$,

$$\Big(\mathbb{E}_{u_{B_r}}[f(X)] - \mathbb{E}_{u_{B_R}}[f(X)]\Big)^2 = \Big(\mathbb{E}_{u_{B_1}}[g(Z)] - \mathbb{E}_{u_{B_q}}[g(Z)]\Big)^2 \leq \frac{q^{d+1}}{d-1}\mathbb{E}_{u_{B_q}}[\|\nabla g(X)\|^2].$$

Then, as $\mathbb{E}_{u_{B_q}}[\|\nabla g(Z)\|^2] = r^2\mathbb{E}_{u_{B_R}}[\|\nabla f(X)\|^2]$, we have the claim. □

Proof [Proof of Proposition 8] To simplify the notations, we define $\bar{\mathcal{B}}_i = B(m_i, R_i)$, $B_i = B(m_i, r_i)$, and $\eta_i(y) = \int f(x,y)\nu_i^x(x)dx$ for $i = 1, 2$. Let $\tilde{Q} = Q + R^2A^2$ and $\tilde{q} = q + r^2a^2$.

Step 1. Replace ν_2^Y with $u_{\bar{\mathcal{B}}_2}$. By Lemma 5, we can control the difference by

$$
\begin{aligned}
&\Big(\mathbb{E}_{\nu_1^X\otimes\nu_2^Y}[f(X,Y)] - \mathbb{E}_{\nu_1^X\otimes u_{\bar{\mathcal{B}}_2}}[f(X,Y)]\Big)^2\\
=&\left(\int \eta_1(y)\nu_2^Y(y)dy - \int \eta_1(y)u_{\bar{\mathcal{B}}_2}(y)dy\right)^2 \qquad\qquad\qquad\qquad .\\
\le& 2\tilde{Q}\int \|\nabla_y \eta_1(y)\|^2\nu_2^Y(y)dy\\
\le& 2\tilde{Q}\int \|\nabla_y f(x,y)\|^2\nu_1^X(x)\nu_2^Y(y)dy \text{ by Jensen's inequality.}
\end{aligned}
$$

Likewise, we change ν_1^X to u_{B_1}. By Lemma 5, we can control the difference by

$$
\begin{aligned}
&\Big(\mathbb{E}_{\nu_1^X\otimes u_{\bar{\mathcal{B}}_2}}[f(X,Y)] - \mathbb{E}_{u_{B_1}\otimes u_{\bar{\mathcal{B}}_2}}[f(X,Y)]\Big)^2\\
=&\left(\int\left(\int f(x,y)\nu_1^X(x)dx - \int f(x,y)u_{B_1}(x)dx\right)u_{\bar{\mathcal{B}}_2}(y)dy\right)^2\\
\le&\int\left(\int f(x,y)\nu_1^X(x)dx - \int f(x,y)u_{B_1}(x)dx\right)^2 u_{\bar{\mathcal{B}}_2}(y)dy \text{ by Jensen's inequality}\\
\le& 2\tilde{q}\int \|\nabla_x f(x,y)\|^2\nu_1^X(x)u_{\bar{\mathcal{B}}_2}(y)dy \le 2\tilde{q}A\int \|\nabla_x f(x,y)\|^2\nu_1^X(x)\nu_2^Y(y)dy.
\end{aligned}
$$

Step 2. Replace $u_{\bar{\mathcal{B}}_2}$ with u_{B_2}. By Lemma 6, we can control the difference by

$$
\begin{aligned}
&\Big(\mathbb{E}_{u_{B_1}\otimes u_{\bar{\mathcal{B}}_2}}[f(X,Y)] - \mathbb{E}_{u_{B_1}\otimes u_{B_2}}[f(X,Y)]\Big)^2\\
\le&\int\left(\int f(x,y)u_{\bar{\mathcal{B}}_2}(y)dy - \int f(x,y)u_{B_2}(y)dy\right)^2 u_{B_1}(x)dx\\
\le&\frac{R^{d+1}}{r^{d-1}}\int \|\nabla_y f(x,y)\|^2 u_{\bar{\mathcal{B}}_2}(y)u_{B_1}(x)dydx\\
\le&\frac{R^{d+1}}{r^{d-1}}aA\int \|\nabla_y f(x,y)\|^2\nu_1^x(x)\nu_2^y(y)dxdy,
\end{aligned}
$$

when $d \ge 2$. If $d = 1$, an additional $\log(R/r)$ is needed.

