arXiv:2511.04413v2 [math.NA] 10 Sep 2026

Mean Square Error Analysis of Stochastic Gradient and Variance-Reduced Sampling Algorithms

Jianfeng Lu JIANFENG@MATH.DUKE.EDU
Departments of Mathematics, Physics, and Chemistry
Duke University
Durham, NC 27708, USA

Xuda Ye YE481@PURDUE.EDU
Department of Mathematics
Purdue University
West Lafayette, IN 47907, USA

Zhennan Zhou ZHOUZHENNAN@WESTLAKE.EDU.CN
Institute for Theoretical Sciences
Westlake University
Hangzhou, 310030, China

Abstract

This paper considers mean square error (MSE) analysis for stochastic gradient sampling algorithms applied to underdamped Langevin dynamics under a global convexity assumption. A novel discrete Poisson equation framework is developed to bound the time-averaged sampling error. For the Stochastic Gradient UBU (SG-UBU) sampler, we derive an explicit MSE bound and establish that the numerical bias exhibits first-order convergence with respect to the step size h, with the leading error coefficient proportional to the variance of the stochastic gradient. The analysis is further extended to variance-reduced algorithms for finite-sum potentials, specifically the SVRG-UBU and SAGA-UBU methods. For these algorithms, we identify a phase transition phenomenon whereby the convergence rate of the numerical bias shifts from first to second order as the step size decreases below a critical threshold. Theoretical findings are validated by numerical experiments. In addition, we compare the computational cost that mini-batch SG-UBU and SVRG-UBU require to reach a prescribed accuracy, which indicates when variance reduction is worthwhile.

Keywords: mean square error, stochastic gradient sampling, discrete Poisson equation, variance reduction

1 Introduction

Sampling from complicated probability distributions is a fundamental task in computational physics and data sciences. In many problems of interest, the potential function (i.e., the negative log-density or log-likelihood) has a natural finite-sum structure. A direct example comes from Bayesian inference: the posterior distribution accumulates log-likelihood contributions from a large dataset (Gelman et al., 1995). This finite-sum structure motivates stochastic gradient sampling methods, pioneered by Stochastic Gradient Langevin Dynamics (SGLD) and related MCMC algorithms (Welling and Teh, 2011; Chen et al., 2014). At each step, a mini-batch gradient, an unbiased estimator computed from a small subset of the data, provides an approximation of the exact gradient. Because of the unbiasedness, the er-

©2025 Jianfeng Lu, Xuda Ye, and Zhennan Zhou.
License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/.

rors introduced into the drift term average out over long simulation horizons; consequently, as the integrator step size vanishes, the dynamics converge to the correct target distribution and yield reliable samples of the posterior. Building on these properties, SGLD has found wide applications in Bayesian deep learning and uncertainty quantification (Welling and Teh, 2011; Li et al., 2016); scalable latent-variable models such as topic models and probabilistic matrix factorization (Patterson and Teh, 2013; Chen et al., 2014); and online or streaming Bayesian inference for large datasets (Ahn et al., 2012).

The necessity of stochastic gradients for sampling extends beyond the data sub-sampling context to problems where the target log-probability gradient is intrinsically stochastic. This situation frequently arises in Bayesian inference for complex generative models, particularly those defined implicitly through stochastic simulators, which render the likelihood function intractable (Cranmer et al., 2020). In such models, common across fields such as systems biology, physics, and econometrics, the gradient of the log-likelihood cannot be computed analytically and must instead be approximated via Monte Carlo estimation, often through repeated simulator runs. As a result, any gradient-based sampler targeting the posterior must operate with noisy gradient estimates. This challenge lies at the core of likelihood-free inference (also known as Approximate Bayesian Computation) and pseudo-marginal methods, where the central task is to perform posterior sampling in the presence of intrinsically noisy gradient information (Andrieu and Roberts, 2009; Meeds and Welling, 2014).

We formulate the stochastic gradient sampling problem as follows. Let $\pi(\boldsymbol{x}) \propto e^{-U(\boldsymbol{x})}$ be the target distribution in $\mathbb{R}^d$. The full gradient $\nabla U(\boldsymbol{x})$ is assumed to be computationally intractable in $\mathcal{O}(1)$ time. Instead, we assume cheap access to a stochastic gradient $b(\boldsymbol{x}, \theta)$ for some random variable $\theta \sim \mathbb{P}_\theta$, and $b(\boldsymbol{x}, \theta)$ is an unbiased estimator of the full gradient:

$$\mathbb{E}_\theta\big[b(\boldsymbol{x}, \theta)\big] = \nabla U(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Two common forms of the unbiased stochastic gradient include:

Additive noise form. The gradient estimator is given by

$$b(\boldsymbol{x}, \theta) = \nabla U(\boldsymbol{x}) + \sigma\theta, \qquad \theta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), \tag{1}$$

with $\sigma > 0$ being a constant noise level.

Finite-sum form. This setting assumes that the potential is an average of N components,

$$U(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} U_i(\boldsymbol{x}), \tag{2}$$

where the stochastic gradient is a mini-batch estimate over a random subset $\theta \subset \{1, \dots, N\}$ of size p, with the batch size p being a key tunable parameter:

$$b(\boldsymbol{x}, \theta) = \frac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}). \tag{3}$$

A stochastic gradient sampling algorithm generates a sequence of samples $(\boldsymbol{X}_k)_{k=0}^{\infty}$ by applying a numerical integrator with a fixed step size $h > 0$, driven by the stochastic gradient

$b(\boldsymbol{x}, \theta)$. For a sufficiently small step size, the distribution of the iterate $\boldsymbol{X}_k$ is expected to provide a close approximation to the target distribution $\pi(\boldsymbol{x})$ as $k \to \infty$.

This paper provides a rigorous analysis of the error for stochastic gradient sampling algorithms. The analysis of such methods is considerably more challenging than their full gradient counterparts, as it must account for the intricate interplay between the noise from the stochastic gradients and the systematic error from the numerical discretization. We focus on a practical error metric by first specifying a test function $f(\boldsymbol{x})$ and the target statistical average $\pi(f) := \int_{\mathbb{R}^d} f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. For the numerical solution $(\boldsymbol{X}_k)_{k=0}^{\infty}$, the sampling error after K steps is quantified by the mean square error (MSE):

$$\mathrm{MSE}_K = \mathbb{E}\left[\left(\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f)\right)^2\right]. \tag{4}$$

Unlike metrics that measure distributional discrepancy (e.g., Wasserstein distance or Kullback–Leibler (KL) divergence), the MSE characterizes the error and variance of the time average obtained from a single numerical trajectory, which more accurately reflects the practice.

The MSE can be conceptually decomposed into three primary sources: (i) the *stochastic fluctuation* arising from the finite sample size, (ii) the *stochastic gradient error* introduced by gradient noise, and (iii) the *discretization error* originating from the numerical integrator. The stochastic fluctuation component is an intrinsic feature of Monte Carlo sampling algorithms and exhibits a conventional decay rate of $\mathcal{O}(1/K)$ as the number of sampling steps K approaches infinity. The discretization error can, in principle, be systematically reduced by employing higher-order integrators. For underdamped Langevin dynamics, second-order schemes like BAOAB (Leimkuhler and Matthews, 2015) and UBU (Sanz-Serna and Zygalakis, 2021), and third-order schemes like the randomized midpoint integrator (Shen and Lee, 2019) and QUICSORT (Scott et al., 2025), are designed for this purpose. However, analyzing these advanced schemes in the stochastic gradient setting is non-trivial. In fact, many existing results for similar algorithms only establish a suboptimal half-order convergence rate in h (Gouraud et al., 2025; Guillin et al., 2024). Our analysis uses the second-order *strong* accuracy of UBU, which allows the local error to be split into a martingale difference and a remainder controlled by the second derivatives of ϕ_h. We expect analogous results for BAOAB, but an integrator lacking this property would require higher-order expansions of ϕ_h, and we do not treat general families of integrators here.

To reduce the stochastic gradient error component of the MSE, variance reduction techniques come into play. By reducing the variance of the stochastic gradient, these techniques have been widely applied in optimization to achieve faster convergence (Bubeck et al., 2015; Allen-Zhu, 2018; Lan et al., 2019; Song et al., 2020). The success of variance reduction in optimization motivates a natural question for sampling: Under what conditions, if any, do these techniques offer a genuine advantage? While methods like SVRG (Johnson and Zhang, 2013; Reddi et al., 2016; Allen-Zhu and Yuan, 2016) and SAGA (Defazio et al., 2014; Chatterji et al., 2018) are well-established for accelerating optimization, their role in sampling is fundamentally different. Optimization seeks an efficient path to a single minimum, whereas sampling must generate a trajectory that characterizes an entire probability distribution. This latter objective requires balancing the *discretization error*, governed by the step size h, against the *stochastic error* originating from gradient noise. Since variance

reduction techniques are designed to mitigate only the stochastic component, their interaction with the discretization error is critical for the total MSE. A key contribution of this paper is to quantify this interplay and establish a practical criterion for the advantageous use of variance-reduced samplers.

1.1 Our contributions

This paper introduces a novel discrete Poisson equation for the MSE analysis of stochastic gradient sampling algorithms. We demonstrate the power of this approach by applying it to the Stochastic Gradient UBU (SG-UBU) sampler (Sanz-Serna and Zygalakis, 2021)—a second-order accurate integrator for underdamped Langevin dynamics—and its variance-reduced variants. Our main contributions are as follows:

- We propose a general discrete Poisson equation framework, extending the ideas in Mattingly et al. (2010), to analyze the MSE for a broad class of numerical integrators. A key advantage of this framework is its modularity: an MSE bound for any specific integrator can be derived by establishing its uniform-in-time moment bounds (stability) and its local error estimate (consistency).
- Applying this framework to SG-UBU, we prove that the numerical bias in computing the statistical average $\pi(f)$ has first-order convergence in the step size h and explicitly show that its leading-order term is proportional to the stochastic gradient variance.
- For the variance-reduced variants, SVRG-UBU and SAGA-UBU, we identify a phase transition: the numerical bias shifts from first to second-order convergence as the step size h decreases below a critical threshold. This accelerated convergence is a unique feature of variance reduction when combined with high-order integrators.
- We derive a practical criterion to guide the choice between standard and variance-reduced samplers, based on the computational cost each requires to reach a prescribed accuracy.

1.2 Related works

The error analysis for stochastic gradient sampling algorithms, which typically discretize either overdamped or underdamped Langevin dynamics, is well studied in the literature. The underdamped approach is often preferred for its computational efficiency, as it exhibits faster mixing times (Cao et al., 2023; Eberle and Lörler, 2024; Altschuler and Chewi, 2024; Altschuler et al., 2025) and permits the use of higher-order integrators that achieve accelerated convergence with respect to step size h. However, this efficiency is coupled with greater analytical challenges due to the hypoellipticity of the underdamped dynamics.

In the following, we summarize four primary analytical frameworks for stochastic gradient samplers, with results covering both overdamped and underdamped Langevin dynamics. These frameworks vary in their methodology and applicability. The most direct approach combines an integrator's recurrence relations with strong or weak error analysis; this is effective for simple schemes like SGLD (Nagapetyan et al., 2017; Dalalyan and Karagulyan, 2019; Brosse et al., 2018; Zou et al., 2019; Cheng et al., 2020; Chada et al., 2023; Guillin et al., 2024; Shaw and Whalley, 2025), but is not suitable for higher-order integrators. A

second approach uses coupling arguments to establish global contractivity in a probability metric, which guarantees a unique invariant distribution but often yields suboptimal convergence rates such as $\mathcal{O}(\sqrt{h})$ (Chak and Monmarché, 2025; Leimkuhler et al., 2024; Schuh and Whalley, 2024; Gouraud et al., 2025). Alternatively, the entropy approach analyzes the associated Fokker–Planck equation to bound the KL divergence; however, its complexity typically limits its use to simple integrators (Li and Wang, 2022; Kinoshita and Suzuki, 2022; Suzuki et al., 2023). Finally, the continuous Poisson equation is a powerful tool for analyzing the MSE, but its application to underdamped systems is hampered by the significant challenge of obtaining the necessary derivative bounds on its solution, especially without assuming convexity (Mattingly et al., 2010; Ross, 2011; Chen et al., 2015; Vollmer et al., 2016; Teh et al., 2016; Xu et al., 2018; Brosse et al., 2019; Chen et al., 2023).

Our approach is based on the discrete Poisson equation, tailored for the direct analysis of numerical integrators. Compared to the continuous Poisson equation, this approach aims at analyzing the one step transition kernel of the numerical scheme instead of the infinitesimal generator of the continuous process. The analysis results in a simple and exact decomposition of the time average error into stochastic gradient, discretization, and martingale components. This decomposition provides a direct path for a sharp MSE analysis, and as in the continuous case, requires establishing bounds on the derivatives of the solution to the Poisson equation. We establish such bounds up to the second order under a global convexity assumption on the potential function $U(\boldsymbol{x})$. The idea of using the discrete Poisson equation draws strong inspiration from early studies (Kontoyiannis and Meyn, 2003; Glynn and Meyn, 1996; Johndrow and Mattingly, 2017). Nevertheless, the error analyses in these previous works mainly utilize Harris' ergodic theorem (Meyn and Tweedie, 2012), whereas our work combines the discrete Poisson equation with local strong error analysis. Working with the one-step transition kernel rather than the generator also lowers the regularity required: only the first and second derivatives of ϕ_h enter our estimates, whereas analyses based on the continuous Poisson equation typically require derivatives of ϕ_0 up to fourth order (Brosse et al., 2019, Lemma 17). Global convexity is used at a single point, in the coupling contractivity of the first variation process (Lemma 17), which in turn drives the Hessian bound of Theorem 1.

Denoting by π_h the invariant distribution of the numerical solution with the step size h, we may quantify the difference between π_h and the target distribution π using a distribution metric, such as Wasserstein or KL divergence, thereby characterizing the effects of stochastic gradient error and discretization error. For overdamped Langevin dynamics, both π_h and π are supported in $\mathbb{R}^d$. For underdamped Langevin dynamics, π_h is a distribution over both position and velocity variables, and is thus supported in $\mathbb{R}^{2d}$. In this case, the difference is measured between π_h and the Gaussian velocity-augmented distribution of π. By setting a test function f supported in $\mathbb{R}^d$, the difference between π_h and π can also be quantified by the numerical bias

$$\pi_h(f) - \pi(f) = \int_{\mathbb{R}^{2d}} f(\boldsymbol{x}) \Big(\pi_h(\boldsymbol{x}, \boldsymbol{v}) - \big(\pi \otimes \mathcal{N}(\boldsymbol{0}, M_2^{-1} \boldsymbol{I}_d)\big)(\boldsymbol{x}, \boldsymbol{v}) \Big) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{v}, \tag{5}$$

whose square is the long-time limit ($K \to \infty$) of the MSE defined in (4).

To situate our contributions within the existing literature, Table 1 summarizes key error analysis results for the stochastic gradient sampling algorithms in overdamped and

underdamped Langevin dynamics. In the table, the second column indicates whether the global convexity of the potential $U(\boldsymbol{x})$ is assumed, with $m > 0$ as the convexity lower bound (see Assumption (P)). The constant C may depend on the spatial dimension d or the convexity bound m, indicated by subscripts, while subscript $*$ indicates dependence on additional parameters for non-convex cases.

Reference	Convex	Integrator	Result
Theorem 9 (Vollmer et al., 2016)	–	SGLD	$\mathcal{W}_1(\pi_h, \pi) \leqslant C_{d,*} h$
Theorem 5 (Dalalyan and Karagulyan, 2019)	Req.	SGLD	$\mathcal{W}_2(\pi_h, \pi) \leqslant C\left(\frac{dh}{m} + \sqrt{\frac{dh}{m}}\right)$
Corollary 3.1 (Li and Wang, 2022)	–	SGLD	$\mathsf{KL}(\pi_h \Vert \pi) \leqslant C_* dh^2$
Theorem 3.3 (Chen et al., 2023)	Req.	SGLD	$\mathcal{W}_1(\pi_h, \pi) \leqslant C_m d(1 + \log \lvert h \rvert)h$
Theorem 3.9 (Shaw and Whalley, 2025)	Req.	SGLD-RR	$\mathcal{W}_2(\pi_h, \pi) \leqslant \frac{Cdh}{m}$
Theorem 3.8 (Zou et al., 2019)	–	SG-EM	$\mathcal{W}_2(\pi_h, \pi) \leqslant C_{d,*} h^{\frac{3}{4}}$
Theorem 1 (Guillin et al., 2024)	Req.	SG-EM	$\mathcal{W}_1(\pi_h, \pi) \leqslant C_m \sqrt{dh}$
Proposition 18 (Gouraud et al., 2025)	Req.	SGgHMC	$\mathcal{W}_2(\pi_h, \pi) \leqslant C_m \sqrt{dh}$
Theorem 9 **(this paper)**	Req.	SG-UBU	$\lvert \pi_h(f) - \pi(f) \rvert \leqslant \frac{Cdh}{m^2}$

Table 1: Numerical bias results for stochastic gradient sampling algorithms. The upper block collects results based on the overdamped Langevin dynamics, and the lower block those based on the underdamped Langevin dynamics. In the "Convex" column, Req. marks a result established under the global convexity of $U(\boldsymbol{x})$, and – marks a result established without it. The constant C may depend on d and m, as indicated by subscripts; subscript $*$ denotes dependence on additional parameters in non-convex cases.

The underdamped results in Table 1 highlight a key contribution of our work. While existing analyses for underdamped samplers often yield suboptimal convergence rates such as $\mathcal{O}(\sqrt{h})$, our analysis demonstrates that the SG-UBU integrator achieves first-order convergence of the numerical bias $\pi_h(f) - \pi(f)$. We further show that the leading-order error coefficient is determined by the stochastic gradient variance and is, remarkably, unaffected by the discretization error.

In the case of finite-sum potentials (2), error analysis for variance-reduced sampling algorithms presents additional challenges. The inherent complexity of methods such as SVRG and SAGA means that the numerical solution is typically no longer a Markov chain. This structural limitation obstructs standard analytical tools such as global contractivity

proofs, and as a result, the literature mostly uses recurrence relations for error analysis (Nagapetyan et al., 2017; Zou et al., 2018b; Chatterji et al., 2018; Baker et al., 2019; Zou et al., 2018a, 2019; Chen et al., 2022). While alternative methods using the continuous Poisson equation (Dubey et al., 2016; Li et al., 2019) or entropy approaches (Kinoshita and Suzuki, 2022) show theoretical promise for non-convex potentials, their extensions to the underdamped case are less explored.

Table 2 summarizes existing error analysis results for these algorithms. The parameter N denotes the number of components in the finite-sum potential defined in (2), and p is the batch size for the mini-batch gradient approximation. Due to the non-Markovian nature of these methods, the numerical bias is measured with respect to $\tilde{\pi}_h$, which represents the long-time limit of the averaged distributions over the trajectory. For a fair comparison across different works, the results cited are normalized such that the assumed upper bound on the Hessian matrix $\|\nabla^2 U(\boldsymbol{x})\|$ is independent of N. For algorithms utilizing an anchor position, the update frequency is set to N/p iterations in the mini-batch setting.

Reference	Convex	Integrator	Result
Theorem 3.5 (Zou et al., 2018b)	Req.	SVRG-LD	$\mathcal{W}_2(\tilde{\pi}_h, \pi) \leqslant C\left(\frac{dh}{m} + \sqrt{\frac{dh}{mp}} \min\left\{1, \sqrt{\frac{Nh}{mp}}\right\}\right)$
Theorem 4.2 (Chatterji et al., 2018)	Req.	SVRG-LD	$\mathcal{W}_2(\tilde{\pi}_h, \pi) \leqslant C\left(\frac{dh}{m} + \sqrt{\frac{dN}{mp^3}} h\right)$
Theorem 4.1 (Chatterji et al., 2018)	Req.	SAGA-LD	$\mathcal{W}_2(\tilde{\pi}_h, \pi) \leqslant C\left(\frac{dh}{m} + \sqrt{\frac{dN}{mp}} h\right)$
Theorem 2.5 (Chen et al., 2022)	–	SVRG-LD	$\mathcal{W}_1(\tilde{\pi}_h, \pi) \leqslant C_{d,*} h^{\frac{1}{4}}$
Theorem 1 (Kinoshita and Suzuki, 2022)	–	SVRG-LD	$\mathsf{KL}(\tilde{\pi}_h \| \pi) \leqslant C_{d,*}\left(1 + \frac{N}{p^2}\right) h$
Theorem 2 (Kinoshita and Suzuki, 2022)	–	SARAH-LD	$\mathsf{KL}(\tilde{\pi}_h \| \pi) \leqslant C_{d,*}\left(1 + \frac{N}{p^2}\right) h$
Theorem 3.3 (Zou et al., 2019)	–	SRVR-HMC	$\mathcal{W}_2(\tilde{\pi}_h, \pi) \leqslant C_{d,*}\left(1 + \frac{N}{p^2}\right)^{\frac{1}{4}} h^{\frac{3}{4}}$
Theorem 12 **(this paper)**	Req.	SVRG-UBU	$\lvert\tilde{\pi}_h(f) - \pi(f)\rvert \leqslant C\left(\frac{dh}{m^2 p} \min\left\{1, \frac{N^2 h^2}{mp^2}\right\} + \frac{dh^2}{m^3}\right)$
Theorem 14 **(this paper)**	Req.	SAGA-UBU	$\lvert\tilde{\pi}_h(f) - \pi(f)\rvert \leqslant C\left(\frac{dh}{m^3} \min\left\{1, N^2 h^2\right\} + \frac{dh^2}{m^3}\right)$

Table 2: Numerical bias results for variance-reduced sampling algorithms. The upper block collects results based on the overdamped Langevin dynamics, and the lower block those based on the underdamped Langevin dynamics. In the "Convex" column, Req. marks a result established under the global convexity of $U(\boldsymbol{x})$, and – marks a result established without it. Results that do not explicitly depend on the batch size p are for the $p = 1$ case. The constant C may depend on d, as indicated by the subscript; subscript $*$ denotes dependence on additional parameters in non-convex cases.

As shown in Table 2, our analysis of SVRG-UBU and SAGA-UBU uncovers a phase transition phenomenon for the rate of convergence. The convergence rate with respect to the step size h transitions from first-order to an accelerated second-order rate as h decreases below a critical threshold. This relationship appears to be a unique feature of combining variance reduction with second-order discretization schemes, a property fulfilled by the UBU-type integrators studied in this work.

The remainder of this paper is organized as follows. Section 2 introduces our notation and the discrete Poisson equation framework. In Section 3, we present our main results for the SG-UBU integrator, followed by the analysis of its variance-reduced counterparts, SVRG-UBU and SAGA-UBU, in Section 4. We provide numerical validation in Section 5 and conclude in Section 6.

2 Preliminaries and methodology

2.1 Notations and assumptions

Throughout the paper, we use $|\cdot|$ to denote the Euclidean norm for vectors. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}$, we use $\|\boldsymbol{A}\|$ for the operator 2-norm (or spectral norm), defined as

$$\|\boldsymbol{A}\| := \sup_{|\boldsymbol{x}|=1, \boldsymbol{x} \in \mathbb{R}^{n_2}} |\boldsymbol{A}\boldsymbol{x}|.$$

This definition extends to a third-order tensor $\boldsymbol{B} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, whose norm is defined via the corresponding multilinear form:

$$\|\boldsymbol{B}\| := \sup_{|\boldsymbol{x}|=1, |\boldsymbol{y}|=1, |\boldsymbol{z}|=1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} B_{ijk} x_i y_j z_k.$$

We adopt standard typesetting conventions where vectors and matrices are denoted by bold fonts (e.g., $\boldsymbol{x}, \boldsymbol{A}$), while scalars are in normal font (e.g., h, σ). The zero matrix and the identity matrix in $\mathbb{R}^{d \times d}$ are indicated by $\boldsymbol{O}_d$ and $\boldsymbol{I}_d$, respectively. For symmetric matrices in $\mathbb{R}^{d \times d}$, we write $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite. Furthermore, we distinguish continuous-time dynamics from their numerical approximations by using lowercase and uppercase letters, respectively. Specifically, the continuous solution to the underdamped Langevin dynamics is denoted by $(\boldsymbol{x}_t, \boldsymbol{v}_t)_{t \geqslant 0}$. Its numerical approximation, generated by an integrator such as SG-UBU with step size h, is denoted by the discrete sequence $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$, where $(\boldsymbol{X}_k, \boldsymbol{V}_k)$ approximates the continuous solution at time $t = kh$.

Our analysis relies on the following assumptions regarding the potential function $U(\boldsymbol{x})$, the stochastic gradient $b(\boldsymbol{x}, \theta)$, and the test function $f(\boldsymbol{x})$.

Assumption (P) *The potential function $U(\boldsymbol{x}) \in C^3(\mathbb{R}^d)$ satisfies $\nabla U(\boldsymbol{0}) = \boldsymbol{0}$ and*

$$m\boldsymbol{I}_d \preccurlyeq \nabla^2 U(\boldsymbol{x}) \preccurlyeq M_2 \boldsymbol{I}_d, \quad \|\nabla^3 U(\boldsymbol{x})\| \leqslant M_3, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

for constants $m, M_2, M_3 > 0$.

Assumption (Sg) *The stochastic gradient $b(\boldsymbol{x}, \theta)$ satisfies*

$$\mathbb{E}_\theta\big[b(\boldsymbol{x}, \theta)\big] = \nabla U(\boldsymbol{x}), \quad \Big(\mathbb{E}_\theta\big[|b(\boldsymbol{x}, \theta) - \nabla U(\boldsymbol{x})|^8\big]\Big)^{\frac{1}{4}} \leqslant \sigma^2 d, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

for the constant $\sigma > 0$, and for almost all $\theta \sim \mathbb{P}_\theta$ there holds

$$|b(\boldsymbol{x}, \theta)| \leqslant M_1 \sqrt{|\boldsymbol{x}|^2 + d}, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

for the constant $M_1 > 0$. Assume $\sigma \leqslant M_1$ for the convenience of analysis.

In the case of finite-sum potentials, Assumption (Sg) is specified as follows.

Assumption (SgN) *For the finite-sum potential $U(\boldsymbol{x})$, the gradients $\{\nabla U_i(\boldsymbol{x})\}_{i=1}^N$ satisfy*

$$\left(\frac{1}{N} \sum_{i=1}^N |\nabla U_i(\boldsymbol{x}) - \nabla U(\boldsymbol{x})|^8 \right)^{\frac{1}{4}} \leqslant \sigma^2 d, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

for the constant $\sigma > 0$, and for all $i \in \{1, \cdots, N\}$ there holds

$$|\nabla U_i(\boldsymbol{x})| \leqslant M_1 \sqrt{|\boldsymbol{x}|^2 + d}, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

for the constant $M_1 > 0$. Assume $\sigma \leqslant M_1$ for the convenience of analysis.

Assumption (T) *The test function $f(\boldsymbol{x}) \in C^2(\mathbb{R}^d)$ satisfies*

$$|\nabla f(\boldsymbol{x})| \leqslant L_1, \quad \left\|\nabla^2 f(\boldsymbol{x})\right\| \leqslant L_2, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

for constants $L_1, L_2 > 0$.

The subscripts on the constants M_i and L_i correspond to the order of the associated derivatives. These assumptions provide several key properties. First, Assumption (P) guarantees $m \leqslant M_2$, and $U(\boldsymbol{x})$ is strongly convex and has a unique minimizer at $\mathbf{0}$. Second, the growth condition and unbiasedness in Assumption (Sg) together imply a linear growth bound on the true gradient, $|\nabla U(\boldsymbol{x})| \leqslant M_1 \sqrt{|\boldsymbol{x}|^2 + d}$. Finally, the eighth moment condition in Assumption (Sg) ensures the necessary regularity for a rigorous MSE analysis.

2.2 Underdamped Langevin dynamics and SG-UBU integrator

To sample from the target distribution $\pi(\boldsymbol{x}) \propto e^{-U(\boldsymbol{x})}$, we consider the discretization of the underdamped Langevin dynamics. Given the Hessian bound $\|\nabla^2 U(\boldsymbol{x})\| \leqslant M_2$, consider the Hamiltonian system

$$H(\boldsymbol{x}, \boldsymbol{v}) = U(\boldsymbol{x}) + \frac{M_2}{2}|\boldsymbol{v}|^2, \quad \boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^d,$$

where the auxiliary velocity variable $\boldsymbol{v}$ extends the state space to $\mathbb{R}^{2d}$.

The underdamped Langevin dynamics for $H(\boldsymbol{x}, \boldsymbol{v})$ evolves according to the following stochastic differential equation:

$$\mathfrak{L} : \begin{cases} \dot{\boldsymbol{x}}_t = \boldsymbol{v}_t, \\ \dot{\boldsymbol{v}}_t = -\dfrac{1}{M_2} \nabla U(\boldsymbol{x}_t) - 2\boldsymbol{v}_t + \dfrac{2}{\sqrt{M_2}} \dot{\boldsymbol{B}}_t, \end{cases} \tag{6}$$

where $\boldsymbol{B}_t$ is the standard Brownian motion in $\mathbb{R}^d$. We choose the damping coefficient as 2 for optimal dissipation characteristics, and the gradient scaling as $1/M_2$ to enable stable numerical integration with $\mathcal{O}(1)$ step sizes.

Next we describe the UBU-type numerical integrators for the underdamped Langevin dynamics (6), including the full gradient and stochastic gradient versions. Applying the Lie–Trotter splitting method, we decompose (6) into two separate dynamics

$$\mathfrak{U}:\begin{cases}\dot{\boldsymbol{x}}_t=\boldsymbol{v}_t,\\ \dot{\boldsymbol{v}}_t=-2\boldsymbol{v}_t+\dfrac{2}{\sqrt{M_2}}\dot{\boldsymbol{B}}_t,\end{cases}\qquad \mathfrak{B}:\begin{cases}\dot{\boldsymbol{x}}_t=\boldsymbol{0},\\ \dot{\boldsymbol{v}}_t=-\dfrac{1}{M_2}\nabla U(\boldsymbol{x}_t),\end{cases}$$

where $\mathfrak{U}$ denotes the linear part and $\mathfrak{B}$ represents a free particle driven by the potential gradient. Their solution flows, denoted by $\Phi_t^{\mathfrak{U}}$ and $\Phi_t^{\mathfrak{B}}$, respectively, are given by

$$\Phi_t^{\mathfrak{U}}:\begin{cases}\boldsymbol{x}_t=\boldsymbol{x}_0+\dfrac{1-e^{-2t}}{2}\boldsymbol{v}_0+\dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^t\frac{1-e^{-2(t-s)}}{2}\mathrm{d}\boldsymbol{B}_s,\\ \boldsymbol{v}_t=e^{-2t}\boldsymbol{v}_0+\dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^t e^{-2(t-s)}\mathrm{d}\boldsymbol{B}_s,\end{cases}\qquad \Phi_t^{\mathfrak{B}}:\begin{cases}\boldsymbol{x}_t=\boldsymbol{x}_0,\\ \boldsymbol{v}_t=\boldsymbol{v}_0-\dfrac{t}{M_2}\nabla U(\boldsymbol{x}_0).\end{cases}$$

The Full Gradient UBU (FG-UBU) integrator implements the following composition:

$$(\boldsymbol{X}_{k+1},\boldsymbol{V}_{k+1})=\big(\Phi_{h/2}^{\mathfrak{U}}\circ\Phi_h^{\mathfrak{B}}\circ\Phi_{h/2}^{\mathfrak{U}}\big)(\boldsymbol{X}_k,\boldsymbol{V}_k),\qquad k=0,1,\cdots. \tag{7}$$

Correspondingly, the Stochastic Gradient UBU (SG-UBU) integrator is constructed by replacing $\nabla U(\boldsymbol{x})$ in FG-UBU with $b(\boldsymbol{x},\theta)$. Define the perturbed solution flow corresponding to the stochastic gradient $b(\boldsymbol{x},\theta)$ by

$$\Phi_t^{\mathfrak{B}}(\theta):\begin{cases}\boldsymbol{x}_t=\boldsymbol{x}_0,\\ \boldsymbol{v}_t=\boldsymbol{v}_0-\dfrac{t}{M_2}b(\boldsymbol{x}_0,\theta).\end{cases}$$

Then the SG-UBU integrator with the step size h becomes

$$(\boldsymbol{X}_{k+1},\boldsymbol{V}_{k+1})=\big(\Phi_{h/2}^{\mathfrak{U}}\circ\Phi_h^{\mathfrak{B}}(\theta_k)\circ\Phi_{h/2}^{\mathfrak{U}}\big)(\boldsymbol{X}_k,\boldsymbol{V}_k),\qquad k=0,1,\cdots, \tag{8}$$

where $(\theta_k)_{k=0}^{\infty}$ are independent samples from $\mathbb{P}_\theta$. Equivalently, we can produce the numerical solution of SG-UBU $(\boldsymbol{X}_k,\boldsymbol{V}_k)_{k=0}^{\infty}$ through Algorithm 1. Note that in SG-UBU, the stochastic gradient is evaluated at the intermediate position $\boldsymbol{Y}_k$, which is the position component of $\Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{X}_k,\boldsymbol{V}_k)$, rather than at $\boldsymbol{X}_k$ itself. Specifically, $\boldsymbol{Y}_k$ is explicitly expressed as

$$\boldsymbol{Y}_k=\boldsymbol{X}_k+\frac{1-e^{-h}}{2}\boldsymbol{V}_k+\frac{2}{\sqrt{M_2}}\int_0^{h/2}\frac{1-e^{-(h-2s)}}{2}\mathrm{d}\boldsymbol{B}_{kh+s}. \tag{9}$$

For the finite-sum potential defined in (2), the SG-UBU integrator can be efficiently implemented using a mini-batch approach. The procedure is detailed in Algorithm 2. The computational cost for each iteration of Algorithm 2 is $\mathcal{O}(p)$, corresponding to the evaluation of p individual gradients that form the mini-batch.

Algorithm 1: Stochastic Gradient UBU (SG-UBU)

Input: initial state $(\boldsymbol{X}_0, \boldsymbol{V}_0)$, step size h
Output: numerical solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$
for $k = 0, 1, \cdots$ **do**
  Evolve $(\boldsymbol{Y}_k, \boldsymbol{V}_k^{(1)}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{X}_k, \boldsymbol{V}_k)$
  Sample a random index $\theta_k \sim \mathbb{P}_\theta$
  Compute the stochastic gradient $\boldsymbol{b}_k = b(\boldsymbol{Y}_k, \theta_k)$
  Update the velocity $\boldsymbol{V}_k^{(2)} = \boldsymbol{V}_k^{(1)} - \frac{h}{M_2}\boldsymbol{b}_k$
  Evolve $(\boldsymbol{X}_{k+1}, \boldsymbol{V}_{k+1}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{Y}_k, \boldsymbol{V}_k^{(2)})$
end

Algorithm 2: Mini-batch Stochastic Gradient UBU (mini-batch SG-UBU)

Input: initial state $(\boldsymbol{X}_0, \boldsymbol{V}_0)$, batch size p, step size h
Output: numerical solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$
for $k = 0, 1, \cdots$ **do**
  Evolve $(\boldsymbol{Y}_k, \boldsymbol{V}_k^{(1)}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{X}_k, \boldsymbol{V}_k)$.
  Sample a random subset $\theta_k \subset \{1, \cdots, N\}$ of size p
  Compute the mini-batch gradient $\boldsymbol{b}_k = \frac{1}{p}\sum_{i\in\theta_k} \nabla U_i(\boldsymbol{Y}_k)$
  Update the velocity $\boldsymbol{V}_k^{(2)} = \boldsymbol{V}_k^{(1)} - \frac{h}{M_2}\boldsymbol{b}_k$
  Evolve $(\boldsymbol{X}_{k+1}, \boldsymbol{V}_{k+1}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{Y}_k, \boldsymbol{V}_k^{(2)})$
end

2.3 The discrete Poisson equation framework

The discrete Poisson equation provides an analytical framework for studying the time average error of numerical solutions to Langevin dynamics. For the underdamped Langevin dynamics (6), the infinitesimal generator is given by

$$\mathcal{L} = \boldsymbol{v}^\top \nabla_{\boldsymbol{x}} - \left(\frac{1}{M_2}\nabla U(\boldsymbol{x}) + 2\boldsymbol{v}\right)^\top \nabla_{\boldsymbol{v}} + \frac{2}{M_2}\Delta_{\boldsymbol{v}}.$$

The discrete Poisson equation with respect to this $\mathcal{L}$ and fixed step size $h > 0$ is defined as

$$\frac{1 - e^{h\mathcal{L}}}{h}\phi_h(\boldsymbol{x}, \boldsymbol{v}) = f(\boldsymbol{x}) - \pi(f), \tag{10}$$

where $\phi_h(\boldsymbol{x}, \boldsymbol{v})$ is known as the discrete Poisson solution.

