# Stochastic Zeroth-order Discretizations of Langevin Diffusions for Bayesian Inference

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

[1]Department of Statistics, University of California, Davis
[2]Tepper School of Business, Carnegie Mellon University
[3]Department of Management Sciences, University of Waterloo

January 19, 2021

**Abstract**

Discretizations of Langevin diffusions provide a powerful method for sampling and Bayesian inference. However, such discretizations require evaluation of the gradient of the potential function. In several real-world scenarios, obtaining gradient evaluations might either be computationally expensive, or simply impossible. In this work, we propose and analyze stochastic zeroth-order sampling algorithms for discretizing overdamped and underdamped Langevin diffusions. Our approach is based on estimating the gradients, based on Gaussian Stein's identities, widely used in the stochastic optimization literature. We provide a comprehensive sample complexity analysis – number noisy function evaluations to be made to obtain an $\epsilon$-approximate sample in Wasserstein distance – of stochastic zeroth-order discretizations of both overdamped and underdamped Langevin diffusions, under various noise models. We also 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 noisy black-box and high-dimensional settings.

## 1 Introduction

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 rates of convergence. However, they are efficiently implementable and widely applicable, as they are based only on exact density function evaluations; see, for example, [BRS93, KLS95, MT96, LS90, LV07, DJ20, MFR20], for more details about such algorithms. Motivated by statistical physics principles, various researchers developed second-generation of sampling algorithms, that leverage geometric information regarding the target density [RR98, Nea11, RT96, ST99a, ST99b, GC11, BBKG18]. Such

*abroy@ucdavis.edu
†lingqins@andrew.cmu.edu. Work done while visiting UC Davis as an exchange student.
‡kbala@ucdavis.edu
§sghadimi@uwaterloo.ca

algorithms are based on gradient-based discretizations of continuous-time underdamped or overdamped Langevin diffusions. Although such algorithms were developed much earlier, recently strong theoretical guarantees have been established for sampling in the works of [DM17, DM$^+$19, Dal17, DK19, CCBJ18, CCAY$^+$18, DCWY19] and several others. Such algorithms typically perform empirically better and exhibit faster rates of convergence compared to the first generation sampling algorithms mentioned above.

In this work, given a density function $\pi : \mathbb{R}^d \to \mathbb{R}$, with potential function $f : \mathbb{R}^d \to \mathbb{R}$, of the form

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

we consider the problem of sampling when we only have access to noisy evaluations of the potential function $f$. We refer to this problem as *stochastic zeroth-order sampling*. Our approach is based on discretizing overdamped and underdamped Langevin diffusions using stochastic zeroth-order oracles, which, when queried returns noisy unbiased evaluations, $F(x, \xi)$, of the function value $f(x)$. That is, we have $\mathbf{E}[F(x, \xi)] = f(x)$, where $\xi$ is the random noise in our function evaluations, which is not necessarily an additive noise. Our motivations for studying such problems are threefold:

- **Computationally Complexity of Gradient-evaluation**: A majority of existing discretizations of Langevin diffusions require computing the gradient of the potential function $f$ in each iteration. It is well-know that for a wide class of functions which could be expressed based on compositions of elementary differentiable functions, the computational cost of evaluating the gradient is 4 to 5 times more than that of evaluating the function; see, for example [GW08]. Furthermore, in order to compute the gradient, it is necessary to store several intermediate gradients, which increases the memory requirement. Hence, for several potential functions, Langevin-discretization based sampling algorithms might end up spending more time and memory for computing and storing gradients in each iteration. To reduce the wall-clock runtimes of such sampling algorithms, it is of interest to develop discretization of Langevin diffusions based only on function evaluations.
- **Non-availability of Analytic form of Potential Function:** In a variety of scientific problems, the potential function $f$ might not even be available in closed form, either due to the sheer size of the dataset (see, for example [STRR15]), or due to the constraints in the physical process underlying the statistical model (see, for example [Bea03, GW11, KDV12]). In these situations, we do not have access to the analytical from of true potential function, let alone its gradients, which are required for discretizing Langevin diffusions. Hence, it is of great interest to develop discretization of Langevin diffusions based on noisy function evaluations to widen the applicability of Bayesian inference. It is worth mentioning here that, in the case of Metropolis-Hastings algorithms, [AR09, STRR15] developed and analyzed the so-called Pseudo-Marginal Metropolis-Hasting algorithms which work with unbiased noisy density evaluations. However, similar algorithms for sampling based on discretizing Langevin diffusions are lacking in the literature, except for the recent work on [ADL16] which considered a pseudo-marginal Hamiltonian Monte Carlo algorithm. Our second motivation for this work is to fill this gap and to develop and analyze a unified framework for stochastic zeroth-order discretization of Langevin diffusions for Bayesian inference.

- **Automating Bayesian Inference:** From a practitioner's perspective, statistical modeling is an inherently iterative process. The probabilistic model is typically refined during the scientific process based on the fit to the data. In the context of sampling, this process could be understood as changing the potential function $f$ in the modeling process. However, each time the function $f$ is changed, it is also invariably required to re-code the sampling algorithm based on the analytically computed gradient of the function $f$ under consideration. Our third motivation in this work is to automate this process, to help the practitioner with quick experimentation. As we will see later, our proposed methodology allows for sampling from a wide variety of density functions in a unified manner, as long as we have an oracle to obtain (noisy) evaluations of the potential function $f$. It is worth mentioning that recently [RGB14, RTCB15, KTR$^+$17] developed related automated Bayesian inference algorithms based on variational inference.

### 1.1 Preliminaries

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

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

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 process in (2) is given by the following Markov chain:

$$x_{n+1} = x_n - h_{n+1}\ \nabla f(x_n) + \sqrt{2h_{n+1}}\varepsilon_{n+1}, \tag{3}$$

for the discrete time index $n = 0, 1, 2 \ldots$. Here $\varepsilon_n \in \mathbb{R}^d$ is a sequence of independent standard Gaussian vectors, $h_n$ denotes the step-size and an initial point $x_0$ 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. As a prelude to the rest of the paper, our main idea in this work is to provide a non-asymptotic analysis of using stochastic zeroth-order gradient estimators (described in details in Section 1.2) in place of the true gradient in (3) and related discretizations.

Denoting the distribution of the random vector $x_n$ by $\varpi_n$, to evaluate the performance of the sampling algorithm, the 2-Wasserstein distance between $\varpi_n$ 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:

$$W_2(p, q) := \left( \inf_{\varrho \in \varrho(p,q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta - \theta'\|_2^2 \, d\varrho(\theta, \theta') \right)^{1/2}, \tag{4}$$

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 $\varpi_n$ and the target density $\pi$, i.e., $W_2(\varpi_n, \pi)$. Specifically, the *iteration complexity* of the algorithm is defined as the number of iterations $N$, required to get $W_2(\varpi_N, \pi) \leq \epsilon$. We also define the notion of *oracle complexity* which is the number of calls to the first-order or stochastic zeroth-order oracle used to obtain $W_2(\varpi_N, \pi) \leq \epsilon$. For the LMC algorithm in (3), as we use only one gradient evaluation in each iteration, the oracle and iteration complexity becomes the same.

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** *Letting $\|\cdot\| = \|\cdot\|_2$ denote the Euclidean norm on $\mathbb{R}^d$, the potential function $f$,*

***A1**: is strongly convex i.e., $f(\theta) - f(\theta') - \nabla f(\theta')^\top(\theta - \theta') \geq \frac{m}{2}\|\theta - \theta'\|^2$, for $m > 0$.*

***A2**: has Lipschitz continuous gradient, i.e., $\|\nabla f(\theta) - \nabla f(\theta')\| \leq M\|\theta - \theta'\|$ for $M > 0$.*

The above assumptions on the potential function in-turn makes the density function $\pi$ strongly log-concave and smooth. Such an assumption is satisfied in several sampling and Bayesian inference problems including sampling from mixture of Gaussian distributions and Bayesian logistic regression. Further assuming access to certain inaccurate gradients, [DK19] provide theoretical guarantees for sampling under Assumption 1.1. Specifically, instead of the true gradient $\nabla f(x_n)$ in each step, it is assumed that we observe $g_n = g(x_n) = \nabla f(x_n) + \zeta_n$, for a sequence of random noise vectors $\zeta_n$ that satisfies certain bias and variance assumption. Then, the noisy LMC updates corresponds to the case of the updates in Equation 3, with $\nabla f(x_n)$ replaced by $g_n$. For such an update, [DK19] have the following non-asymptotic result. Before providing the result, we remark that due to the assumptions on the stochastic gradient made in (5), this setting is referred to as stochastic first-order setting.

**Theorem 1.2** *[DK19] Assume that the bias and variance of $\zeta_n$ satisfies respectively, for all $n = 1, 2, \ldots,$*

$$\mathbf{E}[\|\mathbf{E}(\zeta_n|x_n)\|^2] \leq \delta_b^2 d \qquad \text{and} \qquad \mathbf{E}[\|\zeta_n - \mathbf{E}(\zeta_n|x_n)\|^2] \leq \delta_v^2 d. \tag{5}$$

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

$$W_2(\varpi_n, \pi) \leq (1-mh)^n W_2(\varpi_0, \pi) + 1.65\frac{M}{m}(hd)^{1/2} + \frac{\delta_b\sqrt{d}}{m} + \frac{\delta_v^2(hd)^{1/2}}{1.65M + \sigma\sqrt{m}}.$$

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

$$\mathbf{E}[\|\mathbf{E}(\zeta_n|x_n)\|^2] \leq \delta_b^2 d^\alpha \qquad \text{and} \qquad \mathbf{E}[\|\zeta_n - \mathbf{E}(\zeta_n|x_n)\|^2] \leq \delta_v^2 d^\beta,$$

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

$$\begin{aligned} W_2(\varpi_n, \pi) \leq & (1-mh)^n W_2(\varpi_0, \pi) + \frac{1.65M(hd)^{1/2}}{m} + \frac{\delta_b d^{\alpha/2}}{m} \\ & + \frac{\delta_v^2 h d^\beta}{1.65M(hd)^{1/2} + \delta_b d^{\alpha/2} + \delta_v(mh)^{1/2} d^{\beta/2}}. \end{aligned}$$

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

## 1.2 The Zeroth-order Methodology

The use of zeroth-order information (i.e., noisy function evaluations) for optimizing a function goes back to the works of [KW52, Blu54], that used stochastic version of finite-difference gradient approximation methods for estimating the maximum of a regression function (or equivalently mode of a density function). Since, then zeroth-order optimization has developed into an independent field in itself; see, for example [Spa05, CSV09, AH17, LMW19] for an more up-to-date account of this field. More recently, the focus has been more on developing a non-asymptotic understanding of stochastic zeroth-order optimization [GL13, DJWW15, NS17, BG19]. Despite the fact that stochastic zeroth-order optimization is a well-developed field, to the best of our knowledge, there is no prior work on using related techniques for the closely related problem of zeroth-order discretizations of Langevin diffusions; specifically in terms of non-asymptotic analysis.

We now describe the precise assumption made on the *stochastic zeroth-order oracle* in the first part of this work.

**Assumption 1.3** *For any $\theta \in \mathbb{R}^d$, the stochastic zeroth-order oracle outputs an estimator $F(\theta, \xi)$ of $f(\theta)$ such that, $\mathbf{E}\left[F(\theta,\xi)\right] = f(\theta)$, $\mathbf{E}\left[\nabla F(\theta,\xi)\right] = \nabla f(\theta)$, and $\mathbf{E}\left[\|\nabla F(\theta,\xi) - \nabla f(\theta)\|^2\right] \leq \sigma^2$.*

The assumption above assumes that we have accesses to a stochastic zeroth-order oracle which provides unbiased function evaluations with bounded variance. It is worth noting that in the above, we do not necessarily assume the noise $\xi$ is additive. Our gradient estimator is then constructed by leverage the Gaussian smoothing technique [NS17, GL13, BG19], which is amenable for fine-grained non-asymptotic analysis. Specifically, for a point $\theta \in \mathbb{R}^d$, we define an estimate $g_{\nu,b}(\theta)$, of the gradient $\nabla f(\theta)$ as follows:

$$g_{\nu,b}(\theta) = \frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta + \nu u_i, \xi_i) - F(\theta, \xi_i)}{\nu} u_i \tag{6}$$

where $u_i \sim N(0, I_d)$ and are assumed to be independent and identically distributed. An interpretation of the gradient estimator in (6) as a consequence of Gaussian Stein's identity, popular in the statistics literature [Ste72], was provided in [BG19]. Finally, the parameter $b$ is called as the batch-size parameter. It turns out that in the stochastic zeroth-order setting invariably we require $b > 1$, which in turn leads to the (zeroth-order) oracle complexity being an order $b$ times that of iteration complexity. In Section 2 and 3.1, we use the above gradient estimation technique in the context of discretizing overdamped and underdamped Langevin diffusion and develop their oracle and iteration complexities. In order to establish the results, we will use the following Lemma due to [BG19] which provides an upper bound on the variance of $g_{\nu,b}$.

**Lemma 1.1** *[BG19] Let $g_{\nu,b}$ be defined as in* (6)*. Then under Assumption 1.3, and condition A1 of Assumption 1.1, we have,*

$$\mathbf{E}\left[\|g_{\nu,b}(\theta) - \nabla f_\nu(\theta)\|^2\right] \leq \frac{2(d+5)(\|\nabla f(\theta)\|^2 + \sigma^2)}{b} + \frac{\nu^2 M^2 (d+3)^3}{2b} \tag{7}$$

$$\mathbf{E}\left[\|g_{\nu,b}(\theta) - \nabla f(\theta)\|^2\right] \leq \frac{4(d+5)(\|\nabla f(\theta)\|^2 + \sigma^2)}{b} + \frac{3\nu^2 M^2 (d+3)^3}{2} \tag{8}$$

where $f_\nu(\theta) = \mathbf{E}_u[f(\theta + \nu u)]$.

*One-point versus two-point evaluation:* The gradient estimator in (6) is referred to as the two-point estimator in the literature. The reason is that, for a given random vector $\xi$, it is assumed that the stochastic function in (6) could be evaluated at two points,$F(\theta_1, \xi)$ and $F(\theta_2, \xi)$. Such an assumption is satisfied in several statistics, machine learning and simulation based optimization and sampling; see for example in [Spa05, MP07, Dip03, ADX10, DJWW15, GL13, NS17]. Yet another estimator is the one-point estimator which assumes that for each $\xi$, we observe only one noisy function evaluation $F(\theta, \xi)$. Admittedly, the one-point setting is more challenging than the two-point setting. Specifically, in the one-point feedback setting, Lemma 1.1 no longer holds. From a theoretical point of view, the use of two-point evaluation based gradient estimator is primarily motivated by the sub-optimality (in terms of oracle complexity) of one-point feedback based stochastic zeroth-order optimization methods either in terms of the approximation accuracy or dimension dependency.

The use of one-point feedback for stochastic zeroth-order gradient estimation could be traced back to [NY83]. Motivated by this, there has been several works in the machine learning community focusing on leveraging it for zeroth-order convex optimization. Specifically, considering the class of convex functions (without any further smoothness assumptions) and adversarial noise (i.e., roughly speaking, with noise vectors not necessarily assumed to be independent and identically distributed (i.i.d.)), [BLE17] proposed a polynomial-time algorithm and an oracle complexity of $\mathcal{O}(d^{21}/\epsilon^2)$. This was improved to $\mathcal{O}(d^5/\epsilon^2)$ recently in [Lat20]. Further assuming Lipschitz smooth convex functions, [BLNR15] and [GKL$^+$17], in the i.i.d noise case, obtained an oracle complexity of $\mathcal{O}(d^{7.5}/\epsilon^2)$ and $\mathcal{O}(d/\epsilon^3)$ respectively. The best known lower bound in this case is known to be $\mathcal{O}(d^2/\epsilon^2)$, which was established by [Sha13]. Further assuming $(\beta - 1)$ differentiable derivatives, for $\beta > 2$, [BP16] obtained as oracle complexity of $\mathcal{O}(d^2/\epsilon^{2\beta/(\beta-1)})$ and $\mathcal{O}(d^2/\epsilon^{(\beta+1)/(\beta-1)})$ respectively for the convex and strongly-convex setting, with i.i.d. noise case; see also [APT20]. In contrast to the above discussion, with two-point feedback it is possible to obtain much improved oracle complexities (i.e., linear in dimension and optimal in $\epsilon$) for stochastic zeroth-order optimization, as illustrated in [NS17, GL13, DJWW15, ADX10]. Given this subtle differences between the two-point and one-point evaluation settings for stochastic zeroth-order gradient estimation, in Section 4 we consider the effect of one-point gradient estimation technique for stochastic zeroth-order discretization of overdamped and underdamped Langevin diffusions.

### 1.3 Our Contributions

Under the availability of the stochastic zeroth-order oracles, we make the following contributions to the literature on sampling.

1. We first consider the case of strongly log-concave and smooth densities and analyze a stochastic zeroth-order version of Euler-discretization of overdamped and underdamped Langevin diffusions, under the two-point feedback setting in Section 2. For both cases, we characterize the oracle and iteration complexities to obtain $\epsilon$-approximate samples in term of $W_2$ metric.
2. We next consider in Section 2.3, a stochastic zeroth-order version of the recently proposed Randomized Midpoint Sampling method of the underdamped Langevin diffusion and characterize the oracle and iteration complexities to obtain $\epsilon$-approximate samples in term of $W_2$ metric. We show that for certain range of $\epsilon$, this method achieves improved oracle complexity compared to the above method.

3. While the above contributions are for strongly log-concave densities, in Section 3.1, we consider the more general class of densities satisfying log-Sobolev inequality and establish the oracle and iteration complexities of stochastic zeroth-order discretizations.
4. While all of the above contributions use the two-point stochastic zeroth-order feedback setting, in Section 4, we next consider the case of one-point feedback and characterize the corresponding oracle and iteration complexities for all the above discretizations.
5. Next, in Section 5, we consider variable selection for 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 zeroth-order gradient, which in conjunction with the above discretizations reduces the oracle and iteration complexities to be only poly-logarithmically dependent on the dimensionality $d$ thereby enabling high-dimensional sampling.

Our contributions provide several theoretical insights on the performance of stochastic zeroth-order sampling algorithms, and widen the applicability of theoretically sound Bayesian inference to various practical situations where we do not know the analytical form of the potential function. All proofs are relegated to the appendix.

## 2 Oracle Complexity Results under Strong Log-concavity

We now leverage the stochastic zeroth-order gradient estimation methodology introduced in Section 1.2 for discretizing underdamped and overdamped Langevin diffusions. Throughout this section, we assume the target density is strongly log-concave and smooth (recall Assumption 1.1).