Step 3. The mean difference square in exchanging B_1 and B_2 can be bounded by the additional term in the carre du champ for replica exchange:

$$
\begin{aligned}
&\left(\mathbb{E}_{u_{B_1}\otimes u_{B_2}}[f(X,Y)] - \mathbb{E}_{u_{B_2}\otimes u_{B_1}}[f(X,Y)]\right)^2 \\
\leq &\int (f(x,y)-f(y,x))^2 u_{B_1}(x)u_{B_2}(y)dxdy \quad \text{by Jensen's inequality} \\
\leq &\left(\frac{R}{r}\right)^d \int (f(x,y)-f(y,x))^2 (u_{B_1}(x)u_{\bar{\mathcal{B}}_2}(y) \wedge u_{\bar{\mathcal{B}}_1}(x)u_{B_2}(y))dxdy \text{ as } \tfrac{u_{B_i}(x)}{u_{\bar{\mathcal{B}}_i}(x)} = \tfrac{R_i^d}{r_i^d} \leq \left(\tfrac{R}{r}\right)^d \\
\leq &\left(\frac{R}{r}\right)^d aA \int (f(x,y)-f(y,x))^2 \left(\nu_1^X(x)\nu_2^Y(y) \wedge \nu_2^X(y)\nu_1^Y(x)\right) dxdy.
\end{aligned}
$$

Putting the three steps together, we have

$$
\begin{aligned}
&\left(\mathbb{E}_{\nu_1^X\otimes\nu_2^Y}[f(X,Y)] - \mathbb{E}_{\nu_2^X\otimes\nu_1^Y}[f(X,Y)]\right)^2 \\
\leq &7\left(\mathbb{E}_{\nu_1^X\otimes\nu_2^Y}[f(X,Y)] - \mathbb{E}_{\nu_1^X\otimes u_{\bar{\mathcal{B}}_2}}[f(X,Y)]\right)^2 + 7\left(\mathbb{E}_{\nu_1^X\otimes u_{\bar{\mathcal{B}}_2}}[f(X,Y)] - \mathbb{E}_{u_{B_1}\otimes u_{\bar{\mathcal{B}}_2}}[f(X,Y)]\right)^2 \\
&+ 7\left(\mathbb{E}_{u_{B_1}\otimes u_{\bar{\mathcal{B}}_2}}[f(X,Y)] - \mathbb{E}_{u_{B_1}\otimes u_{B_2}}[f(X,Y)]\right)^2 \\
&+ 7\left(\mathbb{E}_{\nu_2^X\otimes\nu_1^Y}[f(X,Y)] - \mathbb{E}_{\nu_2^X\otimes u_{\bar{\mathcal{B}}_1}}[f(X,Y)]\right)^2 + 7\left(\mathbb{E}_{\nu_2^X\otimes u_{\bar{\mathcal{B}}_1}}[f(X,Y)] - \mathbb{E}_{u_{B_2}\otimes u_{\bar{\mathcal{B}}_1}}[f(X,Y)]\right)^2 \\
&+ 7\left(\mathbb{E}_{u_{B_2}\otimes u_{\bar{\mathcal{B}}_1}}[f(X,Y)] - \mathbb{E}_{u_{B_2}\otimes u_{B_1}}[f(X,Y)]\right)^2 \\
&+ 7\left(\mathbb{E}_{u_{B_1}\otimes u_{B_2}}[f(X,Y)] - \mathbb{E}_{u_{B_2}\otimes u_{B_1}}[f(X,Y)]\right)^2 \\
\leq &\Xi_x \int \|\nabla_x f(x,y)\|^2 (\nu_1^X(x)\nu_2^Y(y) + \nu_2^X(x)\nu_1^Y(y))dxdy \\
&+ \Xi_y \int \|\nabla_y f(x,y)\|^2 (\nu_1^X(x)\nu_2^Y(y) + \nu_2^X(x)\nu_1^Y(y))dxdy \\
&+ \Xi_e \int (f(x,y)-f(y,x))^2 \left(\nu_1^X(x)\nu_2^Y(y) \wedge \nu_2^X(y)\nu_1^Y(x)\right) dxdy,
\end{aligned}
$$

where $\Xi_x = 14\tilde{q}A, \Xi_y = 14\tilde{Q} + 7aA\left(\frac{R^{d+1}}{r^{d-1}}\right)\left(\log\left(\frac{R}{r}\right)\right)^{1_{d=1}}, \Xi_e = 7\left(\frac{R}{r}\right)^d aA.$ □

