Utilising the CLT Structure in Stochastic Gradient based Sampling : Improved Analysis and Faster Algorithms

Aniket Das KETD@GOOGLE.COM
Google Research,
Bangalore, India

Dheeraj Nagaraj DHEERAJNAGARAJ@GOOGLE.COM
Google Research,
Bangalore, India

Anant Raj ANANT.RAJ@INRIA.FR
University of Illinois Urbana-Champaign
& Inria, Ecole Normale Supérieure

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider stochastic approximations of sampling algorithms, such as Stochastic Gradient Langevin Dynamics (SGLD) and the Random Batch Method (RBM) for Interacting Particle Dynamcs (IPD). We observe that the noise introduced by the stochastic approximation is nearly Gaussian due to the Central Limit Theorem (CLT) while the driving Brownian motion is exactly Gaussian. We harness this structure to absorb the stochastic approximation error inside the diffusion process, and obtain improved convergence guarantees for these algorithms. For SGLD, we prove the first stable convergence rate in KL divergence without requiring uniform warm start, assuming the target density satisfies a Log-Sobolev Inequality. Our result implies superior first-order oracle complexity compared to prior works, under significantly milder assumptions. We also prove the first guarantees for SGLD under even weaker conditions such as Hölder smoothness and Poincare Inequality, thus bridging the gap between the state-of-the-art guarantees for LMC and SGLD. Our analysis motivates a new algorithm called covariance correction, which corrects for the additional noise introduced by the stochastic approximation by rescaling the strength of the diffusion. Finally, we apply our techniques to analyze RBM, and significantly improve upon the guarantees in prior works (such as removing exponential dependence on horizon), under minimal assumptions.

Keywords: Langevin Monte Carlo, SGLD, Sampling, CLT, Random Batch Method.

1. Introduction

The task of simulating stochastic systems or sampling from a target distribution in continuous domains via discretizations of Stochastic Differential Equations (SDEs) is a fundamental problem in machine learning, theoretical computer science, statistical physics and scientific computing (Parisi, 1981; Robert et al., 1999; Frenkel and Smit, 2001; Shreve, 2005; Lee and Vempala, 2022). This problem finds applications in several domains such as Bayesian inference (Welling and Teh, 2011), generative modelling (Ho et al., 2020), differential privacy (Gopi et al., 2022) and algorithmic geometry (Kannan et al., 1997). Popular algorithms for this problem include Langevin Monte Carlo (LMC) (Parisi, 1981), Hamiltonian Monte Carlo (HMC) (Neal et al., 2011), Stein Variational Gradient Descent (SVGD) (Liu and Wang, 2016) and Interacting Particle Dynamics (IPD) (Carrillo et al., 2021). In this work, we study the stochastic approximations of Langevin Monte Carlo (LMC) and Interacting Particle Dynamics (IPD). LMC aims to sample from a target distribution over $\mathbb{R}^d$, whose density $\pi^*(\mathbf{x}) \propto \exp(-F(\mathbf{x}))$ is known only upto a normalizing constant. This is achieved via an Euler discretization of the Langevin SDE with a step-size $\eta > 0$ as follows:

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta \nabla F(\mathbf{x}_k) + \sqrt{2\eta}\epsilon_k, \ \epsilon_k \sim \mathcal{N}(0, \mathbf{I}) \tag{LMC}$$

While LMC is a popular algorithm in various statistical applications, there exist several practical settings where computing ∇F might be infeasible, or even intractable, whereas an unbiased estimate of ∇F is easily computable. Perhaps the most popular example is that of problems with a finite-sum structure, where $F(\mathbf{x}) = {}^1\!/_n \sum_{i=1}^n f_i(\mathbf{x})$. Here, a stochastic gradient (or a random batch gradient) can be computed as $\frac{1}{B}\sum_{i=1}^B \nabla f_{I_i}(\mathbf{x})$ for $I_1, \dots, I_B \overset{\text{iid}}{\sim} \mathsf{Uniform}([n])$ and used in place of $\nabla F(\mathbf{x})$ in the LMC update. The resulting algorithm is a stochastic approximation of LMC called Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011), and is widely used in practice in large-scale problems. However, theoretical analysis of its convergence properties is relatively unexplored. Most prior works on analyzing the convergence of SGLD require restrictive assumptions like dissipativity and warm start (Raginsky et al., 2017; Zou et al., 2021), with some being specific to finite-sum problems with smooth components (Kinoshita and Suzuki, 2022). On the contrary, the convergence of LMC is well characterized (Vempala and Wibisono, 2019; Erdogdu and Hosseinzadeh, 2021a; Chewi et al., 2022a) under a variety of isoperimetric conditions (e.g. Poincare Inequality, Log Sobolev Inequality) which are significantly weaker than the assumptions used to analyze SGLD. This situation is in stark contrast to that of optimization where Gradient Descent (GD) and Stochastic Gradient Descent (SGD) are analyzed under (nearly) identical assumptions.

IPD is an algorithm that simulates the aggregation-diffusion dynamics of n particles with pairwise interactions, and is actively used for simulating physical systems, sampling, and optimization (see Appendix A). When implemented naively, IPD incurs a per-step complexity of $O(n^2)$ which can be prohibitive for several problems. To ameliorate this, the Random Batch Method (RBM) performs a stochastic approximation of the inter-particle interactions by only considering the interaction of each particle with a random subset of B particles, thereby reducing the per-step complexity to $O(nB)$.

$$\mathbf{x}^i_{k+1} = \mathbf{x}^i_k + \eta \mathbf{g}^i_k(\mathbf{x}^i_k) + \frac{\eta}{n}\sum_{j=1}^n \mathbf{K}^{ij}_k(\mathbf{x}^i_k, \mathbf{x}^j_k) + \sqrt{\eta}\sigma\epsilon^i_k, \quad \forall\, i \in [n], \epsilon^i_k \sim \mathcal{N}(0, \mathbf{I}) \tag{IPD}$$

For both SGLD and RBM, the stochastic approximation takes the form of a conditionally i.i.d. average, and hence the error produced by the stochastic approximation is nearly Gaussian due to the Central Limit Theorem (CLT). This is in addition to, and independent of the Brownian motion driving the system, which is typically of higher magnitude. The key focus of our work is to utilize the i.i.d average structure (or CLT structure) of the stochastic approximation and apply non-asymptotic CLTs (or CLT-like arguments) to understand the interaction between the two independent sources of stochasticity. To this end, we show that the stochastic approximation noise can be effectively *absorbed inside* the Brownian motion, and use this insight to derive state-of-the-art guarantees for these algorithms. Motivated by our analysis, we also design a novel covariance-correction strategy which compensates for the stochastic approximation error by adaptively rescaling the diffusion term, and show that this leads to faster convergence without increase in computational complexity.

Result	Algorithm	Assumptions	Metric	Complexity
Raginsky et al. (2017)	SGLD	Component Smooth, Dissipative	$\mathcal{W}_2$	$\frac{\mathrm{poly}(d)}{\epsilon^4}$(Unstable)
Zou et al. (2021)	SGLD	Dissipative, Warm Start Component Smooth	TV	$\frac{d^4}{\epsilon^2}$(Unstable)
Kinoshita and Suzuki (2022)	VR-SGLD	Finite-Sum, LSI Component Smooth	$\sqrt{\mathsf{KL}}$	$\frac{d\sqrt{n}}{\epsilon^2}$ (Stable)
Theorem 5	SGLD	Smooth, LSI, 4th moment	$\sqrt{\mathsf{KL}}$	$\frac{d^{1.5}}{\epsilon^2}$ (Stable)
Theorem 6	AB-SGLD	Finite-Sum, Smooth, LSI	$\sqrt{\mathsf{KL}}$	$\frac{d^{1.5}}{\epsilon^2}$ (Stable)
Corollary 32	CC-SGLD	Smooth, LSI, 6th moment	TV	$\frac{d^{4/3}}{\epsilon^2}$(Unstable)
Theorem 7	Averaged SGLD	Smooth, **PI**, 4th moment	TV	$\frac{d^{2.5}}{\epsilon^4}$ (Stable)
Corollary 28	SGLD	Smooth, **PI**, 6th moment	TV	$\frac{d^{3.5}}{\epsilon^2}$ (Unstable)
Corollary 32	CC-SGLD	Smooth, **PI**, 6th moment	TV	$\frac{d^{10/3}}{\epsilon^2}$ (Unstable)

Table 1: Comparison with prior works for SGLD. d, ϵ and n respectively mean the dimension, error and number of components (for finite-sum problems). Note that 1. dissipativity implies LSI and p^{th} moment bounds 2. $\sqrt{\mathsf{KL}}$ is a stronger 'metric' than TV and under LSI it is stronger than $\mathcal{W}_2$

1.1. Contributions

Our work develops non-asymptotic CLTs (our CLT-like arguments) to quantify the approximate Gaussianity of the noise introduced by the random batch-based stochastic approximations used in SGLD and RBM, leading to state-of-the-art convergence guarantees and algorithmic improvements. Our key contributions are summarized as follows.

SOTA Guarantee for Smooth SGLD under LSI When F is smooth and π^* satisfies a Log-Sobolev Inequality (LSI), we prove that the oracle complexity (i.e. number of stochastic gradient evaluations) of SGLD to attain last-iterate ϵ-convergence in KL divergence is $\tilde{O}(d^{1.5}/\epsilon)$, assuming every iterate has $O(d^2)$ 4th moment. When F has a finite-sum structure, we remove the 4th moment assumption by using adaptive batch-sizes. Compared to prior works, our result obtains a significantly improved oracle complexity in a stronger metric, without imposing restrictive assumptions like dissipativity, component smoothness and warm start (see Table 1). Our last-iterate error bounds are stable, i.e. they do not diverge as the number of iterations $k \to \infty$. Our results for SGLD also improve upon the best-known rates for more sophisticated algorithms such as variance-reduced SGLD (VR-SGLD).

Stationarity Guarantee for Smooth SGLD In the above setting, without any isoperimetric assumptions on π^*, we obtain a stable $O(d^{2.5}/\epsilon^2)$ oracle complexity for average-iterate ϵ-convergence of SGLD in Fisher Divergence (FD). In the sampling literature, convergence in FD is considered analogous to first-order stationarity in nonconvex optimization (Balasubramanian et al., 2022; Chewi et al., 2022b), making our result the first known stationarity guarantee for smooth SGLD. As a corollary, we derive a stable oracle complexity of $O(d^{2.5}/\epsilon^4)$ for ϵ-convergence in Total Variation (TV) whenever π^* satisfies a Poincare Inequality (PI), providing the first known rate for SGLD under PI. Prior to this work, such guarantees were only known for LMC (Balasubramanian et al., 2022). Thus, our results significantly bridge the gap between SGLD and LMC.

Result	Algorithm	Assumptions	Metric	Bound
Jin et al. (2020)	RBM	$-\mathbf{g}_k^i(\mathbf{x}) = \nabla V(\mathbf{x})$, $\mathbf{K}_k^{ij}(\mathbf{x},\mathbf{y}) = H(\mathbf{x}-\mathbf{y})$ $\|\nabla H\| \leq L$, $\nabla^2 V \succeq r\mathbf{I}$, $r > 2L$	averaged $\mathcal{W}_2$	$\sqrt{\frac{\eta}{B} + \eta^2}$
Jin et al. (2021)	RBM	poly growth of $\mathbf{g}_k^i, \nabla \mathbf{g}_k^i$ $\mathbf{K}_t^{ij}, \nabla \mathbf{K}_k^{ij}, \nabla^2 \mathbf{K}_k^{ij}$ are bounded	averaged $\mathcal{W}_2$	$e^{CK\eta}\sqrt{\frac{\eta}{B} + \eta^2}$
Theorem 9	RBM	Finite $\mathbf{g}_k^{ij}$, bounded $\mathbf{K}_k^{ij}$	$\sqrt{\mathsf{KL}}$	$\frac{\sqrt{n\eta^2 K}}{B}$
Theorem 11	CC-RBM	Finite $\mathbf{g}_k^{ij}$, bounded $\mathbf{K}_k^{ij}$	$\sqrt{\mathsf{KL}}$	$\frac{\sqrt{n\eta^2 K}}{B^{3/2}}$

Table 2: Comparison with prior works for RBM (prior works consider convergence to continuous time version of IPD). The function H is bounded uniformly. The exponential bound can be inferred by following the proof of (Jin et al., 2021, Theorem 3.1). Averaged $\mathcal{W}_2$ refers to average of $\mathcal{W}_2$ distance of laws of individual particles and KL is the KL divergence between the laws of entire trajectories of all particles.

Statistical Indistinguishability of SGLD and LMC Without imposing any smoothness or isoperimetric assumptions, we obtain trajectory-level KL divergence bounds between SGLD and LMC, which show that the two algorithms are nearly statistically indistinguishable. This result leads to a highly general technique for obtaining last-iterate TV guarantees for SGLD under a variety of settings. As a corollary, we obtain a $\tilde{\Theta}({d^{3.5}}/{\epsilon^2})$ last-iterate oracle complexity for SGLD when F is smooth and π^* satisfies PI.

Covariance Correction We design a novel algorithmic improvement for SGLD and RBM called Covariance Correction, which compensates for the excess noise due to stochastic approximation by appropriately rescaling the diffusion term. We show that this leads to faster convergence without increasing computational complexity. For SGLD with Covariance Correction (CC-SGLD), we obtain last-iterate TV guarantees of $\tilde{\Theta}({d^{4/3}}/{\epsilon^2})$ under smoothness and LSI, and $\tilde{\Theta}({d^{10/3}}/{\epsilon^2})$ under smoothness and PI.

Analysis of RBM Assuming only boundedness of the inter-particle interactions $\mathbf{K}_k^{ij}$, we derive an $\tilde{O}({\eta^2 nK}/{B^2})$ trajectory level upper bound between K iterations of RBM (run with n particles, batch-size B and step-size η) and IPD. We improve this to $\tilde{O}({\eta^2 nK}/{B^3})$ for Covariance Corrected RBM (CC-RBM). Our results (see Table 2) significantly improves upon prior works that either make stringent assumptions on $\mathbf{g}_k, \mathbf{K}_k^{ij}$ and its derivatives (Jin et al., 2020), or suffer an exponential dependence on ηK (Jin et al., 2021).

1.2. Related Work

Sampling algorithms such LMC and IPD have been extensively studied by prior works. Below we give a concise review of the relevant literature and refer to Appendix A for a detailed discussion

LMC and SGLD In typical Bayesian inference settings, the function F in LMC is an empirical average of n sample functions, which can be expensive to evaluate as n is generally very large. For such large-scale problems, SGLD has emerged as the sampling algorithm of choice (Welling and Teh, 2011). While some prior works have investigated the convergence properties of SGLD (Raginsky et al., 2017; Zou et al., 2021), they generally impose stringent assumptions like dissipativity and warm start. On the contrary, the convergence of LMC is well understood under much weaker isoperimetric assumptions (Vempala and Wibisono, 2019; Chewi et al., 2022a).

IPD and RBM IPD can be computationally prohibitive whenever the number of particles n is large. To ameliorate this, the Random Batch Method, which considers the interaction of each particle only with a random subset of other particles, was first proposed in Jin et al. (2020), which also provided convergence guarantees under stringent regularity conditions on the confining and interaction forces (see Table 1.1). These conditions were relaxed in Jin et al. (2021) at the cost of an exponential dependence on the horizon ηK. Subsequent works (Ko et al., 2021; Daus et al., 2021) also analyze RBM in specialized settings.

1.3. Notation and Organization

By $\mathsf{KL}\left(\nu\middle|\middle|\mu\right)$, we denote the KL-divergence between probability measures ν and μ. $\mathsf{FD}\left(\nu\middle|\middle|\mu\right) = \mathbb{E}_\nu\left[\|\nabla\ln(\nu/\mu)\|^2\right]$ denotes the Fisher Divergence between ν and μ. Whenever μ and ν are probability measures over $\mathbb{R}^d$ with finite p^{th} moments, we denote their p-Wasserstein distance with respect to the Euclidean metric by $\mathcal{W}_p(\mu,\nu)$. Whenever $X \sim \mu$ and $Y \sim \nu$, we use $\mathsf{KL}\left(X\middle|\middle|Y\right)$ to denote $\mathsf{KL}\left(\mu\middle|\middle|\nu\right)$ and $\mathcal{W}_p(X,Y)$ to mean $\mathcal{W}_p(\mu,\nu)$. Similarly, given a sigma algebra $\mathcal{G}$ over the same measure space as the random variables X, Y, we use $\mathsf{KL}\left(X\middle|\middle|Y|\mathcal{G}\right)$ to denote the $\mathcal{G}$ measurable random variable $\mathsf{KL}\left(\mathrm{Law}\left(X|\mathcal{G}\right)\middle|\middle|\mathrm{Law}\left(Y|\mathcal{G}\right)\right)$. The first order oracle complexity of any algorithm refers to the number of stochastic gradient computations, i.e., the number of computations of the for, $\nabla f(\mathbf{x},\xi)$ (see Section 2 for a formal construction). We use $\overset{d}{=}$ to denote equality in distribution and $\overset{d}{\approx}$ to denote approximate equality in distribution. We use $[n]$ and (n) to denote the sets $\{1,\dots,n\}$ and $\{0,\dots,n-1\}$ respectively. By C we denote universal constants which can change in every appearance. By $\lesssim$ we denote $\leq$ up-to universal constants.

In Sections 2 and 3, we define the exact problem setup and state the set of assumptions which we use. In Section 4, we discuss our main technical results and in Section 5 we apply them to establish the convergence of SGLD, RBM, CC-SGLD, and CC-RBM and discuss their significance.

2. Preliminaries

LMC and SGLD We consider the problem of sampling from a distribution over $\mathbb{R}^d$ with density $\pi^*(\mathbf{x}) \propto \exp(-F(\mathbf{x}))$. Under minimal assumptions, Langevin SDE $\mathrm{d}\mathbf{x}_t = -\nabla F(\mathbf{x}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$ admits π^* as the stationary distribution (Roberts and Tweedie, 1996). Here, $\mathrm{d}B_t$ denotes the standard Brownian increment in $\mathbb{R}^d$ (Stroock and Varadhan, 1997). The forward Euler discretization of this SDE with step size $\eta > 0$ leads to LMC. Stochastic Gradient Langevin Dynamics (SGLD) is a stochastic approximation of the LMC updates, which arises as follows : For $k \in \mathbb{N}\cup\{0\}$ and $i \in \mathbb{N}$, let $\xi_{k,i} \overset{\text{iid}}{\sim} \mathbb{P}_\xi$ be a sequence of random variables (supported on the domain Ξ) that are sampled independently of $\mathbf{x}_0$, $(\epsilon_k)_{k\in\mathbb{N}\cup\{0\}}$. Given access to a first order stochastic gradient oracle, which inputs $\mathbf{x} \in \mathbb{R}^d$ and batch size $B \in \mathbb{N}$, and outputs stochastic gradients $\nabla f(\mathbf{x},\xi_{k,1}),\dots,\nabla f(\mathbf{x},\xi_{k,B})$ such that $\mathbb{E}\left[\nabla f(\mathbf{x}_k,\xi_{k,1})\middle|\mathbf{x}_k = \mathbf{x}\right] = \nabla F(\mathbf{x})$, the updates of SGLD are given by:

$$\hat{\mathbf{x}}_{k+1} \leftarrow \hat{\mathbf{x}}_k - \frac{\eta}{B}\sum_{i=1}^{B}\nabla f(\hat{\mathbf{x}}_k,\xi_{k,i}) + \sqrt{2\eta}\epsilon_k,\ \epsilon_k \sim \mathcal{N}(0,\mathbf{I}) \tag{SGLD}$$

Covariance Correction When conditioned on $\hat{\mathbf{x}}_k = \mathbf{x}$, the stochastic approximation noise in SGLD, which is given by $\frac{1}{B}\sum_{i=1}^{B}(\nabla f(\hat{\mathbf{x}}_k, \xi_{k,i}) - \nabla F(\hat{\mathbf{x}}_k))$ is close in distribution to $\mathcal{N}(0, \Sigma(\mathbf{x}))$ due to the Central Limit Theorem (CLT). Thus, SGLD updates can be approximately expressed as:

$$\hat{\mathbf{x}}_{k+1} \overset{d}{\approx} \hat{\mathbf{x}}_k - \eta\nabla F(\hat{\mathbf{x}}_k) + \eta\sqrt{\Sigma(\hat{\mathbf{x}}_k)}\tilde{\epsilon}_k + \sqrt{2\eta}\epsilon_k$$

where $\tilde{\epsilon}_k \sim \mathcal{N}(0, \mathbf{I})$ independent of ϵ_k. Since $\sqrt{\eta\frac{\Sigma(\hat{\mathbf{x}}_k)}{2}}\tilde{\epsilon}_k + \epsilon_k \sim \mathcal{N}\left(0, \mathbf{I} + \frac{\eta\Sigma(\hat{\mathbf{x}}_k)}{2}\right)$, we conclude that SGLD approximately resembles LMC but driven with higher noise covariance $\mathbf{I} + \frac{\eta\Sigma(\hat{\mathbf{x}}_k)}{2}$. Thus, given an estimate $\hat{\Sigma}(\hat{\mathbf{x}}_k)$ of $\Sigma(\hat{\mathbf{x}}_k)$ (computed using another random batch of size B) one can correct for this additional noise via a covariance corrected update as follows:

$$\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \frac{\eta}{B}\sum_{i=1}^{B}\nabla f(\hat{\mathbf{x}}_k, \xi_{k,i}) + \sqrt{2\eta\big(\mathbf{I} - \tfrac{\eta\hat{\Sigma}(\hat{\mathbf{x}}_k)}{2}\big)}\epsilon_k$$

The estimator $\hat{\Sigma}(\hat{\mathbf{x}}_k)$ is specified in Section 5.4. To ameliorate the computational expense of computing the matrix square root, we apply the linearization $\sqrt{\mathbf{I} - \eta\hat{\Sigma}(\hat{\mathbf{x}}_k)/2} \approx \mathbf{I} - \eta\hat{\Sigma}(\hat{\mathbf{x}}_k)/4$. The resulting linearized update, called Covariance Corrected SGLD (or CC-SGLD) is as follows,

$$\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \frac{\eta}{B}\sum_{i=1}^{B}\nabla f(\hat{\mathbf{x}}_k, \xi_{k,i}) + \sqrt{2\eta}\left(\mathbf{I} - \frac{\eta\hat{\Sigma}(\hat{\mathbf{x}}_k)}{4}\right)\epsilon_k \tag{CC-SGLD}$$

As we shall show in Section 5.4, the estimator $\hat{\Sigma}(\hat{\mathbf{x}}_k)$ is chosen to be a sum of rank 1 matrices such that $(\mathbf{I} - \frac{\eta\hat{\Sigma}(\hat{\mathbf{x}}_k)}{4})\epsilon_k$ can be evaluated with $O(dB)$ computational cost. Thus, each step of CC-SGLD has the same oracle complexity and computational complexity as that of SGLD.

IPD, RBM and CC-RBM Let $\mathbf{x}_0^1, \ldots, \mathbf{x}_0^n \in \mathbb{R}^d$ be the initial positions of n particles drawn from an arbitrary initial distribution. We denote their positions at time $k \in \mathbb{N} \cup \{0\}$ by $\mathbf{x}_k^i$ for $i \in [n]$. Let $\epsilon_i^k \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$ be independent of the initial positions. Fixing a step-size $\eta > 0$ and diffusion strength $\sigma > 0$, the resulting discrete-time aggregation-diffusion dynamics for the system, driven by an external force $\mathbf{g}_k^i : \mathbb{R}^d \to \mathbb{R}^d$ and interaction forces $\mathbf{K}_k^{ij} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, is known as Intracting Particle Dynamics (IPD). Popular choices for $\mathbf{g}_k^i$ and $\mathbf{K}_k^{ij}$ are $\mathbf{g}_k^i = -\nabla\phi(\mathbf{x}_k^i)$ and $\mathbf{K}_k^{ij}(\mathbf{x}, \mathbf{y}) = -\nabla_{\mathbf{x}}\psi(\|\mathbf{x} - \mathbf{y}\|)$, where ϕ and ψ are the confining and interaction potentials respectively. Implementing K steps of IPD involves $\Theta(n^2K)$ evaluations of $\mathbf{K}_k^{ij}$ which may be prohibitively expensive. To mitigate this, we define the Random Batch Method (RBM) and Covariance Corrected Random Batch Method (CC-RBM). At each time step k, the Random Batch Method draws $I_k^{i1}, \ldots, I_k^{iB} \overset{\text{iid}}{\sim} \mathsf{Uniform}([n])$ for $B \in [n]$, independent of everything else, and deploys a stochastic approximation of the inter-particle interactions as, $\hat{\mathbf{K}}_k^i(\hat{\mathbf{x}}_k) = \frac{1}{B}\sum_{j=1}^{B}\mathbf{K}_k^{iI_k^{ij}}(\hat{\mathbf{x}}_k^i, \hat{\mathbf{x}}_k^{I_k^{ij}})$. The RBM updates are then given by

$$\hat{\mathbf{x}}_{k+1}^i = \hat{\mathbf{x}}_k^i + \eta\mathbf{g}_k^i(\hat{\mathbf{x}}_k^i) + \eta\hat{\mathbf{K}}_k^i(\hat{\mathbf{x}}_k) + \sqrt{\eta}\sigma\epsilon_k^i \tag{RBM}$$

Covariance Corrected RBM additionally samples $J_k^{i1}, \bar{J}_k^{i1}, \ldots, J_k^{iB'}, \bar{J}_k^{iB'} \overset{\text{iid}}{\sim} \mathsf{Uniform}([n])$ for some $B' \in [n]$, and constructs the covariance estimator for $\hat{\mathbf{K}}_k^i$ as follows,

$$\hat{\Sigma}_k^i := \frac{1}{2BB'}\sum_{l=1}^{B'}\left(\mathbf{K}_k^{iJ^{il}}(\hat{\mathbf{x}}_k^i, \hat{\mathbf{x}}_k^{J_k^{il}}) - \mathbf{K}_k^{i\bar{J}^{il}}(\hat{\mathbf{x}}_k^i, \hat{\mathbf{x}}_k^{\bar{J}_k^{il}})\right)\left(\mathbf{K}_k^{iJ^{il}}(\hat{\mathbf{x}}_k^i, \hat{\mathbf{x}}_k^{J_k^{il}}) - \mathbf{K}_k^{i\bar{J}^{il}}(\hat{\mathbf{x}}_k^i, \hat{\mathbf{x}}_k^{\bar{J}_k^{il}})\right)^{\mathsf{T}}$$

The covariance corrected update for RBM is then given by

$$\hat{\mathbf{x}}^i_{k+1} = \hat{\mathbf{x}}^i_k + \eta \mathbf{g}^i_k(\hat{\mathbf{x}}^i_k) + \eta \hat{\mathbf{K}}^i_k(\hat{\mathbf{x}}_k) + \sigma\sqrt{\eta}\sqrt{\mathbf{I} - {}^{\eta}/_{\sigma^2}\hat{\Sigma}^i_k}\,\epsilon^i_k$$

Similar to CC-SGLD, approximating the matrix square-root as $\sqrt{\mathbf{I} - {}^{\eta}/_{\sigma^2}\hat{\Sigma}^i_k} \approx \mathbf{I} - {}^{\eta}/_{2\sigma^2}\hat{\Sigma}^i_k$ gives us the following update for Covariance Corrected RBM (or CC-RBM).

$$\hat{\mathbf{x}}^i_{k+1} = \hat{\mathbf{x}}^i_k + \eta \mathbf{g}^i_k(\hat{\mathbf{x}}^i_k) + \eta \hat{\mathbf{K}}^i_k(\hat{\mathbf{x}}_k) + \sigma\sqrt{\eta}\left(\mathbf{I} - \tfrac{\eta}{2\sigma^2}\hat{\Sigma}^i_k\right)\epsilon^i_k \tag{CC-RBM}$$

Since $\hat{\Sigma}^i_k$ is a sum of rank-1 matrices, the cost of evaluating $(\mathbf{I} - \frac{\eta}{2\sigma^2}\hat{\Sigma}^i_k)\epsilon^i_k$ is $O(dB)$. Thus, CC-RBM has the same computational complexity as that of RBM.

3. Assumptions

Our analysis of SGLD makes the following expected smoothness assumption on the potential F, which is weaker than that of prior works which assume component smoothness (Raginsky et al., 2017; Zou et al., 2021; Kinoshita and Suzuki, 2022), i.e., smoothness of $f(\mathbf{x}, \xi)\ \forall\, \xi \in \Xi$.

Assumption 1 (Expected Hölder Smoothness) *$F(\mathbf{x})$ is s-Hölder smooth (or s-gradient Hölder continuous) for some $s \in (0, 1]$, i.e., there exists a constant $L \geq 0$ such that $\nabla F(\mathbf{x})$ satisfies:*

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|^s, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{s-Hölder}$$

When $s = 1$, $F(\mathbf{x})$ is said to be L-smooth (or L-gradient Lipschitz), i.e.,

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{L-smooth}$$

We use the following growth condition to control the stochastic gradients. As shown in Appendix G.1, this assumption is significantly weaker than that of prior works. (Raginsky et al., 2017; Zou et al., 2021; Kinoshita and Suzuki, 2022).

Assumption 2 (Almost-sure Stochastic Gradient Growth) *For $M, G \geq 0$, the functions $f(\mathbf{x}, \xi)$ satisfy:*

$$\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\| \leq M\|\mathbf{x}\| + G, \quad \forall\, \mathbf{x} \in \mathbb{R}^d, \xi \in \Xi \tag{lin-growth a.s}$$

For the analysis of SGLD in the L-smooth setting, our techniques generalize to the following relaxed assumption on the stochastic gradients which is satisfied for a large class of problems.

Assumption 3 (Subgaussian Stochastic Gradient Growth) *For $M, G \geq 0$, the functions $f(\mathbf{x}, \xi)$ satisfy the following norm-subgaussianity condition (Jin et al., 2019) for the stochastic gradients:*

$$\mathbb{P}\left\{\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\| \geq t \middle| \mathbf{x}\right\} \leq 2\exp\left(-\frac{t^2}{2\left[M\|\mathbf{x}\| + G\right]^2}\right)\ \forall\, \mathbf{x} \in \mathbb{R}^d \tag{lin-growth subG}$$

To establish the convergence of SGLD to the target distribution, we impose the following isoperimetric assumption on π^* proposed by Latała and Oleszkiewicz (2000), which interpolates between the well known Poincare and Log-Sobolev Inequalities (Bakry et al., 2014).