To facilitate error analysis, we introduce a more tractable representation of the discrete Poisson solution $\phi_h(\boldsymbol{x}, \boldsymbol{v})$. First, we define the function

$$u(\boldsymbol{x}, \boldsymbol{v}, t) = \mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})}[f(\boldsymbol{x}_t)] - \pi(f), \tag{11}$$

where the superscript $(\boldsymbol{x}, \boldsymbol{v})$ denotes the initial state for the exact solution $(\boldsymbol{x}_t, \boldsymbol{v}_t)_{t\geqslant 0}$. Using the semigroup operator, this is equivalent to

$$u(\boldsymbol{x}, \boldsymbol{v}, t) = (e^{t\mathcal{L}} f)(\boldsymbol{x}, \boldsymbol{v}) - \pi(f),$$

which leads to a series representation for the discrete Poisson solution:

$$\phi_h(\boldsymbol{x}, \boldsymbol{v}) = h \sum_{k=0}^{\infty} u(\boldsymbol{x}, \boldsymbol{v}, kh). \tag{12}$$

In fact, one can conveniently verify that $\phi_h(\boldsymbol{x}, \boldsymbol{v})$, as defined in (12), satisfies

$$\phi_h(\boldsymbol{x}, \boldsymbol{v}) = h \sum_{k=0}^{\infty} \big((e^{kh\mathcal{L}} f)(\boldsymbol{x}, \boldsymbol{v}) - \pi(f)\big) = h(1 - e^{h\mathcal{L}})^{-1}\big(f(\boldsymbol{x}) - \pi(f)\big),$$

which shows that $\phi_h(\boldsymbol{x}, \boldsymbol{v})$ solves the discrete Poisson equation (10).

As $h \to 0$, the discrete Poisson solution $\phi_h(\boldsymbol{x}, \boldsymbol{v})$ in (12) converges to the integral defining the continuous Poisson solution:

$$\phi_0(\boldsymbol{x}, \boldsymbol{v}) = \int_0^{\infty} u(\boldsymbol{x}, \boldsymbol{v}, t) \mathrm{d}t, \tag{13}$$

which satisfies the conventional Poisson equation $-(\mathcal{L}\phi_0)(\boldsymbol{x}, \boldsymbol{v}) = f(\boldsymbol{x}) - \pi(f)$.

The following theorem establishes global bounds on the gradient and Hessian matrix of the discrete Poisson solution $\phi_h(\boldsymbol{x}, \boldsymbol{v})$.

Theorem 1 *Under Assumptions* (P) *and* (T)*, let the step size $h \leqslant \frac{1}{4}$; then the discrete Poisson solution $\phi_h(\boldsymbol{x}, \boldsymbol{v})$ has a uniformly bounded gradient and Hessian matrix:*

$$\max\big\{|\nabla_{\boldsymbol{x}} \phi_h|, |\nabla_{\boldsymbol{v}} \phi_h|\big\} \leqslant \frac{C}{m},$$
$$\max\big\{\big\|\nabla^2_{\boldsymbol{xx}} \phi_h\big\|, \big\|\nabla^2_{\boldsymbol{xv}} \phi_h\big\|, \big\|\nabla^2_{\boldsymbol{vv}} \phi_h\big\|\big\} \leqslant \frac{C}{m^2},$$

where the constant C depends only on M_2, M_3 and $(L_i)_{i=1}^2$.

The proof of Theorem 1 is given in Appendix A.

Next, we detail how the discrete Poisson equation framework is used to analyze the time average error of the numerical solutions. Given the step size $h > 0$, let $\boldsymbol{Z}_k = (\boldsymbol{X}_k, \boldsymbol{V}_k)$ be the numerical solution of SG-UBU at the k-th step. For each $k = 0, 1, 2, \cdots$, we define the corresponding full-gradient update $\bar{\boldsymbol{Z}}_{k+1} = (\bar{\boldsymbol{X}}_{k+1}, \bar{\boldsymbol{V}}_{k+1})$ as

$$(\bar{\boldsymbol{X}}_{k+1}, \bar{\boldsymbol{V}}_{k+1}) = \big(\Phi^{\mathfrak{U}}_{h/2} \circ \Phi^{\mathfrak{B}}_h \circ \Phi^{\mathfrak{U}}_{h/2}\big)(\boldsymbol{X}_k, \boldsymbol{V}_k), \tag{14}$$

and the exact one-step solution of the Langevin dynamics $\boldsymbol{Z}_k(h) = (\boldsymbol{X}_k(h), \boldsymbol{V}_k(h))$ as

$$(\boldsymbol{X}_k(h), \boldsymbol{V}_k(h)) = \Phi^{\mathfrak{L}}_h(\boldsymbol{X}_k, \boldsymbol{V}_k), \tag{15}$$

where $\Phi^{\mathfrak{L}}_h$ is the solution flow of (6). This allows for a natural decomposition of the one-step evolution of the numerical solution:

$$\boldsymbol{Z}_{k+1} - \boldsymbol{Z}_k = \big(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}\big) + \big(\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)\big) + \big(\boldsymbol{Z}_k(h) - \boldsymbol{Z}_k\big). \tag{16}$$

The three terms on the right-hand side isolate the fundamental sources of error: the first is the unbiased *stochastic gradient error*, the second is the *discretization error* of the FG-UBU integrator, and the third corresponds to the evolution of the exact dynamics, which can be analyzed using its martingale properties.

Applying the discrete Poisson solution $\phi_h(\boldsymbol{x}, \boldsymbol{v})$ to the one-step evolution (16) yields the key error decomposition identity:

$$\frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\boldsymbol{Z}_k)}{h} + f(\boldsymbol{X}_k) - \pi(f) = R_k + S_k + T_k, \tag{17}$$

where the error terms R_k, S_k and T_k are defined by

$$R_k = \frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\bar{\boldsymbol{Z}}_{k+1})}{h}, \qquad S_k = \frac{\phi_h(\bar{\boldsymbol{Z}}_{k+1}) - \phi_h(\boldsymbol{Z}_k(h))}{h},$$
$$T_k = \frac{\phi_h(\boldsymbol{Z}_k(h)) - \phi_h(\boldsymbol{Z}_k)}{h} + f(\boldsymbol{X}_k) - \pi(f).$$

Here, R_k and S_k correspond to the stochastic gradient error and the discretization error, respectively. The third term, T_k, forms a martingale difference sequence with respect to the filtration

$$\mathcal{F}_k = \sigma\big(\boldsymbol{Z}_0, \theta_0, \cdots, \theta_{k-1}, (\boldsymbol{B}_t)_{0 \leqslant t \leqslant kh}\big),$$

representing the history of the algorithm on the time interval $[0, kh]$. This property follows directly from the discrete Poisson equation (10), which ensures that the conditional expectation of T_k is zero:

$$\mathbb{E}\big[T_k \big| \mathcal{F}_k\big] = \frac{(e^{h\mathcal{L}}\phi_h)(\boldsymbol{Z}_k) - \phi_h(\boldsymbol{Z}_k)}{h} + f(\boldsymbol{X}_k) - \pi(f) = 0.$$

The same filtration serves the other two error terms: $\boldsymbol{Z}_k$ is $\mathcal{F}_k$-measurable, while the random index θ_k and the Brownian increment on $[kh, (k+1)h]$ are both independent of $\mathcal{F}_k$.

Finally, summing the identity (17) from $k = 0$ to $K - 1$ establishes a fundamental expression for the time average error:

$$\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f) = \frac{\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_K)}{Kh} + \frac{1}{K}\sum_{k=0}^{K-1}(R_k + S_k + T_k). \tag{18}$$

This representation allows the MSE to be bounded by analyzing the second moments of the cumulative error terms. The analysis is simplified by the orthogonality of the martingale difference sequence $(T_k)_{k=0}^{K-1}$, which implies that its variance is additive:

$$\mathbb{E}\left[\Big(\sum_{k=0}^{K-1} T_k\Big)^2\right] = \sum_{k=0}^{K-1} \mathbb{E}\big[T_k^2\big].$$

Therefore, the MSE estimation reduces to bounding the accumulated stochastic gradient errors (R_k) and discretization errors (S_k), while the martingale terms (T_k) contribute only through the sum of their individual variances.

3 Mean square error and numerical bias of SG-UBU

This section establishes a quantitative estimate for the MSE of SG-UBU by analyzing the second moment of the time average error in (18). Our proof combines two essential components, corresponding to stability and consistency in the standard numerical analysis methodology. We first establish stability by deriving uniform-in-time moment bounds for the SG-UBU solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$. We then analyze consistency by estimating the moments of the local error, which is decomposed into the stochastic gradient error $\boldsymbol{Z}_{k+1}-\bar{\boldsymbol{Z}}_{k+1}$ and the discretization error $\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)$. This decomposition reveals that the stochastic gradient error contributes the dominant first-order term in the step size h, while the discretization error contributes a higher second-order term. This leads to a first-order numerical bias, whose leading-order coefficient is proportional to the variance of the stochastic gradient.

3.1 Stability: Uniform-in-time moment bound

We establish a uniform-in-time eighth moment bound for the SG-UBU solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$. This high-order moment is required to control the corresponding moments of the local error terms in our subsequent analysis. The proof hinges on a key simplification: we first analyze the related SG-BU integrator, whose update rule is given by

$$(\boldsymbol{X}_{k+1}, \boldsymbol{V}_{k+1}) = \big(\Phi_h^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\theta_k)\big)(\boldsymbol{X}_k, \boldsymbol{V}_k), \qquad k = 0, 1, \cdots . \tag{19}$$

Since the SG-UBU integrator (8) differs from SG-BU only by a stable $\Phi_{h/2}^{\mathfrak{U}}$ mapping, their uniform-in-time moment bounds are equivalent (Schuh and Whalley, 2024). This approach is advantageous because the SG-BU scheme evaluates the stochastic gradient at the state $\boldsymbol{X}_k$ itself, making the analysis more tractable than a direct treatment of SG-UBU.

To establish the uniform-in-time moment estimate for SG-BU, we employ a Lyapunov function approach. Thanks to the global convexity of the potential function $U(\boldsymbol{x})$, we define the Lyapunov function based on the quadratic form in Mattingly et al. (2002):

$$\mathcal{V}(\boldsymbol{x}, \boldsymbol{v}) = \big(2|\boldsymbol{x}|^2 + 2\boldsymbol{x}^\top \boldsymbol{v} + |\boldsymbol{v}|^2\big)^4, \qquad \boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{v}) \in \mathbb{R}^{2d},$$

where the fourth power is chosen to control the eighth moment of the numerical solution.

For SG-BU, this choice of $\mathcal{V}$ yields the following one-step contractive property.

Lemma 2 *Under Assumptions* (P) *and* (Sg)*, let the step size $h \leqslant \frac{1}{4}$; for any $\boldsymbol{z}_0 \in \mathbb{R}^{2d}$,*

$$\mathbb{E}\Big[\mathcal{V}\big((\Phi_h^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\theta))\boldsymbol{z}_0\big)\Big] \leqslant \left(1 - \frac{mh}{3M_2}\right)\mathcal{V}(\boldsymbol{z}_0) + \frac{Cd^4h}{m^3},$$

where $\theta \sim \mathbb{P}_\theta$ and the constant C depends only on M_1, M_2.

This Lyapunov condition for the SG-BU integrator provides the foundation for bounding the moments of the desired SG-UBU solution.

Theorem 3 *Under Assumptions* (P) *and* (Sg)*, let the step size $h \leqslant \frac{1}{4}$, and the initial state*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the SG-UBU solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ *satisfies*

$$\sup_{k\geqslant 0} \mathbb{E}\big[|\boldsymbol{Z}_k|^8\big] \leqslant \frac{Cd^4}{m^4},$$

where the constant C *depends only on* M_1, M_2.

The order of parameters d and m in Theorem 3 is optimal, and the equality holds if the target distribution $\pi(\boldsymbol{x})$ is exactly the Gaussian distribution $\mathcal{N}(\boldsymbol{0}, m^{-1}\boldsymbol{I}_d)$.

As a corollary, this uniform-in-time moment bound also applies to the numerical invariant distribution of the SG-UBU integrator, whose existence is guaranteed by the Krylov–Bogolyubov theorem (Hairer, 2006).

Lemma 4 *Under Assumptions* (P) *and* (Sg)*, let the step size* $h \leqslant \frac{1}{4}$*; then SG-UBU has an invariant distribution* $\pi_h(\boldsymbol{x}, \boldsymbol{v})$ *in* $\mathbb{R}^{2d}$*, which satisfies*

$$\int_{\mathbb{R}^{2d}} |\boldsymbol{z}|^8 \pi_h(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \leqslant \frac{Cd^4}{m^4},$$

where the constant C *depends only on* M_1, M_2.

The proofs of the results above are given in Appendix B.

3.2 Consistency: Local error analysis

Building on the one-step error decomposition in (16), we establish fourth moment bounds for both the stochastic gradient error $\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}$ and the discretization error $\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)$. These estimates are a prerequisite for deriving the MSE of SG-UBU. The explicit integral formulas for $\boldsymbol{Z}_{k+1}$, $\bar{\boldsymbol{Z}}_{k+1}$, and $\boldsymbol{Z}_k(h)$ used in this analysis are detailed in Appendix C.

The stochastic gradient error, $\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}$, has the concise form:

$$\begin{cases} \boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1} = -\dfrac{h(1-e^{-h})}{2M_2}(b(\boldsymbol{Y}_k, \theta_k) - \nabla U(\boldsymbol{Y}_k)), \\ \boldsymbol{V}_{k+1} - \bar{\boldsymbol{V}}_{k+1} = -\dfrac{he^{-h}}{M_2}(b(\boldsymbol{Y}_k, \theta_k) - \nabla U(\boldsymbol{Y}_k)), \end{cases} \tag{20}$$

where the gradient is evaluated at the intermediate position $\boldsymbol{Y}_k$ from (9), which is independent of the random variable θ_k. This independence, combined with the unbiased nature of the stochastic gradient, yields the conditional unbiasedness property:

$$\mathbb{E}\big[\boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1}\big|\boldsymbol{Y}_k\big] = \mathbb{E}\big[\boldsymbol{V}_{k+1} - \bar{\boldsymbol{V}}_{k+1}\big|\boldsymbol{Y}_k\big] = \boldsymbol{0}. \tag{21}$$

This confirms that the error sequence $\{\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}\}_{k=0}^{\infty}$ is a martingale difference sequence. The following lemma provides bounds on its fourth moments.

Lemma 5 *Under Assumptions* (P) *and* (Sg)*, let the step size* $h \leqslant \frac{1}{4}$*, and the initial state*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the stochastic gradient error of SG-UBU satisfies

$$\sup_{k\geqslant 0} \sqrt{\mathbb{E}\big|\boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1}\big|^4} \leqslant C\sigma^2 dh^4, \quad \sup_{k\geqslant 0} \sqrt{\mathbb{E}\big|\boldsymbol{V}_{k+1} - \bar{\boldsymbol{V}}_{k+1}\big|^4} \leqslant C\sigma^2 dh^2,$$

where the constant C *depends only on* M_1, M_2.

The analysis of the discretization error, $\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)$, follows the methodology in Theorem 25 of Sanz-Serna and Zygalakis (2021), which employs a martingale decomposition technique. For notational convenience, we define the position and velocity errors as

$$\delta \boldsymbol{X}_k = \bar{\boldsymbol{X}}_{k+1} - \boldsymbol{X}_k(h), \qquad \delta \boldsymbol{V}_k = \bar{\boldsymbol{V}}_{k+1} - \boldsymbol{V}_k(h).$$

This technique decomposes the velocity error $\delta \boldsymbol{V}_k$ into a martingale difference component $\delta_m \boldsymbol{V}_k$ and a higher-order remainder $\delta_h \boldsymbol{V}_k$. The martingale component satisfies the conditional unbiasedness property:

$$\mathbb{E}\big[\delta_m \boldsymbol{V}_k \big| \boldsymbol{Z}_k\big] = \boldsymbol{0}.$$

The following lemma provides fourth moment bounds for each of these error components.

Lemma 6 *Under Assumptions* (P) *and* (Sg)*, let the step size* $h \leqslant \frac{1}{4}$*, and the initial state*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1} \boldsymbol{I}_d);$$

then the discretization error of SG-UBU satisfies

$$\sup_{k \geqslant 0} \sqrt{\mathbb{E}|\delta \boldsymbol{X}_k|^4} \leqslant \frac{Cdh^6}{m}, \quad \sup_{k \geqslant 0} \sqrt{\mathbb{E}|\delta_m \boldsymbol{V}_k|^4} \leqslant Cdh^5, \quad \sup_{k \geqslant 0} \sqrt{\mathbb{E}|\delta_h \boldsymbol{V}_k|^4} \leqslant \frac{Cd^2h^6}{m^2},$$

where the constant C *depends only on* $(M_i)_{i=1}^3$*.*

The proofs of the results above are given in Appendix C.

3.3 Mean square error of SG-UBU

Integrating the stability and consistency results from the preceding sections, we now establish the MSE bound for the SG-UBU integrator. The proof leverages the time average error decomposition (18), which is built upon the discrete Poisson solution $\phi_h(\boldsymbol{x}, \boldsymbol{v})$.

Theorem 7 *Under Assumptions* (P), (Sg) *and* (T)*, let the step size* $h \leqslant \frac{1}{4}$*, and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1} \boldsymbol{I}_d);$$

then the SG-UBU solution satisfies

$$\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f)\right)^2\right] \leqslant C\left(\frac{d}{m^3 K h} + \frac{\sigma^4 d^2 h^2}{m^4} + \frac{d^2 h^4}{m^5}\left(1 + \frac{dh}{m^2}\right)\right),$$

where the constant C *depends only on* $(M_i)_{i=1}^3$ *and* $(L_i)_{i=1}^2$*.*

Sketch of Proof The proof proceeds by bounding the second moments of the cumulative error terms in the time average expression (18):

$$\mathbb{E}\left[\left(\sum_{k=0}^{K-1} R_k\right)^2\right], \quad \mathbb{E}\left[\left(\sum_{k=0}^{K-1} S_k\right)^2\right], \quad \mathbb{E}\left[\left(\sum_{k=0}^{K-1} T_k\right)^2\right].$$

The core technique is to decompose each error term into a principal martingale difference component and a higher-order remainder. The leading-order martingale components for the stochastic error R_k and the discretization error S_k are, respectively,

$$\hat{R}_k = \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \nabla \phi_h(\bar{\boldsymbol{Z}}_{k+1}), \qquad \hat{S}_k = \frac{1}{h}\delta_m \boldsymbol{V}_k^\top \nabla_{\boldsymbol{v}} \phi_h(\boldsymbol{Z}_k).$$

This decomposition allows for a sharp estimation of the second moments. The orthogonality of the martingale sequences ensures that their cumulative variance is the sum of the individual variances, simplifying the analysis and leading to the final MSE bound. ■

Theorem 7 decomposes the MSE into three principal components:

- **Stochastic fluctuation:** Decays with simulation time Kh at a rate of $\mathcal{O}\big(\frac{d}{m^3 Kh}\big)$.
- **Stochastic gradient error:** The first squared-bias component, scaling as $\mathcal{O}\big(\frac{\sigma^4 d^2 h^2}{m^4}\big)$; the bias itself exhibits first-order convergence with respect to the step size h.
- **Discretization error:** The second squared-bias component, scaling as $\mathcal{O}\big(\frac{d^2 h^4}{m^5}(1+\frac{dh}{m^2})\big)$; the bias itself exhibits second-order convergence with respect to the step size h.

In the SG-UBU setting, the numerical bias is dominated by the stochastic gradient error.

The MSE provides a distinct characterization of numerical accuracy compared to metrics such as the Wasserstein distance. While the Wasserstein distance measures the discrepancy between the numerical and target distributions, the MSE quantifies the error and variance of the time-averaged estimate from a single trajectory. As this aligns more closely with how sampling algorithms are used in practice, the MSE often serves as a more direct measure of the computational effort required to achieve a desired level of accuracy.

The $\mathcal{O}(d/(m^3 Kh))$ term in the MSE is mainly driven by the martingale component $\sum_{k=0}^{K-1} T_k$. Its asymptotic behavior can be understood by applying the Itô formula to the discrete Poisson solution. For a sufficiently small step size h, we have

$$\mathrm{d}\phi_h(\boldsymbol{Z}_k(s)) = (\mathcal{L}\phi_h)(\boldsymbol{Z}_k(s))\mathrm{d}s + \frac{2}{\sqrt{M_2}}\nabla_{\boldsymbol{v}}\phi_h(\boldsymbol{Z}_k(s)) \cdot \mathrm{d}\boldsymbol{B}_{kh+s},$$

and integrating over $s \in [0, h]$ yields

$$\begin{aligned}\frac{\phi_h(\boldsymbol{Z}_k(h)) - \phi_h(\boldsymbol{Z}_k)}{h} &\approx (\mathcal{L}\phi_0)(\boldsymbol{Z}_k) + \frac{2}{\sqrt{M_2}h}\int_0^h \nabla_{\boldsymbol{v}}\phi_0(\boldsymbol{Z}_k(s)) \cdot \mathrm{d}\boldsymbol{B}_{kh+s} \\ &= -f(\boldsymbol{X}_k) + \pi(f) + \frac{2}{\sqrt{M_2}h}\int_0^h \nabla_{\boldsymbol{v}}\phi_0(\boldsymbol{Z}_k(s)) \cdot \mathrm{d}\boldsymbol{B}_{kh+s}.\end{aligned}$$

Summing from $k = 0$ to $K - 1$, we find the continuous-time integral limit

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1} T_k &= \frac{1}{K}\sum_{k=0}^{K-1}\left(\frac{\phi_h(\boldsymbol{Z}_k(h)) - \phi_h(\boldsymbol{Z}_k)}{h} + f(\boldsymbol{X}_k) - \pi(f)\right) \\ &\approx \frac{2}{\sqrt{M_2}Kh}\sum_{k=0}^{K-1}\int_0^h \nabla_{\boldsymbol{v}}\phi_0(\boldsymbol{Z}_k(s)) \cdot \mathrm{d}\boldsymbol{B}_{kh+s} \approx \frac{2}{\sqrt{M_2}T}\int_0^T \nabla_{\boldsymbol{v}}\phi_0(\boldsymbol{Z}_0(t)) \cdot \mathrm{d}\boldsymbol{B}_t,\end{aligned}$$

where we identify the total sampling time as $T = Kh$. Consequently, the variance from this stochastic fluctuation is principally governed by

$$\frac{4}{M_2 T^2} \int_0^T \mathbb{E}\big[|\nabla_{\boldsymbol{v}} \phi_0(\boldsymbol{Z}_0(t))|^2\big] \mathrm{d}t,$$

which is independent of the step size h. This reveals a key insight: this component of the sampling error is an intrinsic feature of the underlying underdamped Langevin dynamics (6) and is unaffected by the numerical discretization or the variance of the stochastic gradient.

This framework also provides the MSE estimate for the mini-batch SG-UBU algorithm when applied to the finite-sum potential in (2), as detailed in the following result:

Theorem 8 *Under Assumptions* (P), (SgN) *and* (T), *let the step size* $h \leqslant \frac{1}{4}$, *and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1} \boldsymbol{I}_d);$$

then the mini-batch SG-UBU solution with the batch size p *satisfies*

$$\mathbb{E}\bigg[\bigg(\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f)\bigg)^2\bigg] \leqslant C\bigg(\frac{d}{m^3 K h} + \frac{\sigma^4 d^2 h^2}{m^4 p^2} + \frac{d^2 h^4}{m^5}\bigg(1 + \frac{dh}{m^2}\bigg)\bigg),$$

where the constant C *depends only on* $(M_i)_{i=1}^3$ *and* $(L_i)_{i=1}^2$.

The MSE bound in Theorem 8 differs from the standard SG-UBU result in Theorem 7 only in the stochastic gradient error component. This term is reduced by a factor of p^2, which reflects the variance reduction achieved by using a mini-batch of size p.

The proofs of Theorems 7 and 8 are given in Appendix D.1.

3.4 Numerical bias of SG-UBU

The MSE bound in Theorem 7 is composed of a stochastic fluctuation term that vanishes as the number of iterations $K \to \infty$, alongside persistent terms arising from the stochastic gradient and discretization errors. These latter components constitute the square of the numerical bias defined in (5), which bounds the long-time limit of the MSE:

$$\pi_h(f) - \pi(f) = \int_{\mathbb{R}^{2d}} f(\boldsymbol{x})\Big(\pi_h(\boldsymbol{x}, \boldsymbol{v}) - \big(\pi \otimes \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d)\big)(\boldsymbol{x}, \boldsymbol{v})\Big) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{v}.$$

Here, $\pi_h(\boldsymbol{x}, \boldsymbol{v})$ is an invariant distribution of the SG-UBU algorithm for a step size $h > 0$; its existence is guaranteed by Lemma 4, and uniqueness is neither established nor needed.

Beyond Theorem 7, the discrete Poisson solution $\phi_h(\boldsymbol{x}, \boldsymbol{v})$ also provides a direct expression for the numerical bias. For an arbitrary initial state $\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0)$, the definition of the discrete Poisson equation in (10) implies the identity:

$$\mathbb{E}\bigg[\frac{\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_0(h))}{h}\bigg] = f(\boldsymbol{X}_0) - \pi(f), \tag{22}$$

where the expectation is taken over the Brownian motion driving the exact one-step solution $\boldsymbol{Z}_0(h)$. Taking a further expectation with respect to the invariant distribution $\boldsymbol{Z}_0 \sim \pi_h$, and using the fact that the one-step SG-UBU update

$$\boldsymbol{Z}_1 = \big(\Phi^{\mathfrak{U}}_{h/2} \circ \Phi^{\mathfrak{B}}_h(\theta) \circ \Phi^{\mathfrak{U}}_{h/2}\big)\boldsymbol{Z}_0$$

also follows the distribution π_h, allows us to replace $\mathbb{E}^{\pi_h}[\phi_h(\boldsymbol{Z}_0)]$ with $\mathbb{E}^{\pi_h}[\phi_h(\boldsymbol{Z}_1)]$. This substitution leads to an exact expression for the numerical bias:

$$\pi_h(f) - \pi(f) = \mathbb{E}^{\pi_h}_{\theta}\left[\frac{\phi_h(\boldsymbol{Z}_1) - \phi_h(\boldsymbol{Z}_0(h))}{h}\right], \tag{23}$$

where the total expectation is over $\boldsymbol{Z}_0 \sim \pi_h$, the stochastic gradient index $\theta \sim \mathbb{P}_\theta$, and the Brownian motion $(\boldsymbol{B}_t)_{t\in[0,h]}$.

The right-hand side of (23) corresponds to the local error of the SG-UBU integrator, which can be decomposed into the stochastic gradient error and the discretization error:

$$\pi_h(f) - \pi(f) = \mathbb{E}^{\pi_h}_{\theta}\left[\frac{\phi_h(\boldsymbol{Z}_1) - \phi_h(\bar{\boldsymbol{Z}}_1)}{h}\right] + \mathbb{E}^{\pi_h}_{\theta}\left[\frac{\phi_h(\bar{\boldsymbol{Z}}_1) - \phi_h(\boldsymbol{Z}_0(h))}{h}\right], \tag{24}$$

where $\bar{\boldsymbol{Z}}_1$ is the one-step FG-UBU update

$$\bar{\boldsymbol{Z}}_1 = \big(\Phi^{\mathfrak{U}}_{h/2} \circ \Phi^{\mathfrak{B}}_h \circ \Phi^{\mathfrak{U}}_{h/2}\big)\boldsymbol{Z}_0.$$

The decomposition (24) reveals that the stochastic gradient error term, driven by $\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1$, is the dominant component that dictates the first-order convergence of the numerical bias. The discretization error, driven by $\bar{\boldsymbol{Z}}_1 - \boldsymbol{Z}_0(h)$, contributes to the higher-order terms.

Formalizing this decomposition allows for a precise quantification of the numerical bias, as stated in the following theorem. An additional upper bound of $\frac{m}{d}$ is imposed on the step size h to ensure stability.

Theorem 9 *Under Assumptions* (P), (Sg) *and* (T), *let the step size* $h \leqslant \min\{\frac{1}{4}, \frac{m}{d}\}$*; then the numerical bias* $\pi_h(f) - \pi(f)$ *of SG-UBU satisfies*

$$\left|\pi_h(f) - \pi(f) - \frac{h}{2M_2^2}\mathbb{E}^{\pi_h}_{\theta}\Big[\big(b(\boldsymbol{Y}_0,\theta) - \nabla U(\boldsymbol{Y}_0)\big)^\top \boldsymbol{M}_0\big(b(\boldsymbol{Y}_0,\theta) - \nabla U(\boldsymbol{Y}_0)\big)\Big]\right| \leqslant \frac{Cdh^2}{m^3},$$

where the initial state $(\boldsymbol{X}_0, \boldsymbol{V}_0) \sim \pi_h$*, and* $\theta \sim \mathbb{P}_\theta$*. The intermediate position* $\boldsymbol{Y}_0$ *defined in* (9) *is the position component of* $\Phi^{\mathfrak{U}}_{h/2}\boldsymbol{Z}_0$*,* $\boldsymbol{M}_0$ *is an averaged velocity Hessian of* $\phi_h(\boldsymbol{x}, \boldsymbol{v})$*:*

$$\boldsymbol{M}_0 = \int_0^1 \left(2\lambda \int_0^1 \nabla^2_{\boldsymbol{vv}}\phi_h\big(\bar{\boldsymbol{Z}}_1 + \lambda\varphi(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)\big)\mathrm{d}\varphi\right)\mathrm{d}\lambda,$$

and the constant C *depends only on* $(M_i)_{i=1}^3$ *and* $(L_i)_{i=1}^2$*. Consequently,*

$$|\pi_h(f) - \pi(f)| \leqslant C\left(\frac{\sigma^2 dh}{m^2} + \frac{dh^2}{m^3}\right).$$

Using the inequality $\frac{dh}{m} \leqslant 1$, the numerical bias bound in Theorem 9 simplifies to

$$|\pi_h(f) - \pi(f)| \leqslant \frac{Cdh}{m^2},$$

which recovers the result in Table 1. In Theorem 9, the intermediate position $\boldsymbol{Y}_0$ is independent of the stochastic gradient index $\theta \sim \mathbb{P}_\theta$, and the numerical bias $\pi_h(f) - \pi(f)$ converges at a first-order rate with respect to the step size h. Crucially, its leading-order coefficient is shown to be proportional to the variance of the stochastic gradient.

Theorem 9 can be readily adapted for mini-batch SG-UBU with the batch size p. The use of mini-batching reduces the stochastic gradient variance, scaling the leading-order bias term by $1/p$, thus

$$|\pi_h(f) - \pi(f)| \leqslant C\left(\frac{\sigma^2 dh}{m^2 p} + \frac{dh^2}{m^3}\right).$$

In the limit as $h \to 0$, the leading-order coefficient can be informally identified as

$$\lim_{h\to 0} \frac{\pi_h(f) - \pi(f)}{h} = \frac{1}{2M_2^2}\mathbb{E}_\theta^\pi\Big[\big(b(\boldsymbol{x}_0,\theta) - \nabla U(\boldsymbol{x}_0)\big)^\top \nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0,\boldsymbol{v}_0)\big(b(\boldsymbol{x}_0,\theta) - \nabla U(\boldsymbol{x}_0)\big)\Big], \quad (25)$$

where the expectation is taken over the stochastic gradient index $\theta \sim \mathbb{P}_\theta$, and the stationary distribution $(\boldsymbol{x}_0, \boldsymbol{v}_0) \sim \pi(\boldsymbol{x}) \otimes \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d)$, which is invariant under the underdamped Langevin dynamics (6). This result highlights a significant feature of employing a second-order integrator: the discretization error is suppressed to $\mathcal{O}(h^2)$, allowing the stochastic gradient error alone to determine the leading-order $\mathcal{O}(h)$ bias.

A subtle issue arises in the structure of the leading-order coefficient in (25). The expression involves the Hessian of the continuous Poisson solution, $\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0,\boldsymbol{v}_0)$, and our analysis in Theorem 1 required a bound on $\nabla^2 f(\boldsymbol{x})$ to control this term. This creates an apparent inconsistency, as one might intuitively expect the leading-order bias for an expectation $\pi(f)$ to depend only on the first-order properties of the test function f. The following lemma resolves this apparent mismatch:

Lemma 10 *Under Assumption* (P)*, let the test function $f(\boldsymbol{x})$ satisfy $|\nabla f(\boldsymbol{x})| \leqslant L_1$; then the continuous Poisson solution $\phi_0(\boldsymbol{x},\boldsymbol{v})$ defined in* (13) *satisfies*

$$\mathbb{E}^\pi\Big[\big\|\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0,\boldsymbol{v}_0)\big\|^2\Big] \leqslant \frac{C_d}{m^3},$$

where the expectation is taken over the stationary distribution $(\boldsymbol{x}_0,\boldsymbol{v}_0) \sim \pi(\boldsymbol{x})\otimes\mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d)$, and the constant C_d depends only on d, M_2, M_3, and L_1.

This reconciles the analytical requirements with intuition, confirming that the magnitude of the leading-order bias is controlled by the first-order information of the test function.

In the specific case where the stochastic gradient noise is Gaussian, the leading-order coefficient in the numerical bias admits a more explicit form. Suppose

$$b(\boldsymbol{x},\theta) = \nabla U(\boldsymbol{x}) + \sigma\theta, \qquad \theta \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d),$$

where the parameter $\sigma > 0$ aligns with Assumption (Sg). In this scenario, the expectation over the Gaussian noise θ in (25) can be computed explicitly, reducing the expression to involve the trace of the Hessian, i.e., the Laplacian:

$$\lim_{h\to 0} \frac{\pi_h(f) - \pi(f)}{h} = \frac{\sigma^2}{2M_2^2}\mathbb{E}^\pi[\Delta_{\boldsymbol{v}}\phi_0(\boldsymbol{x}_0, \boldsymbol{v}_0)]. \tag{26}$$

By substituting the continuous Poisson equation into (26), we can express the Laplacian in terms of first-order derivatives of ϕ_0, thereby eliminating the dependence on its Hessian:

$$\lim_{h\to 0} \frac{\pi_h(f) - \pi(f)}{h} = \frac{\sigma^2}{4M_2}\mathbb{E}^\pi\left[\left(\frac{1}{M_2}\nabla U(\boldsymbol{x}_0) + 2\boldsymbol{v}_0\right)^\top \nabla_{\boldsymbol{v}}\phi_0(\boldsymbol{x}_0, \boldsymbol{v}_0) - \boldsymbol{v}_0^\top \nabla_{\boldsymbol{x}}\phi_0(\boldsymbol{x}_0, \boldsymbol{v}_0)\right].$$

Given the uniform bound $|\nabla\phi_0| \leqslant C/m$, the leading-order coefficient is bounded as follows:

$$\left|\lim_{h\to 0} \frac{\pi_h(f) - \pi(f)}{h}\right| \leqslant \frac{C}{m}\mathbb{E}^\pi\left[|\nabla U(\boldsymbol{x}_0)| + |\boldsymbol{v}_0|\right] \leqslant \frac{Cd^{\frac{1}{2}}}{m^{\frac{3}{2}}}. \tag{27}$$

This bound suggests that the numerical bias remains controlled as long as $h = \mathcal{O}(d^{-1/2}m^{3/2})$. This condition on the step size is notably less restrictive than the condition $h = \mathcal{O}(d^{-1}m^2)$ needed to keep the general bias bound of Theorem 9 of order one.