Appendix E. Proof of Theorem 4

In order of handle the mean difference square, which appears as Part (C) in (14), we define $\mathcal{X}_k(\mathbf{x}_{0:K}) = (f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k) - f(\mathbf{w}_k, y_k, x_k, \mathbf{z}_k))^2 s_k(x_k, y_k)$, and $\Gamma_k(\mathbf{x}_{0:K}) = \sum_{l=k}^{K}\left(\tau_l\|\nabla_{x_l} f(\mathbf{x}_{0:K})\|^2 + \rho\mathcal{X}_l(\mathbf{x}_{0:K})\right)$. Denote

$$
\widetilde{\mathbb{E}}_{\nu,k}(f) := \int \tau_l \|\nabla_{x_l} f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k)\|^2 \nu(x_k)\pi_{k+1}(y_k)\pi_k^{\mathbf{Z}}(\mathbf{z}_k)dx_k dy_k d\mathbf{z}_k.
$$

When $\nu = \pi_k$, we simply write $\widetilde{\mathbb{E}}_{\nu,k}$ as $\widetilde{\mathbb{E}}_k$. We also define, for $k = 0, 1, \dots, K-1$, $\Xi_{e_k} = 7\left(\frac{r_{k+1}}{r_k}\right)^d a_k a_{k+1}$, $\Xi_{x_k} = 28 q_k a_{k+1}$,

$$\Xi_{y_k} = 28q_k + 7\frac{(r_{k+1})^{d+1}}{(r_k)^{d-1}} a_k a_{k+1} \left(\log\left(\frac{r_{k+1}}{r_k}\right)\right)^{1_{d=1}}.$$

We first bound the mean difference square.

Proposition 10 *Under Assumptions 4 and 5, for $k \le K-1$,*

$$\sum_{i,j} p_i p_j \left(\theta_{k,i}(\mathbf{w}_k) - \theta_{k,j}(\mathbf{w}_k)\right)^2 \le \Xi_k \mathbb{E}_{k:K}\left[\Gamma_k(\mathbf{w}_k, \mathbf{X}_{k:K})\right],$$

where for any fixed $\alpha, \gamma > 1$ with $\frac{1}{\alpha} + \frac{1}{\gamma} = 1$,

$$\Xi_k = \max\left\{\max_{k+1 \le l \le K-1} (4\alpha)^{l-k-1}\left(\frac{8\alpha\gamma\Xi_{x_l}}{\tau_l} + \frac{2\gamma\Xi_{y_{l-1}}}{\tau_l}\right), \frac{2\gamma\Xi_{x_k}}{\tau_k}, \max_{k \le l \le K-1} (4\alpha)^{l-k}\frac{\gamma\Xi_{e_l}}{\rho}\right\}.$$

Proof We prove the proposition by induction.

We want to show that for any fixed $\mathbf{w}_k$,

$$\begin{aligned}\sum_{i,j} p_i p_j \Big(\theta_{k,i}(\mathbf{w}_k) - \theta_{k,j}(\mathbf{w}_k)\Big)^2 &\le \sum_{l=k+1}^{K-1} (4\alpha)^{l-k-1}\left(\frac{8\alpha\gamma\Xi_{x_l}}{\tau_l} + \frac{2\gamma\Xi_{y_{l-1}}}{\tau_l}\right) \cdot \tau_l \widetilde{\mathbb{E}}_k \|\nabla_{x_l} f\|^2 \\ &+ \frac{2\gamma\Xi_{x_k}}{\tau_k}\tau_k \widetilde{\mathbb{E}}_k\|\nabla_{x_k} f\|^2 + \sum_{l=k}^{K-1}(4\alpha)^{l-k}\frac{\gamma\Xi_{e_l}}{\rho}\rho\widetilde{\mathbb{E}}_k \mathcal{X}_l(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k).\end{aligned} \tag{E.1}$$

For $k = K-1$, (E.1) can be obtained from Proposition 8. **Suppose** (E.1) **holds for $k+1$. Now, for k,** We first note that $\theta_{k,i} = \sum_{h=1}^{I} p_{k+1,h}\zeta_{kih}$ and

$$\begin{aligned}&\sum_{i,j} p_i p_j (\theta_{k,i} - \theta_{k,j})^2 \\ \le &2\alpha \sum_{i,j} p_i p_j \underbrace{(\theta_{k,i} - \zeta_{kij})^2}_{\text{(a)}} + \gamma \sum_{i,j} p_i p_j \underbrace{(\zeta_{kij} - \zeta_{kji})^2}_{\text{(b)}} + 2\alpha \sum_{i,j} p_i p_j \underbrace{(\zeta_{kji} - \theta_{k,j})^2}_{\text{(c)}}.\end{aligned} \tag{E.2}$$