### 2.1 Zeroth-Order Langevin Monte Carlo

Replacing the true gradient in the first-order Langevin Monte Carlo algorithm in (3), with the zeroth-order gradient estimation in (6), we obtain the following Zeroth-Order LMC (ZO-LMC) algorithm:

$$x_{n+1} = x_n - h\ g_{\nu,b}(x_n) + \sqrt{2h}\varepsilon_{n+1} \tag{9}$$

for $n = 0, 1, 2, \cdots, N-1$. Apart from the choice of step-size $h$, the ZO-LMC also requires two additional tuning parameters, the smoothing parameter $\nu$ and the batch-size $b$ of the zeroth-order gradient estimator $b$, that need to be set. Although ZO-LMC could be interpreted as a form of LMC with inaccurate gradient as in [DK19], the corresponding theoretical result from [DK19] cannot be used directly for obtaining the oracle complexity of ZO-LMC, as the variance of the gradient in (6) is not bounded unless we make restrictive assumptions on the true gradient of $f$. We now state the main result of this section, which describes the oracle complexity of ZO-LMC.

**Theorem 2.1** *Let the potential function $f$ satisfy Assumption 1.1. Then, for the ZO-LMC algorithm in* (9)*, under Assumption 1.3, by choosing*

$$h = \frac{\epsilon^2}{d^2}, \quad b = \max(1, \sigma^2)d, \quad \nu = \frac{\epsilon}{\sqrt{d}}, \tag{10}$$

*we have* $W_2(\varpi_N, \pi) \leq \epsilon$ *for* $0 < \epsilon \leq \min\left(d\sqrt{\frac{2}{M+m}}, \sqrt{\frac{m(d+5)}{8M^2}}\right)$*, after*

$$N = \mathcal{O}\left(\frac{d}{\epsilon^2} \cdot \log\left(\frac{d}{\epsilon}\right)\right). \tag{11}$$

*iterations. Hence, the total number of calls to the stochastic zeroth-order oracle is given by,*

$$Nb = \mathcal{O}\left(\frac{\max(1, \sigma^2) \cdot d^2}{\epsilon^2} \cdot \log\left(\frac{d}{\epsilon}\right)\right). \tag{12}$$

**Remark 2** *Recall that for the exact gradient based LMC algorithm, to obtain* $W_2(\varpi_N, \pi) \leq \epsilon$*, we require* $N = \mathcal{O}\left(d/\epsilon^2 \cdot \log(d/\epsilon)\right)$ *(see [DK19]) which matches* (11)*. Thus, ZO-LMC matches the performance of LMC in terms of iteration complexity required to obtain* $W_2(\varpi_N, \pi) \leq \epsilon$*. However, in each iteration of the LMC algorithm, we only require one gradient evaluation. Hence, the total number calls to the first-order oracle is also given by* $\mathcal{O}\left(d/\epsilon^2 \cdot \log(d/\epsilon)\right)$*. For the ZO-LMC, in contrast we require* $b = d$ *calls to the stochastic zeroth-order oracle in each iteration. Hence, the oracle complexity is given by* (12)*. By a straight forward modification of the proof of Theorem 2.1, for the ZO-LMC, if we restrict ourself to* $b = 1$*, the iteration complexity increases to* $N = \mathcal{O}\left(d^2/\epsilon^2 \cdot \log(d/\epsilon)\right)$*, which will then also be the oracle complexity. Thus, the price we pay to match LMC in the absence of true gradient information is* $O(d)$*.*

**Remark 3** *Recently [DCWY19] analyzed the standard Metropolis Random Walk algorithm (MRW), which is a zeroth-order algorithm, in the non-noise setting. Specifically, [DCWY19] showed that to achieve samples that are* $\epsilon$*-close to the target* $\pi$ *in total variation distance, MRW requires* $\mathcal{O}(d^2 \log(1/\epsilon)$ *calls to the non-noisy zeroth-order oracles. Considering the non-noisy setting, the result appears to seemingly have an exponential improvement in terms of* $\epsilon$*. However, this result was obtained under the so-called* warm start *condition on the distribution of the initial vector* $x_o$*, which seems to be an opaque condition hiding the true complexity of the problem. For example, it is not clear how to pick such a warm start distribution for a given target* $\pi$*, in particular in the stochastic zeroth-order setting that we consider in this work. As a way to potentially avoid this opaque warm start condition, [DCWY19] suggests to set* $x_0 \sim N(x^*, \mathbf{I}_d)$*, where* $x^*$ *is the unique minimizer of* $f(x)$ *and* $\mathbf{I}_d$ *is the* $d \times d$ *identity covariance matrix. For this choice of initial vector, to obtain a sample which is* $\epsilon$*-close to the target* $\pi$ *in total variation distance, [DCWY19] showed that MRW requires an oracle complexity of* $\mathcal{O}(d^2 \log(1/\epsilon)$*. However, in the zeroth-order setting, the oracle complexity of finding an* $\epsilon$*-minimizer of a strongly-convex and smooth function* $f(x)$*, is well-studied problem in stochastic optimization – it is upper and lower bounded by* $\mathcal{O}(d/\epsilon)$*; see for example [DJWW15, JNR12, NS17, GL13]. This seems to negate the actual oracle complexity improvements shown in [DCWY19], as it really seems to require extremely careful initial distributions (i.e., knowledge of the exact minimzer), even in the non-noisy setting. Notwithstanding the fact that the results in [DCWY19] are for the non-noisy setting, they are essentially no better than the oracle complexity results established for ZO-LMC algorithm in Theorem 2.1, which also has the advantage that it does not require any opaque warm start conditions or special initial distributions.*

## 2.2 Zeroth-Order Kinetic Langevin Monte Carlo

In the previous section, we consider the stochastic zeroth-order discretizations of the overdamped Langevin diffusions. It is known that in the first-order setting, discretizations of underdamped

Langevin diffusion obtain improved oracle complexities [DRD$^+$20, CCBJ18]. Under Langevin diffusion process (also called as kinetic Langevin diffusion process) is given by the following stochasic differential equation:

$$\begin{aligned} dV_T &= (\gamma V_T + \nabla f(L_T))\, dT + \sqrt{2\gamma} dW_T \\ dL_T &= V_T dT. \end{aligned} \tag{13}$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix. We refer the reader to [EGZ$^+$19, CCBJ18, DRD$^+$20] for more details about the above diffusion process and related theoretical results. Specifically, it was shown in [CCBJ18, DRD$^+$20] that first-order discretizations of the kinetic diffusion process (referred to as KLMC in [CCBJ18]) in (13) have better rates of convergence compared to similar first-order discretizations of the continous-process in (2). Specifically, recall that for the right choice of tuning parameters, LMC (i.e., first-order discretizations of (2)) requires that $N = \mathcal{O}(d/\epsilon^2 \cdot \log(d/\epsilon))$ for $W_2(\varpi_N, \pi) \leq \epsilon$. Whereas, it was shown in [CCBJ18, DRD$^+$20] $N = \mathcal{O}(\sqrt{d}/\epsilon \cdot \log(d/\epsilon))$ suffices ([DRD$^+$20] provides a much sharper result compared to [CCBJ18]). 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 (though the condition number of the sampling density defined as $M/m$), [DRD$^+$20] 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.

The Euler-discretization of the SDE in (13), which is a first-order sampling algorithm is given by the following iterations:

$$\begin{aligned} \tilde{x}_{n+1} &= \psi_0(h)\tilde{x}_n - \psi_1(h)\nabla f(x_n) + \sqrt{2\gamma}\tilde{\epsilon}_{n+1} \\ x_{n+1} &= x_n + \psi_1(h)\tilde{x}_n - \psi_2(h)\nabla f(x_n) + \sqrt{2\gamma}\epsilon_{n+1} \end{aligned} \tag{14}$$

where $(\tilde{\epsilon}_{n+1}, \epsilon_{n+1}) \in \mathbb{R}^{2d}$ 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 $\psi_{n+1} = \int_0^T \psi_n(s)ds$. We refer to this algorithm as KLMC following the terminology of [DRD$^+$20]. Based on this, we now consider the ZO-KLMC updates as:

$$\begin{aligned} \tilde{x}_{n+1} &= \psi_0(h)\tilde{x}_n - \psi_1(h) g_{\nu,b}(x_n) + \sqrt{2\gamma}\tilde{\epsilon}_{n+1} \\ x_{n+1} &= x_n + \psi_1(h)\tilde{x}_n - \psi_2(h) g_{\nu,b}(x_n) + \sqrt{2\gamma}\epsilon_{n+1} \end{aligned} \tag{15}$$

where $g_{\nu,b}$ is the zeroth-order gradient estimator as in (6). In comparison to the ZO-LMC algorithm, the ZO-KLMC algorithm has an additional tuning parameter $\gamma$ that needs to be set. For the ZO-KLMC algorithm, we have the following complexity result.

**Theorem 2.2** *Let the potential function $f$ satisfy Assumption 1.1. If the initial point $(\tilde{x}_0, x_0)$ is chosen such that $\tilde{x}_0 \sim N(0, \boldsymbol{I}_d)$, then, ensuring $\gamma \geq \sqrt{m+M}$, for the ZO-KLMC, under Assumption 1.3, by choosing,*

$$h = \frac{m\epsilon}{12\gamma M\sqrt{d}}, \quad \nu = \frac{\epsilon}{\sqrt{d}}, \quad b = \frac{d^{1.5}\max(1, \sigma^2)}{\epsilon}, \tag{16}$$

*we have $W_2(\varpi_N, \pi) \leq \epsilon$ for $0 < \epsilon \leq \frac{12M\gamma^2\sqrt{d}}{m^2}$ after*

$$N = \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\epsilon}\right) \tag{17}$$

*iterations. Here $\tilde{\mathcal{O}}$ hides poly-logarithmic factors in $1/\epsilon$. Hence, the total number of calls to the stochastic zeroth-order oracle is given by*

$$Nb = \tilde{\mathcal{O}}\left(\frac{d^2 \max(1, \sigma^2)}{\epsilon^2}\right). \tag{18}$$

**Remark 4** *We note that compared to ZO-LMC, while ZO-KLMC obtains improved iteration complexity, the iteration complexity still remains the same. The improvement in the iteration complexity is indeed a consequence of a similar improvement in the first-order setting as demonstrated in [CCBJ18, DRD$^+$20].*

### 2.3 Zeroth Order Randomized Midpoint Method

Given that the ZO-KLMC offers no improvement over ZO-LMC in terms of oracle complexity despite its improved iteration complexity, it is worth examining if there are other discretizations that obtain improvements in oracle complexities. Towards that, in this section we analyze the zeroth-order version of the Randomized Mid-Point discretization of the underdamped Langevin diffusion, proposed in [SL19]. In the first-order setting, [CLW20] recently showed that the Randomized Mid-Point discretization of underdamped Langevin diffusion achieves the information theoretic lower bounds for sampling. See also [HBE20] for additional probabilistic results.

The crux of the randomized midpoint method is based on first representing the kinetic Langevin Monte Carlo in (13) in its integral format, and estimating the integrals based on a randomization technique. We also mention in passing that the randomized midpoint idea shares some similarities to symplectic integration methods [SS92] from the sampling literature and extragradient method [Kor76] from the optimization literature, with the main difference being the randomized choice of step-size which leads to improved oracle complexities. We now provide the algorithm in the zeroth-order setting and the corresponding theoretical result. Let $\epsilon_n^{(i)} \in \mathbb{R}^d$, $i = 1, 2, 3$, be a sequence of Gaussian random vectors generated according to the procedure described in Appendix A of [SL19]. Let $\alpha_n$ be a sequence of uniform random variables supported on the interval $[0, 1]$. Then the zeroth-order Randomized Mid-Point Method (ZO-RMP) is given by the following updates:

$$x_{n+\frac{1}{2}} = x_n + \frac{1 - e^{-2\alpha_n h}}{2} v_n - \frac{u}{2}\left(\alpha_n h - \frac{1 - e^{-2(\alpha_n h)}}{2}\right) g_{\nu,b}(x_n) + \sqrt{u}\epsilon_{n+1}^{(1)} \tag{19}$$

$$x_{n+1} = x_n + \frac{1 - e^{-2h}}{2} v_n - \frac{uh}{2}(1 - e^{-2(h-\alpha_n h)}) g_{\nu,b}\left(x_{n+\frac{1}{2}}\right) + \sqrt{u}\epsilon_{n+1}^{(2)} \tag{20}$$

$$v_{n+1} = v_n e^{-2h} - uhe^{-2(h-\alpha_n h)} g_{\nu,b}\left(x_{n+\frac{1}{2}}\right) + 2\sqrt{u}\epsilon_{n+1}^{(3)}. \tag{21}$$

We remark that we use the same choice of batch-size, $b$, in (19), (20) and (21), as using different batch sizes has no effect on the oracle complexity.

**Theorem 2.3** *Define $\kappa = M/m$ to be the condition number of the potential $f$ which satisfies Assumption 1.1. Furthermore, let the stochastic zeroth-order oracle satisfy Assumption 1.3. Let $x^*$ be the minimizer of $f$, and $x_0$ be such that $\mathbf{E}\left[f(x_0) - f(x^*)\right] = O(d)$, and $v_0 = 0$. Then, for*

$0 \leq \epsilon \leq 1$, *by choosing,*

$$h = C \min\left( \frac{(\epsilon\sqrt{m})^{\frac{1}{3}}}{(d\kappa)^{\frac{1}{6}} \log\left(\frac{1}{\epsilon}\right)^{\frac{1}{6}}}, \min\left( \left(\frac{m}{d}\right)^{\frac{1}{3}}, \left(\frac{Mm}{16\sigma^2}\right)^{\frac{1}{3}}, \sqrt{m} \right) \epsilon^{\frac{2}{3}} \log\left(\frac{1}{\epsilon}\right)^{-\frac{2}{3}} \right) \quad b = \frac{d\kappa}{h^3} \quad \nu = \frac{uh^2}{d^{1.5}} \tag{22}$$

*for the ZO-RMP method described in* (19)-(21)*, with* $u = 1/M$*, we have* $W_2(\varpi_N, \pi) \leq \epsilon$ *after*

$$N = \tilde{O}\left( \max\left( \frac{d^{\frac{1}{6}}\kappa^{\frac{7}{6}}}{(\epsilon\sqrt{m})^{\frac{1}{3}}}, \frac{\kappa \max\left( \left(\frac{d}{m}\right)^{\frac{1}{3}}, \left(\frac{\sigma^2}{Mm}\right)^{\frac{1}{3}}, \frac{1}{\sqrt{m}} \right)}{\epsilon^{\frac{2}{3}}} \right) \right) \tag{23}$$

*iterations. Hence, the total-number of zeroth-order oracle calls are given by*

$$2Nb = \tilde{O}\left( \max\left( \frac{d^{\frac{5}{3}}\kappa^{\frac{8}{3}}}{\epsilon^{\frac{4}{3}}}, \frac{d\kappa^2 \max\left( \left(\frac{d}{m}\right)^{\frac{1}{3}}, \left(\frac{\sigma^2}{Mm}\right)^{\frac{1}{3}}, \frac{1}{\sqrt{m}} \right)^4}{\epsilon^{\frac{8}{3}}} \right) \right). \tag{24}$$

**Remark 5** *The analysis of the randomized midpoint algorithm in [SL19], for the first-order setting, requires access to exact minimizer* $x^*$ *as the initializer. We relax this requirement to the having a point* $x_0$ *satisfying* $\mathbf{E}\left[f(x_0) - f(x^*)\right] = O(d)$*, which is a milder requirement. It is well-known from the stochastic optimization literature, that under Assumption 1.3, and 1.1, in the zeroth-order setting, using the zeroth-order version of stochastic gradient algorith, the oracle complexity of finding a point* $x_0$ *such that* $\mathbf{E}\left[f(x_0) - f(\bar{x})\right] = O(d)$ *where* $\bar{x}$ *is the minimizer of* $f$*, is* $O(\kappa \log d)$ *[DJWW15, NS17].*

**Remark 6** *Note that even though the iteration complexity of ZO-RMP still matches with RMP (except for the dimension dependence which is unavoidable in the zeroth-order setting), and is better than KLMC for all values of* $\epsilon$*, the oracle complexity for ZO-RMP is not uniformly better than ZO-KLMC for all* $\epsilon$*. However, observe that when* $h = C\frac{(\epsilon\sqrt{m})^{\frac{1}{3}}}{(d\kappa)^{\frac{1}{6}} \log\left(\frac{1}{\epsilon}\right)^{\frac{1}{6}}}$*, i.e., when* $\epsilon \geq \max\left( \sqrt{\frac{d}{M}}, \frac{16\sigma^2}{M^{\frac{3}{2}}\sqrt{d}}, \frac{1}{\sqrt{dmM}} \right)$ *the oracle complexity of ZO-RMP is* $\tilde{O}\left( \frac{d^{\frac{5}{3}}\kappa^{\frac{8}{3}}}{\epsilon^{\frac{4}{3}}} \right)$ *which is indeed better compared to* $\tilde{O}\left(\frac{d^2}{\epsilon^2}\right)$ *for ZO-KLMC.*

We end this section by mentioning that developing lower bounds on the oracle complexity of sampling from strongly log-concave densities, in the stochastic zeroth-order setting that we consider is an interesting open problem.

## 3 Oracle Complexity Results under Log-Sobolev Inequality

The algorithms and oracle complexity results in the previous sections were stated for smooth and strongly log-concave densities (i.e., under Assumption 1.1), which covers important classes

of problems in sampling and Bayesian inference. However, the fundamental idea behind the non-asymptotic convergence results of the discretization based sampling algorithm are essentially based on the following facts: (i) the underlying continuous (underdamped or overdamped) Langevin diffusion converges to its equilibrium state (i.e., to the target distribution $\pi$ in this case) exponentially fast in various metrics, and (ii) consequently, the potential function is smooth enough that the error due to discretization is not extremely large. Roughly speaking, condition **A1** and **A2** in Assumption 1.1 corresponds respectively to the above facts, respectively. However, it is well-know that the overdamped Langevin diffusion converges to its equilibrium under much weaker conditions that strong log-concavity; indeed as long as the target density satisfies functional inequalities like Poincare or Log-Sobolev inequalities, the overdamped Langevin diffusion converges to its equilibrium exponentially faster in various metrics; see, for example [BGL13]. Motivated by the above fact, recently [VW19] demonstrated that the LMC algorithm also exhibits rapid convergence to the target density if it has access to the exact gradient evaluations of the potential function $f$. As a consequence, one could sample from densities that are not essentially strongly log-concave, there by extending the applicability of LMC algorithms for a wider class of Bayesian inference problems. In this section, we analyze the performance of stochastic zeroth-order discretization of overdamped Langevin diffusions when the target density satisfies log-Sobolev inequality.