It is important to emphasize that this derivation relies on the informal limit taken in (25), and a rigorous justification of this limit is beyond the scope of our current assumptions. Nonetheless, this analysis motivates the conjecture that the less restrictive condition $h = \mathcal{O}(d^{-1/2}m^{3/2})$ may be sufficient to ensure the stability and first-order accuracy of SG-UBU in general. A proof of this conjecture is left for future investigation.

The proofs of the results above are given in Appendix D.2.

4 Mean square error and numerical bias of variance-reduced SG-UBU

In this section, we extend our analysis to the variance-reduced sampling algorithms designed for finite-sum potentials of the form

$$U(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} U_i(\boldsymbol{x}).$$

We focus on two prominent techniques adapted to the UBU framework: Stochastic Variance Reduced Gradient (SVRG) (Johnson and Zhang, 2013) and SAGA[1] (Defazio et al., 2014). While the numerical bias of the standard SG-UBU integrator is fundamentally limited by a first-order convergence rate due to stochastic gradient noise, the SVRG-UBU and SAGA-UBU variants exhibit a more complex behavior. Our analysis reveals that these methods undergo a phase transition in their convergence rates: the numerical bias shifts from a first-order to a second-order rate as the step size h decreases below a critical threshold. This accelerated convergence is a distinctive feature that emerges from the interplay between variance reduction and a second-order accurate integrator.

Our analysis of the variance-reduced methods requires additional bounds on the individual potential components $\{\nabla U_i(\boldsymbol{x})\}_{i=1}^N$.

1. SAGA is a variant of Stochastic Averaged Gradient (SAG), but not an acronym.

Assumption (SgN$^+$) *For the finite-sum potential $U(\boldsymbol{x})$, there hold*

$$\left(\frac{1}{N}\sum_{i=1}^{N}|\nabla U_i(\boldsymbol{x})-\nabla U(\boldsymbol{x})|^8\right)^{\frac{1}{4}}\leqslant\sigma^2 d,\quad \left(\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla^2 U_i(\boldsymbol{x})-\nabla^2 U(\boldsymbol{x})\right\|^8\right)^{\frac{1}{4}}\leqslant\sigma^2,\quad \forall\boldsymbol{x}\in\mathbb{R}^d$$

for the constant $\sigma>0$, and for all $i\in\{1,\cdots,N\}$, there hold

$$|\nabla U_i(\boldsymbol{x})|\leqslant M_1\sqrt{|\boldsymbol{x}|^2+d},\quad \left\|\nabla^2 U_i(\boldsymbol{x})\right\|\leqslant M_2,\quad \forall\boldsymbol{x}\in\mathbb{R}^d$$

for constants $M_1,M_2>0$. Assume $\sigma\leqslant\min\{M_1,M_2\}$ for the convenience of analysis.

Assumption (SgN$^+$) is a direct strengthening of Assumption (SgN), as it extends the regularity conditions to include bounds on the Hessians of the individual components.

4.1 SVRG-UBU analysis

The SVRG-UBU algorithm adapts the SVRG approach, a control variate technique widely used to accelerate the convergence of Stochastic Gradient Descent (SGD) (Law et al., 2000; Johnson and Zhang, 2013). This adaptation replaces the standard mini-batch gradient estimator in the mini-batch SG-UBU with a variance-reduced counterpart.

Given a mini-batch subset $\theta\subset\{1,\ldots,N\}$ of size p and a fixed anchor position $\boldsymbol{x}_*\in\mathbb{R}^d$, the SVRG estimator is defined as

$$b(\boldsymbol{x},\boldsymbol{x}_*,\theta)=\frac{1}{p}\sum_{i\in\theta}\nabla U_i(\boldsymbol{x})-\frac{1}{p}\sum_{i\in\theta}\nabla U_i(\boldsymbol{x}_*)+\nabla U(\boldsymbol{x}_*).\tag{28}$$

It is straightforward to verify that this estimator remains unbiased for the true gradient:

$$\mathbb{E}_\theta\big[b(\boldsymbol{x},\boldsymbol{x}_*,\theta)\big]=\nabla U(\boldsymbol{x}).$$

The key to the variance reduction lies in the structure of the gradient error. The difference between the SVRG estimator and the true gradient can be expressed as:

$$b(\boldsymbol{x},\boldsymbol{x}_*,\theta)-\nabla U(\boldsymbol{x})=\left[\int_0^1\left(\frac{1}{p}\sum_{i\in\theta}\nabla^2 U_i-\nabla^2 U\right)\big(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{x}_*\big)\mathrm{d}\lambda\right](\boldsymbol{x}-\boldsymbol{x}_*).\tag{29}$$

This expression reveals that as the sampling point $\boldsymbol{x}$ approaches the anchor position $\boldsymbol{x}_*$, the gradient error diminishes. The full gradient $\nabla U(\boldsymbol{x}_*)$ serves as an effective control variate, substantially reducing the variance of the estimator when $\boldsymbol{x}$ is in a neighborhood of $\boldsymbol{x}_*$.

The SVRG-UBU algorithm is implemented in epochs to manage computational overhead. At the start of each epoch, a new anchor position $\boldsymbol{x}_*$ is established, and the corresponding full gradient $\nabla U(\boldsymbol{x}_*)$ is computed. This full gradient is then reused for a fixed number of inner iterations—typically N/p—before being updated. Each inner iteration evaluates p component gradients, and this strategy amortizes the $\mathcal{O}(N)$ cost of the full gradient calculation over the N/p iterations of an epoch, ensuring that the average computational cost per step remains $\mathcal{O}(p)$, on par with the standard mini-batch SG-UBU. For convenience we assume throughout that the epoch length $q=N/p$ is an integer; in practice it can be taken as $\lceil N/p\rceil$. The detailed procedure is presented in Algorithm 3.

The following theorem establishes the MSE bound for the SVRG-UBU integrator, paralleling the analysis for mini-batch SG-UBU in Theorem 7.

Algorithm 3: Stochastic Variance Reduced Gradient UBU (SVRG-UBU)

Input: initial state $(\boldsymbol{X}_0, \boldsymbol{V}_0)$, batch size p, step size h

Output: numerical solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$

for $k = 0, 1, \cdots$ **do**

  Evolve $(\boldsymbol{Y}_k, \boldsymbol{V}_k^{(1)}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{X}_k, \boldsymbol{V}_k)$

  if *k is a multiple of $\frac{N}{p}$* **then**

    Set the anchor position $\boldsymbol{Y}_* = \boldsymbol{Y}_k$

    Compute $\boldsymbol{b}_k = \nabla U(\boldsymbol{Y}_*) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \nabla U_i(\boldsymbol{Y}_*)$

  else

    Sample a random subset $\theta_k \subset \{1, \cdots, N\}$ of size p

    Compute $\boldsymbol{b}_k = \dfrac{1}{p}\displaystyle\sum_{i\in\theta_k} \nabla U_i(\boldsymbol{Y}_k) - \dfrac{1}{p}\sum_{i\in\theta_k} \nabla U_i(\boldsymbol{Y}_*) + \nabla U(\boldsymbol{Y}_*)$

  end

  Update the velocity $\boldsymbol{V}_k^{(2)} = \boldsymbol{V}_k^{(1)} - \frac{h}{M_2}\boldsymbol{b}_k$

  Evolve $(\boldsymbol{X}_{k+1}, \boldsymbol{V}_{k+1}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{Y}_k, \boldsymbol{V}_k^{(2)})$

end

Theorem 11 *Under Assumptions* (P), (SgN^+) *and* (T), *let the step size $h \leqslant \frac{1}{4}$, and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the SVRG-UBU solution with the batch size p satisfies

$$\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f)\right)^2\right] \leqslant C\left(\frac{d}{m^3 K h} + \frac{\sigma^4 d^2 h^2}{m^4 p^2}\min\left\{1, \frac{N^4 h^4}{m^2 p^4}\right\} + \frac{d^2 h^4}{m^5}\left(1 + \frac{dh}{m^2}\right)\right),$$

where the constant C depends only on $(M_i)_{i=1}^3$ and $(L_i)_{i=1}^2$.

Theorem 11 reveals a phase transition in the convergence rate of the numerical bias, which depends on the step size h relative to a critical threshold. For large step sizes, where $h > \sqrt{m}p/N$, the bias is dominated by the stochastic gradient variance and exhibits first-order convergence, mirroring the behavior of the mini-batch SG-UBU. Conversely, when the step size is small, $h < \sqrt{m}p/N$, the variance reduction mechanism becomes effective, suppressing the first-order error term. This allows the underlying second-order accuracy of the UBU integrator to emerge, resulting in an accelerated convergence rate. The transition between these two distinct convergence regimes occurs around the critical step size of $\sqrt{m}p/N$.

Next, we characterize the numerical bias of SVRG-UBU, adapting the approach from Theorem 9 to account for the algorithm's non-Markovian structure. The analysis involves two key distributions. The first is $\pi_h(\boldsymbol{x}, \boldsymbol{v})$, the invariant distribution of the embedded Markovian subsequence $(\boldsymbol{X}_{kN/p}, \boldsymbol{V}_{kN/p})_{k=0}^{\infty}$, whose existence is guaranteed by Lemma 27.

The second, which is used to measure the bias of the full trajectory, is the averaged distribution $\tilde{\pi}_h(\boldsymbol{x}, \boldsymbol{v})$. It represents the average state over a single epoch of $q = N/p$ steps, assuming the process starts from the stationary distribution π_h. For any Borel set $A \subset \mathbb{R}^{2d}$, this averaged distribution is precisely defined as

$$\tilde{\pi}_h(A) = \frac{1}{q} \sum_{k=0}^{q-1} \mathbb{P}\big(\boldsymbol{Z}_k \in A : \boldsymbol{Z}_0 \sim \pi_h\big). \tag{30}$$

We note that only the subsequence of SVRG-UBU iterates sampled at epoch boundaries is Markovian. Consequently, unlike Theorem 9, which represents the local error in a single SG-UBU update, SVRG-UBU must use the numerical solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^q$ within an epoch to characterize the numerical bias, which consists of the stochastic gradient error $\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}$ and the discretization error $\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)$ in the total q steps. Recall that $\bar{\boldsymbol{Z}}_{k+1}$ is the one-step FG-UBU update from $\boldsymbol{Z}_k$, namely,

$$\bar{\boldsymbol{Z}}_{k+1} = (\Phi_{h/2}^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}} \circ \Phi_{h/2}^{\mathfrak{U}}) \boldsymbol{Z}_k.$$

The following theorem provides a precise characterization of the numerical bias associated with the time-averaged distribution $\tilde{\pi}_h$ defined in (30).

Theorem 12 *Under Assumptions* (P), (SgN^+) *and* (T), *let the step size* $h \leqslant \min\{\frac{1}{4}, \frac{m}{d}\}$; *then the numerical bias* $\tilde{\pi}_h(f) - \pi(f)$ *of SVRG-UBU with the batch size* $p = N/q$ *satisfies*

$$\left| \tilde{\pi}_h(f) - \pi(f) - \frac{h}{2M_2^2 q} \mathbb{E}_{\theta_0, \cdots, \theta_{q-1}}^{\pi_h} \left[\sum_{k=0}^{q-1} \big(\boldsymbol{g}(\boldsymbol{Y}_k, \boldsymbol{Y}_0, \theta_k)\big)^\top \boldsymbol{M}_k \big(\boldsymbol{g}(\boldsymbol{Y}_k, \boldsymbol{Y}_0, \theta_k)\big) \right] \right| \leqslant \frac{Cdh^2}{m^3},$$

where $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{x}_*, \theta)$ *is the stochastic gradient deviation with the subset* θ*:*

$$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{x}_*, \theta) = \frac{1}{p} \sum_{i \in \theta} \nabla U_i(\boldsymbol{x}) - \frac{1}{p} \sum_{i \in \theta} \nabla U_i(\boldsymbol{x}_*) + \nabla U(\boldsymbol{x}_*) - \nabla U(\boldsymbol{x}),$$

the numerical solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^q$ *in an epoch is generated by Algorithm 3, the initial state* $(\boldsymbol{X}_0, \boldsymbol{V}_0) \sim \pi_h$, *and* $\theta_0, \cdots, \theta_{q-1}$ *are independent random subsets of* $\{1, \cdots, N\}$ *with size* p. *The intermediate position* $\boldsymbol{Y}_k$ *defined in* (9) *is the position component of* $\Phi_{h/2}^{\mathfrak{U}} \boldsymbol{Z}_k$, $\boldsymbol{M}_k$ *is an averaged velocity Hessian of* $\phi_h(\boldsymbol{x}, \boldsymbol{v})$*:*

$$\boldsymbol{M}_k = \int_0^1 \left(2\lambda \int_0^1 \nabla_{\boldsymbol{vv}}^2 \phi_h\big(\bar{\boldsymbol{Z}}_{k+1} + \lambda\varphi(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})\big) \mathrm{d}\varphi \right) \mathrm{d}\lambda,$$

and the constant C *depends only on* $(M_i)_{i=1}^3$ *and* $(L_i)_{i=1}^2$. *Consequently,*

$$|\tilde{\pi}_h(f) - \pi(f)| \leqslant C \left(\frac{\sigma^2 dh}{m^2 p} \min\left\{1, \frac{N^2 h^2}{mp^2}\right\} + \frac{dh^2}{m^3} \right).$$

In Theorem 12, the intermediate position $\boldsymbol{Y}_k$ depends on the history $\{\theta_0, \cdots, \theta_{k-1}\}$ but does not depend on the current random subset θ_k. Theorem 12 reveals that the numerical bias of SVRG-UBU is primarily determined by the average variance of the stochastic gradient over

a single epoch, with residual errors of order $\mathcal{O}(\frac{dh^2}{m^3})$. A crucial aspect of this result is that the bias is quantified with respect to the averaged distribution $\tilde{\pi}_h$, rather than the invariant distribution π_h of the anchor position subsequence. This choice is deliberate and aligns with the practical goal of measuring the time average error. Since the MSE in Theorem 11 is calculated over the entire numerical trajectory $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$, the averaged distribution $\tilde{\pi}_h$ provides the correct corresponding measure for the systematic bias.

The proofs of the results above are given in Appendix E.

4.2 SAGA-UBU analysis

The SAGA-UBU algorithm adapts the SAGA method, which, unlike SVRG, avoids periodic full gradient computations. Instead, it maintains a table of historical gradients for each of the N potential components, requiring $\mathcal{O}(dN)$ memory to store this information.

The SAGA estimator constructs a control variate from this table of stored gradients. For each component $i \in \{1, \dots, N\}$, the algorithm stores the gradient $\nabla U_i(\boldsymbol{\Phi}_i)$ evaluated at the historical position $\boldsymbol{\Phi}_i$ where it was last computed. At a new sampling position $\boldsymbol{x}$, the SAGA gradient estimator for a mini-batch θ of size p is defined as

$$b\big(\boldsymbol{x}, \{\boldsymbol{\Phi}_i\}_{i=1}^N, \theta\big) = \frac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}) - \frac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{\Phi}_i) + \frac{1}{N}\sum_{i=1}^{N} \nabla U_i(\boldsymbol{\Phi}_i), \tag{31}$$

where θ is a random subset of $\{1, \dots, N\}$. This estimator is unbiased for the full gradient, as can be readily verified by taking the expectation over θ:

$$\mathbb{E}\big[b\big(\boldsymbol{x}, \{\boldsymbol{\Phi}_i\}_{i=1}^N, \theta\big)\big] = \nabla U(\boldsymbol{x}).$$

By replacing the standard stochastic gradient in the UBU framework with this SAGA estimator, we arrive at the SAGA-UBU algorithm.

The SAGA-UBU algorithm generates a non-Markovian solution sequence $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ due to its reliance on a continuously updated table of historical gradients. This structure is more complex than that of SVRG-UBU, which possesses an embedded Markovian subsequence. Quantifying the stochastic gradient error for a general batch size becomes significantly more challenging in the SAGA setting. For analytical tractability, our analysis therefore focuses on the specific case where the batch size $p = 1$. The following theorem establishes the MSE bound for SAGA-UBU under this condition.

Theorem 13 *Under Assumptions* (P), (SgN^+) *and* (T), *let the step size* $h \leqslant c_0 m$, *and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the SAGA-UBU solution with the batch size 1 *satisfies*

$$\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f)\right)^2\right] \leqslant C\left(\frac{d}{m^3 K h} + \frac{d^2h^2}{m^6}\min\left\{1, N^4h^4\right\} + \frac{d^2h^4}{m^5}\left(1 + \frac{dh}{m^2}\right)\right),$$

where c_0 *depends only on* $(M_i)_{i=1}^2$, *and* C *depends only on* $(M_i)_{i=1}^3$ *and* $(L_i)_{i=1}^2$.

Algorithm 4: SAGA-UBU

Input: initial state $(\boldsymbol{X}_0, \boldsymbol{V}_0)$, batch size p, step size h

Output: numerical solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$

Evolve $(\boldsymbol{Y}_0, \boldsymbol{V}_0^{(1)}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{X}_0, \boldsymbol{V}_0)$

Set $\boldsymbol{\Phi}_i = \boldsymbol{Y}_0$ for $i = 1, \cdots, N$

Evolve $\boldsymbol{V}_0^{(2)} = \boldsymbol{V}_0^{(1)} - \frac{h}{M_2}\nabla U(\boldsymbol{Y}_0)$

Evolve $(\boldsymbol{X}_1, \boldsymbol{V}_1) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{Y}_0, \boldsymbol{V}_0^{(2)})$

for $k = 1, 2, \cdots$ **do**

Evolve $(\boldsymbol{Y}_k, \boldsymbol{V}_k^{(1)}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{X}_k, \boldsymbol{V}_k)$

Sample a random subset $\theta_k \subset \{1, \cdots, N\}$ of size p

Compute $\boldsymbol{b}_k = \dfrac{1}{p}\sum_{i\in\theta_k} \nabla U_i(\boldsymbol{Y}_k) - \dfrac{1}{p}\sum_{i\in\theta_k} \nabla U_i(\boldsymbol{\Phi}_i) + \dfrac{1}{N}\sum_{i=1}^{N} \nabla U_i(\boldsymbol{\Phi}_i)$

Set $\boldsymbol{\Phi}_i = \boldsymbol{Y}_k$ for all $i \in \theta_k$

Evolve $\boldsymbol{V}_k^{(2)} = \boldsymbol{V}_k^{(1)} - \frac{h}{M_2}\boldsymbol{b}_k$

Evolve $(\boldsymbol{X}_{k+1}, \boldsymbol{V}_{k+1}) = \Phi_{h/2}^{\mathfrak{U}}(\boldsymbol{Y}_k, \boldsymbol{V}_k^{(2)})$

end

Theorem 13 requires an additional upper bound $c_0 m$ on the step size h. This requirement stems from the modified Lyapunov condition satisfied by SAGA-BU (Lemma 29), which results in a nonlinear recursive inequality. The constant c_0 is then specifically determined in Theorem 30 to ensure the uniform-in-time moments.

SAGA-UBU (Theorem 13) and SVRG-UBU (Theorem 11) exhibit analogous convergence characteristics, including the phase transition concerning the step size h. However, the error bounds derived for SAGA-UBU are weaker than those for SVRG-UBU. This difference stems from the more stringent requirements needed in the analysis to control the moments of the SAGA-UBU solution and the variance of its stochastic gradient estimator, primarily due to its non-Markovian structure and reliance on historical gradient information.

Next, we characterize the numerical bias for SAGA-UBU. Because the SAGA-UBU solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ is non-Markovian, we measure the bias using the limit distribution $\tilde{\pi}_h(\boldsymbol{x}, \boldsymbol{v})$, which is defined as the subsequential weak limit of the time-averaged distributions

$$\tilde{\pi}_h^K(A) = \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{P}\big(\boldsymbol{Z}_k \in A \mid \boldsymbol{X}_0 = \boldsymbol{0},\ \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d)\big)$$

whose existence is guaranteed by Prokhorov's theorem, as stated in Lemma 31. The theorem below presents the bound for the numerical bias $\tilde{\pi}_h(f) - \pi(f)$.

Theorem 14 *Under Assumptions* (P), (SgN^+) *and* (T), *let the step size* $h \leqslant \min\{c_0 m, \frac{m}{d}\}$; *then the numerical bias* $\tilde{\pi}_h(f) - \pi(f)$ *of SAGA-UBU with the batch size* 1 *satisfies*

$$|\tilde{\pi}_h(f) - \pi(f)| \leqslant C\bigg(\frac{dh}{m^3}\min\big\{1, N^2h^2\big\} + \frac{dh^2}{m^3}\bigg),$$

where c_0 depends only on $(M_i)_{i=1}^2$, and C depends only on $(M_i)_{i=1}^3$ and $(L_i)_{i=1}^2$.

The proofs of the results above are given in Appendix F.

5 Numerical verification

In this section, we implement numerical tests to validate our theoretical results on the convergence of stochastic sampling algorithms. For this purpose, we restrict our tests to low-dimensional target distributions.

5.1 First-order convergence of numerical bias for SG-UBU

To verify the leading-order coefficient equality (25) for SG-UBU, we test 1D and 2D potential functions using both Gaussian noise and finite-sum stochastic gradients. Computationally, the limit in the LHS of (25) is evaluated by running SG-UBU at multiple step sizes h. The leading-order coefficient in the RHS is computed by numerically calculating the Hessian matrix $\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x},\boldsymbol{v}) \in \mathbb{R}^{d\times d}$, where the algorithmic details are provided in Appendix G.

1D Example We define the potential function $U(x)$ and test function $f(x)$ as

$$U(x) = \frac{1}{2}x^2 + 0.15\sin(1.6x - 0.5) + 0.1\sin(2.4x + 0.4),$$
$$f(x) = \cos x + 0.5\sin(2.5x) + 0.2\sin(0.5x + 0.4),$$

which are plotted in Figure 1.

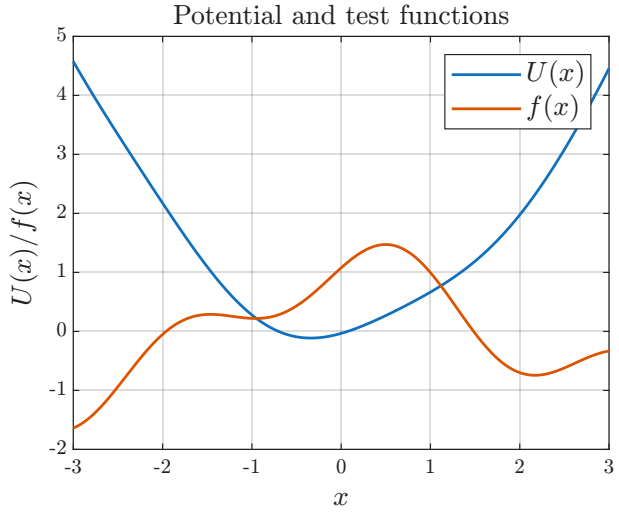


Figure 1: Graphs of the 1D potential function $U(x)$ and the test function $f(x)$.

We test two stochastic gradient models: Gaussian noise and finite-sum cases. In both cases, the numerical bias $\pi_h(f) - \pi(f)$ is computed from SG-UBU simulations with varying step sizes h and a total time $T = 10^6$. The leading-order coefficient is computed from 5×10^7 independent FG-UBU trajectories using $h = 2^{-10}$.

- **Gaussian noise case:** The stochastic gradient is

$$b(x,\theta) = \nabla U(x) + 3\theta, \quad \theta \sim \mathcal{N}(0,1).$$

- **Finite-sum case:** The stochastic gradient is

$$b(x,\theta) = \nabla U(x) + \nabla V_\theta(x)$$

where the random index θ is uniformly distributed on $\{1,\cdots,N\}$, and

$$V_i(x) = a_i \sin x + b_i \cos(1.2x) + c_i \sin(2x) + d_i \cos(2.5x), \quad i = 1,\cdots,N.$$

The parameters $\{a_i, b_i, c_i, d_i\}_{i=1}^N$ are drawn uniformly from $[-6,6]$ and then shifted to be mean-zero, which ensures that $b(x,\theta)$ is an unbiased estimator for $\nabla U(x)$.

Figure 2 plots the computed numerical bias $\pi_h(f) - \pi(f)$ against the step size h on a log-log scale. We compare this to the theoretical linear approximation (i.e., a line with a slope given by the independently computed leading-order coefficient), which shows strong agreement.

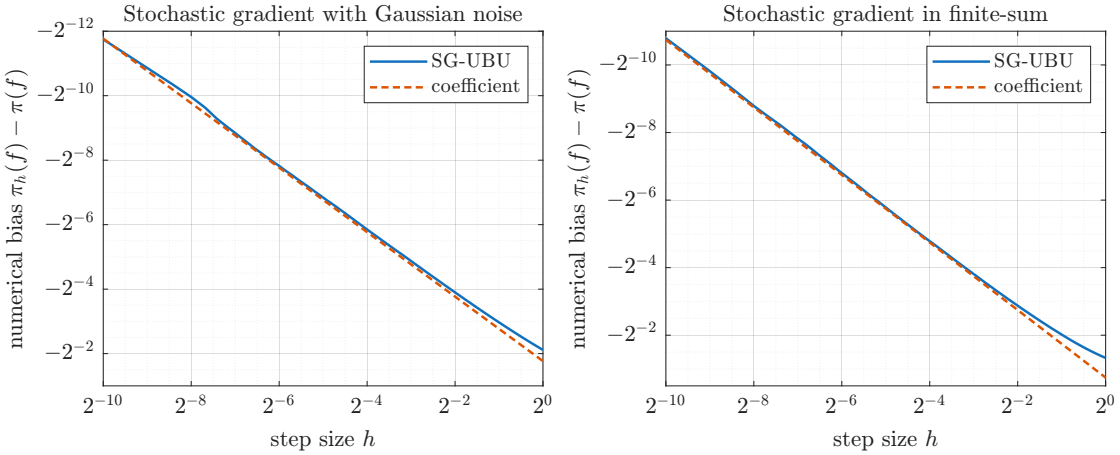


Figure 2: Log-log plot of the SG-UBU numerical bias $\pi_h(f) - \pi(f)$ vs. step size h for the 1D potential. The computed bias (blue solid line) is compared to the theoretical linear approximation (red dashed line). (Left) Gaussian noise case. (Right) Finite-sum case.

2D Example We define the potential function $U(x_1, x_2)$ and test function $f(x_1, x_2)$ as:

$$U(x_1,x_2) = \frac{1}{2}\big(1.4x_1^2 + 0.8x_2^2 + \sin(0.7x_1 - x_2)\cos(0.4x_1 + 0.6x_2)\big),$$
$$f(x_1,x_2) = \cos\big(1.4x_1 - 1.1\sin(1.2x_2)\big),$$

which are plotted in Figure 3.

We test two stochastic gradient models: Gaussian noise and finite-sum cases. In both cases, the numerical bias $\pi_h(f) - \pi(f)$ is computed from SG-UBU simulations, and the parameter settings are the same as in the 1D example.

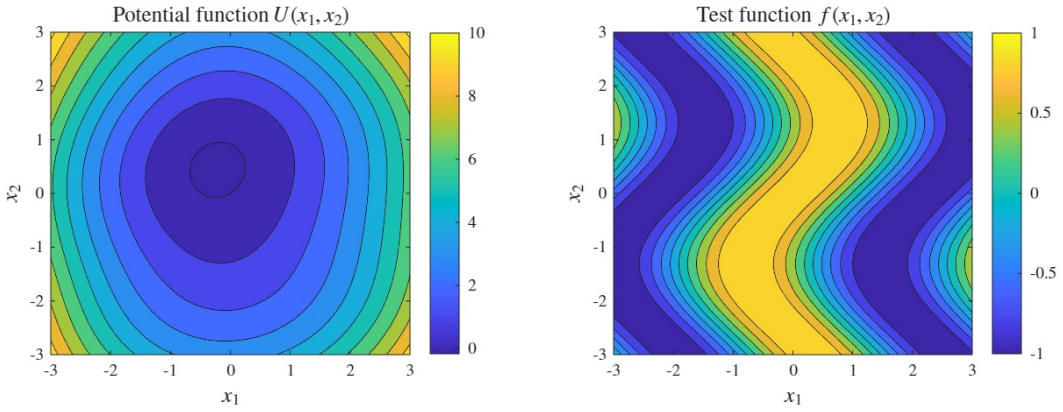


Figure 3: Graphs of the 2D potential function $U(x_1, x_2)$ and the test function $f(x_1, x_2)$.

- **Gaussian noise case:** The stochastic gradient is

$$b(\boldsymbol{x}, \theta) = \nabla U(\boldsymbol{x}) + 3\theta, \quad \theta \sim \mathcal{N}(0, \boldsymbol{I}_2).$$

- **Finite-sum case:** The stochastic gradient is

$$b(\boldsymbol{x}, \theta) = \nabla U(\boldsymbol{x}) + \nabla V_\theta(\boldsymbol{x}),$$

  where the random index θ is uniformly distributed on $\{1, \cdots, N\}$, and

$$V_i(x_1, x_2) = a_i \sin(x_1 + 2x_2) + b_i \cos(1.2x_1 - 0.7x_2) + c_i e^{-\frac{1}{2}x_1^2} + d_i e^{-\frac{1}{3}x_2^2}, \quad i = 1, \cdots, N.$$

  The parameters $\{a_i, b_i, c_i, d_i\}_{i=1}^N$ are drawn uniformly from $[-8, 8]$ and then shifted to be mean-zero, which ensures that $b(x, \theta)$ is an unbiased estimator for $\nabla U(x)$.

Figure 4 plots the computed numerical bias $\pi_h(f) - \pi(f)$ against the step size h on a log-log scale. We compare this to the theoretical linear approximation (i.e., a line with a slope given by the independently computed leading-order coefficient), which shows strong agreement.

Results demonstrate that $\pi_h(f) - \pi(f)$ exhibits first-order convergence in the step size h, with the leading-order coefficient accurately predicting the linear component at moderately small step sizes. This provides numerical validation for the equality (25). However, SG-UBU simulations in practice reveal significant nonlinear bias dependence on h, particularly in the finite-sum case. For the 2D example, the finite-sum bias exhibits approximate linear convergence only when $h \leqslant 2^{-6}$.

5.2 Comparison of variance-reduced sampling algorithms

We compare the numerical performance of three sampling algorithms: mini-batch SG-UBU, SVRG-UBU, and SAGA-UBU. Consider a finite-sum potential $U(\boldsymbol{x})$ in $\mathbb{R}^{10}$ defined as

$$U(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} U_i(\boldsymbol{x}), \quad \text{where} \quad U_i(\boldsymbol{x}) = \frac{1}{6}|\boldsymbol{x}|^2 + D_0(w_i^\top \boldsymbol{x} + b_i), \quad i = 1, \cdots, N,$$

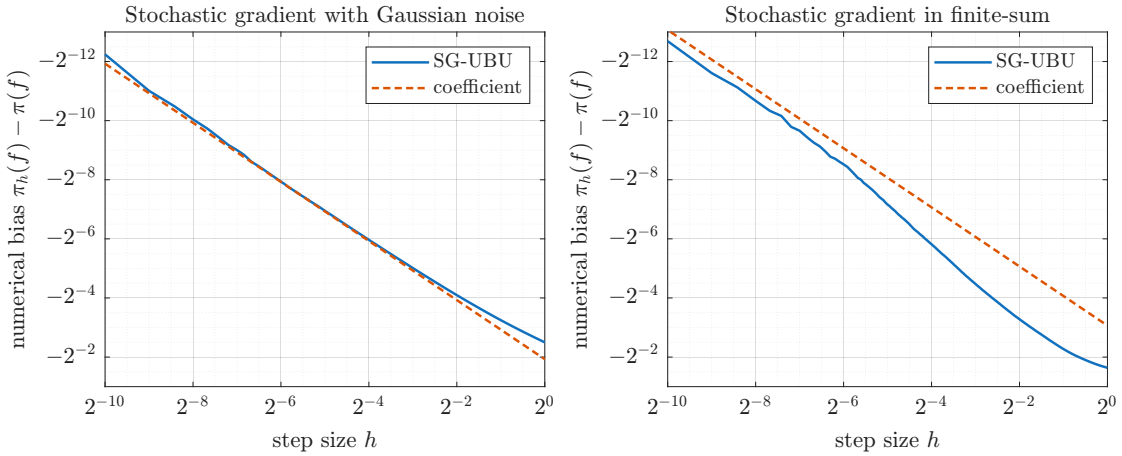


Figure 4: Log-log plot of the SG-UBU numerical bias $\pi_h(f) - \pi(f)$ vs. step size h for the 2D potential. The computed bias (blue solid line) is compared to the theoretical linear approximation (red dashed line). (Left) Gaussian noise case. (Right) Finite-sum case.

where the function $D_0(x) = 16e^{-\frac{1}{2}x^2} - 8\cos(x) - 4\sin(2x)$ introduces non-convexity. The potential $U(\boldsymbol{x})$ is non-convex, so this test lies outside Assumption (P) and probes whether the predicted phase transition persists beyond the regime our theorems cover. The parameters $(w_i, b_i)_{i=1}^N$ are randomly generated as follows:

$$w_{i,j} = \frac{j^{\frac{1}{4}}\xi_{i,j}}{\sqrt{\sum_{l=1}^{10} l^{\frac{1}{2}}\xi_{i,l}^2}}, \quad j = 1, \cdots, 10, \quad \text{and} \quad b_i = \frac{\cos i + \eta_i}{10},$$

where $\xi_{i,j}, \eta_i \sim \mathcal{N}(0,1)$ are independent standard Gaussian random variables.

For convenience, we take the coefficient M_2 appearing in the underdamped Langevin dynamics (6) to be 1, which is a scaling choice for the dynamics rather than a value of the Hessian bound in Assumption (P), resulting in the stochastic differential equation

$$\begin{cases} \dot{\boldsymbol{x}}_t = \boldsymbol{v}_t, \\ \dot{\boldsymbol{v}}_t = -\nabla U(\boldsymbol{x}_t) - 2\boldsymbol{v}_t + 2\dot{\boldsymbol{B}}_t. \end{cases} \tag{32}$$

The stochastic gradient sampling algorithms generate numerical solutions $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ by discretizing (32) with the step size h. To measure the sampling error, we select the vector-valued test function $f : \mathbb{R}^{10} \to \mathbb{R}^{30}$ as

$$f(\boldsymbol{x}) = \left(e^{-\frac{1}{8d}|\boldsymbol{x}|^2 - \frac{1}{4}x_j^2}, e^{-\frac{1}{8d}|\boldsymbol{x}|^2 - \frac{1}{4}(x_j+2)^2}, e^{-\frac{1}{8d}|\boldsymbol{x}|^2 - \frac{1}{4}(x_j+4)^2}\right)_{j=1,\cdots,10} \in \mathbb{R}^{30}.$$

Finally, the numerical performance of the algorithms is characterized by the time average error after $K = T/h$ steps:

$$\text{sampling error} = \left| \frac{1}{K} \sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f) \right|,$$

where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^{30}$. Assumption (T) and the theorems are stated for a scalar test function; they apply to each of the 30 components separately, and we report the Euclidean norm over the components. The sampling error comprises two components: the stochastic fluctuation depending on the total simulation time $T = Kh$, and the numerical bias depending on the step size h.

Sampling error dependence on N. We fix the batch size $p = 1$ and test the performance of the algorithms with various N. Figure 5 shows log-log plots of the sampling error versus the step size h for SG-UBU, SVRG-UBU, and SAGA-UBU. We test different dataset sizes $N \in \{10, 50, 100, 500\}$ with a fixed simulation time $T = 10^7$.

