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

Lingqing Shen*, Krishnakumar Balasubramanian†, and Saeed Ghadimi‡

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

February 4, 2019

Abstract

Euler-Maruyama and Ozaki discretization of a continuous time diffusion process is a popular technique for sampling, that uses (upto) gradient and Hessian information of the density respectively. The Euler-Maruyama discretization has been used particularly for sampling under the name of Langevin Monte Carlo (LMC) 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 problem 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 provide a theoretical analysis of the proposed sampling algorithm 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 contribution 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 $f(\theta) : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function and let $\pi(\theta)$ be a density function defined as follows:

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

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.,

$$
\theta^* = \arg\min_{\theta \in \mathbb{R}^d} f(\theta).
$$

Define the function $f_\tau(\theta) = f(\theta)/\tau$, for some $\tau > 0$ and note that $\theta^*$ is also the minimum point of $f_\tau(\theta)$. Note that 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 $\tau \rightarrow 0$, $\pi_\tau(\theta)$ 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 [3, 18, 21, 22, 23], 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 [25, 31, 32, 35, 36]. 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 [8, 9, 10, 11, 13, 14, 15] 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 [11]. 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, \( dL_T = -\nabla f(L_T) \, dT + \sqrt{2} \, dW_T, T \in \mathbb{R}_+ \), where \( \{W_T: T \in \mathbb{R}_+\} \) 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:

\[
x_{t+1,h} = x_{t,h} - h_{t+1} \nabla f(x_{t,h}) + \sqrt{2 h_{t+1}} \varepsilon_{t+1}
\]

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. Note that 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 \( \pi_t \). To evaluate the performance the sampling algorithm, the 2-Wasserstein distance between \( \pi_t \) 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:

\[
W_2(p, q) \overset{\text{def}}{=} \left( \inf_{\varrho \in \mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\theta - \theta'\|_2^2 \varrho(d\theta, d\theta') \right)^{1/2},
\]

where \( \mathcal{P}(\mathbb{R}^d) \) is the set of joint distribution that has \( p \) and \( q \) as its marginals. The performance of the LMC updates is measured by the above 2-Wasserstein distance between the distribution \( \pi_t \) and the target density \( \pi \), i.e., \( W_2(\pi_t, \pi) \). 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 [11] provided theoretical guarantees for sampling under Assumption 1.1. Specifically, instead of the true gradient \( \nabla f(x_{t,h}) \) in each step, it was 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 3, with \( \nabla f(x_{t,h}) \) replaced by \( g_{t,h} \). For such an update, [11] have the following non-asymptotic result.

**Theorem 1.2** [11] Assume that the bias and variance of \( \zeta_t \) satisfies respectively, for all \( t = 1, 2, \ldots \),

\[
\mathbb{E}[\|\mathbb{E}[\zeta_{t|x_{t,h}}]\|_2^2] \leq \delta^2 d \quad \text{and} \quad \mathbb{E}[\|\zeta_t - \mathbb{E}[\zeta_{t|x_{t,h}}]\|_2^2] \leq \sigma^2 d.
\]

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

\[
W_2(\pi_t, \pi) \leq (1 - mh)^t W_2(\pi_0, \pi) + 1.65 \frac{M}{m} (hd)^{1/2} + \frac{\delta \sqrt{d}}{m} + \frac{\sigma^2 (hd)^{1/2}}{1.65M + \sigma \sqrt{m}}.
\]
Remark 1 We note that more generally, if the bounded bias and variance condition are changed to
\[ E[\|E(\zeta_t|x_{t,h})\|_2^2] \leq \delta^2 d^\alpha \quad \text{and} \quad E[\|\zeta_t - E(\zeta_t|x_{t,h})\|_2^2] \leq \sigma^2 d^\beta, \]
respectively, for some \( \alpha, \beta > 0 \), the conclusion turns into
\[ W_2(\varpi_t, \pi) \leq (1 - mh)^t W_2(\varpi_0, \pi) + \frac{1.65 M (hd)^{1/2}}{m} + \frac{\delta d^{\alpha/2}}{m} + \frac{\sigma^2 hd^\beta}{1.65 M (hd)^{1/2} + \delta d^{\alpha/2} + \sigma (mh)^{1/2} d^\beta/2}. \]
Furthermore, in the case that \( \beta > \max\{1, \alpha\} \), the last term is dominated by \( d^{3/2} \).

Remark 2 [11] 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. Note that \( f_\tau \) 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
\[ W_2(\varpi_t, \pi_\tau) \leq (1 - \frac{m}{M})^t \sqrt{W_2(\delta_0, \pi_\tau) + 1.65 \left( \frac{M}{m} \right) \left( \frac{d\tau}{M} \right)^{1/2}}. \]
As we let \( \tau \to 0 \), we have the LMC updates converging to the standard gradient descent updates and the above bound becomes
\[ \|\theta_t - \theta^*\|_2 \leq (1 - \frac{m}{M})^t \|\theta_0 - \theta^*\|_2. \]

1.2 Our Contributions

Despite the impressive set of theoretical results in [11] and other related works [8, 9, 10, 13, 14, 15], 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. In several statistical machine learning models, the function \( f \) may not be directly accessible. But one could obtain (potentially noisy) evaluations of the function \( f \). Using the idea of Gaussian-smoothing based zeroth-order optimization [1, 16, 27], 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. Here, 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. Finally, we consider Ozaki-discretized LMC updates which involves the Hessian of the function $f(\theta)$. Note that [11] proposed theoretical guarantees for sampling with Ozaki-discretization, from which the corresponding results of the Newton method for optimization could be recovered. But [11] 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 [11] 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.

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 the 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 [37]. 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 Descent

<table>
<thead>
<tr>
<th>Discretization</th>
<th>Information</th>
<th>Structure</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>Exact first-order</td>
<td>GS &amp; SC</td>
<td>[11]</td>
</tr>
<tr>
<td>EM</td>
<td>Inexact first-order</td>
<td>GS &amp; SC</td>
<td>[11]</td>
</tr>
<tr>
<td>Ozaki</td>
<td>Exact first and second order</td>
<td>HS &amp; SC</td>
<td>[11]</td>
</tr>
<tr>
<td>Coordinate-wise EM</td>
<td>Exact first-order</td>
<td>GS &amp; SC</td>
<td>Theorem 2.2</td>
</tr>
<tr>
<td>EM</td>
<td>Exact zeroth-order</td>
<td>GS &amp; SC</td>
<td>Theorem 3.1</td>
</tr>
<tr>
<td>Variable selection+EM</td>
<td>Exact zeroth-order</td>
<td>GS &amp; SC</td>
<td>Theorem 4.3</td>
</tr>
<tr>
<td>Ozaki</td>
<td>Inexact first and second order</td>
<td>HS &amp; SC</td>
<td>Theorem 5.3 &amp; 5.4</td>
</tr>
<tr>
<td>Ozaki</td>
<td>Exact zeroth order</td>
<td>HS &amp; SC</td>
<td>Theorem 6.2</td>
</tr>
</tbody>
</table>

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.
LMC (RCD-LMC) is defined by the following update step:

\[
x_{t+1,h} = x_{t,h} - h [\nabla f(x_{t,h})]_i e_i + \sqrt{2h} (\varepsilon_{t+1})_i e_i.
\] (5)

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 [26, 37].

**Assumption 2.1** In this section, we assume that \(\nabla f\) is coordinate Lipschitz with constants \(M_i\), i.e.,

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

Denote \(M_{\text{max}} = \max_{1 \leq i \leq d} M_i\). Note that \(1 \leq \frac{M}{M_{\text{max}}} \leq d\), as can be seen intuitively by relating Lipschitz constants to the Hessian \(\nabla^2 f\); see also [37]. 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 \(h \leq 2/(m + M)\)

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

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

We now provide several remarks about the above result.

**Remark 3** One can compare the above result with the corresponding bound for the full-gradient based LMC algorithm. Recall that, \(M \leq dM_{\text{max}}\). Hence, we get comparable result in the worst case of \(M = dM_{\text{max}}\). But in the typical case of \(1 \leq M \ll dM_{\text{max}}\), 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 \(\frac{\tau}{M_{\text{max}}}\) respectively, and if we let \(\tau \to 0\), we obtain the following result:

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

This result recovers the corresponding result from the optimization literature for randomized coordinate descent [26, 37], which reads as

\[
\mathbb{E}[f(x_{t,h})] - f(x^*) \leq \left(1 - \frac{m}{dM_{\text{max}}}\right)^t (f(x_0) - f(x^*)).
\]
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 [20, 29, 30] 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, 16, 27], 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:

$$g_{\nu,n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\theta + \nu u_i) - f(\theta)}{\nu} u_i$$  \hspace{1cm} (6)

The gradient estimator in Equation 6 is a consequence of Gaussian Stein’s identity, popular in the statistics literature [34]. 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:

$$x_{t+1,h} = x_{t,h} - h g_{\nu,n}(x_{t,h}) + \sqrt{2}h \varepsilon_{t+1}$$  \hspace{1cm} (7)

for $t = 0, 1, 2, \ldots$. Note that, we 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 [11]. But one can’t leverage the corresponding result in [11] directly as in our setting, as the variance of the gradient is not bounded unless we make restorative assumption 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 $\frac{h n}{n(1-mh)} \leq \frac{m}{2M^2(d+1)}$,

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

$$+ 2\sqrt{2} \frac{\sqrt{M}}{\sqrt{m}} \cdot \frac{1}{\sqrt{n}} h^{1/2}(d+1) + \sqrt{2} \frac{M}{\sqrt{m}} \cdot \nu^2 \frac{2}{\sqrt{n}} h^{1/2}(d+2)^{3/2}.$$  \hspace{1cm} (8)

**Remark 5** Recall that for the exact first-order based LMC algorithm, by the right choice of tuning parameters, $t = \mathcal{O}(d/\varepsilon^2 \cdot \log(1/\varepsilon))$ suffices for $W_2(\varpi_t, \pi) \leq \varepsilon$; see [11]. For the ZO-LMC updates, for $W_2(\varpi_t, \pi) \leq \varepsilon$ when we require $t = \mathcal{O}(d/\varepsilon^2 \cdot \log(1/\varepsilon))$, setting $n = \mathcal{O}(d), \nu = \mathcal{O}\left(\varepsilon/\sqrt{d}\right), h = \mathcal{O}\left(\varepsilon^2/d^2\right)$ suffices. Note that with $n = 1$, it suffices to have $\nu = \mathcal{O}\left(\varepsilon/\sqrt{d}\right), h = \mathcal{O}\left(\varepsilon^2/d^2\right)$. 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

Note that in a practical black-box setting, due to the non-availability of the analytical form of $f(\theta)$, one might potentially over-parametrize $f(\theta)$, in terms of number of features selected to model
the sampling problem. 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 \( S^* \). Note that 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 \( \nabla f_\nu(\theta) = \mathbb{E}_u [\nabla f(\theta + \nu u)] \) for a standard gaussian random vector \( u \) and note that 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:

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

As a consequence, we also have that \( \nabla f_\nu(\theta) \in G_{a,s} \). 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. Note that the above sparsity model on the function \( f \), converts the problem to variable selection in a non-Gaussian sequence model setting:

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

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 [7] 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 on the selected set of coordinates \( \hat{S} \) of \( f(\theta) \).