**Assumption 3.1** *A density $\pi$ is said to satisfy Log-Sobolev Inequality (LSI) with a constant $\lambda > 0$ if for all smooth function $g : \mathbb{R}^d \to \mathbb{R}$ with finite variance,*

$$\int_{\mathbb{R}^d} g^2(\theta) \log g^2(\theta)\pi(\theta)d\theta - \int_{\mathbb{R}^d} g^2(\theta)\pi(\theta)d\theta \log \int_{\mathbb{R}^d} g^2(\theta)\pi(\theta)d\theta \leq \frac{2}{\lambda}\int_{\mathbb{R}^d} \|\nabla g(\theta)\|^2\pi(\theta)d\theta. \quad (25)$$

In Section **??**, we show that mixture of Gaussian densities with unequal covariance satisfies the above assumption, while it does not satisfy condition **A1** of Assumption 1.1, and discuss applications to Bayesian variable selection. The above assumption also leads to the following equivalent formulation. Let $H_\pi(\varpi)$, and $J_\pi(\varpi)$ be the Kullback-Leibler (KL) divergence of $\varpi$ with respect to $\pi$, and the relative Fisher Information respectively which are defined as follows:

$$H_\pi(\varpi) = \int_{\mathbb{R}^d} \varpi(\theta) \log \frac{\varpi(\theta)}{\pi(\theta)} d\theta \quad J_\pi(\varpi) = \int_{\mathbb{R}^d} \varpi(\theta) \left\| \nabla \log \frac{\varpi(\theta)}{\pi(\theta)} \right\|^2 d\theta. \quad (26)$$

One can verify that LSI is equivalent to the following condition by plugging $g^2 = \varpi/\pi$ in (25):

$$H_\pi(\varpi) \leq \frac{1}{2\lambda} J_\pi(\varpi). \quad (27)$$

We also know that when $\pi$ satisfies LSI, Talagrand inequality holds [BGL13], i.e., for all $\varpi$,

$$\frac{\lambda}{2} W_2(\varpi, \pi)^2 \leq H_\pi(\varpi). \quad (28)$$

With this background, we provide our oracle complexity result of ZO-LMC algorithm when the density satisfies LSI and is smooth.

**Theorem 3.2** *Let the target density $\pi$ satisfy Assumption 3.1 and let the potential function $f$ be satisfy condition **A2** in Assumption 1.1. Let $x_0 \sim \varpi_0$ which satisfies $H_\pi(\varpi_0) \leq \infty$. Then for the ZO-LMC update as in* (9)*, under Assumption 1.3, by choosing,*

$$b = \frac{384M^2(d+5)\max(1,\sigma^2)}{h\lambda^2}, \quad \nu = \frac{\sqrt{h}}{d+3}, \quad h = \frac{\epsilon^2}{d}, \quad (29)$$

*we have* $W_2(\varpi_N, \pi) \leq \epsilon$, *for all* $0 \leq \epsilon \leq \frac{\alpha}{4L^2}$, *after* $N$ *iterations where*

$$N = \tilde{O}\left(\frac{d}{\epsilon^2}\right). \tag{30}$$

*Hence, the total number of calls to the stochastic zeroth-order oracle is given by*

$$Nb = \tilde{O}\left(\frac{\max(1, \sigma^2)\ d^3}{\epsilon^4}\right) \tag{31}$$

**Remark 7** *Note that in comparison to condition* ***A1*** *of Assumption 1.1, the assumptions required for the above theorem are weaker. Specifically, in place of condition* ***A1*** *in Assumption 1.1, we have Assumption 3.1. Condition* ***A2*** *is regarding the smoothness is required to handle error that arises due to discretization of continuous time dynamics. For this wider class of densities, the price to pay is that the dependency on both the dimension* $d$ *and* $\epsilon$ *increases in comparison to Theorem 2.1.*

**Remark 8** *Given that ZO-LMC exhibits convergence (albeit with a slightly weaker* $\epsilon$ *and* $d$ *dependency, it is natural to ask if ZO-KLMC also exhibits similar convergence. However, even in the first-order setting this question is open. Indeed, kinetic Langevin diffusions are a class of degenerate diffusions which require a different class of function inequalities (called as hypocoercivity [Vil09]) for them to converge to their equilibrium. It is an open question to show that the discretize sampling algorithm (KLMC or appropriate modifications) also convergence under hypocoercivity and appropriate smoothness assumptions on the potential function* $f$*, either given access to exact first-order oracles or stochastic zeroth-order oracles. We leave this question as future work.*

# 4 One-Point Setting: Independent noise per function evaluation

As discussed in Section 1.2, there are subtle differences between the availability of one and two-point evaluation based stochastic zeroth-order gradients. In this section, we examine this difference in more detail. Recall that while defining the zeroth-order gradient estimator in (6), we assumed that the function can be evaluated at two points, namely, $\theta + \nu u_i$, and $\theta$, with the same noise $\xi_i$. This implies, when the noise is additive, i.e., $F(\theta, \xi) = f(\theta) + \xi$, the gradient estimator is not affected by the noise. Because in that case we have, $F(\theta + \nu u_i, \xi_i) - F(\theta, \xi_i) = f(\theta + \nu u_i) - f(\theta)$. We emphasize that this is our main reason for consider general non-additive noise in the previous sections. For example, under multiplicative noise, consider the case where $F(\theta, \xi) = \xi f(\theta)$, $\mathbf{E}[\xi] = 1$, and $f(\theta)$ is $L$-Lipschitz continuous; then Assumption 1.3 holds.

Now we will examine the one-point setting in that the noise in the two function evaluations of the gradient estimator is not the same. Specifically, first we show that allowing the noise $\xi_i$, and $\xi_i'$ in $F(\theta + \nu u_i, \xi_i)$, and $F(\theta, \xi_i')$ to be independent additive noise, deteriorates the iteration and/or oracle complexities of zeroth-order discretizations considered in the previous settings. Formally, we work under the following assumption in the one-point stochastic zeroth-order setting.

**Assumption 4.1** *The stochastic zeroth-order oracle is such that for each point* $x$*, the observed function evaluation* $F(\theta, \xi)$ *is given by* $F(\theta, \xi) = f(\theta) + \xi$ *where* $\mathbf{E}[\xi] = 0$*, and* $\mathbf{E}\left[\xi^2\right] = \sigma^2$.

Under Assumption 4.1, the upper bound on the variance of the gradient estimator as stated in Lemma 1.1 no longer holds. Instead, we have the following result.

**Lemma 4.1** *Let $g_{\nu,b}(\theta)$ in (6), be defined under the one-point setting. Then under Assumption 4.1 and condition A1 of Assumption 1.1, we have*

$$\mathbf{E}\left[\|g_{\nu,b}(\theta)-\nabla f_\nu(\theta)\|^2\right] \leq \frac{2(d+5)\|\nabla f(\theta)\|^2}{b}+\frac{\nu^2M^2(d+3)^3}{2b}+\frac{2d\sigma^2}{b\nu^2},$$
$$\mathbf{E}\left[\|g_{\nu,b}(\theta)-\nabla f(\theta)\|^2\right] \leq \frac{4(d+5)\left(\|\nabla f(\theta)\|^2+\frac{\sigma^2}{\nu^2}\right)}{b}+\frac{3\nu^2M^2(d+3)^3}{2}+\frac{4d\sigma^2}{b\nu^2}.$$

The main difference in the one-point setting, in terms of the variance of the gradient estimator is the presence of the third term, which is of the order of $1/b\nu^2$. This causes the additional difficulties in terms of setting the parameters $b$ and $\nu$ in the zeroth-order gradient estimator, which in turn causes the oracle complexities to deteriorate. Based on the above result on the variance, we provide the oracle complexity results for ZO-LMC, ZO-KLMC and ZO-RMP under Assumption 4.1 on the stochastic zeroth-order oracle, in Theorem 4.2, 4.3 and 4.4 respectively.

**Theorem 4.2 (ZO-LMC under Strong Log-concavity)** *Let the potential function $f$ satisfy Assumption 1.1. Then, for ZO-LMC algorithms under Assumption 4.1, by choosing*

$$h=\frac{\epsilon^2}{d^2}, \quad b=\frac{\max(1,\sigma^2)\cdot d}{\epsilon^2}, \quad \nu=\frac{\epsilon}{\sqrt{d}}, \tag{32}$$

*we have $W_2(\varpi_N,\pi)\leq\epsilon$ for $0<\epsilon\leq\min\left(d\sqrt{\frac{2}{M+m}},\sqrt{\frac{m(d+5)}{8M^2}}\right)$, after $N$ iterations, where*

$$N=O\left(\frac{d}{\epsilon^2}\log\left(\frac{d}{\epsilon}\right)\right). \tag{33}$$

*Hence, the total number of calls to the stochastic zeroth-order oracle is given by,*

$$Nb=O\left(\frac{\max(1,\sigma^2)\ d^2}{\epsilon^4}\log\left(\frac{d}{\epsilon}\right)\right). \tag{34}$$

**Theorem 4.3 (ZO-KLMC under Strong Log-concavity)** *Let the function $f$ satisfy Assumption 1.1. If the initial point $(\tilde{x}_0,x_0)$ is chosen such that $\tilde{x}_0\sim N(0,\boldsymbol{I}_d)$, then, ensuring $\gamma\geq\sqrt{m+M}$, under Assumption 4.1 for the ZO-KLMC, by choosing,*

$$h=\frac{m\epsilon}{12\gamma M\sqrt{d}} \quad \nu=\frac{\epsilon}{\sqrt{d}} \quad b=\frac{d^{1.5}\max(1,\sigma^2)}{\epsilon^3} \tag{35}$$

*we have $W_2(\varpi_N,\pi)\leq\epsilon$ for $0<\epsilon\leq\frac{12M\gamma^2\sqrt{d}}{m^2}$, after*

$$N=\tilde{O}\left(\frac{\sqrt{d}}{\epsilon}\right) \tag{36}$$

*iterations. Hence, the total number of oracle calls to the stochastic zeroth-order oracle is given by*

$$Nb=O\left(\frac{\max(1,\sigma^2)\ d^2}{\epsilon^4}\right). \tag{37}$$

**Theorem 4.4 (ZO-RMP under Strong Log-concavity)** *Let the potential function satisfy Assumption 1.1 and let $x^*$ be the minimizer of $f$, $x_0$ be such that $\mathbf{E}\left[f(x_0) - f(\bar{x})\right] = O(d)$, and $v_0 = 0$. Let the stochastic zeroth-order oracle satisfy Assumption 4.1. Then, for $0 \leq \epsilon \leq 1$, by choosing,*

$$h = C \min\left(\frac{(\epsilon\sqrt{m})^{\frac{1}{3}}}{(d\kappa)^{\frac{1}{6}}\log\left(\frac{1}{\epsilon}\right)^{\frac{1}{6}}}, \min\left(\left(\frac{m}{d}\right)^{\frac{1}{3}}, \left(\frac{Mm}{16\sigma^2}\right)^{\frac{1}{3}}, \sqrt{m}\right)\epsilon^{\frac{2}{3}}\log\left(\frac{1}{\epsilon}\right)^{-\frac{2}{3}}\right) \quad b = \frac{d^4\kappa}{h^7} \quad \nu = \frac{uh^2}{d^{1.5}} \tag{38}$$

*for the ZO-RMP described in (19)-(21), we have $W_2(\varpi_N, \pi) \leq \epsilon$ after*

$$N = \tilde{O}\left(\max\left(\frac{d^{\frac{1}{6}}\kappa^{\frac{7}{6}}}{(\epsilon\sqrt{m})^{\frac{1}{3}}}, \frac{\kappa\max\left(\left(\frac{d}{m}\right)^{\frac{1}{3}}, \left(\frac{\sigma^2}{Mm}\right)^{\frac{1}{3}}, \frac{1}{\sqrt{m}}\right)}{\epsilon^{\frac{2}{3}}}\right)\right) \tag{39}$$

*iterations. Hence, the total-number of zeroth-order oracle calls are given by*

$$2Nb = \tilde{O}\left(\max\left(\frac{d^{\frac{16}{3}}\kappa^{\frac{10}{3}}}{\epsilon^{\frac{8}{3}}}, \frac{d^4\kappa^2\max\left(\left(\frac{d}{m}\right)^{\frac{1}{3}}, \left(\frac{\sigma^2}{Mm}\right)^{\frac{1}{3}}, \frac{1}{\sqrt{m}}\right)^8}{\epsilon^{\frac{16}{3}}}\right)\right). \tag{40}$$

**Remark 9** *As before, the oracle complexity of ZO-RMP in this setting is not uniformly better than that of ZO-KLMC. We do observe that when $h = C\frac{(\epsilon\sqrt{m})^{\frac{1}{3}}}{(d\kappa)^{\frac{1}{6}}\log\left(\frac{1}{\epsilon}\right)^{\frac{1}{6}}}$, i.e., when $\epsilon \geq \max\left(\sqrt{\frac{d}{M}}, \frac{16\sigma^2}{M^{\frac{3}{2}}\sqrt{d}}, \frac{1}{\sqrt{dmM}}\right)$ the oracle complexity of ZO-RMP is $\tilde{O}\left(\frac{d^{\frac{16}{3}}\kappa^{\frac{10}{3}}}{\epsilon^{\frac{8}{3}}}\right)$ which is worse compared to $\tilde{O}\left(\frac{d^2}{\epsilon^4}\right)$ for ZO-KLMC. However, it is better than that of ZO-KLMC in the opposite regime.*

We now present the corresponding result when the target density is not strongly log-concave but satisfies LSI.

**Theorem 4.5 (ZO-LMC under Log-Sobolev Inequality)** *Let the target density $\pi$ satisfy Assumption 3.1 and let the potential function $f$ satisfy condition **A2** of Assumption 1.1. Let $x_0 \sim \varpi_0(x)$ which satisfies $H_\pi(\varpi_0) \leq \infty$. Then for the ZO-LMC update as in (9), under Assumption 4.1, by choosing,*

$$b = \frac{384M^2(d+5)\max(1,\sigma^2)}{h^2\lambda^2}, \quad \nu = \frac{\sqrt{h}}{d+3}, \quad h = \frac{\epsilon^2}{d}, \tag{41}$$

*we have, $W_2(\varpi_N, \pi) \leq \epsilon$, for all $0 \leq \epsilon \leq \frac{\lambda}{4L^2}$ after*

$$N = \tilde{O}\left(\frac{d}{\epsilon^2}\right) \tag{42}$$

*iterations. Hence, the total number of calls to the zeroth-order oracle is given by*

$$Nb = \tilde{O}\left(\frac{d^4}{\epsilon^6}\right). \tag{43}$$

**Remark 10 (*One-point setting with non-additive noise* )** *Given the above result, it is natural to examine the effect of non-additive noise on the oracle complexities. For this case, we have the following result under an additional smoothness assumption on the stochastic function evaluations $F(x, \xi)$.*

***Lemma 4.2*** *Let the function $F(\theta, \xi)$ be Lipschitz continuous in its second argument, i.e., $|F(\theta, \xi) - F(\theta, \xi')| \leq L|\xi - \xi'|$. Under the above condition, Lemma 4.1 holds. Consequently, all the above complexity results in this section holds.*

**Remark 11 (*Effect of Higher-order smoothness*)** *While the oracle complexities under the one-point evaluation setting are worse than that of the two-point setting, they could be made to approach that of the two-point setting when we make the stronger assumption that the potential function is assumed to the $\beta$-times differentiable and the $(\beta - 1)$-derivatives are Lipschitz continuous. Similar phenomenon has been observed in the case of highly-smooth convex stochastic zeroth-order optimization; see, for example [BP16, APT20]. As the precise statements and the proofs are similar to that of the above theorems, we omit the details.*

# 5 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. Throughout this section, we assume one could observe exact function evaluations, without any noise. We emphasize that we make this assumption purely for technical convenience and to convey the theoretical results insightfully; all results presented in this section extends to the noisy setting in a straightforward manner. Specifically, we make the following assumption on the structure of $f$.

**Assumption 5.1** *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 $\nabla f_\nu(\theta) = \mathbf{E}_u [\nabla f(\theta + \nu u)]$ 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 [BNST18] 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 5.2** *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 5.2) 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 any of the zeroth-order sampling algorithm 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 5.3** *Let $f$ satisfy Assumption 1.1 and the query point selected satisfy Assumption 5.2. Set $\tau = (a - M\nu\sqrt{s})/2$ and assume that $\nu \leq \min\left(\frac{a}{2M\sqrt{s}}, \frac{R}{MC_2\sqrt{s}}\right)$ and*

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

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

**Remark 12** *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 2.1 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 [BNST18]. Furthermore, exploring appropriately defined notions of non-homogenous sparsity assumptions is also challenging.*

# 6 Discussion

In this work, we proposed and analyzed zeroth-order discretizations of overdamped or underdamped Langevin diffusions. We provide a through analysis of the oracle complexity of such sampling algorithms under various noise models on the zeroth-order oracle and provide simulation results corroborating the theory. Recall that our zeroth-order gradient estimators used in this work 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 [BDMP17, BEL18] and in computing volume of convex body [BGVV14], 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, [GM15].

**Proposition 6.1** *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: $\int_{u \in \mathcal{M}} \nabla_u \left(g(u)p(u)\right) du = 0$. 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 2 for with Gaussian random variables. A special case of the above idea, when the space $\mathcal{M}$ is a Riemannian sub-manifold embedded in an Euclidean space was considered in [LBM20] in context of stochastic zeroth-order Riemannian optimization. We postpone a rigorous analysis of the estimation and approximation rates in the general setting, and their applications to black-box sampling on non-Euclidean spaces for future work.

# References


[ADL16] Johan Alenlöv, Arnaud Doucet, and Fredrik Lindsten. Pseudo-marginal Hamiltonian Monte Carlo. *arXiv preprint arXiv:1607.02516*, 2016.

[ADX10] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory*, pages 28–40, 2010.

[AH17] Charles Audet and Warren Hare. Derivative-free and blackbox optimization. 2017.

[APT20] Arya Akhavan, Massimiliano Pontil, and Alexandre Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

[AR09] Christophe Andrieu and Gareth O Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

[BBKG18] 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.

[BDMP17] Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In *Conference on Learning Theory*, 2017.

[Bea03] Mark A Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

[BEL18] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with Projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 2018.

[BFY18] Krishnakumar Balasubramanian, Jianqing Fan, and Zhuoran Yang. Tensor methods for additive index models under discordance and heterogeneity. *arXiv preprint arXiv:1807.06693*, 2018.

[BG19] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order Nonconvex Stochastic Optimization: Handling Constraints, High-Dimensionality and Saddle-Points. *arXiv preprint arXiv:1809.06474v2*, 2019.

[BGL13] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

[BGVV14] Silouanos Brazitikos, Apostolos Giannopoulos, Petros Valettas, and Beatrice-Helen Vritsiou. *Geometry of isotropic convex bodies*, volume 196. American Mathematical Soc., 2014.

[BLE17] Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017.

[BLNR15] Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Conference on Learning Theory*, pages 240–265, 2015.

[Blu54] Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

[BNST18] 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.

[BP16] Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pages 257–283, 2016.

[BRS93] 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.

[CCAY+18] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting. *arXiv preprint arXiv:1805.01648*, 2018.

[CCBJ18] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR, 2018.

[CLW20] Yu Cao, Jianfeng Lu, and Lihan Wang. Complexity of randomized algorithms for underdamped Langevin dynamics. *arXiv preprint arXiv:2003.09906*, 2020.

[CSV09] Andrew Conn, Katya Scheinberg, and Luis Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

[Dal17] 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.

[DCWY19] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast. *Journal of Machine Learning Research*, 20:1–42, 2019.

[Dip03] Jürgen Dippon. Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260–1281, 2003.

[DJ20] David B Dunson and JE Johndrow. The hastings algorithm at fifty. *Biometrika*, 107(1):1–23, 2020.

[DJWW15] John Duchi, Michael Jordan, Martin Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[DK19] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

[DM17] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[DM+19] Alain Durmus, Eric Moulines, et al. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[DRD+20] Arnak S Dalalyan, Lionel Riou-Durand, et al. On sampling from a log-concave density using kinetic langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

[EGZ+19] Andreas Eberle, Arnaud Guillin, Raphael Zimmer, et al. Couplings and quantitative contraction rates for langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

[GC11] 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.

