# Non-asymptotic Results for Langevin Monte Carlo: Coordinate-wise and Black-box Sampling

Lingqing Shen*[1], Krishnakumar Balasubramanian†[2], and Saeed Ghadimi‡[3]

[1]Department of Mathematics, Nanjing University, Jiangsu, China
[2]Department of Statistics, University of California, Davis
[3]Department of Operations Research and Financial Engineering, Princeton University

March 13, 2019

**Abstract**

Discretization of continuous-time diffusion processes, using gradient and Hessian information, is a popular technique for sampling. For example, the Euler-Maruyama discretization of the Langevin diffusion process, called as Langevin Monte Carlo (LMC), is a canonical algorithm for sampling from strongly log-concave densities. In this work, we make several theoretical contributions to the literature on such sampling techniques. Specifically, we first provide a Randomized Coordinate-wise LMC algorithm suitable for large-scale sampling problems and provide a theoretical analysis. We next consider the case of zeroth-order or black-box sampling where one only obtains evaluates of the density. Based on Gaussian Stein's identities we then estimate the gradient and Hessian information and leverage it in the context of black-box sampling. We then provide a theoretical analysis of gradient and Hessian based discretizations of Langevin and kinetic Langevin diffusion processes for sampling, quantifying the non-asymptotic accuracy. We also consider high-dimensional black-box sampling under the assumption that the density depends only on a small subset of the entire coordinates. We propose a variable selection technique based on zeroth-order gradient estimates and establish its theoretical guarantees. Our theoretical contributions extend the practical applicability of sampling algorithms to the large-scale, black-box and high-dimensional settings.

*lingqing.shen@smail.nju.edu.cn. Work done while visiting University of California, Davis as an exchange student.

†kbala@ucdavis.edu

‡sghadimi@princeton.edu

# 1 Introduction

Sampling and optimization are the computational backbones of Bayesian and Frequentist statistics respectively. Motivated by the need to speed-up Bayesian inference for large scale datasets, there has recently been an increased interest on developing faster algorithms for sampling with strong theoretical guarantees. Such techniques are invariably based on techniques from optimization. Indeed there is a strong interplay between the problems of sampling and optimization. Let $f(\theta) : \mathbb{R}^d \to \mathbb{R}$ be a function and let $\pi(\theta)$ be a density function defined as follows:

$$\pi(\theta) = \frac{e^{-f(\theta)}}{\int_{\mathbb{R}^d} e^{-f(r)}\, dr} \tag{1}$$

The problem of sampling involves generating a random vector that is distributed according to the above target density. The closely related optimization problem involves finding a minimum point $\theta^*$ of the function $f(\theta)$, i.e.,

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \;\; f(\theta). \tag{2}$$

Define the function $f_\tau(\theta) = f(\theta)/\tau$, for some $\tau > 0$. Then, $\theta^*$ is also the minimum point of $f_\tau(\theta)$. If we define $\pi_\tau(\theta) \propto e^{-f_\tau(\theta)}$, then as $\tau$ goes to zero: (i) The expectation $\bar{\theta}_\tau = \int_{\mathbb{R}^d} \theta \pi_\tau(\theta)\, d\theta$, converges to the minimum $\theta^*$ and (ii) The distribution $\pi_\tau(d\theta)$ converges to the Dirac measure centered at $\theta^*$. As a straightforward example, let $d = 1$ and consider $f(\theta) = (\theta - a)^2$, for some constant $a > 0$. Then clearly $\theta^* = a$. If we construct the density $\pi_\tau(\theta) \propto e^{-\frac{(\theta-a)^2}{\tau}}$, which is a Gaussian density, then the expectation clearly is $a$. As the variance term $\tau \to 0$, $\pi_\tau(\theta)$ converges to a Dirac measure centered at $a$. This highlights the interplay between sampling and optimization.

First generation sampling algorithms, for example, Metropolis-Hastings algorithm are oblivious to the geometry of the target density as a result of which they suffer from slower rate of convergence. But they are often easy to implement and are just based on function evaluations – hence they could be referred to as *zeroth-order sampling algorithms*. See [4, 24, 27, 28, 30], for more details about such algorithms. Motivated by statistical physics principles, various researchers developed faster sampling algorithms that leverage the geometric information regarding the target density [3, 22, 33, 40, 41, 45, 46]. Such algorithms, for example Langevin and Hamiltonian Monte Carlo, are based on first-order discretization of a continuous-time diffusion process and could be referred to as *first-order sampling algorithms* as they leverage gradient information about the target density. Although such algorithms were developed over a decade ago, recently strong theoretical guarantees have been established for sampling in the works of [11, 12, 13, 14, 16, 17, 19] and several others. Such algorithms achieve significantly faster rates of convergence compared to the zeroth-order sampling techniques. Furthermore, a close connection could be established between the above non-asymptotic results and the corresponding results from the first-order optimization literature, as described in [14] (see also [29] for related interplay between sampling and optimization). In this work, we further explore the connections between optimization with various oracle information and sampling based on various discretizations of continuous-time diffusion process.

### 1.1 Preliminaries

Consider the continuous-time Langevin diffusion process $\{L_T : T \in \mathbb{R}_+\}$ given by the following stochastic differential equation,

$$dL_T = -\nabla f(L_T)dT + \sqrt{2}dW_T \tag{3}$$

where $T \in \mathbb{R}_+$ and $\{W_T : T \in \mathbb{R}_+\}$ is a $d$-dimensional Brownian motion and $\nabla f(\theta) \in \mathbb{R}^d$ denotes the gradient of $f(\theta)$. The Euler-Maruyama discretization of the above process is given by the following Markov chain:

$$x_{t+1,h} = x_{t,h} - h_{t+1}\ \nabla f(x_{t,h}) + \sqrt{2h_{t+1}}\varepsilon_{t+1} \tag{4}$$

for the discrete time index $t = 0, 1, 2 \ldots$. Here $\varepsilon_t \in \mathbb{R}^d$ is a standard Gaussian noise vector, $h > 0$ denotes the step-size and an initial point $x_{0,h}$ is assumed to be given. The above discretization is called as the Langevin Monte Carlo (LMC) sampling algorithm. The update step of the LMC sampling algorithm shares similarity with the standard gradient descent algorithm from the optimization literature. Denote the distribution of the random vector $x_{t,h}$ by $\varpi_t$. To evaluate the performance of the sampling algorithm, the 2-Wasserstein distance between $\varpi_t$ and the target density $\pi(\theta)$ is considered. For measures, $p$ and $q$ defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, the 2-Wasserstein distance is defined as:

$$W_2(p, q) \stackrel{def}{=} \left( \inf_{\varrho \in \varrho(p,q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta - \theta'\|_2^2 \, d\varrho(\theta, \theta') \right)^{1/2}, \tag{5}$$

where $\varrho(p, q)$ is the set of joint distribution that has $p$ and $q$ as its marginals. The performance of the sampling updates is measured by the above 2-Wasserstein distance between the distribution $\varpi_t$ and the target density $\pi$, i.e., $W_2(\varpi_t, \pi)$. In order to obtain theoretical guarantees, a common assumption made in the literature on LMC is that the function $f$ is smooth and strongly convex.

**Assumption 1.1** *Let $\|\cdot\| = \|\cdot\|_2$ denote the Euclidean norm on $\mathbb{R}^d$. Then the function $f$:*

1. ***A1**: has Lipschitz continuous gradient, i.e., $\|\nabla f(\theta) - \nabla f(\theta')\| \leq M\|\theta - \theta'\|$ for $M > 0$.*
2. ***A2:** is strongly convex i.e., $f(\theta) - f(\theta') - \nabla f(\theta')^\top(\theta - \theta') \geq \frac{m}{2}\|\theta - \theta'\|^2$, for $m > 0$.*

Assuming access to inaccurate gradients, [14] provide theoretical guarantees for sampling under Assumption 1.1. Specifically, instead of the true gradient $\nabla f(x_{t,h})$ in each step, it is assumed that we observe $g_{t,h} = g(x_{t,h}) = \nabla f(x_{t,h}) + \zeta_t$, for a sequence of random noise vectors $\zeta_t$ that satisfies certain bias and variance assumption. Then, the noisy LMC updates corresponds to the case of the updates in Equation 4, with $\nabla f(x_{t,h})$ replaced by $g_{t,h}$. For such an update, [14] have the following non-asymptotic result.

**Theorem 1.2** *[14] Assume that the bias and variance of $\zeta_t$ satisfies respectively, for all $t = 1, 2, \ldots$,*

$$\mathbf{E}[\|\mathbf{E}(\zeta_t|x_{t,h})\|_2^2] \leq \delta^2 d \qquad \text{and} \qquad \mathbf{E}[\|\zeta_t - \mathbf{E}(\zeta_t|x_{t,h})\|_2^2] \leq \sigma^2 d.$$

*Let the function $f$ satisfy Assumption 1.1. If $h \leq 2/(m + M)$, the following result holds true.*

$$W_2(\varpi_t, \pi) \leq (1 - mh)^t W_2(\varpi_0, \pi) + 1.65\frac{M}{m}(hd)^{1/2} + \frac{\delta\sqrt{d}}{m} + \frac{\sigma^2(hd)^{1/2}}{1.65M + \sigma\sqrt{m}}.$$

**Remark 1** *More generally, if the bounded bias and variance condition are changed to*

$$\mathbf{E}[\|\mathbf{E}(\zeta_t|x_{t,h})\|_2^2] \leq \delta^2 d^\alpha \qquad \text{and} \qquad \mathbf{E}[\|\zeta_t - \mathbf{E}(\zeta_t|x_{t,h})\|_2^2] \leq \sigma^2 d^\beta,$$

*respectively, for some $\alpha, \beta > 0$, the conclusion turns into*

$$W_2(\varpi_t, \pi) \leq (1 - mh)^t W_2(\varpi_0, \pi) + \frac{1.65M(hd)^{1/2}}{m} + \frac{\delta d^{\alpha/2}}{m} + \frac{\sigma^2 h d^\beta}{1.65M(hd)^{1/2} + \delta d^{\alpha/2} + \sigma (mh)^{1/2} d^{\beta/2}}.$$

*Furthermore, in the case that $\beta > \max\{1, \alpha\}$, the last term is dominated by $d^{\beta/2}$.*

**Remark 2** *[14] One could also recover the optimization corresponding to the standard gradient descent algorithm for minimizing strongly-convex function from Theorem 1.2. In order to see that, consider the function $f_\tau(\theta) = f(\theta)/\tau$ as before. Then, $f_\tau$ also satisfies Assumption 1.1 with $m_\tau = m/\tau$ and $M_\tau = M/\tau$. With the true gradient, (i.e., $\delta = \sigma = 0$), we then have from Theorem 1.2 that*

$$W_2(\varpi_t, \pi_\tau) \leq \left(1 - \frac{m}{M}\right)^t W_2\left(\delta_{\theta_0}, \pi_\tau\right) + 1.65 \left(\frac{M}{m}\right) \left(\frac{d\tau}{M}\right)^{1/2}$$

*As we let $\tau \to 0$, we have the LMC updates converging to the standard gradient descent updates and the above bound becomes*

$$\|\theta_t - \theta^*\|_2 \leq \left(1 - \frac{m}{M}\right)^t \|\theta_0 - \theta^*\|_2.$$

## 1.2 Our Contributions

Despite the impressive set of theoretical results in [11, 12, 13, 14, 15, 16, 17, 18, 19, 37], there are several avenues for improvement to develop practical sampling algorithms with strong guarantees. Motivated by oracle models in optimization, in this work, we make a distinction between the availability of information regarding $f(\theta)$ for sampling. Specifically, in a zeroth-order (or black-box) sampling setting, we only observe (potentially noisy) evaluations of the function $f$. Similarly, in the first- and second-order setting, we observe (potentially noisy) evaluations of the gradient and Hessian of $f(\theta)$ respectively. In this work, we make the following contributions to the literature on sampling.

1. We first consider the first-order LMC sampling and propose and analyze a Randomized Coordinate Descent based LMC (RCD-LMC) update rule for large-scale sampling where updating the entire gradient in each iteration might be computationally demanding. We establish its rate of convergence, from which the corresponding results in the optimization literature for Randomized Coordinate Descent optimization algorithm could be recovered.

2. We next consider the zeroth-order or black-box LMC sampling. Although studied as early as [10] and [32], recently, it has attracted much attention motivated by several statistical machine learning models [26, 38]. Furthermore, in several situations, one might be interested in sampling from a point-wise consistently estimated log-concave density (see [42] for a recent survey on estimating log-concave densities), where naturally the density is not available to us analytically. In the above mentioned situations, one could obtain (potentially noisy)

evaluations of the function $f$ though no analytical expression is available for the function $f$. Using the idea of Gaussian-smoothing based zeroth-order optimization [1, 21, 35], we propose and analyze Zeroth-Order LMC algorithm (ZO-LMC) and establish its theoretical properties.

3. Next, we consider the case of high-dimensional zeroth-order sampling. We specifically assume the unobserved function $f$ is sparse in the sense that it depends only on $s$ of the $d$ coordinates. We provide a variable selection method based on the estimated gradient, which in conjunction with the Zeroth-Order LMC algorithms reduces the rates of convergence to be only poly-logarithmically dependent on the dimensionality $d$ thereby enabling high-dimensional sampling.

4. We next consider Ozaki-discretized LMC updates which involves the Hessian of the function $f(\theta)$. Note that [14] proposed theoretical guarantees for sampling with Ozaki-discretization, from which the corresponding results of the Newton method for optimization could be recovered. But [14] assumed the availability of exact gradients and Hessians. In this work, we first consider the case of inexact gradients and Hessians and extend the results of [14] to this setting. We then consider the case of Zeroth-Order Ozaki discretized LMC (ZOO-LMC) for the case of black-box sampling. Our method is based on a novel technique of estimating the Hessian of a function from just function queries, based on Gaussian Stein's identity proposed recently in [1]. For this case, we also develop corresponding theoretical results and discuss its consequences.

5. Finally, we consider kinetic Langevin diffusions [5, 33] and their first-order and second-order discretization considered in [11, 15] and establish theoretical guarantees with inaccurate gradient and Hessian information. Similar to the previous results, we also establish zeroth-order extensions of the above discretizations and establish the corresponding theoretical properties.

Our results are summarized in Table 1. A list of notations used in the paper is provided in Section A. All proofs are relegated to the appendix Sections B - G.

## 2 Randomized Coordinate Descent LMC Sampling

In this section, we propose and analyze coordinate descent based Langevin Monte Carlo sampling algorithm. In modern large-scale problems, the cost of computing and updating the entire gradient in each update step of LMC algorithm might be prohibitive. Hence, a practical remedy is to update only one coordinate (or a batch of coordinates) at a time. Indeed, such coordinate descent algorithms are popular in the optimization literature to deal with large-scale problems when the function $f$ has special structures [49]. We specifically analyze randomized coordinate descent updates in the context of sampling and provide rates of convergence in 2-Wasserstein distance. For a vector $a \in \mathbb{R}^d$, denote by $a_i$, the $i$-th coordinate. Then the Randomize Coordinate-wise Descent LMC (RCD-LMC) is defined by the following update step:

$$x_{t+1,h} = x_{t,h} - h[\nabla f(x_{t,h})]_{i_t} e_{i_t} + \sqrt{2h}(\varepsilon_{t+1})_{i_t} e_{i_t}. \tag{6}$$

| Discretization | Information | Structure | References |
|---|---|---|---|
| EM | Exact $1^{st}$-order | GS & SC | [14, 18] |
| EM | Inexact $1^{st}$-order | GS & SC | [14] |
| Ozaki | Exact $1^{st}$ and $2^{nd}$ order | HS & SC | [14] |
| Coordinate-wise EM | Exact $1^{st}$-order | GS & SC | Theorem 2.2 |
| EM | Exact $0^{th}$-order | GS & SC | Theorem 3.1 |
| Variable selection+EM | Exact $0^{th}$-order | GS & SC | Theorem 4.3 |
| Ozaki | Inexact $1^{st}$ and $2^{nd}$ order | HS & SC | Theorem 5.3 & 5.4 |
| Ozaki | Exact $0^{th}$ order | HS & SC | Theorem 5.6 |
| Kinetic | Exact $1^{st}$-order | GS & SC | [11, 15] |
| Kinetic-2 | Exact $1^{st}$ and $2^{nd}$ order | HS & SC | [15] |
| Kinetic | Exact $0^{th}$ order/Inexact $1^{st}$ and $2^{nd}$ | GS & SC | Theorem 6.1 |
| Kinetic-2 | Exact $0^{th}$ order/Inexact $1^{st}$ and $2^{nd}$ | HS & SC | Theorem 6.2 |

Table 1: A list of complexity results for sampling based on discretizing continuous-time diffusion processes. EM stands for Euler-Maruyama discretization. GS and HS stands for Gradient- and Hessian-smoothness assumptions respectively. SC stands for Strongly-convex.

In each time step $t$, we randomly pick a coordinate and compute and update the gradient only corresponding to that coordinate. Clearly, this is much faster than computing the full gradient in each step. The choice of distribution over the coordinate based on which the updates are done is typically fixed to be uniform distribution in practice. We also make the following coordinate wise Lipschitz assumption on the function $f$, as is commonly done in the analysis of coordinate descent algorithms in the optimization setting [34, 49].

**Assumption 2.1** *We assume that $\nabla f$ is coordinate Lipschitiz with constants $M_i$, i.e.,*

$$|[\nabla f(\theta + se_i)]_i - [\nabla f(\theta)]_i| \leq M_i s.$$

Denote $M_{\max} = \max_{1 \leq i \leq d} M_i$. Then, $1 \leq \dfrac{M}{M_{\max}} \leq d$, as can be seen intuitively by relating Lipschitz constants to the Hessian $\nabla^2 f$; see also [49]. Under the above assumption, we have the following theorem characterizing the theoretical performance of RCD-LMC.

**Theorem 2.2** *Let the function satisfy part (**A1**) in Assumption 1.1 and Assumption 2.1. Then we have, for all $h \leq 2/(m+M)$*

$$\mathbf{E}[W_2(\varpi_t, \pi)] \leq \left(1 - \frac{m}{2d}h\right)^t W_2(\varpi_0, \pi) + \frac{7\sqrt{2}}{3}\frac{M_{\max}}{m(1-mh)} h^{1/2} d^{3/2},$$

*where the expectation is taken with respect to the choice of coordinates sampled.*

**Remark 3** *One can compare the above result with the corresponding bound for the full-gradient based LMC algorithm (Theorem 1.2 with $\sigma = \delta = 0$). Recall that, $M \leq dM_{\max}$. Hence, we get comparable result in the worst case of $M = dM_{\max}$. But in the typical case of $1 \leq M \ll dM_{\max}$, we see the effect of using updating only one-coordinate of the gradient at a time compared to the true full-gradient.*

**Remark 4** *Consider the function $f_\tau$ from Section 1 and recall that $\theta^*$ is the minimizer of function $f_\tau$ or $f$. If $m$ and $h$ are replaced by $\frac{m}{\tau}$ and $\frac{\tau}{M_{\max}}$ respectively, and if we let $\tau \to 0$, we obtain the following result:*

$$\mathbf{E}[\|x_{t,h} - x^*\|_2] \leq \left(1 - \frac{m}{2dM_{\max}}\right)^t \|x_0 - x^*\|_2.$$

*This result recovers the corresponding result from the optimization literature for randomized coordinate descent [34, 49], which reads as*

$$\mathbf{E}[f(x_{t,h})] - f(x^*) \leq \left(1 - \frac{m}{dM_{\max}}\right)^t (f(x_0) - f(x^*)).$$

## 3 Black-box Sampling via Zeroth-Order LMC

In this section, we consider the problem of zeroth-order or black-box sampling. In this situation, the function $f$ is not observed analytically, but one can obtain (potentially noisy) function evaluation for any query point. This situation occurs, for example, in several statistical machine learning models where describing $f(\theta)$ analytically is prohibitive. We refer the reader to [26, 38, 39] for examples of such problems occurring in practice. In order to proceed, we first estimate the gradient of the function from function queries. We leverage the Gaussian smoothing technique [1, 21, 35], popular in the field of zeroth-order optimization. Specifically, for a point $\theta \in \mathbb{R}^d$, we define an estimate $g_{\nu,n}(\theta)$, of the gradient $\nabla f(\theta)$ as follows:

$$g_{\nu,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\theta + \nu u_i) - f(\theta)}{\nu} u_i \tag{7}$$

where $u_i \sim N(0, I_d)$ and are i.i.d. The gradient estimator in Equation 7 is a consequence of Gaussian Stein's identity, popular in the statistics literature [44]. For a more detailed discussion, we refer the reader to [1]. Based on the above estimate of the gradient, we have the following Zeroth-Order LMC (ZO-LMC) algorithm for black-box sampling, which has the following update steps:

$$x_{t+1,h} = x_{t,h} - h\, g_{\nu,n}(x_{t,h}) + \sqrt{2h}\varepsilon_{t+1} \tag{8}$$

for $t = 0, 1, 2, \ldots$. We then have $g_{\nu,n}(\theta) = \nabla f(\theta) + \zeta$ for some noise vector $\zeta$. This equation has the same form as the form of inaccurate gradient (specifically, Equation above (13)) assumed in [14]. But the corresponding result from [14] cannot be used directly in our setting, as the variance of the gradient is not bounded unless we make restrictive assumptions on the gradient of $f$. We now state the main result of this section.