Note that 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 \( \tau = (a - M\sqrt{s})/2 \) and assume that \( \nu \leq \frac{a}{2M\sqrt{s}} \wedge \frac{R}{MC_2\sqrt{s}} \) and

\[
n \geq \frac{8RC\sqrt{s}}{a} \left( 1 - \log \frac{4d}{\epsilon} \right)^{3/2} \wedge K_1 \frac{8RC\sqrt{s}}{a} \wedge \left( \frac{8RC\sqrt{s}}{a} \right)^4
\]

where \( C, C_2 \) are constants. Then we have \( \Pr \{ \hat{S} \neq S^* \} \leq \epsilon \).
Remark 6: Note that 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 [7].

5 Ozaki-Discretized Langevin Monte Carlo

We now consider the case of Ozaki-discretized LMC. Recall that discrete time LMC updates displayed in Equation 3, 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 [11, 28, 33] corresponds to the discrete time dynamics in Equation 9, that is based on the Hessian of $\pi(\theta)$). Let $H(\cdot) \equiv \nabla^2 f(\cdot) \in \mathbb{R}^{d \times d}$ be the true Hessian of the function $f$. We use the notation, $H_t \equiv H(D_{t,0})$ to denote the Hessian evaluated at point $D_{t,0}$ where $D_{t,0} \sim \pi_t$, the distribution of $x_t$, the $t^{th}$ step of the algorithm. Similarly, we denote by $S(\cdot) \in \mathbb{R}^{d \times d}$ the “inexact” Hessian of the function $f$ and $S_t \equiv S(D_{t,0})$. 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 $\nabla f_t \equiv \nabla f(D_{t,0})$ and $g_t \equiv g(D_{t,0})$.

5.1 OLMC with Inaccurate Gradients and Hessians

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

$$x_{t+1,h} = x_{t,h} - M_t \nabla f(x_{t,h}) + \Sigma_t^{1/2} \epsilon_{t+1}$$

where $M_t = (I_d - e^{-h H_t}) H_t^{-1}$ and $\Sigma_t = (I_d - e^{-h H_t}) H_t^{-1}$. The update steps of the Ozaki-discretized LMC (OLMC) algorithm with true Hessian and gradient information was analyzed in [11] and it was shown to have superior rates of convergence compared to the gradient based LMC algorithm. 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 the gradients are unavailable. Instead we observe a random gradient $g(\cdot)$ and random Hessian matrix $S(\cdot)$. Based on this, the OLMC with inexact information becomes

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

where, we have $\tilde{M}_t \equiv (I_d - e^{-h S_t}) S_t^{-1}$ and $\tilde{\Sigma}_t \equiv (I_d - e^{-2h S_t}) S_t^{-1}$. In the rest of this subsection, we assume that $S_t$ is invertible. The non-invertible case could potentially be handled by defining $\tilde{M}_t \equiv (I_d - e^{-h S_t}) (S_t + \lambda I_d)^{-1}$ and $\tilde{\Sigma}_t \equiv (I_d - e^{-2h S_t}) (S_t + \lambda I_d)^{-1}$ for some $\lambda > 0$. We do not pursue a detailed study of this case in this paper. We emphasize that when $S_t$
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., $\|H(\theta) - H(\theta')\| \leq M_2\|\theta - \theta'\|$, $\forall \theta, \theta' \in \mathbb{R}^d$.

Furthermore, note that gradient smoothness assumption in Assumption 1.1 implies boundedness of the second derivative, i.e., $H(\theta) \preceq M_1 I_d$, $\forall \theta \in \mathbb{R}^d$. Note that Assumption 5.1 and Assumption 1.1 ensure that the true hessian $H(\theta)$ 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 $S_t$ satisfies:

- $S_t, g_t, L_{0,0}$ 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
  \[
  \|g_t - \mathbb{E}[g_t|D_{t,0}]\|_{L_2} \leq C_1(d), \quad \|\nabla f_t - \mathbb{E}[\nabla f_t|D_{t,0}]\|_{L_2} \leq C_2(d),
  \]
  \[
  \|S_t - H_t| D_{t,0}\|_{L_2} \leq C_3(d), \quad \|e^{-TS_t}(S_t - H_t)| D_{t,0}\|_{L_2} \leq C_4(d),
  \]
  \[
  \|e^{-TS_t}S_t^2| D_{t,0}\|_{L_2} \leq C_5(d), \quad \|e^{-TS_t}S_t^2| D_{t,0}\|_{L_2} \leq C_6(d).
  \]

- $\tilde{M} \overset{\text{def}}{=} \sqrt{M\bar{M}} \lor M$ and $\hat{M}$ are constants satisfying the following inequalities.
  - In the case where $S_t \succeq 0$,
    \[
    \|S_t| D_{t,0}\|_{L_2} \leq \bar{M}, \quad \|S_t| D_{t,0}\|_{L_4} \leq \hat{M}.
    \]
  - In the case where $S_t \succeq 0$ does not hold in general,
    \[
    \left\| \|Me^{-S_t/M}\| \lor \|S_t\| \lor \|D_{t,0}\| \right\|_{L_2} \leq \bar{M},
    \]
    \[
    \left\| e^{-TS_t}S_t^2| D_{t,0}\|_{L_2} \leq \hat{M} = C_6(d).
    \]

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 6. 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:
Theorem 5.4  
For the Approximate OLMC updates, under Assumption 1.1, 5.1 and 5.2, we have  
the following guarantees:  

With inexact gradients and Hessian information, we then have  
in Equation 9 by series expansion. Specifically, we consider the following Approximate Ozaki-discretized LMC. Indeed a more practical discretization is obtained by approximating the matrix exponential  
perspective, it suffers from several computational drawbacks. Specifically, we need to compute  
While the Ozaki-discretized LMC is interesting from a theoretical perspective, from a practical  

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 the inverse of a matrix and matrix exponentials [24], both of which are computationally demanding. Indeed a more practical discretization is obtained by approximating the matrix exponential in Equation 9 by series expansion. Specifically, we consider the following Approximate Ozaki-discretized LMC updates considered also in [11]:  

\[ x_{t+1,h} = x_{t,h} - h \left( \mathbf{I}_d - \frac{1}{2} h \mathbf{H}_t \right) \nabla f(x_{t,h}) + \sqrt{2h} \left( \mathbf{I}_d - h \mathbf{H}_t + \frac{1}{3} h^2 \mathbf{H}_t^2 \right)^{1/2} \varepsilon_{t+1} \]  
(10)  

With inexact gradients and Hessian information, we then have  

\[ x_{t+1,h} = x_{t,h} - h \left( \mathbf{I}_d - \frac{1}{2} h \mathbf{S}_t \right) g(x_{t,h}) + \sqrt{2h} \left( \mathbf{I}_d - h \mathbf{S}_t + \frac{1}{3} h^2 \mathbf{S}_t^2 \right)^{1/2} \varepsilon_{t+1} \]  
(11)  

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

• If \( \mathbf{S}_t \) is positive definite, for \( h \leq m/\tilde{M}^2 \),  
  \[ W_2(\varpi_t, \pi) \leq (1 - 0.25mh)^t W_2(\varpi_0, \pi) + 7.2 \frac{M}{m} h(d + 1) \]  
  \[ + \frac{4}{\sqrt{6m}} h^{1/2} \mathcal{C}_1(d) + \frac{4}{m} \mathcal{C}_2(d) + \frac{5.27}{m} h^{3/2} d^{1/2} \mathcal{C}_3(d). \]  

• If \( \mathbf{S}_t \) is not positive definite in general, for \( h \leq m/\tilde{M}^2 \),  
  \[ W_2(\varpi_t, \pi) \leq (1 - 0.25mh)^t W_2(\varpi_0, \pi) + 3.85 \frac{M^2}{m^2} h^2(d + 1) \mathcal{C}_4(d)^2 + \frac{5}{m} (hd)^{1/2} \mathcal{C}_5(d) \]  
  \[ + \left( 3.51 \frac{M}{m} h(d + 1) + \frac{4}{m} (\mathcal{C}_1(d) + \mathcal{C}_2(d)) \right) \mathcal{C}_4(d) \]  
  \[ + \left( \frac{1.85}{m} h^{3/2} d^{1/2} + \frac{2}{3m} h^2 (\mathcal{C}_1(d) + \mathcal{C}_2(d)) \right) \mathcal{C}_6(d). \]
6 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

\[
E \left[ (uu^\top - I_d) g(u) \right] = E[\nabla^2 g(u)],
\]

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 \( f_{\nu}(\theta) = f(\theta + \nu u) \) and note that we have

\[
E \left[ (uu^\top - I_d) f(\theta + \nu u) \right] = E[\nabla^2 f(\theta + \nu u)] = \nabla^2 f_{\nu}(\theta) = H_{f_{\nu}}.
\]

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]:

\[
\begin{align*}
\hat{H}_{f_{\nu}} & \equiv \hat{H}_{f_{\nu}}(\theta, u) = \frac{1}{2\nu^2} \left( uu^\top - I_d \right) \left[ f(\theta + \nu u) - f(\theta) + f(\theta - \nu u) - f(\theta) \right], \\
\hat{H}_{f_{\nu}, n} & = \frac{1}{n} \sum_{i=1}^{n} \hat{H}_{f_{\nu}}(\theta, u_i).
\end{align*}
\]

Hence, ZOO-LMC and approximate ZOO-LMC corresponds to Equation 9 and 11 respectively, with the derivatives estimated based on Stein’s identity. In order to proceed, we note that a consequence of the Hessian smoothness assumption is the following assumption on the function \( f(\theta) \).