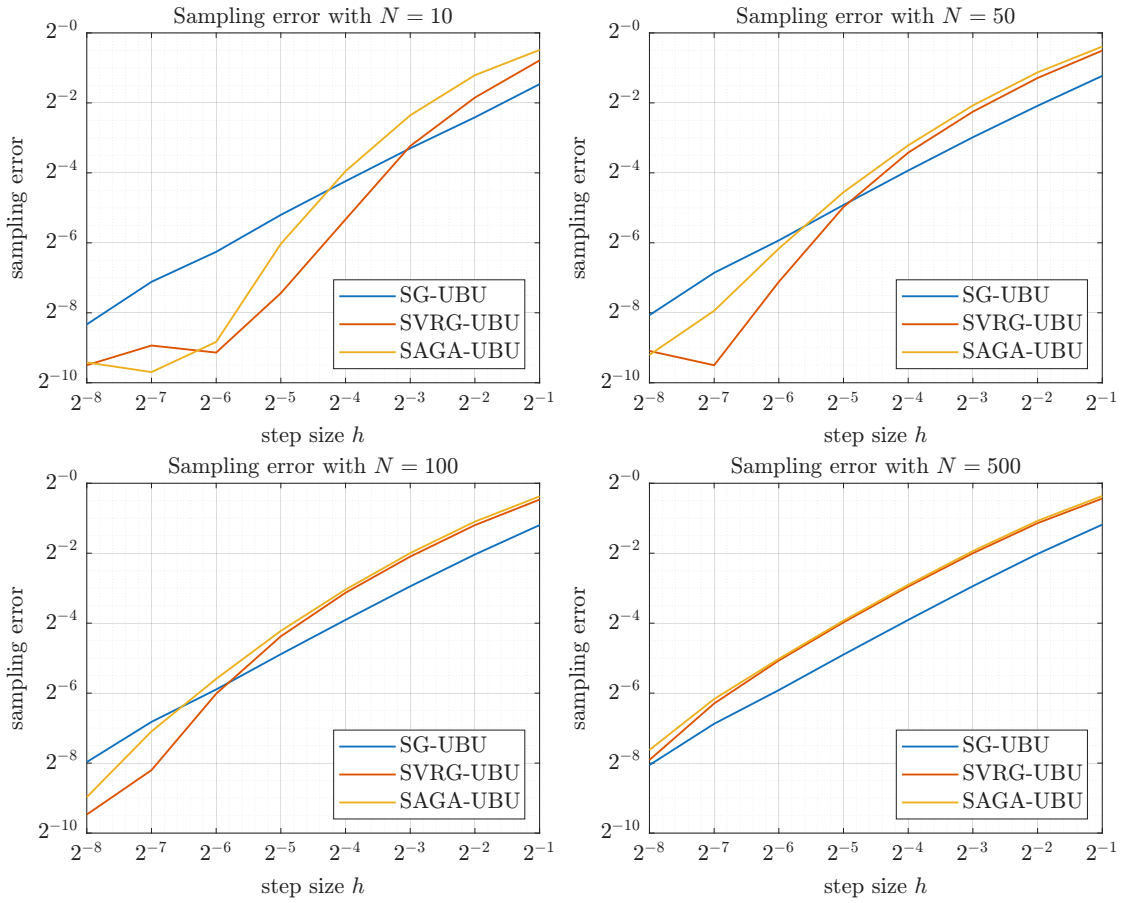


Figure 5: Log-log plot of the sampling error vs. step size h for SG-UBU, SVRG-UBU, and SAGA-UBU for different numbers of components $N \in \{10, 50, 100, 500\}$.

As predicted by Theorems 11 and 13, and observed here even outside their hypotheses, SVRG-UBU and SAGA-UBU exhibit a phase transition in convergence order near the critical step size at $\mathcal{O}(1/N)$. For $N = 10$, the error reduction for SVRG-UBU and SAGA-UBU saturates for $h < 2^{-6}$, likely due to stochastic fluctuations becoming dominant. For $N = 500$, the variance-reduced methods can perform worse than SG-UBU in certain ranges

of h. This occurs because a larger N can increase the variance introduced by the control variate strategy itself when the anchor point is far from the current sampling point.

In summary, SVRG-UBU and SAGA-UBU achieve better error than SG-UBU only when h is sufficiently small, specifically $h < \mathcal{O}(1/N)$. Thus, these methods may underperform SG-UBU when N is large or when the target error tolerance is not sufficiently stringent.

Sampling error dependence on p. We fix $N = 100$ and $T = 10^7$ and compare the performance of mini-batch SG-UBU and SVRG-UBU for different batch sizes $p \in \{2, 4, 8, 16\}$. Table 3 shows the sampling errors of these two algorithms. As predicted by theory, for a fixed T, the sampling error converges to a constant (dominated by stochastic fluctuations) as h becomes sufficiently small, which is evident in each column.

	$p = 2$		$p = 4$	
h	mini-batch SG-UBU	SVRG-UBU	mini-batch SG-UBU	SVRG-UBU
2^{-1}	2.37×10^{-1}	4.27×10^{-1}	1.18×10^{-1}	2.17×10^{-1}
2^{-2}	1.26×10^{-1}	2.28×10^{-1}	6.21×10^{-2}	1.06×10^{-1}
2^{-3}	6.57×10^{-2}	1.11×10^{-1}	3.14×10^{-2}	4.51×10^{-2}
2^{-4}	3.37×10^{-2}	4.75×10^{-2}	1.60×10^{-2}	1.47×10^{-2}
2^{-5}	1.67×10^{-2}	1.55×10^{-2}	7.27×10^{-3}	3.39×10^{-3}
2^{-6}	8.54×10^{-3}	3.35×10^{-3}	3.96×10^{-3}	1.36×10^{-3}
2^{-7}	4.12×10^{-3}	1.30×10^{-3}	2.30×10^{-3}	1.60×10^{-3}
2^{-8}	1.90×10^{-3}	1.65×10^{-3}	1.24×10^{-3}	1.33×10^{-3}
	$p = 8$		$p = 16$	
h	mini-batch SG-UBU	SVRG-UBU	mini-batch SG-UBU	SVRG-UBU
2^{-1}	5.20×10^{-2}	9.31×10^{-2}	1.75×10^{-2}	2.88×10^{-2}
2^{-2}	2.81×10^{-2}	4.12×10^{-2}	1.12×10^{-2}	1.10×10^{-2}
2^{-3}	1.42×10^{-2}	1.40×10^{-2}	6.18×10^{-3}	3.22×10^{-3}
2^{-4}	6.72×10^{-3}	2.87×10^{-3}	3.19×10^{-3}	1.26×10^{-3}
2^{-5}	3.08×10^{-3}	1.09×10^{-3}	1.89×10^{-3}	1.61×10^{-3}
2^{-6}	1.66×10^{-3}	1.00×10^{-3}	1.64×10^{-3}	1.57×10^{-3}
2^{-7}	1.45×10^{-3}	1.27×10^{-3}	1.36×10^{-3}	1.70×10^{-3}
2^{-8}	1.77×10^{-3}	1.36×10^{-3}	1.28×10^{-3}	1.77×10^{-3}

Table 3: Sampling error vs. step size h for mini-batch SG-UBU and SVRG-UBU at $N = 100$ with varying batch sizes $p \in \{2, 4, 8, 16\}$.

To quantify the relative performance, we define the sampling error ratio $R(p, h)$ as

$$R(p,h) = \frac{\text{sampling error of SVRG-UBU with batch size } p \text{ and step size } h}{\text{sampling error of mini-batch SG-UBU with batch size } p \text{ and step size } h}.$$

Figure 6 shows the plot of $R(p, h)$ for $N = 100$ and $p \in \{2, 4, 8, 16\}$.

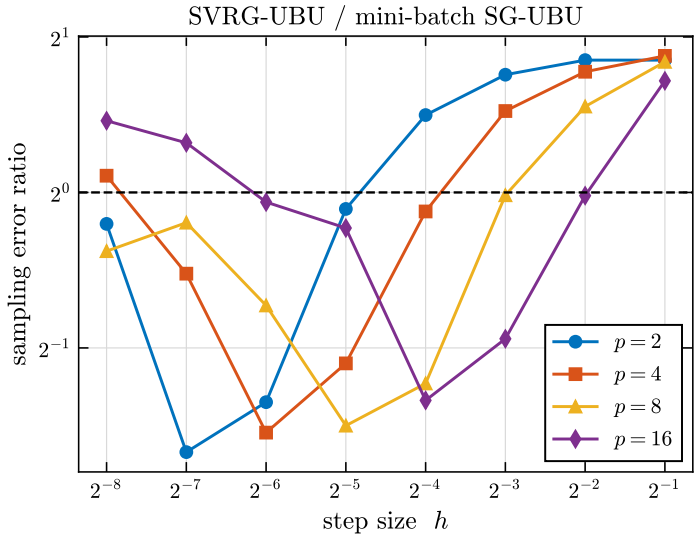


Figure 6: Sampling error ratio $R(p,h)$ (SVRG-UBU error / mini-batch SG-UBU error) vs. step size h at $N=100$ for batch sizes $p\in\{2,4,8,16\}$. Values below 1 (black dashed line) indicate that SVRG-UBU performs better than mini-batch SG-UBU.

Figure 6 illustrates that for each p, the sampling error ratio $R(p,h)$ first decreases and then increases towards 1 as h decreases. The minimum value of $R(p,h)$ occurs near the critical step size $\mathcal{O}(p/N)$, aligning with the prediction from Theorem 11. Based on the MSE bounds within the theorem, for a fixed simulation time T, $R(p,h)$ can be approximated as

$$R(p,h)\sim\frac{\frac{1}{\sqrt{dT}}+\frac{h}{p}\min\left\{1,\frac{N^2h^2}{p^2}\right\}+h^2}{\frac{1}{\sqrt{dT}}+\frac{h}{p}+h^2},\tag{33}$$

where we ignore the constants and dependence on m,σ for simplicity. When $h\leqslant\frac{p}{N}$, the min term becomes small, and (33) implies $R(p,h)\leqslant 1$, indicating that SVRG-UBU outperforms mini-batch SG-UBU. However, when $h\leqslant\frac{p}{\sqrt{dT}}$ we have

$$R(p,h)\geqslant\frac{\frac{1}{\sqrt{dT}}+h^2}{\frac{2}{\sqrt{dT}}+h^2}\geqslant\frac{1}{2},$$

showing that the advantage of SVRG-UBU diminishes. Taken together, the two limits characterize the range of step sizes over which an advantage of SVRG-UBU is observed:

$$\frac{p}{\sqrt{dT}}\leqslant h\leqslant\frac{p}{N}.\tag{34}$$

The two ends of (34) are lost for different reasons. When $h\ll\frac{p}{\sqrt{dT}}$, the sampling error is dominated by the stochastic fluctuation of the finite simulation time T rather than by the

stochastic gradient, so removing gradient variance changes little of the total error. When $h > \frac{p}{N}$, an epoch spans the time Nh/p, over which the trajectory travels a considerable distance before the full gradient is recomputed; the anchor position is then a poor reference for the control variate, and the variance it removes shrinks accordingly. Accordingly, (34) is a qualitative description of where variance reduction is effective, observed in Figure 6, rather than a quantitative prescription.

5.3 Algorithm selection by computational cost

Building upon the MSE bounds derived in Theorems 7 and 11, we propose the following procedure for selecting between mini-batch SG-UBU and SVRG-UBU and determining the simulation parameters. The two algorithms are compared at the same target accuracy: the simulation time T is common to both, only the step size differs, and the choice is made by comparing the resulting computational cost. The range (34) is used only as a reference for where variance reduction is effective.

First, the batch size p is typically chosen based on hardware constraints, such as the parallel processing capabilities of the CPU or GPU. Given p, the procedure is as follows:

1. **Determine the common simulation time T:**

   For a target accuracy level $\varepsilon > 0$ (representing the desired upper bound on the square root of the MSE), set the total simulation time T to control the stochastic fluctuation component, e.g., $T = d/\varepsilon^2$, such that the $\mathcal{O}(d/(Kh))$ term in the MSE is roughly ε^2. This term is common to both algorithms, so T is the same for both.

2. **Determine the step size of each algorithm:**

   The stochastic gradient and discretization bias terms differ between the two algorithms, so each admits its own largest step size. Take h_{SG} and h_{SVRG} to be the largest step sizes satisfying

   $$\begin{cases} \dfrac{dh}{p} \leqslant \varepsilon, \\ dh^2 \leqslant \varepsilon, \end{cases} \qquad \text{and} \qquad \begin{cases} \dfrac{dh}{p} \min\left\{1, \dfrac{N^2h^2}{p^2}\right\} \leqslant \varepsilon, \\ dh^2 \leqslant \varepsilon, \end{cases}$$

   respectively. Both choices give a total error of order ε, and $h_{\mathrm{SVRG}} \geqslant h_{\mathrm{SG}}$, since the min term is at most one.

3. **Compare the computational cost and select:**

   With $K = T/h$ iterations, the number of component gradient evaluations is

   $$\mathcal{C}_{\mathrm{SG}} = \frac{pT}{h_{\mathrm{SG}}}, \qquad \mathcal{C}_{\mathrm{SVRG}} = \frac{pT}{h_{\mathrm{SVRG}}}\left(3 - \frac{2p}{N}\right),$$

   where the factor of roughly 3 counts the two mini-batch gradients and the amortized full gradient of each SVRG-UBU epoch. Select the algorithm with the smaller cost, and set $K = T/h$ accordingly.

Since both algorithms are run to the same accuracy ε, the selection reduces to the cost comparison in Step 3. As ε decreases, h_{SVRG} decreases only as $\varepsilon^{1/2}$ while h_{SG} decreases as ε, so $\mathcal{C}_{\mathrm{SVRG}}/\mathcal{C}_{\mathrm{SG}}$ decreases and SVRG-UBU eventually becomes the cheaper choice.

We verify this procedure on a finite-sum target in $\mathbb{R}^8$ whose invariant expectation is known exactly. Each component potential is a symmetric two-component Gaussian mixture,

$$U_i(\boldsymbol{x}) = -\log\left(e^{-\frac{1}{2}|\boldsymbol{x}-\boldsymbol{a}_i|^2} + e^{-\frac{1}{2}|\boldsymbol{x}+\boldsymbol{a}_i|^2}\right) = \frac{1}{2}|\boldsymbol{x}|^2 - \log\cosh(\boldsymbol{a}_i\cdot\boldsymbol{x}) + \text{const},$$

so that

$$U(\boldsymbol{x}) = \frac{1}{2}|\boldsymbol{x}|^2 - \frac{1}{N}\sum_{i=1}^{N}\log\cosh(\boldsymbol{a}_i\cdot\boldsymbol{x}), \qquad \nabla^2 U(\boldsymbol{x}) = \boldsymbol{I}_d - \frac{1}{N}\sum_{i=1}^{N}\operatorname{sech}^2(\boldsymbol{a}_i\cdot\boldsymbol{x})\,\boldsymbol{a}_i\boldsymbol{a}_i^\top,$$

up to a constant. As $0 < \operatorname{sech}^2 \leqslant 1$, we have $\boldsymbol{I}_d - \frac{1}{N}\sum_i \boldsymbol{a}_i\boldsymbol{a}_i^\top \preccurlyeq \nabla^2 U \preccurlyeq \boldsymbol{I}_d$. We take $N = 100$ and draw the $\boldsymbol{a}_i$ independently, each with direction uniform on the unit sphere of $\mathbb{R}^d$ and length uniform on $[1.05, 1.95]$, giving $\lambda_{\max}\big(\frac{1}{N}\sum_i \boldsymbol{a}_i\boldsymbol{a}_i^\top\big) = 0.44$, so that Assumption (P) holds. Each component U_i is bimodal, with modes at $\pm\boldsymbol{a}_i$, but their average is strongly convex, so the target is log-concave and free of barriers.

Since U is even, π is symmetric about the origin, and the exact value $\pi(f) = 0$ is available for any odd test function without a reference computation; we take $f(\boldsymbol{x}) = \tanh(\boldsymbol{w}\cdot\boldsymbol{x})$ with $\boldsymbol{w}$ a fixed unit vector. To control the stochastic gradient variance independently of the target, we add offsets $\boldsymbol{c}_i$ to ∇U_i alone, drawn independently from $\mathcal{N}(\boldsymbol{0}, 4\boldsymbol{I}_d)$ and then recentered so that $\sum_{i=1}^N \boldsymbol{c}_i = \boldsymbol{0}$. The recentering leaves $\frac{1}{N}\sum_i \nabla U_i = \nabla U$, hence U and π, exactly unchanged, while raising the deviation to $\sigma \approx 5.4$. We use $p = 4$ and average over 64 independent trajectories.

Figure 7 (left) shows that the observed error is proportional to ε for both algorithms over the range $\varepsilon \in [2^{-9}, 2^{-3}]$, with ratios $\sqrt{\mathrm{MSE}}/\varepsilon$ of 0.47 ± 0.05 for mini-batch SG-UBU and 0.50 ± 0.05 for SVRG-UBU. The parameters produced by Steps 1 and 2 therefore deliver the requested accuracy for both, and the selection is governed by cost alone. Figure 7 (right) shows the two costs crossing at $\varepsilon \approx 1.6 \times 10^{-2}$: mini-batch SG-UBU is cheaper by a factor up to 2.9 for larger ε, while SVRG-UBU is cheaper below the crossing, by a factor 4.1 at $\varepsilon = 2^{-9}$ and increasing. The crossing depends only on d, N and p through the two step sizes, and not on σ, which affects only whether the target accuracy is attained.

6 Conclusion

This paper leverages the discrete Poisson equation framework to conduct a rigorous analysis of the mean square error for three stochastic gradient sampling algorithms—SG-UBU, SVRG-UBU, and SAGA-UBU—under the assumption of a globally convex potential function. Our work advances the theoretical understanding of sampling algorithms in several key aspects. First, for the baseline SG-UBU algorithm, we establish that its numerical bias exhibits first-order convergence with respect to the step size, where the leading coefficient is proportional to the stochastic gradient variance. This result quantifies the explicit impact of gradient noise on discretization bias. Second, for the variance-reduced variants SVRG-UBU and SAGA-UBU, we uncover a distinctive phase transition phenomenon: their numerical convergence order shifts between first- and second-order regimes depending on

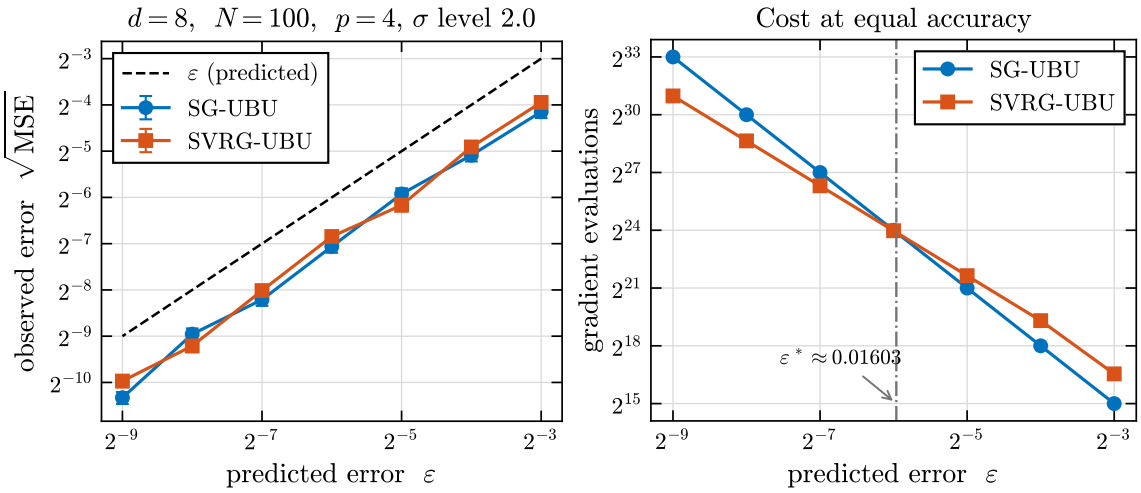


Figure 7: Verification of the selection procedure for the $\mathbb{R}^8$ Gaussian-mixture target with $N = 100$, $p = 4$. (Left) Observed error $\sqrt{\text{MSE}}$ against the target accuracy ε, with the diagonal ε shown dashed; error bars are the sampling uncertainty of a root mean square over 64 trajectories. (Right) Component gradient evaluations required by each algorithm, with the step size of each set by Step 2 and the common simulation time $T = d/\varepsilon^2$. The vertical line marks the target accuracy ε at which the two costs are equal.

the step size magnitude. This behavior highlights the nuanced interplay between variance reduction techniques and discretization schemes in second-order integrators. Third, we derive a practical criterion to guide algorithm selection between mini-batch SG-UBU and mini-batch SVRG-UBU. Given a target accuracy, the two samplers are assigned a common simulation time and their own step sizes, and the one requiring fewer component gradient evaluations is selected; the variance-reduced sampler becomes the cheaper choice once the accuracy demand is sufficiently stringent.

The analytical value of our contributions lies in the precision and generality of the discrete Poisson framework, which enables sharp error decomposition that is difficult to achieve through traditional Wasserstein analysis alone. Specifically, our approach cleanly separates stochastic noise, discretization error, and variance reduction effects, providing a simple bias bound for SG-UBU-type methods under global convexity. For variance-reduced algorithms, the identified step size-dependent phase transition offers new insights into the design of high-precision samplers. These theoretical advances strengthen the foundation for developing efficient sampling algorithms in high-dimensional settings, particularly in Bayesian inference and other applications requiring scalable, accurate integration over complex distributions. The discrete Poisson methodology, with its flexibility in handling time average errors and numerical bias, further establishes itself as a powerful tool for analyzing stochastic sampling algorithms.

However, certain limitations in our analysis warrant mention. First, the explicit dependence of the error bounds on the dimension d and the convexity parameter m may be suboptimal, and our framework does not readily yield convergence rates in standard Wasserstein distances. This constraint arises from the structure of the discrete Poisson

methodology itself, which fundamentally targets mean square error analysis. Additionally, the global convexity assumption is used to establish pointwise bounds on the Hessian of the Kolmogorov solution. Relaxing this assumption is left to future work.

Acknowledgments and Disclosure of Funding

The work of Z. Zhou was partially supported by the National Key R&D Program of China (Project No. 2021YFA1001200), and the National Natural Science Foundation of China (Grant No. 12171013).

The authors would like to thank Xu'an Dou at Peking University for the helpful discussions on the proof of Lemma 10.

References

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1771–1778, 2012.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.

Zeyuan Allen-Zhu and Yang Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning*, pages 1080–1089. PMLR, 2016.

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 71(3):1–55, 2024.

Jason M Altschuler, Sinho Chewi, and Matthew S Zhang. Shifted composition IV: Underdamped Langevin and numerical discretizations with partial acceleration. *arXiv preprint arXiv:2506.23062*, 2025.

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

Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29:599–615, 2019.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yu Cao, Jianfeng Lu, and Lihan Wang. On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. *Archive for Rational Mechanics and Analysis*, 247(5):90, 2023.

Neil K Chada, Benedict Leimkuhler, Daniel Paulin, and Peter A Whalley. Unbiased kinetic Langevin Monte Carlo with inexact gradients. *arXiv preprint arXiv:2311.05025*, 2023.

Martin Chak and Pierre Monmarché. Reflection coupling for unadjusted generalized Hamiltonian Monte Carlo in the nonconvex stochastic gradient case. *IMA Journal of Numerical Analysis*, page draf045, 2025.

Martin Chak, Nikolas Kantas, Tony Lelièvre, and Grigorios A Pavliotis. Optimal friction matrix for underdamped Langevin sampling. *ESAIM: Mathematical Modelling and Numerical Analysis*, 57(6):3335–3371, 2023.

Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018.

Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. *Advances in neural information processing systems*, 28, 2015.

Peng Chen, Jianya Lu, and Lihu Xu. Approximation to stochastic variance reduced gradient Langevin dynamics by stochastic delay differential equations. *Applied Mathematics & Optimization*, 85(2):15, 2022.

Peng Chen, Qi-Man Shao, and Lihu Xu. A probability approximation framework: Markov process approach. *The Annals of Applied Probability*, 33(2):1619–1659, 2023.

Pingyan Chen and Soo Hak Sung. Rosenthal type inequalities for random variables. *Journal of Mathematical Inequalities*, 14(2):305–18, 2020.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. Stochastic gradient and Langevin processes. In *International Conference on Machine Learning*, pages 1810–1819. PMLR, 2020.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Arnak 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.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. Variance reduction in stochastic gradient Langevin dynamics. *Advances in neural information processing systems*, 29, 2016.

Andreas Eberle and Francis Lörler. Non-reversible lifts of reversible diffusion processes and relaxation times. *Probability Theory and Related Fields*, pages 1–31, 2024.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Peter W Glynn and Sean P Meyn. A Liapounov bound for solutions of the Poisson equation. *The Annals of Probability*, pages 916–931, 1996.

Nicolaï Gouraud, Pierre Le Bris, Adrien Majka, and Pierre Monmarché. HMC and underdamped Langevin united in the unadjusted convex smooth case. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):278–303, 2025.

Arnaud Guillin, Yu Wang, Lihu Xu, and Haoran Yang. Error estimates between SGD with momentum and underdamped Langevin diffusion. *arXiv preprint arXiv:2410.17297*, 2024.

Martin Hairer. Ergodic properties of Markov processes. *Lecture notes*, 2006.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

James E Johndrow and Jonathan C Mattingly. Error bounds for approximations of Markov chains used in Bayesian sampling. *arXiv preprint arXiv:1711.05382*, 2017.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer, 2 edition, 1991.

Yuri Kinoshita and Taiji Suzuki. Improved convergence rate of stochastic gradient Langevin dynamics with variance reduction and its application to optimization. *Advances in Neural Information Processing Systems*, 35:19022–19034, 2022.

Ioannis Kontoyiannis and Sean P Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *The Annals of Applied Probability*, 13(1):304–362, 2003.

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Averill M Law, W David Kelton, and W David Kelton. *Simulation modeling and analysis*, volume 3. Mcgraw-hill New York, 2000.

Ben Leimkuhler and Charles Matthews. Molecular dynamics. *Interdisciplinary applied mathematics*, 39(1), 2015.

Benedict Leimkuhler, Daniel Paulin, and Peter A Whalley. Contraction rate estimates of stochastic gradient kinetic langevin integrators. *ESAIM: Mathematical Modelling and Numerical Analysis*, 58(6):2255–2286, 2024.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Lei Li and Yuliang Wang. A sharp uniform-in-time error estimate for Stochastic Gradient Langevin Dynamics. *arXiv preprint arXiv:2207.09304*, 2022.

Zhize Li, Tianyi Zhang, Shuyu Cheng, Jun Zhu, and Jian Li. Stochastic gradient Hamiltonian Monte Carlo with variance reduction for bayesian inference. *Machine Learning*, 108:1701–1727, 2019.

Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

Jonathan C Mattingly, Andrew M Stuart, and Michael V Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM Journal on Numerical Analysis*, 48(2):552–577, 2010.

Edward Meeds and Max Welling. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 593–602, 2014.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Tigran Nagapetyan, Andrew B Duncan, Leonard Hasenclever, Sebastian J Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The true cost of stochastic gradient Langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. *Advances in neural information processing systems*, 26, 2013.

Valentin V Petrov. *Sums of independent random variables*, volume 82. Springer Science & Business Media, 2012.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

Nathan Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

Jesus Maria Sanz-Serna and Konstantinos C Zygalakis. Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *Journal of Machine Learning Research*, 22(242):1–37, 2021.

Katharina Schuh and Peter A Whalley. Convergence of kinetic Langevin samplers for non-convex potentials. *arXiv preprint arXiv:2405.09992*, 2024.

Maximilian Scott, Dáire O'Kane, Andraž Jelinčič, and James Foster. Underdamped Langevin MCMC with third order convergence. *arXiv preprint arXiv:2508.16485*, 2025.

Luke Shaw and Peter A Whalley. Random reshuffling for stochastic gradient Langevin dynamics. *arXiv preprint arXiv:2501.16055*, 2025.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Chaobing Song, Yong Jiang, and Yi Ma. Variance reduction via accelerated dual averaging for finite-sum optimization. *Advances in Neural Information Processing Systems*, 33: 833–844, 2020.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field Langevin dynamics: time-space discretization, stochastic gradient, and variance reduction. *Advances in Neural Information Processing Systems*, 36:15545–15577, 2023.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

Cédric Villani. *Hypocoercivity*, volume 202. American Mathematical Society, 2009.

Sebastian J Vollmer, Konstantinos C Zygalakis, and Yee Whye Teh. Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016.

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

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced Hamilton Monte Carlo methods. In *International Conference on Machine Learning*, pages 6028–6037. PMLR, 2018a.

Difan Zou, Pan Xu, and Quanquan Gu. Subsampled stochastic variance-reduced gradient Langevin dynamics. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018b.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic gradient Hamiltonian Monte Carlo methods with recursive variance reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Contents

1 Introduction **1**
1.1 Our contributions . . . 4
1.2 Related works . . . 4

2 Preliminaries and methodology **8**
2.1 Notations and assumptions . . . 8
2.2 Underdamped Langevin dynamics and SG-UBU integrator . . . 9
2.3 The discrete Poisson equation framework . . . 11

3 Mean square error and numerical bias of SG-UBU **14**
3.1 Stability: Uniform-in-time moment bound . . . 14
3.2 Consistency: Local error analysis . . . 15
3.3 Mean square error of SG-UBU . . . 16
3.4 Numerical bias of SG-UBU . . . 18

4 Mean square error and numerical bias of variance-reduced SG-UBU **21**
4.1 SVRG-UBU analysis . . . 22
4.2 SAGA-UBU analysis . . . 25

5 Numerical verification **27**
5.1 First-order convergence of numerical bias for SG-UBU . . . 27
5.2 Comparison of variance-reduced sampling algorithms . . . 29
5.3 Algorithm selection by computational cost . . . 34

6 Conclusion **35**

A Estimate of discrete Poisson solution **44**
A.1 First variation process: Definition . . . 44
A.2 First variation process: Proof of exponential decay . . . 45
A.3 First variation process: Proof of coupling contractivity . . . 47

B Uniform-in-time moment bound for SG-UBU **51**
B.1 Lyapunov condition for FG-BU integrator . . . 51
B.2 Lyapunov condition for SG-BU integrator . . . 53
B.3 Uniform-in-time moment bound of SG-UBU integrator . . . 55

C Local error analysis for SG-UBU **56**
C.1 Estimate of stochastic gradient error . . . 56
C.2 Estimate of discretization error . . . 57
C.3 Estimate of long-time displacement . . . 60

D Proofs for MSE and numerical bias of SG-UBU **61**
D.1 Mean square error of SG-UBU . . . 61
D.2 Numerical bias of SG-UBU . . . 66

E Proofs for MSE and numerical bias of SVRG-UBU **70**
E.1 Uniform-in-time moment bound of SVRG-UBU 70
E.2 Local error analysis of SVRG-UBU 73
E.3 MSE of SVRG-UBU 75
E.4 Numerical bias of SVRG-UBU 75

F Proofs for MSE and numerical bias of SAGA-UBU **76**
F.1 Uniform-in-time moment bound of SAGA-UBU 77
F.2 Local error analysis of SAGA-UBU 82
F.3 MSE of SAGA-UBU 88
F.4 Numerical bias of SAGA-UBU 88

G Numerical calculation of leading-order coefficient **90**

Appendix A. Estimate of discrete Poisson solution

This section is dedicated to proving Theorem 1, which provides quantitative bounds on the derivatives of the discrete Poisson solution $\phi_h(\boldsymbol{x}, \boldsymbol{v})$ introduced in (12). Our analysis relies on the *first variation process*[2], a technical tool that captures the sensitivity of the underdamped Langevin dynamics (6) to perturbations in its initial state $(\boldsymbol{x}, \boldsymbol{v})$. The idea of utilizing the variation process is inspired by Chak et al. (2023).

A.1 First variation process: Definition

Let $(\boldsymbol{x}_t, \boldsymbol{v}_t)_{t\geqslant 0}$ denote the exact solution to the underdamped Langevin dynamics (6) with the initial state $(\boldsymbol{x}_0, \boldsymbol{v}_0)$. The first variation process $D_{\boldsymbol{v}}\boldsymbol{x}_t \in \mathbb{R}^{d\times d}$ is defined as the Jacobian matrix of $\boldsymbol{x}_t$ with respect to the initial velocity $\boldsymbol{v}_0$, specifically expressed as:

$$D_{\boldsymbol{v}}\boldsymbol{x}_t = \big[\partial_{v_j} x_i\big]_{i,j=1}^{d} = \big[\partial_{v_1}\boldsymbol{x}_t \quad \partial_{v_2}\boldsymbol{x}_t \quad \cdots \quad \partial_{v_d}\boldsymbol{x}_t\big] \in \mathbb{R}^{d\times d}.$$

The companion first variation processes $D_{\boldsymbol{x}}\boldsymbol{x}_t, D_{\boldsymbol{x}}\boldsymbol{v}_t$ and $D_{\boldsymbol{v}}\boldsymbol{v}_t$ admit analogous definitions. Through differentiation of the Kolmogorov solution $u(\boldsymbol{x}_0, \boldsymbol{v}_0, t)$ in (11), we derive the gradient expressions for $u(\boldsymbol{x}_0, \boldsymbol{v}_0, t)$:

$$\begin{aligned}\nabla_{\boldsymbol{x}} u(\boldsymbol{x}_0, \boldsymbol{v}_0, t) &= \mathbb{E}^{(\boldsymbol{x}_0,\boldsymbol{v}_0)}\big[\big(D_{\boldsymbol{x}}\boldsymbol{x}_t\big)^\top \nabla f(\boldsymbol{x}_t)\big],\\ \nabla_{\boldsymbol{v}} u(\boldsymbol{x}_0, \boldsymbol{v}_0, t) &= \mathbb{E}^{(\boldsymbol{x}_0,\boldsymbol{v}_0)}\big[\big(D_{\boldsymbol{v}}\boldsymbol{x}_t\big)^\top \nabla f(\boldsymbol{x}_t)\big].\end{aligned} \tag{35}$$

This formulation reveals that the exponential decay of $\nabla u(\boldsymbol{x}_0, \boldsymbol{v}_0, t)$ originates directly from the analytical properties of the first variation processes $D_{\boldsymbol{x}}\boldsymbol{x}_t$ and $D_{\boldsymbol{v}}\boldsymbol{x}_t$.

By differentiating (6) with respect to the initial state $(\boldsymbol{x}_0, \boldsymbol{v}_0)$, we obtain the governing ordinary differential equations for the first variation processes.

2. This term is also known as the *tangent process* or *Jacobian process* in the literature. We adopt the term *first variation process*, which is common in Malliavin calculus.

Definition 15 *Let $(\boldsymbol{x}_t, \boldsymbol{v}_t)_{t\geqslant 0}$ be the solution to* (6) *with initial state $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ and Brownian motion $(\boldsymbol{B}_t)_{t\geqslant 0}$. Define $(\boldsymbol{Q}_t, \boldsymbol{P}_t)_{t\geqslant 0}$ as the solution to the matrix-valued differential equation*

$$\begin{cases} \dot{\boldsymbol{Q}}_t = \boldsymbol{P}_t, \\ \dot{\boldsymbol{P}}_t = -\dfrac{1}{M_2}\nabla^2 U(\boldsymbol{x}_t)\boldsymbol{Q}_t - 2\boldsymbol{P}_t, \end{cases} \tag{36}$$

with the initial configuration $(\boldsymbol{Q}_0, \boldsymbol{P}_0) \in \mathbb{R}^{2d\times d}$, formally expressed as

$$(\boldsymbol{Q}_t, \boldsymbol{P}_t)_{t\geqslant 0} = \mathrm{FirstVariation}\big(\boldsymbol{Q}_0, \boldsymbol{P}_0, \boldsymbol{x}_0, \boldsymbol{v}_0, (\boldsymbol{B}_t)_{t\geqslant 0}\big).$$

At the initial time $t = 0$, the Jacobian matrix of $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ with respect to $\boldsymbol{x}_0$ yields $(\boldsymbol{I}_d, \boldsymbol{O}_d)$, thus identifying the first variation process

$$(D_{\boldsymbol{x}}\boldsymbol{x}_t, D_{\boldsymbol{x}}\boldsymbol{v}_t)_{t\geqslant 0} = \mathrm{FirstVariation}\big(\boldsymbol{I}_d, \boldsymbol{O}_d, \boldsymbol{x}_0, \boldsymbol{v}_0, (\boldsymbol{B}_t)_{t\geqslant 0}\big).$$

Analogously, the velocity-related variation process satisfies

$$(D_{\boldsymbol{v}}\boldsymbol{x}_t, D_{\boldsymbol{v}}\boldsymbol{v}_t)_{t\geqslant 0} = \mathrm{FirstVariation}\big(\boldsymbol{O}_d, \boldsymbol{I}_d, \boldsymbol{x}_0, \boldsymbol{v}_0, (\boldsymbol{B}_t)_{t\geqslant 0}\big).$$

We proceed to demonstrate the exponential decay property of the first variation process $(\boldsymbol{Q}_t, \boldsymbol{P}_t)_{t\geqslant 0}$. Define the composite matrices

$$\boldsymbol{W}_t = \begin{bmatrix} \boldsymbol{Q}_t \\ \boldsymbol{P}_t \end{bmatrix} \in \mathbb{R}^{2d\times d}, \qquad \boldsymbol{S} = \begin{bmatrix} 2\boldsymbol{I}_d & \boldsymbol{I}_d \\ \boldsymbol{I}_d & \boldsymbol{I}_d \end{bmatrix} \in \mathbb{R}^{2d\times 2d},$$

and consider the *twisted* quadratic form

$$\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t = 2\boldsymbol{Q}_t^\top \boldsymbol{Q}_t + \boldsymbol{Q}_t^\top \boldsymbol{P}_t + \boldsymbol{P}_t^\top \boldsymbol{Q}_t + \boldsymbol{P}_t^\top \boldsymbol{P}_t \in \mathbb{R}^{d\times d}. \tag{37}$$

The twisted quadratic form plays a fundamental role in hypocoercivity analysis (Villani, 2009), particularly for establishing ergodicity and Lyapunov conditions in the underdamped Langevin dynamics (Eberle et al., 2019).