**Theorem 3.1** *Let the function $f$ satisfy Assumption 1.1. Then we have, for $h \leq 2/(m+M)$ and $n$ satisfying $\frac{h}{n(1-mh)} \leq \frac{m}{2M^2(d+1)}$,*

$$W_2(\varpi_t, \pi) \leq (1 - 0.5mh)^t\, W_2(\varpi_0, \pi) + 3.3\frac{M}{m}(hd)^{1/2} + 2\frac{M}{m}\nu d^{1/2} + 2\sqrt{2}\frac{\sqrt{M}}{\sqrt{m}} \cdot \frac{1}{\sqrt{n}} h^{1/2}(d+1) + \sqrt{2}\frac{M}{\sqrt{m}} \cdot \frac{\nu^2}{\sqrt{n}} h^{1/2}(d+2)^{3/2}.$$

**Remark 5** *Recall that for the exact first-order based LMC algorithm, by the right choice of tuning parameters,* $t = \mathcal{O}\left(d/\epsilon^2 \cdot \log(d/\epsilon)\right)$ *suffices for* $W_2(\varpi_t, \pi) \leq \epsilon$*; see [14]. For the ZO-LMC updates, for* $W_2(\varpi_t, \pi) \leq \epsilon$ *when we require* $t = \mathcal{O}\left(d/\epsilon^2 \cdot \log(d/\epsilon)\right)$*, setting* $n = \mathcal{O}(d), \nu = \mathcal{O}\left(\epsilon/\sqrt{d}\right)$*,* $h = \mathcal{O}\left(\epsilon^2/d\right)$ *suffices. With* $n = 1$*, it suffices to have* $\nu = \mathcal{O}\left(\epsilon/\sqrt{d}\right)$*,* $h = \mathcal{O}\left(\epsilon^2/d^2\right)$*. Thus, with the appropriate choice of tuning parameters, ZO-LMC matches the performance of LMC which requires gradient information.*

# 4 Variable Selection for High-dimensional Black-box Sampling

In practical black-box settings, due to the non-availability of the analytical form of $f(\theta)$, one might potentially over-parametrize $f(\theta)$, in terms of number of covariates selected for modeling. Hence, the problem of variable selection, in a zeroth-order setting becomes crucial. To address this issue, in this section, we study variable selection under certain sparsity assumptions on the objective function $f$, to facilitate sampling in high-dimensions. Specifically, we make the following assumption on the structure of $f$.

**Assumption 4.1** *We assume that* $f(\theta) : \mathbb{R}^d \to \mathbb{R}$ *is* $s$ *sparse, i.e., the function* $f$ *depends only on (the same)* $s$ *of the* $d$ *coordinates, for all* $\theta$*, where* $s \ll d$*. We denote the true support set as* $S^*$*. This implies that for any* $\theta \in \mathbb{R}^d$*, we have* $\|\nabla f(\theta)\|_0 \leq s$*, i.e., the gradient is* $s$*-sparse. Furthermore, define* $\nabla f_\nu(\theta) = \mathbf{E}_u\left[\nabla f(\theta + \nu u)\right]$ *for a standard gaussian random vector* $u$*. Then the gradient sparsity assumption also implies that* $\|\nabla f_\nu(\theta)\|_0 \leq s$ *for all* $\theta \in \mathbb{R}^d$*. Furthermore, we assume that the gradient lies in the following set that characterizes the minimal signal strength in the relevant coordinates of the gradient vector:*

$$\mathcal{G}_{a,s} = \left\{ \nabla f(\theta) : \|\nabla f(\theta)\|_0 \leq s \text{ and } \sup_{\theta \in \mathbb{R}^d} \inf_{j \in S^*} |[\nabla f(\theta)]_j| \geq a \right\}$$

As a consequence, we also have that $\nabla f_\nu(\theta) \in \mathcal{G}_{a,s}$. The above assumption makes a *homogenous* sparsity assumption on the sparsity and the minimum signal strength of the gradient. Roughly speaking, $a$ represents the minimum signal strength in the gradient so that efficient estimation of the support $S^*$ is possible in the sample setting. The above sparsity model on the function $f$, converts the problem to variable selection in a non-Gaussian sequence model setting:

$$[g_{\nu,n}]_j = [\nabla f_\nu(\theta)]_j + \zeta_j \qquad j = 1, \ldots, d.$$

Hence, $\zeta_j$ are zero-mean random variables as $[g_{\nu,n}]_j$ is an unbiased estimator of $[\nabla f_\nu(\theta)]_j$. We refer the reader to [9] for recent results on variable selection consistency in Gaussian sequence model setting. We also make the following assumption on the query point selected to estimate the gradient.

**Assumption 4.2** *The query point* $\theta \in \mathbb{R}^d$ *selected is such that* $\|\nabla f(\theta)\|_2 \leq R$*.*

Our algorithm for high-dimensional black-box sampling with variable selection is as follows:

- Pick a point $\theta$ (which is assumed to satisfy Assumption 4.2) and estimate the gradient $g_{\nu,n}$ at that point and compute the estimator $\hat{S}$ of $S^*$ as $\hat{S} = \{j : |[g_{\nu,n}]_j| \geq \tau\}$.

- Run ZO-LMC (or ZO-KLMC from Section 6.1) on the selected set of coordinates $\hat{S}$ of $f(\theta)$.

Here, for the first step, we need to select $n, \tau$ and $\nu$. We separate the set of relevant variables by thresholding $|[g_{\nu,n}]_j|$ at $\tau$. We now provide our result on the probability of erroneous selection.

**Theorem 4.3** *Let $f$ satisfy Assumption 1.1 and the query point selected satisfy Assumption 4.2. Set $\tau = (a - M\nu\sqrt{s})/2$ and assume that $\nu \leq \frac{a}{2M\sqrt{s}} \wedge \frac{R}{MC_2\sqrt{s}}$ and*

$$n \geq \frac{8RC\sqrt{s}}{a}\left(\frac{1}{K_2}\log\frac{4d}{\epsilon}\right)^{3/2} \vee K_1\frac{8RC\sqrt{s}}{a} \vee \left(\frac{8RC\sqrt{s}}{a}\right)^4$$

*where $C, C_2$ are constants. Then we have $\Pr\{\hat{S} \neq S^*\} \leq \epsilon$.*

**Remark 6** *The number of queries $n$ to the function $f$ depends only logarithmically on the dimension $d$ and is a (low-degree) polynomial in the sparsity level $s$. Combining this fact with the result in Theorem 3.1 we see that the total number of queries to the function $f$ (for the sampling error measured in 2-Wasserstein distance) is only poly-logarithmic in the true dimension $d$ and is a low-degree polynomial in the sparsity level $s$. Thus when $s \ll d$, we see the advantage of variable selection in black-box sampling using the two-step approach. The above results assumes that the sparsity level $s$ and signal strength is known. It would be interesting to construct adaptive estimators similar to those for Gaussian sequence model in [9]. Furthermore, exploring appropriately defined notions of non-homogenous sparsity assumptions is also challenging.*

# 5 Ozaki-Discretized Langevin Monte Carlo

We now consider the case of Ozaki-discretized LMC. Recall that the discrete-time LMC updates displayed in Equation 4, corresponds to the Euler-Maruyama discretization of continuous-time diffusion equation and it leverages the first-order gradient information regarding the target density $\pi(\theta)$. One could also potentially leverage higher-order derivative information regarding the target density. Specifically, the discretization proposed by Ozaki [14, 36, 43] corresponds to the discrete-time dynamics in Equation 10, that is based on the Hessian of $f(\theta)$ (and consequently the Hessian of $\pi(\theta)$). Let $\boldsymbol{H}(\cdot) \stackrel{def}{=} \nabla^2 f(\cdot) \in \mathbb{R}^{d\times d}$ be the true Hessian of the function $f$. We use the notation, $\boldsymbol{H}_t \stackrel{def}{=} \boldsymbol{H}(D_{t,0})$ to denote the Hessian evaluated at point $D_{t,0}$ where $D_{t,0} \sim \varpi_t$, the distribution of $x_t$, the $t^{\text{th}}$ step of the algorithm. Similarly, we denote by $\boldsymbol{S}(\cdot) \in \mathbb{R}^{d\times d}$ the "inexact" Hessian of the function $f$ and $\boldsymbol{S}_t \stackrel{def}{=} \boldsymbol{S}(D_{t,0})$. Furthermore, we follow the above conventions for the gradient as well: $\nabla f(\cdot) \in \mathbb{R}^d$ and $g(\cdot) \in \mathbb{R}^d$ denotes the true and "inexact" gradient respectively. Then, we have $\nabla f_t \stackrel{def}{=} \nabla f(D_{t,0})$ and $g_t \stackrel{def}{=} g(D_{t,0})$.

## 5.1 OLMC with Inaccurate Gradients and Hessians

With the above conventions, the Ozaki-discretized LMC corresponds to the following updates:

$$x_{t+1,h} = x_{t,h} - \boldsymbol{M}_t \nabla f(x_{t,h}) + \boldsymbol{\Sigma}_t^{1/2}\varepsilon_{t+1} \tag{9}$$

where $\boldsymbol{M}_t = \left(\boldsymbol{I}_d - e^{-h\boldsymbol{H}_t}\right)\boldsymbol{H}_t^{-1}$ and $\boldsymbol{\Sigma}_t = \left(\boldsymbol{I}_d - e^{-2h\boldsymbol{H}_t}\right)\boldsymbol{H}_t^{-1}$. The update steps of the Ozaki-discretized LMC (OLMC) algorithm with true Hessian and gradient information was analyzed in [14] and it was shown to have superior rates of convergence compared to the gradient based LMC algorithm (specifically, $t = \mathcal{O}(d/\epsilon \cdot \log(d/\epsilon))$ suffices to get $W_2(\varpi_t, \pi) \leq \epsilon$). Furthermore, relationships to the local-quadratic rates of the Newton method for optimization was also established. In this section, we assume that the true Hessian and gradient is unavailable. Instead we observe a random gradient vector $g(\cdot)$ and random Hessian matrix $\boldsymbol{S}(\cdot)$. Based on this, the OLMC with inexact information becomes

$$x_{t+1,h} = x_{t,h} - \tilde{\boldsymbol{M}}_t g(x_{t,h}) + \tilde{\boldsymbol{\Sigma}}_t^{1/2} \varepsilon_{t+1} \tag{10}$$

where, we have $\tilde{\boldsymbol{M}}_t \stackrel{def}{=} \left(\boldsymbol{I}_d - e^{-h\boldsymbol{S}_t}\right) \boldsymbol{S}_t^{-1}$ and $\tilde{\boldsymbol{\Sigma}}_t \stackrel{def}{=} \left(\boldsymbol{I}_d - e^{-2h\boldsymbol{S}_t}\right) \boldsymbol{S}_t^{-1}$ . In the rest of this subsection, we assume that $\boldsymbol{S}_t$ is invertible. The non-invertible case could potentially be handled by defining $\tilde{\boldsymbol{M}}_t \stackrel{def}{=} \left(\boldsymbol{I}_d - e^{-h\boldsymbol{S}_t}\right) (\boldsymbol{S}_t + \lambda \boldsymbol{I}_d)^{-1}$ and $\tilde{\boldsymbol{\Sigma}}_t \stackrel{def}{=} \left(\boldsymbol{I}_d - e^{-2h\boldsymbol{S}_t}\right) (\boldsymbol{S}_t + \lambda \boldsymbol{I}_d)^{-1}$ for some $\lambda > 0$. We do not pursue a detailed study of this case in this paper. We emphasize that when $\boldsymbol{S}_t$ is invertible, it still could be positive definite or not – we make a distinction between these two situations below. We now make the following assumption of the function $f$ and the quality of the approximation of the inexact gradients and Hessians.

**Assumption 5.1** *The hessian of the function $f(\theta)$ has Lipschitz smooth Hessian, i.e., $\|\boldsymbol{H}(\theta) - \boldsymbol{H}(\theta')\| \leq M_2 \|\theta - \theta'\|, \forall \theta, \theta' \in \mathbb{R}^d$.*

Furthermore, the gradient smoothness assumption in Assumption 1.1 implies boundedness of the second derivative, i.e., $\boldsymbol{H}(\theta) \preceq M \boldsymbol{I}_d, \forall \theta \in \mathbb{R}^d$. Assumption 5.1 and Assumption 1.1 ensure that the true hessian $\boldsymbol{H}(\theta)$ is positive definite and hence is invertible. In addition to the above assumption, we also make the following assumption on the quality of approximation of the inexact gradients and Hessians.

**Assumption 5.2** *We assume that the inexact gradient $g_t$ and symmetric inexact Hessian $\boldsymbol{S}_t$ satisfies:*

- *$\boldsymbol{S}_t, g_t, L_{0,0}$ are conditionally independent given $D_{t,0}$, where $L_{0,0}$ is defined as in the proof of Theorem 5.3.*
- *For all $t \in \mathbb{N}$, $T \in [0, h]$, we have*

$$\|g_t - \mathbf{E}[g_t|D_{t,0}]\|_{L_2} \leq \mathcal{C}_1(d) \qquad \|\nabla f_t - \mathbf{E}[g_t|D_{t,0}]\|_{L_2} \leq \mathcal{C}_2(d)$$
$$\|\boldsymbol{S}_t - \boldsymbol{H}_t \mid D_{t,0}\|_{L_2} \leq \mathcal{C}_3(d) \qquad \|e^{-T\boldsymbol{S}_t} \mid D_{t,0}\|_{L_2} \leq \mathcal{C}_4(d)$$
$$\|e^{-T\boldsymbol{S}_t}(\boldsymbol{S}_t - \boldsymbol{H}_t) \mid D_{t,0}\|_{L_2} \leq \mathcal{C}_5(d) \qquad \|e^{-T\boldsymbol{S}_t}\boldsymbol{S}_t^2 \mid D_{t,0}\|_{L_2} \leq \mathcal{C}_6(d)$$

- *$\tilde{M} \stackrel{def}{=} \sqrt{M\bar{M}} \vee M$ and $\hat{M}$ are constants satisfying the following inequalities.*
  - *In the case where $\boldsymbol{S}_t \succeq 0$,*

$$\|\boldsymbol{S}_t \mid D_{t,0}\|_{L_2} \leq \bar{M},$$
$$\|\boldsymbol{S}_t \mid D_{t,0}\|_{L_4} \leq \hat{M}.$$

– *In the case where $\boldsymbol{S}_t \succeq 0$ does not hold in general,*

$$\left\| \|Me^{-\boldsymbol{S}_t/M}\| \vee \|\boldsymbol{S}_t\| \mid D_{t,0} \right\|_{L_2} \le \bar{M},$$
$$\left\| e^{-T\boldsymbol{S}_t} \boldsymbol{S}_t^2 \mid D_{t,0} \right\|_{L_2} \le \hat{M}^2 = \mathcal{C}_6(d).$$

The second part of the above assumption is important as it defines the approximation quantity of the OLMC algorithm with inexact derivatives. A specific instantiation of the above quantities will be calculated for the zeroth-order algorithm presented in section 5.3. We now state our result.

**Theorem 5.3** *For the OLMC with inexact gradient and Hessian information, under Assumption 1.1 (**A1**), 5.1 and 5.2, we have the following guarantees:*

1. *If $\boldsymbol{S}_t$ is positive definite, for $h \le m/\tilde{M}^2$,*

$$W_2(\varpi_t, \pi) \le (1 - 0.25mh)^t W_2(\varpi_0, \pi) + 7.2\frac{M_2}{m} h(d+1) + \frac{4}{\sqrt{5m}} h^{1/2}\mathcal{C}_1(d) + \frac{4}{m}\mathcal{C}_2(d) + \frac{5.27}{m} h^{3/2} d^{1/2} \mathcal{C}_3(d).$$

2. *If $\boldsymbol{S}_t$ is not positive definite in general, for $h \le m/\tilde{M}^2$,*

$$W_2(\varpi_t, \pi) \le (1 - 0.25mh)^t W_2(\varpi_0, \pi) + 4\frac{M^2 M_2}{m^2} h^2 (d+1)\mathcal{C}_4(d)^2 + \left(3.51\frac{M_2}{m} h(d+1) + \frac{4}{\sqrt{5m}} h^{1/2}\mathcal{C}_1(d) + \frac{4}{m}\mathcal{C}_2(d)\right)\mathcal{C}_4(d) + \frac{5.27}{m}(hd)^{1/2}\mathcal{C}_5(d).$$

## 5.2 Approximated Ozaki-discretized LMC

While the Ozaki-discretized LMC is interesting from a theoretical perspective, from a practical perspective, it suffers from several computational drawbacks. Specifically, we need to compute matrix exponentials [31] and inverses, both of which are computationally demanding. Indeed a more practical discretization is obtained by approximating the matrix exponential in Equation 10 by series expansion. Specifically, we consider the following Approximate Ozaki-discretized LMC updates considered also in [14]:

$$x_{t+1,h} = x_{t,h} - h\left(\boldsymbol{I}_d - \frac{1}{2}h\boldsymbol{H}_t\right)\nabla f(x_{t,h}) + \sqrt{2h}\left(\boldsymbol{I}_d - h\boldsymbol{H}_t + \frac{1}{3}h^2\boldsymbol{H}_t^2\right)^{1/2}\varepsilon_{t+1} \quad (11)$$

With inexact gradients and Hessian information, we then have

$$x_{t+1,h} = x_{t,h} - h\left(\boldsymbol{I}_d - \frac{1}{2}h\boldsymbol{S}_t\right)g(x_{t,h}) + \sqrt{2h}\left(\boldsymbol{I}_d - h\boldsymbol{S}_t + \frac{1}{3}h^2\boldsymbol{S}_t^2\right)^{1/2}\varepsilon_{t+1} \quad (12)$$

**Theorem 5.4** *For the Approximate OLMC updates, under Assumption 1.1, 5.1 and 5.2, we have the following guarantees:*

*1. If $\boldsymbol{S}_t$ is positive definite, for $h \le 3m/(4M\hat{M}) \wedge 3m/(4\tilde{M}^2)$,*

$$W_2(\varpi_t, \pi) \le (1 - 0.25mh)^t W_2(\varpi_0, \pi) + 6.15\frac{M_2}{m}h(d+1) + 1.85\frac{\hat{M}^2}{m}h^{3/2}d^{1/2} + \frac{4.38}{m}\left(\mathcal{C}_1(d) + \mathcal{C}_2(d)\right) + \frac{5}{m}(hd)^{1/2}\mathcal{C}_3(d).$$

*2. If $\boldsymbol{S}_t$ is not positive definite in general, for $h \le 3m/(4M\hat{M}) \wedge 3m/(4\tilde{M}^2)$,*

$$\begin{aligned} W_2(\varpi_t, \pi) \le (1 - 0.25mh)^t W_2(\varpi_0, \pi) &+ 3.85\frac{M^2 M_2}{m^2}h^2(d+1)\mathcal{C}_4(d)^2 + \frac{5}{m}(hd)^{1/2}\mathcal{C}_5(d) \\ &+ \left(3.51\frac{M_2}{m}h(d+1) + \frac{4}{m}(\mathcal{C}_1(d) + \mathcal{C}_2(d))\right)\mathcal{C}_4(d) \\ &+ \left(\frac{1.85}{m}h^{3/2}d^{1/2} + \frac{2}{3m}h^2(\mathcal{C}_1(d) + \mathcal{C}_2(d))\right)\mathcal{C}_6(d). \end{aligned}$$

### 5.3 ZOO-LMC: Zeroth-Order OLMC for Black-box Sampling

As discussed in Section 3, in the setting of black-box sampling, access to the function $f(\theta)$ is only through function evaluations. In this section, we extend the Ozaki-Discretized sampling algorithm to the black-box setting, thereby extending their applicability. While, the gradient estimation technique from function queries in Section 3 was based on first-order Gaussian Stein's identity, here we leverage the second-order Gaussian Stein's identity to estimate the Hessian from function queries, as proposed in [1]. Second-order Stein's identity states that for a standard Gaussian vector $u$, we have $\mathbf{E}[(uu^\top - \boldsymbol{I}_d)\ g(u)] = \mathbf{E}[\nabla^2 g(u)]$, for all functions $g$ with well-defined Hessians. Similar to first-order Stein's identity, this naturally relates function queries to Hessians. In order to leverage this, similar to the previous case, we let $f_\nu(\theta) = f(\theta + \nu u)$. Then, we have

$$\mathbf{E}\left[\frac{(uu^\top - \boldsymbol{I}_d)f(\theta + \nu u)}{\nu^2}\right] = \mathbf{E}[\nabla^2 f(\theta + \nu u)] = \nabla^2 f_\nu(\theta) = \boldsymbol{H}_{f_\nu}. \tag{13}$$

This provides a way of approximately estimating the Hessian of the function $f_\nu$ by approximating the expectation on the left hand side using Gaussian samples. Hence, we can leverage this estimate of Hessian of the smoothed function to get an approximate estimate of Hessian of $f$. Specifically, we now have the following estimates of the Hessian, as in [1]:

$$\hat{\boldsymbol{H}}_{f_\nu} \overset{def}{=} \hat{\boldsymbol{H}}_{f_\nu}(\theta, u) = \frac{1}{2\nu^2}\left(uu^\top - \boldsymbol{I}_d\right)\left[f(\theta + \nu u) - f(\theta) + f(\theta - \nu u) - f(\theta)\right], \tag{14}$$

$$\hat{\boldsymbol{H}}_{f_\nu, n} = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{H}}_{f_\nu}(\theta, u_i). \tag{15}$$

Hence, ZOO-LMC and approximate ZOO-LMC corresponds to Equation 10 and 12 respectively, with the derivatives estimated based on Stein's identity. While $f(\theta)$ is strongly convex and so the true Hessian is invertible, there is not guarantee that the above sample Hessian is invertible. For the zeroth-order version of Ozaki discretization to be well-defined the sample Hessian must be invertible. The question, when the above estimator is invertible is a rather delicate question and requires tools from random matrix theory tools (for example, [47]) to be understood. But the approximated Ozaki discretization, which does not involve such inevitability issues also achieves the same performance as the Ozaki discretization. Hence, we postpone a detailed study of inevitability of the zeroth-order Hessian for future work. A consequence of the Hessian smoothness assumption is the following equivalent assumption on the function $f(\theta)$.