**Assumption 6.1** The function \( f \) is assumed to be twice differentiable and satisfy the following smoothness condition: for all points \( \theta, \theta' \in \mathbb{R}^d \),

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

**Lemma 6.1** Let the Hessian estimator be defined in (13) and Assumption 6.1 hold. Then, we have

\[
\|H_{f_{\nu}} - H_f\|_2 \leq M_2 \nu d^{1/2}.
\]

**Lemma 6.2** Under Assumption 6.1, we have

\[
\begin{align*}
\|\hat{H}_{f_{\nu}} - H_f\|_{L_2,2} & \leq \left\|\hat{H}_{f_{\nu}} - H_f\right\|_{L_2,F} \leq \frac{1}{6} M_2 \nu(d + 4)^{5/2} + \frac{1}{2} \|H_f\|_2(d + 3)^2. \\
\|\hat{H}_{f_{\nu}, n} - H_f\|_{L_2,2} & \leq \frac{1}{6\sqrt{n}} M_2 \nu(d + 4)^{5/2} + \frac{1}{2\sqrt{n}} \|H_f\|_2(d + 3)^2 + M_2 \nu d^{1/2}.
\end{align*}
\]
Based on the above result and Theorem 5.3, we have the following guarantees for the ZOO-LMC updates.

**Theorem 6.2** For the ZOO-LMC we have the following guarantees:

\[
W_2(\varpi_t, \pi) \leq (1 - 0.25m h)^t W_2(\varpi_0, \pi)
\]
\[
+ P_{n_H}^{d/2} \left( \frac{4\sqrt{2}}{\sqrt{5m}} \frac{1}{\sqrt{n_g}} h^{1/2} C'_1(d) + \frac{4}{m} C_2(d) + P_{n_H}^5 \frac{5.27}{m} (hd)^{1/2} C_3(d) \right)
\]
\[
+ P_{n_H}^{d/2} \left( \frac{16}{3} P_{n_H}^{d/2} + 3.51 \frac{M_2}{m} h(d+1) + P_{n_H}^5 \frac{5.27M^2}{m} \frac{1}{n_H} (hd)^{3/2} \right).
\]

Furthermore, the Approximate ZOO-LMC updates, we have:

\[
W_2(\varpi_t, \pi) \leq (1 - 0.25m h)^t W_2(\varpi_0, \pi)
\]
\[
+ P_{n_H}^{d/2} \left( \frac{2.51}{\sqrt{m}} \frac{1}{\sqrt{n_g}} h^{1/2} + P_{n_H}^8 \frac{M^2}{6m} h^2 (d+1)^4 C'_1(d) + \frac{4}{m} C_2(d) + P_{n_H}^5 \frac{5}{m} (hd)^{1/2} C_3(d) \right)
\]
\[
+ P_{n_H}^{d/2} \left( 3.89 P_{n_H}^{d/2} + 3.51 \frac{M_2}{m} h(d+1) + P_{n_H}^8 \frac{M^2}{3m} h^{3/2} (d+7)^{9/2} \right).
\]

where subscripts \(g\), \(H\) are used to distinguish parameters of gradient and Hessian estimators respectively. \(P_{n_H} = (1 - \frac{2Mh}{n_H})^{-1/2} \leq \sqrt{5}, P_{n_H}^{d/2} \leq (1 - 2Mhd)^{-1/2} \leq \sqrt{5} \) for \(h \leq \frac{1}{4M\alpha}, \) and

\[
C'_1(d) = \frac{1}{2} M\nu_g^2 (d+2)^{3/2} + \sqrt{M}(d+1), \quad C_2(d) = M\nu_g d^{1/2},
\]
\[
C_3(d) = \frac{1}{6\sqrt{n_H}} M_2 \nu_H (d+4)^{5/2} + \frac{1}{2\sqrt{n_H}} M (d+3)^2 + M_2 \nu_H d^{1/2}.
\]

**Remark 7** For ZOO-LMC in order for \(W_2(\varpi_t, \pi) \leq \epsilon\) to hold, it suffices to have \(h = O(\epsilon/d), \nu_g = O(\epsilon/\sqrt{d}), n_g = O(d/\sqrt{\epsilon}), \nu_H = O(\sqrt{d}/\epsilon), n_H = O(d^4/\epsilon)\) with \(t = O(d/\epsilon)\). For the approximate ZOO-LMC, it suffices to have \(h = O(\epsilon/d) \land O(\epsilon^{2/3}/d^3), \nu_g = O(\epsilon/\sqrt{d}), n_g = O(d/\sqrt{\epsilon}), n_H = O(d^4/\epsilon)\). 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 [19]. Furthermore, it is also interesting to derive zeroth-order versions of sampling using Hessian-based Kinetic Langevin diffusions [12].

### 7 Discussion

Recall that our gradient and Hessian estimators were based on Gaussian Stein’s identity and could be used only 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 [5, 6] and in computing volume of convex body [4], 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 \( S_p : \mathcal{M} \to \mathbb{R} \) associated to density \( p(u) \) defined over \( \mathcal{M} \) is defined as

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

Note that 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, [17].

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

\[
\int_{u \in \mathcal{M}} \nabla_u (g(u)p(u)) \, du = 0.
\]

Then it holds that

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

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**


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 $\|X\|_{L_2} = \mathbb{E}[\|X\|^2_{L_2}]^{1/2}$. The $L_p$ norms of a random matrix $M : \Omega \to \mathbb{R}^{d \times d}$ are defined as follows.

$$\|M\|_{L_p,2} = \mathbb{E}[\|M\|_{L_2}^p]^{1/p},$$
$$\|M\|_{L_p,F} = \mathbb{E}[\|M\|_{F}^p]^{1/p},$$

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 there is no confusion for simplicity.

B Proofs for Section 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

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

for $T \in [0,h]$, $i = 1,2,\ldots,d$, where the initial data is $L_{t,0,i} = x_i - L_{t-1,h,i}-1$, and $W_{t,T}$ is a Brownian motion satisfying $W_{t,h} = \sqrt{h} \epsilon_{t+1}$, 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 $t$. Then

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

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

17
where we note that \( h < \frac{2}{m + M} \). Finally, by definition of Wasserstein distance, we reach the results as desired.

We now 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

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

*Proof.* The proof of part (a) results from the proof of Lemma 4 in [11]. 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

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

\[\blacksquare\]

### 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 $\zeta = g_{\nu,n}(x_t) - \nabla f(x_t)$. First note that by Stein’s identity, $E[g_{\nu,n}(x_t)] = E[\nabla f(x + \nu u)] = \nabla f_{\nu}(x)$, where we denote $f_{\nu}(x) = E[f(x + \nu u)]$. Under Assumption 1.1 on smoothness of $f$, in the case where $n = 1$, we have the following calculation for the bias.

$$\|E[\zeta | x_t]\|^2 = \|E[\nabla f(x_t + \nu u) | x_t] - \nabla f(x_t)\|^2$$

$$\leq E[(M\nu\|u\|^2)]$$

$$\leq M^2\nu^2 d,$$

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

$$\zeta_t - E[\zeta_t | x_t] = \frac{f(x_t + \nu u) - f(x_t)}{\nu} u - \nabla f_{\nu}(x_t)$$

$$= \frac{f(x_t + \nu u) - f(x_t)}{\nu} - \nu u^\top \nabla f(x_t) u + (uu^\top - I) \nabla f(x_t) + (\nabla f(x_t) - \nabla f_{\nu}(x_t))$$

$$\overset{\text{def}}{=} A + B + C.$$

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

$$E[\|A\|^2 | x_t] = E\left[\left\| \frac{f(x_t + \nu u) - f(x_t)}{\nu} - \nu u^\top \nabla f(x_t) u \right\|^2 | x_t \right]$$

$$\leq E\left[\left(\frac{1}{2} M\nu\|u\|^2\right)^2\|u\|^2 | x_t \right]$$

$$\leq \frac{1}{4} M^2\nu^2 (d + 2)^3,$$

$$E[\|B\|^2 | x_t] = E[\|(uu^\top - I) \nabla f(x_t)\|^2 | x_t]$$

$$= \nabla f(x_t)^\top E_{w} \left[\|u\|^2 - 2 uu^\top + I\right] \nabla f(x_t)$$

$$= \nabla f(x_t)^\top E_{w} \left[\|u\|^2 - 2(uu^\top - I) + (\|u\|^2 - 1)I\right] \nabla f(x_t)$$

$$= \nabla f(x_t)^\top (2I + (d - 1)I) \nabla f(x_t)$$

$$= (d + 1)\|\nabla f(x_t)\|^2.$$
Combining with the fact that $C$ is deterministic and Lemma 3 from [27], the variance is bounded by

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

Next, for $n \geq 1$ 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

$$\|E[\zeta_t | x_t]\|^2 = \|E[g_{\nu,n}(x_t) - \nabla f(x_t) | x_t]\|^2 \leq \|E[g_{\nu,1}(x_t) - \nabla f(x_t) | x_t]\|^2 \leq M^2\nu^2 L.$$  

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

$$E[\|\zeta_t - E[\zeta_t | x_t]\|^2] = E[\|g_{\nu,n}(x_t) - f_\nu(x_t)\|^2] = \frac{1}{n} E[\|g_{\nu,1}(x_t) - f_\nu(x_t)\|^2] \leq \frac{1}{n} \left( M\nu(d+2)^{3/2} + (d+1)^{1/2}\|\nabla f(x_t)\|_{L^2} \right)^2.$$  