Assumption 4 (Latała-Oleskiewicz Inequality) *The target π^* satisfies the Latała-Oleskiewicz inequality of order $\alpha \in [1, 2]$, i.e., for every smooth function $g : \mathbb{R}^d \to \mathbb{R}$, π^* satisfies the following inequality for some constant $\lambda_{\mathsf{LO}(\alpha)}$*

$$\sup_{p \in (1,2)} \frac{\mathbb{E}_{\pi^*}\left[g^2\right] - \mathbb{E}_{\pi^*}\left[g^p\right]^{2/p}}{(2-p)^{2-2/\alpha}} \leq \frac{1}{\lambda_{\mathsf{LO}(\alpha)}} \mathbb{E}_{\pi^*}\left[\|\nabla g\|^2\right] \tag{α-LO}$$

When $\alpha = 1$, α-LO is equivalent to the Poincare Inequality with constant $\lambda_{\mathsf{PI}} = \lambda_{\mathsf{LO}(1)}$,

$$\mathbb{E}_{\pi^*}\left[g^2\right] - \mathbb{E}_{\pi^*}[g]^2 \leq \frac{1}{\lambda_{\mathsf{PI}}} \mathbb{E}_{\pi^*}\left[\|\nabla g\|^2\right] \tag{PI}$$

When $\alpha = 2$, α-LO reduces to the Logarithmic Sobolev Inequality with constant $\lambda_{\mathsf{LSI}} = \lambda_{\mathsf{LO}(2)}$:

$$\mathbb{E}_{\pi^*}\left[g^2 \log\left(g^2\right)\right] - \mathbb{E}_{\pi^*}\left[g^2\right] \log\left(\mathbb{E}_{\pi^*}\left[g^2\right]\right) \leq \frac{1}{\lambda_{\mathsf{LSI}}} \mathbb{E}_{\pi^*}\left[\|\nabla g\|^2\right] \tag{LSI}$$

We highlight that the exponent s in s-Hölder and the order α in α-LO must satisfy the relation $\alpha \leq 1 + s$. Finally, we impose the following mild technical condition on the iterates of SGLD.

Assumption 5 (Moment Growth) *The iterates of $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_K$ SGLD satisfy the following p-moment growth condition for some $p \geq 1$.*

$$\mathbb{E}[\|\hat{\mathbf{x}}_k\|^q] \leq C_q d^{q/2}, \ \ \forall\, q \in [0, p] \tag{p-moment growth}$$

where $C_q \geq 0 \ \forall\, q \in [0, p]$ and can be an arbitrary non-negative increasing function of q.

As we shall show in Appendix G.2, the p-moment growth condition is milder than assumptions like dissipativity and strong convexity outside of a compact set used in prior works (Zou et al., 2021; Raginsky et al., 2017; Cheng et al., 2020). On the contrary, we assume p-moment growth only for $p = 4$ (in Theorems 5 and 7) or $p = 6$ (in Theorems 8 and 10), making our assumptions much weaker than that of prior works. Furthermore, for finite-sum problems, we remove this assumption by using adaptive batch-sizes (Theorem 6).

4. Technical Results

Our main results on SGLD and RBM, presented in Section 5, are divided into three major groups:

1. Stable convergence results for SGLD which use a conditional CLT type result (Lemma 2) to establish a differential inequality for the KL divergence to the target distribution (Lemma 3).

2. A sharp convergence analysis of the law of the trajectory of SGLD (and RBM), to the law of the trajectory of LMC (and IPD), without any smoothness or isoperimetric assumptions, using a novel Wasserstein CLT (Lemma 4) and associated entropic CLT.

3. Analysis of CC-SGLD and CC-RBM using the above entropic CLTs.

In this section, we establish our key technical results like Lemmas 1, 2 and 4.

4.1. Conditional CLT Type Analysis and Descent EVI

Using $\mathbf{N}(\hat{\mathbf{x}}_k, \xi_k) := \frac{1}{B}\sum_{i=1}^{n} \nabla f(\hat{\mathbf{x}}_k, \xi_{k,j}) - \nabla F(\hat{\mathbf{x}}_k)$ to denote the stochastic gradient noise, we can express the SGLD updates as follows :

$$\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \eta(\nabla F(\hat{\mathbf{x}}_k) + \mathbf{N}(\hat{\mathbf{x}}_k, \xi_k)) + \sqrt{2\eta}\epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

Adapting the arguments of Vempala and Wibisono (2019) to SGLD, we construct a piecewise-continuous stochastic process $(\mathbf{x}_t)_{t\in[0,K\eta]}$ defined as follows for any d-dimensional Brownian motion B_t (which is independent of $\mathbf{x}_0$):

$$\begin{aligned} &\mathrm{Law}\,(\mathbf{x}_0) = \mathrm{Law}\,(\hat{\mathbf{x}}_0) \\ &\qquad \mathrm{d}\mathbf{x}_t = -\left[\nabla F(\mathbf{x}_{k\eta}) + \mathbf{N}(\mathbf{x}_{k\eta}, \xi_k)\right]\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \ t \in [k\eta, (k+1)\eta], \ k \in [K] \end{aligned}$$

Note that $\mathbf{x}_t \overset{d}{=} \mathbf{x}_{k\eta} - (t - k\eta)(\nabla F(\mathbf{x}_{k\eta}) + \mathbf{N}(\mathbf{x}_{k\eta}, \xi_k)) + \sqrt{2(t-k\eta)}\mathbf{z}_t, \ \mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$ where $t \in [k\eta, (k+1)\eta]$, $k \in (K)$. An inductive argument shows that $\mathrm{Law}\,(\mathbf{x}_{k\eta}) = \mathrm{Law}\,(\hat{\mathbf{x}}_k)$. To this end, we call $(\mathbf{x}_t)_{t\in[0,K\eta]}$ an *interpolating process* for SGLD, and write $\mu_t = \mathrm{Law}\,(\mathbf{x}_t)$. We shorten $\mathbf{N}(\mathbf{x}_{k\eta}, \xi_k)$ to $\mathbf{N}$, whenever clear from the context. Our analysis begins by establishing a differential inequality for the time-evolution of KL divergence along the interpolating process. We present the proof of this lemma in Appendix B.2.

Lemma 1 *Assume* $\mathsf{KL}\,(\mu_0 || \pi^*) < \infty$. *Then, for any* $k \in (K)$ *and* $t \in [k\eta, (k+1)\eta]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\,(\mu_t || \pi^*) \leq -\frac{1}{2}\mathsf{FD}\,(\mu_t || \pi^*) + \mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] + \mathbb{E}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta}, \mathbf{x}_t]\|^2\right]$$

The term $\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right]$ in Lemma 1 corresponds to the discretization error of SGLD, which can be controlled in a straightforward manner. The key technical challenge is to sharply bound the term $\mathbb{E}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta}, \mathbf{x}_t]\|^2\right]$, which corresponds to the error due to stochastic approximation. When conditioned on $\mathbf{x}_{k\eta}$, $\mathbf{N}$ is an average of B i.i.d random vectors, and hence, is approximately Gaussian due to the Central Limit Theorem. In fact, if $\mathbf{N}_j|\mathbf{x}_{k\eta} \sim \mathcal{N}(0, \Sigma)$ were exactly Gaussian, then $\mathbf{N}$ and $\mathbf{x}_t$ would be jointly Gaussian conditioned on $\mathbf{x}_{k\eta}$. In this case, standard results on conditional expectations of Gaussians would imply that: $\mathbb{E}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta}, \mathbf{x}_t]\|^2\right] \leq \frac{4(t-k\eta)(\mathrm{Tr}(\Sigma))^2}{B^2}$, where $\mathbb{E}\left[\mathrm{Tr}(\Sigma)^2\right] \lesssim (M^4 C_4 d^2 + G^4)$ under the lin-growth subG and p-moment growth conditions for $p = 4$. However, since $\mathbf{N}$ need not be Gaussian in general, we require a conditional CLT type result (Dedecker and Merlevède, 2002) to sharply control the stochastic approximation error. To this end, we derive sharp bounds for $\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta}, \mathbf{x}_t]\|^2$ by exploiting the fact that $\mathbf{N}$ is conditionally an i.i.d. sum of zero-mean vectors. The proof of this result is presented in Appendix B.3.

Lemma 2 *Let the lin-growth subG and p-moment growth conditions be satisfied with* $p = 4$. *Then,*

$$\mathbb{E}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta}, \mathbf{x}_t]\|^2\right] \leq \frac{C(t-k\eta)(M^4 C_4 d^2 + G^4)}{B^2}$$

Equipped with Lemmas 1 and 2, we derive the following differential inequality for the time evolution of the KL divergence along the trajectory of the interpolating process. The proof of this result is presented in Appendix B.4

Lemma 3 *Let the L-smooth, lin-growth subG and p-moment growth conditions be satisfied with $p = 4$. Then, for $\eta \leq 1/6L$, the following inequality is satisfied for any $k \in (K)$ and $t \in [k\eta, (k+1)\eta]$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t \middle|\middle| \pi^*\right) \leq -\frac{1}{4}\mathsf{FD}\left(\mu_t \middle|\middle| \pi^*\right) + CL^2(t-k\eta)d + \frac{CL(t-k\eta)(M^2C_2 d + G^2)}{B} + \frac{C(t-k\eta)(M^4C_4d^2 + G^4)}{B^2}$$

4.2. Random Function Representation and Entropic CLT based Analysis

We now describe our technical approach to the analysis of SGLD and RBM via the random function representation and Gaussian-smoothed Wasserstein CLTs. While the description below focuses on SGLD, our analysis of CC-SGLD, RBM and CC-RBM applies similar arguments.

Since LMC is a discrete-time Markov Process (with iterates $\mathbf{x}_0, \mathbf{x}_1, \dots$), it admits a random function representation, i.e., for any $K \in \mathbb{N}$, there exists a measurable function H_K such that $\mathbf{x}_{1:K+1} = H_K(\mathbf{x}_0, \epsilon_0, \dots, \epsilon_K)$. Moreover, from equation (1), we recall that the SGLD iterates can be expressed as $\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \eta\nabla F(\mathbf{x}_k) + \sqrt{2\eta}(\hat{\epsilon}_k + \sqrt{\eta/2}\mathbf{N}_k)$, $\hat{\epsilon}_k \sim \mathcal{N}(0, \mathbf{I})$, $\mathbf{N}_k = \mathbf{N}(\hat{\mathbf{x}}_k, \xi_k)$. Comparing with the update rule of LMC, we conclude that *LMC and SGLD must admit the same random function representation*, i.e., $\hat{\mathbf{x}}_{1:K+1} = H_K(\hat{\mathbf{x}}_0, \hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_K)$ where $\hat{\mathbf{z}}_k = \hat{\epsilon}_k + \sqrt{\eta/2}\mathbf{N}_k$. Applying the data-processing inequality and the chain rule for KL divergence, and using the fact that $\mathrm{Law}\left(\mathbf{x}_0\right) = \mathrm{Law}\left(\hat{\mathbf{x}}_0\right)$, we conclude the following:

$$\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K} \middle|\middle| \mathbf{x}_{1:K}\right) \leq \sum_{k=0}^{K} \mathbb{E}\left[\mathsf{KL}\left(\hat{\mathbf{z}}_k | \mathcal{F}_k \middle|\middle| \epsilon_k\right)\right],$$

where $\mathcal{F}_k = \sigma(\hat{\mathbf{x}}_{0:K}, \hat{\mathbf{z}}_{0:K-1})$. Since $\hat{\mathbf{z}}_k = \hat{\epsilon}_k + \sqrt{\eta/2}\mathbf{N}_k$ where $\mathbf{N}_k|\mathcal{F}_k$ is approximately Gaussian and $\hat{\epsilon}_k$ is exactly Gaussian, controlling the term $\mathsf{KL}\left(\hat{\mathbf{z}}_k|\mathcal{F}_k \middle|\middle| \epsilon_k\right)$ involves establishing an entropic CLT for $\hat{\mathbf{z}}_k|\mathcal{F}_k$. To this end, we prove a reverse T_2-type inequality for Gaussian convolutions to control this term as $\mathsf{KL}\left(\hat{\mathbf{z}}_k|\mathcal{F}_k \middle|\middle| \epsilon_k\right) \leq \mathcal{W}_2^2(\sqrt{\mathbf{I} + \eta\Sigma_k}\mathbf{X}, \mathbf{Y}|\mathcal{F}_k) + \mathcal{W}_2^2(\mathbf{X} + \sqrt{\eta}\mathbf{N}_k, \sqrt{\mathbf{I} + \eta\Sigma_k}\mathbf{Y}|\mathcal{F}_k)$ where $\Sigma_k = \mathbb{E}\left[\mathbf{N}_k\mathbf{N}_k^T|\hat{\mathbf{x}}_k\right]$ and $\mathbf{X}, \mathbf{Y} \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$. The first term is bounded by directly computing the Wasserstein distance between zero-mean Gaussians. To control the second term, we quantify the approximate Gaussianity of $\mathbf{N}_k$ by developing a novel Wasserstein CLT for Gaussian convolutions of bounded random vectors, which may be of independent interest. This result, which is an extension of Zhai (2018), is proved in Appendix C

Lemma 4 ($\mathcal{W}_2$ CLT for Gaussian Convolutions of Bounded RVs) *Let $\mathbf{Y}_1, \dots, \mathbf{Y}_B$ be i.i.d random vectors in $\mathbb{R}^d$ satisfying $\mathbb{E}\left[\mathbf{Y}_i\right] = 0$, $\mathbb{E}\left[\mathbf{Y}_i\mathbf{Y}_i^T\right] = \Sigma_{\mathbf{Y}}$ and $\|\mathbf{Y}_i\| \leq \beta$ almost surely, and let $\mathbf{Y} = \frac{1}{\sqrt{B}}\sum_{i=1}^{B}\mathbf{Y}_i$. Furthermore, let $\mathbf{X}$ and $\mathbf{Z}$ be sampled iid from $\mathcal{N}(0, \mathbf{I})$, independent of $\mathbf{Y}_1, \dots, \mathbf{Y}_B$. Then, the following holds for $\beta^2 \leq 1/5$ and $\|\Sigma_{\mathbf{Y}}\|_2 \leq 1/5d$*

$$\mathcal{W}_2^2(\sqrt{\mathbf{I} - \Sigma_{\mathbf{Y}}}\mathbf{X} + \mathbf{Y}, \mathbf{Z}) \leq \frac{25\beta^6 d(1 + \log(B))^2}{B}$$

Our entropic CLT analysis does not require any smoothness assumptions on F or isoperimetric assumptions on π^*, and only uses the lin-growth a.s and p-moment growth assumptions for $p = 6$.

5. Results for SGLD and RBM

5.1. Convergence of SGLD under Smoothness and LSI

Our first result is a stable convergence guarantee for the last iterate of SGLD with respect to KL divergence when F is smooth and π^* satisfies LSI. The proof, which is presented in Appendix B.5, follows from Lemma 3 and uses the fact that LSI is a gradient dominance condition for $\mathsf{KL}\left(.\middle|\middle|\pi^*\right)$ in the space of measures (Otto and Villani, 2000), i.e., $\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right) \geq 2\lambda_{\mathsf{LSI}}\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right)$.

Theorem 5 (Convergence of SGLD under Smoothness and LSI) *Let the L-smooth, lin-growth subG and p-moment growth conditions be satisfied with $p = 4$ and let π^* satisfy LSI. Then, for $\eta \leq \lambda_{\mathsf{LSI}}/6L^2$, the last iterate of SGLD satisfies the following.*

$$\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_K\right)\middle|\middle|\pi^*\right) \leq e^{-\lambda_{\mathsf{LSI}}\eta K/2}\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right) + \frac{CL^2\eta d}{\lambda_{\mathsf{LSI}}} + \frac{CL\eta(M^2C_2d + G^2)}{B\lambda_{\mathsf{LSI}}} + \frac{C\eta(M^4C_4d^2 + G^4)}{\lambda_{\mathsf{LSI}}B^2}$$

The first term $e^{-\lambda_{\mathsf{LSI}}\eta K/2}\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)$ arises from the error due to initialization which decays exponentially fast. The second term $\frac{CL^2\eta d}{\lambda_{\mathsf{LSI}}}$, which is independent of the batch-size B and diminishes as $\eta \to 0$, corresponds to the inherent bias of LMC (Bernton, 2018; Wibisono, 2018). The remaining terms, which diminish with increasing batch-size B for any fixed $\eta > 0$, encapsulate the error due to stochastic approximation.

Oracle Complexity Lemma 1 of Vempala and Wibisono (2019) shows that $\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right) = O(d)$ is easily ensured via appropriate Gaussian initialization. Thus, for any $\epsilon \leq 1$, Theorem 5 implies that $\eta = O\left(\epsilon/d\right)$, $B = O(\sqrt{d})$ and $K = \tilde{O}\left(d/\epsilon\right)$ ensures $\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_K\right)\middle|\middle|\pi^*\right) \leq \epsilon$. Thus, Theorem 5 implies an oracle complexity of $\tilde{O}\left(d^{1.5}/\epsilon\right)$ for ϵ-convergence in KL divergence, which translates to $\tilde{O}\left(d^{1.5}/\epsilon^2\right)$ oracle complexity for ϵ-convergence in TV and $\mathcal{W}_2$ via Pinsker's inequality and Talagrand's inequality respectively.

Stability and Tightness of the Analysis For any $\eta \leq \lambda_{\mathsf{LSI}}/6L^2$, and any batch size $B \geq 1$ the rate obtained in Theorem 5 is stable, i.e., does not diverge as $K \to \infty$. Moreover, Theorem 5 doesn't require any uniform warm start assumption. In the noise-free setting (i.e., $M, G = 0$) Theorem 5 recovers the result of Vempala and Wibisono (2019) for LMC upto constant factors.

When F admits a finite-sum structure, i.e., $F(\mathbf{x}) = 1/n\sum_{i=1}^{n} f_i(\mathbf{x})$, we propose an adaptive batch-size schedule for SGLD, for which we establish results similar to Theorem 5 without the p-moment growth assumption. The algorithm, called AB-SGLD, is as follows:

1. $B_k := \min\{n, 1 + \lceil M\|\hat{\mathbf{x}}_k\| + G\rceil\}$; Sample indices $i_1, \dots, i_{B_k} \overset{\text{iid}}{\sim} \mathsf{Uniform}([n])$
2. $\hat{\mathbf{x}}_{k+1} \leftarrow \hat{\mathbf{x}}_k - \eta/B_k \sum_{j=1}^{B_k} \nabla f_{i_j}(\hat{\mathbf{x}}_k) + \sqrt{2\eta}\epsilon_k$ where $\epsilon_k \sim \mathcal{N}(0, \mathbf{I})$

We refer to Appendix B.6 for the proof of Theorem 6. For simplicity, we assume $\eta \leq 1/8$ and $M \geq L \geq 1$.

Theorem 6 (Stable Convergence of AB-SGLD) *Let the L-smooth and lin-growth subG conditions be satisfied, and let π^* satisfy LSI. Then, for $\eta \leq \frac{\lambda_{\mathsf{LSI}}^2}{8L(LM+128M^2)}$ and any $K \in \mathbb{N}$:*

$$\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_K\right)\middle|\middle|\pi^*\right) \leq e^{-\lambda_{\mathsf{LSI}}\eta K/4}\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right) + \frac{C\eta}{\lambda_{\mathsf{LSI}}}\left(L^2d + M^2\mathbf{m}_2^2 + G^2\right) + \frac{CL^2\eta^2}{\lambda_{\mathsf{LSI}}}\left(M\mathbf{m}_1 + M + G\right)$$

where $\mathbf{m}_p = \mathbb{E}_{\pi^*}[\|\mathbf{x}\|^p]^{1/p}$*. Furthermore, the amortized batch size* $\bar{B} = {}^1/_K \sum_{k=0}^{K} \mathbb{E}[B_k]$ *satisfies,*

$$\bar{B} \lesssim 2 + G + \frac{M}{\lambda_{\mathsf{LSI}}^{3/2}\eta K}\sqrt{\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)} + \frac{M\sqrt{\eta}}{\lambda_{\mathsf{LSI}}}\left(L\sqrt{d} + M\mathbf{m}_2 + G\right) + \frac{L\eta}{\lambda_{\mathsf{LSI}}}\sqrt{M\mathbf{m}_1 + M + G}$$

Expected Oracle Complexity Since LSI implies subgaussianity (Van Handel, 2014), it is reasonable to consider $\mathbf{m}_1, \mathbf{m}_2 = O(\sqrt{d})$. Moreover, appropriate Gaussian initialization ensures $\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right) = O(d)$. Hence, for any $\epsilon \leq 1$, setting $\eta = O({}^\epsilon/_d)$ and $K = \tilde{O}\left({}^d/_\epsilon\right)$ ensures ϵ-convergence in KL divergence. Under this setting, the amortized batch size of AB-SGLD is $\bar{B} = \tilde{O}(\sqrt{d})$. Hence, we conclude that the expected oracle complexity of AB-SGLD to achieve ϵ-convergence in KL is $K\bar{B} = \tilde{O}\left({}^{d^{1.5}}/_\epsilon\right)$, which further implies $\tilde{O}\left({}^{d^{1.5}}/_{\epsilon^2}\right)$ oracle complexity for ϵ-convergence in TV and $\mathcal{W}_2$.

5.2. First Order Stationarity of Averaged SGLD under Smoothness

Recalling the definition of μ_t from Section 4.1, we denote the averaged law of SGLD as $\bar{\mu}_{K\eta} = {}^1/_{K\eta}\int_0^{K\eta}\mu_t dt$. We prove a stable convergence guarantee for $\mathsf{FD}\left(\bar{\mu}_{K\eta}\middle|\middle|\pi^*\right)$ without imposing any isoperimetry assumptions on π^*. Such guarantees for sampling are considered analogous to first-order stationary point analysis for nonconvex optimization (Balasubramanian et al., 2022; Chewi et al., 2022b). Our result, which follows from Lemma 3 and convexity of Fisher Divergence, implies a stable convergence guarantee in TV if π^* satisfies PI. We present the proof in Appendix B.7.

Theorem 7 (First Order Stationarity of SGLD) *Let the L-smooth, lin-growth subG and p-moment growth conditions be satisfied with* $p = 4$*. For any* $\eta \leq {}^1/_{6L}$*, the following holds*

$$\mathsf{FD}\left(\bar{\mu}_{K\eta}\middle|\middle|\pi^*\right) \lesssim \frac{\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)}{K\eta} + L^2\eta d + \frac{L\eta(M^2C_2d + G^2)}{B} + \frac{\eta(M^4C_4d^2 + G^4)}{B^2}$$

Additionally, the following holds if π^* *satisfies PI,*

$$TV(\bar{\mu}_{K\eta}, \pi^*)^2 \lesssim \frac{\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)}{\lambda_{\mathsf{PI}}K\eta} + \frac{L^2\eta d}{\lambda_{\mathsf{PI}}} + \frac{L\eta(M^2C_2d + G^2)}{\lambda_{\mathsf{PI}}B} + \frac{\eta(M^4C_4d^2 + G^4)}{\lambda_{\mathsf{PI}}B^2}$$

Sampling from $\bar{\mu}_{K\eta}$ It is easy to sample from $\bar{\mu}_{K\eta}$ without any added sample complexity as follows : 1. Sample $t_0 \sim \mathsf{Unif}[0, K\eta]$. Let k_0 be the largest integer such that $\eta k_0 \leq t_0$. 2. Run SGLD for k_0 steps 3. Perform a partial update $\bar{\mathbf{x}} = \hat{\mathbf{x}}_{k_0} - (t_0 - \eta k_0)\mathbf{g}_{k_0} + \sqrt{2(t_0 - \eta k_0)}\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, where $\mathbf{g}_{k_0}$ is the stochastic gradient at step k_0. By the construction of the interpolating process, it is easy to show that $\mathrm{Law}\left(\bar{\mathbf{x}}\right) = \bar{\mu}_{K\eta}$.

Oracle Complexity Appropriate Gaussian initialization ensures that $\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right) = O(d)$ Thus, for $\epsilon \leq 1$, setting $\eta = O\left({}^\epsilon/_d\right)$, $B = O(\sqrt{d})$ and $K = O\left({}^{d^2}/_{\epsilon^2}\right)$ suffices to ensure ϵ-convergence of averaged SGLD in FD, thereby implying an oracle complexity of $O\left({}^{d^{2.5}}/_{\epsilon^2}\right)$. Under the same setting, Theorem 7 also implies an oracle complexity of $O\left({}^{d^{2.5}}/_{\epsilon^4}\right)$ for ϵ-convergence in TV when π^* satisfies PI.

Stability For any $\eta \leq {}^1/_{6L}$ and $B \geq 1$, the rates presented in Theorem 7 are stable as they don't diverge when $K \to \infty$, nor do they require any uniform warm start assumption. Furthermore, in the noise-free setting (i.e. $M, G = 0$), Theorem 7 recovers the result of (Balasubramanian et al., 2022, Theorem 1 and Corollary 5) upto constant factors.

Batch Size Theorems 5, 6 and 7 require a batch size of $B = O(\sqrt{d})$ for ϵ-convergence, which significantly improves upon prior works. In comparison, Zou et al. (2021) require $B = O(d)$ and Raginsky et al. (2017) assume $B = O\left(\mathsf{poly}\left({}^{d}/_{\epsilon}\right)\right)$. The VR-SGLD algorithm of Kinoshita and Suzuki (2022) requires a batch size and inner loop length of $O(\sqrt{n})$. Note that $n \gg d$ for most practical scenarios.

5.3. Analysis of SGLD and RBM via Wasserstein CLT

We use the arguments presented in Section 4.2 to prove convergence bounds between the trajectories of SGLD and LMC (as well as RBM and IPD) under minimal assumptions (assuming only lin-growth a.s and p-moment growth). The proof is presented in Appendix D.2

Theorem 8 (Statistical Indistinguishability of SGLD and LMC) *Let $\mathbf{x}_{1:K}$ and $\hat{\mathbf{x}}_{1:K}$ and denote the iterates of LMC and SGLD respectively, with the same step-size and initialization. Furthermore, let the lin-growth a.s and p-moment growth conditions be satisfied. Then,*

$$\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K} \middle\| \mathbf{x}_{1:K}\right) \lesssim \frac{\eta^2 K}{B^2}\left(M^4 C_4 d^2 + G^4\right) + \frac{\eta^3 K}{B^4}\left(M^6 C_6 d^6 + G^6 d^3\right)(1 + \log B)^2$$

Theorem 8 implies that the convergence of LMC to π^* is nearly sufficient to ensure the convergence of SGLD. To this end, Theorem 8 gives us the following highly general technique to derive last-iterate TV guarantees for SGLD : 1. Using any existing convergence guarantee for LMC, choose η, K such that $\mathsf{TV}(\mathbf{x}_K, \pi^*) \leq \tilde{\Theta}(\epsilon)$ 2. Set B such that $\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K} \middle\| \mathbf{x}_{1:K}\right) \leq \tilde{\Theta}(\epsilon^2)$. Consequently, $\mathsf{TV}(\hat{\mathbf{x}}_K, \pi^*) \leq \tilde{\Theta}(\epsilon)$ by the Data Processing Inequality and Pinsker's inequality.
We quantitatively demonstrate this technique in Appendix D.3, we use Theorem 8 to prove last-iterate TV guarantees for SGLD under the assumptions of s-Hölder and α-LO, obtaining an oracle complexity of $\tilde{\Theta}(\frac{d^\gamma}{\epsilon^{2/s}})$ for ϵ-convergence in TV, where $\gamma = \max\left\{1 + \beta\left(1 + {}^{1}/_{s}\right), {}^{3}/_{2}\left(1 + \beta\right) + {}^{\beta}/_{2s}\right\}$ and $\beta = {}^{2}/_{\alpha} - 1$. For $s = 1$, this result implies an oracle complexity of $\tilde{\Theta}({}^{d^{3.5}}/_{\epsilon^2})$ when π^* satisfies PI (i.e., $\alpha = 1$). This has a better ϵ dependence but worse d dependence than the TV guarantee of Theorem 7. However, unlike Theorems 5, 7 and 6, these guarantees are not stable.

Our techniques apply whenever a sampling algorithm and its stochastic approximation admit the same random function representation, and as such, can be extended to other sampling algorithms such as HMC. In fact, using similar techniques, we obtain the following statistical indistinguishability guarantee between RBM and IPD, whose proof is presented in Appendix F.1.

Theorem 9 (Statistical Indistinguishability of IPD and RBM) *Let $(\mathbf{x}^i_k)_{i \in [n], k \in [K]}$ and $(\hat{\mathbf{x}}^i_k)_{i \in [n], k \in [K]}$ denote the iterates of the interacting particle method and the random batch method respectively, with the same initial distribution and step-size η. Suppose $\left\|K^{ij}_k(\hat{\mathbf{x}}^i_k, \hat{\mathbf{x}}^j_k)\right\| \leq M$ holds almost surely, for every $k \in [K]$ and $i, j \in [n]$. Then, the following holds for any $\eta \leq \frac{B\sigma^2}{40M^2 d}$*

$$\mathsf{KL}\left((\hat{\mathbf{x}}^i_k)_{i \in [n], k \in [K]} \middle\| (\mathbf{x}^i_k)_{i \in [n], k \in [K]}\right) \lesssim \frac{\eta^2 M^4 n K}{B^2 \sigma^4} + \frac{d\eta^3 M^6 n K (1 + \log B)^2}{B^4 \sigma^6}$$

5.4. CC-SGLD and CC-RBM for IPD

We now consider covariance correction introduced in Section 2. For CC-SGLD, we define the estimator $\hat{\Sigma}(\hat{\mathbf{x}}_k)$ as follows:

$$\hat{\Sigma}(\hat{\mathbf{x}}_k) = \begin{cases} 0 \quad \text{if} \quad (M\|\hat{\mathbf{x}}_k\| + G)^2 > {}^{B}/_{5\eta d} \\ {}^{1}/_{2B^2} \sum_{j=1}^{B} \left(\nabla f(\mathbf{x}_k, \xi^{(1)}_{k,j}) - \nabla f(\mathbf{x}_k, \xi^{(2)}_{k,j})\right) \left(\nabla f(\mathbf{x}_k, \xi^{(1)}_{k,j}) - \nabla f(\mathbf{x}_k, \xi^{(2)}_{k,j})\right)^T & \text{o/w} \end{cases}$$

Where $\nabla f(\hat{\mathbf{x}}_k, \xi^{(1)}_{k,1}), \nabla f(\hat{\mathbf{x}}_k, \xi^{(2)}_{k,1}), \dots, \nabla f(\hat{\mathbf{x}}_k, \xi^{(1)}_{k,B}), \nabla f(\hat{\mathbf{x}}_k, \xi^{(2)}_{k,B})$ is a fresh batch of stochastic gradients. We use the entropic CLT approach discussed in Section 4.2 to establish Theorem 10 which is proved in Appendix E.2.