**Assumption 5.5** *The function $f$ is assumed to be twice differentiable and satisfy the following smoothness condition: for all points $\theta, \theta' \in \mathbb{R}^d$,*

$$|f(\theta') - f(\theta) - \nabla^\top f(\theta)(\theta' - \theta) - \frac{1}{2}(\theta' - \theta)^\top \nabla^2 f(\theta)(\theta' - \theta)| \leq \frac{M_2}{6}\|\theta' - \theta\|^3. \tag{16}$$

**Lemma 5.1** *Let the Hessian estimator be defined in* (14) *and Assumption 5.5 hold. Then, we have*

$$\|\boldsymbol{H}_{f_\nu} - \boldsymbol{H}_f\|_2 \leq M_2 \nu d^{1/2} \tag{17}$$

**Lemma 5.2** *Under Assumption 5.5, we have*

$$\|\hat{\boldsymbol{H}}_{f_\nu} - \boldsymbol{H}_f\|_{L_2,2} \leq \|\hat{\boldsymbol{H}}_{f_\nu} - \boldsymbol{H}_f\|_{L_2,F} \leq \frac{1}{6}M_2\nu(d+4)^{5/2} + \frac{1}{2}\|\boldsymbol{H}_f\|_2(d+3)^2. \tag{18}$$

$$\|\hat{\boldsymbol{H}}_{f_\nu,n} - \boldsymbol{H}_f\|_{L_2,2} \leq \frac{1}{6\sqrt{n}}M_2\nu(d+4)^{5/2} + \frac{1}{2\sqrt{n}}\|\boldsymbol{H}_f\|_2(d+3)^2 + M_2\nu d^{1/2}. \tag{19}$$

Based on the above result and Theorem 5.3, we have the following guarantees for the ZOO-LMC updates.

**Theorem 5.6** *Let Assumption 1.1 (**A1**) and 5.1 hold. Then*

1. *For the ZOO-LMC we have the following guarantees:*

$$\begin{aligned} W_2(\varpi_t, \pi) &\leq (1 - 0.25mh)^t W_2(\varpi_0, \pi) \\ &\quad + P_{n_H}^{n_H d/2}\left(\frac{4\sqrt{2}}{\sqrt{5m}}\frac{1}{\sqrt{n_g}}h^{1/2}\mathcal{C}_1'(d) + \frac{4}{m}\mathcal{C}_2(d) + P_{n_H}^5\frac{5.27}{m}(hd)^{1/2}\mathcal{C}_3(d)\right) \\ &\quad + P_{n_H}^{n_H d/2}\left((\frac{16}{3}P_{n_H}^{n_H d/2} + 3.51)\frac{M_2}{m}h(d+1) + P_{n_H}^5\frac{5.27M^2}{m}\frac{1}{n_H}(hd)^{3/2}\right). \end{aligned}$$

2. *For the Approximate ZOO-LMC updates, we have:*

$$\begin{aligned} W_2(\varpi_t, \pi) &\leq (1 - 0.25mh)^t W_2(\varpi_0, \pi) \\ &\quad + P_{n_H}^{n_H d/2}\left((\frac{2.51}{\sqrt{m}}\frac{1}{\sqrt{n_g}}h^{1/2} + P_{n_H}^8\frac{M^2}{6m}h^2(d+7)^4)\mathcal{C}_1'(d) + \frac{4}{m}\mathcal{C}_2(d) + P_{n_H}^5\frac{5}{m}(hd)^{1/2}\mathcal{C}_3(d)\right) \\ &\quad + P_{n_H}^{n_H d/2}\left((3.89P_{n_H}^{n_H d/2} + 3.51)\frac{M_2}{m}h(d+1) + P_{n_H}^8\frac{M^2}{3m}h^{3/2}(d+7)^{9/2}\right), \end{aligned}$$

*where subscripts $g$, $H$ are used to distinguish parameters of gradient and Hessian estimators respectively.* $P_{n_H} = (1 - \frac{2Mh}{n_H})^{-1/2} \leq \sqrt{2}$, $P_{n_H}^{n_H d} \leq (1 - 2Mhd)^{-1/2} \leq \sqrt{2}$ *for* $h \leq \frac{1}{4Md}$, *and*

$$\mathcal{C}_1'(d) = \frac{1}{2}M\nu_g^2(d+2)^{3/2} + \sqrt{M}(d+1), \qquad \mathcal{C}_2(d) = M\nu_g d^{1/2},$$

$$\mathcal{C}_3(d) = \frac{1}{6\sqrt{n_H}}M_2\nu_H(d+4)^{5/2} + \frac{1}{2\sqrt{n_H}}M(d+3)^2 + M_2\nu_H d^{1/2}.$$

**Remark 7** *For ZOO-LMC in order for $W_2(\varpi_t, \pi) \leq \epsilon$ to hold, it suffices to have $h = \mathcal{O}(\epsilon/d)$, $\nu_g = \mathcal{O}\left(\epsilon/\sqrt{d}\right)$, $n_g = \mathcal{O}(d/\sqrt{\epsilon})$, $\nu_H = \mathcal{O}(\sqrt{d/\epsilon})$, $n_H = \mathcal{O}(d^4/\epsilon)$ with $t = \mathcal{O}(d/\epsilon)$. For the approximate ZOO-LMC, it suffices to have $h = \mathcal{O}(\epsilon/d) \wedge \mathcal{O}\left(\epsilon^{2/3}/d^3\right)$, $\nu_g = \mathcal{O}\left(\epsilon/\sqrt{d}\right)$, $n_g = \mathcal{O}(d/\sqrt{\epsilon})$, $\nu_H = \mathcal{O}\left(\sqrt{\epsilon/d}\right)$, $n_H = \mathcal{O}\left(d^4/\epsilon\right)$. Note that depending on the value of $\epsilon$ desired, we get improved rates over ZO-LMC algorithm. It is extremely interesting to obtain better dependence on $d$ under further structural assumptions on the Hessian of $f(\theta)$, for example, when $f$ is a finite-sum as in [25].*

# 6 Kinetic Langevin Monte Carlo Discretizations

In the previous section, we consider several discretizations of the continuous-time diffusion process in Equation 3 based on gradient and Hessian information. We now consider gradient and Hessian based discretizations of Kinetic Langevin diffusion process given below:

$$d\begin{bmatrix} V_T \\ L_T \end{bmatrix} = \begin{bmatrix} (\gamma V_T + \nabla f(L_T)) \\ V_T \end{bmatrix} dT + \sqrt{2\gamma}\begin{bmatrix} \mathbf{I}_d \\ \mathbf{0}_d \end{bmatrix} dW_T, \tag{20}$$

where $\mathbf{I}_d$ is the $d\times d$ identity matrix and $\mathbf{0}_d$ is the all-zero $d\times d$ matrix. We refer the reader to [11, 15, 20] for more details about the above diffusion process and related theoretical results. Specifically, it was shown in [11, 15] that first-order discretizations of the kinetic diffusion process (refered to as KLMC and proposed first in [11]) in Equation 20 have better rates of convergence compared to similar first-order discretizations of the continous-process in Equation 3. Specifically, recall that for the right choice of tuning parameters, vanilla LMC (i.e., first-order discretizations of Equation 3) requires that $t = \mathcal{O}(d/\epsilon^2 \cdot \log(d/\epsilon))$ for $W_2(\varpi_t, \pi) \leq \epsilon$. Whereas, it was shown in [11, 15] $t = \mathcal{O}(\sqrt{d}/\epsilon \cdot \log(d/\epsilon))$ suffices ([15] provides a much sharper result compared to [11]). We emphasize that the above result does not immediately imply that KLMC might be the algorithm to use always (in comparison to LMC); indeed when considering also the dependence of the bound on the strong-convexity and smoothness parameters ( through the condition number of the sampling density defined as $M/m$), [15] precisely characterize when KLMC might be preferred over the vanilla LMC. The bottom line of their analysis is none of the method is uniformly better over the other method. More interestingly, [15] also proposed a second-order discretization of the kinetic diffusion process in Equation 20, denoted as KLMC2, that requires only $t = \mathcal{O}(\sqrt{d/\epsilon} \cdot \log(d/\epsilon))$ under the Hessian-smoothness assumption. Compared to the Ozaki discretization proposed in [14], KLMC2 achieves better rates. In this section, we analyze KLMC and KLMC2 algorithms, under inaccurate gradients and Hessians. Furthermore, we propose the corresponding zeroth-order versions of KLMC and KLMC2 algorithms and establish its theoretical properties.

## 6.1 ZO-KLMC for Black-box Sampling

We first consider the first-order discretization of the SDE in Equation 20, first proposed by [11] and also analyzed in [15]. This KLMC discretization is given by the following updates:

$$\begin{bmatrix} \tilde{x}_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\tilde{x}_t - \psi_1(h)\nabla f(x_t) \\ x_t + \psi_1(h)\tilde{x}_t - \psi_2(h)\nabla f(x_t) \end{bmatrix} + \sqrt{2\gamma}\begin{bmatrix} \tilde{\epsilon}_{t+1} \\ \epsilon_{t+1} \end{bmatrix}, \tag{21}$$

where $(\tilde{\epsilon}_{t+1}, \epsilon_{t+1}) \in \mathbb{R}^{2d}$ is a a sequence of i.i.d standard Normal vectors, independent of $(\tilde{x}_0, x_0)$ and $\psi_0(t) = e^{-\gamma t}$ and $\psi_{t+1} = \int_0^T \psi_t(s)ds$. Based on this, we now consider the following two related update steps, the inaccurate KLMC and ZO-KLMC respectively.

$$\begin{bmatrix} \tilde{x}_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\tilde{x}_t - \psi_1(h)g(x_t) \\ x_t + \psi_1(h)\tilde{x}_t - \psi_2(h)g(x_t) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \tilde{\epsilon}_{t+1} \\ \epsilon_{t+1} \end{bmatrix}, \tag{22}$$

$$\begin{bmatrix} \tilde{x}_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\tilde{x}_t - \psi_1(h)g_{\nu,n}(x_t) \\ x_t + \psi_1(h)\tilde{x}_t - \psi_2(h)g_{\nu,n}(x_t) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \tilde{\epsilon}_{t+1} \\ \epsilon_{t+1} \end{bmatrix}, \tag{23}$$

where $g()$ is any random gradient assumed to satisfy Assumptions 5.2 and $g_{\nu,n}$ is the zeroth-order gradient estimator as in Equation 7.

**Theorem 6.1** *Assume that $f$ is twice differentiable with $m\boldsymbol{I}_d \preceq \nabla^2 f(x) \preceq M\boldsymbol{I}_d$. If the initial point $(\tilde{x}_0, x_0)$ is chosen such that $\tilde{x}_0 \sim N(0, \boldsymbol{I}_d)$, then, for $\gamma \geq \sqrt{m+M}$ and $h \leq \frac{m}{4\gamma M}$, we have the following results:*

1. *For the KLMC with inaccurate gradient and Hessians, we have*

$$W_2(\varpi_t, \pi) \leq \sqrt{2}(1 - \frac{3mh}{4\gamma})^t W_2(\varpi_0, \pi) + \frac{3\sqrt{2}M}{2m} h\sqrt{d} + \frac{9}{2m}(\mathcal{C}_1(d) + \mathcal{C}_2(d)).$$

2. *For the ZO-KLMC, we have, for $n_g \geq \frac{81M^2}{m^2}(d+1)$,*

$$W_2(\varpi_t, \pi) \leq \sqrt{2}(1 - \frac{mh}{2\gamma})^t W_2(\varpi_0, \pi) + \frac{3\sqrt{2}M}{2m} h\sqrt{d} + \frac{9}{2m}(\mathcal{C}_1'(d) + \mathcal{C}_2(d)),$$

*where $\mathcal{C}_1'(d) = \frac{M\nu_g}{\sqrt{n_g}}(d+2)^{3/2} + \frac{\sqrt{M}}{\sqrt{n_g}}(d+1)$ and $\mathcal{C}_2(d) = M\nu_g\sqrt{d}$.*

## 6.2 ZO-KLMC2 for Black-box Sampling

We now consider the second-order discretization of the SDE in Equation 20, called as KLMC2 proposed in [15]:

$$\begin{bmatrix} \tilde{x}_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\tilde{x}_t - \psi_1(h)\nabla f(x_t) - \psi_2(h)\mathbf{H}_t\tilde{x}_t \\ x_t + \psi_1(h)\tilde{x}_t - \psi_2(h)\nabla f(x_t) - \psi_3(h)\mathbf{H}_t\tilde{x}_t \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \tilde{\epsilon}_{t+1} - \mathbf{H}_t\ddot{\epsilon}_{t+1} \\ \check{\epsilon}_{t+1} - \mathbf{H}_t\breve{\epsilon}_{t+1} \end{bmatrix}, \tag{24}$$

where $(\tilde{\epsilon}_{t+1}, \ddot{\epsilon}_{t+1}, \check{\epsilon}_{t+1}, \breve{\epsilon}_{t+1}) \in \mathbb{R}^{4d}$ is a a sequence of i.i.d standard Normal vectors, independent of $(\tilde{x}_0, x_0)$ and $\psi_0(t) = e^{-\gamma t}$ and $\psi_{t+1} = \int_0^T \psi_t(s)ds$. Following the notations, introduced in Section 5, we now introduce the KLMC2 updates with inacurate gradient and Hessian and Zeroth-order KLMC2 (ZO-KLMC2) updates respectively, below:

$$\begin{bmatrix} \tilde{x}_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\tilde{x}_t - \psi_1(h)g(x_t) - \psi_2(h)\mathbf{S}_t\tilde{x}_t \\ x_t + \psi_1(h)\tilde{x}_t - \psi_2(h)g(x_t) - \psi_3(h)\mathbf{S}_t\tilde{x}_t \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \tilde{\epsilon}_{t+1} - \mathbf{S}_t\ddot{\epsilon}_{t+1} \\ \check{\epsilon}_{t+1} - \mathbf{S}_t\breve{\epsilon}_{t+1} \end{bmatrix}, \tag{25}$$

$$\begin{bmatrix} \tilde{x}_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\tilde{x}_t - \psi_1(h)g_{\nu,n}(x_t) - \psi_2(h)\hat{\mathbf{H}}_t\tilde{x}_t \\ x_t + \psi_1(h)\tilde{x}_t - \psi_2(h)g_{\nu,n}(x_t) - \psi_3(h)\hat{\mathbf{H}}_t\tilde{x}_t \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \tilde{\epsilon}_{t+1} - \hat{\mathbf{H}}_t\ddot{\epsilon}_{t+1} \\ \check{\epsilon}_{t+1} - \hat{\mathbf{H}}_t\breve{\epsilon}_{t+1} \end{bmatrix}, \tag{26}$$

where $g()$ and $\mathbf{S}_t$ are any random gradient and Hessian, assumed to satisfy Assumptions 5.2 and $g_{\nu,n}$ is the zeroth-order gradient estimator as in Equation 7 and $\hat{\mathbf{H}}_t$ is the zeroth-order Hessian estimator as in Equation 15.

**Theorem 6.2** *Under Assumption 1.1, 5.1 and 5.2, if the initial point $(\tilde{x}_0, x_0)$ satisfies $\tilde{x}_0 \sim N(0, \boldsymbol{I}_d)$, then, for $\gamma \geq \sqrt{m+M}$ and $h \leq \frac{m}{\gamma(3.5M+1.5\overline{M})} \wedge \frac{m}{4\sqrt{5d}M_2}$, we have the following results:*

1. *For the KLMC2 with inaccurate gradient and Hessians, we have*

$$W_2(\varpi_t, \pi) \leq \sqrt{2}(1 - \frac{mh}{4\gamma})^t W_2(\varpi_0, \pi) + \frac{16M}{m} he^{-d/2} + \frac{2\sqrt{2}M_2}{m} h^2(d+1) + \frac{2\sqrt{2M^3}}{m} h^2 d^{1/2}$$
$$+ \frac{18\sqrt{2}}{m} hd^{1/2}\mathcal{C}_3(d) + \frac{36\sqrt{2}}{m}(\mathcal{C}_1(d) + \mathcal{C}_2(d)).$$

2. *For the ZO-KLMC2, we have, for $n_g \geq \frac{256M^2}{m^2}(d+1)$,*

$$W_2(\varpi_t, \pi) \leq \sqrt{2}(1 - \frac{mh}{8\gamma})^t W_2(\varpi_0, \pi) + \frac{16M}{m} he^{-d/2} + \frac{2\sqrt{2}M_2}{m} h^2(d+1) + \frac{2\sqrt{2M^3}}{m} h^2 d^{1/2}$$
$$+ \frac{18\sqrt{2}}{m} hd^{1/2}\mathcal{C}_3(d) + \frac{36\sqrt{2}}{m}(\mathcal{C}'_1(d) + \mathcal{C}_2(d)).$$

*where*

$$\mathcal{C}'_1(d) = \frac{M\nu_g}{\sqrt{n_g}}(d+2)^{3/2} + \frac{\sqrt{M}}{\sqrt{n_g}}(d+1), \quad \mathcal{C}_2(d) = M\nu_g\sqrt{d},$$
$$\mathcal{C}_3(d) = \frac{1}{6\sqrt{n_H}} M_2\nu_H(d+4)^{5/2} + \frac{1}{2\sqrt{n_H}} M(d+3)^2 + M_2\nu_H d^{1/2}.$$

# 7 Discussion

While our results for black-box sampling are based on exact zeroth-order information, it is straightforward to extend it to the case of inexact (but unbiased) zeroth-order information. The case of biased zeroth-order information is more challenging and we leave it as future work. Furthermore, the choice of the coordinate system for defining the diffusion process and its discretizations are crucial, as pointed out in [22]. We remark that black-box versions of the algorithms proposed in [22] with the transformed coordinate system would immediately follow based on our gradient and Hessian estimators.

Recall that our gradient and Hessian estimators were based on Gaussian Stein's identity and could be used for the case when $f$ is defined on the entire Euclidean space $\mathbb{R}^d$.In several situation, for example, in sampling from densities with compact support [7, 8] and in computing volume of convex body [6], one needs to compute the gradient of the function (and density) supported on $\mathcal{M} \subset \mathbb{R}^d$. For these situations, one can use a version of Stein's identity based on score functions to compute the gradient and Hessian. To explain more, we first recall some definitions. The score function $S_p \colon \mathcal{M} \to \mathbb{R}^d$ associated to density $p(u)$ defined over $\mathcal{M}$ is defined as

$$S_p(u) = -\nabla_u[\log p(u)] = -\nabla_u p(u)/p(u).$$

In the above definition, the derivative is taken with respect to the argument $u$ and not the parameters of the density $p(u)$. Based on the above definition, we have the following versions of Stein's identity; see, for example, [23].

**Proposition 7.1** *Let $U$ be a $\mathcal{M}$-valued random vector with density $p(u)$. Assume that $p\colon \mathcal{M} \to \mathbb{R}$ is differentiable. In addition, let $g : \mathcal{M} \to \mathbb{R}$ be a continuous function such that $\mathbf{E}_U[\nabla g(U)]$ exists and the following is true: $\int_{u \in \mathcal{M}} \nabla_u (g(u)p(u))\, du = 0$. Then it holds that*

$$\mathbf{E}_U[g(U) \cdot S(U)] = \mathbf{E}_U[\nabla g(U)],$$

*where $S(u) = -\nabla p(u)/p(u)$ is the score function of $p(u)$.*

In order to leverage the above identities to estimate the gradient of a given function $f(\theta) : \mathcal{M} \to \mathbb{R}$, consider $g(U) = f(\theta + U)$ where $U \sim p(u)$ is a $\mathcal{M}$-valued random variable and appeal to the above Stein's identity above, as done in Section 3 for with Gaussian random variables. Similar techniques for Hessian estimation could also be used. We postpone a rigorous analysis of the estimation and approximation rates in this case for future work.

## References


[1] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality and saddle-points. *arXiv preprint arXiv:1809.06474v2*, 2019.

[2] Krishnakumar Balasubramanian, Jianqing Fan, and Zhuoran Yang. Tensor methods for additive index models under discordance and heterogeneity. *arXiv preprint arXiv:1807.06693*, 2018.

[3] Alessandro Barp, François-Xavier Briol, Anthony D Kennedy, and Mark Girolami. Geometry and dynamics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 2018.

[4] Claude JP Bélisle, H Edwin Romeijn, and Robert L Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993.

[5] Michael Betancourt, Simon Byrne, Sam Livingstone, and Mark Girolami. The geometric foundations of hamiltonian monte carlo. *Bernoulli*, 23(4A):2257–2298, 2017.

[6] Silouanos Brazitikos, Apostolos Giannopoulos, Petros Valettas, and Beatrice-Helen Vritsiou. *Geometry of isotropic convex bodies*, volume 196. American Mathematical Soc., 2014.

[7] Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *COLT*, 2017.

[8] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete & Computational Geometry*, 2018.

[9] Cristina Butucea, Mohamed Ndaoud, Natalia A Stepanova, and Alexandre B Tsybakov. Variable selection with hamming loss. *The Annals of Statistics*, 46(5):1837–1875, 2018.

[10] Ming-Hui Chen and Bruce Schmeiser. Toward black-box sampling: A random-direction interior-point markov chain approach. *Journal of Computational and Graphical Statistics*, 7(1):1–22, 1998.

[11] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

[12] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

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

[14] Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

[15] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.

[16] Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

[17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[18] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *arXiv preprint arXiv:1802.09188*, 2018.

[19] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! *arXiv preprint arXiv:1801.02309*, 2018.