Note that part (a) and (c) are symmetric. For part (a), we have

$$\sum_{i,j} p_i p_j (\zeta_{kij} - \theta_{k,i})^2 = \sum_{i,j} p_i p_j \left(\sum_{h=1}^{I} p_h (\zeta_{kij} - \zeta_{kih})\right)^2 \le \sum_{i,j,h} p_i p_j p_h \left(\zeta_{kij} - \zeta_{kih}\right)^2.$$

By induction,

$$\begin{aligned}
&\sum_i p_i \sum_{j,h} p_j p_h \left(\zeta_{kij}(\mathbf{w}_k) - \zeta_{kih}(\mathbf{w}_k)\right)^2 \\
\leq &\sum_i p_i \sum_{j,h} p_j p_h \int (\theta_{k+1,j}(\mathbf{w}_k, x_k) - \theta_{k+1,h}(\mathbf{w}_k, x_k))^2 \nu_{k,i}(x_k) dx_k \\
\leq &\sum_{l=k+2}^{K-1} (4\alpha)^{l-k-2} \left(\frac{8\alpha\gamma\Xi_{x_l}}{\tau_l} + \frac{2\gamma\Xi_{y_{l-1}}}{\tau_l} \right) \cdot \tau_l \widetilde{\mathbb{E}}_k \|\nabla_{x_l} f\|^2 \\
&+ \frac{2\gamma\Xi_{x_{k+1}}}{\tau_{k+1}} \widetilde{\mathbb{E}}_k \|\nabla_{y_k} f\|^2 + \sum_{l=k+1}^{K-1} (4\alpha)^{l-k-1} \frac{\gamma\Xi_{e_l}}{\rho} \rho \widetilde{\mathbb{E}}_k \mathcal{X}_l(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k).
\end{aligned}$$

For part (b), recall that $\zeta_{kij} = \int f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k)\nu_{k,i}(x_k)\nu_{k+1,j}(y_k)\pi_k^{\mathbf{Z}}(\mathbf{z}_k) dx_k dy_k d\mathbf{z}_k$ and $\zeta_{kji} = \int f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k)\nu_{k,j}(x_k)\nu_{k+1,i}(y_k)\pi_k^{\mathbf{Z}}(\mathbf{z}_k) dx_k dy_k d\mathbf{z}_k$. By the mean difference square estimate in Proposition 8, we have

$$\begin{aligned}
&(\zeta_{kij}(\mathbf{w}_k) - \zeta_{kji}(\mathbf{w}_k))^2 \\
\leq &\frac{\Xi_{x_k}}{\tau_k} \tau_k \int \|\nabla_{x_k} f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k)\|^2 (\nu_{k,i}\nu_{k+1,j} + \nu_{k,j}\nu_{k+1,i})(x_k, y_k) \pi_k^{\mathbf{Z}}(\mathbf{z}_k) dx_k dy_k d\mathbf{z}_k \\
&+ \frac{\Xi_{y_k}}{\tau_{k+1}} \tau_{k+1} \int \|\nabla_{y_k} f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k)\|^2 (\nu_{k,i}\nu_{k+1,j} + \nu_{k,j}\nu_{k+1,i})(x_k, y_k) \pi_k^{\mathbf{Z}}(\mathbf{z}_k) dx_k dy_k d\mathbf{z}_k \\
&+ \frac{\Xi_{e_k}}{\rho} \rho \int (f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k) - f(\mathbf{w}_k, y_k, x_k, \mathbf{z}_k))^2 \times \\
&\qquad (\nu_{k,i}(x_k)\nu_{k+1,j}(y_k) \wedge \nu_{k,j}(y_k)\nu_{k+1,i}(x_k)) \, \pi_k^{\mathbf{Z}}(\mathbf{z}_k) dx_k dy_k d\mathbf{z}_k.
\end{aligned}$$