Theorem 10 (Statistical Indistinguishability of CC-SGLD and LMC) *Let the lin-growth a.s and p-moment growth conditions be satisfied with $p = 6$. Then, the iterates of CC-SGLD satisfy the following guarantee.*

$$\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K} \middle|\middle| \mathbf{x}_{1:K}\right) \lesssim \frac{\eta^2 K}{B^3}\left(M^4 C_4 d^2 + G^4\right) + \frac{\eta^3 K}{B^3}\left(M^6 C_6 d^3 + G^6\right) + \frac{\eta^5 K d^6}{B^3} + \frac{\eta^3 K d^6 (1 + \log B)^2}{B^4}.$$

The leading order term in Theorem 10 is $O({}^{\eta^2 K}/_{B^3})$ whereas that in Theorem 8 for SGLD is $O({}^{\eta^2 K}/_{B^2})$. This indicates that CC-SGLD is a better stochastic approximation to LMC in comparison to SGLD, and enjoys faster convergence. Using this, we derive an unstable last-iterate TV guarantee for CC-SGLD under s-Hölder and α-LO. The result, which is stated in Appendix E.3, improves upon that implied by Theorem 8 under the same setting. In particular For $s = 1$, we obtain a $\tilde{\Theta}({}^{d^{4/3}}/_{\epsilon^2})$ oracle complexity for CC-SGLD when π^* satisfies LSI, which improves upon our previous guarantees for SGLD and AB-SGLD under smoothness and LSI. When π^* satisfies PI, we obtain an oracle complexity of $\tilde{\Theta}({}^{d^{10/3}}/_{\epsilon^2})$. In a similar way, we derive the following guarantee for covariance corrected RBM, which is proved in Appendix F.2.

Theorem 11 (Statistical Indistinguishability of IPD and CC-RBM) *Let $(\mathbf{x}^i_k)_{i\in[n],k\in[K]}$ and $(\hat{\mathbf{x}}^i_k)_{i\in[n],k\in[K]}$ denote the iterates of the interacting particle method and the covariance corrected random batch method respectively, with the same initial distribution and step-size η. Suppose $\left\|K^{ij}_k(\hat{\mathbf{x}}^i_k, \hat{\mathbf{x}}^j_k)\right\| \leq M$ holds almost surely, for every $k \in [K]$ and $i, j \in [n]$. Then, for any $\eta \leq \frac{B\sigma^2}{40M^2 d}$*

$$\mathsf{KL}\left((\hat{\mathbf{x}}^i_k)_{i\in[n],k\in[K]} \middle|\middle| (\mathbf{x}^i_k)_{i\in[n],k\in[K]}\right) \lesssim \frac{\eta^2 M^4 n K}{B^2 B' \sigma^4} + \frac{\eta^3 M^6 n K}{B^3 \sigma^6} + \frac{d\eta^3 M^6 n K\,(1 + \log B)^2}{B^4 \sigma^6}$$

As per Theorem 9, a batch size of $B \gg (\eta^2 n K)^{1/2}$ suffices to ensure that the trajectories of all the particles of RBM and IPD stay close in distribution. Similarly, Theorem 11 ensures the same for CC-RBM and IPD when $B \gg (\eta^2 n K)^{1/3}$ (assuming $B = B'$). Both these results significantly improve upon prior works on the convergence of RBM (Jin et al., 2020, 2021) that either require highly stringent regularity assumptions on $\mathbf{g}_k$, $\mathbf{K}_k$ and their derivatives, or suffer from an exponential dependence on ηK (See Table 2).

6. Conclusion and Future Work

In this work, we introduced techniques to give sharp analyses of stochastic approximations of sampling algorithms, significantly improving upon prior work. Future work can further extend these techniques to understand convergence properties under the Rényi Divergence, and other algorithms like Hamiltonian Monte Carlo and randomized midpoint methods.

References

Julio Backhoff, Giovanni Conforti, Ivan Gentil, and Christian Léonard. The mean field schrödinger problem: ergodic behavior, entropy estimates and functional inequalities. *Probability Theory and Related Fields*, 178(1):475–530, 2020.

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

Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling:first-order stationarity guarantees for langevin monte carlo. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2896–2923. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/balasubramanian22a.html.

Espen Bernton. Langevin monte carlo and jko splitting. In *Conference On Learning Theory*, pages 1777–1798. PMLR, 2018.

Andrea L Bertozzi, John B Garnett, and Thomas Laurent. Characterization of radially symmetric finite time blowup in multidimensional aggregation equations. *SIAM Journal on Mathematical Analysis*, 44(2):651–681, 2012.

JA Carrillo, F Hoffmann, AM Stuart, and U Vaes. Consensus-based sampling. *Studies in Applied Mathematics*, 2021.

José A Carrillo, Young-Pil Choi, Claudia Totzeck, and Oliver Tse. An analytical framework for consensus-based global optimization method. *Mathematical Models and Methods in Applied Sciences*, 28(06):1037–1066, 2018.

José A Carrillo, Katy Craig, and Yao Yao. Aggregation-diffusion equations: dynamics, asymptotics, and singular limits. In *Active Particles, Volume 2*, pages 65–108. Springer, 2019.

Yongxin Chen. Density control of interacting agent systems. *arXiv preprint arXiv:2108.07342*, 2021.

Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR, 2018.

Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting, 2020.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1–2. PMLR, 02–05 Jul 2022a. URL https://proceedings.mlr.press/v178/chewi22a.html.

Sinho Chewi, Patrik Gerber, Holden Lee, and Chen Lu. Fisher information lower bounds for sampling. *arXiv preprint arXiv:2210.02482*, 2022b.

Alexandre Joel Chorin. Numerical study of slightly viscous flow. *Journal of fluid mechanics*, 57(4): 785–796, 1973.

Katy Craig, Karthik Elamvazhuthi, Matt Haberland, and Olga Turanova. A blob method method for inhomogeneous diffusion with applications to multi-agent control and sampling. *arXiv preprint arXiv:2202.12927*, 2022.

Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Transactions on automatic control*, 52(5):852–862, 2007.

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

Esther S Daus, Markus Fellner, and Ansgar Jüngel. Random-batch method for multi-species stochastic interacting particle systems. *arXiv preprint arXiv:2109.01897*, 2021.

Jérôme Dedecker and Florence Merlevède. Necessary and sufficient conditions for the conditional central limit theorem. *The Annals of Probability*, 30(3):1044–1081, 2002.

Pierre Degond, Jian-Guo Liu, and Robert L Pego. Coagulation–fragmentation model for animal group-size statistics. *Journal of Nonlinear Science*, 27(2):379–424, 2017.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

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

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

Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.

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

Nathael Gozlan and Christian Léonard. Transport inequalities. a survey. *Markov Processes and Related Fields*, 16:635–736, 2010.

Arnaud Guillin, Christian Léonard, Liming Wu, and Nian Yao. Transportation-information inequalities for markov processes. *Probability theory and related fields*, 144(3):669–695, 2009.

Seung-Yeal Ha and Jian-Guo Liu. A simple proof of the cucker-smale flocking dynamics and mean-field limit. *Communications in Mathematical Sciences*, 7(2):297–325, 2009.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm, 2019. URL https://arxiv.org/abs/1902.03736.

Shi Jin, Lei Li, and Jian-Guo Liu. Random batch methods (rbm) for interacting particle systems. *Journal of Computational Physics*, 400:108877, 2020.

Shi Jin, Lei Li, and Jian-Guo Liu. Convergence of the random batch method for interacting particles with disparate species and weights. *SIAM Journal on Numerical Analysis*, 59(2):746–768, 2021.

Ravi Kannan, László Lovász, and Miklós Simonovits. Random walks and an o*(n5) volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997.

Yuri Kinoshita and Taiji Suzuki. Improved convergence rate of stochastic gradient langevin dynamics with variance reduction and its application to optimization, 2022. URL https://arxiv.org/abs/2203.16217.

Dongnam Ko, Seung-Yeal Ha, Shi Jin, and Doheon Kim. Uniform error estimates for the random batch method to the first-order consensus models with antisymmetric interaction kernels. *Studies in Applied Mathematics*, 146(4):983–1022, 2021.

Rafał Latała and Krzysztof Oleszkiewicz. Between sobolev and poincaré. *Geometric aspects of functional analysis*, pages 147–168, 2000.

Yin Tat Lee and Santosh S Vempala. Convergence rate of riemannian hamiltonian monte carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121, 2018.

Yin Tat Lee and Santosh S Vempala. The manifold joys of sampling (invited talk). In *49th International Colloquium on Automata, Languages, and Programming (ICALP 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

László Lovász and Miklós Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings [1990] 31st annual symposium on foundations of computer science*, pages 346–354. IEEE, 1990.

László Lovász and Santosh Vempala. Hit-and-run from a corner. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 310–314, 2004.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

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

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

Steven Shreve. *Stochastic calculus for finance I: the binomial asset pricing model*. Springer Science & Business Media, 2005.

Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes*, volume 233. Springer Science & Business Media, 1997.

Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32: NeurIPS 2019*, pages 8092–8104, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/65a99bb7a3115fdede20da98b08a370f-Abstract.html.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Liming Wu. Uniformly integrable operators and large deviations for markov processes. *Journal of Functional Analysis*, 172(2):301–376, 2000.

Alex Zhai. A high-dimensional clt in w _2 w 2 distance with near optimal convergence rate. *Probability Theory and Related Fields*, 170:821–845, 2018.

Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, pages 1152–1162. PMLR, 2021.

Contents

Appendix A. More References

LMC and SGLD Discretizations of diffusion processes Bakry et al. (2014) such as Langevin Monte Carlo (LMC) and its closely related cousins like Hamiltonian monte-carlo and underdamped LMC have been widely studied in various settings, including large scale (Vempala and Wibisono, 2019; Chewi et al., 2022a; Cheng and Bartlett, 2018; Dalalyan and Karagulyan, 2019; Lovász and Simonovits, 1990; Lovász and Vempala, 2004; Lee and Vempala, 2018). Recent advancements have focused on the sharp study of LMC with non-log-concave densities and have obtained sharp analyses with minimal assumptions under various metrics like the KL divergence and Renyi divergences whenever the target distribution satisfies certain functional inequalities. Vempala and Wibisono (2019) established precise and succinct convergence bounds whenever the target distribution satisfies log-Sobolev inequalities (LSI) and (Chewi et al., 2022a; Erdogdu and Hosseinzadeh, 2021b) considered settings more general settings like the Latała-Oleskiewicz (LO) Inequality and the Modified Log-Sobolev Inequality (MLSI)

Interacting Particle Methods for Sampling and Optimization Equation (IPD) a system which is made up of multiple agents which evolve in time by interacting with the environment and each other via the aggregation-diffusion equations (Carrillo et al., 2019). These models give rise to rich dynamics, and can lead to various phenomenon like swarming, flocking, chemotaxis and vortex formation (Degond et al., 2017; Vicsek et al., 1995; Cucker and Smale, 2007; Bertozzi et al., 2012; Chorin, 1973) and hence are widely used by the scientific community. This has also has generated significant interest from a mathematical perspective (Backhoff et al., 2020; Carrillo et al., 2019; Bertozzi et al., 2012; Ha and Liu, 2009) where precise behavior of popular models, large deviations principles and convergence to mean field limits have been studied. Inspired by these models, various optimization, sampling and control algorithms have been designed (Carrillo et al., 2018, 2021; Duncan et al., 2019; Liu and Wang, 2016; Craig et al., 2022; Chen, 2021).

Appendix B. Stable Convergence of SGLD under Smoothness

B.1. Technical Lemmas

Lemma 12 *Let X be a non-negative random variable such that $\mathbb{E}\left[X^k\right] < \infty$ for some $k > 1$. Then, the following holds for any $\alpha > 0$,*

$$\mathbb{E}\left[X\mathbb{I}_{\{X\geq\alpha\}}\right] \leq \frac{\mathbb{E}\left[X^k\right]}{\alpha^{k-1}}\left(1+\frac{1}{k-1}\right)$$

Proof Since X is a non-negative random variable, the following holds almost surely,

$$\begin{aligned} X &= \int_0^\infty \mathbb{I}_{\{X \geq t\}} \mathrm{d}t \\ X\mathbb{I}_{\{X \geq \alpha\}} &= \int_0^\infty \mathbb{I}_{\{X \geq \alpha\}} \mathbb{I}_{\{X \geq t\}} \mathrm{d}t \\ &= \int_0^\infty \mathbb{I}_{\{X \geq \max\{\alpha, t\}\}} \mathrm{d}t \\ &= \int_0^\alpha \mathbb{I}_{\{X \geq \alpha\}} \mathrm{d}t + \int_\alpha^\infty \mathbb{I}_{\{X \geq t\}} \mathrm{d}t \\ &= \alpha \mathbb{I}_{\{X \geq \alpha\}} + \int_\alpha^\infty \mathbb{I}_{\{X \geq t\}} \mathrm{d}t \end{aligned}$$

Then, by Fubini's Theorem and Markov's Inequality,

$$\begin{aligned} \mathbb{E}\left[X\mathbb{I}_{\{X \geq \alpha\}}\right] &= \alpha \mathbb{P}\left\{X \geq \alpha\right\} + \int_\alpha^\infty \mathbb{P}\left\{X \geq t\right\} \mathrm{d}t \\ &\leq \frac{\mathbb{E}\left[X^k\right]}{\alpha^{k-1}} + \mathbb{E}\left[X^k\right] \int_\alpha^\infty t^{-k} \mathrm{d}t \\ &= \frac{\mathbb{E}\left[X^k\right]}{\alpha^{k-1}} \left(1 + \frac{1}{k-1}\right) \end{aligned}$$

■

Lemma 13 *Let the lin-growth subG condition be satisfied. Then, for any $\mathbf{x} \in \mathbb{R}^d$ and $k \in \mathbb{N}$*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|^{2k} \middle| \mathbf{x}\right] \leq 2^{k+1} k! \left(M\|\mathbf{x}\| + G\right)^{2k}$$

Proof Let $z = \|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|$ and $u = M\|\mathbf{x}\| + G$. Since z is a non-negative random variable, it follows that,

$$\begin{aligned} \mathbb{E}\left[\left(z/u\right)^{2k} \middle| \mathbf{x}\right] &= \int_0^\infty \mathbb{P}\left\{\left(z/u\right)^{2k} \geq y \middle| \mathbf{x}\right\} \mathrm{d}y \\ &= 2k \int_0^\infty v^{2k-1} \mathbb{P}\left\{z/u \geq v \middle| \mathbf{x}\right\} \mathrm{d}v \\ &\leq 4k \int_0^\infty x^{2k-1} e^{-x^2/2} \mathrm{d}x \\ &= k 2^{k+1} \int_0^\infty t^{k-1} e^{-t} \mathrm{d}t \\ &= 2^{k+1} k! \end{aligned}$$

where we use the lin-growth subG condition in step 3. Hence, we conclude that,

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|^{2k} \middle| \mathbf{x}\right] \leq 2^{k+1} k! \left(M\|\mathbf{x}\| + G\right)^{2k}$$

■

Lemma 14 (Otto-Villani Theorem Otto and Villani (2000)) *Let π^* satisfy LSI with constant λ_{LSI}. Then, π^* satisfies Talagrand's inequality T_p for any $p \in [1, 2]$, i.e., for any probability measure μ*

$$\mathcal{W}_p^2(\mu, \pi^*) \leq \frac{2}{\lambda_{\mathsf{LSI}}} \mathsf{KL}\left(\mu \middle|\middle| \pi^*\right), \ \ \forall p \in [1, 2]$$

Lemma 15 ((Guillin et al., 2009, Theorem 3.1)) *Let π^* satisfy PI with constant λ_{PI}. Then, the following holds for any probability measure μ.*

$$TV(\mu, \pi^*)^2 \leq \frac{4}{\lambda_{\mathsf{PI}}} \mathsf{FD}\left(\mu \middle|\middle| \pi^*\right)$$

Lemma 16 ((Chewi et al., 2022a, Lemma 16)) *Assume F satisfies the L-smooth condition. Then, for any arbitrary probability measure μ,*

$$\mathbb{E}_{\mathbf{x} \sim \mu}\left[\|\nabla F(\mathbf{x})\|^2\right] \leq \mathsf{FD}\left(\mu \middle|\middle| \pi^*\right) + 2dL$$

Lemma 17 (Chi Square Between Two Isotropic Gaussians) *Let p be the density of $\mathcal{N}(\mu_1, \sigma^2\mathbf{I})$ and q be the density of $\mathcal{N}(\mu_2, \sigma^2\mathbf{I})$. Then,*

$$\mathbb{E}_{x \sim q}\left[\left(\frac{p}{q}\right)^2 - 1\right] = e^{\|\mu_1 - \mu_2\|^2/\sigma^2} - 1$$

Lemma 18 (Controlling Moments under LSI) *Assume π^* satisfies LSI. Then, for any arbitrary probability measure μ and any $p \in [1, 2]$,*

$$\mathbb{E}_{\mathbf{x} \sim \mu}\left[\|\mathbf{x}\|^p\right] \leq 2^{p-1}\left(\frac{2}{\lambda_{\mathsf{LSI}}} \mathsf{KL}\left(\mu \middle|\middle| \pi^*\right)\right)^{p/2} + 2^{p-1}\mathbf{m}_p^p$$

where $\mathbf{m}_p = \mathbb{E}_{\mathbf{x} \sim \pi^*}\left[\|\mathbf{x}\|^p\right]^{1/p}$

Proof Let Γ denote the $\mathcal{W}_p$ optimal coupling between μ and π^*. Then, by Jensen's inequality

$$\begin{aligned}
\mathbb{E}_{\mathbf{x} \sim \mu}\left[\|\mathbf{x}\|^p\right] &\leq 2^{p-1}\mathbb{E}_{(\mathbf{x}, \mathbf{x}^*) \sim \Gamma}\left[\|\mathbf{x} - \mathbf{x}^*\|^p\right] + 2^{p-1}\mathbb{E}_{\mathbf{x}^* \sim \pi^*}\left[\|\mathbf{x}^*\|^p\right] \\
&\leq 2^{p-1}\mathcal{W}_p^p\left(\mu, \pi^*\right) + 2^{p-1}\mathbf{m}_p^p \\
&\leq 2^{p-1}\left(\frac{2}{\lambda_{\mathsf{LSI}}} \mathsf{KL}\left(\mu \middle|\middle| \pi^*\right)\right)^{p/2} + 2^{p-1}\mathbf{m}_p^p
\end{aligned}$$

where the last inequality follows from Lemma 14. ■

B.2. Proof of Lemma 1

Proof Consider any $k \in (K)$ and $t \in [k\eta, (k+1)\eta]$. Defining $g(\mathbf{x}_{k\eta}, \xi_k) = \nabla F(\mathbf{x}_{k\eta}) + \mathbf{N}(\mathbf{x}_{k\eta}, \xi_k)$, we note that the interpolating process is described by the following SDE for $t \in [k\eta, (k+1)\eta]$,

$$\mathrm{d}\mathbf{x}_t = -g(\mathbf{x}_{k\eta}, \xi_k)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Let $\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)$ denote the density of Law $(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)$. Then, the Fokker Planck Equation for $\mu_{t|k\eta}$ is given by,

$$\frac{\partial}{\partial t}\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k) = \nabla.(g(\mathbf{x}_{k\eta}, \xi_k)\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)) + \Delta\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)$$

Marginalizing out $\mathbf{x}_{k\eta}$ and ξ, we obtain,

$$\mu_t(\mathbf{x}_t) = \int_{\mathbb{R}^d \times \Xi^B} \mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)\mu_{k\eta}(\mathbf{x}_{k\eta})\mathrm{d}\mathbf{x}_{k\eta}\mathrm{d}\mathbb{P}^B_\xi$$

Taking the partial differential with respect to t on both sides, and interchanging the differential and the integral wherever appropriate,

$$\begin{aligned}
\frac{\partial}{\partial t}\mu_t(\mathbf{x}_t) &= \int_{\mathbb{R}^d \times \Xi^B} \left(\frac{\partial}{\partial t}\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)\right)\mu_{k\eta}(\mathbf{x}_{k\eta})\mathrm{d}\mathbf{x}_{k\eta}\mathrm{d}\mathbb{P}^B_\xi \\
&= \int_{\mathbb{R}^d \times \Xi^B} \left(\nabla.(g(\mathbf{x}_{k\eta}, \xi_k)\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k))\right)\mu_{k\eta}(\mathbf{x}_{k\eta})\mathrm{d}\mathbf{x}_{k\eta}\mathrm{d}\mathbb{P}^B_\xi \\
&+ \int_{\mathbb{R}^d \times \Xi^B} \left(\Delta\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)\right)\mu_{k\eta}(\mathbf{x}_{k\eta})\mathrm{d}\mathbf{x}_{k\eta}\mathrm{d}\mathbb{P}^B_\xi \\
&= \nabla.\left(\int_{\mathbb{R}^d \times \Xi^B} g(\mathbf{x}_{k\eta}, \xi_k)\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)\mu_{k\eta}(\mathbf{x}_{k\eta})\mathrm{d}\mathbf{x}_{k\eta}\mathrm{d}\mathbb{P}^B_\xi\right) \\
&+ \Delta\left(\int_{\mathbb{R}^d \times \Xi^B} \mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)\mu_{k\eta}(\mathbf{x}_{k\eta})\mathrm{d}\mathbf{x}_{k\eta}\mathrm{d}\mathbb{P}^B_\xi\right) \\
&= \nabla.\left(\int_{\mathbb{R}^d \times \Xi^B} g(\mathbf{x}_{k\eta}, \xi_k)\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)\mu_{k\eta}(\mathbf{x}_{k\eta})\mathrm{d}\mathbf{x}_{k\eta}\mathrm{d}\mathbb{P}^B_\xi\right) + \Delta\mu_t(\mathbf{x}_t)
\end{aligned}$$

We note that, by disintegration theorem,

$$\begin{aligned}
\int_{\mathbb{R}^d \times \Xi^B} g(\mathbf{x}_{k\eta}, \xi_k)\mu_{t|k\eta}(\mathbf{x}_t|\mathbf{x}_{k\eta}, \xi_k)\mu_{k\eta}(\mathbf{x}_{k\eta})\mathrm{d}\mathbf{x}_{k\eta}\mathrm{d}\mathbb{P}^B_\xi &= \mu_t(\mathbf{x}_t)\int_{\mathbb{R}^d \times \Xi^B} g(\mathbf{x}_{k\eta}, \xi_k)\mathrm{d}\mathbb{P}_{\mathbf{x}_{k\eta}, \xi_k|\mathbf{x}_t}(\mathbf{x}_{k\eta}, \xi_k) \\
&= \mu_t(\mathbf{x}_t)\mathbb{E}\left[g(\mathbf{x}_{k\eta}, \xi_k)|\mathbf{x}_t\right]
\end{aligned}$$

Hence, we obtain the following continuity equation for $\mu_t(\mathbf{x}_t)$

$$\frac{\partial}{\partial t}\mu_t(\mathbf{x}_t) = \nabla.\left(\mu_t(\mathbf{x}_t)\mathbb{E}\left[g(\mathbf{x}_{k\eta}, \xi_k)|\mathbf{x}_t\right]\right) + \Delta\mu_t(\mathbf{x}_t) \qquad (2)$$

We now analyze the time-evolution of $\mathsf{KL}\left(\mu_t \middle|\middle| \pi^*\right)$ using equation (2). Interchanging differentials and integrals wherever appropriate,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t \middle|\middle| \pi^*\right) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} \mu_t(\mathbf{x}) \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{R}^d} \left(\frac{\partial}{\partial t}\mu_t(\mathbf{x})\right) \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) \mathrm{d}\mathbf{x} + \int_{\mathbb{R}^d} \frac{\partial}{\partial t}\mu_t(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{R}^d} \left(\frac{\partial}{\partial t}\mu_t(\mathbf{x})\right) \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) \mathrm{d}\mathbf{x} + \frac{\partial}{\partial t}\int_{\mathbb{R}^d} \mu_t(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{R}^d} \nabla.\left(\mu_t(\mathbf{x})\mathbb{E}\left[g(\mathbf{x}_{k\eta}, \xi_k)|\mathbf{x}_t = \mathbf{x}\right]\right) \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \\
&+ \int_{\mathbb{R}^d} \Delta\mu_t(\mathbf{x}) \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \qquad (3)
\end{aligned}$$

where the last step uses the fact that $\frac{\partial}{\partial t}\int_{\mathbb{R}^d} \mu_t(\mathbf{x})\mathrm{d}\mathbf{x} = 0$ since $\int_{\mathbb{R}^d} \mu_t(\mathbf{x})\mathrm{d}\mathbf{x} = 1$ for any $t \in [k\eta, (k+1)\eta]$. We now note that,

$$\mu_t(\mathbf{x})\nabla \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) = \nabla\mu_t(\mathbf{x}) + \mu_t(\mathbf{x})\nabla F(\mathbf{x})$$

Taking the divergence on both sides, we obtain the following identity,

$$\Delta\mu_t(\mathbf{x}) = \nabla.\left(\mu_t(\mathbf{x})\left(\nabla \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) - \nabla F(\mathbf{x})\right)\right)$$

It follows that,

$$\begin{aligned}
\int_{\mathbb{R}^d} \Delta\mu_t(\mathbf{x}) \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) \mathrm{d}\mathbf{x} &= \int_{\mathbb{R}^d} \nabla.\left(\mu_t(\mathbf{x})\left(\nabla \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) - \nabla F(\mathbf{x})\right)\right) \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{R}^d} \mu_t(\mathbf{x}) \left\langle \nabla F(\mathbf{x}) - \nabla \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right), \nabla \log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right)\right\rangle \mathrm{d}\mathbf{x} \\
&= \mathbb{E}\left[\left\langle \nabla F(\mathbf{x}_t), \nabla \log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] - \mathbb{E}\left[\left\|\nabla \log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\|^2\right] \\
&= -\mathsf{FD}\left(\mu_t \middle|\middle| \pi^*\right) + \mathbb{E}\left[\left\langle \nabla F(\mathbf{x}_t), \nabla \log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] \qquad (4)
\end{aligned}$$

where the second equality follows from integration by parts on $\mathbb{R}^d$, the third follows from the fact that $\mu_t = \mathrm{Law}\left(\mathbf{x}_t\right)$ and the last follows from the definition of Fisher Divergence. Following similar steps, we obtain,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \nabla.\left(\mu_t(\mathbf{x})\mathbb{E}\left[g(\mathbf{x}_{k\eta},\xi_k)|\mathbf{x}_t=\mathbf{x}\right]\right)\log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right)\mathrm{d}\mathbf{x} \\
&= -\int_{\mathbb{R}^d} \mu_t(\mathbf{x})\left\langle \mathbb{E}\left[g(\mathbf{x}_{k\eta},\xi_k)|\mathbf{x}_t=\mathbf{x}\right], \nabla\log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right)\right\rangle \mathrm{d}\mathbf{x} \\
&= -\mathbb{E}\left[\left\langle \mathbb{E}\left[g(\mathbf{x}_{k\eta},\xi_k)|\mathbf{x}_t\right], \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] \\
&= -\mathbb{E}\left[\left\langle g(\mathbf{x}_{k\eta},\xi_k), \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] \\
&= -\mathbb{E}\left[\left\langle \nabla F(\mathbf{x}_{k\eta}) + \mathbf{N}, \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] \qquad (5)
\end{aligned}$$

where $\mathbf{N}$ stands for $\mathbf{N}(\mathbf{x}_{k\eta},\xi_k)$. We note that,

$$\mathbb{E}\left[\left\langle \mathbf{N}, \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] = \mathbb{E}\left[\left\langle \mathbb{E}\left[\mathbf{N}|\mathbf{x}_t,\mathbf{x}_{k\eta}\right], \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right]$$

Substituting the above into equation (5), we obtain,

$$\begin{aligned}
\int_{\mathbb{R}^d} \nabla.\left(\mu_t(\mathbf{x})\mathbb{E}\left[g(\mathbf{x}_{k\eta},\xi_k)|\mathbf{x}_t=\mathbf{x}\right]\right)\log\left(\frac{\mu_t(\mathbf{x})}{\pi^*(\mathbf{x})}\right)\mathrm{d}\mathbf{x} &= -\mathbb{E}\left[\left\langle \nabla F(\mathbf{x}_{k\eta}), \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] \\
&\quad - \mathbb{E}\left[\left\langle \mathbb{E}\left[\mathbf{N}|\mathbf{x}_t,\mathbf{x}_{k\eta}\right], \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right]
\end{aligned} \qquad (6)$$

From equations (3), (4) and (6), we obtain,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t||\pi^*\right) &= -\mathsf{FD}\left(\mu_t||\pi^*\right) + \mathbb{E}\left[\left\langle \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta}), \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] \\
&\quad - \mathbb{E}\left[\left\langle \mathbb{E}\left[\mathbf{N}|\mathbf{x}_t,\mathbf{x}_{k\eta}\right], \nabla\log\left(\frac{\mu_t(\mathbf{x}_t)}{\pi^*(\mathbf{x}_t)}\right)\right\rangle\right] \\
&\leq -\frac{1}{2}\mathsf{FD}\left(\mu_t||\pi^*\right) + \mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] + \mathbb{E}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t]\|^2\right]
\end{aligned}$$

■

B.3. Proof of Lemma 2

Proof Define $u_k = M\|\mathbf{x}_{k\eta}\| + G$ and $h = t - k\eta$. Denoting $\mathbf{N} = {}^1\!/\!{}_B\sum_{j=1}^{B}\mathbf{N}_j$, where $\mathbf{N}_j = \nabla f(\mathbf{x}_{k\eta},\xi_{k,j}) - \nabla F(\mathbf{x}_{k\eta})$, we conclude from lin-growth subG and Lemma 13 that $\mathbb{E}\left[\|\mathbf{N}_j\|^{2m}|\mathbf{x}_{k\eta}\right] \leq 2^{m+1}m!u_k^{2m}$ for any $j \in [B]$ and $m \in \mathbb{N}$.