[20] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for langevin dynamics. *arXiv preprint arXiv:1703.01617*, 2017.

[21] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[22] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[23] Jackson Gorham and Lester Mackey. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

[24] Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(3-4):541–559, 1995.

[25] Avetik Karagulyan. Non-asymptotic guarantees for sampling by stochastic gradient descent. *arXiv preprint arXiv:1811.00781*, 2018.

[26] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18 (1):430–474, 2017.

[27] László Lovász and Miklós Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Foundations of Computer Science, 1990. Proceedings., 31st Annual Symposium on*, pages 346–354. IEEE, 1990.

[28] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[29] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

[30] Kerrie L Mengersen and Richard L Tweedie. Rates of convergence of the hastings and metropolis algorithms. *The annals of Statistics*, 24(1):101–121, 1996.

[31] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1):3–49, 2003.

[32] Radford M Neal. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

[33] Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[34] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

[36] Tohru Ozaki. Statistical identification of storage models with application to stochastic hydrology 1. *JAWRA Journal of the American Water Resources Association*, 21, 1985.

[37] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *COLT*, 2017.

[38] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[39] Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771, 2015.

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

[41] Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[42] Richard J Samworth. Recent progress in log-concave density estimation. *Statistical Science*, 33(4):493–509, 2018.

[43] Isao Shoji and Tohru Ozaki. Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications*, 16(4):733–752, 1998.

[44] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

[45] O Stramer and RL Tweedie. Langevin-type models i: Diffusions with given stationary distributions and their discretizations. *Methodology and Computing in Applied Probability*, 1(3): 283–306, 1999.

[46] O Stramer and RL Tweedie. Langevin-type models ii: Self-targeting candidates for mcmc algorithms. *Methodology and Computing in Applied Probability*, 1(3):307–328, 1999.

[47] Harald Uhlig. On singular wishart and singular multivariate beta distributions. *The Annals of Statistics*, pages 395–405, 1994.

[48] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[49] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

# A Notations

We use $a \wedge b$ and $a \vee b$ to denote the minimum and maximum of $a$ and $b$ respectively. The $L_2$ norm of a random vector $X : \Omega \to \mathbb{R}^d$ is defined to be $\|X\|_{L_2} = \mathbf{E}[\|X\|_2^2]^{1/2}$. The $L_p$ norms of a random matrix $\boldsymbol{M} : \Omega \to \mathbb{R}^{d\times d}$ are defined as follows.

$$\|\boldsymbol{M}\|_{L_p,2} = \mathbf{E}[\|\boldsymbol{M}\|_2^p]^{1/p},$$
$$\|\boldsymbol{M}\|_{L_p,F} = \mathbf{E}[\|\boldsymbol{M}\|_F^p]^{1/p},$$

where $\|\cdot\|_2$ is the spectral norm, and $\|\cdot\|_F$ is the Frobenius norm. For simplicity, we write $\|\cdot\| = \|\cdot\|_2$ and $\|\cdot\|_{L_p} = \|\cdot\|_{L_p,\bullet}$ when there is no ambiguity. Furthermore, we omit the subscript $h$ in $x_{t,h}$ in places where is no confusion for simplicity.

# B Proofs for Section 2

We first state and prove the following Lemma, used in the proof of Theorem 2.2.

**Lemma B.1** *If $f$ is Lipschitz continuous with component Lipschitz constant $M_i$, then*

$$(a)\ \|V_i\|_{L_2} \le \frac{7\sqrt{2}}{6} M_i h^{3/2} \qquad (b)\ \|\tilde{V}\|_{L_2} \le \frac{7\sqrt{2}}{6} M_{\max} h^{3/2} d^{1/2}$$

*Proof.* The proof of part (a) results from the proof of Lemma 4 in [14]. In fact, we can consider $f$ as a function of the $i$-th component only, and apply the lemma in 1-dimensional case. Part (b) follows, as

$$\|\tilde{V}\|_{L_2} = \|\sum_{i=1}^{d} V_i\|_{L_2} = \left(\sum_{i=1}^{d} \|V_i\|_{L_2}^2\right)^{1/2} \le \frac{7\sqrt{2}}{6}\left(\sum_{i=1}^{d} M_i^2\right)^{1/2} h^{3/2} \le \frac{7\sqrt{2}}{6} M_{\max} h^{3/2} d^{1/2}.$$

∎

*Proof.* [Proof of Theorem 2.2] Let $L_{0,0} \sim \pi$ be the random variable that attains the Wasserstein distance $W_2(\varpi_0, \pi) = \|L_{0,0} - x_0\|_{L_2}$. Define a family of random processes inductively by

$$L_{t,T,i} = L_{t,0,i} - \int_0^T [\nabla f(L_{t,s,i})]_i e_i ds + \sqrt{2}(W_{t,T})_i e_i,$$

for $T \in [0, h],\ i = 1, 2, \ldots, d$, where the initial data is $L_{t,0,i} = L_{i-1,h,i_{t-1}}$, and $W_{t,T}$ is a Brownian motion satisfying $W_{t,h} = \sqrt{h}\varepsilon_{t+1}$, the noise term in the LMC algorithm. By Fokker-Planck equation, $\pi$ is the stationary distribution of $L_{t,T,i}$ for each $i = 1, 2, \ldots, d$, which implies $L_{t,T,i} \sim \pi$. Moreover, define $\Delta_{t,i} = L_{t-1,h,i} - x_t$. For simplicity, we drop the last subscript $i$ when it coincides with $i_t$. Then

$$\begin{aligned}
\Delta_{t+1,i} &= \Delta_t - (x_{t+1} - x_t) + (L_{t,h,i} - L_{t,0,i}) \\
&= \Delta_t - \left(-h[\nabla f(x_t)]_i e_i + \sqrt{2h}(\varepsilon_{t+1})_i e_i\right) + \left(-\int_0^h [\nabla f(L_{t,s,i})]_i e_i ds + \sqrt{2}(W_{t,h})_i e_i\right) \\
&= \Delta_t + h[\nabla f(x_t)]_i e_i - \int_0^h [\nabla f(L_{t,s,i})]_i e_i ds \\
&\overset{def}{=} \Delta_t - hU_i - V_i,
\end{aligned}$$

where $U_i = \left([\nabla f(L_{t,0,i})]_i - [\nabla f(x_t)]_i\right) e_i$ and $V_i = \int_0^h \left([\nabla f(L_{t,s,i})]_i - [\nabla f(L_{t,0,i})]_i\right) e_i ds$.

Letting $i = i_t$, and taking expectation w.r.t. $i_t$, we have

$$
\begin{aligned}
\mathbf{E}_{i_t}[\|\Delta_{t+1}\|_{L_2}^2] &= \frac{1}{d}\sum_{i=1}^{d} \|\Delta_{t+1,i}\|_{L_2}^2 \\
&= \frac{1}{d}\sum_{i=1}^{d} \|\Delta_t - hU_i - V_i\|_{L_2}^2 \\
&= \frac{1}{d}\sum_{i=1}^{d} \left(\|(\Delta_t)_{-i}e_{-i}\|_{L_2}^2 + \|(\Delta_t)_i e_i - hU_i - V_i\|_{L_2}^2\right) \\
&= \frac{d-1}{d}\|\Delta_t\|_{L_2}^2 + \frac{1}{d}\|\Delta_t - hU - \tilde{V}\|_{L_2}^2 \\
&\le \frac{d-1}{d}\|\Delta_t\|_{L_2}^2 + \frac{1}{d}\left((1-mh)\|\Delta_t\|_{L_2} + \frac{7\sqrt{2}}{6}M_{\max}h^{3/2}d^{1/2}\right)^2 \\
&\le \left(\frac{d-1}{d} + \frac{1}{d}(1-mh)^2\right)\left(\|\Delta_t\|_{L_2} + \frac{7\sqrt{2}}{6}\frac{M_{\max}}{1-mh}h^{3/2}d^{1/2}\right)^2 \\
&= \left(1 - \frac{mh(2-mh)}{d}\right)\left(\|\Delta_t\|_{L_2} + \frac{7\sqrt{2}}{6}\frac{M_{\max}}{1-mh}h^{3/2}d^{1/2}\right)^2,
\end{aligned}
$$

where $U = \nabla f(L_{t,0,i}) - \nabla f(x_t) = \nabla f(x_t + \Delta_t) - \nabla f(x_t)$, $\tilde{V} = \sum_{i=1}^{d} V_i$. In the above calculation, the first inequality follows from Lemma 2 in [14], that $\|\Delta_t - hU\| \le (1-mh)\|\Delta_t\|$ provided $h < \frac{2}{m+M}$, and from Lemma B.1 below. Taking expectation w.r.t. $i_0, i_1, \ldots, i_{t-1}$, the expected error satisfies the following inequality.

$$
\begin{aligned}
\mathbf{E}\left[\|\Delta_{t+1}\|_{L_2}\right] &\le \mathbf{E}[\|\Delta_{t+1}\|_{L_2}^2]^{\frac{1}{2}} \\
&\le \sqrt{1 - \frac{mh(2-mh)}{d}}\left(\mathbf{E}[\|\Delta_t\|_{L_2}] + \frac{7\sqrt{2}}{6}\frac{M_{\max}h^{3/2}}{1-mh}d^{1/2}\right).
\end{aligned}
$$

Applying it iteratively leads to

$$
\begin{aligned}
\mathbf{E}[\|\Delta_{t+1}\|_{L_2}] &\le \left(1 - \frac{mh(2-mh)}{d}\right)^{\frac{t}{2}}\mathbf{E}[\|\Delta_t\|_{L_2}] \\
&\quad + \left[\left(1 - \frac{mh(2-mh)}{d}\right)^{\frac{1}{2}} + \cdots + \left(1 - \frac{mh(2-mh)}{d}\right)^{\frac{t}{2}}\right]\frac{7\sqrt{2}}{6}\frac{M_{\max}h^{3/2}}{1-mh}d^{1/2} \\
&\le \left(1 - \frac{mh(2-mh)}{d}\right)^{\frac{t}{2}}\|\Delta_0\|_{L_2} + \frac{7\sqrt{2}}{3}\left(1 - \frac{mh(2-mh)}{d}\right)^{\frac{1}{2}}\frac{M_{\max}h^{1/2}}{m(1-mh)}d^{3/2} \\
&\le \left(1 - \frac{m}{2d}h\right)^{t}\|\Delta_0\|_{L_2} + \frac{7\sqrt{2}}{3}\frac{M_{\max}}{m(1-mh)}h^{1/2}d^{3/2},
\end{aligned}
$$

where we have $1 - \frac{mh(2-mh)}{d} \le \left(1 - \frac{mh}{2d}\right)^2$ for $h \le \frac{1}{m}$. Finally, by definition of Wasserstein distance, we reach the results as desired. ■

# C Proofs for Section 3

*Proof.* [Proof of Theorem 3.1] The proof follows by first calculating the bias and variance of the inaccurate gradient in our zeroth-order setting, where the error term $\zeta_t = g_{\nu,n}(x_t) - \nabla f(x_t)$. First, by Stein's identity, $\mathbf{E}[g_{\nu,1}(x,u)] = \mathbf{E}[\nabla f(x+\nu u)] = \nabla f_\nu(x)$, where we denote $f_\nu(x) = \mathbf{E}[f(x+\nu u)]$. Under Assumption 1.1 on smoothness of $f$, in the case where $n = 1$, we have the following calculation for the bias.

$$\begin{aligned}\|\mathbf{E}[\zeta_t \mid x_t]\|^2 &= \|\mathbf{E}[\nabla f(x_t + \nu u) \mid x_t] - \nabla f(x_t)\|^2 \\ &\le \mathbf{E}[(M\nu\|u\|)^2] \\ &\le M^2\nu^2 d,\end{aligned}$$

In order to obtain the variance, we split the centered error term into three parts.

$$\begin{aligned}\zeta_t - \mathbf{E}[\zeta_t \mid x_t] &= \frac{f(x_t+\nu u) - f(x_t)}{\nu} u - \nabla f_\nu(x_t) \\ &= \frac{f(x_t+\nu u) - f(x_t) - \nu u^\top \nabla f(x_t)}{\nu} u + (uu^\top - I)\nabla f(x_t) + (\nabla f(x_t) - \nabla f_\nu(x_t)) \\ &\overset{def}{=} A + B + C.\end{aligned}$$

Note that $\mathbf{E}[\|\zeta_t - \mathbf{E}[\zeta_t \mid x_t]\|^2 \mid x_t] = \mathbf{E}[\|A+B+C\|^2 \mid x_t]$Also, we have the following observation.

$$\begin{aligned}\mathbf{E}[\|A\|^2 \mid x_t] &= \mathbf{E}\left[\left\|\frac{f(x_t+\nu u) - f(x_t) - \nu u^\top \nabla f(x_t)}{\nu} u\right\|^2 \middle| x_t\right] \\ &\le \mathbf{E}[(\tfrac{1}{2}M\nu\|u\|^2)^2\|u\|^2 \mid x_t] \\ &\le \frac{1}{4}M^2\nu^2(d+2)^3, \\ \mathbf{E}[\|B\|^2 \mid x_t] &= \mathbf{E}[\|(uu^\top - I)\nabla f(x_t)\|^2 \mid x_t] \\ &= \nabla f(x_t)^\top \mathbf{E}_u\left[(\|u\|^2 - 2)uu^\top + I\right]\nabla f(x_t) \\ &= \nabla f(x_t)^\top \mathbf{E}_u\left[(\|u\|^2 - 2)(uu^\top - I) + (\|u\|^2 - 1)I\right]\nabla f(x_t) \\ &= \nabla f(x_t)^\top \left(2I + (d-1)I\right)\nabla f(x_t) \\ &= (d+1)\|\nabla f(x_t)\|^2.\end{aligned}$$

Combining with the fact that $C$ is deterministic and Lemma 3 from [35], the variance is bounded by

$$\mathbf{E}[\|\zeta_t - \mathbf{E}[\zeta_t \mid x_t]\|^2] \le \left(M\nu(d+2)^{3/2} + (d+1)^{1/2}\|\nabla f(x_t)\|_{L_2}\right)^2.$$

Next, for $n \ge 1$ in general, $g_{\nu,n}(x) = \frac{1}{n}\sum_{k=1}^n g_{\nu,1}(x,u_k)$, the bias and variance could be calculated as follows. Specifically, for the bias, we have

$$\begin{aligned}\|\mathbf{E}[\zeta_t \mid x_t]\|^2 &= \|\mathbf{E}[g_{\nu,n}(x_t) - \nabla f(x_t) \mid x_t]\|^2 \\ &\le \|\mathbf{E}[g_{\nu,1}(x_t) - \nabla f(x_t) \mid x_t]\|^2 \\ &\le M^2\nu^2 d.\end{aligned}$$

For the variance, by independence of Gaussian sample $u_i$'s, we have the following observation.

$$\begin{aligned}\mathbf{E}[\|\zeta_t - \mathbf{E}[\zeta_t \mid x_t]\|^2] &= \mathbf{E}[\|g_{\nu,n}(x_t) - \nabla f_\nu(x_t)\|^2] \\ &= \frac{1}{n}\mathbf{E}[\|g_{\nu,1}(x_t) - \nabla f_\nu(x_t)\|^2] \\ &\le \frac{1}{n}\left(M\nu(d+2)^{3/2} + (d+1)^{1/2}\|\nabla f(x_t)\|_{L_2}\right)^2.\end{aligned}$$

Next, we follow a similar framework to the proof of Theorem 4 in [14], but with modifications to adapt to the variance that is not uniformly bounded. Recall that $\Delta_t = L_0 - x_t$, $\Delta_{t+1} = L_h - x_{t+1}$, where $L_T = L_0 - \int_0^T \nabla f(L_s)ds + \sqrt{2}W_T$ follows the Langevin diffusion with stationary distribution $\pi$. Moreover, $\|\Delta_t - hU\| = \|\Delta_t - h[\nabla f(x_t+\Delta_t) - \nabla f(x_t)]\| \le (1-mh)\|\Delta_t\|$, $\|V\| = \|\int_0^h [\nabla f(L_s) - \nabla f(L_0)]ds\| \le 1.65M(h^3d)^{1/2}$. Thus,

$$\begin{aligned}\|\Delta_{t+1}\|_{L_2} &= \|\Delta_t - hU - V + h\zeta_t\|_{L_2} \\ &\le \{\|\Delta_t - hU\|_{L_2}^2 + h^2\|\zeta_t - \mathbf{E}[\zeta_t \mid x_t]\|_{L_2}^2\}^{1/2} + \|V\|_{L_2} + h\|\mathbf{E}[\zeta_t \mid x_t]\|_{L_2} \\ &\le \left\{(1-mh)^2\|\Delta_t\|_{L_2}^2 + \frac{h^2}{n}\left(M\nu(d+2)^{3/2} + (\|\nabla f(L_0)\|_{L_2} + M\|\Delta_t\|_{L_2})(d+1)^{1/2}\right)^2\right\}^{1/2} \\ &\quad + 1.65M(h^3d)^{1/2} + M\nu h d^{1/2} \\ &\le \left\{(1-mh)^2\|\Delta_t\|_{L_2}^2 + \frac{2h^2}{n}\left(M\nu(d+2)^{3/2} + \|\nabla f(L_0)\|_{L_2}(d+1)^{1/2}\right)^2\right\}^{1/2} \\ &\quad + \frac{M^2h^2(d+1)}{n(1-mh)}\|\Delta_t\|_{L_2} + 1.65M(h^3d)^{1/2} + M\nu h d^{1/2} \\ &\le \left\{(1-mh)^2\|\Delta_t\|_{L_2}^2 + \frac{2h^2}{n}\left(M\nu(d+2)^{3/2} + \sqrt{M}(d+1)\right)^2\right\}^{1/2} + \frac{1}{2}mh\|\Delta_t\|_{L_2} \\ &\quad + 1.65M(h^3d)^{1/2} + M\nu h d^{1/2}.\end{aligned}$$

Here we use the fact that $\sqrt{a^2+b+c} \le \sqrt{a^2+b} + \frac{c}{2a}$ and $\mathbf{E}[\|\nabla f(L)\|^2] \le Md$. By Lemma 9 in [14], the above inequality leads to

$$\begin{aligned}\|\Delta_t\|_{L_2} &\le (1-0.5mh)^t\|\Delta_0\|_{L_2} + 3.3\frac{M}{m}(hd)^{1/2} + 2\frac{M}{m}\nu d^{1/2} \\ &\quad + 2\sqrt{2}\frac{\sqrt{M}}{\sqrt{m}}\cdot\frac{1}{\sqrt{n}}h^{1/2}(d+1) + \sqrt{2}\frac{M}{\sqrt{m}}\cdot\frac{\nu}{\sqrt{n}}h^{1/2}(d+2)^{3/2}.\end{aligned}$$

Therefore, we obtain the bound in Wasserstein distance.

$$\begin{aligned}W_2(\varpi_t, \pi) &\le (1-0.5mh)^t\, W_2(\varpi_0, \pi) + 3.3\frac{M}{m}(hd)^{1/2} + 2\frac{M}{m}\nu d^{1/2} \\ &\quad + 2\sqrt{2}\frac{\sqrt{M}}{\sqrt{m}}\cdot\frac{1}{\sqrt{n}}h^{1/2}(d+1) + \sqrt{2}\frac{M}{\sqrt{m}}\cdot\frac{\nu}{\sqrt{n}}h^{1/2}(d+2)^{3/2}.\end{aligned}$$

■

Note in particular, in the case where $n = 1$, we have the following non-asymptotic results.

$$W_2(\varpi_t, \pi) \le (1-0.5mh)^t\, W_2(\varpi_0, \pi) + 5.17\frac{M}{m}h^{1/2}(d+1) + 2\frac{M}{m}\nu d^{1/2} + \sqrt{2}\frac{M}{\sqrt{m}}\nu h^{1/2}(d+2)^{3/2}.$$

## D Proofs for Section 4

*Proof.* [Proof of Theorem 4.3] First, we have,

$$\begin{aligned}
\Pr\{\hat{S} \neq S^*\} &= \Pr\{\max_{j\in D\setminus S^*} |[g_{\nu,n}]_j| > \tau \text{ or } \min_{j\in S^*} |[g_{\nu,n}]_j| < \tau\} \\
&\leq \Pr\{\max_{j\in D\setminus S^*} |[g_{\nu,n}]_j| > \tau\} + \Pr\{\min_{j\in S^*} |[g_{\nu,n}]_j| < \tau\} \\
&\leq \sum_{j\in D\setminus S^*} \Pr\{|\zeta_j| > \tau\} + \sum_{j\in S^*} \Pr\{|\zeta_j| > a' - \tau\},
\end{aligned}$$

where $a' = a - M\nu\sqrt{s} \leq a - \|\nabla f(\theta) - \nabla f_\nu(\theta)\|$ is a lower bound for $|[\nabla f_\nu(\theta)]_j|$. Next we utilize concentration inequalities to give a bound for the tail of approximation error $\zeta_j$. Denote $[g_{\nu,1}]_j = \frac{f(\theta+\nu u)-f(\theta)}{\nu} u_j \overset{def}{=} \phi(\nu,u)u_j$, where $\phi(\nu,u)$ is sub-exponential with

$$\begin{aligned}
\|\phi(\nu,u)\|_{\Psi_1} &= \sup_{p\geq 1} p^{-1}(\mathbf{E}[|\phi(\nu,u)|^p])^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}(\mathbf{E}[|\frac{f(\theta+\nu u) - f(\theta) - \nabla f(\theta)^\top \nu u}{\nu}|^p])^{1/p} + \sup_{p\geq 1} p^{-1}(\mathbf{E}[|\nabla f(\theta)^\top u|^p])^{1/p} \\
&\leq \frac{1}{2} M\nu \sup_{p\geq 1} p^{-1}(\mathbf{E}[\|u\|^{2p}])^{1/p} + \|\nabla f(\theta)\| \sup_{p\geq 1} p^{-1}(\mathbf{E}[\|u\|^p])^{1/p} \\
&\leq M\nu\|u\|_{\Psi_2}^2 + \|\nabla f(\theta)\|\|u\|_{\Psi_2} \\
&\leq 2R\|u\|_{\Psi_2},
\end{aligned}$$

where $\|\cdot\|_{\Psi_1} = \sup_{p\geq 1} p^{-1}\mathbf{E}[|\cdot|^p]^{1/p}$ and $\|\cdot\|_{\Psi_2} = \sup_{p\geq 1} p^{-1/2}\mathbf{E}[|\cdot|^p]^{1/p}$ are the sub-exponential and sub-Gaussian norm respectively (see, for example [48] for more details). In the last inequality we require that $\nu \leq \frac{R}{M\|u\|_{\Psi_2}}$. Note that $u \sim N(0, \boldsymbol{I}_d)$ can be replaced by $\sum_{k\in S^*} u_k e_k \sim N(0, \boldsymbol{I}_s)$ due to Assumption 4.1. Moreover, we have the following estimate.

$$\begin{aligned}
\|u_1\|_{\Psi_2} &\leq \inf\{c > 0 : \mathbf{E}\left[\exp\left\{\frac{u_1^2}{c^2}\right\}\right] \leq 2\} = \sqrt{\frac{8}{3}} \overset{def}{=} C_1, \\
\|u\|_{\Psi_2} &\leq \inf\{c > 0 : \mathbf{E}\left[\exp\left\{\frac{\|u\|^2}{c^2}\right\}\right] \leq 2\} \\
&= \sqrt{\frac{2}{1-2^{-2/d}}} \\
&\leq \sqrt{\frac{d}{\log 2(1-\log 2)}} \overset{def}{=} C_2\sqrt{d},
\end{aligned}$$

which implies that $\|\phi(\nu,u)\|_{\Psi_1} \leq 2RC_2\sqrt{s}$, $\|u_1\|_{\Psi_2} \leq C_1$. We now state the following concentration inequality proved in [2].