[GKL+17] Alexander V Gasnikov, Ekaterina A Krymova, Anastasia A Lagunovskaya, Ilnura N Usmanova, and Fedor A Fedorenko. Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case. *Automation and remote control*, 78(2):224–234, 2017.

[GL13] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[GM15] Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

[GW08] Andreas Griewank and Andrea Walther. *Evaluating derivatives: Principles and techniques of algorithmic differentiation*. SIAM, 2008.

[GW11] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface focus*, 1(6):807–820, 2011.

[HBE20] Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method. *Advances in Neural Information Processing Systems*, 33, 2020.

[JNR12] Kevin G Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. *Advances in Neural Information Processing Systems*, 25:2672–2680, 2012.

[KDV12] Jonas Knape and Perry De Valpine. Fitting complex population models by combining particle filters with Markov chain Monte Carlo. *Ecology*, 93(2):256–263, 2012.

[KLS95] 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.

[Kor76] GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[KTR+17] 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.

[KW52] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[Lat20] Tor Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *arXiv preprint arXiv:2006.00475*, 2020.

[LBM20] Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Zeroth-order optimization on riemannian manifolds. *arXiv preprint arXiv:2003.11238*, 2020.

[LMW19] Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[LS90] 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.

[LV07] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[MFR20] Gael M Martin, David T Frazier, and Christian P Robert. Computing bayes: Bayesian computation from 1763 to the 21st century. *arXiv preprint arXiv:2004.06425*, 2020.

[MP07] Abdelkader Mokkadem and Mariane Pelletier. A companion for the kiefer–wolfowitz–blum stochastic approximation algorithm. *The Annals of Statistics*, 35(4):1749–1772, 2007.

[MT96] 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.

[Nea11] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[NS17] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[NY83] A. S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, Philadelphia, 1983.

[RGB14] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[RR98] 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.

[RT96] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[RTCB15] Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771, 2015.

[Sha13] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2100–2111, 2019.

[Spa05] James Spall. *Introduction to stochastic search and optimization: Estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[SS92] Jesus M Sanz-Serna. Symplectic integrators for hamiltonian problems: an overview. *Acta numerica*, 1(243-286):123–124, 1992.

[ST99a] 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.

[ST99b] 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.

[Ste72] 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.

[STRR15] Chris Sherlock, Alexandre H Thiery, Gareth O Roberts, and Jeffrey S Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics*, 43(1):238–275, 2015.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[Vil09] Cedric Villani. *Hypocoercivity*. Number 949-951. American Mathematical Soc., 2009.

[VW19] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pages 8094–8106, 2019.

**Stochastic Zeroth-order Discretizations of Langevin Diffusions for Bayesian Inference: Supplementary Material**

# 7 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.

# 8 Proofs for Section 2

## 8.1 Proofs for Oracle Complexity of ZO-LMC

*Proof.* [of Theorem 2.1] The proof follows by first calculating the bias and variance of the gradient estimator in our zeroth-order setting, where the error term $\zeta_n = g_{\nu,b}(x_n) - \nabla f(x_n)$. 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 $b = 1$, we have the following calculation for the bias.

$$\|\mathbf{E}[\zeta_n \mid x_n]\|^2 = \|\mathbf{E}[\nabla f(x_n+\nu u) \mid x_n] - \nabla f(x_n)\|^2 \le \mathbf{E}[(M\nu\|u\|)^2] \le M^2\nu^2 d. \tag{44}$$

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

$$\|\mathbf{E}[\zeta_n \mid x_n]\|^2 = \|\mathbf{E}[g_{\nu,b}(x_n) - \nabla f(x_n) \mid x_n]\|^2 \le \|\mathbf{E}[g_{\nu,1}(x_n) - \nabla f(x_n) \mid x_n]\|^2 \le M^2\nu^2 d.$$

From Lemma 2.1 of [BG19], we have,

$$\mathbf{E}[\|\zeta_n - \mathbf{E}[\zeta_n \mid x_n]\|^2] \le \frac{\nu^2}{2b}M^2(d+3)^3 + \frac{2(d+5)\left(\sigma^2 + \|\nabla f(x_n)\|_{L_2}^2\right)}{b}.$$

Next, we follow a similar framework to the proof of Theorem 4 in [DK19], but with modifications to adapt to the variance that is not uniformly bounded. Recall that $\Delta_n = L_0 - x_n$, $\Delta_{t+1} = L_h - x_{t+1}$, where $L_n = L_0 - \int_0^T \nabla f(L_s)ds + \sqrt{2}W_n$ follows the Langevin diffusion with stationary distribution $\pi$. Moreover, $\|\Delta_n - hU\| = \|\Delta_n - h[\nabla f(x_n+\Delta_n) - \nabla f(x_n)]\| \le (1-mh)\|\Delta_n\|$, $\|V\| = \|\int_0^h [\nabla f(L_s) - \nabla f(L_0)]ds\| \le 1.65M(h^3 d)^{1/2}$. Thus,

$$\begin{aligned}
\|\Delta_{n+1}\|_{L_2} &= \|\Delta_n - hU - V + h\zeta_n\|_{L_2} \\
&\le \{\|\Delta_n - hU\|_{L_2}^2 + h^2\|\zeta_n - \mathbf{E}[\zeta_n \mid x_n]\|_{L_2}^2\}^{1/2} + \|V\|_{L_2} + h\|\mathbf{E}[\zeta_n \mid x_n]\|_{L_2} \\
&\le \left\{(1-mh)^2\|\Delta_n\|_{L_2}^2 + h^2\left(\frac{\nu^2}{2b}M^2(d+3)^3 + \frac{2(d+5)\left(\sigma^2 + \|\nabla f(x_n)\|_{L_2}^2\right)}{b}\right)\right\}^{1/2} \\
&\quad + 1.65M(h^3 d)^{1/2} + M\nu h d^{1/2} \\
&\le \left\{(1-mh)^2\|\Delta_n\|_{L_2}^2 + h^2\left(\frac{\nu^2}{2b}M^2(d+3)^3 + \frac{2(d+5)\left(\sigma^2 + 2M^2\|\Delta_n\|_{L_2}^2 + 2\|\nabla f(L_0)\|_{L_2}^2\right)}{b}\right)\right\}^{1/2} \\
&\quad + 1.65M(h^3 d)^{1/2} + M\nu h d^{1/2} \\
&\le \left\{(1-mh)^2\|\Delta_n\|_{L_2}^2 + h^2\left(\frac{\nu^2}{2b}M^2(d+3)^3 + \frac{2(d+5)\left(\sigma^2 + 2Md\right)}{b}\right)\right\}^{1/2} \\
&\quad + \frac{4M^2h^2(d+5)}{b(1-mh)}\|\Delta_n\|_{L_2} + 1.65M(h^3 d)^{1/2} + M\nu h d^{1/2} \\
&\le \left\{(1-mh)^2\|\Delta_n\|_{L_2}^2 + h^2\left(\frac{\nu^2}{2b}M^2(d+3)^3 + \frac{2(d+5)\left(\sigma^2 + 2Md\right)}{b}\right)\right\}^{1/2} \\
&\quad + \frac{mh}{2}\|\Delta_n\|_{L_2} + 1.65M(h^3 d)^{1/2} + M\nu h d^{1/2}.
\end{aligned}$$

Here we use the fact that $\sqrt{a^2+b+c} \le \sqrt{a^2+b} + \frac{c}{2a}$, $\mathbf{E}[\|\nabla f(L)\|^2] \le Md$, and that we choose $h$, and $b$ such that $h/\left(b(1-mh)\right) \le m/(8M^2(d+5))$. By Lemma 9 in [DK19], the above inequality leads to

$$\|\Delta_n\|_{L_2} \le (1-0.5mh)^n\|\Delta_0\|_{L_2} + \frac{3.3M\sqrt{hd}}{m} + \frac{2\nu M\sqrt{d}}{m} + \frac{\nu M\sqrt{h}}{2\sqrt{mb}}(d+3)^{\frac{3}{2}} + \frac{3\sqrt{h(d+5)(\sigma^2+2Md)}}{\sqrt{mb}}.$$

Therefore, using the fact $W_2(\varpi_{n+1},\pi) \le \|\Delta_{n+1}\|_{L_2}$, and $W_2(\varpi_0,\pi) = \|\Delta_0\|_{L_2}$, we obtain the bound in Wasserstein distance.

$$W_2(\varpi_n,\pi) \le (1-0.5mh)^n W_2(\varpi_0,\pi) + \frac{3.3M\sqrt{hd}}{m} + \frac{2\nu M\sqrt{d}}{m} + \frac{\nu M\sqrt{h}}{2\sqrt{mb}}(d+3)^{\frac{3}{2}} + \frac{3\sqrt{h(d+5)(\sigma^2+2Md)}}{\sqrt{mb}}. \tag{45}$$

Choosing $h$, $b$, $\nu$, and $N$ as in (10), and (11) we have $W_2(\varpi_N,\pi) \le \epsilon$. ∎

## 8.2 Proofs for Oracle Complexity of ZO-KMLC

*Proof.* [of Theorem 2.2] Let $(V_{n,n}, L_{n,t}),\ t \in [0,h]$ be a stationary kinetic Langevin process for each $n \in \mathbb{N}$, i.e.,

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

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

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

or equivalently,

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

Define a different kinetic Langevin process $(\hat{V}_{n,t}, \hat{L}_{n,t})$ with initial condition $\hat{V}_{n,0} = \tilde{V}_{n,0}, \hat{L}_{n,0} = \tilde{L}_{n,0}$, i.e.,

$$\begin{aligned} d\hat{V}_{n,t} &= -(\gamma \hat{V}_{n,t} + \nabla f(\hat{L}_{n,t}))dt + \sqrt{2\gamma}dW_{n,t}, \\ d\hat{L}_{n,t} &= \hat{V}_{n,t}dt \end{aligned}$$

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 $W_2(\varpi_n,\pi) \le \|\tilde{L}_{n,0} - L_{n,0}\|_{L_2}$.

Now we denote $e_n = \left\| \boldsymbol{P}^{-1} \begin{bmatrix} \tilde{V}_{n,0} - V_{n,0} \\ \tilde{L}_{n,0} - L_{n,0} \end{bmatrix} \right\|_{L_2}$, 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 [DRD+20]. Note that $\|\tilde{L}_{n,0} - L_{n,0}\|_{L_2} \le \sqrt{2}\gamma^{-1}e_n$ and $\|\tilde{V}_{n,0} - V_{n,0}\|_{L_2} \le e_n$. Observe that,

$$\tilde{V}_{n,h} - \hat{V}_{n,h} = \int_0^h e^{-\gamma(h-s)}\left(\nabla f(\hat{L}_{n,s}) - g_{\nu,b}(\hat{L}_{n,0})\right)ds$$

$$
\begin{aligned}
&\leq \int_0^h e^{-\gamma(h-s)} \left( \nabla f(\hat{L}_{n,s}) - \nabla f(\hat{L}_{n,0}) + \nabla f(\hat{L}_{n,0}) - g_{\nu,b}(\hat{L}_{n,0}) \right) ds \\
&\leq \int_0^h e^{-\gamma(h-s)} \left( \nabla f(\hat{L}_{n,s}) - \nabla f(\hat{L}_{n,0}) - \hat{\zeta}_{n,0} + \mathbf{E}\left[\hat{\zeta}_{n,0}|\hat{L}_{n,0}\right] - \mathbf{E}\left[\hat{\zeta}_{n,0}|\hat{L}_{n,0}\right] \right) ds \\
&\leq \underbrace{\int_0^h e^{-\gamma(h-s)} \left( \nabla f(\hat{L}_{n,s}) - \nabla f(\hat{L}_{n,0}) \right) ds}_{A_1} - \underbrace{\int_0^h e^{-\gamma(h-s)} \left( \hat{\zeta}_{n,0} - \mathbf{E}\left[\hat{\zeta}_{n,0}|\hat{L}_{n,0}\right] \right) ds}_{A_2} \\
&- \underbrace{\int_0^h e^{-\gamma(h-s)} \left( \mathbf{E}\left[\hat{\zeta}_{n,0}|\hat{L}_{n,0}\right] \right) ds}_{A_3} \qquad (46)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\tilde{L}_{n,h} - \hat{L}_{n,h} \leq & \underbrace{\int_0^h \int_0^s e^{-\gamma(s-u)} \left( \nabla f(\hat{L}_{n,u}) - \nabla f(\hat{L}_{n,0}) \right) duds}_{B_1} - \underbrace{\int_0^h \int_0^s e^{-\gamma(s-u)} \left( \hat{\zeta}_{n,0} - \mathbf{E}\left[\hat{\zeta}_{n,0}|\hat{L}_{n,0}\right] \right) duds}_{B_2} \\
& - \underbrace{\int_0^h \int_0^s e^{-\gamma(s-u)} \left( \mathbf{E}\left[\hat{\zeta}_{n,0}|\hat{L}_{n,0}\right] \right) duds}_{B_3} \qquad (47)
\end{aligned}
$$

Combining (46), and (47), we have

$$
\begin{aligned}
e_{n+1} &= \left\| \boldsymbol{P}^{-1} \begin{bmatrix} \tilde{V}_{n,h} - V_{n,h} \\ \tilde{L}_{n,h} - L_{n,h} \end{bmatrix} \right\|_{L_2} \\
&\leq \left\| \boldsymbol{P}^{-1} \begin{bmatrix} A_1 - A_2 - A_3 \\ B_1 - B_2 - B_3 \end{bmatrix} + \boldsymbol{P}^{-1} \begin{bmatrix} \hat{V}_{n,h} - V_{n,h} \\ \hat{L}_{n,h} - L_{n,h} \end{bmatrix} \right\|_{L_2} \\
&\leq \left\| \boldsymbol{P}^{-1} \begin{bmatrix} \hat{V}_{n,h} - V_{n,h} \\ \hat{L}_{n,h} - L_{n,h} \end{bmatrix} - \boldsymbol{P}^{-1} \begin{bmatrix} A_2 \\ B_2 \end{bmatrix} \right\|_{L_2} + \left\| \boldsymbol{P}^{-1} \begin{bmatrix} A_1 \\ B_1 \end{bmatrix} \right\|_{L_2} + \left\| \boldsymbol{P}^{-1} \begin{bmatrix} A_3 \\ B_3 \end{bmatrix} \right\|_{L_2} \qquad (48)
\end{aligned}
$$

Now we will upper bound the above three terms. Observe that,

$$
\begin{aligned}
\|A_1\|_{L_2} &= \left\| \int_0^h e^{-\gamma(h-s)} (\nabla f(\hat{L}_{n,s}) - \nabla f(\hat{L}_{n,0})) ds \right\|_{L_2} \leq M \int_0^h \|\hat{L}_{n,s} - \hat{L}_{n,0}\|_{L_2} ds \\
&\leq M \int_0^h \int_0^s \|\hat{V}_{n,u}\|_{L_2} duds \leq \frac{1}{2} M h^2 \max_{u \in [0,h]} \|\hat{V}_{n,u}\|_{L_2}. \qquad (49)
\end{aligned}
$$

$$
\|B_1\|_{L_2} = \left\| \int_0^h \int_0^s e^{-\gamma(s-u)} (\nabla f(\hat{L}_{n,u}) - \nabla f(\hat{L}_{n,0})) ds \right\|_{L_2} \leq \frac{1}{6} M h^3 \max_{u \in [0,h]} \|\hat{V}_{n,u}\|_{L_2}. \qquad (50)
$$

So, combining (49), and (50), and using the fact $\|\hat{V}_{n,u}\|_{L_2} \leq \|V_{n,u}\|_{L_2} + \|\hat{V}_{n,u} - V_{n,u}\|_{L_2} \leq \sqrt{d} + e_n$, and choosing $h \leq \sqrt{2}/(10\gamma)$, we obtain

$$
\left\| \boldsymbol{P}^{-1} \begin{bmatrix} A_1 \\ B_1 \end{bmatrix} \right\|_{L_2} \leq \sqrt{3} \|A_1\|_{L_2} + \sqrt{2}\gamma \|B_1\|_{L_2} \leq \frac{1}{2} M h^2 \left( \sqrt{3} + \frac{\sqrt{2}\gamma h}{3} \right) (\sqrt{d} + e_n) \leq M h^3 (\sqrt{d} + e_n). \qquad (51)
$$

Using (44) we have

$$\left\|\boldsymbol{P}^{-1}\begin{bmatrix}A_3\\B_3\end{bmatrix}\right\|_{L_2} \le \sqrt{3}\|A_3\|_{L_2} + \sqrt{2}\gamma\|B_3\|_{L_2} \le \left(\sqrt{3}h + \frac{\sqrt{2}\gamma h^2}{2}\right) M\nu\sqrt{d} \le 2Mh\nu\sqrt{d}, \tag{52}$$

and

$$\begin{aligned}
&\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{n,h}-V_{n,h}\\ \hat{L}_{n,h}-L_{n,h}\end{bmatrix} - \boldsymbol{P}^{-1}\begin{bmatrix}A_2\\B_2\end{bmatrix}\right\|_{L_2}^2\\
=&\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{n,h}-V_{n,h}\\ \hat{L}_{n,h}-L_{n,h}\end{bmatrix}\right\|_{L_2}^2 + \left\|\boldsymbol{P}^{-1}\begin{bmatrix}A_2\\B_2\end{bmatrix}\right\|_{L_2}^2 - 2\mathbf{E}\left[\begin{bmatrix}\hat{V}_{n,h}-V_{n,h}\\ \hat{L}_{n,h}-L_{n,h}\end{bmatrix}^\top \begin{bmatrix}2\boldsymbol{I}_d & \gamma\boldsymbol{I}_d\\ \gamma\boldsymbol{I}_d & \gamma^2\boldsymbol{I}_d\end{bmatrix}\begin{bmatrix}A_2\\B_2\end{bmatrix}\right].
\end{aligned}$$

Note that,

$$\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{n,t}-V_{n,t}\\ \hat{L}_{n,t}-L_{n,t}\end{bmatrix}\right\|_{L_2} \le e^{-mt/\gamma}\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{n,0}-V_{n,0}\\ \hat{L}_{n,0}-L_{n,0}\end{bmatrix}\right\|_{L_2} = e^{-mt/\gamma}e_n. \tag{53}$$