Our proof begins by obtaining a coarse bound for $\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t]\|^2\right]$. Using the conditional CLT structure of $\mathbf{N}$ (i.e., using the fact that $\mathbf{N}$ is an empirical average of zero-mean random variables that are i.i.d conditioned on $\mathbf{x}_{k\eta}$), we obtain the following via Jensen's inequality.

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t]\|^2\right] &\leq \mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\mathbb{E}[\|\mathbf{N}\|^2|\mathbf{x}_{k\eta},\mathbf{x}_t]\right] = \mathbb{E}\left[\|\mathbf{N}\|^2|\mathbf{x}_{k\eta}\right] \\ &= \tfrac{1}{B}\mathbb{E}\left[\|\mathbf{N}_1\|^2|\mathbf{x}_{k\eta}\right] \leq {}^{4u_k^2}/_{B} \end{aligned} \tag{7}$$

We now proceed to refine this bound as follows. Note that, since $\mathbf{N}_1,\ldots,\mathbf{N}_B$ are i.i.d conditioned on $\mathbf{x}_{k\eta}$ and $\text{Law}\left(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\ldots,\mathbf{N}_B\right) = \mathcal{N}(\mathbf{x}_{k\eta} - h\nabla F(\mathbf{x}_{k\eta}) - {}^{h}/_{B}\sum_{j=1}^{B}\mathbf{N}_j, 2h)$, it is easy to see that they remain i.i.d even after conditioning on $\mathbf{x}_t$, due to the permutation invariance of the function $\sum_{j=1}^{B}\mathbf{N}_j$. Hence $\mathbb{E}\left[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t\right] \overset{d}{=} \mathbb{E}\left[\mathbf{N}_1|\mathbf{x}_{k\eta},\mathbf{x}_t\right]$ by linearity of conditional expectation. It follows by Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t]\|^2\right] &= \mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\|\mathbb{E}\left[\mathbf{N}_1|\mathbf{x}_{k\eta},\mathbf{x}_t\right]\|^2\right] \\ &= \mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\left\|\mathbb{E}_{\mathbf{N}_{2:B}|\mathbf{x}_{k\eta}}\left[\mathbb{E}\left[\mathbf{N}_1|\mathbf{x}_{k\eta},\mathbf{N}_{2:B},\mathbf{x}_t\right]\right]\right\|^2\right] \\ &\leq \mathbb{E}_{\mathbf{N}_{2:B},\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\|\mathbb{E}\left[\mathbf{N}_1|\mathbf{x}_{k\eta},\mathbf{N}_{2:B},\mathbf{x}_t\right]\|^2\right] \end{aligned} \tag{8}$$

Let $\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})$ denote the density of $\text{Law}\left(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B}\right)$ and $\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})$ denote the density of $\text{Law}\left(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B}\right)$.

$$\begin{aligned}\mathbb{E}\left[\mathbf{N}_1|\mathbf{x}_{k\eta},\mathbf{N}_{2:B},\mathbf{x}_t\right] &= \int_{\mathbb{R}^d}\mathbf{N}_1 \mathrm{d}\mathbb{P}_{\mathbf{N}_1|\mathbf{x}_{k\eta},\mathbf{N}_{2:B},\mathbf{x}_t}(\mathbf{N}_1) \\ &= \int_{\mathbb{R}^d}\mathbf{N}_1\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})}\mathrm{d}P_{\mathbf{N}_1|\mathbf{N}_{2:B},\mathbf{x}_{k\eta}}(\mathbf{N}_1) \\ &= \int_{\mathbb{R}^d}\mathbf{N}_1\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})}\mathrm{d}P_{\mathbf{N}_1|\mathbf{x}_{k\eta}}(\mathbf{N}_1) \\ &= \int_{\mathbb{R}^d}\mathbf{N}_1\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})} - 1\right)\mathrm{d}P_{\mathbf{N}_1|\mathbf{x}_{k\eta}}(\mathbf{N}_1) \\ &= \mathbb{E}_{\mathbf{N}_1|\mathbf{x}_{k\eta}}\left[\mathbf{N}_1\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})} - 1\right)\right]\end{aligned}$$

where the third equality uses the fact that $\mathbf{N}_1,\ldots,\mathbf{N}_B$ are i.i.d conditioned on $\mathbf{x}_{k\eta}$ and the fourth equality uses the fact that $\mathbf{N}_1$ is zero-mean conditioned on $\mathbf{x}_{k\eta}$. Now, applying Cauchy-Schwarz inequality,

$$\begin{aligned}\|\mathbb{E}\left[\mathbf{N}_1|\mathbf{x}_{k\eta},\mathbf{N}_{2:B},\mathbf{x}_t\right]\|^2 &\leq \mathbb{E}\left[\|\mathbf{N}_1\|^2|\mathbf{x}_{k\eta}\right]\mathbb{E}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})} - 1\right)^2\right] \\ &\leq 4u_k^2\mathbb{E}_{\mathbf{N}_1|\mathbf{x}_{k\eta}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})}\right)^2 - 1\right]\end{aligned} \tag{9}$$

Let $\tilde{\mathbf{N}}_1$ be an identical and independent copy of $\mathbf{N}_1$ conditioned on $\mathbf{x}_{k\eta}$. It follows that,

$$\mathbb{E}_{\mathbf{N}_1|\mathbf{x}_{k\eta}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})}\right)^2-1\right]=\mathbb{E}_{\mathbf{N}_1|\mathbf{x}_{k\eta}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\mathbb{E}_{\tilde{\mathbf{N}}_1|\mathbf{x}_{k\eta}}\left[\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B})\right]}\right)^2-1\right]$$
$$\leq\mathbb{E}_{\mathbf{N}_1,\tilde{\mathbf{N}}_1|\mathbf{x}_{k\eta}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B})}\right)^2-1\right]$$

where the last step is an application of Jensen's inequality and the convexity of $1/x^2$. Taking expectations wrt $\mathbf{N}_{2:B},\mathbf{x}_t|\mathbf{x}_{k\eta}$ on both sides,

$$\mathbb{E}_{\mathbf{N}_1,\mathbf{N}_{2:B},\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})}\right)^2-1\right]\leq\mathbb{E}_{\mathbf{N}_1,\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B},\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B})}\right)^2-1\right]$$
$$=\mathbb{E}_{\mathbf{N}_1,\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B}|\mathbf{x}_{k\eta}}\left[\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta},\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B})}\right)^2-1\right]\right]$$

Since $\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})=\mathcal{N}(\mathbf{x}_{k\eta}-h\nabla F(\mathbf{x}_{k\eta})-{}^{h}\!/_{B}\mathbf{N}_1-{}^{h}\!/_{B}\sum_{j=2}^{B}\mathbf{N}_j,2h)$, we conclude from Lemma 17 that,

$$\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta},\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B})}\right)^2-1\right]=e^{\frac{h^2\|\mathbf{N}_1-\tilde{\mathbf{N}}_1\|^2}{2B^2}}-1$$

Furthermore, by Jensen's inequality and Lemma 13,

$$\mathbb{E}_{\mathbf{N}_1,\tilde{\mathbf{N}}_1|\mathbf{x}_{k\eta}}\left[\left\|\mathbf{N}_1-\tilde{\mathbf{N}}_1\right\|^{2m}\right]\leq 2^{2m}\mathbb{E}\left[\|\mathbf{N}_1\|^{2m}|\mathbf{x}_{k\eta}\right]\leq 2^{3m+1}m!\;u_k^{2m}\;\;\forall\,m\in\mathbb{N}$$

Hence, the exponential can be controlled as follows

$$\mathbb{E}_{\mathbf{N}_1,\tilde{\mathbf{N}}_1,\mathbf{N}_{2:B}|\mathbf{x}_{k\eta}}[e^{\frac{h\|\mathbf{N}_1-\tilde{\mathbf{N}}_1\|^2}{2B^2}}-1]=\mathbb{E}_{\mathbf{N}_1,\tilde{\mathbf{N}}_1|\mathbf{x}_{k\eta}}[e^{\frac{h\|\mathbf{N}_1-\tilde{\mathbf{N}}_1\|^2}{2B^2}}-1]$$
$$=\sum_{m=1}^{\infty}\frac{h^m}{2^mB^{2m}m!}\mathbb{E}_{\mathbf{N}_1,\tilde{\mathbf{N}}_1|\mathbf{x}_{k\eta}}\left[\left\|\mathbf{N}_1-\tilde{\mathbf{N}}_1\right\|^{2m}\right]$$
$$\leq\sum_{m=1}^{\infty}\frac{h^m2^{2m+1}u_k^{2m}}{B^{2m}}$$
$$\leq\frac{8hu_k^2}{B^2}\sum_{m=0}^{\infty}\left(\tfrac{4hu_k^2}{B^2}\right)^m$$

It follows that,

$$\mathbb{E}_{\mathbf{N}_1,\mathbf{N}_{2:B},\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\left(\frac{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_1,\mathbf{N}_{2:B})}{\rho(\mathbf{x}_t|\mathbf{x}_{k\eta},\mathbf{N}_{2:B})}\right)^2-1\right]\leq\frac{8hu_k^2}{B^2}\sum_{m=0}^{\infty}\left(\tfrac{4hu_k^2}{B^2}\right)^m \tag{10}$$

From equations (8), (9) and (10), we obtain the following finer bound,

$$\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t]\|^2\right] \leq \frac{8hu_k^4}{B^2}\sum_{m=0}^{\infty}\left(\tfrac{4hu_k^2}{B^2}\right)^m \tag{11}$$

Thus, from equations (7) and (11), it follows that,

$$\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t]\|^2\right] \leq \min\left\{\frac{8hu_k^4}{B^2}\sum_{m=0}^{\infty}\left(\tfrac{4hu_k^2}{B^2}\right)^m, \frac{4u_k^2}{B}\right\}$$

Note that, when $u_k^2 \leq {}^{B}/{}_{6h}$, the following holds,

$$\frac{8hu_k^4}{B^2}\sum_{m=0}^{\infty}\left(\tfrac{4hu_k^2}{B^2}\right)^m \leq \frac{8hu_k^4}{B^2}\sum_{m=0}^{\infty}({}^{2}/{}_{3})^m \leq \frac{24hu_k^4}{B^2} \leq \frac{4u_k^2}{B}$$

Hence, the coarse bound and fine bound are combined as,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_{k\eta}}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t]\|^2\right] &\leq \frac{4u_k^2}{B}\mathbb{I}_{\left\{u_k^2 > {}^{B}/{}_{6h}\right\}} + \left[\frac{8hu_k^4}{B^2}\sum_{m=0}^{\infty}\left(\tfrac{4hu_k^2}{B^2}\right)^m\right]\mathbb{I}_{\left\{u_k^2 \leq {}^{B}/{}_{6h}\right\}} \\ &\leq \frac{24hu_k^4}{B^2} + \frac{4u_k^2}{B}\mathbb{I}_{\left\{u_k^2 > {}^{B}/{}_{6h}\right\}}\end{aligned}$$

Since the p-moment growth condition holds with $p = 4$, and Law $(\mathbf{x}_{k\eta}) =$ Law $(\hat{\mathbf{x}}_k)$ by construction of the interpolating process,

$$\mathbb{E}_{\mathbf{x}_{k\eta}}\left[u_k^4\right] \leq 8\left(M^4\mathbb{E}\left[\|\hat{\mathbf{x}}_k\|^4\right] + G^4\right) \leq 8\left(M^4C_4d^2 + G^4\right)$$

Furthermore, applying Lemma 12,

$$\mathbb{E}\left[u_k^2\mathbb{I}_{\left\{u_k^2 > {}^{B}/{}_{6h}\right\}}\right] \leq \frac{12h}{B}\mathbb{E}[u_k^4]$$

Hence, it follows that,

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta},\mathbf{x}_t]\|^2\right] &\leq \frac{24h}{B^2}\mathbb{E}\left[u_k^4\right] + \mathbb{E}\left[\frac{4u_k^2}{B}\mathbb{I}_{\left\{u_k^2 > {}^{B}/{}_{6h}\right\}}\right] \\ &\leq \frac{72h}{B^2}\mathbb{E}\left[u_k^4\right] \\ &\leq \frac{576(t-k\eta)(M^4C_4d^2+G^4)}{B^2}\end{aligned}$$

■

B.4. Proof of Lemma 3

Proof From Lemmas 1 and 2, we infer that,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t||\pi^*\right) \leq -\frac{1}{2}\mathsf{FD}\left(\mu_t||\pi^*\right) + \mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] + \frac{576(t-k\eta)(M^4C_4d^2+G^4)}{B^2} \tag{12}$$

The remainder of this proof controls the discretization error term $\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right]$. To this end, let $h = t - k\eta$. We note that, $0 \leq h \leq \eta \leq 1/6L$. Furthermore, via direct integration of the interpolating process, we infer,

$$\mathbf{x}_t \overset{d}{=} \mathbf{x}_{k\eta} - h(\nabla F(\mathbf{x}_{k\eta}) + \mathbf{N}) + \sqrt{2h}\mathbf{z}_t, \ \mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$$

Hence, by the L-smoothness of F,

$$\begin{aligned}\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] &\leq L^2\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{k\eta}\|^2\right] \\ &\leq L^2h^2\mathbb{E}\left[\|\nabla F(\mathbf{x}_{k\eta})\|^2\right] + L^2h^2\mathbb{E}\left[\|\mathbf{N}\|^2\right] + 2L^2hd\end{aligned}$$

We now rewrite the RHS in terms of $\mathbf{x}_t$. To this end, we note that by the L-lipschitzness of ∇F,

$$\begin{aligned}\|\nabla F(\mathbf{x}_{k\eta})\| &\leq \|\nabla F(\mathbf{x}_t)\| + L\|\mathbf{x}_t - \mathbf{x}_{k\eta}\| \\ &\leq \|\nabla F(\mathbf{x}_t)\| + Lh\|\nabla F(\mathbf{x}_{k\eta})\| + Lh\|\mathbf{N}\| + L\sqrt{2h}\|\mathbf{z}_t\|\end{aligned} \tag{13}$$

Using the fact that $h \leq \eta < 1/3L$ and rearranging,

$$\begin{aligned}\|\nabla F(\mathbf{x}_{k\eta})\| &\leq \frac{3}{2}\|\nabla F(\mathbf{x}_t)\| + \frac{1}{2}\|\mathbf{N}\| + \frac{3L\sqrt{h}}{\sqrt{2}}\|\mathbf{z}_t\| \\ \|\nabla F(\mathbf{x}_{k\eta})\|^2 &\leq \frac{27}{4}\|\nabla F(\mathbf{x}_t)\|^2 + \frac{3}{4}\|\mathbf{N}\|^2 + \frac{27L^2h}{2}\|\mathbf{z}_t\|^2\end{aligned}$$

Substituting the above into (13), we get,

$$\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] \leq \frac{27L^2h^2}{4}\mathbb{E}\left[\|\nabla F(\mathbf{x}_t)\|^2\right] + 2L^2h^2\mathbb{E}\left[\|\mathbf{N}\|^2\right] + \frac{27L^4h^3d}{2} + 2L^2hd$$

Using the fact that $\mathbb{E}\left[\|\nabla F(\mathbf{x}_t)\|^2\right] \leq \mathsf{FD}\left(\mu_t||\pi^*\right) + 2dL$ as per Lemma 16, and $h \leq 1/6L$, we get

$$\begin{aligned}\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] &\leq \frac{27L^2h^2}{4}\mathsf{FD}\left(\rho_t||\gamma\right) + 2L^2h^2\mathbb{E}\left[\|\mathbf{N}\|^2\right] + 5L^2hd \\ &\leq \frac{1}{4}\mathsf{FD}\left(\mu_t||\pi^*\right) + 2L^2h^2\mathbb{E}\left[\|\mathbf{N}\|^2\right] + 5L^2hd\end{aligned}$$

Recall that, due to the lin-growth subG and p-moment growth conditions, the following holds as a consequence of Lemma 13.

$$\mathbb{E}\left[\|\mathbf{N}\|^2\right] \leq \frac{4}{B}\mathbb{E}\left[(M\|\hat{\mathbf{x}}_k\| + G)^2\right] \leq \frac{8}{B}\left(M^2C_2d + G^2\right)$$

Hence,

$$\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] \leq \frac{1}{4}\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right) + \frac{3Lh}{B}\left(M^2 C_2 d + G^2\right) + 5L^2 hd$$

Substituting the above into (12),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right) \leq -\frac{1}{4}\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right) + 5L^2(t-k\eta)d + \frac{3L(t-k\eta)(M^2C_2d+G^2)}{B} \\ + \frac{576(t-k\eta)(M^4C_4d^2+G^4)}{B^2}\end{aligned}$$

$\blacksquare$

B.5. Proof of Theorem 5

Proof Since π^* satisfies LSI, we know that $\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right) \leq \frac{1}{2\lambda_{\mathsf{LSI}}}\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right)$. Thus, from Lemma 3, we conclude,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right) \leq -\frac{\lambda_{\mathsf{LSI}}}{2}\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right) + 5L^2(t-k\eta)d + \frac{3L(t-k\eta)(M^2C_2d+G^2)}{B} \\ + \frac{576(t-k\eta)(M^4C_4d^2+G^4)}{B^2}\end{aligned}$$

Multiplying both sides by $e^{\lambda_{\mathsf{LSI}}(t-k\eta)/2}$ and using the fact that $k\eta \leq t \leq (k+1)\eta$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\lambda_{\mathsf{LSI}}(t-k\eta)/2}\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right)\right) \leq e^{\lambda_{\mathsf{LSI}}(t-k\eta)/2}\left[5L^2\eta d + \frac{3L\eta(M^2C_2d+G^2)}{B} + \frac{576\eta(M^4C_4d^2+G^4)}{B^2}\right]$$

Applying Grönwall's Lemma for $t \in [k\eta, (k+1)\eta]$

$$\begin{aligned}&e^{\lambda_{\mathsf{LSI}}\eta/2}\mathsf{KL}\left(\mu_{(k+1)\eta}\middle|\middle|\pi^*\right) - \mathsf{KL}\left(\mu_{k\eta}\middle|\middle|\pi^*\right) \\ &\leq \frac{2}{\lambda_{\mathsf{LSI}}}\left(e^{\lambda_{\mathsf{LSI}}\eta/2} - 1\right)\left[5L^2\eta d + \frac{3L\eta(M^2C_2d+G^2)}{B} + \frac{576\eta(M^4C_4d^2+G^4)}{B^2}\right] \\ &\leq 5.25L^2\eta^2 d + \frac{3.15L\eta^2(M^2C_2d+G^2)}{B} + \frac{604.8\eta^2(M^4C_4d^2+G^4)}{B^2}\end{aligned}$$

Where the last inequality uses the fact that $\lambda_{\mathsf{LSI}}\eta \leq \nicefrac{\lambda_{\mathsf{LSI}}^2}{6L^2} \leq \nicefrac{1}{6}$, and $e^{\nicefrac{x}{2}} - 1 \leq 0.525x$ for $x \leq \nicefrac{1}{6}$. Furthermore, recalling the fact that $\mu_{k\eta} = \mathrm{Law}\left(\hat{\mathbf{x}}_k\right)$, we obtain the following descent lemma for the discrete-time iterates of SGLD,

$$\begin{aligned}\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_{k+1}\right)\middle|\middle|\pi^*\right) \leq e^{-\lambda_{\mathsf{LSI}}\eta/2}\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_{k+1}\right)\middle|\middle|\pi^*\right) \\ + e^{-\lambda_{\mathsf{LSI}}\eta/2}\left[5.25L^2\eta^2 d + \frac{3.15L\eta^2(M^2C_2d+G^2)}{B} + \frac{604.8\eta^2(M^4C_4d^2+G^4)}{B^2}\right]\end{aligned}$$

Iterating through the above lemma, we get,

$$\begin{aligned}\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_K\right)\middle|\middle|\pi^*\right) \leq e^{-\lambda_{\mathsf{LSI}}\eta K/2}\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right) \\ + \frac{e^{-\lambda_{\mathsf{LSI}}\eta/2}}{1 - e^{-\lambda_{\mathsf{LSI}}\eta/2}}\left[5.25L^2\eta^2 d + \frac{3.15L\eta^2(M^2C_2d+G^2)}{B} + \frac{604.8\eta^2(M^4C_4d^2+G^4)}{B^2}\right]\end{aligned}$$

Using the fact that $e^x \geq 1 + x$, we obtain,

$$\mathsf{KL}\left(\text{Law}\left(\hat{\mathbf{x}}_K\right)\middle|\middle|\pi^*\right) \leq e^{-\lambda_{\mathsf{LSI}}\eta K/2}\mathsf{KL}\left(\text{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right) + \frac{11L^2\eta d}{\lambda_{\mathsf{LSI}}} + \frac{7L\eta(M^2C_2d + G^2)}{B\lambda_{\mathsf{LSI}}} + \frac{1210\eta(M^4C_4d^2 + G^4)}{\lambda_{\mathsf{LSI}}B^2}$$

■

B.6. Analysis of AB-SGLD

Recall that F has a finite-sum structure of the form $F(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})$. Similar to our analysis of smooth SGLD, we construct the following interpolating process $(\mathbf{x}_t)_{t\in[0,K\eta]}$

$$\begin{aligned}\text{Law}\left(\mathbf{x}_0\right) = \text{Law}\left(\hat{\mathbf{x}}_0\right) \quad \xi_{k,1}, \ldots, \xi_{k,n} &\overset{\text{iid}}{\sim} \mathsf{Unif}[n] \;\forall k \in [K] \\ \mathrm{d}\mathbf{x}_t = -\left[\nabla F(\mathbf{x}_{k\eta}) + \mathbf{N}(\mathbf{x}_{k\eta}, \xi_k)\right]\mathrm{d}t &+ \sqrt{2}\mathrm{d}B_t, \; t \in [k\eta, (k+1)\eta], \; k \in [K]\end{aligned}$$

where B_t is a Brownian motion independent of $\mathbf{x}_0$ and $\mathbf{N}(\mathbf{x}_{k\eta}, \xi_k)$ is defined as,

$$\mathbf{N}(\mathbf{x}_{k\eta}, \xi_k) = \frac{1}{B_k}\sum_{j=1}^{B_k} f_{\xi_{k,j}}(\hat{\mathbf{x}}_k) - \nabla F(\hat{\mathbf{x}}_k)$$
$$B_k = \min\{n, 1 + \lceil M\|\hat{\mathbf{x}}_k\| + G\rceil\}$$

Since $\text{Law}\left(\mathbf{x}_0\right) = \text{Law}\left(\hat{\mathbf{x}}_0\right)$, and B_k is a deterministic function of $\mathbf{x}_{k\eta}$, writing a closed form for $\text{Law}\left(\mathbf{x}_t|\mathbf{x}_{k\eta}\right)$ and performing an inductive argument shows that $\text{Law}\left(\mathbf{x}_{k\eta}\right) = \text{Law}\left(\hat{\mathbf{x}}_k\right)$, i.e., the above stochastic process is an interpolating process for AB-SGLD. As before, let $\mu_t = \text{Law}\left(\mathbf{x}_t\right)$. Our first step is to analyze the time-evolution of $\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right)$.

Lemma 19 (ABSGLD : Flow of KL along Interpolating Process) *Assume* $\mathsf{KL}\left(\mu_0\middle|\middle|\pi^*\right) < \infty$. *Then, the following differential inequality is satisfied for any* $k \in (K)$ *and* $t \in [k\eta, (k+1)\eta]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right) \leq -\frac{1}{2}\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right) + \mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] + \mathbb{E}\left[\|\mathbb{E}[\mathbf{N}|\mathbf{x}_{k\eta}, \mathbf{x}_t]\|^2\right]$$

Proof Since B_k is a deterministic function of $\mathbf{x}_{k\eta}$, this lemma can be considered a special case of Lemma 1 with $\Xi = [n]$ and $P_\xi = \mathsf{Unif}[n]$. The proof is completed by applying the same arguments.
■

We shall now use the adaptive batch-size property to control the stochastic gradient error term. Our analysis crucially uses smoothness and LSI.

Lemma 20 (AB-SGLD: Controlling Stochastic Gradient Error) *Let the lin-growth a.s condition be satisfied and let* π^* *satisfy LSI. Then, the following holds for any* $t \in [k\eta, (k+1)\eta], \; k \in (K), \; \eta \leq 1$.

$$\mathbb{E}\left[\|\mathbb{E}\left[\mathbf{N} \mid \mathbf{x}_t, \mathbf{x}_{k\eta}\right]\|^2\right] \leq \frac{128(t - k\eta)M^2}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta}\middle|\middle|\pi^*\right) + 32(t - k\eta)\left(2M^2\mathbf{m}_2^2 + G^2\right)$$

where $\mathbf{m}_2^2 = \mathbb{E}_{\pi^*}\left[\|\mathbf{x}\|^2\right]$

Proof Let $u_k = M\|\mathbf{x}_{k\eta}\| + G$ and $h = t - k\eta$. Denoting $\mathbf{N} = {}^1\!/_B \sum_{j=1}^{B_k} \mathbf{N}_j$, where $\mathbf{N}_j = \nabla f(\mathbf{x}_{k\eta}, \xi_{k,j}) - \nabla F(\mathbf{x}_{k\eta})$, we conclude from lin-growth subG and Lemma 13 that $\mathbb{E}\left[\|\mathbf{N}_j\|^{2m} | \mathbf{x}_{k\eta}\right] \leq 2^{m+1} m! u_k^{2m}$ for any $j \in [B]$ and $m \in \mathbb{N}$. Furthermore, since B_k is a deterministic function of $\mathbf{x}_{k\eta}$ and $\mathbf{N}_1, \dots, \mathbf{N}_{B_k}$ are zero-mean i.i.d conditioned on $\mathbf{x}_{k\eta}$, we can repeat the same arguments as Lemma 2 to obtain the following bound,

$$\mathbb{E}_{\mathbf{x}_t \mid \mathbf{x}_{k\eta}}\left[\|\mathbb{E}\left[\mathbf{N} \mid \mathbf{x}_t, \mathbf{x}_{k\eta}\right]\|^2\right] \leq \frac{8hu_k^4}{B_k^2} \sum_{m=0}^{\infty} \left(\tfrac{4hu_k^2}{B_k^2}\right)^m \leq 16hu_k^2 \leq 32h\left(M^2\|\mathbf{x}_{k\eta}\|^2 + G^2\right)$$

where the second inequality follows from the fact that $B_k \geq u_k$ and $h \leq \eta \leq {}^1\!/_8$. Now, applying Lemma 18, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{E}\left[\mathbf{N} \mid \mathbf{x}_t, \mathbf{x}_{k\eta}\right]\|^2\right] &\leq 32h\left(M^2\mathbb{E}\left[\|\mathbf{x}_{k\eta}\|^2\right] + G^2\right) \\ &\leq \frac{128M^2h}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta}||\pi^*\right) + 32h\left(2M^2\mathbf{m}_2^2 + G^2\right)\end{aligned}$$

■

B.6.1. Proof of Theorem 6

Proof Consider any $k \in (K)$ and $t \in [k\eta, (k+1)\eta]$. Furthermore, let $h = t - k\eta \leq \eta \leq {}^1\!/_{6L}$. From Lemma 19 and Lemma 20, it follows that,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t||\pi^*\right) \leq -\frac{1}{2}\mathsf{FD}\left(\mu_t||\pi^*\right) &+ \mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] \\ &+ \frac{128M^2h}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta}||\pi^*\right) + 32h\left(2M^2\mathbf{m}_2^2 + G^2\right)\end{aligned}$$

To control $\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right]$, we recall that $h \leq {}^1\!/_{6L}$ and follow the same steps as Lemma 3 to obtain,

$$\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] \leq \frac{1}{4}\mathsf{FD}\left(\mu_t||\pi^*\right) + 2L^2h^2\mathbb{E}\left[\|\mathbf{N}\|^2\right] + 5L^2hd$$

We now control $\mathbb{E}\left[\|\mathbf{N}\|^2\right]$. From Lemma 13, we obtain the following.

$$\mathbb{E}\left[\|\mathbf{N}\|^2 | \mathbf{x}_{k\eta}\right] \leq \frac{4(M\|\mathbf{x}_{k\eta}\| + G)^2}{B_k} \leq 4M\|\mathbf{x}_{k\eta}\| + 4G$$

Now, applying Lemma 18,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{N}\|^2\right] &\leq 4M\sqrt{\frac{2}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta}||\pi^*\right)} + 4M\mathbf{m}_1 + 4G \\ &\leq 4M + \frac{2M}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta}||\pi^*\right) + 4M\mathbf{m}_1 + 4G\end{aligned}$$

where the last inequality uses $\sqrt{x} \leq 1 + {}^{x}/_{4}$ for any $x \geq 0$. Hence,

$$\begin{aligned}\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{k\eta})\|^2\right] \leq \frac{1}{4}\mathsf{FD}\left(\mu_t \middle\| \pi^*\right) + \frac{4L^2h^2M}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta} \middle\| \pi^*\right) \\ + 5L^2hd + 8L^2Mh^2 + 8L^2Mh^2\mathbf{m}_1 + 8L^2h^2G\end{aligned}$$

Using $h \leq \eta \leq {}^{1}/_{6L}$, it follows that,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t \middle\| \pi^*\right) &\leq -\frac{1}{4}\mathsf{FD}\left(\mu_t \middle\| \pi^*\right) + \frac{(LM + 128M^2)h}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta} \middle\| \pi^*\right) \\ &\quad + 32h\left(L^2d + 2M^2\mathbf{m}_2^2 + G^2\right) + 8L^2h^2\left(M\mathbf{m}_1 + M + G\right) \\ &\leq -\frac{\lambda_{\mathsf{LSI}}}{2}\mathsf{KL}\left(\mu_t \middle\| \pi^*\right) + \frac{(LM + 128M^2)\eta}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta} \middle\| \pi^*\right) \\ &\quad + 32\eta\left(L^2d + 2M^2\mathbf{m}_2^2 + G^2\right) + 8L^2\eta^2\left(M\mathbf{m}_1 + M + G\right)\end{aligned}$$