**Lemma D.1** *Let $(X_i, Y_i)$, $i = 1,\ldots,n$ be $n$ independent copies of random variables $X$ and $Y$. Let $X$ be a sub-Gaussian random variable with $\|X\|_{\psi_2} \leq \Upsilon_1$, and $Y$ be a sub-exponential random variable with $\|Y\|_{\psi_1} \leq \Upsilon_2$ for some constants $\Upsilon_1$ and $\Upsilon_2$. Then for any $t \geq K \cdot \max\{\Upsilon_1^3, \Upsilon_1\} \cdot \Upsilon_2$,* *we have*

$$Pr\Big\{\Big|\sum_{i=1}^{n}\big[X_i\cdot Y_i-\mathbf{E}(XY)\big]\Big|\geq t\Big\}\leq 4\text{exp}\Big\{-K_1\cdot\min\Big[\Big(\frac{t}{\sqrt{n}\Upsilon_1\cdot\Upsilon_2}\Big)^2,\Big(\frac{t}{\Upsilon_1\cdot\Upsilon_2}\Big)^{2/3}\Big]\Big\},$$

*where $K$ and $K_1$ are absolute constants.*

From Lemma D.1, for $n\geq\max\left\{K_1\frac{2RC\sqrt{s}}{\tau},\left(\frac{2RC\sqrt{s}}{\tau}\right)^4\right\}$, we have:

$$\begin{aligned}\Pr\{|\zeta_j|\geq\tau\}&=\Pr\left\{\left|\frac{1}{n}\sum_{k=1}^{n}g_{\nu,1}^{k}-\mathbf{E}[g_{\nu,1}]\right|\geq\tau\right\}\\&\leq 4\text{exp}\left\{-K_2\left(\frac{n\tau}{\|\phi(\nu,u)\|_{\Psi_1}\|u_1\|_{\Psi_2}}\right)^{2/3}\right\}\\&\leq 4\text{exp}\left\{-K_2\left(\frac{n\tau}{2RC\sqrt{s}}\right)^{2/3}\right\},\end{aligned}$$

where $C=C_1C_2=\sqrt{\frac{8}{3\log 2(1-\log 2)}}, K_1,K_2$ are absolute constants. Therefore, by setting the threshold $\tau=a'/2$, the probability of error is bounded by

$$\begin{aligned}\Pr\{\hat{S}\neq S^*\}&\leq\sum_{j\in D\setminus S^*}\Pr\{|\zeta_j|>\tau\}+\sum_{j\in S^*}\Pr\{|\zeta_j|>a'-\tau\}\\&\leq 4(d-s)\text{exp}\left\{-K_2\left(\frac{n\tau}{2RC\sqrt{s}}\right)^{2/3}\right\}+4s\text{exp}\left\{-K_2\left(\frac{n(a'-\tau)}{2RC\sqrt{s}}\right)^{2/3}\right\}\\&=4d\text{exp}\left\{-K_2\left(\frac{n(a-M\nu\sqrt{s})}{4RC\sqrt{s}}\right)^{2/3}\right\}.\end{aligned}$$

Given a pre-specified error rate $\epsilon>0$, it suffices to have $\nu\leq\frac{a}{2M\sqrt{s}}\wedge\frac{R}{MC_2\sqrt{s}}$ and

$$n\geq\frac{8RC\sqrt{s}}{a}\left(\frac{1}{K_2}\log\frac{4d}{\epsilon}\right)^{3/2}\vee K_1\frac{8RC\sqrt{s}}{a}\vee\left(\frac{8RC\sqrt{s}}{a}\right)^4.$$

■

# E Proofs for Section 5.1

*Proof.* [Proof of Theorem 5.3] Define random processes $D_{t,T}$, $L_{t,T}$ recursively as follows for $t\in\mathbb{N}$, $T\in[0,h]$. First, take $D_{0,0}=x_0$ to be deterministic, and $L_{0,0}\sim\pi$ such that $(D_{0,0},L_{0,0})$ is the optimal coupling that attains the Wasserstein distance $W_2(\varpi_0,\pi)=\|D_{0,0}-L_{0,0}\|_{L_2}$. Next, for each $t\in\mathbb{N}$, let $L_{t,T}$ be the Langevin diffusion driven by the Brownian motion $W_{t,T}$,

$$dL_{t,T}=-\nabla f(L_{t,T})dT+\sqrt{2}dW_{t,T},$$

starting from $L_{t,0} = L_{t-1,h}$. Since $\pi$ is the stationary distribution, we have $L_{t,T} \sim \pi$. $D_{t,T}$ is defined by the SDE

$$dD_{t,T} = -\left[g_t + \boldsymbol{S}_t(D_{t,T} - D_{t,0})\right] dT + \sqrt{2} dW_{t,T}.$$

The Ornstein-Uhlenbeck process can be solved explicitly as

$$D_{t,h} = D_{t,0} - \left(\boldsymbol{I}_d - e^{-h\boldsymbol{S}_t}\right) \boldsymbol{S}_t^{-1} g_t + \left(\left(\boldsymbol{I}_d - e^{-2h\boldsymbol{S}_t}\right) \boldsymbol{S}_t^{-1}\right)^{1/2} N[0, \boldsymbol{I}_d],$$

which indicates that $D_{t,h} = D_{t+1,0} \sim \varpi_{t+1}$. Here we require the common term of noise $W_{k,T}$ to be independent of $\boldsymbol{S}_t$ and $g_t$ conditionally on $D_{t,0}$ for $k \in \mathbb{N}$, $T \in [0, h]$, and moreover, $W_{t,T}$ is independent of $D_{t,0}$ and $L_{t,0}$ for $T \in [0, h]$. To ease the notation, we drop the first subscript when considering the current time step $t$. Define $\Delta_T = L_T - D_T$ and $X_T = (L_T - L_0) - (D_T - D_0) = \Delta_T - \Delta_0$. Then

$$\begin{aligned}
X_T &= -\int_0^T \nabla f(L_s) ds + \int_0^T \left[g_t + \boldsymbol{S}_t(D_s - D_0)\right] ds \\
&= -\int_0^T \left\{\nabla f(L_s) ds - g_t - \boldsymbol{S}_t(L_s - L_0)\right\} ds - \int_0^T \boldsymbol{S}_t X_s ds.
\end{aligned}$$

By Gronwall lemma (see Lemma 5 in [14]), we have

$$\begin{aligned}
X_T &= -\int_0^T e^{-s\boldsymbol{S}_t} \left\{\nabla f(L_s) - g_t - \boldsymbol{S}_t(L_s - L_0)\right\} ds \\
&= \int_0^T e^{-s\boldsymbol{S}_t} ds \left[\nabla f(D_0) - \nabla f(L_0)\right] \\
&\quad + \int_0^T e^{-s\boldsymbol{S}_t} ds \left[g_t - \mathbf{E}[g_t|D_0]\right] + \int_0^T e^{-s\boldsymbol{S}_t} ds \left[\mathbf{E}[g_t|D_0] - \nabla f(D_0)\right] \\
&\quad - \int_0^T e^{-s\boldsymbol{S}_t} \left\{\nabla f(L_s) - \nabla f(L_0) - \nabla^2 f(L_0)(L_s - L_0)\right\} ds \\
&\quad - \int_0^T e^{-s\boldsymbol{S}_t} [\boldsymbol{S}_t - \nabla^2 f(L_0)] \int_0^s \nabla f(L_u) du ds \\
&\quad + \sqrt{2} \int_0^T e^{-s\boldsymbol{S}_t} [\nabla^2 f(D_0) - \nabla^2 f(L_0)] W_s ds \\
&\quad + \sqrt{2} \int_0^T e^{-s\boldsymbol{S}_t} [\boldsymbol{S}_t - \nabla^2 f(D_0)] W_s ds \\
&\stackrel{def}{=} A_T + I_T + J_T - B_T - C_T + P_T + Q_T.
\end{aligned}$$

We now consider the two cases of $\boldsymbol{S}_t$ separately.
**Case 1:** $\boldsymbol{S}_t \succeq 0$.
By calculations similar to that in proof of Theorem 6 in [14], under the independence assumptions,

we have the following bounds for each of the above terms.

$$\begin{aligned}
\|\Delta_0 + A_T\|_{L_2} &\leq (1 - mT + 0.5M\bar{M}T^2)\|\Delta_0\|_{L_2} \\
&= (1 - mT + 0.5\tilde{M}^2T^2)\|\Delta_0\|_{L_2}. \\
\|I_T\|_{L_2} &\leq T\mathcal{C}_1(d). \\
\|J_T\|_{L_2} &\leq T\mathcal{C}_2(d). \\
\|B_T\|_{L_2} &\leq 0.877M_2T^2(d^2 + 2d)^{1/2}. \\
\|C_T\|_{L_2} &\leq \mu\|\Delta_0\|_{L_2} + \frac{1}{16\mu}M^2M_2T^4(d+1) + \frac{1}{2}\sqrt{M}T^2d^{1/2}\mathcal{C}_3(d). \\
\|P_T\|_{L_2}^2 &\leq \frac{2}{3}MM_2T^3d\|\Delta_0\|_{L_2}. \\
\|Q_T\|_{L_2}^2 &\leq \frac{2}{3}T^3d\mathcal{C}_3(d)^2.
\end{aligned}$$

Hence we have, for $h \leq m/\tilde{M}^2$,

$$\begin{aligned}
\|\Delta_h\|_{L_2} &= \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h\|_{L_2} \\
&\leq (\|\Delta_0 + A_h\|_{L_2}^2 + \|I_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|J_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|Q_h\|_{L_2} \\
&\leq \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + h^2\mathcal{C}_1^2(d) + \frac{2}{3}MM_2h^3d\|\Delta_0\|_{L_2}\}^{1/2} \\
&\quad + h\mathcal{C}_2(d) + 0.877M_2h^2(d+1) + \mu\|\Delta_0\|_{L_2} + \frac{1}{16\mu}M^2M_2h^4(d+1) \\
&\quad + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_3(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_3(d) \\
&\leq \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + h^2\mathcal{C}_1^2(d)\}^{1/2} + \frac{MM_2h^3d}{3(1 - mh + 0.5\tilde{M}^2h^2)} \\
&\quad + h\mathcal{C}_2(d) + 0.877M_2h^2(d+1) + \mu\|\Delta_0\|_{L_2} + \frac{1}{16\mu}M^2M_2h^4(d+1) \\
&\quad + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_3(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_3(d) \\
&\leq \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + h^2\mathcal{C}_1^2(d)\}^{1/2} + \frac{2}{3}MM_2h^3d \\
&\quad + h\mathcal{C}_2(d) + 0.877M_2h^2(d+1) + \frac{1}{4}mh\|\Delta_0\|_{L_2} + \frac{M^2M_2}{4m}h^3(d+1) \\
&\quad + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_3(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_3(d) \\
&\leq \{(1 - mh + 0.5\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + h^2\mathcal{C}_1^2(d)\}^{1/2} + 0.25mh\|\Delta_0\|_{L_2} \\
&\quad + 1.8M_2h^2(d+1) + h\mathcal{C}_2(d) + 1.32h^{3/2}d^{1/2}\mathcal{C}_3(d),
\end{aligned}$$

where $\mathbf{E}[(\Delta_0 + A_h)^\top I_h] = \mathbf{E}[(\Delta_0 + A_h)^\top J_h] = \mathbf{E}[I_h^\top J_h] = 0$ due to the independence assumption 5.2. Also, we use the inequality $\sqrt{a^2 + b + c} \leq \sqrt{a^2 + b} + \frac{c}{2a}$, note that $1 - mh + 0.5\tilde{M}^2h^2 \geq 0.5$, and take $\mu = 0.25mh$. Next, by application of Lemma 9 in [14], where $A - D = 0.75mh - 0.5\tilde{M}^2h^2 \geq$

$0.25mh,\ A+D = 1.25mh - 0.5\tilde{M}^2h^2 \le 0.75$, we have

$$\begin{aligned}\|\Delta_{t,0}\|_{L_2} &\le (1-0.25mh)^t\|\Delta_{0,0}\|_{L_2} + \frac{7.18M_2}{m}h(d+1)\\ &\quad + \frac{4}{\sqrt{5m}}h^{1/2}\mathcal{C}_1(d) + \frac{4}{m}\mathcal{C}_2(d) + \frac{5.27}{m}(hd)^{1/2}\mathcal{C}_3(d).\end{aligned}$$

Since $W_2(\varpi_t,\pi) \le \|D_{t,0}-L_{t,0}\|_{L_2} = \|\Delta_{t,0}\|_{L_2}$, and in particular, equality holds for $t=0$ by our choice of $L_{0,0}$, we obatin the bound in Wasserstein distance. Note that it can be reduced to the case of exact oracles, i.e., Equation (17) in Theorem 6 of [14].

**Case 2:** $\boldsymbol{S}_t \succeq 0$ does not hold in general.
Now we have a different estimate for the following bounds.

$$\begin{aligned}\|\Delta_0 + A_T\|_{L_2} &\le (1-mT+0.5\tilde{M}^2T^2)\|\Delta_0\|_{L_2}\\ \|I_T\|_{L_2} &\le T\mathcal{C}_1(d)\mathcal{C}_4(d)\\ \|J_T\|_{L_2} &\le T\mathcal{C}_2(d)\mathcal{C}_4(d)\\ \|B_T\|_{L_2} &\le 0.877M_2T^2(d^2+2d)^{1/2}\mathcal{C}_4(d)\\ \|C_T\|_{L_2} &\le \mu\|\Delta_0\|_{L_2} + \frac{1}{12\mu}M^2M_2T^4(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}T^2d^{1/2}\mathcal{C}_5(d)\\ \|P_T\|_{L_2}^2 &\le \frac{2}{3}MM_2T^3d\|\Delta_0\|_{L_2}\mathcal{C}_4(d)^2\\ \|Q_T\|_{L_2}^2 &\le \frac{2}{3}T^3d\mathcal{C}_5(d)^2.\end{aligned}$$

Hence, for $h \le m/\tilde{M}^2$, we have

$$\begin{aligned}\|\Delta_h\|_{L_2} &= \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h\|_{L_2}\\ &\le (\|\Delta_0 + A_h\|_{L_2}^2 + \|I_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|J_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|Q_h\|_{L_2}\\ &\le \{(1-mh+\frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + h^2\mathcal{C}_1(d)^2\mathcal{C}_4(d)^2 + \frac{2}{3}MM_2h^3d\mathcal{C}_4(d)^2\|\Delta_0\|_{L_2}\}^{1/2}\\ &\quad + h\mathcal{C}_2(d)\mathcal{C}_4(d) + 0.877M_2h^2(d^2+2d)^{1/2}\mathcal{C}_4(d) + \mu\|\Delta_0\|_{L_2}\\ &\quad + \frac{1}{12\mu}M^2M_2h^4(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_5(d)\\ &\le \{(1-mh+\frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + h^2\mathcal{C}_1(d)^2\mathcal{C}_4(d)^2\}^{1/2} + \frac{MM_2h^3d\mathcal{C}_4(d)^2}{3(1-mh+0.5\tilde{M}^2h^2)}\\ &\quad + h\mathcal{C}_2(d)\mathcal{C}_4(d) + 0.877M_2h^2(d^2+2d)^{1/2}\mathcal{C}_4(d) + 0.25mh\|\Delta_0\|_{L_2}\\ &\quad + \frac{M^2M_2}{3m}h^3(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_5(d)\\ &\le \{(1-mh+\frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + h^2\mathcal{C}_4(d)^2\mathcal{C}_1(d)^2\}^{1/2} + 0.25mh\|\Delta_0\|_{L_2}\\ &\quad + \frac{M^2M_2}{m}h^3(d+1)\mathcal{C}_4(d)^2 + \big(0.877M_2h^2(d+1) + h\mathcal{C}_2(d)\big)\mathcal{C}_4(d) + 1.32h^{3/2}d^{1/2}\mathcal{C}_5(d),\end{aligned}$$

where the calculation is similar to case 1. Again application of Lemma 9 in [14] finally leads to

$$\begin{aligned} W_2(\varpi_t, \pi) \leq & (1-0.25mh)^t W_2(\varpi_0, \pi) + 4\frac{M^2 M_2}{m^2} h^2 (d+1) \mathcal{C}_4(d)^2 \\ & + \left(3.51\frac{M_2}{m} h(d+1) + \frac{4}{\sqrt{5m}} h^{1/2} \mathcal{C}_1(d) + \frac{4}{m}\mathcal{C}_2(d)\right) \mathcal{C}_4(d) + \frac{5.27}{m}(hd)^{1/2}\mathcal{C}_5(d). \end{aligned}$$

Note that in this case, the bound is slightly more conservative in the second term by a factor of constant. ■

# F Proof for section 5.2

*Proof.* [Proof of Theorem 5.4] Consider the same settings as in the proof of Theorem 5.3, except that $D_{t,T}$ is now defined by

$$D_{t,T} - D_{t,0} = -\left(T\boldsymbol{I}_d - \frac{1}{2}T^2 \boldsymbol{S}_t\right) g_t + \sqrt{2}\int_0^T \boldsymbol{I}_d - (T-u)\boldsymbol{S}_t dW_{t,u}.$$

From the representation

$$D_{t,h} = D_{t,0} - h\left(\boldsymbol{I}_d - \frac{1}{2}h\boldsymbol{S}_t\right) g_t + \sqrt{2h}\left(\boldsymbol{I}_d - h\boldsymbol{S}_t + \frac{1}{3}h^2\boldsymbol{S}_t^2\right) N[0, \boldsymbol{I}_d],$$

we know that $D_{t+1,0} = D_{t,h} \sim \varpi_{t+1}$, which is the distribution of $x_{t+1}$. On the other hand, $D_{t,T}$ satisfies the following SDE, and can be further written as

$$\begin{aligned} dD_{t,T} =& -(\boldsymbol{I}_d - T\boldsymbol{S}_t) g_t dT - \sqrt{2}\boldsymbol{S}_t W_{t,T} dT + \sqrt{2} dW_{t,T} \\ =& -[g_t + \boldsymbol{S}_t(D_{t,T} - D_{t,0})]dT + \sqrt{2}dW_{t,T} \\ & + T\boldsymbol{S}_t g_t dT - \sqrt{2}\boldsymbol{S}_t W_{t,T} dT \\ & - \boldsymbol{S}_t(T\boldsymbol{I}_d - \frac{1}{2}T^2\boldsymbol{S}_t) g_t dT + \sqrt{2}\int_0^T \boldsymbol{S}_t[\boldsymbol{I}_d - (T-u)\boldsymbol{S}_t] dW_{t,u} dT \\ =& -[g_t + \boldsymbol{S}_t(D_T - D_0)]dT + \sqrt{2}dW_{t,T} \\ & + \frac{1}{2}T^2\boldsymbol{S}_t^2 g_t dT - \sqrt{2}\boldsymbol{S}_t^2 \int_0^T (T-u) dW_{t,u} dT. \end{aligned}$$

Recall $X_T = (L_T - L_0) - (D_T - D_0)$. Now

$$\begin{aligned} X_T =& -\int_0^T \{\nabla f(L_s) ds - g_t - \boldsymbol{S}_t(L_s - L_0)\} ds - \int_0^T \boldsymbol{S}_t X_s ds \\ & + \frac{1}{2}\int_0^T s^2 \boldsymbol{S}_t^2 g_t ds - \sqrt{2}\int_0^T \boldsymbol{S}_t^2 \int_0^s (s-u) dW_u ds. \end{aligned}$$

By Gronwall lemma, $X_T$ is solved as

$$\begin{aligned} X_T =& -\int_0^T e^{-s\boldsymbol{S}_t}\{\nabla f(L_s) - g_t - \boldsymbol{S}_t(L_s - L_0)\} ds \\ & + \frac{1}{2}\int_0^T e^{-s\boldsymbol{S}_t} s^2 ds \boldsymbol{S}_t^2 g_t - \sqrt{2}\boldsymbol{S}_t^2 \int_0^T e^{-s\boldsymbol{S}_t}\int_0^s (s-u) dW_u ds \\ \stackrel{def}{=}& (A_T + I_T + J_T - B_T - C_T + P_T + Q_T) + E_T - F_T, \end{aligned}$$

where the first term coincides with $X_T$ in Theorem 5.3, and the extra terms can be viewed as errors resulting from approximation.