Using Lemma 1.1, we also have,

$$\begin{aligned}
\|A_2\|_{L_2}^2 =& \left\|\int_0^h e^{-\gamma(h-s)}\left(\hat{\zeta}_{n,0} - \mathbf{E}\left[\hat{\zeta}_{n,0}|\hat{L}_{n,0}\right]\right)ds\right\|_{L_2}^2 = \frac{\left(1-e^{-\gamma h}\right)^2}{\gamma^2}\left\|\hat{\zeta}_{n,0} - \mathbf{E}\left[\hat{\zeta}_{n,0}|\hat{L}_{n,0}\right]\right\|_{L_2}^2\\
\le& h^2\left\|g_{\nu,b}(\hat{L}_{n,0}) - \nabla f_\nu(\hat{L}_{n,0})\right\|_{L_2}^2 \le \frac{2h^2(d+5)(\|\nabla f(\hat{L}_{n,0})\|_{L_2}^2+\sigma^2)}{b} + \frac{h^2\nu^2M^2(d+3)^3}{2b}\\
\le& \frac{2h^2(d+5)(2\|\nabla f(\hat{L}_{n,0}) - \nabla f(L_{n,0})\|_{L_2}^2 + 2\|\nabla f(L_{n,0})\|_{L_2}^2+\sigma^2)}{b} + \frac{h^2\nu^2M^2(d+3)^3}{2b}\\
\le& \frac{2h^2(d+5)(2M^2\|\hat{L}_{n,0} - L_{n,0}\|_{L_2}^2 + 2\|\nabla f(L_{n,0})\|_{L_2}^2+\sigma^2)}{b} + \frac{h^2\nu^2M^2(d+3)^3}{2b}.
\end{aligned}$$

Using the fact that $\|\hat{L}_{n,0} - L_{n,0}\|_{L_2}^2 \le 2\gamma^{-2}e_n^2$, and $\|\nabla f(L_{n,0})\|_{L_2}^2 \le Md$, we hence obtain

$$\|A_2\|_{L_2}^2 \le \frac{8M^2h^2(d+5)}{b\gamma^2}e_n^2 + h^2A_4 \tag{54}$$

where $A_4 = \frac{2(d+5)(2Md+\sigma^2)}{b} + \frac{\nu^2M^2(d+3)^3}{2b}$. Similarly, we have

$$\|B_2\|_{L_2}^2 \le \frac{2M^2h^4(d+5)}{b\gamma^2}e_n^2 + \frac{h^4}{4}A_4 \tag{55}$$

So, using (54), and (55), we have

$$\left\|\boldsymbol{P}^{-1}\begin{bmatrix}A_2\\B_2\end{bmatrix}\right\|_{L_2}^2 \le 3\|A_2\|_{L_2}^2 + 2\gamma^2\|B_2\|_{L_2}^2 \le \left(3h^2 + \frac{\gamma^2h^4}{2}\right)\left(\frac{8M^2(d+5)}{b\gamma^2}e_n^2 + A_4\right). \tag{56}$$

Now using (53), (56), and using the facts that, $\mathbf{E}\left[A_2|\hat{L}_{n,0}\right] = 0$, $\mathbf{E}\left[B_2|\hat{L}_{n,0}\right] = 0$, $\hat{V}_{n,h} - V_{n,h}$ is independent of $A_2, B_2$ given $\hat{L}_{n,0}$, and $\hat{L}_{n,h} - L_{n,h}$ is independent of $A_2, B_2$ given $\hat{L}_{n,0}$, we get

$$\left\|\boldsymbol{P}^{-1}\begin{bmatrix}\hat{V}_{n,h}-V_{n,h}\\ \hat{L}_{n,h}-L_{n,h}\end{bmatrix} - \boldsymbol{P}^{-1}\begin{bmatrix}A_2\\B_2\end{bmatrix}\right\|_{L_2}$$

$$\leq \left[\left(8M^2\left(3h^2+\frac{\gamma^2h^4}{2}\right)\frac{d+5}{b\gamma^2}+e^{-\frac{2mh}{\gamma}}\right)e_n^2+4h^2A_4\right]^{\frac{1}{2}}$$
$$\leq \left[\left(\frac{32M^2h^2(d+5)}{b\gamma^2}+e^{-\frac{2mh}{\gamma}}\right)e_n^2+4h^2A_4\right]^{\frac{1}{2}}$$
$$\leq \left[\left(\frac{32M^2h^2(d+5)}{b\gamma^2}+\left(1-\frac{mh}{2\gamma}\right)^2\right)e_n^2+4h^2A_4\right]^{\frac{1}{2}}$$
$$\leq \left[\left(1-\frac{mh}{4\gamma}\right)^2e_n^2+4h^2A_4\right]^{\frac{1}{2}}. \tag{57}$$

The second inequality follows as $h \leq \sqrt{2}/(10\gamma)$, the third inequality follows if we choose $h \leq \min(\gamma/m, \sqrt{2}/(10\gamma))$, and the last inequality follows if we choose $b \geq \frac{512M^2(d+5)}{3m^2}$. Combining, (48), (51), (52), and (57), we get

$$e_{n+1} \leq \left[\left(1-\frac{mh}{4\gamma}\right)^2e_n^2+4h^2A_4\right]^{\frac{1}{2}}+Mh^3e_n+Mh^3\sqrt{d}+2Mh\nu\sqrt{d}.$$

Using Lemma 9 of [DK19], and choosing $h \leq \min(\gamma/m, m/(12\gamma M))$ we have $mh/(4\gamma)-3Mh^2/2 \geq mh/(8\gamma)$, and thus

$$e_{n+1} \leq \left(1-\frac{mh}{8\gamma}\right)^{n+1}e_0+\frac{12M\gamma h\sqrt{d}}{m}+\frac{16M\nu\gamma\sqrt{d}}{m}+\frac{2h\sqrt{A_4}}{\sqrt{\frac{mh}{8\gamma}\left(2-\frac{mh}{4\gamma}-\frac{3Mh^2}{2}\right)}}$$
$$\leq \left(1-\frac{mh}{8\gamma}\right)^{n+1}e_0+\frac{12M\gamma h\sqrt{d}}{m}+\frac{16M\nu\gamma\sqrt{d}}{m}+\frac{4\sqrt{h}}{\sqrt{m}}\left(\frac{\sqrt{2(d+5)}(\sqrt{2Md}+\sigma)}{\sqrt{b}}+\frac{\nu M(d+3)^{\frac{3}{2}}}{\sqrt{2b}}\right)$$

Then we obtain

$$W_2(\varpi_n,\pi)$$
$$\leq \sqrt{2}\gamma^{-1}e_n$$
$$\leq \sqrt{2}\gamma^{-1}\left(1-\frac{mh}{8\gamma}\right)^{n+1}W_2(\varpi_0,\pi)+\frac{24Mh\sqrt{d}}{m}+\frac{32M\nu\sqrt{d}}{m}$$
$$+\frac{4\sqrt{h}}{\gamma\sqrt{m}}\left(\frac{2\sqrt{(d+5)}(\sqrt{2Md}+\sigma)}{\sqrt{b}}+\frac{\nu M(d+3)^{\frac{3}{2}}}{\sqrt{b}}\right)$$

Now, choosing $h$, $\nu$, $b$, and $N$ as in (16), we get (17), and (18). ∎

### 8.3 Proofs for Oracle Complexity of ZO-RMP

Before proceeding, we also recall that $(x_n^*(t), v_n^*(t))$ when $t \in [0,h]$ is the true solution to the underdamped Langevin diffusion with the initial point $(x_n, v_n)$ coupled with $x_{n+\frac{1}{2}}$ through a shared Brownian motion defined as follows:

$$x_n^*(t) = x_n+\frac{1-e^{-2t}}{2}v_n-\frac{u}{2}\int_0^t\left(1-e^{-2(t-s)}\right)\nabla f(x_n^*(s))ds+\sqrt{u}\int_0^t\left(1-e^{-2(t-s)}\right)dB_s \tag{58}$$

$$v_n^*(t) = v_n e^{-2t} - u\left(\int_0^t e^{-2(t-s)}\nabla f(x_n^*(s))ds\right) + 2\sqrt{u}\int_0^t e^{-2(t-s)}dB_s. \tag{59}$$

We also recall some preliminary results from [SL19].

**Lemma 8.1 (Lemma 6[SL19])** *Let $\{x(t)\}_{t\in[0,h]}$, and $\{v(t)\}_{t\in[0,h]}$ be the true solution to the underdamped Langevin diffusion* (58)*, and* (59) *on $t \in [0,h]$. Then for $h \le 1/20$, and $u = 1/M$, we have*

$$\mathbf{E}\left[\sup_{t\in[0,h]} \|x(0) - x(t)\|^2\right] \le O\left(h^2\|v(0)\|^2 + u^2h^4\|\nabla f(x(0))\|^2 + udh^3\right) \tag{60}$$

$$\mathbf{E}\left[\sup_{t\in[0,h]} \|\nabla f(x_t)\|^2\right] \le O(\|\nabla f(x(0))\|^2 + M^2h^2\|v(0)\|^2 + Mdh^3) \tag{61}$$

$$\mathbf{E}\left[\sup_{t\in[0,h]} \|v(t)\|^2\right] \le O(\|v(0)\|^2 + u^2h^2\|\nabla f(x(0))\|^2 + udh) \tag{62}$$

**Lemma 8.2** *Let $\alpha_n$ be sampled uniformly randomly from $[0,1]$ at iteration $n$. Let $x_{n+\frac{1}{2}}$ be the intermediate value at step $n$. Let $\{x_n^*(t)\}_{t\in[0,h]}$ be the true solution to* (58)*, and* (59) *with the initial point $x_n^*(0) = x_n$ coupled to $x_{n+\frac{1}{2}}$ through a shared Brownian motion. Then, under Assumption 1.3 and Assumption 1.1, for $h \le 1/20$, we have*

$$\begin{aligned}\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\|^2\right] \le& O\Big(M^2h^6\mathbf{E}\left[\|v_n\|^2\right] + (h^8 + h^7\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] \\ &+ Mdh^7 + h^7\kappa^{-1}\sigma^2 + h^8\Big).\end{aligned} \tag{63}$$

*Proof.* [of Lemma 8.2] For notational simplicity, we drop the subscript $n$ from $\alpha_n$ below. First, note that we have

$$\begin{aligned}
&\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\|^2\right] \\
\le& M^2\mathbf{E}\left[\|x_{n+\frac{1}{2}} - x_n^*(\alpha h)\|^2\right] \\
\le& M^2\mathbf{E}\left[\|\frac{u}{2}\int_0^{\alpha h}(1 - e^{-2(\alpha h - s)})(g_{\nu,b}(x_n) - \nabla f(x_n^*(s))ds\|^2\right] \\
\le& \frac{u^2M^2}{4}\mathbf{E}\left[\int_0^{\alpha h}(1 - e^{-2(\alpha h-s)})^2 ds\int_0^{\alpha h}\|g_{\nu,b}(x_n) - \nabla f(x_n^*(s))\|^2 ds\right] \\
\le& h^3\mathbf{E}\left[\int_0^{\alpha h}\|g_{\nu,b}(x_n) - \nabla f(x_n^*(s))\|^2 ds\right] \\
\le& 2h^3\mathbf{E}\left[\int_0^{\alpha h}\left(\|g_{\nu,b}(x_n) - \nabla f(x_n)\|^2 + \|\nabla f(x_n) - \nabla f(x_n^*(s))\|^2\right) ds\right] \\
\le& 2h^3\mathbf{E}\left[\int_0^{\alpha h}\left(\|g_{\nu,b}(x_n) - \nabla f(x_n)\|^2 + M^2\|x_n - x_n^*(s)\|^2\right) ds\right] \\
\le& 2h^3\mathbf{E}\Bigg[\int_0^{\alpha h}\left(\frac{3\nu^2}{2}M^2(d+3)^3 + \frac{4(d+5)\left(\sigma^2 + \|\nabla f(x_n)\|^2\right)}{b}\right.
\end{aligned}$$

$$+ M^2 O\left(h^2\|v_n\|^2 + u^2h^4\|\nabla f(x_n)\|^2 + udh^3\right)\Big)ds\Big]$$
$$=2h^4\mathbf{E}\left[\frac{3\nu^2}{2}M^2(d+3)^3 + \frac{4(d+5)\left(\sigma^2 + \|\nabla f(x_n)\|^2\right)}{b} + M^2O\left(h^2\|v_n\|^2 + u^2h^4\|\nabla f(x_n)\|^2 + udh^3\right)\right]$$

The first and the sixth inequality follows from the first condition of Assumption 1.1, the second inequality follows from (19), and (58), the third inequality follows from Cauchy-Schwarz inequality, the fourth inequality follows from choosing $u = 1/M$ and the fact that $1 - e^{-2(\alpha h - s)} \leq 2h$, the fifth inequality follows from Young's inequality, the seventh inequality follows from Lemma 1.1 and Lemma 8.1. Choosing $b$, and $\nu$ as in (22), we have,

$$\begin{aligned}&\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\|^2\right]\\ \leq& O\left(M^2h^6\mathbf{E}\left[\|v_n\|^2\right] + (h^8 + h^7\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + Mdh^7 + h^7\kappa^{-1}\sigma^2 + h^8\right)\end{aligned}$$

■

**Lemma 8.3** *Let $g_{\nu,b}(x_n)$ be defined as in* (6)*. Then under the conditions of Lemma 8.2, we have*

$$\begin{aligned}&\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - g_{\nu,b}(x_{n+\frac{1}{2}})\|^2\right]\\ \leq& O\left(M^2h^5\kappa^{-1}\mathbf{E}\left[\|v_n\|^2\right] + h^3\kappa^{-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + h^4 + Mdh^6\kappa^{-1} + h^3\kappa^{-1}\sigma^2\right)\end{aligned} \tag{64}$$

*Proof.* [of Lemma 8.3] Using Lemma 1.1 and Young's inequality, we have

$$\begin{aligned}&\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - g_{\nu,b}(x_{n+\frac{1}{2}})\|^2\right] \leq \frac{3\nu^2}{2}M^2(d+3)^3 + \frac{4(d+5)\left(\sigma^2 + \mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}})\|^2\right]\right)}{b}\\ \leq& \frac{3\nu^2}{2}M^2(d+3)^3 + \frac{4(d+5)\left(\sigma^2 + 2\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\|^2\right] + 2\mathbf{E}\left[\|\nabla f(x_n^*(\alpha h))\|^2\right]\right)}{b}.\end{aligned}$$

Furthermore, using Lemma 8.2, and (61), and the fact that $h$ is small, we get

$$\begin{aligned}&\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\|^2\right] + \mathbf{E}\left[\|\nabla f(x_n^*(\alpha h))\|^2\right]\\ \leq& O\left(M^2h^2\mathbf{E}\left[\|v_n\|^2\right] + \mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + Mdh^3 + h^7\kappa^{-1}\sigma^2 + h^8\right).\end{aligned}$$

Hence, we have

$$\begin{aligned}&\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - g_{\nu,b}(x_{n+\frac{1}{2}})\|^2\right]\\ \leq& O\left(M^2h^5\kappa^{-1}\mathbf{E}\left[\|v_n\|^2\right] + h^3\kappa^{-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + h^4 + Mdh^6\kappa^{-1} + h^3\kappa^{-1}\sigma^2\right)\end{aligned}$$

■

**Lemma 8.4** *Let $\mathbf{E}_\alpha$ denote the expectation with respect to $\alpha$ at each iteration $n$. Let $\mathbf{E}[\cdot]$ be the expectation with respect to other randomness present in iteration $n$. Let $\{x_n^*(t)\}_{t\in[0,h]}$ be the true*

*solution to* (58)*, and* (59) *with the initial point* $x_n^*(0) = x_n$ *coupled to* $x_{n+\frac{1}{2}}$*,* $v_n$*, and* $x_{n+1}$ *through a shared Brownian motion. Then, under Assumption 1.3–1.1, for* $h \leq 1/20$*, and* $u = 1/M$*, we have*

$$\begin{aligned}
\mathbf{E}\left[\|\mathbf{E}_\alpha x_{n+1} - x_n^*(h)\|^2\right] \leq & O\left((h^{10} + h^9\kappa^{-1})\mathbf{E}\left[\|v_n\|^2\right] + u^2(h^{12} + h^7\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\right. \\
& \left. + ud(h^{11} + h^{10}\kappa^{-1}) + u^2h^7\kappa^{-1}\sigma^2 + u^2h^8\right) \qquad (65a) \\
\mathbf{E}\left[\|\mathbf{E}_\alpha v_{n+1} - v_n^*(h)\|^2\right] \leq & O\left((h^7\kappa^{-1} + h^8)\mathbf{E}\left[\|v_n\|^2\right] + u^2(h^{10} + h^5\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\right. \\
& \left. + u^2h^6 + u^2h^5\kappa^{-1}\sigma^2 + ud(h^9 + h^8\kappa^{-1})\right) \qquad (65b) \\
\mathbf{E}\left[\|x_{n+1} - x_n^*(h)\|^2\right] \leq & O\left(h^6\mathbf{E}\left[\|v_n\|^2\right] + u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^8 + udh^7 + u^2h^7\kappa^{-1}\sigma^2\right) \qquad (65c) \\
\mathbf{E}\left[\|v_{n+1} - v_n^*(h)\|^2\right] \leq & O\left(h^4\mathbf{E}\left[\|v_n\|^2\right] + u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^8 + udh^5 + u^2h^7\kappa^{-1}\sigma^2\right) \qquad (65d)
\end{aligned}$$

*Proof.* [of Lemma 8.4]

a) Using Lemma 8.2, and 8.3, we have

$$\begin{aligned}
& \mathbf{E}\left[\|\mathbf{E}_\alpha x_{n+1} - x_n^*(h)\|^2\right] \\
\leq & \mathbf{E}\left[\|\frac{uh}{2}\mathbf{E}_\alpha(1 - e^{-2(h-\alpha h)})g_{\nu,b}(x_{n+\frac{1}{2}}) - \frac{u}{2}\int_0^h (1 - e^{-2(h-s)})\nabla f(x_n^*(s))ds\|^2\right] \\
\leq & 2\mathbf{E}\left[\|\frac{uh}{2}\mathbf{E}_\alpha(1 - e^{-2(h-\alpha h)})(g_{\nu,b}(x_{n+\frac{1}{2}}) - \nabla f(x_{n+\frac{1}{2}}))\|^2\right] \\
& + 2\mathbf{E}\left[\|\frac{uh}{2}\mathbf{E}_\alpha(1 - e^{-2(h-\alpha h)})\nabla f(x_{n+\frac{1}{2}}) - \frac{u}{2}\int_0^h (1 - e^{-2(h-s)})\nabla f(x_n^*(s))ds\|^2\right] \\
\leq & 2u^2h^4\mathbf{E}\left[\|\mathbf{E}_\alpha(g_{\nu,b}(x_{n+\frac{1}{2}}) - \nabla f(x_{n+\frac{1}{2}}))\|^2\right] \\
& + 2\mathbf{E}\left[\|\frac{uh}{2}\mathbf{E}_\alpha(1 - e^{-2(h-\alpha h)})\nabla f(x_{n+\frac{1}{2}}) - \frac{u}{2}\int_0^h (1 - e^{-2(h-s)})\nabla f(x_n^*(s))ds\|^2\right] \\
\leq & O\left(h^9\kappa^{-1}\mathbf{E}\left[\|v_n\|^2\right] + u^2h^7\kappa^{-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^8 + udh^{10}\kappa^{-1} + u^2h^7\kappa^{-1}\sigma^2\right) \\
& + 2\mathbf{E}\left[\|\frac{uh}{2}\mathbf{E}_\alpha(1 - e^{-2(h-\alpha h)})(\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))) + \frac{uh}{2}\mathbf{E}_\alpha(1 - e^{-2(h-\alpha h)})\nabla f(x_n^*(\alpha h))\right. \\
& \left. - \frac{u}{2}\int_0^h (1 - e^{-2(h-s)})\nabla f(x_n^*(s))ds\|^2\right] \\
\leq & O\left(h^9\kappa^{-1}\mathbf{E}\left[\|v_n\|^2\right] + u^2h^7\kappa^{-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^8 + udh^{10}\kappa^{-1} + u^2h^7\kappa^{-1}\sigma^2\right) \\
& + 2u^2h^4\mathbf{E}\left[\|\mathbf{E}_\alpha(\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h)))\|^2\right] \\
\leq & O\left(h^9\kappa^{-1}\mathbf{E}\left[\|v_n\|^2\right] + u^2h^7\kappa^{-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^8 + udh^{10}\kappa^{-1} + u^2h^7\kappa^{-1}\sigma^2\right) \\
& + O\left(h^{10}\mathbf{E}\left[\|v_n\|^2\right] + u^2(h^{12} + h^{11}\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + udh^{11} + u^2h^{11}\kappa^{-1}\sigma^2 + u^2h^{12}\right) \\
\leq & O\left((h^{10} + h^9\kappa^{-1})\mathbf{E}\left[\|v_n\|^2\right] + u^2(h^{12} + h^7\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + ud(h^{11} + h^{10}\kappa^{-1})\right. \\
& \left. + u^2h^7\kappa^{-1}\sigma^2 + u^2h^8\right)
\end{aligned}$$