A.2 First variation process: Proof of exponential decay

We show that the exponential decay of the quadratic form $\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t$ can be rigorously established under the global convexity of the potential function $U(\boldsymbol{x})$.

Lemma 16 *Under Assumption* (P) *with the initial configuration*

$$(\boldsymbol{Q}_0, \boldsymbol{P}_0) = (\boldsymbol{I}_d, \boldsymbol{O}_d) \text{ or } (\boldsymbol{O}_d, \boldsymbol{I}_d),$$

the first variation process $\boldsymbol{W}_t$ satisfies the exponential decay property

$$\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t \preccurlyeq 2e^{-\frac{mt}{M_2}}\boldsymbol{I}_d, \quad \forall t \geqslant 0.$$

Proof Define the auxiliary matrices

$$\boldsymbol{H}_t = \frac{1}{M_2}\nabla^2 U(\boldsymbol{x}_t) \in \mathbb{R}^{d\times d}, \qquad \boldsymbol{A}_t = \begin{bmatrix} \boldsymbol{O}_d & -\boldsymbol{I}_d \\ \boldsymbol{H}_t & 2\boldsymbol{I}_d \end{bmatrix} \in \mathbb{R}^{2d\times 2d}, \tag{38}$$

then the first variation process in (36) admits the matrix form $\dot{\boldsymbol{W}}_t = -\boldsymbol{A}_t\boldsymbol{W}_t$. Note that Assumption (P) yields the matrix eigenvalue bound

$$\frac{m}{M_2}\boldsymbol{I}_d \preccurlyeq \boldsymbol{H}_t \preccurlyeq \boldsymbol{I}_d,$$

and we obtain the following matrix inequality as the coercivity condition:

$$\boldsymbol{A}_t^\top \boldsymbol{S} + \boldsymbol{S}\boldsymbol{A}_t = \begin{bmatrix} 2\boldsymbol{H}_t & \boldsymbol{H}_t \\ \boldsymbol{H}_t & 2\boldsymbol{I}_d \end{bmatrix} \succcurlyeq \begin{bmatrix} 2\boldsymbol{H}_t & \boldsymbol{H}_t \\ \boldsymbol{H}_t & 2\boldsymbol{H}_t \end{bmatrix} \succcurlyeq \frac{m}{M_2}\begin{bmatrix} 2\boldsymbol{I}_d & \boldsymbol{I}_d \\ \boldsymbol{I}_d & 2\boldsymbol{I}_d \end{bmatrix} \succcurlyeq \frac{m}{M_2}\boldsymbol{S}. \tag{39}$$

Differentiating the quadratic form $\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t$ with respect to t yields

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t) = -\boldsymbol{W}_t^\top(\boldsymbol{A}_t^\top \boldsymbol{S} + \boldsymbol{S}\boldsymbol{A}_t)\boldsymbol{W}_t \preccurlyeq -\frac{m}{M_2}(\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t),$$

then Grönwall's inequality generates the exponential decay:

$$\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t \preccurlyeq e^{-\frac{m}{M_2}t}\boldsymbol{W}_0^\top \boldsymbol{S}\boldsymbol{W}_0 \preccurlyeq 2e^{-\frac{m}{M_2}t}\boldsymbol{I}_d,$$

yielding the desired result. ■

Next, we derive the coupling contractivity for the underdamped Langevin dynamics (6).

Lemma 17 *Under Assumption* (P)*, let* $(\boldsymbol{x}_t, \boldsymbol{v}_t)_{t\geqslant 0}$ *and* $(\boldsymbol{x}_t', \boldsymbol{v}_t')_{t\geqslant 0}$ *be the strong solutions to* (6) *driven by the same Brownian motion* $(\boldsymbol{B}_t)_{t\geqslant 0}$*. Then we have almost surely*

$$|\boldsymbol{x}_t' - \boldsymbol{x}_t| \leqslant \sqrt{2}e^{-\frac{mt}{2M_2}}\big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|\big), \quad \forall t \geqslant 0.$$

Proof Lemma 16 has established the matrix inequality

$$\boldsymbol{Q}_t^\top \boldsymbol{Q}_t \preccurlyeq \boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t \preccurlyeq 2e^{-\frac{mt}{M_2}}\boldsymbol{I}_d, \quad \forall t \geqslant 0, \tag{40}$$

implying $\|\boldsymbol{Q}_t\| \leqslant \sqrt{2}e^{-\frac{mt}{2M_2}}$ for all $t \geqslant 0$. This consequently establishes

$$\max\big\{\|D_{\boldsymbol{x}}\boldsymbol{x}_t\|, \|D_{\boldsymbol{v}}\boldsymbol{x}_t\|\big\} \leqslant \sqrt{2}e^{-\frac{mt}{2M_2}}, \quad \forall t \geqslant 0. \tag{41}$$

Consider coupled solutions $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ and $(\boldsymbol{x}_t', \boldsymbol{v}_t')$ driven by the same Brownian motion $(\boldsymbol{B}_t)_{t\geqslant 0}$. For $s \in [0,1]$, define their interpolation as

$$(\boldsymbol{x}_0^s, \boldsymbol{v}_0^s) = (1-s)(\boldsymbol{x}_0, \boldsymbol{v}_0) + s(\boldsymbol{x}_0', \boldsymbol{v}_0') = (\boldsymbol{x}_0, \boldsymbol{v}_0) + s(\boldsymbol{x}_0' - \boldsymbol{x}_0, \boldsymbol{v}_0' - \boldsymbol{v}_0),$$

and let the process $(\boldsymbol{x}_t^s, \boldsymbol{v}_t^s)_{t\geqslant 0}$ solve the underdamped Langevin dynamics (6) with the initial state $(\boldsymbol{x}_0^s, \boldsymbol{v}_0^s)$. Through the chain rule differentiation of $\boldsymbol{x}_t^s$ with respect to the parameter s, we derive the sensitivity equation

$$\frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{x}_t^s = D_{\boldsymbol{x}}\boldsymbol{x}_t^s(\boldsymbol{x}_0' - \boldsymbol{x}_0) + D_{\boldsymbol{v}}\boldsymbol{x}_t^s(\boldsymbol{v}_0' - \boldsymbol{v}_0).$$

Integration over $s \in [0,1]$ provides the error decomposition

$$\boldsymbol{x}_t' - \boldsymbol{x}_t = \int_0^1 \Big(D_{\boldsymbol{x}} \boldsymbol{x}_t^s (\boldsymbol{x}_0' - \boldsymbol{x}_0) + D_{\boldsymbol{v}} \boldsymbol{x}_t^s (\boldsymbol{v}_0' - \boldsymbol{v}_0) \Big) \mathrm{d}s. \tag{42}$$

Finally, applying (41) and (42) generates the contractivity bound:

$$|\boldsymbol{x}_t' - \boldsymbol{x}_t| \leqslant \sqrt{2} e^{-\frac{mt}{2M_2}} \int_0^1 \big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0| \big) \mathrm{d}s = \sqrt{2} e^{-\frac{mt}{2M_2}} \big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0| \big),$$

which completes the proof of Lemma 17. ■

The companion bound on $|\boldsymbol{v}_t' - \boldsymbol{v}_t|$ is not needed. The coupling enters the subsequent estimates only through the differences of $\nabla^2 U$ and ∇f evaluated at $\boldsymbol{x}_t$ and $\boldsymbol{x}_t'$, which the position bound above already controls. An analogous exponential bound for $|\boldsymbol{v}_t' - \boldsymbol{v}_t|$ can be derived, and we omit it in Lemma 17 for convenience.

While the exponential decay of the first variation process $\boldsymbol{W}_t$ is sufficient for bounding the gradient of $u(\boldsymbol{x}, \boldsymbol{v}, t)$, estimating its Hessian requires analyzing the second-order sensitivity of the underdamped Langevin dynamics (6).

A.3 First variation process: Proof of coupling contractivity

We now prove the contractivity of the first variation process $\boldsymbol{W}_t$ with respect to perturbations in the initial state $(\boldsymbol{x}_0, \boldsymbol{v}_0)$. This property provides the necessary second-order information for analyzing the underdamped Langevin dynamics (6).

Lemma 18 *Under Assumption* (P) *with the initial configuration*

$$(\boldsymbol{Q}_0, \boldsymbol{P}_0) = (\boldsymbol{I}_d, \boldsymbol{O}_d) \text{ or } (\boldsymbol{O}_d, \boldsymbol{I}_d),$$

for the first variation processes

$$\begin{aligned} \boldsymbol{W}_t &= \mathrm{FirstVariation}\big(\boldsymbol{Q}_0, \boldsymbol{P}_0, \boldsymbol{x}_0, \boldsymbol{v}_0, (\boldsymbol{B}_t)_{t \geqslant 0}\big), \\ \boldsymbol{W}_t' &= \mathrm{FirstVariation}\big(\boldsymbol{Q}_0, \boldsymbol{P}_0, \boldsymbol{x}_0', \boldsymbol{v}_0', (\boldsymbol{B}_t)_{t \geqslant 0}\big), \end{aligned}$$

their discrepancy satisfies almost surely the contractive bound

$$(\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top \boldsymbol{S} (\boldsymbol{W}_t' - \boldsymbol{W}_t) \preccurlyeq \frac{32 M_3^2}{m^2} e^{-\frac{mt}{4M_2}} \big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0| \big)^2 \boldsymbol{I}_d, \quad \forall t \geqslant 0.$$

Proof Building upon the analytical framework established in Lemma 16, define

$$\boldsymbol{F}_t = (\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top \boldsymbol{S} (\boldsymbol{W}_t' - \boldsymbol{W}_t) \in \mathbb{R}^{d \times d}.$$

Computing the time derivative of $\boldsymbol{F}_t$ with respect to t yields the evolution equation:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{F}_t &= -(\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top \boldsymbol{S} (\boldsymbol{A}_t' \boldsymbol{W}_t' - \boldsymbol{A}_t \boldsymbol{W}_t) + \text{transpose} \\ &= -(\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top \boldsymbol{S} \boldsymbol{A}_t (\boldsymbol{W}_t' - \boldsymbol{W}_t) - (\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top \boldsymbol{S} (\boldsymbol{A}_t' - \boldsymbol{A}_t) \boldsymbol{W}_t' + \text{transpose} \\ &= -\boldsymbol{F}_t^{(1)} - \boldsymbol{F}_t^{(2)}, \end{aligned} \tag{43}$$

with constituent matrices defined as

$$\begin{aligned}\boldsymbol{F}_t^{(1)} &= (\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top (\boldsymbol{A}_t^\top \boldsymbol{S} + \boldsymbol{S}\boldsymbol{A}_t)(\boldsymbol{W}_t' - \boldsymbol{W}_t),\\ \boldsymbol{F}_t^{(2)} &= (\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top \boldsymbol{S}(\boldsymbol{A}_t' - \boldsymbol{A}_t)\boldsymbol{W}_t' + \text{transpose}.\end{aligned}$$

Applying the matrix inequality (39) directly to $\boldsymbol{F}_t^{(1)}$ yields the coercive bound

$$\boldsymbol{F}_t^{(1)} \succcurlyeq \frac{m}{M_2}(\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top \boldsymbol{S}(\boldsymbol{W}_t' - \boldsymbol{W}_t) = \frac{m}{M_2}\boldsymbol{F}_t. \tag{44}$$

For $\boldsymbol{F}_t^{(2)}$, we first decompose it using the expressions of $\boldsymbol{A}_t$ and $\boldsymbol{A}_t'$ in (38),

$$\boldsymbol{F}_t^{(2)} = (\boldsymbol{Q}_t' - \boldsymbol{Q}_t + \boldsymbol{P}_t' - \boldsymbol{P}_t)^\top (\boldsymbol{H}_t' - \boldsymbol{H}_t)\boldsymbol{Q}_t' + \text{transpose}.$$

Applying Young's inequality with parameter $k > 0$ generates

$$\boldsymbol{F}_t^{(2)} \succcurlyeq -\|\boldsymbol{H}_t' - \boldsymbol{H}_t\|\bigg(k(\boldsymbol{Q}_t' - \boldsymbol{Q}_t + \boldsymbol{P}_t' - \boldsymbol{P}_t)^\top(\boldsymbol{Q}_t' - \boldsymbol{Q}_t + \boldsymbol{P}_t' - \boldsymbol{P}_t) + \frac{1}{k}(\boldsymbol{Q}_t')^\top \boldsymbol{Q}_t'\bigg), \tag{45}$$

where Lemma 17 provides the distance bound of $\boldsymbol{H}_t$ and $\boldsymbol{H}_t'$:

$$\|\boldsymbol{H}_t' - \boldsymbol{H}_t\| \leqslant \frac{M_3}{M_2}|\boldsymbol{x}_t' - \boldsymbol{x}_t| \leqslant \frac{\sqrt{2}M_3}{M_2} e^{-\frac{mt}{2M_2}}\big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|\big).$$

Furthermore, the quadratic term in (45) admits the upper bound

$$(\boldsymbol{Q}_t' - \boldsymbol{Q}_t + \boldsymbol{P}_t' - \boldsymbol{P}_t)^\top(\boldsymbol{Q}_t' - \boldsymbol{Q}_t + \boldsymbol{P}_t' - \boldsymbol{P}_t) \preccurlyeq (\boldsymbol{W}_t' - \boldsymbol{W}_t)^\top \boldsymbol{S}(\boldsymbol{W}_t' - \boldsymbol{W}_t) = \boldsymbol{F}_t,$$

and (40) implies the inequality

$$(\boldsymbol{Q}_t')^\top \boldsymbol{Q}_t' \preccurlyeq 2e^{-\frac{mt}{M_2}}\boldsymbol{I}_d.$$

Substituting the estimates above into (45) yields the estimate of $\boldsymbol{F}_t^{(2)}$:

$$\boldsymbol{F}_t^{(2)} \succcurlyeq -\frac{\sqrt{2}M_3}{M_2} e^{-\frac{mt}{2M_2}}\big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|\big)\bigg(k\boldsymbol{F}_t + \frac{2}{k}e^{-\frac{mt}{M_2}}\boldsymbol{I}_d\bigg). \tag{46}$$

We synthesize the estimates of $\boldsymbol{F}_t^{(1)}$ and $\boldsymbol{F}_t^{(2)}$ from (44) and (46) to derive

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}_t \preccurlyeq -\frac{m}{M_2}\boldsymbol{F}_t + \frac{\sqrt{2}M_3}{M_2} e^{-\frac{mt}{2M_2}}\delta\bigg(k\boldsymbol{F}_t + \frac{2}{k}e^{-\frac{mt}{M_2}}\boldsymbol{I}_d\bigg), \tag{47}$$

where $\delta = |\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|$ denotes the initial state discrepancy parameter. Given $\delta > 0$, we choose k to be moderately small to ensure $\boldsymbol{F}_t$ in (47) has a negative coefficient:

$$\frac{m}{2M_2} = \frac{\sqrt{2}M_3}{M_2} e^{-\frac{mt}{2M_2}}\delta k \implies k = \frac{m}{2\sqrt{2}M_3\delta} e^{\frac{mt}{2M_2}}.$$

Substituting this k value into (47) yields the refined differential inequality

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}_t \preccurlyeq -\frac{m}{2M_2}\boldsymbol{F}_t + \frac{8(M_3\delta)^2}{M_2 m}e^{-\frac{2mt}{M_2}}\boldsymbol{I}_d, \qquad \forall t \geqslant 0.$$

To establish the exponential decay of $\boldsymbol{F}_t$, we introduce the exponentially weighted matrix $\boldsymbol{G}_t = e^{\frac{mt}{4M_2}}\boldsymbol{F}_t$, which satisfies the evolution inequality

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{G}_t &= \frac{m}{4M_2}e^{\frac{mt}{4M_2}}\boldsymbol{F}_t + e^{\frac{mt}{4M_2}}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}_t \\ &\preccurlyeq e^{\frac{mt}{4M_2}}\bigg(-\frac{m}{4M_2}\boldsymbol{F}_t + \frac{8(M_3\delta)^2}{M_2 m}e^{-\frac{2mt}{M_2}}\boldsymbol{I}_d\bigg) \\ &\preccurlyeq -\frac{m}{4M_2}\boldsymbol{G}_t + \frac{8(M_3\delta)^2}{M_2 m}\boldsymbol{I}_d.\end{aligned}$$

Since the two processes start from the same initial data, $\boldsymbol{W}_0' = \boldsymbol{W}_0$ and hence $\boldsymbol{F}_0 = \boldsymbol{G}_0 = \boldsymbol{O}_d$. This Grönwall's inequality thus generates the upper bound

$$\boldsymbol{G}_t \preccurlyeq \frac{32M_3^2}{m^2}\delta^2\boldsymbol{I}_d \implies \boldsymbol{F}_t \preccurlyeq \frac{32M_3^2}{m^2}e^{-\frac{mt}{4M_2}}\delta^2\boldsymbol{I}_d,$$

thereby completing the proof of Lemma 18. ■

Applying Lemma 18, we can now provide a quantitative estimate of the gradient and Hessian for the Kolmogorov solution $u(\boldsymbol{x}, \boldsymbol{v}, t)$ defined in (11).

Theorem 19 *Under Assumptions* (P) *and* (T)*, the Kolmogorov solution $u(\boldsymbol{x}, \boldsymbol{v}, t)$ satisfies*

$$\begin{aligned}|u(\boldsymbol{x}, \boldsymbol{v}, t)| &\leqslant Ce^{-\frac{mt}{2M_2}}\bigg(|\boldsymbol{x}| + |\boldsymbol{v}| + \sqrt{\frac{d}{m}}\bigg), \\ \max\big\{|\nabla_{\boldsymbol{x}}u(\boldsymbol{x}, \boldsymbol{v}, t)|, |\nabla_{\boldsymbol{v}}u(\boldsymbol{x}, \boldsymbol{v}, t)|\big\} &\leqslant Ce^{-\frac{mt}{2M_2}}, \\ \max\big\{\big\|\nabla^2_{\boldsymbol{x}\boldsymbol{x}}u(\boldsymbol{x}, \boldsymbol{v}, t)\big\|, \big\|\nabla^2_{\boldsymbol{x}\boldsymbol{v}}u(\boldsymbol{x}, \boldsymbol{v}, t)\big\|, \big\|\nabla^2_{\boldsymbol{v}\boldsymbol{v}}u(\boldsymbol{x}, \boldsymbol{v}, t)\big\|\big\} &\leqslant \frac{C}{m}e^{-\frac{mt}{8M_2}},\end{aligned}$$

where the constant C depends only on M_2, M_3 and $(L_i)_{i=1}^2$.

Proof First applying Lemma 16 with the decay estimate (41), we establish the exponential decay of the first variation processes:

$$\max\big\{\|D_{\boldsymbol{x}}\boldsymbol{x}_t\|, \|D_{\boldsymbol{v}}\boldsymbol{x}_t\|\big\} \leqslant \sqrt{2}e^{-\frac{mt}{2M_2}}, \qquad \forall t \geqslant 0.$$

Combining this with the gradient representation $\nabla_{\boldsymbol{x}}u(\boldsymbol{x}_0, \boldsymbol{v}_0, t)$ in (35), we deduce

$$|\nabla_{\boldsymbol{x}}u(\boldsymbol{x}_0, \boldsymbol{v}_0, t)| \leqslant \mathbb{E}^{(\boldsymbol{x}_0, \boldsymbol{v}_0)}\big[\|D_{\boldsymbol{x}}\boldsymbol{x}_t\||\nabla f(\boldsymbol{x}_t)|\big] \leqslant Ce^{-\frac{mt}{2M_2}}.$$

The same argument bounds $\nabla_{\boldsymbol{v}}u(\boldsymbol{x}, \boldsymbol{v}, t)$: by (35) it is represented through $D_{\boldsymbol{v}}\boldsymbol{x}_t$ in place of $D_{\boldsymbol{x}}\boldsymbol{x}_t$, and the decay estimate covers both. This gives the gradient bound of Theorem 19.

To establish the Hessian estimates, consider two distinct initial states $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ and $(\boldsymbol{x}_0', \boldsymbol{v}_0')$ and define the corresponding solutions $(\boldsymbol{x}_t, \boldsymbol{v}_t)_{t\geqslant 0}$ and $(\boldsymbol{x}_t', \boldsymbol{v}_t')_{t\geqslant 0}$ to the underdamped Langevin dynamics (6) driven by the same Brownian motion $(\boldsymbol{B}_t)_{t\geqslant 0}$. Then

$$\nabla_{\boldsymbol{x}} u(\boldsymbol{x}_0', \boldsymbol{v}_0', t) - \nabla_{\boldsymbol{x}} u(\boldsymbol{x}_0, \boldsymbol{v}_0, t) = \mathbb{E}\Big[(D_{\boldsymbol{x}}\boldsymbol{x}_t')^\top \nabla f(\boldsymbol{x}_t') - (D_{\boldsymbol{x}}\boldsymbol{x}_t)^\top \nabla f(\boldsymbol{x}_t)\Big] = I_1 + I_2, \quad (48)$$

where the expectation is taken over $(\boldsymbol{B}_t)_{t\geqslant 0}$, and the constituent terms decompose as

$$I_1 = \mathbb{E}\Big[(D_{\boldsymbol{x}}\boldsymbol{x}_t' - D_{\boldsymbol{x}}\boldsymbol{x}_t)^\top \nabla f(\boldsymbol{x}_t')\Big], \qquad I_2 = \mathbb{E}\Big[(D_{\boldsymbol{x}}\boldsymbol{x}_t)^\top (\nabla f(\boldsymbol{x}_t') - \nabla f(\boldsymbol{x}_t))\Big].$$

For I_1 estimation, Lemma 18 yields the exponential decay property

$$\big\|D_{\boldsymbol{x}}\boldsymbol{x}_t' - D_{\boldsymbol{x}}\boldsymbol{x}_t\big\| \leqslant \frac{C}{m} e^{-\frac{mt}{8M_2}} \big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|\big),$$

which leads to the inequality

$$|I_1| \leqslant \mathbb{E}\Big[\big\|D_{\boldsymbol{x}}\boldsymbol{x}_t' - D_{\boldsymbol{x}}\boldsymbol{x}_t\big\| |\nabla f(\boldsymbol{x}_t')|\Big] \leqslant \frac{C}{m} e^{-\frac{mt}{8M_2}} \big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|\big). \quad (49)$$

The I_2 term is controlled through Lemma 17 via the chain of inequalities

$$\begin{aligned} |I_2| &\leqslant \mathbb{E}\Big[\|D_{\boldsymbol{x}}\boldsymbol{x}_t\| |\nabla f(\boldsymbol{x}_t') - \nabla f(\boldsymbol{x}_t)|\Big] \leqslant C \cdot \mathbb{E}\Big[\|D_{\boldsymbol{x}}\boldsymbol{x}_t\| |\boldsymbol{x}_t' - \boldsymbol{x}_t|\Big] \\ &\leqslant C \cdot e^{-\frac{mt}{2M_2}} \cdot e^{-\frac{mt}{2M_2}} \big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|\big) \\ &\leqslant C e^{-\frac{mt}{M_2}} \big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|\big). \end{aligned} \quad (50)$$

Synthesizing the estimates in (48)(49)(50), we establish the composite bound

$$\big|\nabla_{\boldsymbol{x}} u(\boldsymbol{x}_0', \boldsymbol{v}_0', t) - \nabla_{\boldsymbol{x}} u(\boldsymbol{x}_0, \boldsymbol{v}_0, t)\big| \leqslant \frac{C}{m} e^{-\frac{mt}{8M_2}} \big(|\boldsymbol{x}_0' - \boldsymbol{x}_0| + |\boldsymbol{v}_0' - \boldsymbol{v}_0|\big).$$

The arbitrariness of initial states $(\boldsymbol{x}_0, \boldsymbol{v}_0)$ and $(\boldsymbol{x}_0', \boldsymbol{v}_0')$ implies the Hessian norm control:

$$\big\|\nabla^2_{\boldsymbol{xx}} u(\boldsymbol{x}, \boldsymbol{v}, t)\big\| \leqslant \frac{C}{m} e^{-\frac{mt}{8M_2}}.$$

The mixed derivative $\nabla^2_{\boldsymbol{xv}} u(\boldsymbol{x}, \boldsymbol{v}, t)$ and the velocity Hessian $\nabla^2_{\boldsymbol{vv}} u(\boldsymbol{x}, \boldsymbol{v}, t)$ follow from the same decomposition with $D_{\boldsymbol{v}}\boldsymbol{x}_t$ in place of $D_{\boldsymbol{x}}\boldsymbol{x}_t$, which (41) bounds identically.

It remains to bound $u(\boldsymbol{x}, \boldsymbol{v}, t)$ itself. Write

$$\pi(\boldsymbol{x}, \boldsymbol{v}) = \pi(\boldsymbol{x}) \left(\frac{M_2}{2\pi}\right)^{\frac{d}{2}} e^{-\frac{M_2}{2}|\boldsymbol{v}|^2} \propto \exp\left(-\frac{M_2}{2}|\boldsymbol{v}|^2 - U(\boldsymbol{x})\right)$$

for the invariant density of (6) on $\mathbb{R}^{2d}$, whose position marginal is $\pi(\boldsymbol{x})$. Then

$$\begin{aligned} |u(\boldsymbol{x}, \boldsymbol{v}, t)| &= \left| \int_{\mathbb{R}^{2d}} \big(u(\boldsymbol{x}, \boldsymbol{v}, t) - u(\boldsymbol{x}', \boldsymbol{v}', t)\big) \pi(\boldsymbol{x}', \boldsymbol{v}') \mathrm{d}\boldsymbol{x}' \mathrm{d}\boldsymbol{v}' \right| \\ &\leqslant C e^{-\frac{mt}{2M_2}} \int_{\mathbb{R}^{2d}} \big(|\boldsymbol{x} - \boldsymbol{x}'| + |\boldsymbol{v} - \boldsymbol{v}'|\big) \pi(\boldsymbol{x}', \boldsymbol{v}') \mathrm{d}\boldsymbol{x}' \mathrm{d}\boldsymbol{v}' \\ &\leqslant C e^{-\frac{mt}{2M_2}} \left(|\boldsymbol{x}| + |\boldsymbol{v}| + \int_{\mathbb{R}^{2d}} \big(|\boldsymbol{x}'| + |\boldsymbol{v}'|\big) \pi(\boldsymbol{x}', \boldsymbol{v}') \mathrm{d}\boldsymbol{x}' \mathrm{d}\boldsymbol{v}'\right) \\ &\leqslant C e^{-\frac{mt}{2M_2}} \left(|\boldsymbol{x}| + |\boldsymbol{v}| + \sqrt{\frac{d}{m}}\right), \end{aligned}$$

where the last inequality uses $\int_{\mathbb{R}^{2d}} |\boldsymbol{x}|\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leqslant \sqrt{d/m}$. ■

Theorem 19 shows that the series defining the discrete Poisson solution

$$\phi_h(\boldsymbol{x},\boldsymbol{v}) = h\sum_{k=0}^{\infty} u(\boldsymbol{x},\boldsymbol{v},kh)$$

and its derivatives $\nabla\phi_h, \nabla^2\phi_h$ converge absolutely at every $(\boldsymbol{x},\boldsymbol{v})$. Summing the resulting geometric series under the condition $h \leqslant \frac{1}{4}$, and using $\frac{h}{1-e^{-ah}} \leqslant \frac{2}{a}$ for $ah \leqslant 1$ with $a = \frac{m}{2M_2}$ and $a = \frac{m}{8M_2}$ respectively, gives the bounds $\frac{C}{m}$ and $\frac{C}{m^2}$ of Theorem 1, concluding its proof.

Appendix B. Uniform-in-time moment bound for SG-UBU

For convenience, introduce the notations

$$\boldsymbol{z} = \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{v}\end{bmatrix} \in \mathbb{R}^{2d}, \quad \boldsymbol{Z}_k = \begin{bmatrix}\boldsymbol{X}_k\\ \boldsymbol{V}_k\end{bmatrix} \in \mathbb{R}^{2d}, \quad \boldsymbol{S} = \begin{bmatrix}2\boldsymbol{I}_d & \boldsymbol{I}_d\\ \boldsymbol{I}_d & \boldsymbol{I}_d\end{bmatrix} \in \mathbb{R}^{2d\times 2d},$$

then the Lyapunov function $\mathcal{V}(\boldsymbol{x},\boldsymbol{v}) = \big(2|\boldsymbol{x}|^2 + 2\boldsymbol{x}^\top\boldsymbol{v} + |\boldsymbol{v}|^2\big)^4$ can be compactly written as

$$\mathcal{V}(\boldsymbol{z}) = (\boldsymbol{z}^\top\boldsymbol{S}\boldsymbol{z})^4. \tag{51}$$

B.1 Lyapunov condition for FG-BU integrator

To begin with, we verify the Lyapunov condition for the FG-BU integrator.

Lemma 20 *Under Assumption* (P)*, let the step size $h \leqslant \frac{1}{4}$; for any $\boldsymbol{z}_0 \in \mathbb{R}^{2d}$,*

$$\mathbb{E}\Big[\mathcal{V}\big((\Phi_h^{\mathfrak{U}}\circ\Phi_h^{\mathfrak{B}})\boldsymbol{z}_0\big)\Big] \leqslant \Big(1 - \frac{mh}{2M_2}\Big)\mathcal{V}(\boldsymbol{z}_0) + \frac{Cd^4h}{m^3},$$

where the constant C depends only on M_1, M_2.

Proof From the expressions of the flow solutions $\Phi_h^{\mathfrak{U}}$ and $\Phi_h^{\mathfrak{B}}$ in Section 2.2, the one-step FG-BU update $\bar{\boldsymbol{z}}_1 = (\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{v}}_1) \in \mathbb{R}^{2d}$ can be explicitly represented as

$$\begin{cases}\bar{\boldsymbol{x}}_1 = \boldsymbol{x}_0 + \dfrac{1-e^{-2h}}{2}\boldsymbol{v}_0 - \dfrac{h(1-e^{-2h})}{2M_2}\nabla U(\boldsymbol{x}_0) + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h \frac{1-e^{-2(h-s)}}{2}\mathrm{d}\boldsymbol{B}_s,\\ \bar{\boldsymbol{v}}_1 = e^{-2h}\boldsymbol{v}_0 - \dfrac{he^{-2h}}{M_2}\nabla U(\boldsymbol{x}_0) + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h e^{-2(h-s)}\mathrm{d}\boldsymbol{B}_s.\end{cases}$$

Using $\nabla U(\boldsymbol{0}) = \boldsymbol{0}$ and the mean value theorem, it holds

$$\frac{1}{M_2}\nabla U(\boldsymbol{x}_0) = \boldsymbol{H}\boldsymbol{x}_0, \qquad \text{where } \boldsymbol{H} = \frac{1}{M_2}\int_0^1 \nabla^2 U(s\boldsymbol{x}_0)\mathrm{d}s \in \mathbb{R}^{d\times d}$$

is positive definite with $\frac{m}{M_2}\boldsymbol{I}_d \preccurlyeq \boldsymbol{H} \preccurlyeq \boldsymbol{I}_d$. Thus the FG-BU update $\bar{\boldsymbol{z}}_1 = (\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{v}}_1)$ reads

$$\bar{\boldsymbol{z}}_1 = \boldsymbol{A}\boldsymbol{z}_0 + \sqrt{h}\boldsymbol{w}, \tag{52}$$

where we employ the notations

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_d - \dfrac{h(1-e^{-2h})}{2}\boldsymbol{H} & \dfrac{1-e^{-2h}}{2}\boldsymbol{I}_d \\ -he^{-2h}\boldsymbol{H} & e^{-2h}\boldsymbol{I}_d \end{bmatrix}, \quad \boldsymbol{w} = \frac{2}{\sqrt{M_2 h}} \begin{bmatrix} \displaystyle\int_0^h \frac{1-e^{-2(h-s)}}{2}\mathrm{d}\boldsymbol{B}_s \\ \displaystyle\int_0^h e^{-2(h-s)}\mathrm{d}\boldsymbol{B}_s \end{bmatrix}.$$

It is clear that $\boldsymbol{w}$ is a Gaussian random variable with

$$\mathbb{E}[\boldsymbol{w}] = \boldsymbol{0}, \qquad \mathbb{E}\big[|\boldsymbol{w}|^8\big] \leqslant Cd^4,$$

where the constant C depends only on M_1, M_2.

Next, we prove that for any step size $h \leqslant \frac{1}{4}$, the following matrix inequality holds:

$$\boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A} \preccurlyeq \left(1 - \frac{mh}{2M_2}\right)\boldsymbol{S}. \tag{53}$$

The inequality (53) serves as the coercivity condition for the FG-BU integrator. Direct calculation yields the upper bound of $\boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A}$:

$$\begin{aligned} \boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A} &= \begin{bmatrix} 2\boldsymbol{I}_d - 2h\boldsymbol{H} + \dfrac{1+e^{-4h}}{2}h^2\boldsymbol{H}^2 & \boldsymbol{I}_d - \dfrac{1+e^{-4h}}{2}h\boldsymbol{H} \\ \boldsymbol{I}_d - \dfrac{1+e^{-4h}}{2}h\boldsymbol{H} & \dfrac{1+e^{-4h}}{2}\boldsymbol{I}_d \end{bmatrix} \\ &\preccurlyeq \begin{bmatrix} 2\boldsymbol{I}_d - \dfrac{7}{4}h\boldsymbol{H} & \boldsymbol{I}_d - \dfrac{1+e^{-4h}}{2}h\boldsymbol{H} \\ \boldsymbol{I}_d - \dfrac{1+e^{-4h}}{2}h\boldsymbol{H} & \dfrac{1+e^{-4h}}{2}\boldsymbol{I}_d \end{bmatrix}. \qquad (\text{using } h \leqslant \tfrac{1}{4} \text{ and } \boldsymbol{H} \preccurlyeq \boldsymbol{I}_d) \end{aligned}$$

Thus using the expression of $\boldsymbol{S} \in \mathbb{R}^{2d\times 2d}$, there holds

$$\begin{aligned} \boldsymbol{S} - \boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A} &\succcurlyeq \begin{bmatrix} \dfrac{7}{4}h\boldsymbol{H} & \dfrac{1+e^{-4h}}{2}h\boldsymbol{H} \\ \dfrac{1+e^{-4h}}{2}h\boldsymbol{H} & \dfrac{1-e^{-4h}}{2}\boldsymbol{I}_d \end{bmatrix} \\ &\succcurlyeq \begin{bmatrix} \dfrac{7}{4}h\boldsymbol{H} & \dfrac{1+e^{-4h}}{2}h\boldsymbol{H} \\ \dfrac{1+e^{-4h}}{2}h\boldsymbol{H} & h\boldsymbol{H} \end{bmatrix} \qquad (\text{using } \frac{1-e^{-4h}}{2} \geqslant h \text{ and } \boldsymbol{I}_d \succcurlyeq \boldsymbol{H}) \\ &\succcurlyeq \frac{mh}{M_2}\begin{bmatrix} \dfrac{7}{4}\boldsymbol{I}_d & \dfrac{1+e^{-4h}}{2}\boldsymbol{I}_d \\ \dfrac{1+e^{-4h}}{2}\boldsymbol{I}_d & \boldsymbol{I}_d \end{bmatrix} \succcurlyeq \frac{mh}{2M_2}\boldsymbol{S}, \qquad (\text{using } \boldsymbol{H} \succcurlyeq \frac{m}{M_2}I_d) \end{aligned}$$

where the last inequality follows from

$$\begin{bmatrix} \dfrac{7}{4} & \dfrac{1+e^{-4h}}{2} \\ \dfrac{1+e^{-4h}}{2} & 1 \end{bmatrix} \succcurlyeq \frac{1}{2}\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \iff \begin{bmatrix} \dfrac{3}{2} & e^{-4h} \\ e^{-4h} & 1 \end{bmatrix} \succcurlyeq 0 \iff \frac{3}{2} - e^{-8h} \geqslant 0.$$

This completes the proof of inequality (53).