**Case 1:** $\boldsymbol{S}_t \succeq 0$.

By the independence assumptions, we have the following estimate for the extra terms $E_T$ and $F_T$.

$$\|E_T\|_{L_2} \le \frac{1}{6}\hat{M}^2T^3\left(\sqrt{M}d^{1/2} + M\|\Delta_0\|_{L_2} + \|g_t - \nabla f_t\|_{L_2}\right).$$

$$\|F_T\|_{L_2} \le \frac{1}{\sqrt{10}}\hat{M}^2T^{5/2}d^{1/2}.$$

Proceeding as before, for $h \le 3m/(4M\hat{M}) \wedge 3m/(4\tilde{M}^2)$, we have

$$\begin{aligned}
\|\Delta_h\|_{L_2} &= \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h + E_h - F_h\|_{L_2} \\
&\le (\|\Delta_0 + A_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|I_h + J_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} \\
&\qquad + \|Q_h\|_{L_2} + \|E_h\|_{L_2} + \|F_h\|_{L_2} \\
&\le \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{2}{3}MM_2dh^3\|\Delta_0\|_{L_2}\}^{1/2} \\
&\quad + h\left(\mathcal{C}_1(d) + \mathcal{C}_2(d)\right) + 0.877M_2h^2(d+1) + \mu\|\Delta_0\|_{L_2} + \frac{1}{16\mu}M^2M_2h^4(d+1) \\
&\quad + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_3(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_3(d) \\
&\quad + \frac{1}{6}\hat{M}^2h^3\left(\sqrt{M}d^{1/2} + M\|\Delta_0\|_{L_2} + \mathcal{C}_1(d) + \mathcal{C}_2(d)\right) + \frac{1}{\sqrt{10}}\hat{M}^2h^{5/2}d^{1/2} \\
&\le (1 - mh + 0.5\tilde{M}^2h^2)\|\Delta_0\|_{L_2} + \frac{MM_2h^3d}{3(1 - mh + 0.5\tilde{M}^2h^2)} + h\left(\mathcal{C}_1(d) + \mathcal{C}_2(d)\right) \\
&\quad + 0.877M_2h^2(d+1) + \frac{1}{4}mh\|\Delta_0\|_{L_2} + \frac{M^2M_2}{4m}h^3(d+1) + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_3(d) \\
&\quad + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_3(d) + \frac{1}{6}\hat{M}^2h^3\left(\sqrt{M}d^{1/2} + M\|\Delta_0\|_{L_2} + \mathcal{C}_1(d) + \mathcal{C}_2(d)\right) + \frac{1}{\sqrt{10}}\hat{M}^2h^{5/2}d^{1/2} \\
&\le (1 - 0.25mh)\|\Delta_0\|_{L_2} + 1.54M_2h^2(d+1) + 0.47\hat{M}^2h^{5/2}d^{1/2} \\
&\quad + 1.10h\left(\mathcal{C}_1(d) + \mathcal{C}_2(d)\right) + 1.25h^{3/2}d^{1/2}\mathcal{C}_3(d),
\end{aligned}$$

where we use the inequality $\sqrt{a^2+b} \le a + \frac{b}{2a}$. Note also that $1 - mh + \frac{1}{2}\tilde{M}^2h^2 \ge \frac{17}{32}$, and $1 - \frac{3}{4}mh + \frac{1}{2}\tilde{M}^2h^2 + \frac{1}{6}M\hat{M}^2h^3 \le 1 - \frac{1}{4}mh$. Recursively applying the above result, we obtain

$$\begin{aligned}
W_2(\varpi_t, \pi) &\le (1 - 0.25mh)^tW_2(\varpi_0, \pi) + 6.15\frac{M_2}{m}h(d+1) + 1.85\frac{\hat{M}^2}{m}h^{3/2}d^{1/2} \\
&\qquad + \frac{4.38}{m}\left(\mathcal{C}_1(d) + \mathcal{C}_2(d)\right) + \frac{5}{m}(hd)^{1/2}\mathcal{C}_3(d).
\end{aligned}$$

**Case 2:** $\boldsymbol{S}_t \succeq 0$ does not hold in general.

For $\hat{M}$ defined in the corresponding case, now the bounds for $E_T$ and $F_T$ are

$$\|E_T\|_{L_2} \le \frac{1}{6}\hat{M}^2T^3\left(\sqrt{M}d^{1/2} + M\|\Delta_0\|_{L_2} + \|g_t - \nabla f_t\|_{L_2}\right).$$

$$\|F_T\|_{L_2} \le \frac{1}{\sqrt{10}}T^{5/2}d^{1/2}\mathcal{C}_6(d).$$

The following calculation goes, for $h \le 3m/(4M\hat{M}) \wedge 3m/(4\tilde{M}^2)$,

$$\begin{aligned}
\|\Delta_h\|_{L_2} &= \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h + E_h - F_h\|_{L_2} \\
&\le (\|\Delta_0 + A_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|I_h + J_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|Q_h\|_{L_2} + \|E_h\|_{L_2} + \|F_h\|_{L_2} \\
&\le \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{2}{3}MM_2h^3d\mathcal{C}_4(d)^2\|\Delta_0\|_{L_2}\}^{1/2} \\
&\quad + h\left(\mathcal{C}_1(d) + \mathcal{C}_2(d)\right)\mathcal{C}_4(d) + 0.877M_2h^2(d^2+2d)^{1/2}\mathcal{C}_4(d) + \mu\|\Delta_0\|_{L_2} \\
&\quad + \frac{1}{12\mu}M^2M_2h^4(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_5(d) \\
&\quad + \frac{1}{6}\hat{M}^2h^3\left(\sqrt{M}d^{1/2} + M\|\Delta_0\|_{L_2} + \mathcal{C}_1(d) + \mathcal{C}_2(d)\right) + \frac{1}{\sqrt{10}}h^{5/2}d^{1/2}\mathcal{C}_6(d) \\
&\le (1 - mh + 0.5\tilde{M}^2h^2)\|\Delta_0\|_{L_2} + \frac{MM_2h^3d}{3(1 - mh + 0.5\tilde{M}^2h^2)}\mathcal{C}_4(d)^2 + h\left(\mathcal{C}_1(d) + \mathcal{C}_2(d)\right)\mathcal{C}_4(d) \\
&\quad + 0.877M_2h^2(d+1)\mathcal{C}_4(d) + \frac{1}{4}mh\|\Delta_0\|_{L_2} + \frac{M^2M_2}{3m}h^3(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_5(d) \\
&\quad + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_5(d) + \frac{1}{6}\hat{M}^2h^3\left(\sqrt{M}d^{1/2} + M\|\Delta_0\|_{L_2} + \mathcal{C}_1(d) + \mathcal{C}_2(d)\right) + \frac{1}{\sqrt{10}}h^{5/2}d^{1/2}\mathcal{C}_6(d) \\
&\le (1 - 0.25mh)\|\Delta_0\|_{L_2} + 0.97\frac{M^2M_2}{m}h^3(d+1)\mathcal{C}_4(d)^2 + 1.25h^{3/2}d^{1/2}\mathcal{C}_5(d) \\
&\quad + \left(0.877M_2h^2(d+1) + h(\mathcal{C}_1(d) + \mathcal{C}_2(d))\right)\mathcal{C}_4(d) \\
&\quad + \left(0.47h^{5/2}d^{1/2} + \frac{1}{6}h^3(\mathcal{C}_1(d) + \mathcal{C}_2(d))\right)\mathcal{C}_6(d).
\end{aligned}$$

Therefore, we end up with

$$\begin{aligned}
W_2(\varpi_t, \pi) \le (1 - 0.25mh)^tW_2(\varpi_0, \pi)& \\
+ 3.85\frac{M^2M_2}{m^2}h^2(d+1)\mathcal{C}_4(d)^2 &+ \frac{5}{m}(hd)^{1/2}\mathcal{C}_5(d) \\
+ \left(3.51\frac{M_2}{m}h(d+1) + \frac{4}{m}(\mathcal{C}_1(d) + \mathcal{C}_2(d))\right)&\mathcal{C}_4(d) \\
+ \left(\frac{1.85}{m}h^{3/2}d^{1/2} + \frac{2}{3m}h^2(\mathcal{C}_1(d) + \mathcal{C}_2(d))\right)&\mathcal{C}_6(d).
\end{aligned}$$

■

# G Proofs for section 5.3

*Proof.* [Proof of Lemma 5.1] Under Assumption 5.1 that $f$ has Lipschitz smooth Hessian, we have

$$\begin{aligned}
\|\boldsymbol{H}_{f_\nu} - \boldsymbol{H}_f\|_2 &\le \|\mathbf{E}[\nabla^2 f(\theta + \nu u)] - \nabla^2 f(\theta)\|_2 \\
&\le \mathbf{E}[\|\nabla^2 f(\theta + \nu u) - \nabla^2 f(\theta)\|_2] \\
&\le \mathbf{E}[M_2\nu\|u\|] \le M_2\nu d^{1/2}.
\end{aligned}$$

■

*Proof.* [Proof of Lemma 5.2] Taking $\theta' = \theta + \nu u$ in Equation (16),

$$|f(\theta + \nu u) - f(\theta) - \nu \nabla^\top f(\theta) u - \frac{\nu^2}{2} u^\top \nabla^2 f(\theta) u| \leq \frac{M_2 \nu^3}{6} \|u\|^3. \tag{27}$$

Note also that $\|uu^\top - \boldsymbol{I}_d\|_2^2 \leq (\|u\|^2 - 1)^2 + 1$, $\|uu^\top - \boldsymbol{I}_d\|_F^2 = (\|u\|^2 - 1)^2 + d - 1$ and $\mathbf{E}[\|u\|^{2p}] = \frac{(d+2p-2)!!}{(d-2)!!}$. To apply (27), we split the error into two terms,

$$\begin{aligned}
\hat{\boldsymbol{H}}_{f_\nu} - \boldsymbol{H}_f &= \frac{1}{2\nu^2}(uu^\top - \boldsymbol{I}_d)[f(\theta + \nu u) - f(\theta) + f(\theta - \nu u) - f(\theta)] - \boldsymbol{H}_f \\
&= \frac{1}{2\nu^2}(uu^\top - \boldsymbol{I}_d)[f(\theta + \nu u) - f(\theta) - \frac{\nu^2}{2} u^\top \boldsymbol{H}_f u + f(\theta - \nu u) - f(\theta) - \frac{\nu^2}{2} u^\top \boldsymbol{H}_f u] \\
&\quad + (uu^\top - \boldsymbol{I}_d)\frac{1}{2} u^\top \boldsymbol{H}_f u - \boldsymbol{H}_f \\
&\overset{def}{=} \boldsymbol{A} + \boldsymbol{B},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{E}[\|\boldsymbol{A}\|_F^2] &\leq \left(\tfrac{M_2 \nu}{6}\right)^2 \mathbf{E}[\|uu^\top - \boldsymbol{I}_d\|_F^2 \|u\|^6] \\
&\leq \left(\tfrac{M_2 \nu}{6}\right)^2 (d+4)^5, \\
\mathbf{E}[\|\boldsymbol{B}\|_F^2] &\leq \mathbf{E}[\|(uu^\top - \boldsymbol{I}_d)\frac{1}{2} u^\top \boldsymbol{H}_f u\|_F^2] \\
&\leq \left(\tfrac{1}{2}\|\boldsymbol{H}_f\|_2\right)^2 \mathbf{E}[\|uu^\top - \boldsymbol{I}_d\|_F^2 \|u\|^4] \\
&\leq \left(\tfrac{1}{2}\|\boldsymbol{H}_f\|_2\right)^2 (d+3)^4.
\end{aligned}$$

Therefore,

$$\|\hat{\boldsymbol{H}}_{f_\nu} - \boldsymbol{H}_f\|_{L_2,F} \leq \|\boldsymbol{A}\|_{L_2,F} + \|\boldsymbol{B}\|_{L_2,F} \leq \frac{1}{6} M_2 \nu (d+4)^{5/2} + \frac{1}{2}\|\boldsymbol{H}_f\|_2 (d+3)^2.$$

For the sample mean estimator $\hat{\boldsymbol{H}}_{f_\nu,n}$, we have the following observation.

$$\begin{aligned}
\|\hat{\boldsymbol{H}}_{f_\nu,n} - \boldsymbol{H}_f\|_{L_2,2} &\leq \|\hat{\boldsymbol{H}}_{f_\nu,n} - \boldsymbol{H}_{f_\nu}\|_{L_2,F} + \|\boldsymbol{H}_{f_\nu} - \boldsymbol{H}_f\|_{L_2,2} \\
&= \frac{1}{\sqrt{n}}\|\hat{\boldsymbol{H}}_{f_\nu} - \boldsymbol{H}_{f_\nu}\|_{L_2,F} + \|\boldsymbol{H}_{f_\nu} - \boldsymbol{H}_f\|_{L_2,2} \\
&\leq \frac{1}{\sqrt{n}}\|\hat{\boldsymbol{H}}_{f_\nu} - \boldsymbol{H}_f\|_{L_2,F} + \|\boldsymbol{H}_{f_\nu} - \boldsymbol{H}_f\|_{L_2,2}.
\end{aligned}$$

Applying previous lemmas leads to

$$\|\hat{\boldsymbol{H}}_{f_\nu,n} - \boldsymbol{H}_f\|_{L_2,2} \leq \frac{1}{6\sqrt{n}} M_2 \nu (d+4)^{5/2} + \frac{1}{2\sqrt{n}}\|\boldsymbol{H}_f\|_2 (d+3)^2 + M_2 \nu d^{1/2}.$$

■

*Proof.* [Proof of Theorem 5.6] We first derive the bounds $\mathcal{C}_4(d)$, $\mathcal{C}_5(d)$, $\mathcal{C}_6(d)$ as defined in Assumption 5.2. First, consider the Hessian estimator $\boldsymbol{S}_t = \hat{\boldsymbol{H}}_{f_\nu}$ from a single Gaussian sample. Write $\hat{\boldsymbol{H}}_{f_\nu} = (uu^\top - \boldsymbol{I}_d)\phi(\nu, u)$ where

$$\begin{aligned}\phi(\nu,u) &\overset{def}{=} \frac{1}{2\nu^2}[f(\theta+\nu u) - f(\theta) + f(\theta-\nu u) - f(\theta)] \\ &= \frac{1}{2}u^\top \nabla^2 f(\theta+\iota u)u \in [0, \frac{1}{2}M\|u\|^2].\end{aligned}$$

Indeed, if we set $F(\nu) = f(\theta+\nu u) - f(\theta) + f(\theta-\nu u) - f(\theta)$, then $F'(\nu) = \nabla f(\theta+\nu u)^\top u - \nabla f(\theta-\nu u)^\top u$, and thus $F''(\nu) = u^\top\nabla^2 f(\theta+\nu u)u + u^\top \nabla^2 f(\theta-\nu u)u$. By Taylor's expansion, $F(\nu) = F(0)+F'(0)\nu+\frac{1}{2}F''(\iota')\nu^2 = \frac{1}{2}\nu^2[u^\top\nabla^2 f(\theta+\iota' u)u+u^\top\nabla^2 f(\theta-\iota' u)u]$. If $\nabla^2 f$ is continuous, then $u^\top\nabla^2 f(\theta+\iota' u)u$ is continuous w.r.t. $\iota' \in \mathbb{R}$. By intermediate value theorem, there exists $\iota \in (-\iota', \iota')$ such that $\frac{1}{2}[u^\top\nabla^2 f(\theta+\iota' u)u + u^\top\nabla^2 f(\theta-\iota' u)u] = u^\top\nabla^2 f(\theta+\iota u)u$. Therefore $\phi(\nu,u) = \frac{1}{2\nu^2}F(\nu) = \frac{1}{2\nu^2}\nu^2 u^\top\nabla^2 f(\theta+\iota u)u = \frac{1}{2}u^\top\nabla^2 f(\theta+\iota u)u$. The eigenvalues of $\hat{\boldsymbol{H}}_{f_\nu}$ are $\lambda_i = -\phi(\nu,u)$, $\lambda_d = (\|u\|^2-1)\phi(\nu,u)$, $i = 1,\ldots,d-1$. For $h < \frac{1}{2M}$, $n \in \mathbb{N}$, denote $P_n \overset{def}{=} (1-\frac{2hM}{n})^{-1/2}$. Then we have the following calculation.

$$\begin{aligned}\|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}\|_{L_2}^2 &= \mathbf{E}[\|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}\|_2^2] \\ &\le \mathbf{E}[e^{TM\|u\|^2}] = (1-2TM)^{-d/2} \le P_1^d, \\ \|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}\hat{\boldsymbol{H}}_{f_\nu}^2\|_{L_2}^2 &= \mathbf{E}[\|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}\hat{\boldsymbol{H}}_{f_\nu}^2\|_2^2] \\ &\le \mathbf{E}[e^{TM\|u\|^2}\|\hat{\boldsymbol{H}}_{f_\nu}\|_2^4] \\ &\le \mathbf{E}[e^{TM\|u\|^2}\|u\|^8\left(\tfrac{1}{2}M\|u\|^2\right)^4\mathbf{1}_{\|u\|^2\ge 2}] + \mathbf{E}[e^{TM\|u\|^2}\left(\tfrac{1}{2}M\|u\|^2\right)^4\mathbf{1}_{\|u\|^2\le 2}] \\ &\le (\tfrac{M}{2})^4\left(\mathbf{E}[e^{TM\|u\|^2}\|u\|^{16}] + \mathbf{E}[e^{TM\|u\|^2}\|u\|^8(1-\|u\|^8)\mathbf{1}_{\|u\|^2\le 2}]\right) \\ &\le (\tfrac{M}{2})^4 P_1^{d+16}(d+7)^8,\end{aligned}$$

and moreover,

$$\begin{aligned}\|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}(\hat{\boldsymbol{H}}_{f_\nu} - \boldsymbol{H}_f)\|_{L_2} &\le \|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}(\hat{\boldsymbol{H}}_{f_\nu} - \hat{\boldsymbol{H}}_q)\|_{L_2} + \|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}\hat{\boldsymbol{H}}_q\|_{L_2} + \|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}\|_{L_2}\|\boldsymbol{H}_f\|_2 \\ &\le \frac{1}{6}M_2\nu P_1^{d/2+5}(d+4)^{5/2} + \frac{1}{2}\|\boldsymbol{H}_f\|_2 P_1^{d/2+4}(d+3)^2,\end{aligned}$$

where $\hat{\boldsymbol{H}}_q = (uu^\top - \boldsymbol{I}_d)\frac{1}{2}u^\top \boldsymbol{H}_f u$. The second inequality results from estimate as follows.

$$
\begin{aligned}
\mathbf{E}[\|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}(\hat{\boldsymbol{H}}_{f_\nu} - \hat{\boldsymbol{H}}_q)\|_2^2] &\leq \left(\tfrac{1}{6}M_2\nu\right)^2 \mathbf{E}[e^{TM\|u\|^2}\|uu^\top - \boldsymbol{I}_d\|_2^2\|u\|^6] \\
&\leq \left(\tfrac{1}{6}M_2\nu\right)^2 \left(\mathbf{E}[e^{TM\|u\|^2}\|u\|^{10}\mathbf{1}_{\|u\|^2\geq 2}] + \mathbf{E}[e^{TM\|u\|^2}\|u\|^6\mathbf{1}_{\|u\|^2\leq 2}]\right) \\
&= \left(\tfrac{1}{6}M_2\nu\right)^2 \left(\mathbf{E}[e^{TM\|u\|^2}\|u\|^{10}] + \mathbf{E}[e^{TM\|u\|^2}\|u\|^6(1-\|u\|^4)\mathbf{1}_{\|u\|^2\leq 2}]\right) \\
&\leq \left(\tfrac{1}{6}M_2\nu\right)^2 P_1^{d+10}(d+4)^5, \\
\mathbf{E}[\|e^{-T\hat{\boldsymbol{H}}_{f_\nu}}\hat{\boldsymbol{H}}_q\|_2^2] &\leq \left(\tfrac{1}{2}\|\boldsymbol{H}_f\|_2\right)^2 \mathbf{E}[e^{TM\|u\|^2}\|uu^\top - \boldsymbol{I}_d\|_2^2\|u\|^4] \\
&\leq \left(\tfrac{1}{2}\|\boldsymbol{H}_f\|_2\right)^2 \left(\mathbf{E}[e^{TM\|u\|^2}\|u\|^8\mathbf{1}_{\|u\|^2\geq 2}] + \mathbf{E}[e^{TM\|u\|^2}\|u\|^4\mathbf{1}_{\|u\|^2\leq 2}]\right) \\
&= \left(\tfrac{1}{2}\|\boldsymbol{H}_f\|_2\right)^2 \left(\mathbf{E}[e^{TM\|u\|^2}\|u\|^8] + \mathbf{E}[e^{TM\|u\|^2}\|u\|^4(1-\|u\|^4)\mathbf{1}_{\|u\|^2\leq 2}]\right) \\
&\leq \left(\tfrac{1}{2}\|\boldsymbol{H}_f\|_2\right)^2 P_1^{d+8}\left((d+3)^2-2\right)^2.
\end{aligned}
$$