Next, we follow a similar framework to the proof of Theorem 4 in [11], 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, $\|\Delta_t - hU\| = \|\Delta_t - h[\nabla f(x_t + \Delta_t) - \nabla f(x_t)]\| \leq (1 - mh)\|\Delta_t\|$, $\|V\| = \|\int_0^h [\nabla f(L_s) - \nabla f(L_0)] ds\| \leq 1.65 M(h^3 d)^{1/2}$. Thus,

$$\|\Delta_{t+1}\|_{L^2} = \|\Delta_t - hU - V + h\zeta_t\|_{L^2} \leq \left\{ \|\Delta_t - hU\|^2_{L^2} + h^2\|\zeta_t - E[\zeta_t | x_t]\|^2_{L^2} \right\}^{1/2} + \|V\|_{L^2} + h\|E[\zeta_t | x_t]\|_{L^2}$$

$$\leq \left\{ (1 - mh)^2\|\Delta_t\|^2_{L^2} + \frac{h^2}{n} \left( M\nu(d+2)^{3/2} + \|\nabla f(L_0)\|_{L^2} + M\|\Delta_t\|_{L^2}(d+1)^{1/2} \right)^2 \right\}^{1/2} + 1.65 M(h^3 d)^{1/2} + M\nu h d^{1/2}$$

$$\leq \left\{ (1 - mh)^2\|\Delta_t\|^2_{L^2} + \frac{2h^2}{n} \left( M\nu(d+2)^{3/2} + \|\nabla f(L_0)\|_{L^2}(d+1)^{1/2} \right)^2 \right\}^{1/2} + \frac{M^2 h^2 (d+1)}{n(1-mh)} \|\Delta_t\|_{L^2} + 1.65 M(h^3 d)^{1/2} + M\nu h d^{1/2}$$

$$\leq \left\{ (1 - mh)^2\|\Delta_t\|^2_{L^2} + \frac{2h^2}{n} \left( M\nu(d+2)^{3/2} + \sqrt{M}(d+1) \right)^2 \right\}^{1/2} + \frac{1}{2} mh\|\Delta_t\|_{L^2} + 1.65 M(h^3 d)^{1/2} + M\nu h d^{1/2}.$$  

Here we use the fact that $\sqrt{a^2 + b + c} \leq \sqrt{a^2 + b} + \frac{c}{2a}$ and $E[\|\nabla f(L)\|^2] \leq Md$. By Lemma 9 in [11], the above inequality leads to

$$\|\Delta_t\|_{L^2} \leq (1 - 0.5mh)^2\|\Delta_0\|_{L^2} + 3.3\frac{M}{m}(hd)^{1/2} + 2\frac{M}{m}\nu d^{1/2}$$

$$+ 2\sqrt{\frac{M}{n}} \cdot \frac{1}{\sqrt{n}} h^{1/2}(d+1) + \sqrt{\frac{2}{n}} \cdot \nu h^{1/2}(d+2)^{3/2}.$$
Therefore, we obtain the bound in Wasserstein distance.

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

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

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

**D  Proofs for Section 4**

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

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

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

\[
\|\phi(\nu, u)\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}[|\phi(\nu, u)|^p])^{1/p} \\
\leq \sup_{p \geq 1} p^{-1} (\mathbb{E}[\|f(\theta + \nu u) - f(\theta) - \nabla f(\theta)^\top \nu u\|^p])^{1/p} + \sup_{p \geq 1} p^{-1} (\mathbb{E}[\|\nabla f(\theta)^\top u\|^p])^{1/p} \\
\leq \frac{1}{2} M\nu \sup_{p \geq 1} p^{-1} (\mathbb{E}[\|u\|^{2p}])^{1/p} + \|\nabla f(\theta)\| \sup_{p \geq 1} p^{-1} (\mathbb{E}[\|u\|^p])^{1/p} \\
\leq M\nu \|u\|_{\psi_2}^2 + \|\nabla f(\theta)\| \|u\|_{\psi_2} \\
\leq 2R \|u\|_{\psi_2}^2,
\]

where \( \|\cdot\|_{\psi_1} = \sup_{p \geq 1} p^{-1} \mathbb{E}[\cdot]^p]^{1/p} \) and \( \|\cdot\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} \mathbb{E}[|\cdot|^p]^{1/p} \) are the sub-exponential and sub-Gaussian norm respectively. In the last inequality we require that \( \nu \leq \frac{R}{M\|u\|_{\psi_2}}. \) Note that \( u \sim N(0, I_d) \) can be replaced by \( \sum_{k \in S^*} u_k e_k \sim N(0, I_s) \) due to Assumption 4.1. Moreover,
we have the following estimate.

\[ \|u_1\|_{\Psi_2} \leq \inf\{c > 0 : \mathbb{E} \left[ \exp \left\{ \frac{u_1^2}{c^2} \right\} \right] \leq 2 \} = \sqrt{\frac{8}{3}} \overset{\text{def}}{=} C_1, \]

\[ \|u\|_{\Psi_2} \leq \inf\{c > 0 : \mathbb{E} \left[ \exp \left\{ \|u\|^2 / c^2 \right\} \right] \leq 2 \}
\]

\[ = \sqrt{\frac{2}{1 - 2^{-2/d}}} \overset{\text{def}}{=} C_2 \sqrt{d}, \]

which implies that \( \|\phi(\nu, u)\|_{\Psi_1} \leq 2RC_2 \sqrt{s} \), \( \|u_1\|_{\Psi_2} \leq C_1 \). Thus the following concentration inequality follows by an application of Lemma 3 from [2], for \( n \geq \max \left\{ K_1 \frac{2RC_2 \sqrt{s}}{d}, \left( \frac{2RC_2 \sqrt{s}}{d} \right)^4 \right\} : \)

\[ \text{Pr}\{ \|z_j\| \geq \tau \} = \text{Pr}\left\{ \left\| \frac{1}{n} \sum_{k=1}^{n} g_{\nu,1}^k - \mathbb{E}[g_{\nu,1}] \right\| \geq \tau \right\}
\]

\[ \leq 4 \exp \left\{ -K_2 \left( \frac{n^2}{\|\phi(\nu, u)\|_{\Psi_1}\|u_1\|_{\Psi_2}} \right)^{2/3} \right\}
\]

\[ \leq 4 \exp \left\{ -K_2 \left( \frac{n^2}{2RC_2 \sqrt{s}} \right)^{2/3} \right\}, \]

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

\[ \text{Pr}\{ \hat{S} \neq S^* \} \leq \sum_{j \in D \setminus S^*} \text{Pr}\{ |z_j| > \tau \} + \sum_{j \in S^*} \text{Pr}\{ |z_j| > a' - \tau \}
\]

\[ \leq 4(d - s) \exp \left\{ -K_2 \left( \frac{n^2}{2RC_2 \sqrt{s}} \right)^{2/3} \right\} + 4s \exp \left\{ -K_2 \left( \frac{n(a' - \tau)}{2RC_2 \sqrt{s}} \right)^{2/3} \right\}
\]

\[ = 4 \exp \left\{ -K_2 \left( \frac{n(a - M\nu \sqrt{s})}{4RC_2 \sqrt{s}} \right)^{2/3} \right\}. \]

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

\[ n \geq \frac{8RC_2 \sqrt{s}}{a} \left( \frac{1}{K_2^2} \log \frac{4d}{\epsilon} \right)^{3/2} \vee K_1 \frac{8RC_2 \sqrt{s}}{a} \vee \left( \frac{8RC_2 \sqrt{s}}{a} \right)^4. \]

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

\[
dD_{t,T} = -[g_t + S_t(D_{t,T} - D_{t,0})]dT + \sqrt{2}dW_{t,T}.
\]

The Ornstein-Uhlenbeck process can be solved explicitly as

\[
D_{t,h} = D_{t,0} - \left( I_d - e^{-hS_t} \right) S_t^{-1} g_t + \left( \left( I_d - e^{-2hS_t} \right) S_t^{-1} \right)^{1/2} N[0, I_d],
\]

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 \( S_k \) and \( g_k \) 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

\[
X_T = -\int_0^T \nabla f(L_s)ds + \int_0^T [g_t + S_t(D_s - D_0)] ds
= -\int_0^T \{ \nabla f(L_s)ds - g_t - S_t(L_s - L_0) \} ds - \int_0^T S_t X_s ds.
\]

By Gronwall lemma (see Lemma 5 in [11]), we have

\[
X_T = -\int_0^T e^{-sS_t} \{ \nabla f(L_s) - g_t - S_t(L_s - L_0) \} ds
= \int_0^T e^{-sS_t} ds \{ \nabla f(D_0) - \nabla f(L_0) \}
+ \int_0^T e^{-sS_t} ds \{ g_t - E[g_t|D_0] \}
+ \int_0^T e^{-sS_t} ds \{ E[g_t|D_0] - \nabla f(D_0) \}
- \int_0^T e^{-sS_t} ds \{ \nabla f(L_s) - \nabla f(L_0) - \nabla^2 f(L_0)(L_s - L_0) \} ds
- \int_0^T e^{-sS_t} ds [S_t - \nabla^2 f(L_0)] \int_0^s \nabla f(L_u) du ds
+ \sqrt{2} \int_0^T e^{-sS_t} ds [\nabla^2 f(L_0) - \nabla^2 f(L_0)] W_s ds
+ \sqrt{2} \int_0^T e^{-sS_t} ds [S_t - \nabla^2 f(D_0)] W_s ds
def = A_T + I_T + J_T - B_T - C_T + P_T + Q_T.
\]

We now consider the two cases of \( S_t \) separately.

**Case 1**: \( S_t \geq 0 \).
By calculations similar to that in proof of Theorem 6 in [11], under the independence assumptions, we have the following bounds for each of the above terms.

\[
\| \Delta_0 + A_T \|_{L_2} \leq (1 - mT + 0.5MT^2) \| \Delta_0 \|_{L_2}
\]
\[
= (1 - mT + 0.5M^2T^2) \| \Delta_0 \|_{L_2}.
\]
\[
\| I_T \|_{L_2} \leq TC_1(d).
\]
\[
\| J_T \|_{L_2} \leq TC_2(d).
\]
\[
\| B_T \|_{L_2} \leq 0.877M^2T^2(d^2 + 2d)^{1/2}.
\]
\[
\| C_T \|_{L_2} \leq \mu \| \Delta_0 \|_{L_2} + \frac{1}{16\mu} M^2M_2T^2(d + 1) + \frac{1}{2} \sqrt{MT^2d^{1/2}C_3(d)}.
\]
\[
\| P_T \|_{L_2} \leq \frac{2}{3} MM_2T^3d \| \Delta_0 \|_{L_2}.
\]
\[
\| Q_T \|_{L_2} \leq \frac{2}{3} T^3dC_3(d)^2.
\]