Multiplying both sides by $e^{\lambda_{\mathsf{LSI}}(t-k\eta)/2}$, applying Grönwall's Lemma for $t \in [k\eta, (k+1)\eta]$ and using the fact that $e^{x/2} - 1 \leq x$ for $x \leq {}^{1}/_{6}$,

$$\begin{aligned}e^{\lambda_{\mathsf{LSI}}\eta/2}\mathsf{KL}\left(\mu_{(k+1)\eta} \middle\| \pi^*\right) - \mathsf{KL}\left(\mu_{k\eta} \middle\| \pi^*\right) \leq \frac{(2LM + 256M^2)\eta^2}{\lambda_{\mathsf{LSI}}}\mathsf{KL}\left(\mu_{k\eta} \middle\| \pi^*\right) \\ + 64\eta^2\left(L^2d + 2M^2\mathbf{m}_2^2 + G^2\right) + 16L^2\eta^3\left(M\mathbf{m}_1 + M + G\right)\end{aligned}$$

It follows that

$$\begin{aligned}\mathsf{KL}\left(\mu_{(k+1)\eta} \middle\| \pi^*\right) &\leq e^{-\lambda_{\mathsf{LSI}}\eta/2}\left(1 + \frac{(2LM + 256M^2)\eta^2}{\lambda_{\mathsf{LSI}}}\right)\mathsf{KL}\left(\mu_{k\eta} \middle\| \pi^*\right) \\ &\quad + e^{-\lambda_{\mathsf{LSI}}\eta/2}\left[64\eta^2\left(L^2d + 2M^2\mathbf{m}_2^2 + G^2\right) + 16L^2\eta^3\left(M\mathbf{m}_1 + M + G\right)\right] \\ &\leq e^{-\lambda_{\mathsf{LSI}}\eta/4}\mathsf{KL}\left(\mu_{k\eta} \middle\| \pi^*\right) \\ &\quad\quad + e^{-\lambda_{\mathsf{LSI}}\eta/4}\left[64\eta^2\left(L^2d + 2M^2\mathbf{m}_2^2 + G^2\right) + 16L^2\eta^3\left(M\mathbf{m}_1 + M + G\right)\right]\end{aligned}$$

Where the last inequality follows from the fact that $\eta \leq \frac{\lambda_{\mathsf{LSI}}^2}{8L(LM+128M^2)}$, implies $1 + \frac{(2LM+256M^2)\eta^2}{\lambda_{\mathsf{LSI}}} \leq 1 + \frac{\eta\lambda_{\mathsf{LSI}}}{4} \leq e^{\eta\lambda_{\mathsf{LSI}}/4}$. Iterating through the above recurrence, we conclude that for any $k \in (K)$,

$$\begin{aligned}\mathsf{KL}\left(\mu_{(k+1)\eta} \middle\| \pi^*\right) &\leq e^{-\lambda_{\mathsf{LSI}}\eta k/4}\mathsf{KL}\left(\mu_0 \middle\| \pi^*\right) \\ &\quad + \frac{e^{-\lambda_{\mathsf{LSI}}\eta/4}}{1 - e^{-\lambda_{\mathsf{LSI}}\eta/4}}\left[64\eta^2\left(L^2d + 2M^2\mathbf{m}_2^2 + G^2\right) + 16L^2\eta^3\left(M\mathbf{m}_1 + G\right)\right] \\ &\leq e^{-\lambda_{\mathsf{LSI}}\eta k/4}\mathsf{KL}\left(\mu_0 \middle\| \pi^*\right) + \frac{256\eta}{\lambda_{\mathsf{LSI}}}\left(L^2d + 2M^2\mathbf{m}_2^2 + G^2\right) + \frac{64L^2\eta^2}{\lambda_{\mathsf{LSI}}}\left(M\mathbf{m}_1 + M + G\right)\end{aligned} \tag{14}$$

Recalling that $\mu_{k\eta} = \text{Law}\left(\hat{\mathbf{x}}_k\right)$, the desired last iterate guarantee is,

$$\mathsf{KL}\left(\text{Law}\left(\hat{\mathbf{x}}_{K+1}\right) \middle\| \pi^*\right) \leq e^{-\lambda_{\mathsf{LSI}}\eta K/4}\mathsf{KL}\left(\mu_0 \middle\| \pi^*\right) + \frac{256\eta}{\lambda_{\mathsf{LSI}}}\left(L^2d + 2M^2\mathbf{m}_2^2 + G^2\right) + \frac{64L^2\eta^2}{\lambda_{\mathsf{LSI}}}\left(M\mathbf{m}_1 + M + G\right)$$

We now control the expected amortized batch size $\bar{B} = {}^1\!/\!{}_K \sum_{k=1}^{K} \mathbb{E}\left[B_k\right]$. We note that,

$$\begin{aligned}\mathbb{E}\left[B_k\right] &\leq 1 + \mathbb{E}\left[1 + \lceil M\|\hat{\mathbf{x}}_k\| + G\rceil\right] \leq 2 + M\mathbb{E}\left[\|\mathbf{x}_{k\eta}\|\right] + G \\ &\leq 2 + \frac{M\sqrt{2}}{\sqrt{\lambda_{\mathsf{LSI}}}}\sqrt{\mathsf{KL}\left(\mu_{k\eta}\middle|\middle|\pi^*\right)} + G\end{aligned}$$

From the above inequality and (14), it follows that,

$$\begin{aligned}\bar{B} &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[B_k\right] \\ &\leq 2 + G + \frac{M\sqrt{2\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)}}{K\sqrt{\lambda_{\mathsf{LSI}}}}\sum_{k=1}^{K} e^{-\lambda_{\mathsf{LSI}}\eta(k-1)/8} \\ &+ \frac{28M\sqrt{\eta}}{\lambda_{\mathsf{LSI}}}\left(L\sqrt{d} + M\mathbf{m}_2 + G\right) + \frac{8L\eta}{\lambda_{\mathsf{LSI}}}\sqrt{2M\mathbf{m}_1 + G} \\ &\leq 2 + G + \frac{50M}{\lambda_{\mathsf{LSI}}^{3/2}\eta K}\sqrt{\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)} + \frac{28M\sqrt{\eta}}{\lambda_{\mathsf{LSI}}}\left(L\sqrt{d} + M\mathbf{m}_2 + G\right) + \frac{8L\eta}{\lambda_{\mathsf{LSI}}}\sqrt{2M\mathbf{m}_1 + 2M + 2G}\end{aligned}$$

■

B.7. Proof of Theorem 7

Proof From Lemma 3, and using the fact that $k\eta \leq t \leq (k+1)\eta$, we obtain,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\mu_t\middle|\middle|\pi^*\right) \leq -\frac{1}{4}\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right) + 5L^2\eta d + \frac{3L\eta(M^2C_2d + G^2)}{B} + \frac{576\eta(M^4C_4d^2 + G^4)}{B^2}$$

Integrating from $t \in [k\eta, (k+1)\eta]$, and rearranging, we get,

$$\begin{aligned}\int_{k\eta}^{(k+1)\eta}\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right) &\leq 4\left[\mathsf{KL}\left(\mu_{k\eta}\middle|\middle|\pi^*\right) - \mathsf{KL}\left(\mu_{(k+1)\eta}\middle|\middle|\pi^*\right)\right] \\ &\qquad + 20L^2\eta^2 d + \frac{12L\eta^2(M^2C_2d + G^2)}{B} + \frac{2304\eta^2(M^4C_4d^2 + G^4)}{B^2}\end{aligned}$$

Averaging the above inequality for $k \in (K)$ and using the fact that $\mu_{k\eta} = \mathrm{Law}\left(\hat{\mathbf{x}}_k\right)$, we obtain.

$$\frac{1}{K\eta}\int_0^{K\eta}\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right) \leq \frac{4\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)}{K\eta} + 20L^2\eta d + \frac{12L\eta(M^2C_2d + G^2)}{B} + \frac{2304\eta(M^4C_4d^2 + G^4)}{B^2}$$

Using the fact that $\mathsf{FD}\left(\mu\middle|\middle|\pi\right)$ is a convex functional of μ for any probability measure π (Wu, 2000), we conclude that $\mathsf{FD}\left(\bar{\mu}_{K\eta}\middle|\middle|\pi^*\right) \leq \frac{1}{K\eta}\int_0^{K\eta}\mathsf{FD}\left(\mu_t\middle|\middle|\pi^*\right)$. Hence, we obtain the following guarantee.

$$\mathsf{FD}\left(\bar{\mu}_{K\eta}\middle|\middle|\pi^*\right) \leq \frac{4\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)}{K\eta} + 20L^2\eta d + \frac{12L\eta(M^2C_2d + G^2)}{B} + \frac{2304\eta(M^4C_4d^2 + G^4)}{B^2}$$

We now consider the case when π^* satisfies PI. From Lemma 15, we conclude that,

$$\mathsf{TV}(\bar{\mu}_{K\eta}, \pi^*)^2 \leq \frac{16\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_0\right)\middle|\middle|\pi^*\right)}{\lambda_{\mathsf{PI}}K\eta} + \frac{80L^2\eta d}{\lambda_{\mathsf{PI}}} + \frac{48L\eta(M^2C_2d + G^2)}{\lambda_{\mathsf{PI}}B} + \frac{9216\eta(M^4C_4d^2 + G^4)}{\lambda_{\mathsf{PI}}B^2}$$

■

Appendix C. Improved Wasserstein CLT with Gaussian Convolutions

We now prove the $\mathcal{W}_2$ CLT for Gaussian convolutions of bounded random vectors, as stated in Lemma 4. Our proof is an extension of the high-dimensional $\mathcal{W}_2$ CLT of Zhai (2018), whose proof structure we closely follow and adapt. In particular, the presence of the Gaussian convolution leads to improved regularity, which allows us to derive sharper bounds than Zhai (2018) by adapting their proof. Without loss of generality, we begin by assuming that $\Sigma_{\mathbf{Y}}$ is diagonal with entries ς_i^2 for $i \in [d]$. Under our assumptions, $\varsigma_i^2 \leq 1/5d$ for every $i \in [d]$, a fact which is crucial to our proof. Let $\hat{\mathbf{X}} = \sqrt{\mathbf{I} - \Sigma_{\mathbf{Y}}}\mathbf{X} + \mathbf{Y}$. Then, $\hat{\mathbf{X}}$ can be written as,

$$\hat{\mathbf{X}} \overset{d}{=} \frac{1}{\sqrt{B}} \sum_{k=1}^{B} \sqrt{\mathbf{I} - \Sigma_{\mathbf{Y}}}\mathbf{X}_k + \mathbf{Y}_k$$

where $\mathbf{X}_k \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$ are sampled independently of $\mathbf{Y}_1, \dots, \mathbf{Y}_B$. Our proof relies on the following key technical lemma, which is analogous to Lemma 1.6 of Zhai (2018). We present a proof of this lemma in Appendix C.2

Lemma 21

$$\mathcal{W}_2\left(\mathbf{X}_1, \sqrt{I - \frac{\Sigma_Y}{k}}\mathbf{X}_1 + \frac{1}{\sqrt{k}}\mathbf{Y}_1\right) \leq \sqrt{\frac{25d\beta^6}{k^3}}$$

C.1. Proof of Lemma 4

Proof The proof proceeds similar to the proof of Theorem 1.1 in Zhai (2018), using Lemma 21 as the key ingredient (analogous to Lemma 1.6 in Zhai (2018)). First, by the property that a sum of independent Gaussian random variables is also a Gaussian random variable, we have:

$$\begin{aligned}\mathcal{W}_2(\sum_{j=1}^{k}\mathbf{X}_j, \sum_{j=1}^{k-1}\mathbf{X}_j + \sqrt{\mathbf{I} - \Sigma_Y}\mathbf{Z}_k + \mathbf{Y}_k) &= \mathcal{W}_2\left(\sqrt{k}\mathbf{Z}_1, \sqrt{k}\sqrt{\mathbf{I} - \frac{\Sigma_Y}{k}}\mathbf{Z}_1 + \mathbf{Y}_k\right) \\ &= \sqrt{k}\mathcal{W}_2\left(\mathbf{Z}_1, \sqrt{\mathbf{I} - \frac{\Sigma_Y}{k}}\mathbf{Z}_1 + \frac{\mathbf{Y}_k}{\sqrt{k}}\right) \\ &\leq \frac{\sqrt{25\beta^6 d}}{k} \qquad (15)\end{aligned}$$

In the third step, we have applied Lemma 21. Now, by triangle inequality, we must have:

$$\begin{aligned}&\sqrt{B}\mathcal{W}_2\left(\mathbf{X}, \hat{\mathbf{X}}\right) = \mathcal{W}_2(\sum_{j=1}^{B}\hat{\mathbf{X}}_j, \sum_{j=1}^{B}\sqrt{\mathbf{I} - \Sigma_Y}\mathbf{X}_j + \mathbf{Y}_j) \\ &\leq \sum_{k=1}^{B}\mathcal{W}_2\left(\sum_{j=1}^{k}\mathbf{X}_j + \sum_{j=k+1}^{B}\sqrt{\mathbf{I} - \Sigma_Y}\mathbf{X}_j + \mathbf{Y}_j, \sum_{j=1}^{k-1}\mathbf{X}_j + \sum_{j=k}^{B}\sqrt{\mathbf{I} - \Sigma_Y}\mathbf{X}_j + \mathbf{Y}_j\right) \\ &\leq \sum_{k=1}^{B}\mathcal{W}_2\left(\sum_{j=1}^{k}\mathbf{X}_j, \sum_{j=1}^{k-1}\mathbf{X}_j + \sqrt{\mathbf{I} - \Sigma_Y}\mathbf{X}_k + \mathbf{Y}_k\right) \leq \sum_{k=1}^{B}\frac{5\beta^3\sqrt{d}}{k} \\ &\leq 5\beta^3\sqrt{d}(1 + \log B) \qquad (16)\end{aligned}$$

In the third step, we have used the fact that $\mathcal{W}_2(\mathbf{Z}+\mathbf{A},\mathbf{Z}+\mathbf{B}) \le \mathcal{W}_2(\mathbf{A},\mathbf{B})$ whenever $\mathbf{Z},\mathbf{A}$ and $\mathbf{Z},\mathbf{B}$ are independent random variables. Hence, we have proved that $\mathcal{W}_2^2\left(\mathbf{X},\hat{\mathbf{X}}\right) \le {}^{25\beta^6 d}/_{B}\,(1+\log B)^2$ ■

C.2. Proof of Lemma 21

Define $n_i := \frac{k}{\varsigma_i^2}$. Suppose $\mathbf{Y},\mathbf{Z}$ are independent random variables such that $\mathbf{Z} \sim \mathcal{N}(0,\mathbf{I})$ and $\mathbf{Y}$ has the same distribution as any one of $\mathbf{Y}_1,\dots,\mathbf{Y}_B$ used in the statement of Lemma 4. In this section only, for any vector $\mathbf{x} \in \mathbb{R}^d$, we will let $\mathbf{x}_i$ denote its component along the i-th standard basis vector. Notice that the i-th co-ordinate of $\sqrt{\mathbf{I}-\frac{\Sigma_{\mathbf{Y}}}{k}}\mathbf{Z} + \frac{1}{\sqrt{k}}\mathbf{Y}$ is given by $\sqrt{1-\frac{1}{n_i}}\mathbf{Z}_i + \bar{\mathbf{Y}}_i$, where, $\bar{\mathbf{Y}} := \frac{\mathbf{Y}}{\sqrt{k}}$. Clearly, $\|\bar{\mathbf{Y}}\| \le \frac{\beta}{\sqrt{k}}$, a fact which we will use heavily below. We will also use the observation that $\sum_i \varsigma_i^2 \le \beta^2$ and $\beta^2 n_j \ge k$.

Let $f(x) := \frac{\tau(\mathbf{x})}{\rho(\mathbf{x})}$ where τ is the density of $\sqrt{\mathbf{I}-\frac{\Sigma_{\mathbf{Y}}}{k}}\mathbf{Z}_1 + \bar{\mathbf{Y}}_1$ and ρ is the density of $\mathbf{Z}_1$. The proof of the following lemma is identical to that of Lemma 4.1 of Zhai (2018), with n replaced with the co-ordinate dependent n_i, and σ_i replaced with 1. The result relies on the fact that $\varsigma_i^2 \le {}^{1}/_{5d} < 1$.

Lemma 22

$$\mathbb{E}\left[f(\mathbf{Z}_1)^2\right] = \mathbb{E}_{\bar{\mathbf{Y}},\bar{\mathbf{Y}}'}\left[\exp\left(\sum_{i=1}^{d}\frac{2n_i^2\bar{\mathbf{Y}}_i\bar{\mathbf{Y}}_i' - n_i\bar{\mathbf{Y}}_i^2 - n_i(\bar{\mathbf{Y}}_i')^2+1}{2(n_i^2-1)} - r(n_i)\right)\right]$$

Where $\bar{\mathbf{Y}}_i'$ is an independent copy of $\bar{\mathbf{Y}}_i$ and $r(n) := \frac{1}{2(n^2-1)} - \frac{1}{2}\log\left(1+\frac{1}{n^2-1}\right)$

Proceeding similarly, we let $Q_i := \frac{2n_i^2\bar{\mathbf{Y}}_i\bar{\mathbf{Y}}_i' - n_i\bar{\mathbf{Y}}_i^2 - n_i(\bar{\mathbf{Y}}_i')^2+1}{2(n_i^2-1)} - r(n_i)$ and $Q := \sum_{i=1}^{d} Q_i$. Let $f_{(i)}(\mathbf{x})$ be the ratio $\frac{\tau_{(i)}(\mathbf{x})}{\gamma_{(i)}(\mathbf{x})}$, where $\tau_{(i)}(\mathbf{x})$ denotes the marginal density under τ of all co-ordinates other than the i-th co-ordinate. The following lemma is a rewriting of Lemmas 4.4 and 4.5 in Zhai (2018)

Lemma 23 $\sup_i 5\varsigma_i^2 d < 1$, $5\beta^2 \le 1$. *Then, the following bounds hold:*

1. $|Q_i| \le \frac{n_i^2|\bar{Y}_i||\bar{Y}_i'|}{n_i^2-1} + \frac{1}{2n_i}$, $|Q| \le 1$, $|Q-Q_i| \le 1$

2. $$\mathbb{E}Q_i = -\frac{1}{2(n_i^2-1)} - r(n_i)$$

3. $$\mathbb{E}Q_iQ_j \le \frac{n_i^2\delta_{ij}}{(n_i^2-1)^2} + \frac{n_in_j\mathbb{E}\bar{Y}_i^2\bar{Y}_j^2}{2(n_i^2-1)(n_j^2-1)} + \frac{1}{2(n_i^2-1)(n_j^2-1)}$$

4. $$\mathbb{E}Q_i^2 \le \frac{2n_i^2+n_i+1}{2(n_i^2-1)^2}$$

5.
$$\mathbb{E}(Q - Q_i)Q_i \le \sup_j \frac{k(d-1) + \beta^2 n_j}{2k(n_j^2-1)^2}\,.$$

6.
$$\mathbb{E}Q^2 \le \sup_j \frac{\beta^2 n_j d + 3n_j^2 kd}{2k(n_j^2-1)^2} \le \sup_j \frac{2n_j^2 d}{(n_j^2-1)^2}$$

Proof Items 1 - 4 and item 6 can be shown by essentially the same methods used in Lemmas 4.4 and 4.5 of Zhai (2018). For item 5, we have:

$$\begin{aligned}
\mathbb{E}(Q - Q_i)Q_i = \sum_{j\neq i} Q_j Q_i &\le \sup_l \frac{d-1}{2(n_l^2-1)^2} + \sum_{j\neq i} \frac{n_i n_j \mathbb{E}\bar{Y}_i^2 \bar{Y}_j^2}{2(n_i^2-1)(n_j^2-1)} \\
&\le \sup_l \frac{d-1}{2(n_l^2-1)^2} + \sup_l \frac{n_l}{2(n_l^2-1)^2} \sum_{j\neq i} n_i \mathbb{E}\bar{Y}_i^2 \bar{Y}_j^2 \\
&\le \sup_l \frac{d-1}{2(n_l^2-1)^2} + \sup_l \frac{n_l}{2(n_l^2-1)^2} \frac{\beta^2}{k} n_i \mathbb{E}\bar{Y}_i^2 \\
&= \sup_l \frac{n_l \beta^2 + k(d-1)}{2k(n_l^2-1)^2} \qquad (17)
\end{aligned}$$

In the last step we have used the fact that $\mathbb{E}\bar{Y}_i^2 = \frac{1}{n_i}$. ■

The proof of Lemma 21 now follows by using the bounds established above along with the proof of Lemma 1.6 in Zhai (2018) and by noting that $\beta^2 n_i^2 \ge k$ for every $i \in [d]$.

Appendix D. Analysis of SGLD via Entropic CLT

D.1. Technical Lemmas

Lemma 24 (Chain Rule for KL Divergence) *Suppose ν is a distribution over some Polish space Ξ^T and μ be a product distribution over Ξ^T given as $\mu = \otimes_{t=1}^T \mu_t$. Let $\nu_t(\cdot|X_{<t})$ denote the conditional distribution of the t-th co-ordinate conditioned on the co-ordinates $1, \dots, t-1$ (and the marginal of the first co-ordinate under ν when $t = 1$) and let $\nu_{<t}$ denote the joint marginal law of the co-ordinates $1, \dots, t-1$ under the measure ν.*

$$\mathsf{KL}\left(\nu || \mu\right) = \sum_{t=1}^T \mathbb{E}_{X_{<t}\sim\nu_{<t}} \mathsf{KL}\left(\nu_t(\cdot|X_{<t}) || \mu_t\right)$$

Proof Lemma 4.18 in Van Handel (2014) ■

Lemma 25 (Tensorization of T_2) *Suppose P is the law of $\mathcal{N}(0, \Sigma)$ for some non-singular Σ. Then for any probability measure Q over $\mathbb{R}^d$, we have:*

$$\mathcal{W}_2^2(P, Q) \le 2\lambda_{\max}(\Sigma)\mathsf{KL}\left(Q || P\right)$$

Proof Proposition 1.8 in Gozlan and Léonard (2010) ■

Lemma 26 (Reverse T_2 for Gaussian Convolutions) *Suppose $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, $\mathbf{A}, \mathbf{B}$ are random variables independent of $\mathbf{Z}$. Then, $\mathsf{KL}\left(\mathbf{Z}+\mathbf{A} \middle|\middle| \mathbf{Z}+\mathbf{B}\right) \leq \frac{1}{2\sigma^2}\mathcal{W}_2^2(\mathbf{A},\mathbf{B})$*

Proof Let Γ be a $\mathcal{W}_2$ optimal coupling between the laws of $\mathbf{A}$ and $\mathbf{B}$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right)$. The density of $Z + A$ with respect to the Lebesgue measure is given by $P(\mathbf{x}) = \mathbb{E} f(\mathbf{x} - \mathbf{A})$ and similarly the density of $Z + B$ is given by $Q(\mathbf{x}) = \mathbb{E} f(\mathbf{x} - \mathbf{B})$. Therefore, the KL divergence can be written as:

$$\begin{aligned}
\mathsf{KL}\left(\mathbf{Z}+\mathbf{A} \middle|\middle| \mathbf{Z}+\mathbf{B}\right) &= \int f(\mathbf{x}-\mathbf{A}) \log\left(\frac{\int f(\mathbf{x}-\mathbf{A}) d\Gamma(\mathbf{A}\times\mathbf{B})}{\int f(\mathbf{x}-\mathbf{B}) d\Gamma(\mathbf{A}\times\mathbf{B})}\right) d\Gamma(\mathbf{A}\times\mathbf{B}) d\mathbf{x} \\
&\leq \int f(\mathbf{x}-\mathbf{A}) \log\left(\frac{f(\mathbf{x}-\mathbf{A})}{f(\mathbf{x}-\mathbf{B})}\right) d\Gamma(\mathbf{A}\times\mathbf{B}) d\mathbf{x} \\
&= \frac{1}{2\sigma^2}\int f(\mathbf{x}-\mathbf{A}) \left[\|\mathbf{x}-\mathbf{B}\|^2 - \|\mathbf{x}-\mathbf{A}\|^2\right] d\Gamma(\mathbf{A}\times\mathbf{B}) d\mathbf{x} \\
&= \frac{1}{2\sigma^2}\int \|\mathbf{A}-\mathbf{B}\|^2 d\Gamma(\mathbf{A}\times\mathbf{B}) = \frac{1}{2\sigma^2}\mathcal{W}_2^2(\mathbf{A},\mathbf{B}) \qquad (18)
\end{aligned}$$

The second step above follows from the log-sum inequality. The third step follows from the definition of f. In the fourth step we have used Fubini's theorem to integrate out $\mathbf{x}$, noting that $f(\mathbf{x} - \mathbf{A})$ is the density of a Gaussian with covariance $\sigma^2\mathbf{I}$ and mean $\mathbf{A}$. We have finally used the fact that Γ is a $\mathcal{W}_2$ optimal coupling between $\mathbf{A}$ and $\mathbf{B}$. ■

Under the (low probability) event that the conditions in Lemma 4 are not satisfied, we use the Wasserstein CLT of Zhai (2018) to quantify the approximate subgaussianity of the stochastic gradient noise.

Lemma 27 (Zhai (2018)) *Let $\mathbf{Y} = {}^1\!/\!\sqrt{B}\sum_{i=1}^{B} \mathbf{Y}_i$ where $\mathbf{Y}_1, \dots, \mathbf{Y}_B \in \mathbb{R}^d$ are zero-mean i.i.d random vectors with covariance matrix Σ such that $\|\mathbf{Y}_i\| \leq \beta$ holds almost surely. Let $\mathbf{Z} \sim \mathcal{N}(0, \Sigma)$ be sampled independently of $\mathbf{Y}_i$. Then,*

$$\mathcal{W}_2^2(\mathbf{Y}, \mathbf{Z}) \leq \frac{25\beta^2 d\left(1+\log(B)\right)^2}{B}$$

D.2. Proof of Theorem 8

Proof Denoting $\mathbf{N}_k = \mathbf{N}(\hat{\mathbf{x}}_k, \xi_k)$, we note that the iterates of SGLD and LMC (with the same step-size and initialization) can be written as,

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta\nabla F(\mathbf{x}_k) + \sqrt{2\eta}\mathbf{z}_k \\
\hat{\mathbf{x}}_{k+1} &= \hat{\mathbf{x}}_k - \eta\nabla F(\hat{\mathbf{x}}_k) + \sqrt{2\eta}\hat{\mathbf{z}}_k
\end{aligned}$$

where $\mathbf{z}_k$ and $\hat{\mathbf{z}}_k$ are defined as,

$$\begin{aligned}
\mathbf{z}_k &= \epsilon_k \sim \mathcal{N}(0, \mathbf{I}), \\
\hat{\mathbf{z}}_k &= \sqrt{\eta/2}\mathbf{N}_k + \hat{\epsilon}_k, \ \ \hat{\epsilon}_k \sim \mathcal{N}(0, \mathbf{I})
\end{aligned}$$

Defining the filtration $\mathcal{F}_k = \sigma(\hat{\mathbf{x}}_0, \dots, \hat{\mathbf{x}}_k, \hat{\mathbf{z}}_0, \dots, \hat{\mathbf{z}}_{k-1})$, we observe that SGLD and LMC admit the same random function representation, i.e., there exists a measurable function H_K such that,

$$(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{K+1}) = H_K(\hat{\mathbf{x}}_0, \hat{\mathbf{z}}_0, \hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_K)$$
$$(\mathbf{x}_1, \dots, \mathbf{x}_{K+1}) = H_K(\mathbf{x}_0, \mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_K)$$

Since $\text{Law}(\hat{\mathbf{x}}_0) = \text{Law}(\mathbf{x}_0)$, we use the data processing inequality and Lemma 24 to conclude the following.

$$\mathsf{KL}\left(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{K+1} \middle|\middle| \mathbf{x}_1, \dots, \mathbf{x}_{K+1}\right) = \sum_{k=0}^{K} \mathbb{E}\left[\mathsf{KL}\left(\text{Law}(\hat{\mathbf{z}}_k \mid \mathcal{F}_k) \middle|\middle| \mathbf{z}_k\right)\right]$$

We shall now control each term in the above summation. To this end, let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{W} \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$. It follows that,

$$\begin{aligned}
\mathsf{KL}\left(\text{Law}(\hat{\mathbf{z}}_k \mid \mathcal{F}_k) \middle|\middle| \mathbf{z}_k\right) &= \mathsf{KL}\left(\mathbf{X}_1 + \sqrt{\eta/2}\mathbf{N}_k \middle|\middle| \mathbf{Z}_1 \middle| \mathcal{F}_k\right) \\
&= \mathsf{KL}\left(\sqrt{1/2}\mathbf{X}_2 + \sqrt{1/2}\mathbf{X} + \sqrt{\eta/2}\mathbf{N}_k \middle|\middle| \sqrt{1/2}\mathbf{Z}_2 + \sqrt{1/2}\mathbf{Z} \middle| \mathcal{F}_k\right) \\
&\leq \frac{1}{2}\mathcal{W}_2^2\left(\mathbf{X} + \sqrt{\eta}\mathbf{N}_k, \mathbf{Z} \middle| \mathcal{F}_k\right)
\end{aligned}$$

Where the last inequality follows from Lemma 26. Now, let $\mathbf{Y} = \sqrt{\eta}\mathbf{N}_k$ and define $\Sigma_{\mathbf{Y}} = \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^T \middle| \mathcal{F}_k\right]$. It follows that,

$$\begin{aligned}
\mathsf{KL}\left(\text{Law}(\hat{\mathbf{z}}_k \mid \mathcal{F}_k) \middle|\middle| \mathbf{z}_k\right) &\leq \frac{1}{2}\mathcal{W}_2^2\left(\mathbf{X} + \sqrt{\eta}\mathbf{N}_k, \mathbf{Z} \middle| \mathcal{F}_k\right) \\
&\leq \underbrace{\mathcal{W}_2^2\left(\mathbf{X} + \mathbf{Y}, \sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z} \middle| \mathcal{F}_k\right)}_{\text{Wasserstein CLT Term}} + \underbrace{\mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}, \mathbf{W} \middle| \mathcal{F}_k\right)}_{\text{Covariance Mismatch Term}} \qquad (19)
\end{aligned}$$