Next, consider the sample mean estimator $\boldsymbol{S}_t = \hat{\boldsymbol{H}}_{f_\nu,n}$. Utilizing the results for a single sample version, $\mathcal{C}_4(d)$ and $\mathcal{C}_6(d)$ can be readily obtained.

$$
\begin{aligned}
\|e^{-T\hat{\boldsymbol{H}}_{f_\nu,n}}\|_{L_2} &= \mathbf{E}[\|e^{-T\hat{\boldsymbol{H}}_{f_\nu,n}}\|_2^2]^{1/2} \\
&\leq \mathbf{E}[\|e^{-\frac{T}{n}\hat{\boldsymbol{H}}_{f_\nu}}\|_2^2]^{n/2} \\
&\leq \left(1-\tfrac{2TM}{n}\right)^{-nd/4} \leq P_n^{nd/2} = \mathcal{C}_4(d), \\
\|e^{-T\hat{\boldsymbol{H}}_{f_\nu,n}}\hat{\boldsymbol{H}}_{f_\nu,n}^2\|_{L_2} &= \mathbf{E}[\|e^{-T\hat{\boldsymbol{H}}_{f_\nu,n}}\hat{\boldsymbol{H}}_{f_\nu,n}^2\|_2^2]^{1/2} \\
&= \|e^{-\frac{T}{2}\hat{\boldsymbol{H}}_{f_\nu,n}}\hat{\boldsymbol{H}}_{f_\nu,n}\|_{L_4}^2 \\
&\leq \|e^{-\frac{T}{2}\hat{\boldsymbol{H}}_{f_\nu,n}}\hat{\boldsymbol{H}}_{f_\nu}\|_{L_4}^2 \\
&\leq \mathbf{E}[\|e^{-\frac{T}{n}\hat{\boldsymbol{H}}_{f_\nu}}\|_2^2]^{(n-1)/2}\mathbf{E}[\|e^{-\frac{T}{n}\hat{\boldsymbol{H}}_{f_\nu}}\hat{\boldsymbol{H}}_{f_\nu}^2\|_2^2]^{1/2} \\
&\leq (\tfrac{M}{2})^2 P_n^{nd/2+8}(d+7)^4 = \mathcal{C}_6(d).
\end{aligned}
$$

Finally, we focus on estimating the bound $\mathcal{C}_5(d)$.

$$
\begin{aligned}
\|e^{-T\hat{\boldsymbol{H}}_{f_\nu,n}}(\hat{\boldsymbol{H}}_{f_\nu,n} - \boldsymbol{H}_f)\|_{L_2} &\leq \|e^{-T\hat{\boldsymbol{H}}_{f_\nu,n}}(\hat{\boldsymbol{H}}_{f_\nu,n} - \bar{\boldsymbol{H}})\|_{L_2} + \|e^{-T\hat{\boldsymbol{H}}_{f_\nu,n}}\|_{L_2}\|\bar{\boldsymbol{H}} - \boldsymbol{H}_f\|_2 \qquad (28) \\
&\leq \frac{1}{\sqrt{n}}P_n^{nd/2+5}\left(\frac{1}{6}M_2\nu(d+4)^{5/2} + \frac{1}{2}\|\boldsymbol{H}_f\|_2(d+3)^2\right) \\
&\quad + P_n^{nd/2}\left(P_n^5 M_2\nu d^{1/2} + P_n^4 M^2 n^{-1}hd\right) \\
&\leq P_n^{nd/2+5}\left(\frac{1}{6\sqrt{n}}M_2\nu(d+4)^{5/2} + \frac{1}{2\sqrt{n}}M(d+3)^2 + M_2\nu d^{1/2} + \frac{1}{n}M^2hd\right).
\end{aligned}
$$

Here $\bar{\boldsymbol{H}}$ is the expectation of $\hat{\boldsymbol{H}}_{f_\nu}$ after change of variable, replacing $u$ by $P_n u$, such that

$\mathbf{E}[\|\hat{\boldsymbol{H}}_{f_\nu,n}(P_n u) - \bar{\boldsymbol{H}}\|_F^2] = \frac{1}{n}\mathbf{E}[\|\hat{\boldsymbol{H}}_{f_\nu}(P_n u) - \bar{\boldsymbol{H}}\|_F^2]$ holds. Then

$$\begin{aligned}\mathbf{E}[\|e^{-T\hat{\boldsymbol{H}}_{f_\nu,n}}(\hat{\boldsymbol{H}}_{f_\nu,n} - \bar{\boldsymbol{H}})\|_2^2] &\le \mathbf{E}[e^{\frac{T}{n}\sum_i M\|u_i\|^2}\|\hat{\boldsymbol{H}}_{f_\nu,n}(u) - \bar{\boldsymbol{H}}\|_2^2] \\ &= P_n^{nd}\mathbf{E}[\|\hat{\boldsymbol{H}}_{f_\nu,n}(P_n u) - \bar{\boldsymbol{H}}\|_2^2] \\ &\le \frac{1}{n}P_n^{nd}\mathbf{E}[\|\hat{\boldsymbol{H}}_{f_\nu}(P_n u) - \bar{\boldsymbol{H}}\|_F^2] \\ &\le \frac{1}{n}P_n^{nd}\mathbf{E}[\|\hat{\boldsymbol{H}}_{f_\nu}(P_n u)\|_F^2].\end{aligned}$$

Note that $\hat{\boldsymbol{H}}_{f_\nu}(P_n u)$ is not a Hessian estimator of the form (14). However, the bound for $\|\hat{\boldsymbol{H}}_{f_\nu}(P_n u)\|_{L_2,F}$ can still be given in a similar fashion.

$$\begin{aligned}\mathbf{E}[\|\hat{\boldsymbol{H}}_{f_\nu}(P_n u)\|_F^2] &= \mathbf{E}[\|(P_n^2 uu^\top - \boldsymbol{I}_d)\frac{f(x+\nu P_n u) - f(x) + f(x - \nu P_n u) - f(x)}{2\nu^2}\|_F^2] \\ &\le \left(\mathbf{E}[\|P_n^2 uu^\top - \boldsymbol{I}_d\|_F^2(\tfrac{1}{6}M_2\nu P_n^3\|u\|^3)^2]^{1/2} + \mathbf{E}[\|(P_n^2 uu^\top - \boldsymbol{I}_d)\tfrac{1}{2}P_n^2 u^\top\nabla^2 f(x)u\|_F^2]^{1/2}\right)^2 \\ &\le \left(\tfrac{1}{6}P_n^3 M_2\nu\mathbf{E}[\|P_n^2 uu^\top - \boldsymbol{I}_d\|_F^2\|u\|^6]^{1/2} + \tfrac{1}{2}P_n^2\|\boldsymbol{H}_f\|_2\mathbf{E}[\|P_n^2 uu^\top - \boldsymbol{I}_d\|_F^2\|u\|^4]^{1/2}\right)^2 \\ &\le \left(\tfrac{1}{6}P_n^5 M_2\nu(d+4)^{5/2} + \tfrac{1}{2}P_n^4\|\boldsymbol{H}_f\|_2[(d+3)^2 - 5]\right)^2\end{aligned}$$

To write out $\bar{\boldsymbol{H}}$ explicitly,

$$\begin{aligned}\bar{\boldsymbol{H}} &= \mathbf{E}[(P_n^2 uu^\top - \boldsymbol{I}_d)\phi(\nu, P_n u)] \\ &= P_n^4\mathbf{E}[(uu^\top - \boldsymbol{I}_d)\phi(P_n\nu, u)] + (P_n^2 - 1)P_n^2\mathbf{E}[\phi(P_n\nu, u)]\boldsymbol{I}_d \\ &= P_n^4\boldsymbol{H}_{f_{P_n\nu}} + (P_n^2 - 1)P_n^2\mathbf{E}[\phi(P_n\nu, u)]\boldsymbol{I}_d.\end{aligned}$$

Then the second term in (28) can be estimated by

$$\begin{aligned}\|\bar{\boldsymbol{H}} - \boldsymbol{H}_f\|_2 &= \|P_n^4\left(\boldsymbol{H}_{f_{P_n\nu}} - \boldsymbol{H}_f\right) + (P_n^2 - 1)P_n^2\mathbf{E}[\phi(P_n\nu, u)]\boldsymbol{I}_d + (P_n^4 - 1)\boldsymbol{H}_f\|_2 \\ &\le P_n^4\|\boldsymbol{H}_{f_{P_n\nu}} - \boldsymbol{H}_f\|_2 + P_n^4(1 - P_n^{-2})\mathbf{E}[\tfrac{1}{2}M\|u\|^2] + P_n^4(1 - P_n^{-4})\|\boldsymbol{H}_f\|_2 \\ &\le P_n^5 M_2\nu d^{1/2} + P_n^4 M^2 n^{-1}hd + 4\|\boldsymbol{H}_f\|_2 P_n^4 M n^{-1}h.\end{aligned}$$

Assuming that $\frac{h}{\sqrt{n}} \le \frac{1}{2M}$, combining the above results gives rise to $\mathcal{C}_5(d)$ in (28).

Since $\|g_t - \mathbf{E}[g_t|D_0]\|_{L_2}$ is not bounded by a global constant, Theorem 5.3 and 5.4 does not apply directly. Therefore, we need to modify the proofs specifically. For the ZOOLMC,

$$
\begin{aligned}
\|\Delta_h\|_{L_2} &= \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h\|_{L_2} \\
&\le (\|\Delta_0 + A_h\|_{L_2}^2 + \|I_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|J_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|Q_h\|_{L_2} \\
&\le \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{h^2}{n_g}\left(\mathcal{C}_1'(d) + M(d+1)^{1/2}\|\Delta_0\|_{L_2}\right)^2 \mathcal{C}_4(d)^2 \\
&\quad + \frac{2}{3}MM_2h^3d\mathcal{C}_4(d)^2\|\Delta_0\|_{L_2}\}^{1/2} + h\mathcal{C}_2(d)\mathcal{C}_4(d) + 0.877M_2h^2(d^2+2d)^{1/2}\mathcal{C}_4(d) \\
&\quad + \mu\|\Delta_0\|_{L_2} + \frac{1}{12\mu}M^2M_2h^4(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_5(d) \\
&\le \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{2h^2}{n_g}\mathcal{C}_1'(d)^2\mathcal{C}_4(d)^2\}^{1/2} + \frac{M^2h^2(d+1)\mathcal{C}_4(d)^2\|\Delta_0\|_{L_2}}{n_g(1 - mh + 0.5\tilde{M}^2h^2)} \\
&\quad + \frac{MM_2h^3d\mathcal{C}_4(d)^2}{3(1 - mh + 0.5\tilde{M}^2h^2)} + h\mathcal{C}_2(d)\mathcal{C}_4(d) + 0.877M_2h^2(d^2+2d)^{1/2}\mathcal{C}_4(d) \\
&\quad + \mu\|\Delta_0\|_{L_2} + \frac{1}{12\mu}M^2M_2h^4(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_5(d) \\
&\le \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{2h^2}{n_g}\mathcal{C}_1'(d)^2\mathcal{C}_4(d)^2\}^{1/2} + \frac{2}{n}M^2h^2(d+1)\mathcal{C}_4(d)^2\|\Delta_0\|_{L_2} \\
&\quad + \frac{2}{3}MM_2h^3d\mathcal{C}_4(d)^2 + h\mathcal{C}_2(d)\mathcal{C}_4(d) + 0.877M_2h^2(d^2+2d)^{1/2}\mathcal{C}_4(d) \\
&\quad + \frac{1}{8}mh\|\Delta_0\|_{L_2} + \frac{2M^2M_2}{3m}h^3(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_5(d) \\
&\le \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{2h^2}{n_g}\mathcal{C}_1'(d)^2\mathcal{C}_4(d)^2\}^{1/2} + \frac{1}{4}mh\|\Delta_0\|_{L_2} \\
&\quad + \frac{4M^2M_2}{3m}h^3(d+1)\mathcal{C}_4(d)^2 + \left(0.877M_2h^2(d+1) + h\mathcal{C}_2(d)\right)\mathcal{C}_4(d) + 1.32h^{3/2}d^{1/2}\mathcal{C}_5(d),
\end{aligned}
$$

where $\mathcal{C}_1'(d) = \frac{1}{2}M\nu_g^2(d+2)^{3/2} + \sqrt{M}(d+1)$, $\mathcal{C}_2(d) = M\nu_g\sqrt{d}$ as in Theorem 3.1, and $\frac{h}{n_g} \le \frac{m}{16P_{n_H}^{n_H d}M^2(d+1)}$. The Wasserstein bound is obtained as

$$
\begin{aligned}
W_2(\varpi_t, \pi) &\le (1 - 0.25mh)^tW_2(\varpi_0, \pi) + \frac{16M^2M_2}{3m^2}h^2(d+1)\mathcal{C}_4(d)^2 \\
&\quad + \left(3.51\frac{M_2}{m}h(d+1) + \frac{4\sqrt{2}}{\sqrt{5m}}\frac{1}{\sqrt{n_g}}h^{1/2}\mathcal{C}_1'(d) + \frac{4}{m}\mathcal{C}_2(d)\right)\mathcal{C}_4(d) + \frac{5.27}{m}(hd)^{1/2}\mathcal{C}_5(d) \\
&\le (1 - 0.25mh)^tW_2(\varpi_0, \pi) \\
&\quad + P_{n_H}^{n_H d/2}\left(\frac{4\sqrt{2}}{\sqrt{5m}}\frac{1}{\sqrt{n_g}}h^{1/2}\mathcal{C}_1'(d) + \frac{4}{m}\mathcal{C}_2(d) + P_{n_H}^5\frac{5.27}{m}(hd)^{1/2}\mathcal{C}_3(d)\right) \\
&\quad + P_{n_H}^{n_H d/2}\left((\frac{16}{3}P_{n_H}^{n_H d/2} + 3.51)\frac{M_2}{m}h(d+1) + P_{n_H}^5\frac{5.27M^2}{m}\frac{1}{n_H}(hd)^{3/2}\right).
\end{aligned}
$$

Further, we have

$$
\begin{aligned}
\|E_T\|_{L_2} &= \left\| \frac{1}{2} \int_0^T e^{-s\boldsymbol{S}_t} s^2 ds \boldsymbol{S}_t^2 g_t \right\|_{L_2} \\
&\le \frac{1}{6} \hat{M}^2 T^3 \|g_t\|_{L_2} \\
&\le \frac{1}{6} \hat{M}^2 T^3 \left( \frac{1}{2} M\nu (d+2)^{3/2} + (d+2)^{1/2} \|\nabla f(x_t)\| \right) \\
&\le \frac{1}{6} \hat{M}^2 T^3 \left( \frac{1}{2} M\nu (d+2)^{3/2} + \sqrt{M}(d+1) + M(d+2)^{1/2} \|\Delta_0\|_{L_2} \right) \\
&= \frac{1}{6} \hat{M}^2 T^3 \left( \mathcal{C}_1'(d) + M(d+2)^{1/2} \|\Delta_0\|_{L_2} \right).
\end{aligned}
$$

Thus, for the Approximate ZOOLMC,

$$
\begin{aligned}
\|\Delta_h\|_{L_2} &= \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h + E_h - F_h\|_{L_2} \\
&\le (\|\Delta_0 + A_h\|_{L_2}^2 + \|I_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|J_h\|_{L_2} + \|B_h\|_{L_2} \\
&\quad + \|C_h\|_{L_2} + \|Q_h\|_{L_2} + \|E_h\|_{L_2} + \|F_h\|_{L_2} \\
&\le \{(1 - mh + \frac{1}{2}\tilde{M}^2 h^2)^2 \|\Delta_0\|_{L_2}^2 + \frac{h^2}{n_g} \left( \mathcal{C}_1'(d) + M(d+1)^{1/2} \|\Delta_0\|_{L_2} \right)^2 \mathcal{C}_4(d)^2 \\
&\quad + \frac{2}{3} M M_2 h^3 d \mathcal{C}_4(d)^2 \|\Delta_0\|_{L_2}\}^{1/2} + h\mathcal{C}_2(d)\mathcal{C}_4(d) + 0.877 M_2 h^2 (d^2 + 2d)^{1/2} \mathcal{C}_4(d) \\
&\quad + \mu \|\Delta_0\|_{L_2} + \frac{1}{12\mu} M^2 M_2 h^4 (d+1) \mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M} h^2 d^{1/2} \mathcal{C}_5(d) + \frac{\sqrt{6}}{3} h^{3/2} d^{1/2} \mathcal{C}_5(d) \\
&\quad + \frac{1}{6} \hat{M}^2 h^3 \left( \mathcal{C}_1'(d) + M(d+2)^{1/2} \|\Delta_0\|_{L_2} \right) + \frac{1}{\sqrt{10}} h^{5/2} d^{1/2} \mathcal{C}_6(d).
\end{aligned}
$$

Continuing the calculation, we have

$$\begin{aligned}
\|\Delta_h\|_{L_2} \leq& \{(1-mh+\frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{2h^2}{n_g}\mathcal{C}_1'(d)^2\mathcal{C}_4(d)^2\}^{1/2} + \frac{M^2h^2(d+1)\mathcal{C}_4(d)^2\|\Delta_0\|}{n_g(1-mh+0.5\tilde{M}^2h^2)} \\
&+ \frac{MM_2h^3d\mathcal{C}_4(d)^2}{3(1-mh+0.5\tilde{M}^2h^2)} + h\mathcal{C}_2(d)\mathcal{C}_4(d) + 0.877M_2h^2(d^2+2d)^{1/2}\mathcal{C}_4(d) \\
&+ \frac{1}{8}mh\|\Delta_0\|_{L_2} + \frac{2M^2M_2}{3m}h^3(d+1)\mathcal{C}_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}\mathcal{C}_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}\mathcal{C}_5(d) \\
&+ \frac{1}{6}\hat{M}^2h^3\left(\mathcal{C}_1'(d) + M(d+2)^{1/2}\|\Delta_0\|_{L_2}\right) + \frac{1}{\sqrt{10}}h^{5/2}d^{1/2}\mathcal{C}_6(d) \\
\leq& \{(1-mh+\frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{2h^2}{n_g}\mathcal{C}_1'(d)^2\mathcal{C}_4(d)^2\}^{1/2} + \frac{1}{4}mh\|\Delta_0\|_{L_2} \\
&+ \frac{1}{6}M\hat{M}^2h^3(d+2)^{1/2}\|\Delta_0\|_{L_2} + 1.3\frac{M^2M_2}{m}h^3(d+1)\mathcal{C}_4(d)^2 \\
&+ \left(0.877M_2h^2(d+1) + h\mathcal{C}_2(d)\right)\mathcal{C}_4(d) \\
&+ 1.25h^{3/2}d^{1/2}\mathcal{C}_5(d) + \left(\frac{1}{\sqrt{10}}h^{5/2}d^{1/2} + \frac{1}{6}h^3\mathcal{C}_1'(d)\right)\mathcal{C}_6(d).
\end{aligned}$$

Here we assume that $\frac{h}{n_g} \leq \frac{17m}{256P_{n_H}^{n_H d}M^2(d+1)}$, and $h \leq \frac{3m}{4M\hat{M}(d+1)^{1/4}}$. Denote $A = mh - \frac{1}{2}\tilde{M}^2h^2$, $D = \frac{1}{4}mh - \frac{1}{6}M\hat{M}^2h^3(d+2)^{1/2}$. Then $A - D \geq \frac{1}{4}mh$, $A + D \leq \frac{3}{4}\cdot\frac{31}{32}$. Therefore,

$$\begin{aligned}
W_2(\varpi_t, \pi) \leq& (1-0.25mh)^tW_2(\varpi_0,\pi) + 5.18\frac{M^2M_2}{m^2}h^2(d+1)\mathcal{C}_4(d)^2 \\
&+ \left(3.51\frac{M_2}{m}h(d+1) + \frac{2.51}{\sqrt{m}}\frac{1}{\sqrt{n_g}}h^{1/2}\mathcal{C}_1'(d) + \frac{4}{m}\mathcal{C}_2(d)\right)\mathcal{C}_4(d) \\
&+ \left(\frac{4}{\sqrt{10}m}h^{3/2}d^{1/2} + \frac{2}{3m}h^2\mathcal{C}_1'(d)\right)\mathcal{C}_6(d) + \frac{5}{m}(hd)^{1/2}\mathcal{C}_5(d).
\end{aligned}$$

Plugging in $\mathcal{C}_i(d)$'s gives the results as desired, i.e.,

$$\begin{aligned}
W_2(\varpi_t, \pi) \leq& (1-0.25mh)^tW_2(\varpi_0,\pi) \\
&+ P_{n_H}^{n_H d/2}\left((\frac{2.51}{\sqrt{m}}\frac{1}{\sqrt{n_g}}h^{1/2} + P_{n_H}^8\frac{M^2}{6m}h^2(d+7)^4)\mathcal{C}_1'(d) + \frac{4}{m}\mathcal{C}_2(d) + P_{n_H}^5\frac{5}{m}(hd)^{1/2}\mathcal{C}_3(d)\right) \\
&+ P_{n_H}^{n_H d/2}\left((3.89P_{n_H}^{n_H d/2} + 3.51)\frac{M_2}{m}h(d+1) + P_{n_H}^8\frac{M^2}{3m}h^{3/2}(d+7)^{9/2}\right).
\end{aligned}$$