Note that $\sum_{i,j} p_i p_j \nu_{k,i}(x_k)\nu_{k+1,j}(y_k) = \pi_k(x_k)\pi_{k+1}(y_k)$. Because $a \wedge c + b \wedge d \leq (a+b) \wedge (c+d)$,

$$\sum_{i,j} (p_i p_j \nu_{k,i}(x_k)\nu_{k+1,j}(y_k)) \wedge (p_i p_j \nu_{k,j}(y_k)\nu_{k+1,i}(x_k)) \leq (\pi_k(x_k)\pi_{k+1}(y_k)) \wedge (\pi_k(y_k)\pi_{k+1}(x_k)).$$

Thus,

$$\begin{aligned}
\sum_{i,j} p_i p_j (\zeta_{kij}(\mathbf{w}_k) - \zeta_{kji}(\mathbf{w}_k))^2 \leq &2\Xi_{x_k} \widetilde{\mathbb{E}}_k \|\nabla_{x_k} f\|^2 + 2\Xi_{y_k} \widetilde{\mathbb{E}}_k \|\nabla_{y_k} f\|^2 \\
&+ \frac{\Xi_{e_k}}{\rho} \rho \int (f(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k) - f(\mathbf{w}_k, y_k, x_k, \mathbf{z}_k))^2 \times \\
&\qquad (\pi_k(x_k)\pi_{k+1}(y_k) \wedge \pi_k(y_k)\pi_{k+1}(x_k)) \, \pi_k^{\mathbf{Z}}(\mathbf{z}_k) dx_k dy_k d\mathbf{z}_k.
\end{aligned}$$

Combining parts (a) – (c), we have

$$
\begin{aligned}
&\sum_{i,j} p_i p_j \left(\theta_{i,k}(\mathbf{w}_k) - \theta_{j,k}(\mathbf{w}_k)\right)^2 \\
\leq &\sum_{l=k+1}^{K-1} (4\alpha)^{l-k-1} \left(\frac{8\alpha\gamma\Xi_{x_l}}{\tau_l} + \frac{2\gamma\Xi_{y_{l-1}}}{\tau_l} \right) \tau_l \widetilde{\mathbb{E}}_k \|\nabla_{x_l} f\|^2 \\
&+ \frac{2\gamma\Xi_{x_k}}{\tau_k} \tau_k \widetilde{\mathbb{E}}_k \|\nabla_{x_k} f\|^2 + \sum_{l=k}^{K-1} (4\alpha)^{l-k} \frac{\gamma\Xi_{e_l}}{\rho} \rho \widetilde{\mathbb{E}}_k \mathcal{X}_l(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k).
\end{aligned}
$$

Setting

$$
\Xi_k = \max\left\{ \max_{k+1 \leq l \leq K-1} (4\alpha)^{l-k-1} \left(\frac{8\alpha\gamma\Xi_{x_l}}{\tau_l} + \frac{2\gamma\Xi_{y_{l-1}}}{\tau_l} \right),\ \frac{2\gamma\Xi_{x_k}}{\tau_k},\ \max_{k \leq l \leq K-1} (4\alpha)^{l-k} \frac{\gamma\Xi_{e_l}}{\rho} \right\},
$$

we have the result. □

Proof [Proof of Theorem 4] Recall the upper bound (14).

For part (A) By Assumption 4 and Proposition 3, we have

$$
\begin{aligned}
&\sum_{i=1}^{I} p_i \int \left(g_k(\mathbf{w}_k, x_k, y_k) - \eta_{k,i}(\mathbf{w}_k, y_k)\right)^2 \nu_{k,i}(x_k) dx_k \pi_{k+1}(y_k) dy_k \\
\leq &\sum_{i=1}^{I} p_i q_k \int \|\nabla_{x_k} g_k(\mathbf{w}_k, x_k, y_k)\|^2 \nu_{k,i}(x_k) \pi_{k+1}(y_k) dx_k dy_k \\
\leq &q_k \int \|\nabla_{x_k} g_k(\mathbf{w}_k, x_k, y_k)\|^2 \pi_k(x_k) \pi_{k+1}(y_k) dx_k dy_k.
\end{aligned}
$$

For part (B) By Assumption 4 and Proposition 3, we have

$$
\begin{aligned}
&\sum_{i=1}^{I} p_i \int (\eta_{k,i}(\mathbf{w}_k, y_k) - \theta_{k,i}(\mathbf{w}_k))^2 \pi_{k+1}(y_k) dy_k \\
\leq &\sum_{i=1}^{I} p_i q_{k+1} \int \|\nabla_{y_k} g_k(\mathbf{w}_k, x_k, y_k)\|^2 \nu_{k,i}(x_k) \pi_{k+1}(y_k) dx_k dy_k \\
\leq &q_{k+1} \int \|\nabla_{y_k} g_k(\mathbf{w}_k, x_k, y_k)\|^2 \pi_k(x_k) \pi_{k+1}(y_k) dx_k dy_k.
\end{aligned}
$$