We first bound the covariance mismatch term via direct computation of the Wasserstein distance between two zero-mean Gaussians. It follows that,

$$\mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}, \mathbf{W} \middle| \mathcal{F}_k\right) = \text{Tr}\left(2\mathbf{I} + \Sigma_{\mathbf{Y}} - 2\sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\right)$$

Let $\lambda_1, \dots, \lambda_d \geq 0$ denote the eigenvalues of $\Sigma_{\mathbf{Y}}$. Then,

$$\begin{aligned}
\mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}, \mathbf{W} \middle| \mathcal{F}_k\right) &= \sum_{i=1}^{d}\left(2 + \lambda_i - 2\sqrt{1 + \lambda_i}\right) \\
&\leq \sum_{i=1}^{d} \lambda_i^2/4 \leq 1/4\left(\sum_{i=1}^{d}\lambda_i\right)^2 = \text{Tr}(\Sigma_{\mathbf{Y}})^2/4
\end{aligned}$$

Define $u_k = M\|\hat{\mathbf{x}}_k\| + G$. From the definition of $\mathbf{Y}$ and the lin-growth a.s condition, it follows that

$$\mathrm{Tr}\left(\Sigma_{\mathbf{Y}}\right) = {}^{\eta}\!/\!{}_{B}\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}_k, \xi_1) - \nabla F(\mathbf{x}_k)\|^2 \mid \hat{\mathbf{x}}_k\right] \leq \frac{\eta u_k^2}{B}$$

Hence, the covariance mismatch term is bounded as follows

$$\mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}, \mathbf{W}\middle|\mathcal{F}_k\right) \leq \frac{\eta^2 u_k^4}{4B^2} \tag{20}$$

For controlling the Wasserstein CLT term, we observe that $\mathbf{Y} = {}^{1}\!/\!{}_{\sqrt{B}}\sum_{i=1}^{B}\mathbf{Y}_i$ where $\mathbf{Y}_i = \sqrt{{}^{\eta}\!/\!{}_{B}}\mathbf{N}(\hat{\mathbf{x}}_k, \xi_{k,i})$. We note that $\|\mathbf{Y}_i\| \leq u_k\sqrt{{}^{\eta}\!/\!{}_{B}}$. Our proof proceeds by considering two cases

Case 1 : $u_k^2 > {}^{B}\!/\!{}_{5\eta d}$ In this case, the conditions required to apply Lemma 4 are not satisfied, and hence, we control this term using the Wasserstein CLT of Zhai (2018).

$$\begin{aligned}\mathcal{W}_2^2\left(\mathbf{X} + \mathbf{Y}, \sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}\middle|\mathcal{F}_k\right) &= \mathcal{W}_2^2\left(\mathbf{X} + \mathbf{Y}, \mathbf{Z}_1 + \sqrt{\Sigma_{\mathbf{Y}}}\mathbf{Z}_2\middle|\mathcal{F}_k\right)\\ &= \mathcal{W}_2^2\left(\mathbf{Y}, \sqrt{\Sigma_{\mathbf{Y}}}\mathbf{Z}_2\middle|\mathcal{F}_k\right) \leq \frac{25\eta d(1+\log(B))^2 u_k^2}{B^2}\end{aligned}$$

where the last inequality follows from Lemma 27

Case 2 : $u_k^2 \leq {}^{B}\!/\!{}_{5\eta d}$ We first note that, $\mathrm{Tr}\left(\Sigma_{\mathbf{Y}}\right) \leq {}^{\eta u_k^2}\!/\!{}_{B} \leq {}^{1}\!/\!{}_{5d}$. We then define $\mathbf{G}_i = (\mathbf{I} + \Sigma_{\mathbf{Y}})^{-1/2}\mathbf{Y}_i$ and $\mathbf{G} = {}^{1}\!/\!{}_{\sqrt{B}}\sum_{i=1}^{B}\mathbf{G}_i$. Moreover, let $\Gamma_{\mathbf{G}} = \mathbb{E}\left[\mathbf{G}_i\mathbf{G}_i^T\right]$. We note that, since $(\mathbf{I} + \Sigma_{\mathbf{Y}})^{-1/2} \preceq \mathbf{I}$, $\|\mathbf{G}_i\|^2 \leq \|\mathbf{Y}_i\|^2 \leq {}^{1}\!/\!{}_{5}$. Furthermore,

$$\Gamma_{\mathbf{G}} = \left(\mathbf{I} + \Sigma_{\mathbf{Y}}\right)^{-1}\Sigma_{\mathbf{Y}} = \mathbf{I} - \left(\mathbf{I} + \Sigma_{\mathbf{Y}}\right)^{-1}$$

Moreover, let $\lambda_1, \dots, \lambda_d$ denote the eigenvalues of $\Sigma_{\mathbf{Y}}$ and $\mu_1, \dots, \mu_d$ denote the eigenvalues of $\Gamma_{\mathbf{G}}$. From the above identity, we note that, $\mu_i = {}^{\lambda_i}\!/\!{}_{1+\lambda_i}$ hence $\|\Gamma_{\mathbf{G}}\| \leq \|\Sigma_{\mathbf{Y}}\| \leq {}^{1}\!/\!{}_{5d}$ and $\mathrm{Tr}\left(\Gamma_{\mathbf{G}}\right) \leq \mathrm{Tr}\left(\Sigma_{\mathbf{Y}}\right)$. Hence, the conditions required to apply Lemma 4 are satisfied. It follows that,

$$\begin{aligned}\mathcal{W}_2^2\left(\mathbf{X} + \mathbf{Y}, \sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}\middle|\mathcal{F}_k\right) &\leq \|\mathbf{I} + \Sigma_{\mathbf{Y}}\|\mathcal{W}_2^2\left(\left(\mathbf{I} + \Sigma_{\mathbf{Y}}\right)^{-1/2}\mathbf{X} + \mathbf{G}, \mathbf{Z}\middle|\mathcal{F}_k\right)\\ &\leq \frac{6}{5}\mathcal{W}_2^2\left(\sqrt{\mathbf{I} - \Gamma_{\mathbf{G}}}\mathbf{X} + \mathbf{G}, \mathbf{Z}\middle|\mathcal{F}_k\right)\end{aligned}$$

where the first inequality follows from the fact that $\|\mathbf{I} + \Sigma_{\mathbf{Y}}\| = 1 + \|\Sigma_{\mathbf{Y}}\| \leq 1 + {}^{1}\!/\!{}_{5d} \leq {}^{6}\!/\!{}_{5}$. Now, using the sharper Wasserstein CLT of Lemma 4, we conclude that,

$$\begin{aligned}\mathcal{W}_2^2\left(\mathbf{X} + \mathbf{Y}, \sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}\middle|\mathcal{F}_k\right) &\leq \frac{6}{5}\mathcal{W}_2^2\left(\sqrt{\mathbf{I} - \Gamma_{\mathbf{G}}}\mathbf{X} + \mathbf{G}, \mathbf{Z}\middle|\mathcal{F}_k\right)\\ &\leq \frac{30\eta^3 d\left(1 + \log(B)\right)^2 u_k^6}{B^4}\end{aligned}$$

From Case 1 and Case 2, it follows that,

$$\begin{aligned}\mathcal{W}_2^2\left(\mathbf{X}+\mathbf{Y}, \sqrt{\mathbf{I}+\Sigma_{\mathbf{Y}}}\mathbf{Z}\middle|\mathcal{F}_k\right) &\leq \frac{25\eta d(1+\log(B))^2 u_k^2}{B^2}\mathbb{I}_{\left\{u_k^2 > B/5\eta d\right\}} \\ &+ \frac{30\eta^3 d\left(1+\log(B)\right)^2 u_k^6}{B^4}\mathbb{I}_{\left\{u_k^2 \leq B/5\eta d\right\}} \\ &\leq \frac{25\eta d(1+\log(B))^2 u_k^2}{B^2}\mathbb{I}_{\left\{u_k^2 > B/5\eta d\right\}} + \frac{30\eta^3 d\left(1+\log(B)\right)^2 u_k^6}{B^4}\end{aligned} \tag{21}$$

Thus, from (19), (20) and (21), it follows that,

$$\begin{aligned}\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{z}}_k \mid \mathcal{F}_k\right) \middle|\middle| \mathbf{z}_k\right) &\leq \frac{\eta^2 u_k^4}{4B^2} + \frac{30\eta^3 d\left(1+\log(B)\right)^2 u_k^6}{B^4} \\ &+ \frac{25\eta d(1+\log(B))^2 u_k^2}{B^2}\mathbb{I}_{\left\{u_k^2 > B/5\eta d\right\}}\end{aligned}$$

Recalling that $u_k = M\|\hat{\mathbf{x}}_k\| + G$, we use the lin-growth a.s condition for $p = 6$ to conclude that,

$$\begin{aligned}\mathbb{E}\left[u_k^4\right] &\leq 8\left(M^4\mathbb{E}\left[\|\hat{\mathbf{x}}\|^4\right] + G^4\right) \leq 8\left(M^4 d^2 + G^4\right) \\ \mathbb{E}\left[u_k^6\right] &\leq 32\left(M^6\mathbb{E}\left[\|\hat{\mathbf{x}}\|^6\right] + G^6\right) \leq 32\left(M^6 d^3 + G^6\right) \\ \mathbb{E}\left[u_k^2\mathbb{I}_{\left\{u_k^2 > B/5\eta d\right\}}\right] &\leq \frac{75\eta^2 d^2}{2B^2}\mathbb{E}\left[u_k^6\right] \leq \frac{1200\eta^2 d^2}{B^2}\left(M^6 d^3 + G^6\right)\end{aligned}$$

where the last inequality uses Lemma 12. From the above inequalities, we obtain

$$\begin{aligned}\mathbb{E}\left[\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{z}}_k \mid \mathcal{F}_k\right) \middle|\middle| \mathbf{z}_k\right)\right] &\leq \frac{2\eta^2}{B^2}\left(M^4 d^2 + G^4\right) + \frac{960\eta^3}{B^4}\left(M^6 d^4 + G^6 d\right)(1+\log B)^2 \\ &+ \frac{2000\eta^3}{B^4}\left(M^6 d^{3+3} + G^6 d^3\right)(1+\log B)^2 \\ &\leq \frac{2\eta^2}{B^2}\left(M^4 d^2 + G^4\right) + \frac{3000\eta^3}{B^4}\left(M^6 d^{3+3} + G^6 d^3\right)(1+\log B)^2\end{aligned}$$

Thus, we finally obtain the following statistical indistinguishability guarantee

$$\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K} \middle|\middle| \mathbf{x}_{1:K}\right) \leq \frac{2\eta^2 K}{B^2}\left(M^4 d^2 + G^4\right) + \frac{3000\eta^3 K}{B^4}\left(M^6 d^6 + G^6 d^3\right)(1+\log B)^2$$

■

D.3. Convergence of Non-Smooth SGLD under α-LO

In this section, we use $\mathcal{R}_q\left(\mu \middle|\middle| \nu\right)$ to denote the Rényi divergence of order q between two measures μ and ν.

Corollary 28 (Convergence of SGLD under LO) *Let the s-Hölder, lin-growth a.s, and p-moment growth be satisfied with $p = 6$. Furthermore, assume the target π^* satisfies α-LO for some $\alpha \in [1, 2]$ and define $\beta = {}^2\!/\!_\alpha - 1$. Then, for $\epsilon \leq {}^1\!/\!_{\mathsf{poly}(d)}$, the last iterate of SGLD, under appropriate Gaussian initialization, requires N stochastic gradient oracle calls to ensure TV$(\text{Law}\left(\hat{\mathbf{x}}_K\right), \pi^*) \leq \epsilon$, where*

$$N = \tilde{\Theta}\left(\frac{d^{\max\{1+\beta(1+1/s), 3/2(1+\beta)+\beta/2s\}}}{\epsilon^{2/s}}\right)$$

Proof Our proof relies on the recent result of Chewi et al. (2022a) where the authors establish an unstable convergence guarantee for LMC under Hölder smoothness and Latała-Oleskiewicz inequality. In particular, Theorem 7 of Chewi et al. (2022a) implies that, under appropriate Gaussian initialization such that $\Delta_0 = \mathcal{R}_3\left(\text{Law}\left(\mathbf{x}_0\right) \middle|\middle| \pi^*\right) = O(d)$, it suffices to set η and K as follows to achieve $\text{TV}\left(\text{Law}\left(\mathbf{x}_{K+1}\right), \pi^*\right) \leq \epsilon$

$$\eta = \tilde{\Theta}\left(\frac{\epsilon^{2/s}}{d\Delta_0^{\beta/s}}\right)$$
$$K = \tilde{\Theta}\left(\frac{d\Delta_0^{\beta(1+1/s)}}{\epsilon^{2/s}}\right)$$
$$T = \eta K = \tilde{\Theta}\left(\Delta_0^{\beta}\right)$$

where $\beta = {}^2\!/\!_\alpha - 1$. Under this choice of η and T, Theorem 8 suggests that $\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K} \middle|\middle| \mathbf{x}_{1:K}\right) \leq O({}^{\eta T d^2}\!/\!_{B^2})$. Thus, to ensure $\text{TV}(\text{Law}\left(\hat{\mathbf{x}}_{K+1}\right), \text{Law}\left(\mathbf{x}_{K+1}\right)) \leq {}^\epsilon\!/\!_2$, it suffices to set $B \geq \tilde{O}\left(\max\left\{1, d^{1/2}\Delta_0^{\beta/2(1-1/s)}\right\}\right)$. Hence, $\text{TV}(\text{Law}\left(\hat{\mathbf{x}}_{K+1}\right), \pi^*) \leq \epsilon$ by subadditivity of Total Variation. The required stochastic gradient complexity is,

$$N = KB = \tilde{\Theta}\left(\max\left\{\frac{d\Delta_0^{\beta(1+1/s)}}{\epsilon^{2/s}}, \frac{d^{3/2}\Delta_0^{3\beta/2+\beta/2s}}{\epsilon^{2/s}}\right\}\right)$$
$$= \tilde{\Theta}\left(\frac{d^{\max\{1+\beta(1+1/s), 3/2(1+\beta)+\beta/2s\}}}{\epsilon^{2/s}}\right)$$

where we use the fact that $\Delta_0 = O(d)$. ■

Rates Under LSI and Smoothness We recall that the LO inequality of order $\alpha = 2$ is equivalent to the Logarithmic Sobolev Inequality. Moreover, as a consequence of the fact $\alpha - 1 \leq s \leq 1$, we note that LSI and Holder continuity together imply smoothness. To this end, we observe that Corollary 28 implies an oracle complexity of $\tilde{\Theta}({}^{d^{3/2}}\!/\!_{\epsilon^2})$ to ensure ϵ-convergence in TV for the last iterate. We note that this matches the guarantee implied by Theorem 5 via Pinsker's inequality. However, we emphasize that Theorem 5 is a stronger result as it establishes convergence in KL divergence, which automatically implies convergence in TV (due to Pinsker's inequality) and $\mathcal{W}_2$ (due to the Otto-Villani theorem). Furthermore, unlike Corollary 28, Theorem 5 is a stable convergence guarantee.

Last-Iterate Convergence Under Poincare Inequality From Assumption 4, we recall that the LO inequality of order $\alpha = 1$ is equivalent to the Poincare Inequality. In this setting, we observe that that Corollary 28 implies an oracle complexity of $\tilde{\Theta}\left(\frac{d^{\max\{2+1/s, 3+1/2s\}}}{\epsilon^{2/s}}\right)$ to ensure ϵ-convergence in TV for the last iterate. To the best of our knowledge, this is the first known result for the last-iterate convergence of SGLD under the Poincare Inequality, as well as the first analysis of SGLD under Poincare Inequality that does not assume smoothness. For the smooth case, i.e., when $s = 1$, we note that the implied oracle complexity is $\tilde{\Theta}\left(\frac{d^{3.5}}{\epsilon^2}\right)$ for last-iterate convergence. When compared to the $\tilde{O}\left(\frac{d^{2.5}}{\epsilon^4}\right)$ guarantee of Theorem 7 for *average-iterate convergence in TV*, we note that the rates implied by Corollary 28 exhibit an improved ϵ-dependence at the cost of a worse d dependence. However, we highlight that, unlike Corollary 28, Theorem 7 is a stable convergence guarantee

Appendix E. Analysis of Covariance Corrected SGLD

E.1. Technical Lemmas

Lemma 29 *Let* $\mathbf{Z}, \mathbf{Z}_2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$ *and let* $\mathbf{N}$ *be a zero-mean random vector independent of* $\mathbf{Z}, \mathbf{Z}_2$ *such that* $\Sigma = \mathsf{Cov}\left[\mathbf{N}\right] \prec \mathbf{I}/2$. *Let* $\hat{\Sigma}$ *be an arbitrary PSD matrix with* $\hat{\Sigma} \prec \mathbf{I}/2$ *and let* $\bar{\mathbf{Z}} = (\mathbf{I} - \hat{\Sigma}/2)\mathbf{Z} + \mathbf{N}$. *Then,*

$$\mathsf{KL}\left(\bar{\mathbf{Z}}||\mathbf{Z}\right) \leq 2\mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2 + \mathbf{N}\right) \\ + 8\lambda_{\max}(\tfrac{\mathbf{I}}{2} - \Sigma)\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2 || \sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2\right) + 2\,\mathrm{Tr}\left(\hat{\Sigma}\right)^3$$

Proof Since $\hat{\Sigma} \prec \mathbf{I}/2$, $(\mathbf{I} - \hat{\Sigma}/2)^2 \succ \mathbf{I}/2$, and thus $\mathbf{B} = \sqrt{(\mathbf{I} - \hat{\Sigma}/2)^2 - \mathbf{I}/2}$ is a well defined symmetric positive definite matrix. Furthermore, if $\mathbf{Z}_1, \mathbf{Z}_2$ are i.i.d. isotropic Gaussians over $\mathbb{R}^d$, we can write: $\mathbf{Z} \overset{d}{=} \frac{1}{\sqrt{2}}\mathbf{Z}_1 + \frac{1}{\sqrt{2}}\mathbf{Z}_2$ and $\bar{\mathbf{Z}} \overset{d}{=} \frac{1}{\sqrt{2}}\mathbf{Z}_1 + \mathbf{B}\mathbf{Z}_2 + \mathbf{N}$, where $\mathbf{N}$ is independent of both $\mathbf{Z}_1$ and $\mathbf{Z}_2$. Applying Lemma 26 with $\sigma = \frac{1}{\sqrt{2}}$ and $\mathbf{Z}$ replaced with $\mathbf{Z}_1$, we have:

$$\begin{aligned}
&\mathsf{KL}\left(\bar{\mathbf{Z}}||\mathbf{Z}\right) \leq \mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \mathbf{B}\mathbf{Z}_2 + \mathbf{N}\right) \\
&\leq 2\mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2 + \mathbf{N}\right) + 2\mathcal{W}_2^2\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2 + \mathbf{N}, \mathbf{B}\mathbf{Z}_2 + \mathbf{N}\right) \\
&\leq 2\mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2 + \mathbf{N}\right) + 2\mathcal{W}_2^2\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2, \mathbf{B}\mathbf{Z}_2\right) \\
&\leq 2\mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2 + \mathbf{N}\right) + 4\mathcal{W}_2^2\left(\mathbf{B}\mathbf{Z}_2, \sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2\right) + 4\mathcal{W}_2^2\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2, \sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2\right) \\
&\leq 2\mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2 + \mathbf{N}\right) + 8\lambda_{\max}(\tfrac{\mathbf{I}}{2} - \Sigma)\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2 || \sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2\right) \\
&+ 4\mathcal{W}_2^2\left(\mathbf{B}\mathbf{Z}_2, \sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2\right),
\end{aligned} \tag{22}$$

where the second step uses the triangle inequality for $\mathcal{W}_2$ along with the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, the third step follows from the definition of $\mathcal{W}_2$ by restricting to the $\mathcal{W}_2$-optimal coupling between $\sqrt{\frac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2$ and $\mathbf{B}\mathbf{Z}_2$, and the last step follows from an application of Lemma 25. The remainder of our proof controls the term $\mathcal{W}_2^2\left(\mathbf{B}\mathbf{Z}_2, \sqrt{\frac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2\right)$. To this end, we define the matrices $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ as follows,

$$\begin{aligned}\mathbf{A}_1 &= \tfrac{1}{2}\mathbf{I} - \hat{\Sigma} \\ \mathbf{A}_2 &= \mathbf{B}^2 = \left(\mathbf{I} - \tfrac{1}{2}\hat{\Sigma}\right)^2 - \tfrac{1}{2}\mathbf{I} \\ \mathbf{A}_3 &= \mathbf{A}_1 + \mathbf{A}_2 - 2\left(\mathbf{A}_1^{1/2}\mathbf{A}_2\mathbf{A}_1^{1/2}\right)^{1/2}\end{aligned}$$

Since $0 \preceq \hat{\Sigma} \prec \mathbf{I}/2$, $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \hat{\Sigma}$ are simultaneously diagonalizable PSD matrices. To this end, let $\lambda_1, \dots, \lambda_d$ denote the eigenvalues of $\hat{\Sigma}$. Clearly, $0 \leq \lambda_i < 1/2$ for all $i \in [d]$. Moreover, the corresponding eigenvalue $\mu_1, \dots, \mu_d$ of $\mathbf{A}_3$ is given by

$$\mu_i = (1 - \lambda_i/2)^2 - \lambda_i - 2\sqrt{(1/2 - \lambda_i)\left[(1 - \lambda_i/2)^2 - 1/2\right]} \leq \lambda_i^3/2$$

where the last inequality follows from a Taylor expansion. By direct computation of the Wasserstein distance between zero-mean Gaussians, we have,

$$\mathcal{W}_2^2\left(\mathbf{B}\mathbf{Z}_2, \sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2\right) = \mathrm{Tr}(\mathbf{A}_3) = \sum_{i=1}^{d}\mu_i \leq \sum_{i=1}^{d}\lambda_i^3/2 \leq \frac{1}{2}\left(\sum_{i=1}^{d}\lambda_i\right)^3 = \frac{1}{2}\,\mathrm{Tr}\left(\hat{\Sigma}\right)^3$$

Substituting the above inequality into (22), we obtain,

$$\begin{aligned}\mathsf{KL}\left(\bar{\mathbf{Z}}||\mathbf{Z}\right) &\leq 2\mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2 + \mathbf{N}\right) \\ &\quad + 8\lambda_{\max}(\tfrac{\mathbf{I}}{2} - \Sigma)\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2||\sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2\right) + 2\,\mathrm{Tr}\left(\hat{\Sigma}\right)^3\end{aligned}$$

■

Lemma 30 *For any PSD matrices $\Sigma, \hat{\Sigma} \prec \mathbf{I}/2$*

$$\begin{aligned}\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2||\sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2\right) &= \mathrm{tr}\left((\mathbf{I} - 2\Sigma)^{-1}\left(\Sigma - \hat{\Sigma}\right)\right) \\ &\quad + \sum_{k=1}^{\infty} 2^{k-1}\frac{\left[\mathrm{tr}\left(\hat{\Sigma}^k\right) - \mathrm{tr}\left(\Sigma^k\right)\right]}{k}\end{aligned} \tag{23}$$

Proof From standard formula for KL divergence between two multi-variate Gaussians, it follows that:

$$\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \hat{\Sigma}}\mathbf{Z}_2||\sqrt{\tfrac{1}{2}\mathbf{I} - \Sigma}\mathbf{Z}_2\right) = \frac{1}{2}\left[\log\frac{\det\left(\frac{\mathbf{I}}{2} - \Sigma\right)}{\det\left(\frac{\mathbf{I}}{2} - \hat{\Sigma}\right)} - d + \mathrm{tr}\left((\tfrac{\mathbf{I}}{2} - \Sigma)^{-1}(\tfrac{\mathbf{I}}{2} - \hat{\Sigma})\right)\right]$$

We first note that by basic algebraic manipulation,

$$\operatorname{tr}\left((\tfrac{\mathbf{I}}{2} - \Sigma)^{-1}(\tfrac{\mathbf{I}}{2} - \hat{\Sigma})\right) - d = 2\operatorname{tr}\left((\mathbf{I} - 2\Sigma)^{-1}\left(\Sigma - \hat{\Sigma}\right)\right)$$

Now, consider $\log(\det(A))$ for a $d \times d$ PSD matrix $0 \prec A \prec \mathbf{I}$. Taking $\lambda_1, \dots, \lambda_d$ to be the eigenvalues of A, we have

$$\begin{aligned}\log(\det(A)) = \sum_{i=1}^{d} \log(\lambda_i) = \sum_{i=1}^{d} \log(1 + \lambda_i - 1) &= \sum_{i=1}^{d}\sum_{k=1}^{\infty}(-1)^{k-1}\frac{(\lambda_i - 1)^k}{k} \\ &= -\sum_{k=1}^{\infty}\frac{\operatorname{tr}\left[(\mathbf{I} - A)^k\right]}{k} \qquad (24)\end{aligned}$$

Therefore, $\log\frac{\det\left(\frac{\mathbf{I}}{2} - \Sigma\right)}{\det\left(\frac{\mathbf{I}}{2} - \hat{\Sigma}\right)} = \log\frac{\det(\mathbf{I} - 2\Sigma)}{\det\left(\mathbf{I} - 2\hat{\Sigma}\right)} = \sum_{k=1}^{\infty}\frac{2^k}{k}\left[\operatorname{tr}\left(\hat{\Sigma}^k\right) - \operatorname{tr}\left(\Sigma^k\right)\right]$, which proves the result. ■

Lemma 31 (Covariance Estimation Guarantees) *Consider any $B' \geq 1$. Let $\mathbf{y}_1, \dots, \mathbf{y}_{B'}$ and $\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_{B'}$ be i.i.d samples from some probability measure P supported on $\mathbb{R}^d$, with covariance matrix $\tilde{\Sigma}$. Furthermore, assume $\|\mathbf{y} - \mathbb{E}[\mathbf{y}]\| \leq M$ holds almost surely for any $\mathbf{y} \sim P$. Define Σ as $\tilde{\Sigma}/B$ and $\hat{\Sigma}$ as follows,*

$$\hat{\Sigma} = \frac{1}{2BB'}\sum_{j=1}^{B'}\left(\mathbf{y}_j - \tilde{\mathbf{y}}_j\right)\left(\mathbf{y}_j - \tilde{\mathbf{y}}_j\right)^T$$

Then, the following holds,

1. $\mathbb{E}\left[\hat{\Sigma}\right] = \Sigma$
2. $\mathbb{E}\left[\operatorname{Tr}\left(\hat{\Sigma}^2\right)\right] - \operatorname{Tr}\left(\Sigma^2\right) \leq \frac{4M^4}{B^2B'}$
3. $\operatorname{Tr}\left(\hat{\Sigma}^k\right) \leq \left(\frac{2M^2}{B}\right)^k$

Proof The first property can be verified directly by taking expectations and using the fact that $\mathbf{y}_{1:B'}, \tilde{\mathbf{y}}_{1:B'}$ are independent. To prove the third property, we note that $\|\mathbf{y} - \mathbb{E}[\mathbf{y}]\| \leq M$ implies that, $\operatorname{Tr}(\Sigma), \operatorname{Tr}\left(\hat{\Sigma}\right) \leq 2M^2/B$. Hence, $\operatorname{Tr}\left(\hat{\Sigma}^k\right) \leq \operatorname{Tr}\left(\hat{\Sigma}\right)^k \leq \left(\frac{2M^2}{B}\right)^k$. Finally, to prove the second property, define $\mathbf{A}_j = \left(\mathbf{y}_j - \tilde{\mathbf{y}}_j\right)\left(\mathbf{y}_j - \tilde{\mathbf{y}}_j\right)^T$. Then,

$$\hat{\Sigma}^2 = \frac{1}{4B^2(B')^2}\sum_{j=1}^{B'}\mathbf{A}_j^2 + \frac{1}{4B^2(B')^2}\sum_{k,l\in[B'],k\neq l}\mathbf{A}_k\mathbf{A}_l$$

Since $\mathbf{y}_{1:B'}, \tilde{\mathbf{y}}_{1:B'}$ are sampled i.i.d, it follows that, $\mathbb{E}[\mathbf{A}_k\mathbf{A}_l] = 4\Sigma_i^2$ whenever $k \neq l$. Hence, we can conclude that,

$$\mathbb{E}\left[\hat{\Sigma}^2\right] - \Sigma^2 = \frac{\mathbb{E}\left[\mathbf{A}_1^2\right]}{B^2B'} - \frac{\Sigma^2}{B^2B'}$$

The final result follows by taking the trace on both sides and observing that $\operatorname{Tr}\left(\mathbf{A}_1^2\right) \leq 16M^4$ ■

E.2. Proof of Theorem 10

Proof Define $u_k = M\|\hat{\mathbf{x}}_k\| + G$. Denoting $\mathbf{N}_k = \mathbf{N}(\hat{\mathbf{x}}_k, \xi_k)$ and $\Sigma_k = \mathbb{E}\left[\mathbf{N}_k \mathbf{N}_k^T \mid \hat{\mathbf{x}}_k\right]$, we note that the iterates of CC-SGLD and LMC (with the same step-size and initialization) can be written as,

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta \nabla F(\mathbf{x}_k) + \sqrt{2\eta}\mathbf{z}_k \\ \hat{\mathbf{x}}_{k+1} &= \hat{\mathbf{x}}_k - \eta \nabla F(\hat{\mathbf{x}}_k) + \sqrt{2\eta}\hat{\mathbf{z}}_k\end{aligned}$$

where $\mathbf{z}_k$ and $\hat{\mathbf{z}}_k$ are defined as,

$$\begin{aligned}\mathbf{z}_k &= \epsilon_k \sim \mathcal{N}(0, \mathbf{I}), \\ \hat{\mathbf{z}}_k &= \sqrt{\eta/2}\mathbf{N}_k + \hat{\epsilon}_k \mathbb{I}_{\{u_k^2 > B/5\eta d\}} + \left(\mathbf{I} - \tfrac{\eta}{4}\hat{\Sigma}_k\right)\hat{\epsilon}_k \mathbb{I}_{\{u_k^2 \leq B/5\eta d\}}, \quad \hat{\epsilon}_k \sim \mathcal{N}(0, \mathbf{I})\end{aligned}$$

Defining the filtration $\mathcal{F}_k = \sigma(\hat{\mathbf{x}}_0, \dots, \hat{\mathbf{x}}_k, \hat{\mathbf{z}}_0, \dots, \hat{\mathbf{z}}_{k-1})$, we observe that CC-SGLD and LMC admit the same random function representation, i.e., there exists a measurable function H_K such that,

$$\begin{aligned}(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{K+1}) &= H_K(\hat{\mathbf{x}}_0, \hat{\mathbf{z}}_0, \hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_K) \\ (\mathbf{x}_1, \dots, \mathbf{x}_{K+1}) &= H_K(\mathbf{x}_0, \mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_K)\end{aligned}$$

Since $\text{Law}(\mathbf{x}_0) = \text{Law}(\hat{\mathbf{x}}_0)$, we use the data processing inequality and Lemma 24 to obtain,

$$\mathsf{KL}\left(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{K+1} \middle|\middle| \mathbf{x}_1, \dots, \mathbf{x}_{K+1}\right) = \sum_{k=0}^{K} \mathbb{E}\left[\mathsf{KL}\left(\text{Law}(\hat{\mathbf{z}}_k \mid \mathcal{F}_k) \middle|\middle| \mathbf{z}_k\right)\right]$$

Define the filtration $\mathcal{G}_k = \mathcal{F}_k \vee \sigma(\xi_{k,j}^{(1)}, \xi_{k,j}^{(2)} \mid i \in \{1, 2\}, j \in [B])$. By Jensen's inequality, it follows that,

$$\mathbb{E}\left[\mathsf{KL}\left(\text{Law}(\hat{\mathbf{z}}_k \mid \mathcal{F}_k) \middle|\middle| \mathbf{z}_k\right)\right] \leq \mathbb{E}\left[\mathsf{KL}\left(\text{Law}(\hat{\mathbf{z}}_k \mid \mathcal{G}_k) \middle|\middle| \mathbf{z}_k\right)\right]$$

Let E denote the event $E = \{u_k^2 > B/5\eta d\}$. We note that $\hat{\Sigma}_k$ is $\mathcal{G}_k$ measurable and E is $\mathcal{F}_k$-measurable. Furthermore, let $\mathbf{X}, \mathbf{W}, \mathbf{Z} \sim \mathcal{N}(0, \mathbf{I})$. Hence,

$$\begin{aligned}\mathbb{E}\left[\mathsf{KL}\left(\text{Law}(\hat{\mathbf{z}}_k \mid \mathcal{G}_k) \middle|\middle| \mathbf{z}_k\right)\right] &= \mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \mathbf{X}\mathbb{I}_{\{E\}} + \left(\mathbf{I} - \eta/4\hat{\Sigma}_k\right)\mathbf{W}\mathbb{I}_{\{E^C\}} \middle|\middle| \mathbf{Z} \middle| \mathcal{G}_k\right) \\ &= \underbrace{\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \mathbf{X} \middle|\middle| \mathbf{Z} \middle| \mathcal{G}_k\right)\mathbb{I}_{\{E\}}}_{\text{Uncorrected KL}} + \underbrace{\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \left(\mathbf{I} - \eta/4\hat{\Sigma}_k\right)\mathbf{W} \middle|\middle| \mathbf{Z} \middle| \mathcal{G}_k\right)\mathbb{I}_{\{E^C\}}}_{\text{Covariance Corrected KL}}\end{aligned}$$

We proceed by controlling the uncorrected and covariance corrected KL terms separately.