Hence we have, for \( h \leq m/\bar{M}^2 \),

\[
\| \Delta_h \|_{L_2} = \| \Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h \|_{L_2}
\]
\[
\leq (\| \Delta_0 \|_{L_2}^2 + \| P_h \|_{L_2}^2 + \| M_h \|_{L_2}^2)^{1/2} + \| J_h \|_{L_2} + \| B_h \|_{L_2} + \| C_h \|_{L_2} + \| Q_h \|_{L_2}
\]
\[
\leq \{ (1 - mh + \frac{1}{2} \bar{M}^2h^2)^2 \Delta_0 \|_{L_2}^2 + h^2C_2^2(d) + \frac{2}{3} MM_2h^3d \Delta_0 \|_{L_2} \}^{1/2}
\]
\[
+ hC_2(d) + 0.877M^2h^2(d + 1) + \mu \| \Delta_0 \|_{L_2} + \frac{1}{16\mu} M^2M_2h^4(d + 1)
\]
\[
+ \frac{1}{2} \sqrt{Mh^2d^{1/2}C_3(d)} + \frac{\sqrt{6}}{3} h^{3/2}d^{1/2}C_3(d)
\]
\[
\leq \{ (1 - mh + \frac{1}{2} \bar{M}^2h^2)^2 \Delta_0 \|_{L_2}^2 + h^2C_2^2(d) \}^{1/2} + \frac{MM_2h^3d}{3(1 - mh + 0.5M^2h^2)}
\]
\[
+ hC_2(d) + 0.877M^2h^2(d + 1) + \mu \| \Delta_0 \|_{L_2} + \frac{1}{16\mu} M^2M_2h^4(d + 1)
\]
\[
+ \frac{1}{2} \sqrt{Mh^2d^{1/2}C_3(d)} + \frac{\sqrt{6}}{3} h^{3/2}d^{1/2}C_3(d)
\]
\[
\leq \{ (1 - mh + \frac{1}{2} \bar{M}^2h^2)^2 \Delta_0 \|_{L_2}^2 + h^2C_2^2(d) \}^{1/2} + \frac{2}{3} MM_2h^3d
\]
\[
+ hC_2(d) + 0.877M^2h^2(d + 1) + \frac{1}{4} m h \| \Delta_0 \|_{L_2} + \frac{M^2M_2}{4m} h^3(d + 1)
\]
\[
+ \frac{1}{2} \sqrt{Mh^2d^{1/2}C_3(d)} + \frac{\sqrt{6}}{3} h^{3/2}d^{1/2}C_3(d)
\]
\[
\leq \{ (1 - mh + 0.5\bar{M}^2h^2)^2 \| \Delta_0 \|_{L_2}^2 + h^2C_2^2(d) \}^{1/2} + 0.25 mh \| \Delta_0 \|_{L_2}
\]
\[
+ 1.8M^2h^2(d + 1) + hC_2(d) + 1.32h^{3/2}d^{1/2}C_3(d),
\]

where \( \mathbf{E}[\Delta_0 + A_h] = \mathbf{E}[\Delta_0 + A_h] = \mathbf{E}[I_h J_h] = 0 \) due to the independence assumption 5.2. Also, we use the inequality \( \sqrt{a^2 + b + c} \leq \sqrt{a^2 + b + \frac{c}{2}} \), note that \( 1 - mh + 0.5\bar{M}^2h^2 \geq 0.5 \), and take \( \mu = 0.25 mh \). Next, by application of Lemma 9 in [11], where \( A - D = 0.75 mh - 0.5\bar{M}^2h^2 \geq 0 \),
0.25mh, $A + D = 1.25mh - 0.5\tilde{M}^2h^2 \leq 0.75$, we have

$$\|\Delta t,0\|_{L_2} \leq (1 - 0.25mh)^t\|\Delta_{0,0}\|_{L_2} + \frac{7.18M_2}{m}h(d + 1) + \frac{4}{\sqrt{5m}}h^{1/2}C_1(d) + \frac{4}{m}C_2(d) + \frac{5.27}{m}(h(d))^{1/2}C_3(d).$$

Since $W_2(\pi, \pi) \leq \|D_{t,0} - L_{t,0}\|_{L_2} = \|\Delta t,0\|_{L_2}$, and in particular, equality holds for $t = 0$ by our choice of $L_{0,0}$, we obtain 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 [11].

**Case 2:** $S_t \geq 0$ does not hold in general.

Now we have a different estimate for the following bounds.

$$\|\Delta_0 + AT\|_{L_2} \leq (1 - mT + 0.5\tilde{M}^2T^2)\|\Delta_0\|_{L_2}$$

$$\|IT\|_{L_2} \leq TC_1(d)C_4(d)$$

$$\|JT\|_{L_2} \leq TC_2(d)C_4(d)$$

$$\|BT\|_{L_2} \leq 0.877M_2T^2(d^2 + 2d)^{1/2}C_4(d)$$

$$\|CT\|_{L_2} \leq \mu\|\Delta_0\|_{L_2} + \frac{1}{12\mu}M^2M_2T^4(d + 1)C_4(d)^2 + \frac{1}{2}\sqrt{\tilde{M}^2T^2d^{1/2}C_5(d)}$$

$$\|P_T\|_{L_2}^2 \leq \frac{2}{3}M_2T^3d\|\Delta_0\|_{L_2}C_4(d)^2$$

$$\|Q_T\|_{L_2}^2 \leq \frac{2}{3}T^3dC_5(d)^2.$$
where the calculation is similar to case 1. Again application of Lemma 9 in [11] finally leads to
\[
W_2(\varpi_t, \pi) \leq (1 - 0.25mh)^2W_2(\varpi_0, \pi) + 4 \frac{M^2M_2}{m^2}h^2(d + 1)C_4(d)^2 \\
+ \left( 3.51 \frac{M_2}{m}h(d + 1) + \frac{4}{\sqrt{5m}}h^{1/2}C_1(d) + \frac{4}{m}C_2(d) \right) C_4(d) + \frac{5.27}{m}(hd)^{1/2}C_5(d)
\]
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
\[
D_{t,T} - D_{t,0} = - \left( TI_d - \frac{1}{2} T^2 S_t \right) g_t + \sqrt{2} \int_0^T I_d - (T - u)S_t dW_{t,u}.
\]
From the representation
\[
D_{t,h} = D_{t,0} - h \left( I_d - \frac{1}{2} h S_t \right) g_t + \sqrt{2h} \left( I_d - h S_t + \frac{1}{3} h^2 S_t^2 \right) N[0, I_d],
\]
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
\[
dD_{t,T} = - (I_d - TS_t) g_t dt - \sqrt{2} S_t W_{t,T} dt + \sqrt{2} dW_{t,T}
\]
\[
= -[g_t + S_t(D_{t,T} - D_{t,0})] dt + \sqrt{2} dW_{t,T}
\]
\[
+ TS_t g_t dt - \sqrt{2} S_t W_{t,T} dt
\]
\[
= -[g_t + S_t(D_{t,T} - D_{t,0})] dt + \sqrt{2} dW_{t,T}
\]
\[
+ \frac{1}{2} T^2 S_t^2 g_t dt - \sqrt{2} S_t^2 \int_0^T (T - u) dW_{t,u} dt.
\]
Recall \(X_T = (L_T - L_0) - (D_T - D_0)\). Now
\[
X_T = - \int_0^T \{ \nabla f(L_s) ds - g_t - S_t(L_s - L_0) \} ds - \int_0^T S_t X_s ds
\]
\[
+ \frac{1}{2} \int_0^T s^2 S_t^2 g_t ds - \sqrt{2} \int_0^T S_t^2 \int_0^s (s - u) dW_u ds.
\]
By Gronwall lemma, \(X_T\) is solved as
\[
X_T = - \int_0^T e^{-sS_t} \{ \nabla f(L_s) - g_t - S_t(L_s - L_0) \} ds
\]
\[
+ \frac{1}{2} \int_0^T e^{-sS_t} s^2 ds S_t^2 g_t - \sqrt{2} S_t^2 \int_0^T e^{-sS_t} \int_0^s (s - u) dW_u ds
\]
\[
\overset{\text{def}}{=} (A_T + I_T + J_T - B_T - C_T + P_T + Q_T) + E_T - F_T.
\]
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: $S_t \geq 0$.**

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

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

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

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

\[
\|\Delta_h\|_{L_2} = \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h - E_h - F_h\|_{L_2}
\]

\[
\leq (\|\Delta_0 + A_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|I_h + J_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|E_h\|_{L_2} + \|F_h\|_{L_2}
\]

\[
\leq \left\{ (1 - mh + \frac{1}{2} \hat{M}^2 h^2)^2 \|\Delta_0\|_{L_2}^2 + \frac{2}{3} M M_2 d h^3 \|\Delta_0\|_{L_2} \right\}^{1/2}
\]

\[
+ h (C_1(d) + C_2(d)) + 0.877 M_2 h^2 (d + 1) + \mu \|\Delta_0\|_{L_2} + \frac{1}{16\mu} M_2^2 M_2 h^4 (d + 1)
\]

\[
+ \frac{1}{2} \sqrt{M} h^2 d^{1/2} C_3(d) + \frac{\sqrt{6}}{3} h^{3/2} d^{1/2} C_3(d)
\]

\[
+ \frac{1}{6} \hat{M}^2 h^3 \left( \sqrt{M} d^{1/2} + M \|\Delta_0\|_{L_2} + C_1(d) + C_2(d) \right) + \frac{1}{\sqrt{10}} \hat{M}^2 h^{5/2} d^{1/2}
\]

\[
\leq (1 - mh + 0.5 \hat{M}^2 h^2) \|\Delta_0\|_{L_2} + \frac{M M_2 h^3 d}{3(1 - mh + 0.5 \hat{M}^2 h^2)} + h (C_1(d) + C_2(d))
\]

\[
+ 0.877 M_2 h^2 (d + 1) + \frac{1}{4} m h \|\Delta_0\|_{L_2} + \frac{M_2^2 M_2}{4 m} h^3 (d + 1) + \frac{1}{2} \sqrt{M} h^2 d^{1/2} C_3(d)
\]

\[
+ \frac{\sqrt{6}}{3} h^{3/2} d^{1/2} C_3(d) + \frac{1}{6} \hat{M}^2 h^3 \left( \sqrt{M} d^{1/2} + M \|\Delta_0\|_{L_2} + C_1(d) + C_2(d) \right) + \frac{1}{\sqrt{10}} \hat{M}^2 h^{5/2} d^{1/2}
\]

\[
\leq (1 - 0.25 mh) \|\Delta_0\|_{L_2} + 1.54 M_2 h^2 (d + 1) + 0.47 \hat{M}^2 h^{5/2} d^{1/2}
\]

\[
+ 1.10 h (C_1(d) + C_2(d)) + 1.25 h^{3/2} d^{1/2} C_3(d),
\]

where we use the inequality $\sqrt{a^2 + b} \leq a + \frac{b}{2a}$. Note also that $1 - mh + \frac{1}{2} \hat{M}^2 h^2 \geq \frac{17}{32}$. and $1 - \frac{3}{4} mh + \frac{1}{2} \hat{M}^2 h^2 + \frac{1}{6} M M_2 h^3 \leq 1 - \frac{1}{4} mh$. Recursively applying the above result, we obtain

\[
W_2(\varpi, \pi) \leq (1 - 0.25 mh)^4 W_2(\varpi_0, \pi) + 6.15 \frac{M^2}{m} h(d + 1) + 1.85 \frac{\hat{M}^2}{m} h^{3/2} d^{1/2}
\]

\[
+ \frac{4.38}{m} (C_1(d) + C_2(d)) + \frac{5}{m} (hd)^{1/2} C_3(d).
\]