The second inequality follows from Young's inequality, the third and fifth inequality uses the fact $1-e^{-2(\alpha-\alpha h)} \leq 2h$, the fifth inequality follows from the fact $\frac{uh}{2}\mathbf{E}_\alpha(1-e^{-2(h-\alpha h)})\nabla f(x_n^*(\alpha h))-\frac{u}{2}\int_0^h(1-e^{-2(h-s)})\nabla f(x_n^*(s))ds = 0$.

b) Next, note that

$$
\begin{aligned}
&\mathbf{E}\left[\|\mathbf{E}_\alpha v_{n+1} - v_n^*(h)\|^2\right]\\
=&\mathbf{E}\left[\|\mathbf{E}_\alpha uhe^{-2(h-\alpha h)}g_{\nu,b}(x_{n+\frac{1}{2}}) - u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))ds\|^2\right]\\
=&\mathbf{E}\Bigg[\|\mathbf{E}_\alpha uhe^{-2(h-\alpha h)}(g_{\nu,b}(x_{n+\frac{1}{2}}) - \nabla f(x_{n+\frac{1}{2}}) + \nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h)) + \nabla f(x_n^*(\alpha h)))\\
&-u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))ds\|^2\Bigg]\\
\leq&2u^2h^2\mathbf{E}\left[\|g_{\nu,b}(x_{n+\frac{1}{2}}) - \nabla f(x_{n+\frac{1}{2}})\|^2\right] + 2u^2h^2\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\|^2\right]\\
\leq&2u^2h^2O\left(M^2h^5\kappa^{-1}\mathbf{E}\left[\|v_n\|^2\right] + h^3\kappa^{-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + h^4 + Mdh^6\kappa^{-1} + h^3\kappa^{-1}\sigma^2\right)\\
&+2u^2h^2O\left(M^2h^6\mathbf{E}\left[\|v_n\|^2\right] + (h^8+h^7\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + Mdh^7 + h^7\kappa^{-1}\sigma^2 + h^8\right)\\
\leq&O\Big((h^7\kappa^{-1}+h^8)\mathbf{E}\left[\|v_n\|^2\right] + u^2(h^{10}+h^5\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^6\\
&+u^2h^5\kappa^{-1}\sigma^2 + ud(h^9+h^8\kappa^{-1})\Big)
\end{aligned}
$$

The first inequality follows from using, $\mathbf{E}_\alpha\nabla f(x_n^*(\alpha h))) - u\int_0^h e^{-2(h-s)}\nabla f(x_n^*(s))ds = 0$, and $e^{-2(h-\alpha h)} \leq 1$, and the second inequality follows from Lemma 8.2, and 8.3.

c) For the next part, note that we have

$$
\begin{aligned}
&\mathbf{E}\left[\|x_{n+1} - x_n^*(h)\|^2\right]\\
\leq&\mathbf{E}\left[\|\frac{uh}{2}(1-e^{-2(h-\alpha h)})g_{\nu,b}(x_{n+\frac{1}{2}}) - \frac{u}{2}\int_0^h(1-e^{-2(h-s)})\nabla f(x_n^*(s))ds\|^2\right]\\
\leq&\mathbf{E}\Bigg[\|\frac{uh}{2}(1-e^{-2(h-\alpha h)})(g_{\nu,b}(x_{n+\frac{1}{2}}) - \nabla f(x_{n+\frac{1}{2}}) + \nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\\
&+\nabla f(x_n^*(\alpha h))) - \frac{u}{2}\int_0^h(1-e^{-2(h-\alpha h)})\nabla f(x_n^*(s))ds + \frac{u}{2}\int_0^h(1-e^{-2(h-\alpha h)})\nabla f(x_n^*(s))ds\\
&-\frac{u}{2}\int_0^h(1-e^{-2(h-s)})\nabla f(x_n^*(s))ds\|^2\Bigg]\\
\leq&4u^2h^4\mathbf{E}\left[\|g_{\nu,b}(x_{n+\frac{1}{2}}) - \nabla f(x_{n+\frac{1}{2}})\|^2\right] + 4u^2h^4\mathbf{E}\left[\|\nabla f(x_{n+\frac{1}{2}}) - \nabla f(x_n^*(\alpha h))\|^2\right]\\
&+\mathbf{E}\left[\|uh(1-e^{-2(h-\alpha h)})\nabla f(x_n^*(\alpha h)) - u\int_0^h(1-e^{-2(h-\alpha h)})\nabla f(x_n^*(s))ds\|^2\right]\\
&+u^2\mathbf{E}\left[\|\int_0^h(1-e^{-2(h-\alpha h)})\nabla f(x_n^*(s))ds - \int_0^h(1-e^{-2(h-s)})\nabla f(x_n^*(s))ds\|^2\right]\\
\leq&O\left(h^9\kappa^{-1}\mathbf{E}\left[\|v_n\|^2\right] + u^2h^7\kappa^{-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^8 + udh^{10}\kappa^{-1} + u^2h^7\kappa^{-1}\sigma^2\right)\\
&+O\left(h^{10}\mathbf{E}\left[\|v_n\|^2\right] + u^2(h^{12}+h^{11}\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + udh^{11} + u^2h^{11}\kappa^{-1}\sigma^2 + u^2h^{12}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+16h^4\mathbf{E}\left[\sup_{t\in[0,h]}\|x_n^*(0)-x_n^*(t)\|^2\right]+4u^2h^4\mathbf{E}\left[\sup_{t\in[0,h]}\|\nabla f(x_n^*(t))\|^2\right]\\
&\leq O\big((h^{10}+h^9\kappa^{-1})\mathbf{E}\left[\|v_n\|^2\right]+u^2(h^7\kappa^{-1}+h^{12})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\\
&+u^2h^8+(udh^{10}\kappa^{-1}+udh^{11})+u^2h^7\kappa^{-1}\sigma^2\big)\\
&+O\left(h^6\mathbf{E}\left[\|v_n\|^2\right]+u^2h^8\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+udh^7\right)\\
&+O(h^6\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+udh^7)\\
&\leq O\big((h^{10}+h^9\kappa^{-1})\mathbf{E}\left[\|v_n\|^2\right]+u^2(h^7\kappa^{-1}+h^{12})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\\
&+u^2h^8+(udh^{10}\kappa^{-1}+udh^{11})+u^2h^7\kappa^{-1}\sigma^2\big)\\
&+O(h^6\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+udh^7)\\
&\leq O\left(h^6\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+u^2h^8+udh^7+u^2h^7\kappa^{-1}\sigma^2\right)
\end{aligned}
$$

The third inequality follows from Young's inequality, the fourth inequality follows from Lemma 8.2, and 8.3, and the fact $1-e^{-2(\alpha-\alpha h)}\leq 2h$, and the fifth inequality follows from (60),and (61).

d) Finally, note that we have

$$
\begin{aligned}
&\mathbf{E}\left[\|v_{n+1}-v_n^*(h)\|^2\right]\\
&\leq 2u^2h^4\mathbf{E}\left[\|g_{\nu,b}(x_{n+\frac{1}{2}})-\nabla f(x_{n+\frac{1}{2}})\|^2\right]+O(h^4\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+udh^5)\\
&\leq O\left(h^9\kappa^{-1}\mathbf{E}\left[\|v_n\|^2\right]+u^2h^7\kappa^{-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+u^2h^8+udh^{10}\kappa^{-1}+u^2h^7\kappa^{-1}\sigma^2\right)\\
&+O(h^4\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+udh^5)\\
&\leq O\left(h^4\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+u^2h^8+udh^5+u^2h^7\kappa^{-1}\sigma^2\right)
\end{aligned}
$$

The first inequality follows from Lemma 2 of [SL19], and the second inequality follows from Lemma 8.3.

■

**Lemma 8.5** *Under conditions of Lemma 8.4,*

$$
\mathbf{E}\left[f(x_{n+1}(0))-f(x_n(h))\right]\leq O\left(Mh^5\mathbf{E}\left[\|v_n\|^2\right]+uh^3\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+dh^6+uh^4\kappa^{-1}\sigma^2+uh^5\right) \tag{66}
$$

*Proof.* [of Lemma 8.5] Note that, we have

$$
\begin{aligned}
&\mathbf{E}\left[f(x_{n+1}(0))-f(x_n(h))\right]\\
&\leq uh^3\mathbf{E}\left[\|\nabla f(x_n(h))\|^2\right]+\frac{M}{h^3}\mathbf{E}\left[\|\mathbf{E}_\alpha x_{n+1}(0)-x_n(h)\|^2\right]+\frac{M}{2}\mathbf{E}\left[\|x_{n+1}(0)-x_n(h)\|^2\right]\\
&\leq uh^3O(M^2h^2\mathbf{E}\left[\|v_n\|^2\right]+\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+Mdh^3)\\
&+\frac{M}{h^3}O\left((h^{10}+h^9\kappa^{-1})\mathbf{E}\left[\|v_n\|^2\right]+u^2(h^{12}+h^7\kappa^{-1})\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+ud(h^{11}+h^{10}\kappa^{-1})+u^2h^7\kappa^{-1}\sigma^2+u^2h^8\right)\\
&+\frac{M}{2}O\left(h^6\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+u^2h^8+udh^7+u^2h^7\kappa^{-1}\sigma^2\right)\\
&\leq O\left(Mh^5\mathbf{E}\left[\|v_n\|^2\right]+uh^3\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+dh^6+uh^4\kappa^{-1}\sigma^2+uh^5\right)
\end{aligned}
$$

■

**Lemma 8.6** *At iteration $n$, with the initial point $(x_n, v_n)$, for the updates* (19), (20), *and* (21), *we have*

$$\sum_{n=0}^{N-1} \mathbf{E}\left[\|v_n\|^2\right] \leq O\left(u^2 h \sum_{n=0}^{N-1} \mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + Ndu + Nu^2h^3\kappa^{-1}\sigma^2 + Nu^2h^4\right) \tag{67}$$

*Proof.* [of Lemma 8.6] From Lemma 11 of [SL19], we have

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{2u}\|v_n(h)\|^2 + f(x_n(h))\right] \\
\leq &\mathbf{E}\left[\frac{1}{2u}\|v_n\|^2 + f(x_n)\right] - \frac{2}{3}hM\mathbf{E}\left[\|v_n\|^2\right] + O\left(uh^3\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + dh\right)
\end{aligned} \tag{68}$$

From Lemma 11 we also have,

$$\begin{aligned}
&\mathbf{E}\left[\|v_{n+1}\|^2 - \|v_n(h)\|^2\right] \\
\leq &\frac{2}{h^2}\mathbf{E}\left[\|v_{n+1} - v_n(h)\|^2\right] + 4h^2\mathbf{E}\left[\|v_n(h)\|^2\right] \\
\leq &\frac{2}{h^2}O\left(h^4\mathbf{E}\left[\|v_n\|^2\right] + u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^8 + udh^5 + u^2h^7\kappa^{-1}\sigma^2\right) \\
&+4h^2O(\mathbf{E}\left[\|v_n\|^2\right] + u^2h^2\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + udh) \\
\leq &O\left(h^2\mathbf{E}\left[\|v_n\|^2\right] + u^2h^2\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + u^2h^6 + udh^3 + u^2h^5\kappa^{-1}\sigma^2\right)
\end{aligned} \tag{69}$$

The second inequality above follows from (62). From Lemma 8.5, we have

$$\mathbf{E}\left[f(x_{n+1}(0)) - f(x_n(h))\right] \leq O\left(Mh^5\mathbf{E}\left[\|v_n\|^2\right] + uh^3\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + dh^6 + uh^4\kappa^{-1}\sigma^2 + uh^5\right)$$

Now, from (68), (69) and Lemma 8.5, we get

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{2u}\|v_{n+1}\|^2 + f(x_{n+1})\right] = &\mathbf{E}\left[\frac{1}{2u}(\|v_{n+1}\|^2 - \|v_n(h)\|^2) + f(x_{n+1}) - f(x_n(h)\right] \\
&+ \mathbf{E}\left[\frac{1}{2u}\|v_n(h)\|^2 + f(x_n(h))\right] \\
\leq &O\left(Mh^2\mathbf{E}\left[\|v_n\|^2\right] + uh^2\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + uh^6 + dh^3 + uh^5\kappa^{-1}\sigma^2\right) \\
&+O\left(Mh^5\mathbf{E}\left[\|v_n\|^2\right] + uh^3\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + dh^6 + uh^4\kappa^{-1}\sigma^2 + uh^5\right) \\
&+\mathbf{E}\left[\frac{1}{2u}\|v_n\|^2 + f(x_n)\right] - \frac{2}{3}hM\mathbf{E}\left[\|v_n\|^2\right] + O\left(uh^3\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + dh\right)
\end{aligned}$$

Choosing $h$ such that, $\frac{1}{3}hM \geq Mh^2$, i.e., $h \leq \frac{1}{3}$, we get

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{2u}\|v_{n+1}\|^2 + f(x_{n+1})\right] \\
\leq &O\left(uh^2\mathbf{E}\left[\|\nabla f(x_n)\|^2\right] + dh + uh^4\kappa^{-1}\sigma^2 + uh^5\right) + \mathbf{E}\left[\frac{1}{2u}\|v_n\|^2 + f(x_n)\right] - \frac{1}{3}hM\mathbf{E}\left[\|v_n\|^2\right]
\end{aligned}$$

Summing both sides from $n=0$ to $N-1$, we get

$$\begin{aligned}\sum_{n=0}^{N-1}\mathbf{E}\left[\frac{1}{2u}\|v_{n+1}\|^2+f(x_{n+1})\right]\leq& O\left(uh^2\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+Ndh+Nuh^4\kappa^{-1}\sigma^2+Nuh^5\right)\\&+\mathbf{E}\left[\frac{1}{2u}\sum_{n=0}^{N-1}\left(\|v_n\|^2+f(x_n)\right)\right]-\frac{1}{3}hM\sum_{n=0}^{N-1}\mathbf{E}\left[\|v_n\|^2\right]\end{aligned}$$

Since, $\|v_0\|=0$, and $\mathbf{E}\left[f(x_0)\right]-f(x^*)\leq O(d)$, and consequently, $\mathbf{E}\left[f(x_0)-f(x_N)\right]\leq O(d)$, we have

$$\begin{aligned}\frac{1}{3}hM\sum_{n=0}^{N-1}\mathbf{E}\left[\|v_n\|^2\right]&\leq O\left(uh^2\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+Ndh+Nuh^4\kappa^{-1}\sigma^2+Nuh^5\right)\\\sum_{n=0}^{N-1}\mathbf{E}\left[\|v_n\|^2\right]&\leq O\left(u^2h\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+Ndu+Nu^2h^3\kappa^{-1}\sigma^2+Nu^2h^4\right)\end{aligned}$$

■

**Lemma 8.7** *At iteration $n$, with the initial point $(x_n,v_n)$, for the updates* (19), (20), *and* (21), *we have*

$$\begin{aligned}\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]&\leq O\left(\frac{M}{h}\left|\mathbf{E}\left[\nabla f(x_N)^\top v_N\right]\right|+MNd+Nh^3\kappa^{-1}\sigma^2+Nh^4\right)\\\sum_{n=0}^{N-1}\mathbf{E}\left[\|v_n\|^2\right]&\leq O\left(u\left|\mathbf{E}\left[\nabla f(x_N)^\top v_N\right]\right|+Ndu+Nu^2h^3\kappa^{-1}\sigma^2+Nu^2h^4\right)\end{aligned}$$

*Proof.* [of Lemma 8.7] From (15) in Lemma 12 of [SL19] we have,

$$\begin{aligned}&\mathbf{E}\left[\nabla f(x_n(h))^\top v_n(h)\right]\\\leq&\mathbf{E}\left[\nabla f(x_n)^\top v_n\right]-\frac{1}{6}uh\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+O\left(Mh\mathbf{E}\left[\|v_n\|^2\right]+uh^3\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+dh^2\right)\end{aligned}\tag{70}$$

From Lemma 12 of [SL19] we also have,

$$\begin{aligned}&\mathbf{E}\left[\nabla f(x_{n+1})^\top v_{n+1}-\nabla f(x_n(h))^\top v_n(h)\right]\\\leq&\frac{2u}{h}\mathbf{E}\left[\|\nabla f(x_{n+1})-\nabla f(x_n(h))\|^2\right]+\frac{2M}{h^2}\mathbf{E}\left[\|v_{n+1}-v_n(h)\|^2\right]+uh^2\mathbf{E}\left[\|\nabla f(x_n(h))\|^2\right]+Mh\mathbf{E}\left[\|v_n(h)\|^2\right]\\\leq&\frac{2M}{h}\mathbf{E}\left[\|x_{n+1}-x_n(h)\|^2\right]+\frac{2M}{h^2}\mathbf{E}\left[\|v_{n+1}-v_n(h)\|^2\right]+uh^2\mathbf{E}\left[\|\nabla f(x_n(h))\|^2\right]+Mh\mathbf{E}\left[\|v_n(h)\|^2\right]\end{aligned}$$

Now from (65c), (65d), and Lemma 8.1 we have,

$$\begin{aligned}&\mathbf{E}\left[\nabla f(x_{n+1})^\top v_{n+1}-\nabla f(x_n(h))^\top v_n(h)\right]\\\leq&\frac{2M}{h}O\left(h^6\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+u^2h^8+udh^7+u^2h^7\kappa^{-1}\sigma^2\right)\end{aligned}$$

$$
\begin{aligned}
&+\frac{2M}{h^2}O\left(h^4\mathbf{E}\left[\|v_n\|^2\right]+u^2h^4\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+u^2h^8+udh^5+u^2h^7\kappa^{-1}\sigma^2\right)\\
&+uh^2O(M^2h^2\|v_n\|^2+\|\nabla f(x_n)\|^2+Mdh^3)+MhO(\|v_n\|^2+u^2h^2\|\nabla f(x_n)\|^2+udh)\\
&\leq O\left(Mh\mathbf{E}\left[\|v_n\|^2\right]+uh^2\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+dh^2+uh^6+uh^5\kappa^{-1}\sigma^2\right) \qquad (71)
\end{aligned}
$$