Bounding the Uncorrected KL To control the uncorrected KL, we essentially repeat the same arguments as Theorem 8. In particular, let $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3 \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$. From Lemma 26, it follows that,

$$\begin{aligned}\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \mathbf{X}\middle|\middle|\mathbf{Z}\middle|\mathcal{G}_k\right) &= \mathsf{KL}\left(\sqrt{1/2}\mathbf{X}_1 + \sqrt{1/2}\mathbf{X}_2 + \sqrt{\eta/2}\mathbf{N}_k\middle|\middle|\sqrt{1/2}\mathbf{Z}_1 + \sqrt{1/2}\mathbf{Z}_2\middle|\mathcal{G}_k\right) \\ &\leq \frac{1}{2}\mathcal{W}_2^2\left(\mathbf{X}_1 + \sqrt{\eta}\mathbf{N}_k, \mathbf{Z}_1\middle|\mathcal{G}_k\right) \\ &\leq \mathcal{W}_2^2\left(\mathbf{X}_1 + \sqrt{\eta}\mathbf{N}_k, \sqrt{\mathbf{I} + \eta\Sigma_k}\mathbf{Z}_3\middle|\mathcal{G}_k\right) + \mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \eta\Sigma_k}\mathbf{Z}_3, \mathbf{Z}_2\middle|\mathcal{G}_k\right)\end{aligned}$$

Hence,

$$\begin{aligned}\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \mathbf{X}\middle|\middle|\mathbf{Z}\middle|\mathcal{G}_k\right)\mathbb{I}_{\{u_k^2 > B/5\eta d\}} &\leq \mathcal{W}_2^2\left(\mathbf{X}_1 + \sqrt{\eta}\mathbf{N}_k, \sqrt{\mathbf{I} + \eta\Sigma_k}\mathbf{Z}_3\middle|\mathcal{G}_k\right)\mathbb{I}_{\{u_k^2 > B/5\eta d\}} \\ &+ \mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \eta\Sigma_k}\mathbf{Z}_3, \mathbf{Z}_2\middle|\mathcal{G}_k\right)\mathbb{I}_{\{u_k^2 > B/5\eta d\}}\end{aligned}$$

We recall from our analysis of Theorem 8 that the first term corresponds to the Wasserstein CLT term whereas the second term corresponds to the covariance mismatch. Thus, repeating the same arguments as Theorem 8, we conclude that,

$$\mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \eta\Sigma_k}\mathbf{Z}_3, \mathbf{Z}_2\middle|\mathcal{G}_k\right) \leq \frac{\eta^2 u_k^4}{4B^2}$$

For the Wasserstein CLT term, we note that $\mathbf{N}_k = 1/\sqrt{B}\sum_{j=1}^{B} \mathbf{N}_{k,j}/\sqrt{B}$ where $\|\mathbf{N}_{k,j}\| \leq u_k$. Thus, from the Wasserstein CLT of Zhai (2018), it follows that,

$$\mathcal{W}_2^2\left(\mathbf{X}_1 + \sqrt{\eta}\mathbf{N}_k, \sqrt{\mathbf{I} + \eta\Sigma_k}\mathbf{Z}_3\middle|\mathcal{G}_k\right) = \eta\mathcal{W}_2^2\left(\mathbf{N}_k, \sqrt{\Sigma_k}\mathbf{Z}_3\middle|\mathcal{G}_k\right) \leq \frac{25\eta d(1 + \log(B))^2 u_k^2}{B^2}$$

Hence, we obtain,

$$\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \mathbf{X}\middle|\middle|\mathbf{Z}\middle|\mathcal{G}_k\right)\mathbb{I}_{\{u_k^2 > B/5\eta d\}} \leq \frac{\eta^2 u_k^4}{4B^2}\mathbb{I}_{\{u_k^2 > B/5\eta d\}} + \frac{25\eta d(1 + \log(B))^2 u_k^2}{B^2}\mathbb{I}_{\{u_k^2 > B/5\eta d\}}$$

From Lemma 12, we note that,

$$\begin{aligned}\mathbb{E}\left[u_k^2\mathbb{I}_{\{u_k^2 > B/5\eta d\}}\right] &\leq \frac{75\eta^2 d^2}{B^2}\mathbb{E}\left[u_k^6\right] \\ \mathbb{E}\left[u_k^4\mathbb{I}_{\{u_k^2 > B/5\eta d\}}\right] &= \mathbb{E}\left[u_k^4\mathbb{I}_{\{u_k^4 > (B/5\eta d)^2\}}\right] \\ &\leq \frac{375\eta^3 d^3}{B^3}\mathbb{E}\left[u_k^6\right]\end{aligned}$$

It follows that,

$$\mathbb{E}\left[\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \mathbf{X}\middle|\middle|\mathbf{Z}\middle|\mathcal{G}_k\right)\mathbb{I}_{\{u_k^2 > B/5\eta d\}}\right] \leq \frac{100\eta^5 d^3}{B^5}\mathbb{E}\left[u_k^6\right] + \frac{1875\eta^3 d^3\left(1 + \log(B)\right)^2}{B^4}\mathbb{E}\left[u_k^6\right]$$

From the lin-growth a.s and p-moment growth conditions, we know that, $\mathbb{E}\left[u_k^6\right] \leq 32\left(M^6 C_6 d^3 + G^6\right)$. Hence, the uncorrected KL term is bounded as follows,

$$\mathbb{E}\left[\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \mathbf{X}\middle|\middle|\mathbf{Z}\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E\}}\right] \leq \left(\frac{3200\eta^5}{B^3} + \frac{1875\eta^3\left(1 + \log(B)\right)^2}{B^4}\right)\left(M^6 C_6 d^6 + G^6 d^3\right) \tag{25}$$

Bounding the Covariance Corrected KL Let $\mathbf{Z}_1, \mathbf{Z}_2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$. We note that the presence of the indicator $\mathbb{I}_{\{E^c\}} = \mathbb{I}_{\{u_k^2 \leq B/5\eta d\}}$ ensures $\Sigma_k, \hat{\Sigma}_k \prec \mathbf{I}/2$. Hence, applying Lemma 29,

$$\begin{aligned}
\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \left(I - \eta/4\hat{\Sigma}_k\right)\mathbf{W}||\mathbf{Z}\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}} &\leq 2\mathcal{W}_2^2\left(\sqrt{1/2\mathbf{I} - \eta/2\Sigma_k}\mathbf{Z}_1 + \sqrt{\eta/2}\mathbf{N}_k, \frac{\mathbf{Z}_2}{\sqrt{2}}\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}} \\
&+ 4\|\mathbf{I} - \eta\Sigma_k\|\mathsf{KL}\left(\sqrt{1/2\mathbf{I} - \eta/2\hat{\Sigma}_k}\mathbf{Z}_1||\sqrt{1/2\mathbf{I} - \eta/2\Sigma_k}\mathbf{Z}_2\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}} \\
&+ \tfrac{\eta^3}{4}\operatorname{Tr}\left(\hat{\Sigma}_k\right)^3\mathbb{I}_{\{E^C\}} \\
&\leq \mathcal{W}_2^2\left(\sqrt{\mathbf{I} - \eta\Sigma_k}\mathbf{Z}_1 + \sqrt{\eta}\mathbf{N}_k, \mathbf{Z}_2\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}} \\
&+ 4\mathsf{KL}\left(\sqrt{1/2\mathbf{I} - \eta/2\hat{\Sigma}_k}\mathbf{Z}_1||\sqrt{1/2\mathbf{I} - \eta/2\Sigma_k}\mathbf{Z}_2\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}} \\
&+ \frac{\eta^3}{4}\operatorname{Tr}\left(\hat{\Sigma}_k\right)^3\mathbb{I}_{\{E^C\}}
\end{aligned}$$

It follows that,

$$\begin{aligned}
\mathbb{E}\left[\mathsf{KL}\left(\sqrt{\eta/2}\mathbf{N}_k + \sqrt{I - \eta/2\hat{\Sigma}_k}\mathbf{W}||\mathbf{Z}\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}}\right] &\leq \mathbb{E}\left[\mathcal{W}_2^2\left(\sqrt{\mathbf{I} - \eta\Sigma_k}\mathbf{Z}_1 + \sqrt{\eta}\mathbf{N}_k, \mathbf{Z}_2\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}}\right] \\
&+ \frac{\eta^3}{4}\mathbb{E}\left[\operatorname{Tr}\left(\hat{\Sigma}_k\right)^3\mathbb{I}_{\{E^C\}}\right] \\
&+ \mathbb{E}\left[\mathsf{KL}\left(\sqrt{1/2\mathbf{I} - \eta/2\hat{\Sigma}_k}\mathbf{Z}_1||\sqrt{1/2\mathbf{I} - \eta/2\Sigma_k}\mathbf{Z}_2\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}}\right]
\end{aligned}$$

As before, the first term corresponds to the Wasserstein CLT, the second term corresponds to the error due to linearization of the matrix square root, and the last term corresponds to the covariance mismatch of CC-SGLD, which depends upon how well $\hat{\Sigma}_k$ approximates Σ_k. We note that the error due to linearization can be easily controlled via the lin-growth a.s and p-moment growth conditions as follows:

$$\frac{\eta^3}{4}\mathbb{E}\left[\operatorname{Tr}\left(\hat{\Sigma}_k\right)^3\mathbb{I}_{\{E^C\}}\right] \leq \frac{\eta^3}{4B^3}\mathbb{E}\left[u_k^6\right] \leq \frac{8\eta^3}{B^3}\left(M^6 C_6 d^3 + G^6\right)$$

where we use the fact that $\operatorname{Tr}\left(\hat{\Sigma}_k\right) \leq \frac{u_k^2}{B}$ and $\mathbb{E}\left[u_k^6\right] \leq 32\left(M^6 C_6 d^3 + G^6\right)$ as per the lin-growth a.s and p-moment growth conditions We then control the Wasserstein CLT term via Lemma 4. We note that, similar to Theorem 8, the presence of the indicator $\mathbb{I}_{\{E^c\}} = \mathbb{I}_{\{u_k^2 \leq B/5\eta d\}}$ ensures that all conditions of Lemma 4 are satisfied. Hence, repeating similar arguments, we obtain the following bound

$$\mathcal{W}_2^2\left(\sqrt{\mathbf{I} - \eta\Sigma_k}\mathbf{Z}_1 + \sqrt{\eta}\mathbf{N}_k, \mathbf{Z}_2\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}} \leq \frac{25\eta^3 d\left(1 + \log(B)\right)^2 u_k^6}{B^4}$$

From the lin-growth a.s and p-moment growth conditions, we know that, $\mathbb{E}\left[u_k^6\right] \leq 32\left(M^6 C_6 d^3 + G^6\right)$. Hence,

$$\mathbb{E}\left[\mathcal{W}_2^2\left(\sqrt{\mathbf{I} - \eta\Sigma_k}\mathbf{Z}_1 + \sqrt{\eta}\mathbf{N}_k, \mathbf{Z}_2\middle|\mathcal{G}_k\right)\mathbb{I}_{\{E^C\}}\right] \leq \frac{800\eta^3 d}{B^4}\left(1 + \log(B)\right)^2\left(M^6 C_6 d^{3+1} + G^6 d\right)$$

We now control the covariance mismatch term by applying Lemma 30. We note that, since $\mathrm{Tr}(\Sigma), \mathrm{Tr}\left(\hat{\Sigma}\right) \leq {}^{u_k^2}/_{B}$, the presence of the indicator $\mathbb{I}_{\{u_k^2 \leq {}^{B}/_{5\eta d}\}}$ ensures that $\eta\Sigma_k, \eta\hat{\Sigma}_k \prec {}^{\mathbf{I}}/_{2}$. Hence, it follows that,

$$\mathsf{KL}\left(\sqrt{{}^{1}/_{2}\mathbf{I} - {}^{\eta}/_{2}\hat{\Sigma}_k}\mathbf{Z}_1 \middle|\middle| \sqrt{{}^{1}/_{2}\mathbf{I} - {}^{\eta}/_{2}\Sigma_k}\mathbf{Z}_2 \middle| \mathcal{G}_k\right)\mathbb{I}_{\{E^C\}} \leq \frac{1}{2}\mathrm{Tr}\left((\mathbf{I} - \eta\Sigma_k)^{-1}\left(\hat{\Sigma}_k - \Sigma_k\right)\right)\mathbb{I}_{\{E^c\}} + \frac{1}{2}\sum_{j=1}^{\infty}\frac{\eta^j}{j}\left[\mathrm{Tr}\left(\hat{\Sigma}_k^j\right) - \mathrm{Tr}\left(\hat{\Sigma}_k^j\right)\right]\mathbb{I}_{\{E^c\}}$$

We now control this quantity by first taking conditional expectation with respect to $\mathcal{F}_k$. From Property 1 of Lemma 31, we know that $\mathbb{E}\left[\hat{\Sigma}_k \mid \mathcal{F}_k\right] = \Sigma_k$. Furthermore $\mathbb{I}_{\{E\}}$ is $\mathcal{F}_k$ measurable. Hence,

$$\mathbb{E}\left[\mathsf{KL}\left(\sqrt{{}^{1}/_{2}\mathbf{I} - {}^{\eta}/_{2}\hat{\Sigma}_k}\mathbf{Z}_1 \middle|\middle| \sqrt{{}^{1}/_{2}\mathbf{I} - {}^{\eta}/_{2}\Sigma_k}\mathbf{Z}_2 \middle| \mathcal{G}_k\right) | \mathcal{F}_k\right]\mathbb{I}_{\{E^C\}} \leq \frac{\eta^2}{4}\left[\mathbb{E}\left[\mathrm{Tr}\left(\hat{\Sigma}_k^2\right) \mid \mathcal{F}_k\right] - \mathrm{Tr}\left(\hat{\Sigma}_k^2\right)\right] + \frac{1}{6}\sum_{j=3}^{\infty}\eta^j\,\mathrm{Tr}\left(\hat{\Sigma}_k^j\right)\mathbb{I}_{\{E^c\}}$$

Applying Property 2 and Property 3 of Lemma 31 wherever appropriate,

$$\begin{aligned}\frac{\eta^2}{4}\left[\mathbb{E}\left[\mathrm{Tr}\left(\hat{\Sigma}_k^2\right) \mid \mathcal{F}_k\right] - \mathrm{Tr}\left(\hat{\Sigma}_k^2\right)\right] &\leq \frac{\eta^2 u_k^4}{B^3} \\ \sum_{j=3}^{\infty}\eta^j\,\mathrm{Tr}\left(\hat{\Sigma}_k^j\right)\mathbb{I}_{\{E^c\}} &= \left(\frac{2\eta u_k^2}{B}\right)^3\sum_{j=0}^{\infty}\left(\frac{2\eta u_k^2}{B}\right)^j\mathbb{I}_{\{u_k^2 \leq {}^{B}/_{5\eta d}\}} \\ &\leq \frac{8\eta^3 u_k^6}{B^3}\sum_{j=0}^{\infty}({}^{2}/_{5})^j \leq \frac{40\eta^3 u_k^6}{3B^3}\end{aligned}$$

Hence,

$$\mathbb{E}\left[\mathsf{KL}\left(\sqrt{{}^{1}/_{2}\mathbf{I} - {}^{\eta}/_{2}\hat{\Sigma}_k}\mathbf{Z}_1 \middle|\middle| \sqrt{{}^{1}/_{2}\mathbf{I} - {}^{\eta}/_{2}\Sigma_k}\mathbf{Z}_2 \middle| \mathcal{G}_k\right) | \mathcal{F}_k\right]\mathbb{I}_{\{E^C\}} \leq \frac{\eta^2 u_k^4}{B^3} + \frac{3\eta^3 u_k^6}{B^3}$$

Recall that, as per the lin-growth a.s and p-moment growth conditions, $\mathbb{E}\left[u_k^4\right] \leq 8\left(M^4 C_4 d^2 + G^4\right)$ and $\mathbb{E}\left[u_k^6\right] \leq 32\left(M^6 C_6 d^3 + G^6\right)$. Hence, the covariance corrected KL term is controlled as,

$$\mathbb{E}\left[\mathsf{KL}\left(\sqrt{{}^{1}/_{2}\mathbf{I} - {}^{\eta}/_{2}\hat{\Sigma}_k}\mathbf{Z}_1 \middle|\middle| \sqrt{{}^{1}/_{2}\mathbf{I} - {}^{\eta}/_{2}\Sigma_k}\mathbf{Z}_2 \middle| \mathcal{G}_k\right)\mathbb{I}_{\{E^C\}}\right] \leq \frac{8\eta^2}{B^3}\left(M^4 C_4 d^2 + G^4\right) + \frac{96\eta^3}{B^3}\left(M^6 C_6 d^3 + G^6\right) \tag{26}$$

From (25) and (26), we finally obtain the following statistical indistinguishability guarantee,

$$\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{x}}_{1:K}\right)\middle|\middle|\mathrm{Law}\left(\mathbf{x}_{1:K}\right)\right) \leq \frac{8\eta^2 K}{B^3}\left(M^4 C_4 d^2 + G^4\right) + \frac{96\eta^3 K}{B^3}\left(M^6 C_6 d^3 + G^6\right) + \left(\frac{3200\eta^5 K}{B^3} + \frac{1875\eta^3 K\left(1 + \log(B)\right)^2}{B^4}\right)\left(M^6 C_6 d^6 + G^6 d^3\right)$$

■

E.3. Convergence of Non-Smooth CC-SGLD under LO

As before, we use $\mathcal{R}_q\left(\mu\middle|\middle|\nu\right)$ to denote the Rényi divergence of order q between two measures μ and ν.

Corollary 32 (Convergence of CC-SGLD under LO) *Let the s-Hölder, lin-growth a.s, and p-moment growth be satisfied with $p = 6$. Furthermore, assume the target π^* satisfies α-LO for some $\alpha \in [1, 2]$ and define $\beta = {}^2/_\alpha - 1$. Then, for $\epsilon \leq {}^1/_{\mathsf{poly}(d)}$, the last iterate of CC-SGLD, under appropriate Gaussian initialization, requires N stochastic gradient oracle calls to ensure TV$(\mathrm{Law}\left(\hat{\mathbf{x}}_K\right), \pi^*) \leq \epsilon$, where*

$$N = \tilde{\Theta}\left(\frac{d^{\max\{1+\beta(1+1/s), 4/3(1+\beta+\beta/2s)\}}}{\epsilon^{2/s}}\right)$$

Proof The proof of this result closely resembles that of Corollary 28. In particular, we know from Theorem 7 of Chewi et al. (2022a) that under appropriate Gaussian initialization such that $\Delta_0 = \mathcal{R}_3\left(\mathrm{Law}\left(\mathbf{x}_0\right)\middle|\middle|\pi^*\right) = O(d)$, the following choice of η and K suffices to ensure $\mathrm{TV}\left(\mathrm{Law}\left(\mathbf{x}_{K+1}\right), \pi^*\right) \leq \epsilon$

$$\eta = \tilde{\Theta}\left(\frac{\epsilon^{2/s}}{d\Delta_0^{\beta/s}}\right)$$
$$K = \tilde{\Theta}\left(\frac{d\Delta_0^{\beta(1+1/s)}}{\epsilon^{2/s}}\right)$$
$$T = \eta K = \tilde{\Theta}\left(\Delta_0^{\beta}\right)$$

where $\beta = {}^2/_\alpha - 1$. Since $\epsilon \leq {}^1/_{\mathsf{poly}(d)}$, under this choice of η and T, Theorem 10 suggests that $\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K}\middle|\middle|\mathbf{x}_{1:K}\right) \leq O({}^{\eta T d^2}/_{B^3})$. Thus, to ensure $\mathrm{TV}(\mathrm{Law}\left(\hat{\mathbf{x}}_{K+1}\right), \mathrm{Law}\left(\mathbf{x}_{K+1}\right)) \leq {}^\epsilon/_2$, it suffices to set $B \geq \tilde{O}\left(\max\left\{1, d^{1/3}\Delta_0^{\beta/3(1-1/s)}\right\}\right)$. Hence, $\mathrm{TV}(\mathrm{Law}\left(\hat{\mathbf{x}}_{K+1}\right), \mathrm{Law}\left(\mathbf{x}_{K+1}\right)) \leq \epsilon$ by subadditivity of Total Variation. The required stochastic gradient complexity is,

$$\begin{aligned} N = KB &= \tilde{\Theta}\left(\max\left\{\frac{d\Delta_0^{\beta(1+1/s)}}{\epsilon^{2/s}}, \frac{d^{4/3}\Delta_0^{4\beta/3+2\beta/3s}}{\epsilon^{2/s}}\right\}\right) \\ &= \tilde{\Theta}\left(\frac{d^{\max\{1+\beta(1+1/s), 4/3(1+\beta+\beta/2s)\}}}{\epsilon^{2/s}}\right) \end{aligned}$$

where we use the fact that $\Delta_0 = O(d)$. ■

We observe that for $s \geq {}^1/_2$, the oracle complexity of CC-SGLD strictly improves upon SGLD (whereas for $s \leq {}^1/_2$ both have the same oracle complexity since $B = O(1)$ in that setting). We elucidate this improvement for the special cases of LSI and PI as follows.

Rates Under Smoothness and LSI Recall that α-LO of order $\alpha = 2$ is equivalent to LSI. Furthermore, LSI and Holder continuity together imply smoothness. To this end, Corollary 32 implies an oracle complexity of $\tilde{\Theta}(\frac{d^{4/3}}{\epsilon^2})$ for attaining last iterate ϵ-convergence in TV. We note that this improves upon all our prior guarantees for SGLD under this setting. However, we highlight that the guarantee implied by Corollary 32 is unstable.

Rates Under PI Since α-LO of order $\alpha = 1$ is equivalent to PI, Corollary 32 implies an oracle complexity of $\tilde{\Theta}\left(\frac{d^{\max\{2+1/s, 8/3+2/3s\}}}{\epsilon^{2/s}}\right)$ under this setting for last-iterate ϵ-convergence in TV. For $s = 1$, the implied oracle complexity is $\tilde{\Theta}\left(\frac{d^{10/3}}{\epsilon^2}\right)$. This strictly improves upon the last-iterate guarantee implied by Corollary 28 in this setting. When compared to the $\tilde{O}\left(\frac{d^{2.5}}{\epsilon^4}\right)$ average-iterate guarantee implied by Theorem 7, we note an improved dependence on ϵ at the cost of a sublinear additional d dependence

Appendix F. Analysis of RBM and CC-RBM

F.1. Proof of Theorem 9

Proof The proof of this result is similar to that of Theorem 8 and uses the same technical tools. We first note that the iterates of IPD and RBM, with the same initialization and step-size can be expressed as,

$$\mathbf{x}_{k+1}^i = \mathbf{x}_k^i + \eta \mathbf{g}_k^i(\mathbf{x}_k^i) + \frac{\eta}{n}\sum_{j=1}^{n} \mathbf{K}_k^{ij}(\mathbf{x}_k^i, \mathbf{x}_k^j) + \sqrt{\eta}\sigma \mathbf{z}_k^i$$

$$\hat{\mathbf{x}}_{k+1}^i = \hat{\mathbf{x}}_k^i + \eta \mathbf{g}_k^i(\hat{\mathbf{x}}_k^i) + \frac{\eta}{n}\sum_{j=1}^{n} \mathbf{K}_k^{ij}(\hat{\mathbf{x}}_k^i, \hat{\mathbf{x}}_k^j) + \sqrt{\eta}\sigma \hat{\mathbf{z}}_k^i$$

where $\mathbf{z}_k^i$ and $\hat{\mathbf{z}}_k^i$ are defined as follows for any $k \in [K], n \in [n]$

$$\mathbf{z}_k^i = \epsilon_k^i \sim \mathcal{N}(0, \mathbf{I}),$$

$$\hat{\mathbf{z}}_k^i = \frac{\sqrt{\eta}}{\sigma}\mathbf{N}_k^i + \hat{\epsilon}_k^i, \quad \hat{\epsilon}_k^i \sim \mathcal{N}(0, \mathbf{I})$$

$$\mathbf{N}_k^i = \frac{1}{B}\sum_{j=1}^{B}\left[\mathbf{K}_k^{iI_k^{ij}}(\hat{\mathbf{x}}_k^i, \hat{\mathbf{x}}_k^{I_k^{ij}}) - \frac{1}{n}\sum_{l=1}^{n}\mathbf{K}_k^{il}(\hat{\mathbf{x}}_k^i, \hat{\mathbf{x}}_k^l)\right]$$

where $I_k^{i1}, \dots, I_k^{iB} \overset{\text{iid}}{\sim} \mathsf{Uniform}([n])$ for every $i \in [n]$. Define the filtration $\mathcal{F}_k = \sigma(\hat{\mathbf{x}}_0^i, \dots, \hat{\mathbf{x}}_k^i, \hat{\mathbf{z}}_0^i, \dots, \hat{\mathbf{z}}_{k-1}^i | i \in [n])$. Clearly, $\mathbf{N}_k^i$ is an empirical average of zero-mean i.i.d random variables conditioned on $\mathcal{F}_k$. Furthermore, we note that IPD and RBM admit the same random function representation, i.e., there exists a measurable function H_K such that,

$$(\hat{\mathbf{x}}_{1:K+1}^i | i \in [n]) = H_K(\hat{\mathbf{x}}_0^i, \hat{\mathbf{z}}_0^i, \hat{\mathbf{z}}_1^i, \dots, \hat{\mathbf{z}}_K^i | i \in [n])$$

$$(\mathbf{x}_{1:K+1}^i | i \in [n]) = H_K(\mathbf{x}_0^i, \mathbf{z}_0^i, \mathbf{z}_1^i, \dots, \mathbf{z}_K^i | i \in [n])$$

Since $\text{Law}\left(\hat{\mathbf{x}}_0^i\right) = \text{Law}\left(\mathbf{x}_0^i\right)$ for every $i \in [n]$, we use the data processing inequality and Lemma 24 to conclude the following.

$$\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K+1}^i | i \in [n] || \mathbf{x}_{1:K+1}^i | i \in [n]\right) = \sum_{k=0}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\mathsf{KL}\left(\text{Law}\left(\hat{\mathbf{z}}_k^i \mid \mathcal{F}_k\right) || \mathbf{z}_k^i\right)\right]$$

We shall now control each term in the above summation by following the same steps as Theorem 8. To this end, let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{W} \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$. It follows that,

$$\begin{aligned}
\mathsf{KL}\left(\text{Law}\left(\hat{\mathbf{z}}_k^i \mid \mathcal{F}_k\right) \middle|\middle| \mathbf{z}_k^i\right) &= \mathsf{KL}\left(\mathbf{X}_1 + {\sqrt{\eta}}/{\sigma}\mathbf{N}_k^i \middle|\middle| \mathbf{Z}_1 \middle| \mathcal{F}_k\right) \\
&= \mathsf{KL}\left(\sqrt{1/2}\mathbf{X}_2 + \sqrt{1/2}\mathbf{X} + {\sqrt{\eta}}/{\sigma}\mathbf{N}_k^i \middle|\middle| \sqrt{1/2}\mathbf{Z}_2 + \sqrt{1/2}\mathbf{Z} \middle| \mathcal{F}_k\right) \\
&\leq \frac{1}{2}\mathcal{W}_2^2\left(\mathbf{X} + {\sqrt{2\eta}}/{\sigma}\mathbf{N}_k^i, \mathbf{Z} \middle| \mathcal{F}_k\right)
\end{aligned}$$