**Case 2: $S_t \geq 0$ does not hold in general.**

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

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

\[
\|F_T\|_{L_2} \leq \frac{1}{\sqrt{10}} T^{5/2} d^{1/2} C_6(d).
\]
The following calculation goes, for $h \leq 3m/(4M\tilde{M})$ and $3m/(4\tilde{M}^2)$,

$$\|\Delta_h\|_{L_2} = \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h + E_h - F_h\|_{L_2}$$

$$\leq (\|\Delta_0 + A_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|I_h + J_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|Q_h\|_{L_2} + \|E_h\|_{L_2} + \|F_h\|_{L_2}$$

$$\leq \{(1 - mh + \frac{1}{2}\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2}^2 + \frac{2}{3}MM_2h^3dC_4(d)^2\|\Delta_0\|_{L_2}\}^{1/2}$$

$$+ h(C_1(d) + C_2(d)) C_4(d) + 0.877M_2h^2(d^2 + 2d)^{1/2}C_4(d) + \frac{\nu d}{\nu d^2}C_5(d)$$

$$+ \frac{1}{6}\tilde{M}^2h^3 \left(\sqrt{M}d^{1/2} + M\|\Delta_0\|_{L_2} + C_1(d) + C_2(d)\right) + \frac{1}{\nu 10}h^{5/2}d^{1/2}C_6(d)$$

$$\leq (1 - mh + 0.5\tilde{M}^2h^2)^2\|\Delta_0\|_{L_2} + \frac{MM_2h^3d}{3(1 - mh + 0.5\tilde{M}^2h^2)}C_4(d)^2 + h(C_1(d) + C_2(d)) C_4(d)$$

$$+ 0.877M_2h^2(d + 1)C_4(d) + \frac{1}{4}mh\|\Delta_0\|_{L_2} + \frac{M^2M_2}{3m}h^3(d + 1)C_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}C_5(d)$$

$$+ \frac{\nu d}{\nu d^2}C_5(d) + \frac{1}{6}\tilde{M}^2h^3 \left(\sqrt{M}d^{1/2} + M\|\Delta_0\|_{L_2} + C_1(d) + C_2(d)\right) + \frac{1}{\nu 10}h^{5/2}d^{1/2}C_6(d)$$

$$\leq (1 - 0.25mh)\|\Delta_0\|_{L_2} + 0.97\frac{M^2M_2}{m}h^3(d + 1)C_4(d)^2 + 1.25h^{3/2}d^{1/2}C_5(d)$$

$$+ \left(0.47h^{5/2}d^{1/2} + \frac{1}{6}h^3(C_1(d) + C_2(d))\right) C_6(d).$$

Therefore, we end up with

$$W_2(\omega_t, \pi) \leq (1 - 0.25mh)^tW_2(\omega_0, \pi)$$

$$+ 3.85\frac{M^2M_2}{m^2}h^2(d + 1)C_4(d)^2 + \frac{5}{m}(hd)^{1/2}C_5(d)$$

$$+ \left(3.51\frac{M_2}{m}h(d + 1) + \frac{4}{m}(C_1(d) + C_2(d))\right) C_4(d)$$

$$+ \left(1.85\frac{M^2}{m}h^{3/2}d^{1/2} + \frac{2}{3m}h^2(C_1(d) + C_2(d))\right) C_6(d).$$

\[\square\]

**G Proofs for section 6**

*Proof. [Proof of Lemma 6.1] Under Assumption 5.1 that $f$ has Lipschitz smooth Hessian, we have*

$$\|H_{f,c} - H_f\|_2 \leq \|E[\nabla^2 f(x + \nu u)] - \nabla^2 f(x)\|_2$$

$$\leq E[\|\nabla^2 f(x + \nu u) - \nabla^2 f(x)\|_2]$$

$$\leq E[M_2\nu \|u\|] \leq M_2\nu d^{1/2}.$$
Proof. [Proof of Lemma 6.2] Taking $y = x + \nu u$ in Equation (15),

$$|f(x + \nu u) - f(x) - \nu \nabla^\top f(x)u - \frac{\nu^2}{2} u^\top \nabla^2 f(x)u| \leq \frac{M_2 \nu^3}{6} \|u\|^3. \quad (19)$$

Note also that $\|uu^\top - I_d\|^2 \leq (\|u\|^2 - 1)^2 + 1$, $\|uu^\top - I_d\|_F^2 = (\|u\|^2 - 1)^2 + d - 1$ and $E[\|u\|^{2p}] = \frac{(d + 2p - 2)!}{(d - 2)!}$. To apply (19), we split the error into two terms,

$$\begin{align*}
\hat{H}_{f_\nu} - H_f &= \frac{1}{2\nu^2}(uu^\top - I_d)[f(x + \nu u) - f(x) + f(x - \nu u) - f(x)] - H_f \\
&= \frac{1}{2\nu^2}(uu^\top - I_d)[f(x + \nu u) - f(x) - \frac{\nu^2}{2} u^\top H_f u + f(x - \nu u) + f(x) - \frac{\nu^2}{2} u^\top H_f u] \\
&\quad + (uu^\top - I_d)\frac{1}{2} u^\top H_f u - H_f \\
&\overset{d_{f_\nu}}{=} A + B,
\end{align*}$$

where

$$\begin{align*}
E[\|A\|^2_{\hat{f}_\nu}] &\leq \left(\frac{M_2 \nu^3}{6}\right)^2 E[\|uu^\top - I_d\|^2_{\hat{f}_\nu}\|u\|^6] \\
&\leq \left(\frac{M_2 \nu^3}{6}\right)^2 (d + 4)^5, \\
E[\|B\|^2_{\hat{f}_\nu}] &\leq E[\|(uu^\top - I_d)\frac{1}{2} u^\top H_f u\|^2_{\hat{f}_\nu}] \\
&\leq \left(\frac{1}{2}\|H_f\|_2\right)^2 E[\|uu^\top - I_d\|^2_{\hat{f}_\nu}\|u\|^4] \\
&\leq \left(\frac{1}{2}\|H_f\|_2\right)^2 (d + 3)^4.
\end{align*}$$

Therefore,

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

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

$$\begin{align*}
\|\hat{H}_{f_\nu,n} - H_f\|_{L_2,F} &\leq \|\hat{H}_{f_\nu,n} - H_{f_\nu}\|_{L_2,F} + \|H_{f_\nu} - H_f\|_{L_2,F} \\
&= \frac{1}{\sqrt{n}} \|\hat{H}_{f_\nu} - H_{f_\nu}\|_{L_2,F} + \|H_{f_\nu} - H_f\|_{L_2,F} \\
&\leq \frac{1}{\sqrt{n}} \|\hat{H}_{f_\nu} - H_f\|_{L_2,F} + \|H_{f_\nu} - H_f\|_{L_2,F}.
\end{align*}$$

Applying previous lemmas leads to

$$\|\hat{H}_{f_\nu,n} - H_f\|_{L_2,F} \leq \frac{1}{6\sqrt{n}} M_2 \nu (d + 4)^{5/2} + \frac{1}{2\sqrt{n}} \|H_f\|_2 (d + 3)^2 + M_2 \nu d^{1/2}. \quad \blacksquare$$
Proof. [Proof of Theorem 6.2] We first derive the bounds $C_4(d)$, $C_5(d)$, $C_6(d)$ as defined in Assumption 5.2. First, consider the Hessian estimator $\hat{S}_t = \hat{H}_{f_{\nu}}$ from a single Gaussian sample. Write $\hat{H}_{f_{\nu}} = (uu^T - I_d)\phi(\nu, u)$ where

$$\phi(\nu, u) \overset{def}{=} \frac{1}{2\nu^2} [f(x + \nu u) - f(x) + f(x - \nu u) - f(x)]$$

$$= \frac{1}{2} u^T \nabla^2 F(x + \nu u) u \in [0, \frac{1}{2} M\|u\|^2].$$

The eigenvalues of $\hat{H}_{f_{\nu}}$ 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.

$$\|e^{-T\hat{H}_{\nu}}\|_{L_2}^2 = \mathbf{E}[\|e^{-T\hat{H}_{\nu}}\|_2^2]$$

$$\leq \mathbf{E}[e^{TM\|u\|^2}] = (1 - 2TM)^{-d/2} \leq P_1^d,$$

$$\|e^{-T\hat{H}_{\nu}}\hat{H}_{f_{\nu}}\|_{L_2}^2 = \mathbf{E}[\|e^{-T\hat{H}_{\nu}}\hat{H}_{f_{\nu}}\|_2^2]$$

$$\leq \mathbf{E}[e^{TM\|u\|^2}\|\hat{H}_{f_{\nu}}\|_2^2]$$

$$\leq \mathbf{E}[e^{TM\|u\|^2}\|u\|^8 (\frac{1}{2} M\|u\|^2)^4 1_{\|u\|^2 \geq 2} + \mathbf{E}[e^{TM\|u\|^2}(\frac{1}{2} M\|u\|^2)^4 1_{\|u\|^2 \leq 2}]
$$

$$\leq (\frac{M}{2})^4 \left( \mathbf{E}[e^{TM\|u\|^2}\|u\|^{10}] + \mathbf{E}[e^{TM\|u\|^2}\|u\|^8(1 - \|u\|^4) 1_{\|u\|^2 \leq 2}] \right)
$$

$$\leq (\frac{M}{2})^4 P_1^{d+16}(d + 7)^8,$$

and moreover,

$$\|e^{-T\hat{H}_{\nu}}(\hat{H}_{f_{\nu}} - H_f)\|_{L_2} \leq \|e^{-T\hat{H}_{\nu}}(\hat{H}_{f_{\nu}} - \hat{H}_q)\|_{L_2} + \|e^{-T\hat{H}_{\nu}}\hat{H}_q\|_{L_2} + \|e^{-T\hat{H}_{\nu}}\|_{L_2}\|H_f\|_2$$

$$\leq \frac{1}{6} M_2 \nu P_1^{d/2 + 5}(d + 4)^{5/2} + \frac{1}{2} \|H_f\|_2 P_1^{d/2 + 4}(d + 3)^2,$$

where $\hat{H}_q = (uu^T - I_d)^{1/2} u^T H_f u$. The second inequality results from estimate as follows.