Combining (70), and (71), we get

$$
\begin{aligned}
\mathbf{E}\left[\nabla f(x_{n+1})^\top v_{n+1}\right]\leq&\mathbf{E}\left[\nabla f(x_n)^\top v_n\right]-\frac{1}{6}uh\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\\
&+O\left(Mh\mathbf{E}\left[\|v_n\|^2\right]+uh^2\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+dh^2+uh^6+uh^5\kappa^{-1}\sigma^2\right).
\end{aligned}
$$

Summing both sides from $n=0$ to $N-1$, and using Lemma 8.6, we get

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\mathbf{E}\left[\nabla f(x_{n+1})^\top v_{n+1}\right]\\
&\leq\sum_{n=0}^{N-1}\mathbf{E}\left[\nabla f(x_n)^\top v_n\right]-\frac{1}{6}uh\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\\
&+O\left(Mh\sum_{n=0}^{N-1}\mathbf{E}\left[\|v_n\|^2\right]+uh^2\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+Ndh^2+Nuh^6+Nuh^5\kappa^{-1}\sigma^2\right)\\
&\leq\sum_{n=0}^{N-1}\mathbf{E}\left[\nabla f(x_n)^\top v_n\right]-\frac{1}{6}uh\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\\
&+O\left(MhO\left(u^2h\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+Ndu+Nu^2h^3\kappa^{-1}\sigma^2+Nu^2h^4\right)\right.\\
&\quad\left.+uh^2\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+Ndh^2+Nuh^6+Nuh^5\kappa^{-1}\sigma^2\right)\\
&\leq\sum_{n=0}^{N-1}\mathbf{E}\left[\nabla f(x_n)^\top v_n\right]-\frac{1}{6}uh\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\\
&+O\left(uh^2\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]+Ndh+Nuh^4\kappa^{-1}\sigma^2+Nuh^5\right)
\end{aligned}
$$

Now choosing $\frac{1}{24}uh\geq uh^2$, and $v_0=0$, we have,

$$
\begin{aligned}
&\frac{1}{8}uh\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\leq O\left(\left|\mathbf{E}\left[\nabla f(x_N)^\top v_N\right]\right|+Ndh+Nuh^4\kappa^{-1}\sigma^2+Nuh^5\right)\\
&\sum_{n=0}^{N-1}\mathbf{E}\left[\|\nabla f(x_n)\|^2\right]\leq O\left(\frac{M}{h}\left|\mathbf{E}\left[\nabla f(x_N)^\top v_N\right]\right|+MNd+Nh^3\kappa^{-1}\sigma^2+Nh^4\right)
\end{aligned}
$$

Using Lemma 8.6, we have,

$$
\sum_{n=0}^{N-1}\mathbf{E}\left[\|v_n\|^2\right]\leq O\left(u\left|\mathbf{E}\left[\nabla f(x_N)^\top v_N\right]\right|+Ndu+Nu^2h^3\kappa^{-1}\sigma^2+Nu^2h^4\right)
$$

■

*Proof.* [of Theorem 2.3] From Theorem 3 in [SL19], we have

$$\begin{aligned} q_N \leq & e^{-\frac{Nh}{2\kappa}} q_0 + \sum_{n=1}^{N} \frac{2\kappa}{h} \left( 2\mathbf{E}\left[ \|\mathbf{E}_\alpha v_{n+1} - v_n^*(h)\|^2 \right] + 3\mathbf{E}\left[ \|\mathbf{E}_\alpha x_{n+1} - x_n^*(h)\|^2 \right] \right) \\ & + \sum_{n=1}^{N} \left( 2\mathbf{E}\left[ \|v_{n+1} - v_n^*(h)\|^2 \right] + 3\mathbf{E}\left[ \|x_{n+1} - x_n^*(h)\|^2 \right] \right) \end{aligned} \tag{72}$$

where $q_N = \mathbf{E}\left[ \|x_N - y_N\|^2 + \|x_N + v_N - y_N - w_N\|^2 \right]$. We also have,

$$e^{-\frac{Nh}{2\kappa}} q_0 \leq \frac{\epsilon^2 d}{4m} \tag{73}$$

From Lemma 8.4,

$$\begin{aligned} & \sum_{n=1}^{N} \frac{2\kappa}{h} \left( 2\mathbf{E}\left[ \|\mathbf{E}_\alpha v_{n+1} - v_n^*(h)\|^2 \right] + 3\mathbf{E}\left[ \|\mathbf{E}_\alpha x_{n+1} - x_n^*(h)\|^2 \right] \right) \\ \leq & O\left( (h^6 + \kappa h^7) \sum_{n=1}^{N} \mathbf{E}\left[ \|v_n\|^2 \right] + u^2(\kappa h^9 + h^4) \sum_{n=1}^{N} \mathbf{E}\left[ \|\nabla f(x_n)\|^2 \right] \right. \\ & \left. + N\kappa u^2 h^5 + N u^2 h^4 \sigma^2 + N u d(\kappa h^8 + h^7) \right) \end{aligned} \tag{74}$$

$$\begin{aligned} & \sum_{n=1}^{N} \left( 2\mathbf{E}\left[ \|v_{n+1} - v_n^*(h)\|^2 \right] + 3\mathbf{E}\left[ \|x_{n+1} - x_n^*(h)\|^2 \right] \right) \\ \leq & \sum_{n=1}^{N} \left( O\left( h^4 \mathbf{E}\left[ \|v_n\|^2 \right] + u^2 h^4 \mathbf{E}\left[ \|\nabla f(x_n)\|^2 \right] + u^2 h^8 + u d h^5 + u^2 h^7 \kappa^{-1} \sigma^2 \right) \right. \\ & \left. O\left( h^6 \mathbf{E}\left[ \|v_n\|^2 \right] + u^2 h^4 \mathbf{E}\left[ \|\nabla f(x_n)\|^2 \right] + u^2 h^8 + u d h^7 + u^2 h^7 \kappa^{-1} \sigma^2 \right) \right) \\ \leq & O\left( h^4 \sum_{n=1}^{N} \mathbf{E}\left[ \|v_n\|^2 \right] + u^2 h^4 \sum_{n=1}^{N} \mathbf{E}\left[ \|\nabla f(x_n)\|^2 \right] + N u^2 h^8 + N u d h^5 + N u^2 h^7 \kappa^{-1} \sigma^2 \right) \end{aligned} \tag{75}$$

Combining (74), and (75), we have

$$\begin{aligned} & \sum_{n=1}^{N} \frac{2\kappa}{h} \left( 2\mathbf{E}\left[ \|\mathbf{E}_\alpha v_{n+1} - v_n^*(h)\|^2 \right] + 3\mathbf{E}\left[ \|\mathbf{E}_\alpha x_{n+1} - x_n^*(h)\|^2 \right] \right) \\ & + \sum_{n=1}^{N} \left( 2\mathbf{E}\left[ \|v_{n+1} - v_n^*(h)\|^2 \right] + 3\mathbf{E}\left[ \|x_{n+1} - x_n^*(h)\|^2 \right] \right) \\ \leq & O\left( (h^4 + \kappa h^7) \sum_{n=1}^{N} \mathbf{E}\left[ \|v_n\|^2 \right] + u^2(h^4 + \kappa h^9 + h^4) \sum_{n=1}^{N} \mathbf{E}\left[ \|\nabla f(x_n)\|^2 \right] \right. \\ & \left. + N u d h^5 + N u d(\kappa h^8 + h^7) + N u^2 h^4 \sigma^2 + N\kappa u^2 h^5 \right) \end{aligned} \tag{76}$$

From the proof of Theorem 3 of [SL19] we have,

$$\left\|\mathbf{E}\left[\nabla f(x_N)^\top v_N\right]\right\| \leq 4d + 6Mq_N$$

Then we have,

$$\sum_{n=0}^{N-1} \mathbf{E}\left[\|v_n\|^2\right] \leq O\left(q_N + Ndu + Nu^2h^3\kappa^{-1}\sigma^2 + Nu^2h^4\right) \tag{77}$$

and

$$\sum_{n=0}^{N-1} \mathbf{E}\left[\|\nabla f(x_n)\|^2\right] \leq O\left(\frac{dM}{h} + \frac{M^2}{h}q_N + MNd + Nh^3\kappa^{-1}\sigma^2 + Nh^4\right) \tag{78}$$

From (76), (77), and (78), and setting $N = \frac{2\kappa}{h}\log\left(\frac{20}{\epsilon^2}\right)$ we have

$$\begin{aligned}
&\sum_{n=1}^{N} \frac{2\kappa}{h}\left(2\mathbf{E}\left[\|\mathbf{E}_\alpha v_{n+1} - v_n^*(h)\|^2\right] + 3\mathbf{E}\left[\|\mathbf{E}_\alpha x_{n+1} - x_n^*(h)\|^2\right]\right)\\
&+\sum_{n=1}^{N}\left(2\mathbf{E}\left[\|v_{n+1} - v_n^*(h)\|^2\right] + 3\mathbf{E}\left[\|x_{n+1} - x_n^*(h)\|^2\right]\right)\\
&\leq O\left((h^3+\kappa h^7)q_N + Ndu(h^4+\kappa h^7) + duh^3 + duh^3 + Nu^2h^4\sigma^2 + Nu^2\kappa h^5\right)\\
&\leq O\left((h^3+\kappa h^7)q_N + \frac{d}{m}(h^3+\kappa h^6)\log\left(\frac{1}{\epsilon}\right) + \frac{h^3\kappa^{-1}}{m^2}\sigma^2\log\left(\frac{1}{\epsilon}\right) + \frac{h^4}{m^2}\log\left(\frac{1}{\epsilon}\right)\right)
\end{aligned}$$

From (72), and (73),

$$q_N \leq \frac{\epsilon^2 d}{4m} + O\left((h^3+\kappa h^7)q_N + \left(\frac{d}{m}(h^3+\kappa h^6) + \frac{h^3}{Mm}\sigma^2 + \frac{h^4}{m^2}\right)\log\left(\frac{1}{\epsilon}\right)\right)$$

Using, $(h^3+\kappa h^7) \leq 1/2$, we have,

$$\frac{q_N}{2} \leq \frac{\epsilon^2 d}{4m} + O\left(\left(\frac{d}{m}(h^3+\kappa h^6) + \frac{h^3}{Mm}\sigma^2 + \frac{h^4}{m^2}\right)\log\left(\frac{1}{\epsilon}\right)\right)$$

Choosing, $h = C\min\left(\frac{(\epsilon\sqrt{m})^{\frac{1}{3}}}{(d\kappa)^{\frac{1}{6}}\log\left(\frac{1}{\epsilon}\right)^{\frac{1}{6}}}, \min\left(\left(\frac{m}{d}\right)^{\frac{1}{3}}, \left(\frac{Mm}{16\sigma^2}\right)^{\frac{1}{3}}, \sqrt{m}\right)\epsilon^{\frac{2}{3}}\log\left(\frac{1}{\epsilon}\right)^{-\frac{2}{3}}\right)$, we get,

$$\mathbf{E}\left[\|x_N - y_N\|^2\right] \leq q_N \leq \frac{\epsilon^2 d}{m}$$

So, the iteration complexity is given by,

$$N = \tilde{O}\left(\max\left(\frac{d^{\frac{1}{6}}\kappa^{\frac{7}{6}}}{(\epsilon\sqrt{m})^{\frac{1}{3}}}, \frac{\kappa\max\left(\left(\frac{d}{m}\right)^{\frac{1}{3}}, \left(\frac{\sigma^2}{Mm}\right)^{\frac{1}{3}}, \frac{1}{\sqrt{m}}\right)}{\epsilon^{\frac{2}{3}}}\right)\right)$$

The total number of zeroth-order oracle calls are given by,

$$Nb = \tilde{O}\left(\max\left(\frac{d^{\frac{5}{3}}\kappa^{\frac{8}{3}}}{\epsilon^{\frac{4}{3}}}, \frac{d\kappa^2 \max\left(\left(\frac{d}{m}\right)^{\frac{1}{3}}, \left(\frac{\sigma^2}{Mm}\right)^{\frac{1}{3}}, \frac{1}{\sqrt{m}}\right)^4}{\epsilon^{\frac{8}{3}}}\right)\right)$$

■

# 9 Proofs for Section 3.1

*Proof.* [of Theorem 3.2] Let us define the following continuous time SDE with the initial point $\hat{x}_0$:

$$\hat{x}_t = -g_{\nu,b}(\hat{x}_0)dt + \sqrt{2}dW_n$$

Observe that $\hat{x}_h$ has the same distribution as $x_{n+1}$ when $\hat{x}_0 = x_n$. Let $z$ denote $(\{u_i\}_{i=1}^b, \{\xi_i\}_{i=1}^b)$. To show the dependence of $g_{\nu,b}(\hat{\hat{x}}_0)$ on $z$ we will use $g_{\nu,b}(\hat{x}_0, z)$ to denote $g_{\nu,b}(\hat{x}_0)$ just for this proof. Let $\rho_{t0z}(\hat{x}_t, \hat{x}_0, z)$ be the joint distribution of $\hat{x}_t$, $\hat{x}_0$, and $z$. Observe that conditioned on $\hat{x}_0$, and $z$, $g_{\nu,b}(\hat{x}_0, z)$ is deterministic. Then by Fokker-Plank equation, we have

$$\frac{\partial \rho_{t|0,z}(\hat{x}_t|\hat{x}_0, z)}{\partial t} = \nabla \cdot \left(\rho_{t|0,z}(\hat{x}_t|\hat{x}_0, z) g_{\nu,b}(\hat{\hat{x}}_0, z)\right) + \Delta \rho_{t|0,z}(\hat{x}_t|\hat{x}_0, z)$$

Then the time evolution of $\rho_t(x)$ is given by

$$\begin{aligned}
&\frac{\partial \rho_t(x)}{\partial t} = \mathbf{E}_{x_0,z}\left[\frac{\partial \rho_{t|0,z}(x|x_0, z)}{\partial t}\right] \\
\leq &\mathbf{E}_{x_0,z}\left[\nabla \cdot \left(\rho_{t|0,z}(\hat{x}_t|\hat{x}_0, z) g_{\nu,b}(\hat{\hat{x}}_0, z)\right) + \Delta \rho_{t|0,z}(\hat{x}_t|\hat{x}_0, z)\right] \\
= &\mathbf{E}_{x_0,z}\left[\nabla \cdot \left(\rho_{t|0,z}(\hat{x}_t|\hat{x}_0, z) g_{\nu,b}(\hat{\hat{x}}_0, z)\right)\right] + \Delta \rho_t(\hat{x}_t) \\
= &\int_{\mathbb{R}^d}\int_{\mathbb{R}^{2d}} \nabla \cdot \left(\rho_{t,0,z}(\hat{x}_t, \hat{x}_0, z) g_{\nu,b}(\hat{\hat{x}}_0, z)\right) dx_0 dz + \Delta \rho_t(\hat{x}_t) \\
= &\nabla \cdot \left(\rho_t(x) \mathbf{E}_{0,z|t}\left[g_{\nu,b}(\hat{x}_0, z)|\hat{x}_t = x\right]\right) + \Delta \rho_t(\hat{x}_t)
\end{aligned} \tag{79}$$

Now, as shown in [VW19] we have,

$$\frac{\partial H_\pi(\rho_t(x))}{\partial t} = \int_{\mathbb{R}^d} \frac{\partial \rho_t(x)}{\partial t} \log\left(\frac{\rho_t(x)}{\pi(x)}\right) dx$$

Then using (79), we have

$$\begin{aligned}
&\frac{\partial H_\pi(\rho_t(x))}{\partial t} \\
= &\int_{\mathbb{R}^d} \left(\nabla \cdot \left(\rho_t(x) \mathbf{E}_{0,z|t}\left[g_{\nu,b}(\hat{x}_0, z)|\hat{x}_t = x\right]\right) + \Delta \rho_t(\hat{x}_t)\right) \log\left(\frac{\rho_t(x)}{\pi(x)}\right) dx \\
= &\int_{\mathbb{R}^d} \nabla \cdot \left(\left(\rho_t(x) \mathbf{E}_{0,z|t}\left[g_{\nu,b}(\hat{x}_0, z)|\hat{x}_t = x\right]\right) + \nabla \rho_t(\hat{x}_t)\right) \log\left(\frac{\rho_t(x)}{\pi(x)}\right) dx
\end{aligned}$$

$$= \int_{\mathbb{R}^d} \nabla \cdot \left( \rho_t(x) \left( \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) + \mathbf{E}_{0,z|t} \left[ g_{\nu,b}(\hat{x}_0, z) | \hat{x}_t = x \right] - \nabla f(x) \right) \right) \log \left( \frac{\rho_t(x)}{\pi(x)} \right) dx$$

Now we use the fact that $\nabla \cdot (ax) = ax \cdot \nabla a + a \nabla \cdot x$ where $a$ is a scalar, and $x$ is a vector:

$$\begin{aligned}
&\frac{\partial H_\pi(\rho_t(x))}{\partial t} \\
&= \int_{\mathbb{R}^d} \nabla \cdot \left( \rho_t(x) \left( \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) + \mathbf{E}_{0,z|t} \left[ g_{\nu,b}(\hat{x}_0, z) | \hat{x}_t = x \right] - \nabla f(x) \right) \log \left( \frac{\rho_t(x)}{\pi(x)} \right) \right) dx \\
&- \int_{\mathbb{R}^d} \rho_t(x) \left\langle \left( \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) + \mathbf{E}_{0,z|t} \left[ g_{\nu,b}(\hat{x}_0, z) | \hat{x}_t = x \right] - \nabla f(x) \right), \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) \right\rangle dx
\end{aligned}$$

Now as $\rho_t(x) \left( \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) + \mathbf{E}_{0,z|t} \left[ g_{\nu,b}(\hat{x}_0, z) | \hat{x}_t = x \right] - \nabla f(x) \right) \log \left( \frac{\rho_t(x)}{\pi(x)} \right)$ decays to 0 as $x$ goes to infinity, we have,

$$\int_{\mathbb{R}^d} \nabla \cdot \left( \rho_t(x) \left( \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) + \mathbf{E}_{0,z|t} \left[ g_{\nu,b}(\hat{x}_0, z) | \hat{x}_t = x \right] - \nabla f(x) \right) \log \left( \frac{\rho_t(x)}{\pi(x)} \right) \right) dx = 0$$

Then we get,

$$\begin{aligned}
&\frac{\partial H_\pi(\rho_t(x))}{\partial t} \\
&= - \int_{\mathbb{R}^d} \rho_t(x) \left\langle \left( \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) + \mathbf{E}_{0,z|t} \left[ g_{\nu,b}(\hat{x}_0, z) | \hat{x}_t = x \right] - \nabla f(x) \right), \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) \right\rangle dx \\
&= - J_\pi(\rho_t(x)) - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^{2d}} \rho_t(x, \hat{x}_0, z) \left\langle g_{\nu,b}(\hat{x}_0, z) - \nabla f(x), \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) \right\rangle dz dx d\hat{x}_0 \\
&= - J_\pi(\rho_t(x)) + \mathbf{E}_{t0z} \left[ \left\langle \nabla f(\hat{x}_t) - g_{\nu,b}(\hat{x}_0, z), \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) \right\rangle \right] \qquad (80)
\end{aligned}$$