The positive definite matrix $\boldsymbol{S}$ admits a unique symmetric square root

$$\sqrt{\boldsymbol{S}} = \frac{1}{\sqrt{5}} \begin{bmatrix} 3\boldsymbol{I}_d & \boldsymbol{I}_d \\ \boldsymbol{I}_d & 2\boldsymbol{I}_d \end{bmatrix} \in \mathbb{R}^{2d\times 2d}.$$

This gives the representation $\boldsymbol{z}^\top \boldsymbol{S}\boldsymbol{z} = \big|\sqrt{\boldsymbol{S}}\boldsymbol{z}\big|^2$ for $\boldsymbol{z} \in \mathbb{R}^{2d}$. Using the identity $\bar{\boldsymbol{z}}_1 = \boldsymbol{A}\boldsymbol{z}_0 + \sqrt{h}\boldsymbol{w}$, we obtain the moment bound for the one-step FG-BU update $\boldsymbol{z}_1$:

$$\begin{aligned}
\mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^8 &= \mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0 + \sqrt{h\boldsymbol{S}}\boldsymbol{w}\big|^8 \\
&\leqslant \big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^8 + 28\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^6 h\cdot\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{w}\big|^2 + 70\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^4 h^2\cdot\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{w}\big|^4 \\
&\quad + 28\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^2 h^3\cdot\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{w}\big|^6 + h^4\cdot\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{w}\big|^8 \\
&\leqslant \big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^8 + Cdh\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^6 + Cd^2h^2\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^4 + Cd^3h^3\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^2 + Cd^4h^4 \\
&\leqslant \big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^8 + \frac{mh}{2M_2}\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|^8 + \frac{Cd^4h}{m^3},
\end{aligned}$$

where we have utilized the binomial theorem, the vanishing of odd powers of $\boldsymbol{w}$ in the expansion, and the Cauchy–Schwarz inequality to bound the surviving inner products by $\big|\sqrt{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{z}_0\big|\big|\sqrt{\boldsymbol{S}}\boldsymbol{w}\big|$. Then the contractivity inequality (53) yields the moment estimate

$$\begin{aligned}
\mathbb{E}\big[(\bar{\boldsymbol{z}}_1^\top \boldsymbol{S}\bar{\boldsymbol{z}}_1)^4\big] &\leqslant \left(1 + \frac{mh}{2M_2}\right)\mathbb{E}\big[(\boldsymbol{z}_0^\top \boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A}\boldsymbol{z}_0)^4\big] + \frac{Cd^4h}{m^3} \\
&\leqslant \left(1 + \frac{mh}{2M_2}\right)\left(1 - \frac{mh}{2M_2}\right)^4 (\boldsymbol{z}_0^\top \boldsymbol{S}\boldsymbol{z}_0)^4 + \frac{Cd^4h}{m^3} \\
&\leqslant \left(1 - \frac{mh}{2M_2}\right)(\boldsymbol{z}_0^\top \boldsymbol{S}\boldsymbol{z}_0)^4 + \frac{Cd^4h}{m^3},
\end{aligned}$$

concluding the proof of the Lyapunov condition for the FG-BU integrator. ■

Notably, Lemma 20 establishes the Lyapunov condition directly for the FG-BU integrator. While a corresponding condition can be derived for the continuous underdamped Langevin dynamics (6), our analytical framework bypasses its continuous counterpart, operating entirely at the discrete level.

B.2 Lyapunov condition for SG-BU integrator

Based on Lemma 20 and the fact that SG-BU employs an unbiased stochastic gradient, we can conveniently establish the Lyapunov condition for the SG-BU integrator. This approach also applies to variance-reduced stochastic gradient sampling algorithms, as the unbiasedness property still holds (see Lemmas 25 and 29).

Proof of Lemma 2 Following the proof of Lemma 20, we denote $\bar{\boldsymbol{z}}_1 = (\Phi_h^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}})\boldsymbol{z}_0$ and $\boldsymbol{z}_1 = (\Phi_h^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\theta))\boldsymbol{z}_0$ as the one-step updates of FG-BU and SG-BU. Then

$$\begin{cases}
\boldsymbol{x}_1 = \boldsymbol{x}_0 + \dfrac{1-e^{-2h}}{2}\boldsymbol{v}_0 - \dfrac{h(1-e^{-2h})}{2M_2}b(\boldsymbol{x}_0,\theta) + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h \frac{1-e^{-2(h-s)}}{2}\mathrm{d}\boldsymbol{B}_s, \\
\boldsymbol{v}_1 = e^{-2h}\boldsymbol{v}_0 - \dfrac{he^{-2h}}{M_2}b(\boldsymbol{x}_0,\theta) + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h e^{-2(h-s)}\mathrm{d}\boldsymbol{B}_s.
\end{cases}$$

Thus we can write the stochastic gradient error as

$$\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1 = h\boldsymbol{\Delta}_{\boldsymbol{x}}, \qquad \boldsymbol{v}_1 - \bar{\boldsymbol{v}}_1 = h\boldsymbol{\Delta}_{\boldsymbol{v}},$$

where the error terms are given by

$$\boldsymbol{\Delta}_{\boldsymbol{x}} = -\frac{1-e^{-2h}}{2M_2}(b(\boldsymbol{x}_0,\theta) - \nabla U(\boldsymbol{x}_0)), \qquad \boldsymbol{\Delta}_{\boldsymbol{v}} = -\frac{e^{-2h}}{M_2}(b(\boldsymbol{x}_0,\theta) - \nabla U(\boldsymbol{x}_0)).$$

The proof below uses only the unbiasedness and the eighth moment bound of $b(\boldsymbol{x},\theta)-\nabla U(\boldsymbol{x})$ in Assumption (Sg), not the almost-sure growth bound $|b(\boldsymbol{x},\theta)| \leqslant M_1\sqrt{|\boldsymbol{x}|^2+d}$. From Assumption (Sg) we obtain the unbiased condition and the moment bound

$$\mathbb{E}[\boldsymbol{\Delta}_{\boldsymbol{x}}] = \mathbb{E}[\boldsymbol{\Delta}_{\boldsymbol{v}}] = \mathbf{0}, \quad \mathbb{E}\big[|\boldsymbol{\Delta}_{\boldsymbol{x}}|^8\big] \leqslant Cd^4h^8, \quad \mathbb{E}\big[|\boldsymbol{\Delta}_{\boldsymbol{v}}|^8\big] \leqslant Cd^4.$$

Define $\boldsymbol{\Delta}_{\boldsymbol{z}} = (\boldsymbol{\Delta}_{\boldsymbol{x}}, \boldsymbol{\Delta}_{\boldsymbol{v}}) \in \mathbb{R}^{2d}$, giving $\boldsymbol{z}_1 = \bar{\boldsymbol{z}}_1 + h\boldsymbol{\Delta}_{\boldsymbol{z}}$ with

$$\mathbb{E}[\boldsymbol{\Delta}_{\boldsymbol{z}}] = \mathbf{0}, \quad \mathbb{E}\big[|\boldsymbol{\Delta}_{\boldsymbol{z}}|^8\big] \leqslant Cd^4.$$

Exploiting the Lyapunov condition of the FG-BU integrator in Lemma 20, we have

$$\begin{aligned}
\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{z}_1\big|^8 &= \mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1 + h\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_{\boldsymbol{z}}\big|^8 \\
&\leqslant \mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^8 + 28h^2\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^6\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_{\boldsymbol{z}}\big|^2\Big) \\
&\qquad + 56h^3\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^5\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_{\boldsymbol{z}}\big|^3\Big) + 70h^4\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^4\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_{\boldsymbol{z}}\big|^4\Big) \\
&\qquad + 56h^5\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^3\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_{\boldsymbol{z}}\big|^5\Big) + 28h^6\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^2\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_{\boldsymbol{z}}\big|^6\Big) \\
&\qquad + 8h^7\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_{\boldsymbol{z}}\big|^7\Big) + h^8\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_{\boldsymbol{z}}\big|^8 \\
&\leqslant \mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^8 + \frac{mh}{12M_2}\mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{z}}_1\big|^8 + \frac{Ch}{m^3}\mathbb{E}\big[|\boldsymbol{\Delta}_{\boldsymbol{z}}|^8\big] \\
&\leqslant \left(1+\frac{mh}{12M_2}\right)\left(\left(1-\frac{mh}{2M_2}\right)\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{z}_0\big|^8 + \frac{Cd^4h}{m^3}\right) + \frac{Ch}{m^3}\mathbb{E}\big[|\boldsymbol{\Delta}_{\boldsymbol{z}}|^8\big] \\
&\leqslant \left(1-\frac{mh}{3M_2}\right)\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{z}_0\big|^8 + \frac{Cd^4h}{m^3}.
\end{aligned}$$

The term linear in $\boldsymbol{\Delta}_{\boldsymbol{z}}$ is dropped from the binomial expansion because $\mathbb{E}[\boldsymbol{\Delta}_{\boldsymbol{z}}] = \mathbf{0}$, and the remaining terms are absorbed into the first and last ones by Young's inequality. Recalling the Lyapunov function $\mathcal{V}(\boldsymbol{z}) = (\boldsymbol{z}^\top\boldsymbol{S}\boldsymbol{z})^4$, the inequality above implies

$$\mathbb{E}\big[\mathcal{V}(\boldsymbol{z}_1)\big] \leqslant \left(1-\frac{mh}{3M_2}\right)\mathcal{V}(\boldsymbol{z}_0) + \frac{Cd^4h}{m^3},$$

which completes the proof of Lemma 2. ■

As noted at the start of its proof, Lemma 2 applies to any unbiased stochastic gradient whose error satisfies the eighth moment bound of Assumption (Sg), and it is in this form that it is invoked for the variance-reduced integrators.

B.3 Uniform-in-time moment bound of SG-UBU integrator

Next, it is convenient to derive the uniform-in-time moment bound for the SG-UBU integrator by simply applying an additional $\Phi_{h/2}^{\mathfrak{U}}$ flow on the SG-BU integrator.

Proof of Theorem 3 For the step size $h \leqslant \frac{1}{4}$, the following bounds hold for any $\boldsymbol{z}_0 \in \mathbb{R}^{2d}$:

$$\mathbb{E}\big[\mathcal{V}\big(\Phi_{h/2}^{\mathfrak{U}}\boldsymbol{z}_0\big)\big] \leqslant C\mathcal{V}(\boldsymbol{z}_0) + Cd^4, \quad \mathbb{E}\big[\mathcal{V}\big(\Phi_h^{\mathfrak{B}}(\theta)\boldsymbol{z}_0\big)\big] \leqslant C\mathcal{V}(\boldsymbol{z}_0) + Cd^4,$$

where C depends only on M_1, M_2. These inequalities can be easily found from the explicit expressions of the solution flows $\Phi_{h/2}^{\mathfrak{U}}$ and $\Phi_h^{\mathfrak{B}}(\theta)$. Then for $\boldsymbol{X}_0 = \boldsymbol{0}$ and $\boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d)$, we have the moment bound

$$\mathbb{E}\big[\mathcal{V}\big(\Phi_{h/2}^{\mathfrak{U}}\boldsymbol{Z}_0\big)\big] \leqslant \frac{Cd^4}{m^4}.$$

For $k \geqslant 2$, the SG-UBU solution is written as

$$\boldsymbol{Z}_k = \Phi_{h/2}^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\theta_{k-1}) \circ \bigg(\prod_{l=0}^{k-2} \Phi_h^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\theta_l)\bigg) \circ \Phi_{h/2}^{\mathfrak{U}}\boldsymbol{Z}_0.$$

The Lyapunov condition of SG-BU provided in Lemma 2 thus yields the bound for $k \geqslant 2$, while $k = 0$ and $k = 1$ are covered by the two one-step inequalities above; hence

$$\sup_{k\geqslant 0} \mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_k)\big] \leqslant \frac{Cd^4}{m^4},$$

which completes the proof of Theorem 3. ■

Finally, we employ Theorem 3 to show the existence and moment bound for the invariant distribution of SG-UBU.

Proof of Lemma 4 Employing the Krylov–Bogolyubov theorem (Theorem 4.21, Hairer (2006)), when the step size $h \leqslant \frac{1}{4}$, the Lyapunov condition in Lemma 2 implies that SG-BU has an invariant distribution $\hat{\pi}_h(\boldsymbol{x}, \boldsymbol{v})$ satisfying

$$\int_{\mathbb{R}^{2d}} |\boldsymbol{z}|^8 \hat{\pi}_h(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \leqslant \frac{Cd^4}{m^4}. \tag{54}$$

Define distribution semigroups for $\Phi_t^{\mathfrak{U}}$ and $\Phi_t^{\mathfrak{B}}(\theta)$:

$$\begin{aligned} P_t^{\mathfrak{U}}\mu &:= \text{distribution of } \Phi_t^{\mathfrak{U}}\boldsymbol{z}_0 \ \text{ with } \boldsymbol{z}_0 \sim \mu, \\ P_t^{\mathfrak{B}}\mu &:= \mathbb{E}_\theta\Big[\text{distribution of } \Phi_t^{\mathfrak{B}}(\theta)\boldsymbol{z}_0 \text{ with } \boldsymbol{z}_0 \sim \mu\Big]. \end{aligned}$$

Then the invariant distribution $\hat{\pi}_h$ of SG-BU satisfies:

$$P_h^{\mathfrak{U}} P_h^{\mathfrak{B}} \hat{\pi}_h = \hat{\pi}_h.$$

Now define the distribution π_h in $\mathbb{R}^{2d}$ as

$$\pi_h = P_{h/2}^{\mathfrak{U}} P_h^{\mathfrak{B}} \hat{\pi}_h.$$

This π_h satisfies the same moment bound (54), and the identity

$$\pi_h = P_{h/2}^{\mathfrak{U}} P_h^{\mathfrak{B}} P_h^{\mathfrak{U}} P_h^{\mathfrak{B}} \hat{\pi}_h = P_{h/2}^{\mathfrak{U}} P_h^{\mathfrak{B}} P_{h/2}^{\mathfrak{U}} \pi_h,$$

confirming that π_h is an invariant distribution of SG-UBU. ■

It should be noted that while Lemma 4 establishes the existence of an invariant distribution, it does not guarantee uniqueness, a property that is not required for the subsequent analysis.

Appendix C. Local error analysis for SG-UBU

This section provides fourth moment estimates for the stochastic gradient error, $\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}$, and the discretization error, $\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)$. The analysis requires comparing the explicit formulas for the one-step evolution of the exact solution, the full gradient update, and the stochastic gradient update, which we present below for clarity.

Exact solution in one-step: $\boldsymbol{Z}_k(h)$

$$\begin{cases} \boldsymbol{X}_k(h) = \boldsymbol{X}_k + \dfrac{1-e^{-2h}}{2}\boldsymbol{V}_k - \dfrac{1}{M_2}\displaystyle\int_0^h \frac{1-e^{-2(h-s)}}{2}\nabla U(\boldsymbol{X}_k(s))\mathrm{d}s \\ \qquad\qquad + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h \frac{1-e^{-2(h-s)}}{2}\mathrm{d}\boldsymbol{B}_{kh+s}, \\ \boldsymbol{V}_k(h) = e^{-2h}\boldsymbol{V}_k - \dfrac{1}{M_2}\displaystyle\int_0^h e^{-2(h-s)}\nabla U(\boldsymbol{X}_k(s))\mathrm{d}s + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h e^{-2(h-s)}\mathrm{d}\boldsymbol{B}_{kh+s}. \end{cases}$$

One-step of FG-UBU update: $\bar{\boldsymbol{Z}}_{k+1}$

$$\begin{cases} \bar{\boldsymbol{X}}_{k+1} = \boldsymbol{X}_k + \dfrac{1-e^{-2h}}{2}\boldsymbol{V}_k - \dfrac{h(1-e^{-h})}{2M_2}\nabla U(\boldsymbol{Y}_k) + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h \frac{1-e^{-2(h-s)}}{2}\mathrm{d}\boldsymbol{B}_{kh+s}, \\ \bar{\boldsymbol{V}}_{k+1} = e^{-2h}\boldsymbol{V}_k - \dfrac{he^{-h}}{M_2}\nabla U(\boldsymbol{Y}_k) + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h e^{-2(h-s)}\mathrm{d}\boldsymbol{B}_{kh+s}. \end{cases}$$

One-step of SG-UBU update: $\boldsymbol{Z}_{k+1}$

$$\begin{cases} \boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \dfrac{1-e^{-2h}}{2}\boldsymbol{V}_k - \dfrac{h(1-e^{-h})}{2M_2}b(\boldsymbol{Y}_k,\theta_k) + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h \frac{1-e^{-2(h-s)}}{2}\mathrm{d}\boldsymbol{B}_{kh+s}, \\ \boldsymbol{V}_{k+1} = e^{-2h}\boldsymbol{V}_k - \dfrac{he^{-h}}{M_2}b(\boldsymbol{Y}_k,\theta_k) + \dfrac{2}{\sqrt{M_2}}\displaystyle\int_0^h e^{-2(h-s)}\mathrm{d}\boldsymbol{B}_{kh+s}. \end{cases}$$

These expressions are derived by applying the $\Phi_t^{\mathfrak{U}}$ and $\Phi_t^{\mathfrak{B}}$ solution flows in the UBU integrator, and can also be found directly in Example 8 of Sanz-Serna and Zygalakis (2021). Recall that $\boldsymbol{Y}_k$ is the intermediate position defined in (9), and the stochastic gradient indices $(\theta_k)_{k=0}^{\infty}$ are generated independently from $\mathbb{P}_\theta$.

C.1 Estimate of stochastic gradient error

With the explicit expression for the stochastic gradient error established in (20), we now prove the corresponding moment bounds in Lemma 5.

Proof of Lemma 5 Under Assumption (Sg), the gradient expression (20) gives

$$\mathbb{E}\big|\boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1}\big|^4 \leqslant Ch^8\mathbb{E}\big|b(\boldsymbol{Y}_k, \theta_k) - \nabla U(\boldsymbol{Y}_k)\big|^4 \leqslant C\sigma^4 d^2 h^8,$$

yielding the position stochastic gradient error bound

$$\sqrt{\mathbb{E}\big|\boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1}\big|^4} \leqslant C\sigma^2 dh^4.$$

Similarly, the velocity error bound follows

$$\sqrt{\mathbb{E}\big|\boldsymbol{V}_{k+1} - \bar{\boldsymbol{V}}_{k+1}\big|^4} \leqslant C\sigma^2 dh^2,$$

which completes the proof. ■

Notably, the position and velocity components of the stochastic gradient error converge at different rates with respect to the step size h. This discrepancy in order is a direct consequence of their respective coefficients in the explicit form (20).

C.2 Estimate of discretization error

Our analysis of the fourth moment of the discretization error relies on the explicit expressions for the local errors $\delta\boldsymbol{X}_k$ and $\delta\boldsymbol{V}_k$ provided in Equations (40) and (42) of Sanz-Serna and Zygalakis (2021). The proof relies on a delicate decomposition of the discretization error using integration by parts.

Proof of Lemma 6 Utilizing the explicit expressions of $\bar{\boldsymbol{Z}}_{k+1}$ and $\boldsymbol{Z}_k(h)$, the discretization errors $\delta\boldsymbol{X}_k = \bar{\boldsymbol{X}}_{k+1} - \boldsymbol{X}_k(h)$ and $\delta\boldsymbol{V}_k = \bar{\boldsymbol{V}}_{k+1} - \boldsymbol{V}_k(h)$ are written as

$$\delta\boldsymbol{X}_k = \frac{h(1-e^{-h})}{2M_2}\big(\nabla U(\boldsymbol{X}_k(h/2)) - \nabla U(\boldsymbol{Y}_k)\big) + \boldsymbol{I}_6 + \boldsymbol{I}_7, \tag{55a}$$

$$\delta\boldsymbol{V}_k = \frac{he^{-h}}{M_2}\big(\nabla U(\boldsymbol{X}_k(h/2)) - \nabla U(\boldsymbol{Y}_k)\big) + \boldsymbol{I}_1 + \boldsymbol{I}_2 + \boldsymbol{I}_3 + \boldsymbol{I}_4 + \boldsymbol{I}_5, \tag{55b}$$

with the error terms $\boldsymbol{I}_1$ to $\boldsymbol{I}_7$ given by

$$\boldsymbol{I}_1 = \frac{4}{M_2}\int_{h/2}^{h}\mathrm{d}s\int_{h/2}^{s}\mathrm{d}s'\int_{h-s'}^{s'} e^{-2(h-s'')}\nabla U(\boldsymbol{X}_k(s''))\mathrm{d}s'',$$

$$\boldsymbol{I}_2 = \frac{2}{M_2}\int_{h/2}^{h}\mathrm{d}s\int_{h/2}^{s}\mathrm{d}s'\int_{h-s'}^{s'} e^{-2(h-s'')}\nabla^2 U(\boldsymbol{X}_k(s''))\boldsymbol{V}_k(s'')\mathrm{d}s'',$$

$$\boldsymbol{I}_3 = \frac{1}{M_2}\int_{h/2}^{h}\mathrm{d}s\int_{h/2}^{s}\mathrm{d}s'\int_{h-s'}^{s'} e^{-2(h-s'')}\boldsymbol{V}_k(s'')^\top\nabla^3 U(\boldsymbol{X}_k(s''))\boldsymbol{V}_k(s'')\mathrm{d}s'',$$

$$\boldsymbol{I}_4 = -\frac{1}{M_2^2}\int_{h/2}^{h}\mathrm{d}s\int_{h/2}^{s}\mathrm{d}s'\int_{h-s'}^{s'} e^{-2(h-s'')}\nabla^2 U(\boldsymbol{X}_k(s''))\nabla U(\boldsymbol{X}_k(s''))\mathrm{d}s'',$$

$$\boldsymbol{I}_5 = \frac{2}{M_2^{3/2}} \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathrm{d}s' \int_{h-s'}^{s'} e^{-2(h-s'')} \nabla^2 U(\boldsymbol{X}_k(s'')) \mathrm{d}\boldsymbol{B}_{kh+s''},$$
$$\boldsymbol{I}_6 = \frac{1}{M_2} \int_{0}^{h} \mathrm{d}s \int_{h/2}^{s} \frac{1 - e^{-2(h-s')}}{2} \nabla^2 U(\boldsymbol{X}_k(s')) \boldsymbol{V}_k(s') \mathrm{d}s',$$
$$\boldsymbol{I}_7 = -\frac{1}{M_2} \int_{0}^{h} \mathrm{d}s \int_{h/2}^{s} e^{-2(h-s')} \nabla U(\boldsymbol{X}_k(s')) \mathrm{d}s'.$$

For convenience, we define the deviation of $\boldsymbol{Y}_k$ by

$$\delta \boldsymbol{Y}_k = \boldsymbol{Y}_k - \boldsymbol{X}_k(h/2) = \frac{1}{M_2} \int_{0}^{h/2} \frac{1 - e^{-(h-2s)}}{2} \nabla U(\boldsymbol{X}_k(s)) \mathrm{d}s,$$

then its norm $|\delta \boldsymbol{Y}_k|$ admits the upper bound

$$|\delta \boldsymbol{Y}_k| \leqslant Ch \int_{0}^{h/2} |\nabla U(\boldsymbol{X}_k(s))| \mathrm{d}s.$$

From the uniform-in-time moment bound in Theorem 3 and Hölder's inequality we derive

$$\mathbb{E}|\delta \boldsymbol{Y}_k|^4 \leqslant Ch^7 \int_{0}^{h/2} \mathbb{E}|\nabla U(\boldsymbol{X}_k(s))|^4 \mathrm{d}s \leqslant Ch^7 \int_{0}^{h/2} \mathbb{E}\big(|\boldsymbol{X}_k(s)|^2 + d\big)^2 \mathrm{d}s \leqslant \frac{Cd^2h^8}{m^2}.$$

Applying the same approach, we establish fourth moment bounds of $\boldsymbol{I}_1$ to $\boldsymbol{I}_4$:

$$\mathbb{E}|\boldsymbol{I}_1|^4 \leqslant Ch^9 \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathrm{d}s' \int_{h-s'}^{s'} \mathbb{E}|\nabla U(\boldsymbol{X}_k(s''))|^4 \mathrm{d}s'' \leqslant \frac{Cd^2h^{12}}{m^2},$$
$$\mathbb{E}|\boldsymbol{I}_2|^4 \leqslant Ch^9 \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathrm{d}s' \int_{h-s'}^{s'} \mathbb{E}|\boldsymbol{V}_k(s'')|^4 \mathrm{d}s'' \leqslant \frac{Cd^2h^{12}}{m^2},$$
$$\mathbb{E}|\boldsymbol{I}_3|^4 \leqslant Ch^9 \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathrm{d}s' \int_{h-s'}^{s'} \mathbb{E}|\boldsymbol{V}_k(s'')|^8 \mathrm{d}s'' \leqslant \frac{Cd^4h^{12}}{m^4},$$
$$\mathbb{E}|\boldsymbol{I}_4|^4 \leqslant Ch^9 \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathrm{d}s' \int_{h-s'}^{s'} \mathbb{E}\big|\nabla U(\boldsymbol{X}_k(s''))\big|^4 \mathrm{d}s'' \leqslant \frac{Cd^2h^{12}}{m^2}.$$

For the stochastic integral $\boldsymbol{I}_5$, its second moment is controlled via

$$\begin{aligned}
\mathbb{E}|\boldsymbol{I}_5|^2 &\leqslant Ch^2 \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathbb{E}\bigg| \int_{h-s'}^{s'} e^{-2(h-s'')} \nabla^2 U(\boldsymbol{X}_k(s'')) \mathrm{d}\boldsymbol{B}_{kh+s''} \bigg|^2 \mathrm{d}s' \\
&\leqslant Ch^2 \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathrm{d}s' \int_{h-s'}^{s'} \mathbb{E}\big\|\nabla^2 U(\boldsymbol{X}_k(s''))\big\|_{\mathrm{F}}^2 \mathrm{d}s'' \\
&\leqslant Ch^2 \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathrm{d}s' \int_{h-s'}^{s'} M_2^2 d \, \mathrm{d}s'' \leqslant Cdh^5,
\end{aligned}$$

where the Itô isometry produces the Frobenius norm of the Hessian, bounded by $\|\nabla^2 U\|_{\mathrm{F}}^2 \leqslant d \|\nabla^2 U\|^2 \leqslant M_2^2 d$ under Assumption (P). Writing $\boldsymbol{J}(s') = \int_{h-s'}^{s'} e^{-2(h-s'')} \nabla^2 U(\boldsymbol{X}_k(s'')) \mathrm{d}\boldsymbol{B}_{kh+s''}$, the Burkholder–Davis–Gundy inequality (Karatzas and Shreve, 1991) gives

$$\mathbb{E}|\boldsymbol{J}(s')|^4 \leqslant C \, \mathbb{E}\bigg(\int_{h-s'}^{s'} \big\|\nabla^2 U(\boldsymbol{X}_k(s''))\big\|_{\mathrm{F}}^2 \mathrm{d}s'' \bigg)^2 \leqslant C\big(M_2^2 dh\big)^2,$$

and Cauchy's inequality on the outer double integral yields the fourth moment

$$\mathbb{E}|\boldsymbol{I}_5|^4 \leqslant Ch^6 \int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathbb{E}|\boldsymbol{J}(s')|^4 \mathrm{d}s' \leqslant Cd^2h^{10}. \tag{56}$$

Therefore, in the discretization error $\delta \boldsymbol{V}_k = \bar{\boldsymbol{V}}_{k+1} - \boldsymbol{V}_k(h)$, we select the martingale component $\delta_m \boldsymbol{V}_k = \boldsymbol{I}_5$ with the high-order term $\delta_h \boldsymbol{V}_k$ as

$$\delta_h \boldsymbol{V}_k = \frac{he^{-h}}{M_2}\big(\nabla U(\boldsymbol{X}_k(h/2)) - \nabla U(\boldsymbol{Y}_k)\big) + \boldsymbol{I}_1 + \boldsymbol{I}_2 + \boldsymbol{I}_3 + \boldsymbol{I}_4.$$

The fourth moment of $\delta_h \boldsymbol{V}_k$ then satisfies

$$\mathbb{E}|\delta_h \boldsymbol{V}_k|^4 \leqslant Ch^4 \mathbb{E}|\delta \boldsymbol{Y}_k|^4 + C\sum_{i=1}^{4} \mathbb{E}|\boldsymbol{I}_i|^4 \leqslant \frac{Cd^4h^{12}}{m^4}. \tag{57}$$

Finally, we bound the fourth moment of $\boldsymbol{I}_6$ and $\boldsymbol{I}_7$. Observing the inequality

$$|\boldsymbol{I}_6| \leqslant Ch \int_0^h \mathrm{d}s \left| \int_{h/2}^{s} |\boldsymbol{V}_k(s')| \mathrm{d}s' \right|,$$

then Hölder's inequality yields the moment bound

$$\mathbb{E}|\boldsymbol{I}_6|^4 \leqslant Ch^{10} \int_0^h \mathrm{d}s \left| \int_{h/2}^{s} \mathbb{E}|\boldsymbol{V}_k(s')|^4 \mathrm{d}s' \right| \leqslant \frac{Cd^2h^{12}}{m^2}.$$

Next, from the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big(e^{2s}\nabla U(\boldsymbol{X}_k(s))\Big) = e^{2s}\Big(2\nabla U(\boldsymbol{X}_k(s)) + \nabla^2 U(\boldsymbol{X}_k(s))\boldsymbol{V}_k(s)\Big),$$

we deduce the equality

$$\boldsymbol{I}_7 = -\frac{1}{M_2}\int_{h/2}^{h} \mathrm{d}s \int_{h/2}^{s} \mathrm{d}s' \int_{h-s'}^{s'} \frac{\mathrm{d}}{\mathrm{d}s''}\Big(e^{-2(h-s'')}\nabla U(\boldsymbol{X}_k(s''))\Big)\mathrm{d}s'' = -\frac{1}{2}(\boldsymbol{I}_1 + \boldsymbol{I}_2).$$

Hence the fourth moment of $\boldsymbol{I}_7$ is bounded by

$$\mathbb{E}|\boldsymbol{I}_7|^4 \leqslant \frac{Cd^2h^{12}}{m^2}.$$

In total, the moment bound in the position variable reads

$$\mathbb{E}|\delta \boldsymbol{X}_k|^4 \leqslant C\big(h^8\,\mathbb{E}|\delta \boldsymbol{Y}_k|^4 + \mathbb{E}|\boldsymbol{I}_6|^4 + \mathbb{E}|\boldsymbol{I}_7|^4\big) \leqslant \frac{Cd^2h^{12}}{m^2}. \tag{58}$$

Inequalities (56)–(58) complete the proof of Lemma 6. ■

We note that the estimate for the term $\boldsymbol{I}_3$ in the preceding proof requires both the bound on the third derivative, $\nabla^3 U(\boldsymbol{x})$, and the uniform eighth moment bound of the numerical solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$. For this reason, Assumption (Sg) requires an eighth moment bound on the stochastic gradient deviation $b(\boldsymbol{x}, \theta) - \nabla U(\boldsymbol{x})$.

C.3 Estimate of long-time displacement

At the K-th step, the state displacement $\boldsymbol{Z}_K - \boldsymbol{Z}_0$ can be estimated as follows.