$$\mathbf{E}[\|e^{-T\hat{H}_{\nu}}(\hat{H}_{f_{\nu}} - \hat{H}_q)\|_2^2] \leq (\frac{1}{6} M_2 \nu)^2 \mathbf{E}[e^{TM\|u\|^2}\|uu^T - I_d\|_2^2\|u\|^6]
$$

$$\leq (\frac{1}{6} M_2 \nu)^2 \left( \mathbf{E}[e^{TM\|u\|^2}\|u\|^{10} 1_{\|u\|^2 \geq 2} + \mathbf{E}[e^{TM\|u\|^2}\|u\|^6 1_{\|u\|^2 \leq 2}] \right)
$$

$$= (\frac{1}{6} M_2 \nu)^2 \left( \mathbf{E}[e^{TM\|u\|^2}\|u\|^{10}] + \mathbf{E}[e^{TM\|u\|^2}\|u\|^6(1 - \|u\|^4) 1_{\|u\|^2 \leq 2}] \right)
$$

$$\leq (\frac{1}{6} M_2 \nu)^2 P_1^{d+10}(d + 4)^5,$$

$$\mathbf{E}[\|e^{-TH_{\nu}}\hat{H}_q\|_2^2] \leq (\frac{1}{2} \|H_f\|_2)^2 \mathbf{E}[e^{TM\|u\|^2}\|uu^T - I_d\|_2^2\|u\|^4]
$$

$$\leq (\frac{1}{2} \|H_f\|_2)^2 \left( \mathbf{E}[e^{TM\|u\|^2}\|u\|^{10} 1_{\|u\|^2 \geq 2} + \mathbf{E}[e^{TM\|u\|^2}\|u\|^4 1_{\|u\|^2 \leq 2}] \right)
$$

$$= (\frac{1}{2} \|H_f\|_2)^2 \left( \mathbf{E}[e^{TM\|u\|^2}\|u\|^{10}] + \mathbf{E}[e^{TM\|u\|^2}\|u\|^4(1 - \|u\|^4) 1_{\|u\|^2 \leq 2}] \right)
$$

$$\leq (\frac{1}{2} \|H_f\|_2)^2 P_1^{d+8}((d + 3)^2 - 2)^2.$$
version, $C_4(d)$ and $C_0(d)$ can be readily obtained.

$$
\|e^{-T\hat{H}_{f,v,n}}\|_{L_2} = E[\|e^{-\frac{T}{n}\hat{H}_{f,v,n}}\|_2^2]^{1/2}
\leq E[\|e^{-\frac{T}{n}\hat{H}_{f,v}}\|_2^2]
\leq (1 - \frac{2TM}{n})^{-nd/4} \leq P_n^{nd/2} = C_4(d),
$$

$$
\|e^{-T\hat{H}_{f,v,n}}\hat{H}_{f,v,n}^2\|_{L_2} = E[\|e^{-T\hat{H}_{f,v,n}}\hat{H}_{f,v,n}^2\|_2^2]^{1/2}
= \|e^{-\frac{T}{n}\hat{H}_{f,v,n}}\hat{H}_{f,v,n}\|_{L_4}^2
\leq \|e^{-\frac{T}{n}\hat{H}_{f,v,n}}\hat{H}_{f,v}\|_{L_4}^2
\leq E[\|e^{-\frac{T}{n}\hat{H}_{f,v}}\|_2(n-1)/2]E[\|e^{-\frac{T}{n}\hat{H}_{f,v}}\hat{H}_{f,v}^2\|_2^2]^{1/2}
\leq (\frac{M}{2})^2 P_n^{nd/2+8}(d + 7)^4 = C_0(d).
$$

Finally, we focus on estimating the bound $C_5(d)$.

$$
\|e^{-T\hat{H}_{f,v,n}}(\hat{H}_{f,v,n} - H_f)\|_{L_2} \leq \|e^{-T\hat{H}_{f,v,n}}(\hat{H}_{f,v,n} - \bar{H})\|_{L_2} + \|e^{-T\hat{H}_{f,v,n}}\|_{L_2} \|\bar{H} - H_f\|_2
\leq \frac{1}{\sqrt{n}} P_n^{nd/2+5}
\left\{ \frac{1}{2} M_2 \nu (d + 4)^{5/2} + \frac{1}{2} \|H_f\|_2^2 (d + 3)^2 \right\}
+ P_n^{nd/2} \left( P_n^5 M_2 \nu d^{1/2} + P_n^4 M^2 n^{-1} \right)
\leq P_n^{nd/2+5}
\frac{1}{6\sqrt{n}} M_2 \nu (d + 4)^{5/2} + \frac{1}{2} \nu^2 M (d + 3)^2 + M_2 \nu d^{1/2} + \frac{1}{n} M^2 h d
\right).
$$

Here $\bar{H}$ is the expectation of $\hat{H}_{f,v}$ after change of variable, replacing $u$ by $P_n u$, such that $E[\|\hat{H}_{f,v,n}(P_n u) - \bar{H}\|_{F}^2] = \frac{1}{n} E[\|\hat{H}_{f,v}(P_n u) - \bar{H}\|_{F}^2]$ holds. Then

$$
E[\|e^{-T\hat{H}_{f,v,n}}(\hat{H}_{f,v,n} - \bar{H})\|_2^2] \leq E[\|e^{-\frac{T}{n}\sum_{i=1}^M M_i u_i^2}\|\hat{H}_{f,v,n}(u) - \bar{H}\|_2^2]
\leq P_n^{nd} E[\|\hat{H}_{f,v,n}(P_n u) - \bar{H}\|_2^2]
\leq \frac{1}{n} P_n^{nd} E[\|\hat{H}_{f,v}(P_n u) - \bar{H}\|_F^2]
\leq \frac{1}{n} P_n^{nd} E[\|\hat{H}_{f,v}(P_n u)\|_F^2].
$$

Note that $\hat{H}_{f,v}(P_n u)$ is not a Hessian estimator of the form (13). However, the bound for $\|\hat{H}_{f,v}(P_n u)\|_{L_2,F}$ can still be given in a similar fashion.

$$
E[\|\hat{H}_{f,v}(P_n u)\|_F^2] = E[\|(P_n^2 u u^\top - I_d)(f(x + \nu P_n u) - f(x) + f(x - \nu P_n u) - f(x))\|_F^2]
\leq \left( E[\|P_n^2 u u^\top - I_d\|_F^2] \left( \frac{1}{2} M_2 \nu P_n^2 \|u\|_F^2 \right)^2 \right)^{1/2} + E[\|(P_n^2 u u^\top - I_d)(P_n^2 u u^\top - I_d)\|_F^2]
\leq \left( \frac{1}{2} P_n^2 M_2 \nu E[\|P_n^2 u u^\top - I_d\|_F^2] \|u\|_F^2 \right)^{1/2} + \frac{1}{2} P_n^2 \nu \|H_f\|_F^2 E[\|P_n^2 u u^\top - I_d\|_F^2] \|u\|_F^2 \|u\|_F^2 \right)^{1/2}
\leq \left( \frac{1}{2} P_n^2 M_2 \nu (d + 4)^{5/2} + \frac{1}{2} P_n^4 \nu \|H_f\|_F^2 (d + 3)^2 - 5 \right)^{1/2}
$$
To write out $\tilde{H}$ explicitly,
\[
\tilde{H} = \mathbf{E}[(P_n^2 uu^\top - I_d)\phi(\nu, P_nu)] \\
= P_n^4\mathbf{E}[(uu^\top - I_d)\phi(P_n\nu, u)] + (P_n^2 - 1)P_n^2\mathbf{E}[\phi(P_n\nu, u)]I_d \\
= P_n^4H_{f \nu} + (P_n^2 - 1)P_n^2\mathbf{E}[\phi(P_n\nu, u)]I_d.
\]

Then the second term in (20) can be estimated by
\[
\|\tilde{H} - H_f\|_2 = \|P_n^4(H_{f\nu} - H_f) + (P_n^2 - 1)P_n^2\mathbf{E}[\phi(P_n\nu, u)]I_d + (P_n^4 - 1)H_f\|_2 \\
\leq P_n^4\|H_{f\nu} - H_f\|_2 + P_n^4(1 - P_n^{-2})\mathbf{E}[\|M\|u\|^2] + P_n^4(1 - P_n^{-4})\|H_f\|_2 \\
\leq P_n^5M_2\nu d^{1/2} + P_n^4M^2n^{-1}hd + 4\|H_f\|_2P_n^4M n^{-1}h.
\]

Assuming that $\frac{h}{\nu n} \leq \frac{1}{3M}$, combining the above results gives rise to $C_5(d)$ in (20).