The second equality above follows from (26), and in the last line we have substituted $x_t$ in place of $x$. Now we will upper bound the second term above.

$$\begin{aligned}
&\mathbf{E}_{t0z} \left[ \left\langle \nabla f(\hat{x}_t) - g_{\nu,b}(\hat{x}_0, z), \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) \right\rangle \right] \\
\leq &\mathbf{E}_{t0z} \left[ \| \nabla f(\hat{x}_t) - g_{\nu,b}(\hat{x}_0, z) \|^2 \right] + \frac{1}{4} \mathbf{E}_{t0z} \left[ \left\| \nabla \log \left( \frac{\rho_t(x)}{\pi(x)} \right) \right\|^2 \right] \\
\leq &2M^2 \mathbf{E}_{t0} \left[ \| \hat{x}_t - \hat{x}_0 \|^2 \right] + 2\mathbf{E}_{0z} \left[ \| \nabla f(\hat{x}_0) - g_{\nu,b}(\hat{x}_0, z) \|^2 \right] + \frac{1}{4} J_\pi(\rho_t(x)) \qquad (81)
\end{aligned}$$

Now, from Lemma 1.1, we have,

$$\mathbf{E}_{0z} \left[ \| \nabla f(\hat{x}_0) - g_{\nu,b}(\hat{x}_0, z) \|^2 \right] \leq \frac{4(d+5) \mathbf{E}_{0z} \left[ \| \nabla f(\hat{x}_0) \|^2 \right]}{b} + C_1 \qquad (82)$$

where $C_1 = \frac{4(d+5)\sigma^2}{b} + \frac{3\nu^2 M^2 (d+3)^3}{2}$. We also have, with $\tau_0 \sim N(0, \mathbf{I_d})$

$$\mathbf{E}_{t0} \left[ \| \hat{x}_t - \hat{x}_0 \|^2 \right]$$

$$
\begin{aligned}
&= \mathbf{E}_{t0}\left[\| - t g_{\nu,b}(\hat{x}_0, z) + \sqrt{2t}\tau_0\|^2\right] \\
&\leq 2dt + 2t^2 \mathbf{E}_{0z}\left[\|\nabla f(\hat{x}_0) - g_{\nu,b}(\hat{x}_0, z)\|^2\right] + 2t^2 \mathbf{E}_0\left[\|\nabla f(\hat{x}_0)\|^2\right] \qquad (83)
\end{aligned}
$$

Combining (81), (82), and (83), for $t \leq 1/(2M)$ we get

$$
\begin{aligned}
&\mathbf{E}_{t0z}\left[\left\langle \nabla f(\hat{x}_t) - g_{\nu,b}(\hat{x}_0, z), \nabla \log\left(\frac{\rho_t(x)}{\pi(x)}\right)\right\rangle\right] \\
\leq &\frac{1}{4} J_\pi(\rho_t(x)) + 4M^2 td + (2 + 4M^2 t^2)\mathbf{E}_{0z}\left[\|\nabla f(\hat{x}_0) - g_{\nu,b}(\hat{x}_0, z)\|^2\right] + 4M^2 t^2 \mathbf{E}_0\left[\|\nabla f(\hat{x}_0)\|^2\right] \\
\leq &\frac{1}{4} J_\pi(\rho_t(x)) + 4M^2 td + 3\left(\frac{4(d+5)\mathbf{E}_0\left[\|\nabla f(\hat{x}_0)\|^2\right]}{b} + C_1\right) + 4M^2 t^2 \mathbf{E}_0\left[\|\nabla f(\hat{x}_0)\|^2\right] \\
\leq &\frac{1}{4} J_\pi(\rho_t(x)) + 4M^2 td + 3C_1 + \left(\frac{12(d+5)}{b} + 4M^2 t^2\right)\mathbf{E}_0\left[\|\nabla f(\hat{x}_0)\|^2\right] \\
\leq &\frac{1}{4} J_\pi(\rho_t(x)) + 4M^2 td + 3C_1 + \left(\frac{12(d+5)}{b} + 4M^2 t^2\right)\left(\frac{4M^2}{\lambda} H_\pi(\rho_0(x)) + 2Md\right) \qquad (84)
\end{aligned}
$$

We get the last inequality using Lemma 12 of [VW19]. Now combining, (80), and (84), we get,

$$
\begin{aligned}
&\frac{\partial H_\pi(\rho_t(x))}{\partial t} \\
&\leq -\frac{3}{4} J_\pi(\rho_t(x)) + 4M^2 td + 3C_1 + \left(\frac{12(d+5)}{b} + 4M^2 t^2\right)\left(\frac{4M^2}{\lambda} H_\pi(\rho_0(x)) + 2Md\right)
\end{aligned}
$$

Using (27), we get

$$
\begin{aligned}
&\frac{\partial H_\pi(\rho_t(x))}{\partial t} \\
&\leq -\frac{3\lambda}{2} H_\pi(\rho_t(x)) + 4M^2 td + 3C_1 + \left(\frac{12(d+5)}{b} + 4M^2 t^2\right)\left(\frac{4M^2}{\lambda} H_\pi(\rho_0(x)) + 2Md\right) \qquad (85)
\end{aligned}
$$

Taking $t \leq h$, we get,

$$
\begin{aligned}
&\frac{\partial H_\pi(\rho_t(x))}{\partial t} \\
&\leq -\frac{3\lambda}{2} H_\pi(\rho_t(x)) + 4M^2 hd + 3C_1 + \left(\frac{12(d+5)}{b} + 4M^2 h^2\right)\left(\frac{4M^2}{\lambda} H_\pi(\rho_0(x)) + 2Md\right)
\end{aligned}
$$

Multiplying both sides with $e^{\frac{3\lambda t}{2}}$, and integrating from $t = 0$ to $h$, we get

$$
\begin{aligned}
&e^{\frac{3\lambda h}{2}} H_\pi(\rho_h(x)) - H_\pi(\rho_0(x)) \\
\leq &\frac{2(e^{\frac{3\lambda h}{2}} - 1)}{3\lambda}\left(4M^2 hd + 3C_1 + \left(\frac{12(d+5)}{b} + 4M^2 h^2\right)\left(\frac{4M^2}{\lambda} H_\pi(\rho_0(x)) + 2Md\right)\right) \\
\leq &2h\left(4M^2 hd + 3C_1 + \left(\frac{12(d+5)}{b} + 4M^2 h^2\right)\left(\frac{4M^2}{\lambda} H_\pi(\rho_0(x)) + 2Md\right)\right) \\
= &\left(8M^2 h^2 d + 3hC_1 + \left(\frac{24(d+5)h}{b} + 8M^2 h^3\right) Md\right) + \frac{4M^2}{\lambda}\left(\frac{24(d+5)h}{b} + 8M^2 h^3\right) H_\pi(\rho_0(x))
\end{aligned}
$$

As in [VW19], in the penultimate step we use the fact $e^a \leq 1+2a$ for $0 < a = \frac{3\lambda h}{2}$, and $h \leq \frac{2}{3\lambda}$. Hence, we have

$$\begin{aligned} H_\pi(\rho_h(x)) \leq & e^{-\frac{3\lambda h}{2}} \left(1 + \frac{4M^2}{\lambda}\left(\frac{24(d+5)h}{b} + 8M^2h^3\right)\right) H_\pi(\rho_0(x)) \\ & + e^{-\frac{3\lambda h}{2}} \left(8M^2h^2d + 3hC_1 + \left(\frac{24(d+5)h}{b} + 8M^2h^3\right) Md\right). \end{aligned}$$

Choosing $b \geq \frac{384M^2(d+5)}{\lambda^2}$, and $h \leq \frac{\lambda}{12M^2}$, we get,

$$1 + \frac{4M^2}{\lambda}\left(\frac{24(d+5)h}{b} + 8M^2h^3\right) \leq 1 + \frac{\lambda h}{2} \leq e^{\frac{\lambda h}{2}}.$$

Then we have,

$$H_\pi(\rho_h(x)) \leq e^{-\lambda h} H_\pi(\rho_0(x)) + \left(8M^2h^2d + 3hC_1 + \left(\frac{24(d+5)h}{b} + 8M^2h^3\right) Md\right)$$

Observe that when $\hat{x}_0 = x_n$, $\rho_0$ is same as $\varpi_n$, and then $\rho_h(x)$ is same as $\varpi_{n+1}$. Then

$$\begin{aligned} & H_\pi(\varpi_{n+1}) \\ \leq & e^{-\lambda h} H_\pi(\varpi_n) + \left(8M^2h^2d + 3hC_1 + \left(\frac{24(d+5)h}{b} + 8M^2h^3\right) Md\right) \\ \leq & e^{-(n+1)\lambda h} H_\pi(\varpi_0) + \frac{1}{1-e^{-\lambda h}} \left(8M^2h^2d + 3hC_1 + \left(\frac{24(d+5)h}{b} + 8M^2h^3\right) Md\right) \end{aligned}$$

Choosing $n = N = \frac{1}{\lambda h} \log\left(\frac{\epsilon^2}{H_\pi(\varpi_0)}\right)$, and using $1 - e^{-\lambda h} \geq \frac{\lambda h}{2}$, for $h \leq \frac{1}{\lambda}$, we get

$$\begin{aligned} & H_\pi(\varpi_N) \\ \leq & \epsilon^2 + \left(\frac{16M^2hd}{\lambda} + \frac{6}{\lambda}\left(\frac{4(d+5)\sigma^2}{b} + \frac{3\nu^2M^2(d+3)^3}{2}\right) + \left(\frac{48(d+5)}{b} + 16M^2h^2\right)\frac{Md}{\lambda}\right) \end{aligned}$$

Choosing $b$, $\nu$, and $h$ as in (29), we get

$$H_\pi(\varpi_N) = O(\epsilon^2)$$

Using, (28), we get,

$$W_2(\varpi_N, \pi) = O(\epsilon)$$

■

# 10 Proofs for Section 4

*Proof.* [of Lemma 4.1] First note that, we have

$$g_{\nu,b}(\theta) - \nabla f_\nu(\theta) = \frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta + \nu u_i, \xi_i) - F(\theta, \xi_i')}{\nu} u_i - \nabla f_\nu(\theta)$$

$$=\frac{1}{b}\sum_{i=1}^{b}\frac{f(\theta+\nu u_i)-f(\theta)}{\nu}u_i-\nabla f_\nu(\theta)+\frac{1}{b}\sum_{i=1}^{b}\frac{\xi_i-\xi_i'}{\nu}u_i.$$

Hence, we have

$$\begin{aligned}\mathbf{E}\left[\|g_{\nu,b}(\theta)-\nabla f_\nu(\theta)\|^2\right]=&\mathbf{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\frac{f(x+\nu u_i)-f(\theta)}{\nu}u_i-\nabla f_\nu(\theta)\right\|^2\right]+\mathbf{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\frac{\xi_i-\xi_i'}{\nu}u_i\right\|^2\right]\\&+2\mathbf{E}\left[\left\langle\left(\frac{1}{b}\sum_{i=1}^{b}\frac{f(x+\nu u_i)-f(\theta)}{\nu}u_i-\nabla f_\nu(\theta)\right),\left(\frac{1}{b}\sum_{i=1}^{b}\frac{\xi_i-\xi_i'}{\nu}u_i\right)\right\rangle\right].\end{aligned}$$

Now note that, using independence of $\xi_i, \xi_i'$, and $u_i$, we have $\forall\ i$

$$\begin{aligned}&\mathbf{E}\left[\left\langle\left(\frac{f(x+\nu u_i)-f(\theta)}{\nu}u_i-\nabla f_\nu(\theta)\right),\left(\frac{\xi_i-\xi_i'}{\nu}u_i\right)\right\rangle\right]\\=&\mathbf{E}\left[\left\langle\frac{f(x+\nu u_i)-f(\theta)}{\nu}u_i-\nabla f_\nu(\theta),u_i\right\rangle\right]\mathbf{E}\left[\frac{\xi_i-\xi_i'}{\nu}\right]=0\end{aligned}$$

We also have, $\forall\ i\neq j$,

$$\begin{aligned}&\mathbf{E}\left[\left\langle\left(\frac{f(x+\nu u_i)-f(\theta)}{\nu}u_i-\nabla f_\nu(\theta)\right),\left(\frac{\xi_j-\xi_j'}{\nu}u_j\right)\right\rangle\right]\\=&\mathbf{E}\left[\left\langle\frac{f(x+\nu u_i)-f(\theta)}{\nu}u_i-\nabla f_\nu(\theta),u_j\right\rangle\right]\mathbf{E}\left[\frac{\xi_j-\xi_j'}{\nu}\right]=0\end{aligned}\tag{86}$$

Using, Lemma 1.1, we hence have,

$$\mathbf{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\frac{f(x+\nu u_i)-f(\theta)}{\nu}u_i-\nabla f_\nu(\theta)\right\|^2\right]\leq\frac{2(d+5)\|\nabla f(\theta)\|^2}{b}+\frac{\nu^2M^2(d+3)^3}{2b}.\tag{87}$$

Furthermore, we have

$$\mathbf{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\frac{\xi_i-\xi_i'}{\nu}u_i\right\|^2\right]=\frac{1}{b^2}\sum_{i=1}^{b}\mathbf{E}\left[\frac{(\xi_i-\xi_i')^2}{\nu^2}\right]\mathbf{E}\left[\|u_i\|^2\right]=\frac{2d\sigma^2}{b\nu^2}.\tag{88}$$

Combining, (87), (86), and (88), we obtain Lemma 4.1. ∎

*Proof.* [of Theorem 4.2] Using Lemma 4.1, (45) changes to,

$$\begin{aligned}W_2(\varpi_n,\pi)\leq&(1-0.5mh)^nW_2(\varpi_0,\pi)+\frac{3.3M\sqrt{hd}}{m}+\frac{2\nu M\sqrt{d}}{m}+\frac{\nu M\sqrt{h}}{2\sqrt{mb}}(d+3)^{\frac{3}{2}}\\&+\frac{3\sqrt{h(d+5)(\frac{\sigma^2}{\nu^2}+2Md)}}{\sqrt{mb}}.\end{aligned}$$

Now the last term involves $\nu$ in the denominator. To counter the effect we have to increase the sample size to $b=\frac{d}{\epsilon^2}$. ∎

*Proof.* [of Lemma 4.2] First note that, we have

$$\begin{aligned} g_{\nu,b}(\theta) - \nabla f_\nu(\theta) =& \frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta+\nu u_i, \xi_i) - F(\theta, \xi_i')}{\nu} u_i(\theta) - \nabla f_\nu(\theta) \\ =& \frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta+\nu u_i, \xi_i) - F(\theta, \xi_i)}{\nu} u_i - \nabla f_\nu(\theta) + \frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta, \xi_i) - F(\theta, \xi_i')}{\nu} u_i \end{aligned}$$

Hence, we have

$$\begin{aligned} \mathbf{E}\left[\|g_{\nu,b}(\theta) - \nabla f_\nu(\theta)\|^2\right] =& 2\mathbf{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta+\nu u_i, \xi_i) - F(\theta, \xi_i)}{\nu} u_i - \nabla f_\nu(\theta)\right\|^2\right] \\ &+ 2\mathbf{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta, \xi_i) - F(\theta, \xi_i')}{\nu} u_i\right\|^2\right]. \end{aligned}$$

Using, Lemma 1.1, we have,

$$\mathbf{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta+\nu u_i, \xi_i) - F(\theta, \xi_i)}{\nu} u_i - \nabla f_\nu(\theta)\right\|^2\right] \leq \frac{2(d+5)(\|\nabla f(\theta)\|^2 + \sigma^2)}{b} + \frac{\nu^2 M^2 (d+3)^3}{2b}. \tag{89}$$

Furthermore, note that

$$\begin{aligned} &\mathbf{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b} \frac{F(\theta, \xi_i) - F(\theta, \xi_i')}{\nu} u_i\right\|^2\right] = \frac{1}{b^2}\sum_{i=1}^{b} \mathbf{E}\left[\frac{(F(\theta, \xi_i) - F(\theta, \xi_i'))^2}{\nu^2}\right] \mathbf{E}\left[\|u_i\|^2\right] \\ \leq& \frac{L^2}{b^2}\sum_{i=1}^{b} \mathbf{E}\left[\frac{(\xi_i - \xi_i')^2}{\nu^2}\right] \mathbf{E}\left[\|u_i\|^2\right] = \frac{2dL^2\sigma^2}{b\nu^2}. \end{aligned} \tag{90}$$

Combining, (89), and (90), we get the result stated in Lemma 4.1. ∎

# 11 Proofs for Section 5

*Proof.* [of Theorem 5.3] First, we have,

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

where $a' = a - M\nu\sqrt{s} \le a - \|\nabla f(\theta) - \nabla f_\nu(\theta)\|$ 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

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

where $\|\cdot\|_{\Psi_1} = \sup_{p\ge1} p^{-1}\mathbf{E}[|\cdot|^p]^{1/p}$ and $\|\cdot\|_{\Psi_2} = \sup_{p\ge1} p^{-1/2}\mathbf{E}[|\cdot|^p]^{1/p}$ are the sub-exponential and sub-Gaussian norm respectively (see, for example [Ver18] for more details). In the last inequality we require that $\nu \le \frac{R}{M\|u\|_{\Psi_2}}$. 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 5.1. Moreover, we have the following estimate.

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

which implies that $\|\phi(\nu,u)\|_{\Psi_1} \le 2RC_2\sqrt{s}$, $\|u_1\|_{\Psi_2} \le C_1$. We now state the following concentration inequality proved in [BFY18].

**Lemma 11.1** *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 $\|X\|_{\psi_2} \le \Upsilon_1$, and $Y$ be a sub-exponential random variable with $\|Y\|_{\psi_1} \le \Upsilon_2$ for some constants $\Upsilon_1$ and $\Upsilon_2$. Then for any $t \ge K \cdot \max\{\Upsilon_1^3, \Upsilon_1\} \cdot \Upsilon_2$, we have*

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

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

From Lemma 11.1, for $n \ge \max\left\{K_1 \frac{2RC\sqrt{s}}{\tau}, \left(\frac{2RC\sqrt{s}}{\tau}\right)^4\right\}$, we have:

$$\Pr\{|\zeta_j| \ge \tau\} = \Pr\left\{\left|\frac{1}{n}\sum_{k=1}^n g_{\nu,1}^k - \mathbf{E}[g_{\nu,1}]\right| \ge \tau\right\}$$

$$\leq 4\exp\left\{-K_2\left(\frac{n\tau}{\|\phi(\nu,u)\|_{\Psi_1}\|u_1\|_{\Psi_2}}\right)^{2/3}\right\}$$
$$\leq 4\exp\left\{-K_2\left(\frac{n\tau}{2RC\sqrt{s}}\right)^{2/3}\right\},$$

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

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

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

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

■