Lemma 21 *Under Assumptions* (P) *and* (Sg)*, let the step size* $h \leqslant \frac{1}{4}$*, and the initial state*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the SG-UBU solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ *satisfies*

$$\mathbb{E}|\boldsymbol{Z}_K - \boldsymbol{Z}_0|^2 \leqslant \frac{CdKh}{m},$$

where the constant C *depends only on* M_1, M_2*.*

Proof We write the one-step SG-UBU update $\boldsymbol{Z}_{k+1}$ as

$$\boldsymbol{X}_{k+1} - \boldsymbol{X}_k = \frac{1-e^{-2h}}{2}\boldsymbol{V}_k - \frac{h(1-e^{-h})}{2M_2}b(\boldsymbol{Y}_k, \theta_k) + \sqrt{h}\boldsymbol{M}_k, \tag{59a}$$

$$\boldsymbol{V}_{k+1} - \boldsymbol{V}_k = -(1-e^{-2h})\boldsymbol{V}_k - \frac{he^{-h}}{M_2}b(\boldsymbol{Y}_k, \theta_k) + \sqrt{h}\boldsymbol{N}_k, \tag{59b}$$

where the martingales $\boldsymbol{M}_k$ and $\boldsymbol{N}_k$ are defined as

$$\boldsymbol{M}_k = \frac{2}{\sqrt{M_2 h}}\int_0^h \frac{1-e^{-2(h-s)}}{2}\mathrm{d}\boldsymbol{B}_{kh+s}, \quad \boldsymbol{N}_k = \frac{2}{\sqrt{M_2 h}}\int_0^h e^{-2(h-s)}\mathrm{d}\boldsymbol{B}_{kh+s},$$

and their second moments satisfy

$$\mathbb{E}|\boldsymbol{M}_k|^2 \leqslant Cd, \quad \mathbb{E}|\boldsymbol{N}_k|^2 \leqslant Cd.$$

Applying Cauchy's inequality to (59) yields

$$\begin{aligned}\mathbb{E}\left|\boldsymbol{X}_K - \boldsymbol{X}_0 - \sqrt{h}\sum_{k=0}^{K-1}\boldsymbol{M}_k\right|^2 &= \mathbb{E}\left|\sum_{k=0}^{K-1}\left(\boldsymbol{X}_{k+1} - \boldsymbol{X}_k - \sqrt{h}\boldsymbol{M}_k\right)\right|^2 \\ &\leqslant K\sum_{k=0}^{K-1}\mathbb{E}\left|\frac{1-e^{-2h}}{2}\boldsymbol{V}_k - \frac{h(1-e^{-h})}{2M_2}b(\boldsymbol{Y}_k, \theta_k)\right|^2 \leqslant \frac{CdK^2h^2}{m},\end{aligned}$$

and similarly,

$$\begin{aligned}\mathbb{E}\left|\boldsymbol{V}_K - \boldsymbol{V}_0 - \sqrt{h}\sum_{k=0}^{K-1}\boldsymbol{N}_k\right|^2 &= \mathbb{E}\left|\sum_{k=0}^{K-1}\left(\boldsymbol{V}_{k+1} - \boldsymbol{V}_k - \sqrt{h}\boldsymbol{N}_k\right)\right|^2 \\ &\leqslant K\sum_{k=0}^{K-1}\mathbb{E}\left|-(1-e^{-2h})\boldsymbol{V}_k - \frac{he^{-h}}{M_2}b(\boldsymbol{Y}_k, \theta_k)\right|^2 \leqslant \frac{CdK^2h^2}{m}.\end{aligned}$$

Since the $\boldsymbol{M}_k$ are independent and centered, we obtain the estimate

$$\begin{aligned}\mathbb{E}|\boldsymbol{X}_K - \boldsymbol{X}_0|^2 &\leqslant 2h\mathbb{E}\left|\sum_{k=0}^{K-1}\boldsymbol{M}_k\right|^2 + \frac{CdK^2h^2}{m} \\ &= 2h\sum_{k=0}^{K-1}\mathbb{E}|\boldsymbol{M}_k|^2 + \frac{CdK^2h^2}{m} \leqslant \frac{Cd}{m}(Kh + K^2h^2).\end{aligned}$$

Similarly, for $\boldsymbol{V}_k - \boldsymbol{V}_0$ we have the same inequality

$$\mathbb{E}|\boldsymbol{Z}_K - \boldsymbol{Z}_0|^2 \leqslant \frac{Cd}{m}(Kh + K^2h^2). \tag{60}$$

On the other hand, from the uniform-in-time moments in Theorem 3, we have

$$\mathbb{E}|\boldsymbol{Z}_K - \boldsymbol{Z}_0|^2 \leqslant 2\mathbb{E}\big(|\boldsymbol{Z}_K|^2 + |\boldsymbol{Z}_0|^2\big) \leqslant \frac{Cd}{m}. \tag{61}$$

Integrating (60), (61), and the algebraic inequality $\min\{x + x^2, 1\} \leqslant 2x$ we obtain

$$\mathbb{E}|\boldsymbol{Z}_K - \boldsymbol{Z}_0|^2 \leqslant \frac{CdKh}{m},$$

which completes the proof of Lemma 21. ■

Appendix D. Proofs for MSE and numerical bias of SG-UBU

Recall that the time average error of SG-UBU is expressed as in (18):

$$\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f) = \frac{\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_K)}{Kh} + \frac{1}{K}\sum_{k=0}^{K-1}(R_k + S_k + T_k),$$

where the error terms R_k, S_k, and T_k are given by

$$R_k = \frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\bar{\boldsymbol{Z}}_{k+1})}{h}, \qquad S_k = \frac{\phi_h(\bar{\boldsymbol{Z}}_{k+1}) - \phi_h(\boldsymbol{Z}_k(h))}{h},$$
$$T_k = \frac{\phi_h(\boldsymbol{Z}_k(h)) - \phi_h(\boldsymbol{Z}_k)}{h} + f(\boldsymbol{X}_k) - \pi(f).$$

D.1 Mean square error of SG-UBU

Proof of Theorem 7 According to (18), we only need to estimate the second moments:

$$\mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} R_k\Big)^2\bigg], \quad \mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} S_k\Big)^2\bigg], \quad \mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} T_k\Big)^2\bigg].$$

First, the stochastic gradient error R_k can be written as

$$R_k = \frac{1}{h}\big(\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\bar{\boldsymbol{Z}}_{k+1})\big) = \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \int_0^1 \nabla\phi_h\big(\lambda\boldsymbol{Z}_{k+1} + (1-\lambda)\bar{\boldsymbol{Z}}_{k+1}\big)\mathrm{d}\lambda,$$

we decompose $R_k = R_{k,1} + R_{k,2}$, where $R_{k,1}$ is a martingale and $R_{k,2}$ is a high-order remainder:

$$R_{k,1} = \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \nabla\phi_h(\bar{\boldsymbol{Z}}_{k+1}),$$
$$R_{k,2} = \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \int_0^1 \Big(\nabla\phi_h\big(\lambda\boldsymbol{Z}_{k+1} + (1-\lambda)\bar{\boldsymbol{Z}}_{k+1}\big) - \nabla\phi_h(\bar{\boldsymbol{Z}}_{k+1})\Big)\mathrm{d}\lambda.$$

The bounds of $\nabla\phi_h(\boldsymbol{x},\boldsymbol{v})$ and $\nabla^2\phi_h(\boldsymbol{x},\boldsymbol{v})$ in Theorem 1 imply the inequalities

$$|R_{k,1}| \leqslant \frac{C}{mh}|\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}|, \qquad |R_{k,2}| \leqslant \frac{C}{m^2h}|\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}|^2.$$

Hence utilizing the stochastic gradient error estimate in Lemma 5, we obtain

$$\begin{aligned}\mathbb{E}[R_{k,1}^2] &\leqslant \frac{C}{m^2h^2}\mathbb{E}|\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}|^2 \leqslant \frac{C\sigma^2 d}{m^2},\\ \mathbb{E}[R_{k,2}^2] &\leqslant \frac{C}{m^4h^2}\mathbb{E}|\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}|^4 \leqslant \frac{C\sigma^4 d^2h^2}{m^4}.\end{aligned}$$

Since $R_{k,1}$ is a martingale difference with $\mathbb{E}[R_{k,1}|\mathcal{G}_k] = 0$, where

$$\mathcal{G}_k = \mathcal{F}_k \vee \sigma\big((\boldsymbol{B}_t)_{kh\leqslant t\leqslant (k+1)h}\big)$$

enlarges $\mathcal{F}_k$ by the Brownian increment over the k-th step, so that $\bar{\boldsymbol{Z}}_{k+1}$ is $\mathcal{G}_k$-measurable while θ_k stays independent of $\mathcal{G}_k$, and $R_{j,1}$ is $\mathcal{G}_k$-measurable for $j < k$,

$$\mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} R_{k,1}\Big)^2\bigg] = \sum_{k=0}^{K-1}\mathbb{E}[R_{k,1}^2] \leqslant \frac{C\sigma^2 dK}{m^2}.$$

On the other hand, by Cauchy's inequality

$$\mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} R_{k,2}\Big)^2\bigg] \leqslant K\sum_{k=0}^{K-1}\mathbb{E}[R_{k,2}^2] \leqslant \frac{C\sigma^4 d^2K^2h^2}{m^4}.$$

Combining the estimates of $\sum_{k=0}^{K-1} R_{k,1}$ and $\sum_{k=0}^{K-1} R_{k,2}$ yields

$$\frac{1}{K^2}\mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} R_k\Big)^2\bigg] \leqslant C\bigg(\frac{\sigma^2 d}{m^2K} + \frac{\sigma^4 d^2h^2}{m^4}\bigg). \tag{62}$$

The discretization error S_k analysis parallels the approach for R_k. Adopting the analytical framework in Lemma 6, we decompose the local error $\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)$ as

$$\bar{\boldsymbol{X}}_{k+1} - \boldsymbol{X}_k(h) = \delta\boldsymbol{X}_k, \quad \bar{\boldsymbol{V}}_{k+1} - \boldsymbol{V}_k(h) = \delta_m\boldsymbol{V}_k + \delta_h\boldsymbol{V}_k,$$

where $\delta_m\boldsymbol{V}_k$ is a martingale and $\delta_h\boldsymbol{V}_k$ is a higher-order term. Now S_k decomposes as

$$\begin{aligned}S_{k,1} &= \frac{1}{h}\Big(\delta\boldsymbol{X}_k^\top\nabla_{\boldsymbol{x}}\phi_h(\boldsymbol{Z}_k) + \delta_h\boldsymbol{V}_k^\top\nabla_{\boldsymbol{v}}\phi_h(\boldsymbol{Z}_k)\Big),\\ S_{k,2} &= \frac{1}{h}\delta_m\boldsymbol{V}_k^\top\nabla_{\boldsymbol{v}}\phi_h(\boldsymbol{Z}_k),\\ S_{k,3} &= \frac{1}{h}(\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h))^\top\int_0^1\Big(\nabla\phi_h\big(\lambda\bar{\boldsymbol{Z}}_{k+1} + (1-\lambda)\boldsymbol{Z}_k(h)\big) - \nabla\phi_h(\boldsymbol{Z}_k)\Big)\mathrm{d}\lambda.\end{aligned}$$

Theorem 1 yields the bounds of $S_{k,1}$ and $S_{k,2}$:

$$|S_{k,1}| \leqslant \frac{C}{mh}\big(|\delta\boldsymbol{X}_k| + |\delta_h\boldsymbol{V}_k|\big), \qquad |S_{k,2}| \leqslant \frac{C}{mh}|\delta_m\boldsymbol{V}_k|.$$

Hence leveraging the discretization error estimates in Lemma 6 produces

$$\mathbb{E}[S_{k,1}^2] \leqslant \frac{Cd^2h^4}{m^4}, \qquad \mathbb{E}[S_{k,2}^2] \leqslant \frac{Cdh^3}{m^2}. \tag{63}$$

The martingale property of $S_{k,2}$ gives

$$\mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} S_{k,2}\Big)^2\bigg] = \sum_{k=0}^{K-1} \mathbb{E}[S_{k,2}^2] \leqslant \frac{CdKh^3}{m^2}. \tag{64}$$

By Cauchy's inequality, $\sum_{k=0}^{K-1} S_{k,1}$ has the moment bound

$$\mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} S_{k,1}\Big)^2\bigg] \leqslant K\sum_{k=0}^{K-1} \mathbb{E}[S_{k,1}^2] \leqslant \frac{Cd^2K^2h^4}{m^4}. \tag{65}$$

Exploiting the expressions of $\bar{\boldsymbol{Z}}_{k+1}$ and $\boldsymbol{Z}_k(h)$ in Appendix C, it is easy to derive

$$\mathbb{E}|\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k|^4 \leqslant \frac{Cd^2h^2}{m^2}, \quad \mathbb{E}|\boldsymbol{Z}_k(h) - \boldsymbol{Z}_k|^4 \leqslant \frac{Cd^2h^2}{m^2}.$$

Applying Theorem 1 we obtain

$$|S_{k,3}| \leqslant \frac{C}{m^2h}\sqrt{|\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)|^2\big(|\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k|^2 + |\boldsymbol{Z}_k(h) - \boldsymbol{Z}_k|^2\big)},$$

and thus by the discretization error moment estimate in Lemma 6,

$$\begin{aligned}\mathbb{E}[S_{k,3}^2] &\leqslant \frac{C}{m^4h^2}\sqrt{\mathbb{E}|\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k(h)|^4\mathbb{E}\big(|\bar{\boldsymbol{Z}}_{k+1} - \boldsymbol{Z}_k|^4 + |\boldsymbol{Z}_k(h) - \boldsymbol{Z}_k|^4\big)} \\ &\leqslant \frac{C}{m^4h^2}\bigg(\frac{d^2h^6}{m^2} + dh^5\bigg)\frac{dh}{m} = C\bigg(\frac{d^2h^4}{m^5} + \frac{d^3h^5}{m^7}\bigg).\end{aligned} \tag{66}$$

Then by Cauchy's inequality:

$$\mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} S_{k,3}\Big)^2\bigg] \leqslant C\bigg(\frac{d^2K^2h^4}{m^5} + \frac{d^3K^2h^5}{m^7}\bigg). \tag{67}$$

Synthesizing the inequalities (64), (65), and (67) gives

$$\frac{1}{K^2}\mathbb{E}\bigg[\Big(\sum_{k=0}^{K-1} S_k\Big)^2\bigg] \leqslant C\bigg(\frac{dh^3}{m^2K} + \frac{d^2h^4}{m^5} + \frac{d^3h^5}{m^7}\bigg). \tag{68}$$

Finally, decompose the martingale T_k as $T_k = T_{k,1} + T_{k,2}$ with

$$T_{k,1} = \frac{1}{h}\big(\phi_h(\boldsymbol{Z}_k(h)) - \phi_h(\boldsymbol{Z}_k)\big), \qquad T_{k,2} = f(\boldsymbol{X}_k) - \pi(f).$$

Applying Theorem 1 produces the bound

$$|T_{k,1}| \leqslant \frac{C}{mh}|\boldsymbol{Z}_k(h) - \boldsymbol{Z}_k| \implies \mathbb{E}[T_{k,1}^2] \leqslant \frac{C}{m^2h^2}\mathbb{E}|\boldsymbol{Z}_k(h) - \boldsymbol{Z}_k|^2 \leqslant \frac{Cd}{m^3h}.$$

For $T_{k,2}$, take $\boldsymbol{X}_* \sim \pi$ independent of $\boldsymbol{X}_k$, then $\mathbb{E}|\boldsymbol{X}_*|^2 \leqslant \frac{Cd}{m}$ and

$$\mathbb{E}[T_{k,2}^2] \leqslant \mathbb{E}|f(\boldsymbol{X}_k) - f(\boldsymbol{X}_*)|^2 \leqslant C\mathbb{E}\big(|\boldsymbol{X}_k|^2 + |\boldsymbol{X}_*|^2\big) \leqslant \frac{Cd}{m}.$$

Combining these moment estimates, we derive

$$\mathbb{E}[T_k^2] \leqslant C\big(\mathbb{E}[T_{k,1}^2] + \mathbb{E}[T_{k,2}^2]\big) \leqslant C\bigg(\frac{d}{m^3 h} + \frac{d}{m}\bigg) \leqslant \frac{Cd}{m^3 h}.$$

Leveraging the martingale property of T_k, we establish the time-averaged moment bound

$$\frac{1}{K^2}\mathbb{E}\bigg[\bigg(\sum_{k=0}^{K-1} T_k\bigg)^2\bigg] \leqslant \frac{Cd}{m^3 K h}. \tag{69}$$

Combining the estimates (62), (68), and (69), we derive

$$\mathbb{E}\bigg[\bigg(\frac{1}{K}\sum_{k=0}^{K-1}(R_k + S_k + T_k)\bigg)^2\bigg] \leqslant C\bigg(\frac{d}{m^3 K h} + \frac{\sigma^4 d^2 h^2}{m^4} + \frac{d^2 h^4}{m^5} + \frac{d^3 h^5}{m^7}\bigg). \tag{70}$$

On the other hand, Lemma 21 provides the displacement moment estimate:

$$\mathbb{E}\bigg[\bigg(\frac{\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_K)}{Kh}\bigg)^2\bigg] \leqslant \frac{C}{m^2 K^2 h^2}\mathbb{E}|\boldsymbol{Z}_0 - \boldsymbol{Z}_K|^2 \leqslant \frac{Cd}{m^3 K h}. \tag{71}$$

Integrating (70), (71) with the time average error expression (18) yields

$$\mathbb{E}\bigg[\bigg(\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f)\bigg)^2\bigg] \leqslant C\bigg(\frac{d}{m^3 K h} + \frac{\sigma^4 d^2 h^2}{m^4} + \frac{d^2 h^4}{m^5} + \frac{d^3 h^5}{m^7}\bigg),$$

completing the proof of Theorem 7. ■

The MSE analysis for mini-batch SG-UBU requires a quantitative bound on the high-order moments of the stochastic gradient. The following lemma establishes the key estimate.

Lemma 22 *Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be vectors with bounded eighth moments:*

$$\frac{1}{N}\sum_{i=1}^{N} |x_i|^8 \leqslant 1.$$

Denote $\bar{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i \in \mathbb{R}^d$. Let θ be a random subset of $\{1, \ldots, N\}$ of size p. Then

$$\mathbb{E}_\theta\bigg[\bigg|\frac{1}{p}\sum_{i\in\theta} x_i - \bar{x}_N\bigg|^8\bigg] \leqslant \frac{C}{p^4},$$

where the constant C is absolute.

The proof mainly relies on Hoeffding's inequality (Hoeffding, 1963) and Rosenthal's inequality (Petrov, 2012; Chen and Sung, 2020), which are stated below.

Theorem 23 (Hoeffding) *Let $\Pi = \{x_1, \dots, x_N\}$ be a finite set. Let $(X_1, \dots, X_p)$ be a sample drawn without replacement from Π, and let $(Y_1, \dots, Y_p)$ be a sample drawn with replacement from Π. If f is a continuous convex function on $\mathbb{R}^d$,*

$$\mathbb{E}\bigg[f\Big(\sum_{j=1}^{p} X_j\Big)\bigg] \leqslant \mathbb{E}\bigg[f\Big(\sum_{j=1}^{p} Y_j\Big)\bigg].$$

Theorem 24 (Rosenthal) *Let $Z_1, \dots, Z_p$ be independent mean-zero random vectors in $\mathbb{R}^d$. For $k \geqslant 2$, there exists a constant C_k depending only on k such that*

$$\mathbb{E}\bigg[\Big|\sum_{j=1}^{p} Z_j\Big|^k\bigg] \leqslant C_k \max\bigg(\sum_{j=1}^{p} \mathbb{E}\big[|Z_j|^k\big], \Big(\sum_{j=1}^{p} \mathbb{E}\big[|Z_j|^2\big]\Big)^{\frac{k}{2}}\bigg).$$

Proof of Lemma 22 Since $|\bar{x}_N|^8 \leqslant \frac{1}{N}\sum_{i=1}^{N} |x_i|^8 \leqslant 1$, replacing x_i by $x_i - \bar{x}_N$ preserves the hypothesis up to the absolute factor 2^8, and the constant C below is absolute; we may therefore assume $\bar{x}_N = 0$. Let $X_1, \cdots, X_p$ denote independent samples from $\{x_1, \cdots, x_N\}$ (with replacement), then by Hoeffding's inequality we have

$$\mathbb{E}_\theta\bigg[\Big|\frac{1}{p}\sum_{i\in\theta} x_i\Big|^8\bigg] \leqslant \mathbb{E}\bigg[\Big|\frac{1}{p}\sum_{j=1}^{p} X_j\Big|^8\bigg]. \tag{72}$$

Since $X_1, \cdots, X_p$ are independent, applying Rosenthal's inequality with $k = 8$,

$$\begin{aligned}
\mathbb{E}\bigg[\Big|\sum_{j=1}^{p} X_j\Big|^8\bigg] &\leqslant C \max\bigg(\sum_{j=1}^{p} \mathbb{E}\big[|X_j|^8\big], \Big(\sum_{j=1}^{p} \mathbb{E}\big[|X_j|^2\big]\Big)^4\bigg) \\
&\leqslant C\Big(p\,\mathbb{E}\big[|X_1|^8\big] + p^4\Big(\mathbb{E}\big[|X_1|^2\big]\Big)^4\Big) \\
&\leqslant Cp^4\mathbb{E}\big[|X_1|^8\big] = \frac{Cp^4}{N}\sum_{i=1}^{N} |x_i|^8 \leqslant Cp^4.
\end{aligned}$$

Hence we derive the upper bound

$$\mathbb{E}\bigg[\Big|\frac{1}{p}\sum_{j=1}^{p} X_j\Big|^8\bigg] \leqslant \frac{C}{p^4}. \tag{73}$$

Combining the estimates (72) and (73) yields the result of Lemma 22. ■

By Lemma 22, Assumption (SgN) implies the following eighth moment control of the mini-batch gradient error: for any $\boldsymbol{x} \in \mathbb{R}^d$ and random subset $\theta \subset \{1, \dots, N\}$ with size p,

$$\bigg(\mathbb{E}_\theta\bigg[\Big|\frac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}) - \nabla U(\boldsymbol{x})\Big|^8\bigg]\bigg)^{\frac{1}{4}} \leqslant \frac{C\sigma^2 d}{p}.$$

Hence the stochastic gradient in mini-batch SG-UBU satisfies all items of Assumption (Sg) with σ^2 replaced by $C\sigma^2/p$. Therefore, Theorem 8 follows directly from Theorem 7.

D.2 Numerical bias of SG-UBU

Proof of Theorem 9 To streamline the notation, expectations are taken with respect to the numerical invariant distribution π_h and the stochastic gradient index θ, which we omit in the proof. From Lemma 4, $\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0)$ drawn from π_h satisfies the moment bound:

$$\mathbb{E}\big[|\boldsymbol{Z}_0|^8\big] \leqslant \frac{Cd^4}{m^4}.$$

According to (24), the numerical bias is the sum of the expected stochastic gradient error, $\mathbb{E}[R_0]$, and the expected discretization error, $\mathbb{E}[S_0]$, where

$$R_0 = \frac{\phi_h(\boldsymbol{Z}_1) - \phi_h(\bar{\boldsymbol{Z}}_1)}{h}, \qquad S_0 = \frac{\phi_h(\bar{\boldsymbol{Z}}_1) - \phi_h(\boldsymbol{Z}_0(h))}{h}.$$

Since π_h satisfies the same uniform-in-time moment bound by Lemma 4, the local error estimates of Lemma 5 and Lemma 6, together with (63) and (66), remain valid here: all of them depend on the state only through its eighth moment.

We begin with the stochastic error term $\mathbb{E}[R_0]$. A Taylor expansion of $\phi_h(\boldsymbol{Z}_1)$ around $\bar{\boldsymbol{Z}}_1$ using the mean value theorem gives

$$\mathbb{E}[R_0] = \frac{1}{h}\mathbb{E}\bigg[(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)^\top \int_0^1 \nabla\phi_h\big(\lambda\boldsymbol{Z}_1 + (1-\lambda)\bar{\boldsymbol{Z}}_1\big)\mathrm{d}\lambda\bigg].$$

The key insight is that the first-order term of this expansion vanishes in expectation due to the martingale property of the stochastic gradient error established in (21):

$$\mathbb{E}\big[(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)^\top \nabla\phi_h(\bar{\boldsymbol{Z}}_1)\big] = 0.$$

This simplification leaves only the second-order terms:

$$\begin{aligned}
\mathbb{E}[R_0] &= \frac{1}{h}\mathbb{E}\bigg[(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)^\top \int_0^1 \Big(\nabla\phi_h\big(\lambda\boldsymbol{Z}_1 + (1-\lambda)\bar{\boldsymbol{Z}}_1\big) - \nabla\phi_h(\bar{\boldsymbol{Z}}_1)\Big)\mathrm{d}\lambda\bigg] \\
&= \frac{1}{h}\mathbb{E}\bigg[(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)^\top \int_0^1 \bigg(\lambda\int_0^1 \nabla^2\phi_h\big(\bar{\boldsymbol{Z}}_1 + \lambda\varphi(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)\big)\mathrm{d}\varphi\bigg)\mathrm{d}\lambda\,(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)\bigg].
\end{aligned}$$

Recall that the components of the local error $\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1$ converge at different rates. Lemma 5 provides the moment bounds:

$$\sqrt{\mathbb{E}\big|\boldsymbol{X}_1 - \bar{\boldsymbol{X}}_1\big|^4} \leqslant C\sigma^2 dh^4, \qquad \sqrt{\mathbb{E}\big|\boldsymbol{V}_1 - \bar{\boldsymbol{V}}_1\big|^4} \leqslant C\sigma^2 dh^2.$$

Since the position error $\boldsymbol{X}_1 - \bar{\boldsymbol{X}}_1$ is of a higher order in h than the velocity error, its contribution to the bias is smaller. Using the Hessian bound $\|\nabla^2\phi_h\| \leqslant C/m^2$ from Theorem 1, we can bound these higher-order terms:

$$\begin{aligned}
&\frac{1}{h}\mathbb{E}\bigg[(\boldsymbol{X}_1 - \bar{\boldsymbol{X}}_1)^\top \int_0^1 \bigg(\lambda\int_0^1 \nabla^2_{\boldsymbol{xv}}\phi_h\big(\bar{\boldsymbol{Z}}_1 + \lambda\varphi(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)\big)\mathrm{d}\varphi\bigg)\mathrm{d}\lambda\,(\boldsymbol{V}_1 - \bar{\boldsymbol{V}}_1)\bigg] \leqslant \frac{Cdh^2}{m^2}, \\
&\frac{1}{h}\mathbb{E}\bigg[(\boldsymbol{X}_1 - \bar{\boldsymbol{X}}_1)^\top \int_0^1 \bigg(\lambda\int_0^1 \nabla^2_{\boldsymbol{xx}}\phi_h\big(\bar{\boldsymbol{Z}}_1 + \lambda\varphi(\boldsymbol{Z}_1 - \bar{\boldsymbol{Z}}_1)\big)\mathrm{d}\varphi\bigg)\mathrm{d}\lambda\,(\boldsymbol{X}_1 - \bar{\boldsymbol{X}}_1)\bigg] \leqslant \frac{Cdh^3}{m^2}.
\end{aligned}$$

The dominant contribution to the bias therefore comes from the velocity error component. Using the definition of $\boldsymbol{M}_0$ from the theorem statement, we arrive at the estimate:

$$\left|\mathbb{E}[R_0] - \frac{1}{2h}\mathbb{E}\Big[(\boldsymbol{V}_1 - \bar{\boldsymbol{V}}_1)^\top \boldsymbol{M}_0(\boldsymbol{V}_1 - \bar{\boldsymbol{V}}_1)\Big]\right| \leqslant \frac{Cdh^2}{m^2}. \tag{74}$$

Finally, we substitute the explicit expression for the velocity error from (20):

$$\boldsymbol{V}_1 - \bar{\boldsymbol{V}}_1 = -\frac{he^{-h}}{M_2}\big(b(\boldsymbol{Y}_0, \theta) - \nabla U(\boldsymbol{Y}_0)\big).$$

This substitution into (74), along with the linear approximation $e^{-2h} = 1 + \mathcal{O}(h)$, yields the leading-order term for the stochastic error's contribution to the bias:

$$\left|\mathbb{E}[R_0] - \frac{h}{2M_2^2}\mathbb{E}\Big[\big(b(\boldsymbol{Y}_0, \theta) - \nabla U(\boldsymbol{Y}_0)\big)^\top \boldsymbol{M}_0\big(b(\boldsymbol{Y}_0, \theta) - \nabla U(\boldsymbol{Y}_0)\big)\Big]\right| \leqslant \frac{Cdh^2}{m^2}.$$

For the discretization error component, $\mathbb{E}[S_0]$, we directly follow the decomposition in Theorem 7 and write $S_0 = S_{0,1} + S_{0,2} + S_{0,3}$, where

$$\begin{aligned}
S_{0,1} &= \frac{1}{h}\Big(\delta \boldsymbol{X}_0^\top \nabla_{\boldsymbol{x}}\phi_h(\boldsymbol{Z}_0) + \delta_h \boldsymbol{V}_0^\top \nabla_{\boldsymbol{v}}\phi_h(\boldsymbol{Z}_0)\Big),\\
S_{0,2} &= \frac{1}{h}\delta_m \boldsymbol{V}_0^\top \nabla_{\boldsymbol{v}}\phi_h(\boldsymbol{Z}_0),\\
S_{0,3} &= \frac{1}{h}(\bar{\boldsymbol{Z}}_1 - \boldsymbol{Z}_0(h))^\top \int_0^1 \Big(\nabla\phi_h\big(\lambda\bar{\boldsymbol{Z}}_1 + (1-\lambda)\boldsymbol{Z}_0(h)\big) - \nabla\phi_h(\boldsymbol{Z}_0)\Big)\mathrm{d}\lambda.
\end{aligned}$$

The martingale property of $\delta_m \boldsymbol{V}_0$ stated in Lemma 6 implies $\mathbb{E}[S_{0,2}] = 0$, and thus the expectation simplifies to $\mathbb{E}[S_0] = \mathbb{E}[S_{0,1}] + \mathbb{E}[S_{0,3}]$. Utilizing the moment estimates of $S_{0,1}$ and $S_{0,3}$ in (63) and (66), and the step size constraint $h \leqslant \frac{m}{d}$, we obtain

$$\mathbb{E}|S_{0,1}| \leqslant \frac{Cdh^2}{m^2}, \qquad \mathbb{E}|S_{0,3}| \leqslant \frac{Cdh^2}{m^3}.$$

Combining these estimates provides a bound on the expected total local error:

$$\big|\mathbb{E}[S_0]\big| \leqslant \mathbb{E}|S_{0,1}| + \mathbb{E}|S_{0,3}| \leqslant \frac{Cdh^2}{m^3}. \tag{75}$$

Synthesizing the estimates for the stochastic error in (74) and the discretization error in (75), we arrive at a complete characterization of the numerical bias:

$$\left|\pi_h(f) - \pi(f) - \frac{h}{2M_2^2}\mathbb{E}\Big[\big(b(\boldsymbol{Y}_0, \theta) - \nabla U(\boldsymbol{Y}_0)\big)^\top \boldsymbol{M}_0\big(b(\boldsymbol{Y}_0, \theta) - \nabla U(\boldsymbol{Y}_0)\big)\Big]\right| \leqslant \frac{Cdh^2}{m^3}.$$

This expression isolates the leading-order term of the bias, which is proportional to the variance of the stochastic gradient. By applying the established bounds $\|\boldsymbol{M}_0\| \leqslant C/m^2$ and $\mathbb{E}\big[|b(\boldsymbol{Y}_0, \theta) - \nabla U(\boldsymbol{Y}_0)|^2\big] \leqslant C\sigma^2 d$, we can bound this leading term, confirming that the overall numerical bias exhibits first-order convergence:

$$|\pi_h(f) - \pi(f)| \leqslant C\left(\frac{\sigma^2 dh}{m^2} + \frac{dh^2}{m^3}\right).$$

This completes the proof of Theorem 9. ■

Proof of Lemma 10 We begin by recalling the continuous Poisson equation in (13):

$$\boldsymbol{v}^\top \nabla_{\boldsymbol{x}} \phi_0(\boldsymbol{x},\boldsymbol{v}) - \left(\frac{1}{M_2}\nabla U(\boldsymbol{x}) + 2\boldsymbol{v}\right)^\top \nabla_{\boldsymbol{v}} \phi_0(\boldsymbol{x},\boldsymbol{v}) + \frac{2}{M_2}\Delta_{\boldsymbol{v}} \phi_0(\boldsymbol{x},\boldsymbol{v}) = \pi(f) - f(\boldsymbol{x}). \quad (76)$$

Theorem 19 provides an exponential decay estimate for the Kolmogorov solution $u(\boldsymbol{x},\boldsymbol{v},t)$:

$$\max\left\{|\nabla_{\boldsymbol{x}} u(\boldsymbol{x},\boldsymbol{v},t)|, |\nabla_{\boldsymbol{v}} u(\boldsymbol{x},\boldsymbol{v},t)|\right\} \leqslant C L_1 e^{-\frac{mt}{2M_2}}.$$

By integrating this inequality with respect to t over $(0,\infty)$ and using the integral representation $\phi_0(\boldsymbol{x},\boldsymbol{v}) = \int_0^\infty u(\boldsymbol{x},\boldsymbol{v},t)\mathrm{d}t$, we obtain uniform bounds on the derivatives of $\phi_0(\boldsymbol{x},\boldsymbol{v})$:

$$\max\left\{|\nabla_{\boldsymbol{x}} \phi_0(\boldsymbol{x},\boldsymbol{v})|, |\nabla_{\boldsymbol{v}} \phi_0(\boldsymbol{x},\boldsymbol{v})|\right\} \leqslant \frac{C L_1}{m}.$$

With these derivative bounds established, we can rearrange the Poisson equation (76) to isolate the Laplacian term. Since the potential gradient $\nabla U(\boldsymbol{x})$ and the test function $f(\boldsymbol{x})$ both exhibit at most linear growth in $|\boldsymbol{x}|$, all terms in the equation other than the Laplacian grow at most linearly in $|\boldsymbol{x}|$ and $|\boldsymbol{v}|$. This allows us to establish a growth bound on the Laplacian of the velocity component of ϕ_0:

$$|\Delta_{\boldsymbol{v}} \phi_0(\boldsymbol{x},\boldsymbol{v})| \leqslant \frac{C}{m}\sqrt{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d}, \qquad \forall \boldsymbol{x},\boldsymbol{v} \in \mathbb{R}^d. \quad (77)$$

To complete the proof, we need to bound the second moment of the Hessian of the Poisson solution, averaged over the equilibrium velocity distribution. Thus we define the quantity $\mathcal{Q}(\boldsymbol{x})$ as this velocity-averaged integral:

$$\mathcal{Q}(\boldsymbol{x}) = \left(\frac{M_2}{2\pi}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} \left\|\nabla^2_{\boldsymbol{v}\boldsymbol{v}} \phi_0(\boldsymbol{x},\boldsymbol{v})\right\|^2 e^{-\frac{M_2}{2}|\boldsymbol{v}|^2} \mathrm{d}\boldsymbol{v}.$$

The goal is to show that the expectation of $\mathcal{Q}(\boldsymbol{x})$ with respect to the target distribution $\pi(\boldsymbol{x})$ is appropriately bounded:

$$\mathbb{E}^\pi\left[\mathcal{Q}(\boldsymbol{x})\right] \leqslant \frac{C_d}{m^3}. \quad (78)$$

Our strategy is to leverage local elliptic regularity estimates. We partition the velocity space $\mathbb{R}^d$ into unit hypercubes centered at integer vectors $\boldsymbol{j} = (j_1, \cdots, j_d) \in \mathbb{Z}^d$:

$$B_{\boldsymbol{j}} = \left(j_1 - \frac{1}{2}, j_1 + \frac{1}{2}\right) \times \cdots \times \left(j_d - \frac{1}{2}, j_d + \frac{1}{2}\right),$$
$$\tilde{B}_{\boldsymbol{j}} = (j_1 - 1, j_1 + 1) \times \cdots \times (j_d - 1, j_d + 1).$$

Standard local L^2-estimates for elliptic equations allow us to bound the integral of the Hessian over a block $B_{\boldsymbol{j}}$ by the integral of the Laplacian and the function itself over the

larger block $\tilde{B}_{\boldsymbol{j}}$:

$$\begin{aligned}\int_{B_{\boldsymbol{j}}} \left\|\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x},\boldsymbol{v})\right\|^2 \mathrm{d}\boldsymbol{v} &\leqslant C_d \int_{\tilde{B}_{\boldsymbol{j}}} \Big(|\Delta_{\boldsymbol{v}}\phi_0(\boldsymbol{x},\boldsymbol{v})|^2 + |\phi_0(\boldsymbol{x},\boldsymbol{v})|^2\Big)\mathrm{d}\boldsymbol{v} \\ &\leqslant \frac{C_d}{m^2}\int_{\tilde{B}_{\boldsymbol{j}}} (|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d)\mathrm{d}\boldsymbol{v},\end{aligned}$$

where we have used the growth bound (77) for the Laplacian and a similar bound for the solution ϕ_0.

For any $\boldsymbol{v} \in \tilde{B}_{\boldsymbol{j}}$, its squared norm is related to the squared norm of the center point $\boldsymbol{j}$. This relationship allows us to bound the Gaussian weight within each block. For $\boldsymbol{v} \in B_{\boldsymbol{j}}$, we can bound the exponential term from below in terms of $|\boldsymbol{j}|^2$, which simplifies the integral. Combining these estimates, we can bound the contribution to $\mathcal{Q}(\boldsymbol{x})$ from each block $B_{\boldsymbol{j}}$:

$$\begin{aligned}&\left(\frac{M_2}{2\pi}\right)^{\frac{d}{2}} \int_{B_{\boldsymbol{j}}} \left\|\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x},\boldsymbol{v})\right\|^2 e^{-\frac{M_2}{2}|\boldsymbol{v}|^2}\mathrm{d}\boldsymbol{v} \\ &\leqslant C_d \exp\Big(-\frac{M_2}{4}\sum_{i=1}^d j_i^2\Big)\int_{B_{\boldsymbol{j}}} \left\|\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x},\boldsymbol{v})\right\|^2 \mathrm{d}\boldsymbol{v} \\ &\leqslant \frac{C_d}{m^2}\exp\Big(-\frac{M_2}{4}\sum_{i=1}^d j_i^2\Big)\int_{\tilde{B}_{\boldsymbol{j}}} (|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d)\mathrm{d}\boldsymbol{v} \\ &\leqslant \frac{C_d}{m^2}\exp\Big(-\frac{M_2}{4}\sum_{i=1}^d j_i^2\Big)\Big(|\boldsymbol{x}|^2 + \sum_{i=1}^d j_i^2 + C_d\Big) \\ &\leqslant \frac{C_d}{m^2}\exp\Big(-\frac{M_2}{8}\sum_{i=1}^d j_i^2\Big)(|\boldsymbol{x}|^2 + 1).\end{aligned}$$

Summing over all blocks $\boldsymbol{j} \in \mathbb{Z}^d$ yields a bound on $\mathcal{Q}(\boldsymbol{x})$. The sum over the exponential terms is a convergent geometric series, resulting in an upper bound for $\mathcal{Q}(\boldsymbol{x})$ that exhibits at most quadratic growth in $|\boldsymbol{x}|$:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{x}) &= \sum_{\boldsymbol{j}\in\mathbb{Z}^d}\left(\frac{M_2}{2\pi}\right)^{\frac{d}{2}} \int_{B_{\boldsymbol{j}}} \left\|\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x},\boldsymbol{v})\right\|^2 e^{-\frac{M_2}{2}|\boldsymbol{v}|^2}\mathrm{d}\boldsymbol{v} \\ &\leqslant \frac{C_d}{m^2}(|\boldsymbol{x}|^2+1)\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\exp\Big(-\frac{M_2}{8}\sum_{i=1}^d j_i^2\Big) \leqslant \frac{C_d}{m^2}(|\boldsymbol{x}|^2+1).\end{aligned}$$

Finally, taking the expectation of this inequality with respect to $\pi(\boldsymbol{x})$ yields the desired result. Since the potential is strongly convex, $\pi(\boldsymbol{x})$ has finite second moments, giving:

$$\mathbb{E}^\pi\big[\mathcal{Q}(\boldsymbol{x})\big] \leqslant \frac{C_d}{m^2}\mathbb{E}^\pi[|\boldsymbol{x}|^2+1] \leqslant \frac{C_d}{m^3}.$$

This completes the proof of Lemma 10. ■

The idea of using hypercubes to derive the L^2-estimate comes from Xu'an Dou.