Since $\|g\_t - \mathbf{E}[g|D_0]\|_{L_2}$ 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,

\[
\|\Delta_h\|_{L_2} = \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h\|_{L_2} \\
\leq (\|\Delta_0\|_{L_2}^2 + \|I_h\|_{L_2}^2 + \|P_h\|_{L_2}^2)^{1/2} + \|J_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|Q_h\|_{L_2} \\
\leq \left\{(1 - mh + \frac{1}{2}M^2h^2)^{1/2}\|\Delta_0\|_{L_2}^2 + \frac{h^2}{n_{g}}(C'_1(d) + M(d + 1)^{1/2}\|\Delta_0\|_{L_2})^2\right\}C_4(d)^2 \\
+ \frac{2}{3}MM_2h^3dC_4(d)^2\|\Delta_0\|_{L_2}^{1/2} + hC_2(d)C_4(d) + 0.877M_2h^2(d + 2)^{1/2}C_4(d) \\
+ \mu\|\Delta_0\|_{L_2} + \frac{1}{12\mu}M^2M_2h^4(d + 1)C_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}C_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}C_5(d) \\
\leq \left\{(1 - mh + \frac{1}{2}M^2h^2)^{1/2}\|\Delta_0\|_{L_2}^2 + \frac{h^2}{n_{g}}C'_1(d)^2C_4(d)^2\right\}^{1/2} + \frac{M^2h^2(d + 1)C_4(d)^2}\|\Delta_0\|_{L_2} \\
+ \frac{MM_2h^3dC_4(d)^2}{3(1 - mh + 0.5M^2h^2)} + hC_2(d)C_4(d) + 0.877M_2h^2(d + 2)^{1/2}C_4(d) \\
+ \mu\|\Delta_0\|_{L_2} + \frac{1}{12\mu}M^2M_2h^4(d + 1)C_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}C_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}C_5(d) \\
\leq \left\{(1 - mh + \frac{1}{2}M^2h^2)^{1/2}\|\Delta_0\|_{L_2}^2 + \frac{h^2}{n_{g}}C'_1(d)^2C_4(d)^2\right\}^{1/2} + \frac{2}{n}M^2h^2(d + 1)^{1/2}C_4(d)^2\|\Delta_0\|_{L_2} \\
+ \frac{2}{3}MM_2h^3dC_4(d)^2 + hC_2(d)C_4(d) + 0.877M_2h^2(d + 2)^{1/2}C_4(d) \\
+ \frac{1}{8}mh\|\Delta_0\|_{L_2} + \frac{2M^2M_2}{3m}h^3(d + 1)C_4(d)^2 + \frac{1}{2}\sqrt{M}h^2d^{1/2}C_5(d) + \frac{\sqrt{6}}{3}h^{3/2}d^{1/2}C_5(d) \\
\leq \left\{(1 - mh + \frac{1}{2}M^2h^2)^{1/2}\|\Delta_0\|_{L_2}^2 + \frac{h^2}{n_{g}}C'_1(d)^2C_4(d)^2\right\}^{1/2} + \frac{1}{4}mh\|\Delta_0\|_{L_2} \\
\leq \frac{4M^2M_2h^3(d + 1)C_4(d)^2}{3m} + (0.877M_2h^2(d + 1) + hC_2(d))C_4(d) + 1.32h^3/2d^1/2C_5(d),
\]

where $C'_1(d) = \frac{1}{2}M^2\nu_0^2(d + 2)^{3/2} + \sqrt{M}(d + 1)$, $C_2(d) = M\nu_0\sqrt{d}$ as in Theorem 3.1, and $\frac{h}{n_{g}} \leq
Thus, for the Approximate ZOOLMC, 

\[
W_2(\varpi, \pi) \leq (1 - 0.25mh)^\epsilon W_2(\varpi_0, \pi) + \frac{16M^2M_2}{3m^2}h^2(d + 1)C_4(d)^2 \\
+ \left( 3.51 \frac{M_2}{m} h(d + 1) + \frac{4\sqrt{2}}{\sqrt{5m}} \frac{1}{\sqrt{n_g}} h^{1/2} C_4'(d) + \frac{4}{m} C_2(d) \right) C_4(d) + \frac{5.27}{m} (hd)^{1/2} C_5(d) \\
\leq (1 - 0.25mh)^\epsilon W_2(\varpi_0, \pi) \\
+ P_{nh}^{n_{h}d/2} \left( \frac{4\sqrt{2}}{\sqrt{5m}} \frac{1}{\sqrt{n_g}} h^{1/2} C_4'(d) + \frac{4}{m} C_2(d) + P_{nh}^5 \frac{5.27}{m} (hd)^{1/2} C_5(d) \right) \\
+ P_{nh}^{n_{h}d/2} \left( \frac{16}{3} P_{nh}^{n_{h}d/2} + 3.51 \frac{M_2}{m} h(d + 1) + P_{nh}^5 \frac{5.27M^2}{n_H} \frac{1}{n_H} (hd)^{3/2} \right).
\]

Further note that

\[
\|E_T\|_{L_2} = \left\| \frac{1}{2} \int_0^T e^{-s|s^2 ds^2 S_t g_t|} \right\|_{L_2} \\
\leq \frac{1}{6} M^2 T^3 \|g_t\|_{L_2} \\
\leq \frac{1}{6} M^2 T^3 \left( \frac{1}{2} M\nu(d + 2)^{3/2} + (d + 2)^{1/2} \|\nabla f(x_t)\| \right) \\
\leq \frac{1}{6} \tilde{M}^2 T^3 \left( \frac{1}{2} M\nu(d + 2)^{3/2} + \sqrt{M} d + 1 + M(d + 2)^{1/2} \|\Delta_0\|_{L_2} \right) \\
= \frac{1}{6} \tilde{M}^2 T^3 \left( C_4'(d) + M(d + 2)^{1/2} \|\Delta_0\|_{L_2} \right).
\]

Thus, for the Approximate ZOOLMC,

\[
\|\Delta_h\|_{L_2} = \|\Delta_0 + A_h + I_h + J_h - B_h - C_h + P_h + Q_h + E_h - F_h\|_{L_2} \\
\leq \left\{ (1 - mh + \frac{1}{2} \tilde{M}^2 h^2)^2 \|\Delta_0\|_{L_2}^2 + \frac{h^2}{n_g} \left( C_4'(d) + M(d + 1)^{1/2} \|\Delta_0\|_{L_2} \right)^2 C_4(d)^2 \right\}^{1/2} \\
+ \frac{2}{3} MM_2 h^3 d C_4(d)^2 \|\Delta_0\|_{L_2}^{1/2} + hC_2(d) C_4(d) + 0.877 M_2 h^2 (d^2 + 2d)^{1/2} C_4(d) \\
+ \mu \|\Delta_0\|_{L_2} + \frac{1}{12\mu} M^2 M_2 h^4 (d + 1) C_4(d)^2 + \frac{1}{2} \sqrt{M} h^2 d^{1/2} C_5(d) + \frac{\sqrt{6}}{3} h^{3/2} d^{1/2} C_5(d) \\
+ \frac{1}{6} \tilde{M}^2 h^3 \left( C_4'(d) + M(d + 2)^{1/2} \|\Delta_0\|_{L_2} \right) + \frac{1}{\sqrt{10}} h^{5/2} d^{1/2} C_6(d).
\]
Continuing the calculation, we have

\[
\|\Delta_h\|_{L_2} \leq \left\{ (1 - mh + \frac{1}{2} \hat{M}^2 h^2)^2 \|\Delta_0\|_{L_2} \leq 2 \frac{h^2}{n_0} c_1'(d) c_4(d)^2 \right\}^{1/2} + \frac{M^2 h^2 (d + 1) c_4(d)^2}{n_0 (1 - mh + 0.5 M^2 h^2)} \\
+ \frac{M M_2 h^3 d c_4(d)^2}{3 (1 - mh + 0.5 M^2 h^2)} + h c_2(d) c_4(d) + 0.877 M_2 h^2 (d^2 + 2d)^{1/2} c_4(d) \\
+ \frac{1}{8} m h \|\Delta_0\|_{L_2} + \frac{2 M^2 M_2}{3 m} h^3 (d + 1) c_4(d)^2 + \frac{1}{2} \sqrt{M} h^2 d^{1/2} c_5(d) + \frac{\sqrt{6}}{3} h^3 d^{1/2} c_5(d) \\
+ \frac{1}{6} \hat{M}^2 h^3 \left( c_1'(d) + M (d + 2)^{1/2} \|\Delta_0\|_{L_2} \right) + \frac{1}{\sqrt{10}} h^{5/2} d^{1/2} c_6(d) \\
\leq \left\{ (1 - mh + \frac{1}{2} \hat{M}^2 h^2)^2 \|\Delta_0\|_{L_2} \leq 2 \frac{h^2}{n_0} c_1'(d) c_4(d)^2 \right\}^{1/2} + \frac{1}{4} m h \|\Delta_0\|_{L_2} \\
+ \frac{1}{6} M \hat{M}^2 h^3 (d + 2)^{1/2} \|\Delta_0\|_{L_2} + 1.3 \frac{M^2 M_2}{m} h^3 (d + 1) c_4(d)^2 \\
+ (0.877 M_2 h^2 (d + 1) + h c_2(d)) c_4(d) \\
+ 1.25 h^{3/2} d^{1/2} c_5(d) + \left( \frac{1}{\sqrt{10}} h^{5/2} d^{1/2} + \frac{1}{6} h^3 c_1'(d) \right) c_6(d).
\]

Here we assume that \( \frac{h}{n_0} \leq \frac{17 m}{256 h^2 M^2 (d + 1)} \), and \( h \leq \frac{3 m}{4 M M_2 (d + 1)^{1/4}} \). Denote \( A = mh - \frac{1}{2} \hat{M}^2 h^2, \)
\( D = \frac{1}{4} mh - \frac{1}{6} M \hat{M}^2 h^3 (d + 2)^{1/2} \). Then \( A - D \geq \frac{1}{4} mh, A + D \leq \frac{3}{4} \cdot \frac{31}{32} \). Therefore,

\[
W_2(\varpi_t, \pi) \leq (1 - 0.25 mh)^4 W_2(\varpi_0, \pi) + 5.18 \frac{M^2 M_2}{m_0^2} h^2 (d + 1) c_4(d)^2 \\
+ \left( 3.51 \frac{M_2}{m} h (d + 1) + 2.51 \frac{1}{\sqrt{m}} \frac{1}{\sqrt{n_0}} h^{1/2} c_1'(d) + \frac{4}{m} c_2(d) \right) c_4(d) \\
+ \left( \frac{4}{\sqrt{10} m} h^{3/2} d^{1/2} + \frac{2}{3 m} h^2 c_1'(d) \right) c_6(d) + \frac{5}{m} (hd)^{1/2} c_5(d).
\]

Plugging in \( c_i(d) \)'s gives the results as desired, i.e.,

\[
W_2(\varpi_t, \pi) \leq (1 - 0.25 mh)^4 W_2(\varpi_0, \pi) \\
+ P_{n_0}^{n_0 h^{4d/2}} \left( 2.51 \frac{1}{\sqrt{m}} \frac{1}{\sqrt{n_0}} h^{1/2} + P_{n_0}^{8 M^2 M_2} h^2 (d + 7)^4 c_1'(d) + \frac{4}{m} c_2(d) + P_{n_0}^{5} \frac{5}{m} (hd)^{1/2} c_3(d) \right) \\
+ P_{n_0}^{n_0 h^{4d/2}} \left( 3.89 P_{n_0}^{4 M^2} + 3.51 \frac{M_2}{m} h (d + 1) + P_{n_0}^{4 M_2} \frac{M_2}{m} h^2 (d + 7)^{9/2} \right).
\]