For part (C) From Proposition 10,

$$\begin{aligned}
&\sum_{i,j} p_i p_j \left(\theta_{i,k}(\mathbf{w}_k) - \theta_{j,k}(\mathbf{w}_k)\right)^2 \\
&\le \sum_{l=k+1}^{K-1} (4\alpha)^{l-k-1} \left(8\alpha\gamma\Xi_{x_l} + 2\gamma\Xi_{y_{l-1}}\right) \widetilde{\mathbb{E}}_k \|\nabla_{x_l} f\|^2 + 2\gamma\Xi_{x_k} \widetilde{\mathbb{E}}_k \|\nabla_{x_k} f\|^2 \\
&\quad + \sum_{l=k}^{K-1} (4\alpha)^{l-k} \gamma\Xi_{e_l} \widetilde{\mathbb{E}}_k \mathcal{X}_l(\mathbf{w}_k, x_k, y_k, \mathbf{z}_k).
\end{aligned}$$

Putting parts (A) – (C) together, we have

$$\begin{aligned}
&\mathbb{E}_{0:K}\left[\mathbb{E}_k\left[\left(\mathbb{E}_{(k+1):K} f(\mathbf{X}_{0:K}) - \mathbb{E}_{k:K} f(\mathbf{X}_{0:K})\right)^2\right]\right] \\
&\le \sum_{l=k+2}^{K-1} \frac{3(4\alpha)^{l-k-1}}{\tau_l} \left(8\alpha\gamma\Xi_{x_l} + 2\gamma\Xi_{y_{l-1}}\right) \int \tau_l \|\nabla_{x_l} f(\mathbf{x}_{0:K})\|^2 \pi_{0:K}(\mathbf{x}_{0:K}) d\mathbf{x}_{0:K} \\
&\quad + \frac{3}{\tau_{k+1}} \left(8\alpha\gamma\Xi_{x_{k+1}} + 2\gamma\Xi_{y_k} + q_{k+1}\right) \int \|\nabla_{x_{k+1}} f(\mathbf{x}_{0:K})\|^2 \pi_{0:K}(\mathbf{x}_{0:K}) d\mathbf{x}_{0:K} \\
&\quad + \frac{3}{\tau_k} \left(2\gamma\Xi_{x_k} + q_k\right) \int \tau_k \|\nabla_{x_k} f(\mathbf{x}_{0:K})\|^2 \pi_{0:K}(\mathbf{x}_{0:K}) d\mathbf{x}_{0:K} \\
&\quad + \sum_{l=k}^{K-1} 3(4\alpha)^{l-k} \frac{\gamma\Xi_{e_l}}{\rho} \rho \int \mathcal{X}_l(\mathbf{x}_{0:K}) \pi_{0:K}(\mathbf{x}_{0:K}) d\mathbf{x}_{0:K}.
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{E}_{0:K}\left[\left(f(\mathbf{X}_{0:K}) - \mathbb{E} f(\mathbf{X}_{0:K})\right)^2\right] \\
\le &\sum_{k=0}^{K} \mathbb{E}_{0:K}\left[\mathbb{E}_k\left[\left(\mathbb{E}_{(k+1):K} f(\mathbf{X}_{0:K}) - \mathbb{E}_{k:K} f(\mathbf{X}_{0:K})\right)^2\right]\right] \le \kappa \mathbb{E}_{0:K}[\Gamma_R(f(\mathbf{X}_{0:K}))],
\end{aligned}$$

where

$$\begin{aligned}
\kappa = \max_{0\le k\le K-1} \max \Bigg\{ &\sum_{h=2}^{k-2} \frac{3(4\alpha)^{k-h+1}}{\tau_k} \left(8\alpha\gamma\Xi_{x_k} + 2\gamma\Xi_{y_{k-1}}\right) \\
&+ \frac{3}{\tau_k} \left((8\alpha\gamma + 2\gamma)\Xi_{x_k} + 2\gamma\Xi_{y_{k-1}} + 2q_k\right), \sum_{h=0}^{k} \frac{3(4\alpha)^{k-h+2}}{\rho} \gamma\Xi_{e_k} \Bigg\}.
\end{aligned}$$

□