∎

# H Proof for Section 6.1

*Proof.* Let $(V_{t,T}, L_{t,T})$, $T \in [0,h]$ be a stationary kinetic Langevin process for each $t \in \mathbb{N}$, i.e.,

$$\begin{aligned}
dV_{t,T} &= -(\gamma V_{t,T} + \nabla f(L_{t,T}))dT + \sqrt{2\gamma}dW_{t,T}, \\
dL_{t,T} &= V_{t,T}dT,
\end{aligned}$$

starting from $V_{0,0} \sim N(0, \boldsymbol{I}_d)$, $L_{0,0} \sim \pi$, and satisfying $V_{t,h} = V_{t+1,0}, L_{t,h} = L_{t+1,h}$. Define $(\tilde{V}_{t,T}, \tilde{L}_{t,T})$ by the following discretized version of kinetic Langevin diffusion,

$$
\begin{aligned}
d\tilde{V}_{t,T} &= -(\gamma \tilde{V}_{t,T} + g(\tilde{L}_{t,0}))dT + \sqrt{2\gamma} dW_{t,T}, \\
d\tilde{L}_{t,T} &= \tilde{V}_{t,T} dT,
\end{aligned}
$$

or equivalently,

$$
\begin{aligned}
\tilde{V}_{t,T} &= e^{-\gamma T} \tilde{V}_{t,0} - \int_0^T e^{-\gamma(T-s)} ds \cdot g(\tilde{L}_{t,0}) + \sqrt{2\gamma} \int_0^T e^{-\gamma(T-s)} dW_{t,T}, \\
\tilde{L}_{t,T} &= \tilde{L}_{t,0} + \int_0^T \tilde{V}_{t,s} ds.
\end{aligned}
$$

Assume that $(\tilde{V}_{0,0}, \tilde{L}_{0,0})$ is chosen such that $\tilde{V}_{0,0} = V_{0,0}$ and $W_2(\varpi_0, \pi) = \|\tilde{L}_{0,0} - L_{0,0}\|_{L_2}$. By definition of Wasserstein distance, we have $W_2(\varpi_t, \pi) \leq \|\tilde{L}_{t,0} - L_{t,0}\|_{L_2}$.
Now we denote $e_t = \left\| \boldsymbol{P}^{-1} \begin{bmatrix} \tilde{V}_{t,0} - V_{t,0} \\ \tilde{L}_{t,0} - L_{t,0} \end{bmatrix} \right\|_{L_2}$, where $\boldsymbol{P}^{-1} = \begin{bmatrix} \boldsymbol{I}_d & \gamma \boldsymbol{I}_d \\ -\boldsymbol{I}_d & \boldsymbol{0} \end{bmatrix}$, $\boldsymbol{P} = \gamma^{-1} \begin{bmatrix} 0 & -\gamma \boldsymbol{I}_d \\ \boldsymbol{I}_d & \boldsymbol{I}_d \end{bmatrix}$ corresponds to the contraction to the kinetic Langevin process. See [15]. Note that $\|\tilde{L}_{t,0} - L_{t,0}\|_{L_2} \leq \sqrt{2}\gamma^{-1} e_t$ and $\|\tilde{V}_{t,0} - V_{t,0}\|_{L_2} \leq e_t$. Define a different kinetic Langevin process $(\hat{V}_{t,T}, \hat{L}_{t,T})$ with initial condition $\hat{V}_{t,0} = \tilde{V}_{t,0}, \hat{L}_{t,0} = \tilde{L}_{t,0}$. Then

$$
\begin{aligned}
\left\| \boldsymbol{P}^{-1} \begin{bmatrix} \hat{V}_{t,T} - V_{t,T} \\ \hat{L}_{t,T} - L_{t,T} \end{bmatrix} \right\|_{L_2} &\leq e^{-mT/\gamma} \left\| \boldsymbol{P}^{-1} \begin{bmatrix} \hat{V}_{t,0} - V_{t,0} \\ \hat{L}_{t,0} - L_{t,0} \end{bmatrix} \right\|_{L_2}, \\
&= e^{-mT/\gamma} e_t.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\|\tilde{V}_{t,T} - \hat{V}_{t,T}\|_{L_2} &= \left\| \int_0^T e^{-\gamma(T-s)} (\nabla f(\hat{L}_{t,s}) - g(\tilde{L}_{t,0})) ds \right\|_{L_2} \\
&\leq \left\| \int_0^T e^{-\gamma(T-s)} (\nabla f(\hat{L}_{t,s}) - \nabla f(\hat{L}_{t,0})) ds \right\|_{L_2} + \left\| \int_0^T e^{-\gamma(T-s)} (\nabla f(\tilde{L}_{t,0}) - g(\tilde{L}_{t,0})) ds \right\|_{L_2} \\
&\leq M \int_0^T \|\hat{L}_{t,s} - \hat{L}_{t,0}\|_{L_2} ds + \int_0^T \|\nabla f(\tilde{L}_{t,0}) - g(\tilde{L}_{t,0})\|_{L_2} ds \\
&\leq M \int_0^T \int_0^s \|\hat{V}_{t,u}\|_{L_2} du ds + \int_0^T \|\nabla f(\tilde{L}_{t,0}) - g(\tilde{L}_{t,0})\|_{L_2} ds \\
&\leq \frac{1}{2} M T^2 \max_{u \in [0,h]} \|\hat{V}_{t,u}\|_{L_2} + T(\mathcal{C}_1(d) + \mathcal{C}_2(d)),
\end{aligned}
$$

$$
\begin{aligned}
\|\tilde{L}_{t,T} - \hat{L}_{t,T}\|_{L_2} &= \left\| \int_0^T \tilde{V}_{t,s} - \hat{V}_{t,s} ds \right\|_{L_2} \\
&\leq \frac{1}{6} M T^3 \max_{u \in [0,h]} \|\hat{V}_{t,u}\|_{L_2} + \frac{1}{2} T^2 (\mathcal{C}_1(d) + \mathcal{C}_2(d)).
\end{aligned}
$$

Thus we have

$$\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\tilde{V}_{t,T}-\hat{V}_{t,T}\\ \tilde{L}_{t,T}-\hat{L}_{t,T}\end{bmatrix}\right\|_{L_2} \le \sqrt{(1+\gamma T/3)^2+1}\cdot\frac{1}{2}MT^2(\sqrt{d}+e_t)+\sqrt{(1+\gamma T/2)^2+1}\cdot T(\mathcal{C}_1(d)+\mathcal{C}_2(d))$$
$$\le \frac{3}{4}MT^2(\sqrt{d}+e_t)+\frac{9\sqrt{2}}{8}T(\mathcal{C}_1(d)+\mathcal{C}_2(d)),$$

where we use the fact that $\|\hat{V}_{t,u}\|_{L_2}\le\|V_{t,u}\|_{L_2}+\|\hat{V}_{t,u}-V_{t,u}\|_{L_2}\le\sqrt{d}+e_t$. Combining the above results, we have

$$\begin{aligned}
e_{t+1} &= \left\|\boldsymbol{P}^{-1}\begin{bmatrix}\tilde{V}_{t,h}-V_{t,h}\\ \tilde{L}_{t,h}-L_{t,h}\end{bmatrix}\right\|_{L_2}\\
&\le \left\|\boldsymbol{P}^{-1}\begin{bmatrix}\tilde{V}_{t,h}-\hat{V}_{t,h}\\ \tilde{L}_{t,h}-\hat{L}_{t,h}\end{bmatrix}\right\|_{L_2}+\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{t,h}-V_{t,h}\\ \hat{L}_{t,h}-L_{t,h}\end{bmatrix}\right\|_{L_2}\\
&\le \frac{3}{4}Mh^2(\sqrt{d}+e_t)+\frac{9\sqrt{2}}{8}h(\mathcal{C}_1(d)+\mathcal{C}_2(d))+e^{-mh/\gamma}e_t\\
&\le (1-\frac{3mh}{4\gamma})e_t+\frac{3}{4}Mh^2\sqrt{d}+\frac{9\sqrt{2}}{8}h(\mathcal{C}_1(d)+\mathcal{C}_2(d))\\
&\le (1-\frac{3mh}{4\gamma})e_t+\frac{3}{4}Mh^2\sqrt{d}+\frac{9\sqrt{2}}{8}h(\mathcal{C}_1'(d)+\mathcal{C}_2(d)+C\|\tilde{L}_{t,0}-L_{t,0}\|_{L_2})\\
&\le (1-\frac{3mh}{4\gamma})e_t+\frac{3}{4}Mh^2\sqrt{d}+\frac{9\sqrt{2}}{8}h(\mathcal{C}_1'(d)+\mathcal{C}_2(d))+\frac{9Ch}{4\gamma}e_t\\
&\le (1-\frac{mh}{2\gamma})e_t+\frac{3}{4}Mh^2\sqrt{d}+\frac{9\sqrt{2}}{8}h(\mathcal{C}_1'(d)+\mathcal{C}_2(d)),
\end{aligned}$$

where we assume that $C=Mn^{-1/2}(d+1)^{1/2}\le m/9$, and note that

$$\begin{aligned}
\|g(\tilde{L}_{t,0})-\nabla f(\tilde{L}_{t,0})\|_{L_2} &\le \mathcal{C}_1(d)+\mathcal{C}_2(d)\\
&= \frac{M\nu}{\sqrt{n}}(d+2)^{3/2}+\frac{1}{\sqrt{n}}(d+1)^{1/2}\|\nabla f(\tilde{L}_{t,0})\|_{L_2}+M\nu\sqrt{d}\\
&\le \frac{M\nu}{\sqrt{n}}(d+2)^{3/2}+\frac{1}{\sqrt{n}}(d+1)^{1/2}(\sqrt{Md}+\|\nabla f(\tilde{L}_{t,0})-\nabla f(L_{t,0})\|_{L_2})+M\nu\sqrt{d}\\
&\le \frac{M\nu}{\sqrt{n}}(d+2)^{3/2}+\frac{\sqrt{M}}{\sqrt{n}}(d+1)+M\nu\sqrt{d}+\frac{M}{\sqrt{n}}(d+1)^{1/2}\|\tilde{L}_{t,0}-L_{t,0}\|_{L_2}\\
&\overset{def}{=} \mathcal{C}_1'(d)+\mathcal{C}_2(d)+C\|\tilde{L}_{t,0}-L_{t,0}\|_{L_2}.
\end{aligned}$$

Applying the inequality recursively, we obtain

$$\begin{aligned}
W_2(\varpi_t,\pi) &\le \sqrt{2}\gamma^{-1}e_t\\
&\le \sqrt{2}\gamma^{-1}(1-\frac{mh}{2\gamma})^t e_0+\frac{3\sqrt{2}M}{2m}h\sqrt{d}+\frac{9}{2m}(\mathcal{C}_1'(d)+\mathcal{C}_2(d))\\
&= \sqrt{2}(1-\frac{mh}{2\gamma})^t W_2(\varpi_0,\pi)+\frac{3\sqrt{2}M}{2m}h\sqrt{d}+\frac{9}{2m}(\mathcal{C}_1'(d)+\mathcal{C}_2(d)).
\end{aligned}$$

This completes the proof of part 1. The proof of part 2 immediately follows from the statement in part 1 and the estimates in Theorem 5.6. ∎

# I Proof for Section 6.2

*Proof.* Denote $\psi_0(t) = e^{-\gamma t}, \psi_{i+1}(t) = \int_0^t \psi_i(s)ds$ and $\phi_{i+1}(t) = \int_0^t e^{-\gamma(t-s)}\psi_i(s)ds$. Consider the settings as in the previous proof, except that $(\tilde{V}_{t,T}, \tilde{L}_{t,T})$ is now defined by an alternative discretization of the kinetic Langevin diffusion. The explicit form is as follow.

$$\begin{aligned}
\tilde{V}_{t,T} &= e^{-\gamma T}\tilde{V}_{t,0} - \psi_1(t)\nabla f(\tilde{L}_{t,0}) - \phi_2(t)\nabla^2 f(\tilde{L}_{t,0})\tilde{V}_{t,0} \\
&\qquad \sqrt{2\gamma}\int_0^T e^{-\gamma(T-s)}dW_{t,s} - \sqrt{2\gamma}\nabla^2 f(\tilde{L}_{t,0})\int_0^T \phi_2(T-s)dW_{t,s}, \\
\tilde{L}_{t,T} &= \tilde{L}_{t,0} + \psi_1(t)\tilde{V}_{t,0} - \psi_2(t)\nabla f(\tilde{L}_{t,0}) - \phi_3(t)\nabla^2 f(\tilde{L}_{t,0})\tilde{V}_{t,0} \\
&\qquad \sqrt{2\gamma}\int_0^T \psi_1(T-s)dW_{t,s} - \sqrt{2\gamma}\nabla^2 f(\tilde{L}_{t,0})\int_0^T \phi_3(T-s)dW_{t,s},
\end{aligned}$$

Moreover, $(\hat{V}_{t,T}, \hat{L}_{t,T})$ is defined by the same discretization, with $\hat{V}_{t,0} = V_{t,0}, \hat{L}_{t,0} = L_{t,0}$. Then for the discretization error we have

$$\begin{aligned}
\|\hat{V}_{t,T} - V_{t,T}\|_{L_2} &= \|A - B + D + E\|_{L_2} \\
&\leq \frac{1}{6}M_2T^3(d+1) + \frac{1}{6}T^3\sqrt{M^3 d} + T(\mathcal{C}_1(d) + \mathcal{C}_2(d)) + \frac{1}{2}T^2\sqrt{d}\,\mathcal{C}_3(d), \\
\|\hat{L}_{t,T} - L_{t,T}\|_{L_2} &\leq \frac{1}{24}M_2T^4(d+1) + \frac{1}{24}T^4\sqrt{M^3 d} + \frac{1}{2}T^2(\mathcal{C}_1(d) + \mathcal{C}_2(d)) + \frac{1}{6}T^3\sqrt{d}\,\mathcal{C}_3(d).
\end{aligned}$$

Here

$$\begin{aligned}
A &= \int_0^T e^{-\gamma(T-s)}[\nabla f(L_{t,s}) - \nabla f(L_{t,0}) - \nabla^2 f(L_{t,0})(L_{t,s} - L_{t,0})]ds, \\
B &= \nabla^2 f(L_{t,0})\int_0^T\int_0^s\int_0^r e^{-\gamma(T-s)}e^{-\gamma(r-w)}\nabla f(L_{t,w})dwdrds, \\
D &= \int_0^T e^{-\gamma(T-s)}[\nabla f(L_0) - g(L_0)]ds, \\
E &= [\boldsymbol{S}(L_{t,0}) - \nabla^2 f(L_{t,0})]\int_0^T\int_0^s e^{-\gamma(T-s)}V_r drds.
\end{aligned}$$

See [15]. Thus we have, for $T \leq \frac{1}{5\gamma}$,

$$\begin{aligned}
\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{t,T} - V_{t,T} \\ \hat{L}_{t,T} - L_{t,T}\end{bmatrix}\right\|_{L_2} &\leq \sqrt{(1+\gamma T/4)^2 + 1}\left(\frac{1}{6}M_2T^3(d+1) + \frac{1}{6}T^3\sqrt{M^3 d}\right) \\
&\quad + \sqrt{(1+\gamma T/2)^2 + 1}\cdot T(\mathcal{C}_1(d) + \mathcal{C}_2(d)) + \sqrt{(1+\gamma T/3)^2 + 1}\cdot\frac{1}{2}T^2\sqrt{d}\,\mathcal{C}_3(d) \\
&\leq \frac{1}{4}M_2T^3(d+1) + \frac{1}{4}\sqrt{M^3}T^3 d^{1/2} + \frac{3}{2}T(\mathcal{C}_1(d) + \mathcal{C}_2(d)) + \frac{5}{6}T^2 d^{1/2}\mathcal{C}_3(d).
\end{aligned}$$

On the other hand,

$$\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{t,T}-\tilde{V}_{t,T}\\ \hat{L}_{t,T}-\tilde{L}_{t,T}\end{bmatrix} = \left(\boldsymbol{I}_{2d}-\psi_1(t)\boldsymbol{P}^{-1}\boldsymbol{R}\boldsymbol{P}-\boldsymbol{P}^{-1}\boldsymbol{E}'\boldsymbol{P}\right)\boldsymbol{P}^{-1}\begin{bmatrix}V_{t,0}-\tilde{V}_{t,0}\\ L_{t,0}-\tilde{L}_{t,0}\end{bmatrix}$$
$$+\boldsymbol{P}^{-1}\begin{bmatrix}\phi_2(t)(\boldsymbol{S}(L_{t,0})-\boldsymbol{S}(\tilde{L}_{t,0}))V_{t,0}\\ \phi_3(t)(\boldsymbol{S}(L_{t,0})-\boldsymbol{S}(\tilde{L}_{t,0}))V_{t,0}\end{bmatrix}$$
$$+\boldsymbol{P}^{-1}\begin{bmatrix}\psi_1(t)(\nabla f(\tilde{L}_{t,0})-g(\tilde{L}_{t,0})-\nabla f(L_{t,0})+g(L_{t,0}))\\ \psi_2(t)(\nabla f(\tilde{L}_{t,0})-g(\tilde{L}_{t,0})-\nabla f(L_{t,0})+g(L_{t,0}))\end{bmatrix},$$

where $\boldsymbol{R}=\begin{bmatrix}\gamma\boldsymbol{I}_d & \boldsymbol{H}_0\\ -\boldsymbol{I}_d & \boldsymbol{0}_{d\times d}\end{bmatrix}$, $\boldsymbol{E}'=\begin{bmatrix}\phi_2(t)\boldsymbol{S}(\tilde{L}_{t,0}) & \boldsymbol{0}_{d\times d}\\ \phi_3(t)\boldsymbol{S}(\tilde{L}_{t,0}) & -\psi_2(t)\boldsymbol{H}_0\end{bmatrix}$. See [15]. Since

$$\|\boldsymbol{I}_{2d}-\psi_1(t)\boldsymbol{P}^{-1}\boldsymbol{R}\boldsymbol{P}-\boldsymbol{P}^{-1}\boldsymbol{E}'\boldsymbol{P}\|_{L_2}\le 1-\frac{mT}{\gamma}+\frac{2+\sqrt{2}}{2}MT^2+\frac{\sqrt{2}}{2}\bar{M}T^2,$$

and

$$\|(\boldsymbol{S}(L_{t,0})-\boldsymbol{S}(\tilde{L}_{t,0}))V_{t,0}\|_{L_2}\le\|(\nabla^2 f(L_{t,0})-\nabla^2 f(\tilde{L}_{t,0}))V_{t,0}\|_{L_2}+2\|(\boldsymbol{S}(L_{t,0})-\nabla^2 f(L_{t,0}))V_{t,0}\|_{L_2}$$
$$\le\frac{\sqrt{2a}M_2}{\gamma}e_t+2(M-m)e^{-(a-p)/8}+2\sqrt{d}\,\mathcal{C}_3(d)$$

for $a\ge 5d$, plugging them into the previous expression, for $T\le m\gamma^{-1}/(3.5M+1.5\bar{M})$, we obtain

$$\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{t,T}-\tilde{V}_{t,T}\\ \hat{L}_{t,T}-\tilde{L}_{t,T}\end{bmatrix}\right\|_{L_2}\le(1-\frac{mT}{2\gamma}+\frac{\sqrt{a}M_2T^2}{\gamma})e_t+\sqrt{2}(M-m)e^{-(a-p)/8}T^2$$
$$+T^2\sqrt{2d}\,\mathcal{C}_3(d)+2\sqrt{2}T(\mathcal{C}_1(d)+\mathcal{C}_2(d)).$$

Therefore, combining the results above, we have, for $C=\frac{M}{\sqrt{n}}(d+1)^{1/2}\le\frac{m}{16}$ and $a=\frac{m^2}{(4M_2h)^2}$,

$$e_{t+1}\le\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{t,h}-\tilde{V}_{t,h}\\ \hat{L}_{t,h}-\tilde{L}_{t,h}\end{bmatrix}\right\|_{L_2}+\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{t,h}-V_{t,h}\\ \hat{L}_{t,h}-L_{t,h}\end{bmatrix}\right\|_{L_2}$$
$$\le(1-\frac{mh}{2\gamma}+\frac{\sqrt{a}M_2h^2}{\gamma})e_t+\sqrt{2}(M-m)e^{-(a-p)/8}h^2+\frac{1}{4}M_2h^3(d+1)+\frac{1}{4}\sqrt{M^3}h^3d^{1/2}$$
$$+\sqrt{2}h^2d^{1/2}\mathcal{C}_3(d)+2\sqrt{2}h(\mathcal{C}_1(d)+\mathcal{C}_2(d))+\frac{3}{2}h(\mathcal{C}_1(d)+\mathcal{C}_2(d))+\frac{5}{6}h^2d^{1/2}\mathcal{C}_3(d)$$
$$\le(1-\frac{mh}{4\gamma}+\frac{2Ch}{\gamma})e_t+\sqrt{2}(M-m)e^{-\frac{m^2}{160M_2^2h^2}}h^2+\frac{1}{4}M_2h^3(d+1)+\frac{1}{4}\sqrt{M^3}h^3d^{1/2}$$
$$+\frac{9}{4}h^2d^{1/2}\mathcal{C}_3(d)+\frac{9}{2}h(\mathcal{C}_1'(d)+\mathcal{C}_2(d)).$$

Finally, we obtain

$$W_2(\varpi_t,\pi)\le\sqrt{2}(1-\frac{mh}{8\gamma})^tW_2(\varpi_0,\pi)+\frac{16M}{m}he^{-\frac{m^2}{160M_2^2h^2}}+\frac{2\sqrt{2}M_2}{m}h^2(d+1)+\frac{2\sqrt{2M^3}}{m}h^2d^{1/2}$$
$$+\frac{18\sqrt{2}}{m}hd^{1/2}\mathcal{C}_3(d)+\frac{36\sqrt{2}}{m}(\mathcal{C}_1'(d)+\mathcal{C}_2(d)).$$

This completes the proof of part 1. The proof of part 2 immediately follows from the statement in part 1 and the estimates in Theorem 5.6. ∎