Where the last inequality follows from Lemma 26. Now, let $\mathbf{Y} = {\sqrt{2\eta}}/{\sigma}\mathbf{N}_k^i$ and define $\Sigma_{\mathbf{Y}} = \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^T \middle| \mathcal{F}_k\right]$. It follows that,

$$\begin{aligned}
\mathsf{KL}\left(\text{Law}\left(\hat{\mathbf{z}}_k^i \mid \mathcal{F}_k\right) \middle|\middle| \mathbf{z}_k^i\right) &\leq \frac{1}{2}\mathcal{W}_2^2\left(\mathbf{X} + \sqrt{\eta}\mathbf{N}_k, \mathbf{Z} \middle| \mathcal{F}_k\right) \\
&\leq \underbrace{\mathcal{W}_2^2\left(\mathbf{X} + \mathbf{Y}, \sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z} \middle| \mathcal{F}_k\right)}_{\text{Wasserstein CLT Term}} + \underbrace{\mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}, \mathbf{W} \middle| \mathcal{F}_k\right)}_{\text{Covariance Mismatch Term}} \qquad (27)
\end{aligned}$$

We note that, $\text{Tr}(\Sigma_{\mathbf{Y}}) \leq {4\eta M^2}/{B\sigma^2}$. Controlling the covariance mismatch term in a manner similar to Theorem 8 (i.e., by direct computation of the Wasserstein distance), we obtain,

$$\mathcal{W}_2^2\left(\sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z}, \mathbf{W} \middle| \mathcal{F}_k\right) \leq \frac{\text{Tr}(\Sigma_{\mathbf{Y}})^2}{4} \leq \frac{4\eta^2 M^4}{\sigma^4 B^2} \qquad (28)$$

To control the Wasserstein CLT term, we observe that the CLT structure of $\mathbf{N}_k^i$ allows us to express $\mathbf{Y}$ as $\mathbf{Y} = {1}/{\sqrt{B}}\sum_{j=1}^B \mathbf{Y}^{(j)}$ where $\left\|\mathbf{Y}^{(j)}\right\| \leq \frac{2\sqrt{2\eta}}{\sigma\sqrt{B}}M$. We note that the condition $\eta \leq {B\sigma^2}/{40\eta M^2 d}$ ensures that $\left\|\mathbf{Y}^{(j)}\right\|^2 \leq {1}/{5}$ and $\text{Tr}(\Sigma_{\mathbf{Y}}) \leq {1}/{5d}$. Thus, all the conditions required to apply Lemma 4 are satisfied. Applying the same arguments used to control the Wasserstein CLT term in Case 2 of Theorem 8, we conclude that the Wasserstein CLT term is bounded as follows for some universal constant C_{CLT},

$$\mathcal{W}_2^2\left(\mathbf{X} + \mathbf{Y}, \sqrt{\mathbf{I} + \Sigma_{\mathbf{Y}}}\mathbf{Z} \middle| \mathcal{F}_k\right) \leq \frac{C_{\mathsf{CLT}}\eta^3 M^6 d(1 + \log(B))^2}{\sigma^6 B^4} \qquad (29)$$

Thus, from equations (27), (28) and (29), we obtain the following statistical indistinguishability guarantee,

$$\mathsf{KL}\left((\hat{\mathbf{x}}_k^i)_{i \in [n], k \in [K]} \middle|\middle| (\mathbf{x}_k^i)_{i \in [n], k \in [K]}\right) \leq C_{\mathsf{Cov}}\frac{\eta^2 M^4 nK}{B^2 \sigma^4} + C_{\mathsf{CLT}}\frac{d\eta^3 M^6 nK(1 + \log(B))^2}{B^4 \sigma^6}$$

where C_{Cov} and C_{CLT} are universal constants. ■

F.2. Proof of Theorem 11

Proof The proof of this result follows a structure similar to Theorem 9 and uses the same techniques as Theorem 10. We first note that the iterates of IPD and CC-RBM, with the same initialization and step-size can be expressed as,

$$\mathbf{x}_{k+1}^{i} = \mathbf{x}_{k}^{i} + \eta \mathbf{g}_{k}^{i}(\mathbf{x}_{k}^{i}) + \frac{\eta}{n}\sum_{j=1}^{n}\mathbf{K}_{k}^{ij}(\mathbf{x}_{k}^{i}, \mathbf{x}_{k}^{j}) + \sqrt{\eta}\sigma \mathbf{z}_{k}^{i}$$

$$\hat{\mathbf{x}}_{k+1}^{i} = \hat{\mathbf{x}}_{k}^{i} + \eta \mathbf{g}_{k}^{i}(\hat{\mathbf{x}}_{k}^{i}) + \frac{\eta}{n}\sum_{j=1}^{n}\mathbf{K}_{k}^{ij}(\hat{\mathbf{x}}_{k}^{i}, \hat{\mathbf{x}}_{k}^{j}) + \sqrt{\eta}\sigma \hat{\mathbf{z}}_{k}^{i}$$

where $\mathbf{z}_k^i$ and $\hat{\mathbf{z}}_k^i$ are defined as follows for any $k \in [K], n \in [n]$

$$\begin{aligned}
\mathbf{z}_{k}^{i} &= \epsilon_{k}^{i} \sim \mathcal{N}(0, \mathbf{I}), \\
\hat{\mathbf{z}}_{k}^{i} &= \frac{\sqrt{\eta}}{\sigma}\mathbf{N}_{k}^{i} + \left(\mathbf{I} - {}^{\eta}\!/\!{}_{2\sigma^2}\hat{\Sigma}_{k}^{i}\right)\hat{\epsilon}_{k}^{i}, \;\; \hat{\epsilon}_{k}^{i} \sim \mathcal{N}(0, \mathbf{I}) \\
\mathbf{N}_{k}^{i} &= \frac{1}{B}\sum_{j=1}^{B}\left[\mathbf{K}_{k}^{iI_{k}^{ij}}(\hat{\mathbf{x}}_{k}^{i}, \hat{\mathbf{x}}_{k}^{I_{k}^{ij}}) - \frac{1}{n}\sum_{l=1}^{n}\mathbf{K}_{k}^{il}(\hat{\mathbf{x}}_{k}^{i}, \hat{\mathbf{x}}_{k}^{l})\right]
\end{aligned}$$

where $I_k^{i1}, \dots, I_k^{iB} \overset{\text{iid}}{\sim} \mathsf{Uniform}([n])$ for every $i \in [n]$. Define the filtration $\mathcal{F}_k = \sigma(\hat{\mathbf{x}}_0^i, \dots, \hat{\mathbf{x}}_k^i, \hat{\mathbf{z}}_0^i, \dots, \hat{\mathbf{z}}_{k-1}^i | i \in [n])$. Clearly, $\mathbf{N}_k^i$ is an empirical average of zero-mean i.i.d random variables conditioned on $\mathcal{F}_k$. Furthermore, we note that IPD and RBM admit the same random function representation, i.e., there exists a measurable function H_K such that,

$$\begin{aligned}
(\hat{\mathbf{x}}_{1:K+1}^{i} | i \in [n]) &= H_K(\hat{\mathbf{x}}_0^i, \hat{\mathbf{z}}_0^i, \hat{\mathbf{z}}_1^i, \dots, \hat{\mathbf{z}}_K^i | i \in [n]) \\
(\mathbf{x}_{1:K+1}^{i} | i \in [n]) &= H_K(\mathbf{x}_0^i, \mathbf{z}_0^i, \mathbf{z}_1^i, \dots, \mathbf{z}_K^i | i \in [n])
\end{aligned}$$

Since Law $\left(\hat{\mathbf{x}}_0^i\right) =$ Law $\left(\mathbf{x}_0^i\right)$ for every $i \in [n]$, we use the data processing inequality and Lemma 24 to conclude the following.

$$\mathsf{KL}\left(\hat{\mathbf{x}}_{1:K+1}^{i} | i \in [n] \middle|\middle| \mathbf{x}_{1:K+1}^{i} | i \in [n]\right) = \sum_{k=0}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{z}}_k^i \mid \mathcal{F}_k\right) \middle|\middle| \mathbf{z}_k^i\right)\right]$$

We now control each term in the above summation by following the same steps as Theorem 10. To this end, we define the filtration $\mathcal{G}_k$ as $\mathcal{G}_k = \mathcal{F}_k \vee \sigma(J_k^{ij}, \bar{J}_k^{ij} : i \in [n], j \in [B'])$, where $J_k^{ij}, \bar{J}_k^{ij}$ are the additional random variables used in the estimator $\hat{\Sigma}_k^i$. We note that by Jensen's inequality $\mathbb{E}\left[\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{z}}_k^i \mid \mathcal{F}_k\right)\middle|\middle|\mathbf{z}_k^i\right)\right] \leq \mathbb{E}\left[\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{z}}_k^i \mid \mathcal{G}_k\right)\middle|\middle|\mathbf{z}_k^i\right)\right]$. Furthermore the condition $\eta \leq {}^{B\sigma^2}\!/\!{}_{40M^2 d}$ ensures that $\frac{\eta}{\sigma^2}\Sigma_k^i, \frac{\eta}{\sigma^2}\hat{\Sigma}_k^i \prec {}^{\mathbf{I}}\!/\!{}_{2}$. To this end, let $\mathbf{Z}_2 \sim \mathcal{N}(0, \mathbf{I})$ be sampled independent of everything else. Hence, from, Lemma 24,

$$\begin{aligned}
\mathsf{KL}\left(\mathrm{Law}\left(\hat{\mathbf{z}}_k^i \mid \mathcal{G}_k\right)\middle|\middle|\mathbf{z}_k^i\right) &\leq 2\mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \sqrt{\tfrac{1}{2}\mathbf{I} - \frac{\eta}{\sigma^2}\Sigma_k^i}\mathbf{Z}_2 + \frac{\sqrt{\eta}}{\sigma}\mathbf{N}_k^i\middle|\mathcal{G}_k\right) + \frac{2\eta^3}{\sigma^6}\operatorname{Tr}\left(\hat{\Sigma}_k\right)^3 \\
&\quad + 8\lambda_{\max}\left(\tfrac{\mathbf{I}}{2} - \tfrac{\eta}{\sigma^2}\Sigma\right)\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \tfrac{\eta}{\sigma^2}\hat{\Sigma}_k^i}\mathbf{Z}_2\middle|\middle|\sqrt{\tfrac{1}{2}\mathbf{I} - \tfrac{\eta}{\sigma^2}\Sigma_k^i}\mathbf{Z}_2\middle|\mathcal{G}_k\right) \\
&\leq 2\mathcal{W}_2^2\left(\frac{\mathbf{Z}_2}{\sqrt{2}}, \sqrt{\tfrac{1}{2}\mathbf{I} - \frac{\eta}{\sigma^2}\Sigma_k^i}\mathbf{Z}_2 + \frac{\sqrt{\eta}}{\sigma}\mathbf{N}_k^i\middle|\mathcal{G}_k\right) + \frac{2\eta^3}{\sigma^6}\operatorname{Tr}\left(\hat{\Sigma}_k\right)^3 \\
&\quad + 4\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \tfrac{\eta}{\sigma^2}\hat{\Sigma}_k^i}\mathbf{Z}_2\middle|\middle|\sqrt{\tfrac{1}{2}I - \tfrac{\eta}{\sigma^2}\Sigma_k^i}\mathbf{Z}_2\middle|\mathcal{G}_t\right) \\
&= \mathcal{W}_2^2\left(\mathbf{Z}_2, \sqrt{I - \frac{2\eta}{\sigma^2}\Sigma_k^i}\mathbf{Z}_2 + \frac{\sqrt{2\eta}}{\sigma}\mathbf{N}_k^i\middle|\mathcal{G}_k\right) + \frac{16\eta^3 M^6}{\sigma^6 B^3} \\
&\quad + 4\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \tfrac{\eta}{\sigma^2}\hat{\Sigma}_k^i}\mathbf{Z}_2\middle|\middle|\sqrt{\tfrac{1}{2}I - \tfrac{\eta}{\sigma^2}\Sigma_k^i}\mathbf{Z}_2\middle|\mathcal{G}_k\right) \qquad (30)
\end{aligned}$$

We note that the first term is the Wasserstein CLT term which is controlled in a manner similar to Theorem 9. Thus, for some universal constant C_{CLT},

$$\mathcal{W}_2^2\left(\mathbf{Z}_2, \sqrt{I - \frac{2\eta}{\sigma^2}\Sigma_k^i}\mathbf{Z}_2 + \frac{\sqrt{2\eta}}{\sigma}\mathbf{N}_k^i\middle|\mathcal{G}_k\right) \leq \frac{C_{\mathsf{CLT}}\eta^3 M^6 d(1+\log B)^2}{\sigma^6 B^4} \qquad (31)$$

To control the covariance mismatch term, we use the same techniques as Theorem 10. In fact, using Lemma 30 and the properties of the covariance estimator established in Lemma 31, we obtain,

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{KL}\left(\sqrt{\tfrac{1}{2}\mathbf{I} - \tfrac{\alpha}{\sigma^2}\hat{\Sigma}_k^i}\mathbf{Z}_2\middle|\middle|\sqrt{\tfrac{1}{2}\mathbf{I} - \tfrac{\alpha}{\sigma^2}\Sigma_k^i}\mathbf{Z}_2\middle|\mathcal{G}_k\right)\right] \\
&= \mathbb{E}\operatorname{tr}\left(\left(\mathbf{I} - 2\frac{\eta\Sigma_k^i}{\sigma^2}\right)^{-1}\left(\frac{\eta\Sigma_k^i}{\sigma^2} - \frac{\eta\hat{\Sigma}_k^i}{\sigma^2}\right)\right) + \sum_{k=1}^{\infty}\frac{2^{k-1}\eta^k}{\sigma^{2k}}\mathbb{E}\frac{\left[\operatorname{tr}\left((\hat{\Sigma}_k^i)^k\right) - \operatorname{tr}\left((\Sigma_k^i)^k\right)\right]}{k} \\
&= \sum_{k=2}^{\infty}\frac{2^{k-1}\eta^k}{\sigma^{2k}}\mathbb{E}\frac{\left[\operatorname{tr}\left((\hat{\Sigma}_k^i)^k\right) - \operatorname{tr}\left((\Sigma_k^i)^k\right)\right]}{k} \\
&\leq \frac{4M^4\eta^2}{B^2B'\sigma^4} + \sum_{k=3}^{\infty}\frac{2^{k-1}\eta^k}{\sigma^{2k}}\mathbb{E}\frac{\left[\operatorname{tr}\left((\hat{\Sigma}_k^i)^k\right) - \operatorname{tr}\left((\Sigma_k^i)^k\right)\right]}{k} \\
&\leq \frac{4M^4\eta^2}{B^2B'\sigma^4} + \frac{32\eta^3M^6}{3\sigma^6B^3(1 - \frac{4\eta M^2}{\sigma^2 B})} \leq C_{\mathsf{Cov}}\left[\frac{M^4\eta^2}{B^2B'\sigma^4} + \frac{\eta^3M^6}{\sigma^6B^3}\right] \qquad (32)
\end{aligned}$$

where C_{Cov} is some universal constant. Hence, from equations (30), (31) and (32), we obtain the following guarantee

$$\mathsf{KL}\left((\hat{\mathbf{x}}_k^i)_{i\in[n],k\in[K]}\middle|\middle|(\mathbf{x}_k^i)_{i\in[n],k\in[K]}\right) \lesssim \frac{\eta^2M^4nK}{B^2B'\sigma^4} + \frac{\eta^3M^6nK}{B^3\sigma^6} + \frac{d\eta^3M^6nK(1+\log B)^2}{B^4\sigma^6}$$

■

Appendix G. Discussion on Assumptions

G.1. Stochastic Gradient Growth

We discuss how Assumption 2 is actually weaker than the assumptions made in Raginsky et al. (2017) and Zou et al. (2021). Beyond this, we also demonstrate how, unlike our result, Raginsky et al. (2017) requires the stochastic gradient noise to become very small (i.e. $\tilde{O}(\epsilon^4)$) to ensure ϵ-convergence.

G.1.1. Assumption 2 and Raginsky et al. (2017)

We first compare Assumption 2 to the assumptions made in Raginsky et al. (2017). Our work analyzes SGLD with the random batch stochastic approximation, (i.e. SGLD run using mini-batch stochastic gradient estimate $\frac{1}{B}\sum_{j=1}^{B}\nabla f(\hat{\mathbf{x}}_k, \xi_{k,j})$), which is the most commonly used stochastic approximation in the literature. Moreover, both our work and Raginsky et al. (2017) consider the general stochastic problem $F(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$ (i.e. both are more general than finite-sum problems)

Under setting, we note that Raginsky et al. (2017) considers the following assumptions

- **Boundedness at a Point** $\|\nabla f(0, \xi)\| \leq B \;\forall\; \xi \in \Xi$, as implied in Assumption A.1 of Raginsky et al. (2017)
- **Component Smoothness** $\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{y}, \xi)\| \leq M\|\mathbf{x} - \mathbf{y}\|, \;\forall\; \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \xi \in \Xi$ as implied in Assumption A.2 of Raginsky et al. (2017)

Applying Assumption A.2 with $\mathbf{y} = 0$ shows that $\|\nabla f(\mathbf{x}, \xi)\| \leq M\|\mathbf{x}\| + B \;\forall\; \mathbf{x} \in \mathbb{R}^d$. Moreover, since $\nabla F(\mathbf{x}) = \mathbb{E}_\xi[\nabla f(\mathbf{x}, \xi)]$, Jensen's inequality implies that $\|\nabla F(\mathbf{x})\| = \|\mathbb{E}_\xi[\nabla f(\mathbf{x}, \xi)]\| \leq \mathbb{E}_\xi[\|\nabla f(\mathbf{x}, \xi)\|] \leq M\|\mathbf{x}\| + B$. Finally, from the triangle inequality, we conclude that $\|\nabla F(\mathbf{x}) - \nabla f(\mathbf{x}, \xi)\| \leq \|\nabla F(\mathbf{x})\| + \|\nabla f(\mathbf{x}, \xi)\| \leq 2M\|\mathbf{x}\| + 2B$. This is identical to Assumption 2 (with $M \leftrightarrow 2M$ and $G \leftrightarrow 2B$). Thus, Assumption A.1 and A.2 of Raginsky et al. (2017) imply Assumption 2.

Along similar lines, one can show that Assumption 2 is more general than that of Zou et al. (2021). Firstly, **Zou et al. (2021) consider the finite-sum problem**, i.e., $F(\mathbf{x}) = {}^1\!/\!_n \sum_{i\in[n]} f_i(\mathbf{x})$, whereas we consider the general stochastic problem $F(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$. Moreover, Assumption 4.4 of Zou et al. (2021) assumes **component smoothness**, i.e., $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \;\forall\; i \in [n], \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Lastly, Zou et al. (2021) uniformly bound the gradient at a point, i.e., their proof relies on the fact that there exists a constant G such that $\|\nabla f_i(0)\| \leq G \;\forall\; i \in [n]$ (see Lemma 6.2 in Zou et al. (2021) and also their proof of Theorem 4.5 in Appendix A.1 which uses Lemma 6.2).

Under this setting, one can apply the same arguments as above to show that Assumption 2 of our work is weaker than the assumptions in Zou et al. (2021). In particular, the component smoothness and boundedness at a point assumptions imply $\|\nabla f_i(\mathbf{x})\| \leq L\|\mathbf{x}\| + G \;\forall\; i \in [n], \mathbf{x} \in \mathbb{R}^d$. Then, one can apply triangle inequality and Jensen's inequality to show that $\|\nabla F(\mathbf{x}) - \nabla f_i(\mathbf{x})\| \leq 2L\|\mathbf{x}\| + 2G \;\forall\; i \in [n], \mathbf{x} \in \mathbb{R}^d$.

We note that, unlike Raginsky et al. (2017); Zou et al. (2021) our work does not make any component smoothness or uniform boundedness at a point assumptions.

G.1.2. MAGNITUDE OF STOCHASTIC GRADIENT NOISE IN RAGINSKY ET AL. (2017)

In Assumption A.4 of their paper, Raginsky et al. (2017) assume that, for any $\mathbf{x} \in \mathbb{R}^d$, **the variance of the stochastic gradient oracle at x is bounded by** $2\delta(M^2\|\mathbf{x}\|^2 + B^2)$**.**

Beyond Assumption A.4, Raginsky et al. (2017) also assume component dissipativity (i.e., dissipativity of each $f(\mathbf{x}, \xi)$, see Assumption A.3), component smoothness and boundedness at a point (discussed above), to prove the following convergence guarantee for SGLD in Wasserstein-2 distance (see Proposition 3.3 of Raginsky et al. (2017)).

$$\mathcal{W}_2(\mathrm{Law}\left(\hat{\mathbf{x}}_K\right), \pi^*) \leq C_0 K\eta\delta^{1/4} + C_1 K\eta^{5/4} + C_2 e^{-K\eta/\beta c_{\mathsf{LS}}} \tag{33}$$

where $C_0, C_1, C_2, c_{\mathsf{LS}}$ are problem dependent constants. Consider any $\epsilon \geq 0$. To ensure $\mathcal{W}_2(\mathrm{Law}\left(\hat{\mathbf{x}}_K\right), \pi^*) \leq O(\epsilon)$, one must ensure each term in (33) is $O(\epsilon)$. To set $C_2 e^{-K\eta/\beta c_{\mathsf{LS}}} \leq O(\epsilon)$, η and K must satisfy $\eta K \gtrsim \ln(1/\epsilon)$. This implies that $K\eta\delta^{1/4} \gtrsim \delta^{1/4}\ln(1/\epsilon)$. However, to ensure ϵ-convergence, $K\eta\delta^{1/4}$ must be $O(\epsilon)$ which implies $\delta^{1/4}\ln(1/\epsilon) \leq O(\epsilon)$ i.e. $\delta \leq O(\epsilon^4)$

Note that since δ effectively controls the variance of the stochastic gradient oracle, $\delta \leq O(\epsilon^4)$ means that the strength of the stochastic gradient noise must be very small to ensure convergence. On the contrary, in Assumption 2 (and in its relaxation in Assumption 3), we allow $M = O(1), G = O(\sqrt{d})$ and only require a constant $O(\sqrt{d})$ (or $O(d^{1/3})$ for CC-SGLD) batch size in all our results. Hence, our results ensure ϵ-convergence without needing the strength of the stochastic gradient noise to diminish with ϵ.

G.2. Moment Bounds

We now demonstrate that dissipativity and smoothness of F along with the growth condition on stochastic gradients implies p-moment growth. This result is a straightforward adaptation of Lemma 3.2 of Raginsky et al. (2017) under relaxed conditions (as highlighted above, the growth condition is weaker than Assumptions A.1 and A.2 of Raginsky et al. (2017). Moreover. Raginsky et. al. assume dissipativity for each component $f(\mathbf{x}, \xi)$ while we only assume it for F).

Suppose F is L-smooth and (m, b) dissipative, i.e., $\langle \nabla F(\mathbf{x}), \mathbf{x}\rangle \geq m\|\mathbf{x}\|^2 - b$. Also assume that lin-growth a.s is satisfied (the proof holds even with lin-growth subG). For convenience, assume $\nabla F(0) = 0$, i.e., 0 is a stationary point of F (the result holds even if we consider any arbitrary stationary point $\mathbf{x}^* \neq 0$). Recall that the trajectory $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_K$ follows the update rule

$$\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \eta\mathbf{g}_k + \sqrt{2\eta}\epsilon_k, \quad \mathbf{g}_k = \tfrac{1}{B}\sum_{j=1}^{B}\nabla f(\hat{\mathbf{x}}_k, \xi_{k,j}) \quad \epsilon_k \sim \mathcal{N}(0, \mathbf{I})$$

Let $\hat{\mathbf{y}}_k = \hat{\mathbf{x}}_k - \eta\mathbf{g}_k$. Using the fact that $\epsilon_k \sim \mathcal{N}(0, \mathbf{I})$

$$\mathbb{E}[\|\hat{\mathbf{x}}_{k+1}\|^2 | \hat{\mathbf{x}}_k] = \mathbb{E}\left[\|\hat{\mathbf{y}}_k\|^2 + \sqrt{2\eta}\left\langle \epsilon_k, \mathbf{g}_k\right\rangle + 2\eta\|\epsilon_k\|^2 | \hat{\mathbf{x}}_k\right] = \mathbb{E}[\|\hat{\mathbf{y}}_k\|^2 | \hat{\mathbf{x}}_k] + 2\eta d \tag{34}$$

Let $\mathbf{N}_k = \mathbf{g}_k - \nabla F(\hat{\mathbf{x}}_k)$. It follows that, $\|\hat{\mathbf{y}}_k\|^2 = \|\hat{\mathbf{x}}_k\|^2 - 2\eta \langle \mathbf{g}_k, \hat{\mathbf{x}}_k \rangle + \eta^2 \|\mathbf{g}_k\|^2$. Moreover, $\mathbb{E}[\mathbf{g}_k|\hat{\mathbf{x}}_k] = \nabla F(\hat{\mathbf{x}}_k)$, and, by the L-smooth and lin-growth a.s conditions,

$$\mathbb{E}[\|\mathbf{g}_k\|^2|\hat{\mathbf{x}}_k] \leq \|\nabla F(\hat{\mathbf{x}}_k)\|^2 + 2(M^2\|\hat{\mathbf{x}}_k\|^2 + G^2) \leq (L^2 + 2M^2)\|\hat{\mathbf{x}}_k\|^2 + 2G^2$$

Hence,

$$\begin{aligned} \mathbb{E}[\|\hat{\mathbf{y}}_k\|^2|\hat{\mathbf{x}}_k] &\leq (1 + \eta^2 L^2 + 2\eta^2 M^2)\|\hat{\mathbf{x}}_k\|^2 - 2\eta \langle \nabla F(\hat{\mathbf{x}}_k), \hat{\mathbf{x}}_k \rangle + 2\eta^2 G^2 \\ &\leq (1 - 2\eta m + \eta^2 L^2 + 2\eta^2 M^2)\|\hat{\mathbf{x}}_k\|^2 + 2\eta^2 G^2 + 2\eta b \end{aligned} \tag{35}$$

where the last inequality follows from dissipativity. Now, setting $\eta \leq \frac{m}{L^2 + 2M^2}$, we conclude from (34) and (35),

$$\mathbb{E}[\|\hat{\mathbf{x}}_{k+1}\|^2] \leq (1 - \eta m)\mathbb{E}[\|\hat{\mathbf{x}}_k\|^2] + 2\eta d + 2\eta b + 2\eta^2 G^2$$

Unrolling the above recurrence and using we obtain,

$$\mathbb{E}[\|\hat{\mathbf{x}}_k\|^2] \leq 2d/m + 2b/m + 2\eta G^2/m \lesssim d \; \forall \; k \in [K]$$

From the above and (35) it also follows that $\mathbb{E}[\|\hat{\mathbf{y}}_k\|^2] \lesssim d \; \forall \; k \in [K]$. Thus, we have established p-moment growth for $p = 2$. We follow a similar procedure for $p = 4$. In particular, expanding powers and taking expectations wrt ϵ_k gives us

$$\begin{aligned} \mathbb{E}[\|\hat{\mathbf{x}}_{k+1}\|^4] &\leq \mathbb{E}[\|\hat{\mathbf{y}}_k\|^4] + 8\eta \mathbb{E}[\|\hat{\mathbf{y}}_k\|^2] + 8\eta^2 d^2 + 4\eta d \mathbb{E}[\|\hat{\mathbf{y}}_k\|]^2 \\ &\leq \mathbb{E}[\|\hat{\mathbf{y}}_k\|^4] + C\left(\eta d + \eta d^2 + \eta^2 d^2\right) \end{aligned} \tag{36}$$

where $C \geq 0$ is an absolute numerical constant. The last inequality uses the fact that $\mathbb{E}[\|\hat{\mathbf{y}}_k\|^2] \lesssim d$. As before, using $\hat{\mathbf{y}}_k = \hat{\mathbf{x}}_k - \eta \nabla F(\hat{\mathbf{x}}_k) - \eta \mathbf{N}_k$, using the dissipativity and moment growth conditions and expanding powers, we get,

$$\|\hat{\mathbf{y}}_k\|^4 \leq \left[1 - 4\eta m + c_1 \eta^2 (L^2 + M^2) + c_2 \eta^4 (L^2 + M^2)^2\right] \|\hat{\mathbf{x}}_k\|^4 + (c_3 \eta^2 G^2 + c_4 \eta b)\|\hat{\mathbf{x}}_k\|^2 + c_5 \eta^4 G^4$$

where $c_1, \dots, c_5$ are universal constants. Taking expctations and using the fact that $\mathbb{E}[\|\hat{\mathbf{x}}_k\|^2] \lesssim d$, we conclude from the above inequality and (36) that,

$$\mathbb{E}[\|\hat{\mathbf{x}}_{k+1}\|^4] \leq \left[1 - 4\eta m + c_1 \eta^2 (L^2 + M^2) + c_2 \eta^4 (L^2 + M^2)^2\right] \mathbb{E}[\|\hat{\mathbf{x}}_k\|^4] + C(\eta d + \eta d^2 + \eta^2 d^2 + \eta^4 G^4)$$

Setting $\eta \leq \frac{m}{c_6 (L^2 + M^2)}$ for some large enough universal constant c_6 and unrolling the above recurrence, we conclude $\mathbb{E}[\|\hat{\mathbf{x}}_k\|^4] \lesssim d^2 \; \forall k \in [K]$. A similar procedure can be followed for $p = 6, 8, \dots$ to show that for any even p, the p-moment growth condition holds under some appropriate choice of η. Extension to odd p follows by an application of Jensen's inequality (i.e. $\mathbb{E}[\|\hat{\mathbf{x}}_k\|^{2m-1}] \leq (\mathbb{E}[\|\hat{\mathbf{x}}_k\|^{2m}])^{2m - 1/2m} \lesssim d^{m - 1/2}$)

Finally, since dissipativity is a weaker condition than strong convexity outside of a compact set (as per the discussion in Cheng et al. (2020), Section 2.1), the above proof also establishes that strong convexity outside of a compact set implies p-moment growth