Appendix E. Proofs for MSE and numerical bias of SVRG-UBU

The analysis of SVRG-UBU follows the same structure as that for mini-batch SG-UBU, establishing stability through uniform-in-time moment bounds and consistency via local error estimates. The discrete Poisson equation is then applied to derive the MSE and numerical bias. A critical distinction, however, is that the SVRG-UBU solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ is non-Markovian. To address this, we define the invariant distribution $\pi_h(\boldsymbol{x}, \boldsymbol{v})$ with respect to the embedded Markovian subsequence $(\boldsymbol{X}_{kN/p}, \boldsymbol{V}_{kN/p})_{k\geqslant 0}$. This subsequence is constructed from the anchor positions, which are updated at the end of each epoch.

Throughout this section, we let $q = N/p$ denote the number of inner iterations in each epoch. To formulate the SVRG-UBU integrator compactly, we define a solution flow for the variance-reduced gradient update, analogous to $\Phi_t^{\mathfrak{B}}$ and $\Phi_t^{\mathfrak{B}}(\theta)$. For a given anchor position $\boldsymbol{x}_* \in \mathbb{R}^d$, this flow is defined as

$$\Phi_t^{\mathfrak{B}}(\boldsymbol{x}_*, \theta) : \begin{cases} \boldsymbol{x}_t = \boldsymbol{x}_0, \\ \boldsymbol{v}_t = \boldsymbol{v}_0 - \dfrac{t}{M_2}\Big(\dfrac{1}{p}\displaystyle\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}_0) - \dfrac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}_*) + \nabla U(\boldsymbol{x}_*)\Big). \end{cases}$$

Then SVRG-UBU from Algorithm 3 can be expressed as a single update rule. Since the anchor position is recomputed every q steps, the integrator takes the form:

$$\boldsymbol{Z}_{k+1} = \big(\Phi_{h/2}^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\boldsymbol{Y}_{q[k/q]}, \theta_k) \circ \Phi_{h/2}^{\mathfrak{U}}\big)\boldsymbol{Z}_k, \qquad k = 0, 1, \cdots, \tag{79}$$

where $\boldsymbol{Y}_k$ is the position component of $\Phi_{h/2}^{\mathfrak{U}}\boldsymbol{Z}_k$. The term $\boldsymbol{Y}_{q[k/q]}$ represents the anchor position established at the beginning of the current epoch, which remains fixed for q iterations. The random subsets $(\theta_k)_{k=0}^{\infty}$ of size p are drawn independently at each step.

E.1 Uniform-in-time moment bound of SVRG-UBU

Our first step is to establish a uniform-in-time moment bound for the SVRG-UBU solution, which is a prerequisite for proving the existence of an invariant distribution for its embedded Markovian subsequence. Following the analytical approach used for SG-UBU, we begin by analyzing the corresponding BU-type integrator, SVRG-BU. Its update rule is given by

$$\boldsymbol{Z}_{k+1} = \big(\Phi_h^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\boldsymbol{X}_{q[k/q]}, \theta_k)\big)\boldsymbol{Z}_k, \qquad k = 0, 1, \cdots. \tag{80}$$

The analysis of SVRG-BU closely parallels that of the standard SG-BU integrator. By directly adapting the proof of Lemma 2, we establish an analogous Lyapunov condition for SVRG-BU, as stated below.

Lemma 25 *Under Assumptions* (P) *and* (SgN^+)*, let the step size* $h \leqslant \frac{1}{4}$*; for any initial state* $\boldsymbol{z}_0 \in \mathbb{R}^{2d}$ *and the anchor position* $\boldsymbol{x}_* \in \mathbb{R}^d$*,*

$$\mathbb{E}\Big[\mathcal{V}\big((\Phi_h^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\boldsymbol{x}_*, \theta))\boldsymbol{z}_0\big)\Big] \leqslant \left(1 - \frac{mh}{3M_2}\right)\mathcal{V}(\boldsymbol{z}_0) + \frac{Cd^4h}{m^3},$$

where θ *is a random subset of* $\{1, \cdots, N\}$ *of size* p*, and* C *depends only on* M_1, M_2*.*

Proof The proof proceeds by verifying that the SVRG estimator $b(\boldsymbol{x}, \boldsymbol{x}_*, \theta)$ is unbiased and that its error carries the eighth moment required by Lemma 2. The almost-sure growth bound of Assumption (Sg) is not used here.

First, the unbiasedness property, $\mathbb{E}_\theta[b(\boldsymbol{x}, \boldsymbol{x}_*, \theta)] = \nabla U(\boldsymbol{x})$, is readily verified from the definition of the estimator. To check the eighth moment condition in Assumption (Sg), we apply the triangle inequality:

$$|b(\boldsymbol{x}, \boldsymbol{x}_*, \theta) - \nabla U(\boldsymbol{x})| \leqslant \left|\frac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}) - \nabla U(\boldsymbol{x})\right| + \left|\frac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}_*) - \nabla U(\boldsymbol{x}_*)\right|.$$

Computing the eighth moment of both sides yields

$$\begin{aligned}
&\mathbb{E}_\theta\big[|b(\boldsymbol{x}, \boldsymbol{x}_*, \theta) - \nabla U(\boldsymbol{x})|^8\big] \\
&\leqslant C\,\mathbb{E}_\theta\left[\left|\frac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}) - \nabla U(\boldsymbol{x})\right|^8 + \left|\frac{1}{p}\sum_{i\in\theta} \nabla U_i(\boldsymbol{x}_*) - \nabla U(\boldsymbol{x}_*)\right|^8\right] \\
&\leqslant \frac{C}{N}\sum_{i=1}^N |\nabla U_i(\boldsymbol{x}) - \nabla U(\boldsymbol{x})|^8 + \frac{C}{N}\sum_{i=1}^N |\nabla U_i(\boldsymbol{x}_*) - \nabla U(\boldsymbol{x}_*)|^8 \leqslant C(\sigma^2 d)^4,
\end{aligned}$$

where we have applied Jensen's inequality. This exactly produces the first two conditions of Assumption (Sg)

$$\Big(\mathbb{E}_\theta\big[|b(\boldsymbol{x}, \boldsymbol{x}_*, \theta) - \nabla U(\boldsymbol{x})|^8\big]\Big)^{\frac{1}{4}} \leqslant C\sigma^2 d$$

for an arbitrary anchor position $\boldsymbol{x}_*$. Therefore, Lemma 2 establishes the Lyapunov condition of the SVRG-BU integrator, regardless of the choice of anchor positions. ■

The Lyapunov condition established for SVRG-BU in Lemma 25 provides the foundation for bounding the moments of the SVRG-UBU solution. Since the SVRG-UBU integrator differs from its BU-type counterpart only by the stable $\Phi^{\mathfrak{U}}_{h/2}$ mapping, their uniform-in-time moment bounds are equivalent. This leads directly to the following theorem.

Theorem 26 *Under Assumptions* (P) *and* (SgN^+)*, let the step size* $h \leqslant \frac{1}{4}$*, and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the SVRG-UBU solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^\infty$ *satisfies*

$$\sup_{k\geqslant 0} \mathbb{E}\big[|\boldsymbol{Z}_k|^8\big] \leqslant \frac{Cd^4}{m^4},$$

where the constant C *depends only on* M_1, M_2*.*

By applying the same techniques from the proof of Lemma 4, we can prove the existence of and establish a moment bound for the invariant distribution of SVRG-UBU's embedded Markovian subsequence. This subsequence, $(\boldsymbol{X}_{kq}, \boldsymbol{V}_{kq})_{k=0}^\infty$, is formed by the states at the end of each epoch (recall that the anchor position is updated every $q = N/p$ iterations). The result is described in the following lemma.

Lemma 27 *Under Assumptions* (P) *and* (SgN^+)*, let the step size* $h \leqslant \frac{1}{4}$*; the SVRG-UBU subsequence* $(\boldsymbol{X}_{kq}, \boldsymbol{V}_{kq})_{k=0}^{\infty}$ *has an invariant distribution* $\pi_h(\boldsymbol{x}, \boldsymbol{v})$ *in* $\mathbb{R}^{2d}$*, which satisfies*

$$\int_{\mathbb{R}^{2d}} |\boldsymbol{z}|^8 \pi_h(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \leqslant \frac{Cd^4}{m^4},$$

where the constant C *depends only on* M_1, M_2*.*

Proof To formalize the analysis, we first define a semigroup for the variance-reduced gradient update that explicitly depends on a fixed anchor position $\boldsymbol{x}_* \in \mathbb{R}^d$. Analogous to the previous definitions, we have

$$P_t^{\mathfrak{B}}(\boldsymbol{x}_*)\mu := \mathbb{E}_\theta\big[\text{distribution of } \Phi_t^{\mathfrak{B}}(\boldsymbol{x}_*, \theta)\boldsymbol{z}_0 \text{ with } \boldsymbol{z}_0 \sim \mu\big].$$

The Lyapunov condition established in Lemma 25, combined with the Krylov–Bogolyubov theorem, guarantees that the SVRG-BU subsequence $(\boldsymbol{X}_{kq}, \boldsymbol{V}_{kq})_{k=0}^{\infty}$ admits an invariant distribution $\hat{\pi}_h(\boldsymbol{x}, \boldsymbol{v})$ in $\mathbb{R}^{2d}$. This distribution satisfies the moment bound:

$$\int_{\mathbb{R}^{2d}} |\boldsymbol{z}|^8 \hat{\pi}_h(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \leqslant \frac{Cd^4}{m^4}. \tag{81}$$

By definition of the SVRG-BU integrator, an epoch consists of $q = N/p$ inner iterations where the anchor point is fixed. The invariance of $\hat{\pi}_h$ across these epochs is expressed by the relation

$$\int_{\mathbb{R}^{2d}} \Big(\big(P_h^{\mathfrak{U}} P_h^{\mathfrak{B}}(\boldsymbol{x})\big)^q \delta_{\boldsymbol{x},\boldsymbol{v}} \Big) \hat{\pi}_h(\boldsymbol{x}, \boldsymbol{v}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{v} = \hat{\pi}_h. \tag{82}$$

Here, the integral signifies an expectation over the invariant distribution $\hat{\pi}_h$. For each state $(\boldsymbol{x}, \boldsymbol{v})$ drawn from $\hat{\pi}_h$, its position component $\boldsymbol{x}$ becomes the anchor for an entire epoch of q steps. The use of the Dirac measure $\delta_{\boldsymbol{x},\boldsymbol{v}}$ indicates that this epoch starts from the state $(\boldsymbol{x}, \boldsymbol{v})$ itself, ensuring the anchor position remains fixed for all subsequent $\mathfrak{B}$ steps.

Next, we construct the invariant distribution for the SVRG-UBU subsequence, $\pi_h(\boldsymbol{x}, \boldsymbol{v})$, using the invariant distribution of the SVRG-BU subsequence, $\hat{\pi}_h$. Let the SVRG-BU epoch start from $(\boldsymbol{x}, \boldsymbol{v}) \sim \hat{\pi}_h$, set $\boldsymbol{Z}_1 = \big(\Phi_{h/2}^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\boldsymbol{x}, \theta_0)\big)(\boldsymbol{x}, \boldsymbol{v})$, and let $(\boldsymbol{Z}_k)_{k=1}^q$ be the SVRG-UBU solution continued from $\boldsymbol{Z}_1$ with the same anchor $\boldsymbol{x}$ and the same Brownian motion, so that $\Phi_{h/2}^{\mathfrak{U}} \boldsymbol{Z}_k$ is the SVRG-BU state after k steps. We define π_h to be the law of $\boldsymbol{Z}_q$:

$$\pi_h = \int_{\mathbb{R}^{2d}} \Big(\big(P_{h/2}^{\mathfrak{U}} P_h^{\mathfrak{B}}(\boldsymbol{x}) P_{h/2}^{\mathfrak{U}}\big)^{q-1} \big(P_{h/2}^{\mathfrak{U}} P_h^{\mathfrak{B}}(\boldsymbol{x})\big) \delta_{\boldsymbol{x},\boldsymbol{v}} \Big) \hat{\pi}_h(\boldsymbol{x}, \boldsymbol{v}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{v}. \tag{83}$$

The rightmost factor $P_{h/2}^{\mathfrak{U}} P_h^{\mathfrak{B}}(\boldsymbol{x})$ carries the SVRG-BU state $(\boldsymbol{x}, \boldsymbol{v})$ to $\boldsymbol{Z}_1$, and the remaining $q-1$ factors are full SVRG-UBU steps. Note that the anchor $\boldsymbol{x}$ is the position component of the state drawn from $\hat{\pi}_h$, and stays fixed throughout the epoch. Then π_h inherits the same moment bound given in (81).

Applying one further half-step of the $\mathfrak{U}$ flow to (83) recovers the SVRG-BU state after q steps, whose law is $\hat{\pi}_h$ by (82):

$$P_{h/2}^{\mathfrak{U}} \pi_h = \int_{\mathbb{R}^{2d}} \Big(\big(P_h^{\mathfrak{U}} P_h^{\mathfrak{B}}(\boldsymbol{x})\big)^q \delta_{\boldsymbol{x},\boldsymbol{v}} \Big) \hat{\pi}_h(\boldsymbol{x}, \boldsymbol{v}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{v} = \hat{\pi}_h,$$

where the semigroup property $P^{\mathfrak{U}}_{h/2}P^{\mathfrak{U}}_{h/2} = P^{\mathfrak{U}}_h$ has been used to merge the half-steps. Therefore, substituting $\hat{\pi}_h = P^{\mathfrak{U}}_{h/2}\pi_h$ into (83) gives

$$\pi_h = \int_{\mathbb{R}^{2d}} \Big(\big(P^{\mathfrak{U}}_{h/2}P^{\mathfrak{B}}_h(\boldsymbol{x})P^{\mathfrak{U}}_{h/2}\big)^{q-1} \big(P^{\mathfrak{U}}_{h/2}P^{\mathfrak{B}}_h(\boldsymbol{x})\big)\delta_{\boldsymbol{x},\boldsymbol{v}}\Big)(P^{\mathfrak{U}}_{h/2}\pi_h)(\boldsymbol{x},\boldsymbol{v})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{v}$$

showing π_h is invariant for the SVRG-UBU subsequence $(\boldsymbol{X}_{kq}, \boldsymbol{V}_{kq})_{k=0}^{\infty}$. ■

E.2 Local error analysis of SVRG-UBU

We now turn to the local error analysis for the SVRG-UBU integrator. The discretization error is identical to that for SG-UBU, as this term is independent of the gradient estimator. Our focus therefore shifts to bounding the stochastic gradient error, which is where the effect of the variance reduction mechanism becomes apparent. The following lemma, analogous to Lemma 5, establishes the corresponding fourth moment bounds.

Lemma 28 *Under Assumptions* (P) *and* (SgN^+)*, let the step size* $h \leqslant \frac{1}{4}$*, and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the stochastic gradient error of SVRG-UBU satisfies

$$\sup_{k\geqslant 0}\sqrt{\mathbb{E}\big|\boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1}\big|^4} \leqslant \frac{C\sigma^2 dh^4}{p}\min\left\{1, \frac{N^2h^2}{mp^2}\right\},$$
$$\sup_{k\geqslant 0}\sqrt{\mathbb{E}\big|\boldsymbol{V}_{k+1} - \bar{\boldsymbol{V}}_{k+1}\big|^4} \leqslant \frac{C\sigma^2 dh^2}{p}\min\left\{1, \frac{N^2h^2}{mp^2}\right\},$$

where the constant C *depends only on* M_1, M_2*.*

Proof Analogous to the expression for SG-UBU, the stochastic gradient error components of the SVRG-UBU integrator are given by:

$$\boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1} = -\frac{h(1-e^{-h})}{2M_2}\boldsymbol{g}_k(\theta_k), \qquad \boldsymbol{V}_{k+1} - \bar{\boldsymbol{V}}_{k+1} = -\frac{he^{-h}}{M_2}\boldsymbol{g}_k(\theta_k), \tag{84}$$

where $\boldsymbol{g}_k(\theta_k) \in \mathbb{R}^d$ is the deviation of the SVRG estimator from the true gradient $\nabla U(\boldsymbol{Y}_k)$. According to (29), this deviation is explicitly written as

$$\begin{aligned}\boldsymbol{g}_k(\theta_k) &= \left(\frac{1}{p}\sum_{i\in\theta_k}\nabla U_i(\boldsymbol{Y}_k) - \frac{1}{p}\sum_{i\in\theta_k}\nabla U_i(\boldsymbol{Y}_*) + \nabla U(\boldsymbol{Y}_*)\right) - \nabla U(\boldsymbol{Y}_k)\\ &= \int_0^1 \left(\frac{1}{p}\sum_{i\in\theta_k}\nabla^2 U_i - \nabla^2 U\right)\big(\lambda\boldsymbol{Y}_k + (1-\lambda)\boldsymbol{Y}_*\big)\mathrm{d}\lambda\,(\boldsymbol{Y}_k - \boldsymbol{Y}_*),\end{aligned} \tag{85}$$

where $\boldsymbol{Y}_* = \boldsymbol{Y}_{q[k/q]}$ is the anchor position.

To bound the fourth moment of $\boldsymbol{g}_k(\theta_k)$, we leverage its integral representation. For a fixed $\lambda \in [0,1]$, we define a set of vectors:

$$a_i(\lambda) = \big(\nabla^2 U_i - \nabla^2 U\big)\big(\lambda\boldsymbol{Y}_k + (1-\lambda)\boldsymbol{Y}_*\big)(\boldsymbol{Y}_k - \boldsymbol{Y}_*), \qquad i = 1, \cdots, N.$$

This sequence of vectors has a zero mean, namely, $\frac{1}{N}\sum_{i=1}^{N} a_i(\lambda) = \mathbf{0}$. We can bound the average eighth moment of $a_i(\lambda)$ using Assumption (SgN^+):

$$\frac{1}{N}\sum_{i=1}^{N} |a_i(\lambda)|^8 \leqslant |\boldsymbol{Y}_k - \boldsymbol{Y}_*|^8 \cdot \frac{C}{N}\sum_{i=1}^{N} \big\|(\nabla^2 U_i - \nabla^2 U)(\lambda \boldsymbol{Y}_k + (1-\lambda)\boldsymbol{Y}_*)\big\|^8 \leqslant C\sigma^8 |\boldsymbol{Y}_k - \boldsymbol{Y}_*|^8.$$

This confirms that the conditions of Lemma 22 are met. Applying the lemma to the mini-batch average of these vectors gives:

$$\mathbb{E}_{\theta_k}\left[\left|\frac{1}{p}\sum_{i\in\theta_k} a_i(\lambda)\right|^8\right] \leqslant \frac{C\sigma^8}{p^4}|\boldsymbol{Y}_k - \boldsymbol{Y}_*|^8. \tag{86}$$

By integrating over $\lambda \in [0,1]$ and applying Jensen's inequality, we obtain a bound on the eighth moment of $\boldsymbol{g}_k(\theta_k)$, which implies a bound on its fourth moment:

$$\mathbb{E}_{\theta_k}\big[|\boldsymbol{g}_k(\theta_k)|^8\big] \leqslant \frac{C\sigma^8}{p^4}|\boldsymbol{Y}_k - \boldsymbol{Y}_*|^8 \quad \Longrightarrow \quad \mathbb{E}_{\theta_k}\big[|\boldsymbol{g}_k(\theta_k)|^4\big] \leqslant \frac{C\sigma^4}{p^2}|\boldsymbol{Y}_k - \boldsymbol{Y}_*|^4. \tag{87}$$

Taking the full expectation over the numerical solution yields:

$$\mathbb{E}\big[|\boldsymbol{g}_k(\theta_k)|^4\big] \leqslant \frac{C\sigma^4}{p^2}\mathbb{E}\big|\boldsymbol{Y}_k - \boldsymbol{Y}_{q[k/q]}\big|^4 \leqslant \frac{C\sigma^4 q^3}{p^2}\sum_{s=q[k/q]}^{k-1}\mathbb{E}|\boldsymbol{Y}_{s+1} - \boldsymbol{Y}_s|^4. \tag{88}$$

The single-step deviation can be bounded using SVRG-UBU from Algorithm 3, which implies $(\boldsymbol{Y}_{s+1}, \boldsymbol{V}_{s+1}^{(1)}) = \Phi_h^{\mathfrak{U}}(\boldsymbol{Y}_s, \boldsymbol{V}_s^{(2)})$. The uniform-in-time moment bounds then give

$$\mathbb{E}|\boldsymbol{Y}_{s+1} - \boldsymbol{Y}_s|^4 \leqslant Ch^4\big(\mathbb{E}|\boldsymbol{Y}_s|^4 + \mathbb{E}|\boldsymbol{V}_s^{(2)}|^4 + d^2\big) \leqslant \frac{Cd^2h^4}{m^2}.$$

Substituting this into (88) and noting that the sum has at most q terms, we arrive at the final bound for the stochastic gradient deviation:

$$\sqrt{\mathbb{E}\big[|\boldsymbol{g}_k(\theta_k)|^4\big]} \leqslant \frac{C\sigma^2 dq^2h^2}{mp}. \tag{89}$$

Alternatively, a simpler bound on the stochastic gradient deviation $\boldsymbol{g}_k(\theta_k)$ that is independent of the distance between the sampling point $\boldsymbol{Y}_k$ and the anchor point $\boldsymbol{Y}_*$ can be obtained via the triangle inequality:

$$\begin{aligned}\mathbb{E}_{\theta_k}\big[|\boldsymbol{g}_k(\theta_k)|^8\big] &\leqslant C\left(\left|\frac{1}{p}\sum_{i\in\theta_k}\nabla U_i(\boldsymbol{Y}_k) - \nabla U(\boldsymbol{Y}_k)\right|^8 + \left|\frac{1}{p}\sum_{i\in\theta_k}\nabla U_i(\boldsymbol{Y}_*) - \nabla U(\boldsymbol{Y}_*)\right|^8\right)\\ &\leqslant \frac{C}{p^4}\left(\frac{1}{N}\sum_{i=1}^{N}\big|\nabla U_i(\boldsymbol{Y}_k) - \nabla U(\boldsymbol{Y}_k)\big|^8 + \frac{1}{N}\sum_{i=1}^{N}\big|\nabla U_i(\boldsymbol{Y}_*) - \nabla U(\boldsymbol{Y}_*)\big|^8\right) \leqslant \frac{C\sigma^8 d^4}{p^4}.\end{aligned}$$

Then we arrive at a uniform bound on the fourth moment:

$$\sqrt{\mathbb{E}\big[|\boldsymbol{g}_k(\theta_k)|^4\big]} \leqslant \frac{C\sigma^2 d}{p}. \tag{90}$$

Combining this uniform bound with the distance-dependent bound from (89) yields a tighter composite estimate using the min function:

$$\sqrt{\mathbb{E}\big[|\boldsymbol{g}_k(\theta_k)|^4\big]} \leqslant \frac{C\sigma^2 d}{p} \min\left\{1, \frac{q^2h^2}{m}\right\} = \frac{C\sigma^2 d}{p} \min\left\{1, \frac{N^2h^2}{mp^2}\right\}. \tag{91}$$

Finally, substituting the comprehensive bound from (91) into the local error expressions in (84) yields the result and completes the proof. ■

E.3 MSE of SVRG-UBU

The uniform-in-time moment bound from Theorem 26 and the local stochastic gradient error estimate from Lemma 28 provide the necessary stability and consistency results for the SVRG-UBU integrator. Consequently, the MSE bound in Theorem 11 follows directly by applying the analytical procedure detailed in the proof of Theorem 7.

Two points deserve mention. The anchor $\boldsymbol{Y}_*$ is $\mathcal{G}_k$-measurable, so the martingale property $\mathbb{E}[R_{k,1}|\mathcal{G}_k] = 0$ and the resulting orthogonality carry over verbatim. Moreover, Lemma 6 and Lemma 21 are stated under Assumption (Sg), while their proofs use only Assumption (P), the uniform-in-time eighth moment, and the growth bound in Assumption (SgN^+), all of which SVRG-UBU satisfies.

E.4 Numerical bias of SVRG-UBU

Proof of Theorem 12 Let $q = N/p$ be the length of one epoch, and let $\pi_h(\boldsymbol{x}, \boldsymbol{v})$ be the invariant distribution of the SVRG-UBU subsequence $(\boldsymbol{X}_{kq}, \boldsymbol{V}_{kq})_{k=0}^{\infty}$. Let the initial state $\boldsymbol{Z}_0 \sim \pi_h$, then the state after one full epoch is also distributed according to π_h, i.e., $\boldsymbol{Z}_q \sim \pi_h$.

Using this property with the discrete Poisson equation (10), we can express the bias of the initial state distribution as follows:

$$\begin{aligned}\mathbb{E}^{\pi_h}[f(\boldsymbol{Z}_0)] - \pi(f) &= \mathbb{E}^{\pi_h}\left[\left(\frac{1 - e^{h\mathcal{L}}}{h}\phi_h\right)(\boldsymbol{Z}_0)\right] = \mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_0(h))}{h}\right] \\ &= \mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_q) - \phi_h(\boldsymbol{Z}_0(h))}{h}\right]. \end{aligned} \tag{92}$$

To connect this expression to the bias of the intermediate steps within the epoch, we decompose the term $\phi_h(\boldsymbol{Z}_q) - \phi_h(\boldsymbol{Z}_0(h))$ into a sum over the q steps:

$$\begin{aligned}\mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_q) - \phi_h(\boldsymbol{Z}_0(h))}{h}\right] &= \sum_{k=0}^{q-1} \mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\boldsymbol{Z}_k(h))}{h}\right] + \sum_{k=1}^{q-1} \mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_k(h)) - \phi_h(\boldsymbol{Z}_k)}{h}\right] \\ &= \sum_{k=0}^{q-1} \mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\boldsymbol{Z}_k(h))}{h}\right] + \sum_{k=1}^{q-1} \mathbb{E}^{\pi_h}\left[\frac{e^{h\mathcal{L}}\phi_h(\boldsymbol{Z}_k) - \phi_h(\boldsymbol{Z}_k)}{h}\right] \\ &= \sum_{k=0}^{q-1} \mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\boldsymbol{Z}_k(h))}{h}\right] - \sum_{k=1}^{q-1} \big(\mathbb{E}[f(\boldsymbol{Z}_k)] - \pi(f)\big).\end{aligned}$$

By combining this decomposition with (92) and rearranging, we obtain an expression for the average bias over one epoch:

$$\frac{1}{q}\sum_{k=0}^{q-1}\mathbb{E}[f(\boldsymbol{Z}_k)] - \pi(f) = \frac{1}{q}\sum_{k=0}^{q-1}\mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\boldsymbol{Z}_k(h))}{h}\right]. \tag{93}$$

By the definition of the averaged distribution $\tilde{\pi}_h$, the left-hand side is precisely the numerical bias $\tilde{\pi}_h(f) - \pi(f)$. This yields the final expression

$$\tilde{\pi}_h(f) - \pi(f) = \frac{1}{q}\sum_{k=0}^{q-1}\mathbb{E}^{\pi_h}\left[\frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\boldsymbol{Z}_k(h))}{h}\right]. \tag{94}$$

Following the same decomposition strategy used in the proof of Theorem 9, we separate each local error term on the right-hand side of (94) into its corresponding stochastic gradient error component, R_k, and discretization error component, S_k. The expected values of these components satisfy bounds analogous to those derived for the standard SG-UBU case:

$$\left|\mathbb{E}[R_k] - \frac{h}{2M_2^2}\mathbb{E}_{\theta_k}^{\pi_h}\left[\big(\boldsymbol{g}(\boldsymbol{Y}_k, \boldsymbol{Y}_0, \theta_k)\big)^\top \boldsymbol{M}_k\big(\boldsymbol{g}(\boldsymbol{Y}_k, \boldsymbol{Y}_0, \theta_k)\big)\right]\right| \leqslant \frac{Cdh^2}{m^2}, \quad \big|\mathbb{E}[S_k]\big| \leqslant \frac{Cdh^2}{m^3}.$$

By substituting these estimates into (94) and summing the error bounds over the epoch from $k = 0$ to $q - 1$, we arrive at the explicit characterization of the numerical bias:

$$\left|\tilde{\pi}_h(f) - \pi(f) - \frac{h}{2M_2^2 q}\mathbb{E}_{\theta_0,\cdots,\theta_{q-1}}^{\pi_h}\left[\sum_{k=0}^{q-1}\big(\boldsymbol{g}(\boldsymbol{Y}_k, \boldsymbol{Y}_0, \theta_k)\big)^\top \boldsymbol{M}_k\big(\boldsymbol{g}(\boldsymbol{Y}_k, \boldsymbol{Y}_0, \theta_k)\big)\right]\right| \leqslant \frac{Cdh^2}{m^3}. \tag{95}$$

Next, since the numerical solution $(\boldsymbol{Y}_k)_{k=0}^{q-1}$ has bounded eighth moments, we can apply the bound from (91) to obtain the following estimate on the second moment:

$$\mathbb{E}\Big[\big|\boldsymbol{g}(\boldsymbol{Y}_k, \boldsymbol{Y}_0, \theta_k)\big|^2\Big] \leqslant \frac{C\sigma^2 d}{p}\min\left\{1, \frac{N^2h^2}{mp^2}\right\}. \tag{96}$$

Finally, combining (95) and (96) yields the inequality

$$|\tilde{\pi}_h(f) - \pi(f)| \leqslant C\left(\frac{\sigma^2 dh}{m^2 p}\min\left\{1, \frac{N^2h^2}{mp^2}\right\} + \frac{dh^2}{m^3}\right),$$

thus completing the proof of Theorem 12. ■

Appendix F. Proofs for MSE and numerical bias of SAGA-UBU

The MSE analysis for SAGA-UBU proceeds by establishing its stability and consistency. Stability is addressed by deriving a modified Lyapunov condition for the corresponding SAGA-BU integrator, while consistency is analyzed by bounding the moments of the stochastic gradient error. A key challenge is the non-Markovian nature of the SAGA-UBU solution

$(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$, which arises from its dependence on the evolving table of historical positions $\{\boldsymbol{\Phi}_i\}_{i=1}^N$. The position-velocity marginal $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ is therefore not itself a Markov chain, so no invariant distribution of that marginal chain is available. Consequently, the numerical bias, $\tilde{\pi}_h(f) - \pi(f)$, is measured with respect to a limit distribution $\tilde{\pi}_h(\boldsymbol{x}, \boldsymbol{v})$, defined as the subsequential limit of the time-averaged distributions of the SAGA-UBU trajectory.

To formulate a compact representation for the SAGA-UBU integrator, we first define a solution flow that incorporates the SAGA gradient estimator. For a given table of historical positions $\{\boldsymbol{\Phi}_i\}_{i=1}^N$ and a single randomly drawn index $\theta \in \{1, \ldots, N\}$ (corresponding to the batch size $p = 1$ case), this flow is defined as:

$$\Phi_t^{\mathfrak{B}}(\{\boldsymbol{\Phi}_i\}_{i=1}^N, \theta) : \begin{cases} \boldsymbol{x}_t = \boldsymbol{x}_0, \\ \boldsymbol{v}_t = \boldsymbol{v}_0 - \dfrac{t}{M_2}\left(\nabla U_\theta(\boldsymbol{x}_0) - \nabla U_\theta(\boldsymbol{\Phi}_\theta) + \dfrac{1}{N}\displaystyle\sum_{i=1}^N \nabla U_i(\boldsymbol{\Phi}_i)\right). \end{cases}$$

Then the SAGA-UBU integrator can be expressed as a two-part update rule for each iteration k. The first equation describes the evolution of the state $\boldsymbol{Z}_k$ using the UBU splitting scheme, while the second equation describes the update to the historical position table for the sampled index θ_k:

$$\begin{cases} \boldsymbol{Z}_{k+1} = \left(\Phi_{h/2}^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\{\boldsymbol{\Phi}_i\}_{i=1}^N, \theta_k) \circ \Phi_{h/2}^{\mathfrak{U}}\right) \boldsymbol{Z}_k, \\ \boldsymbol{\Phi}_{\theta_k} = \boldsymbol{Y}_k, \end{cases} \qquad k = 0, 1, \cdots, \tag{97}$$

where the historical positions $\{\boldsymbol{\Phi}_i\}_{i=1}^N$ are initialized at $\boldsymbol{Y}_0$, and $(\theta_k)_{k=0}^{\infty}$ are selected independently and uniformly from $\{1, \cdots, N\}$ at each step.

Specifically, at step k of SAGA-UBU, the historical position $\boldsymbol{\Phi}_i$ for each component i is set to the intermediate state $\boldsymbol{Y}_{k_i}$, where k_i is the index of the most recent iteration prior to k in which the gradient $\nabla U_i(\boldsymbol{x})$ was evaluated. This relationship is defined formally as

$$\boldsymbol{\Phi}_i = \boldsymbol{Y}_{k_i}, \qquad k_i = \max\Big\{0 \leqslant l \leqslant k-1 : \nabla U_i(\boldsymbol{x}) \text{ was computed at step } l\Big\}. \tag{98}$$

By convention, if the gradient for a component i has not been computed in any of the steps from 1 to $k-1$, its historical index is taken to be the initial step, $k_i = 0$.

F.1 Uniform-in-time moment bound of SAGA-UBU

Following our analytical strategy, we begin by examining the SAGA-BU integrator, which is the BU-type counterpart to SAGA-UBU. Its update rule, similar to (80), is given by

$$\begin{cases} \boldsymbol{Z}_{k+1} = \left(\Phi_h^{\mathfrak{U}} \circ \Phi_h^{\mathfrak{B}}(\{\boldsymbol{\Phi}_i\}_{i=1}^N, \theta_k)\right) \boldsymbol{Z}_k, \\ \boldsymbol{\Phi}_{\theta_k} = \boldsymbol{X}_k, \end{cases} \qquad k = 0, 1, \cdots, \tag{99}$$

where the table of historical positions $\{\boldsymbol{\Phi}_i\}_{i=1}^N$ is initialized with $\boldsymbol{X}_0$. At each step k, the historical position for a given component i is determined by

$$\boldsymbol{\Phi}_i = \boldsymbol{X}_{k_i}, \qquad k_i = \max\Big\{0 \leqslant l \leqslant k-1 : \nabla U_i(\boldsymbol{x}) \text{ was computed at step } l\Big\}. \tag{100}$$

To establish a uniform-in-time moment bound, we derive a Lyapunov condition for this integrator. Unlike the condition for SVRG-BU, the non-Markovian nature of SAGA introduces a dependency on the maximum moment over the entire history of the process. This modified Lyapunov condition is stated in the following lemma.

Lemma 29 *Under Assumptions* (P) *and* (SgN^+)*, let the step size* $h \leqslant \frac{1}{4}$*; then the SAGA-BU solution* $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ *with the batch size* 1 *satisfies*

$$\mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_{k+1})\big] \leqslant \left(1 - \frac{mh}{3M_2}\right)\mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_k)\big] + \frac{Cd^4h}{m^3} + \frac{Ch^5}{m^3}\max_{0\leqslant l\leqslant k}\mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_l)\big],$$

where the constant C *depends only on* M_1, M_2*.*

Proof The Lyapunov condition for the SAGA-BU integrator can be validated by adapting the proof for the full-gradient case from Lemma 20. The core of the argument is to analyze the stochastic gradient error, which is the difference between the SAGA-BU and FG-BU updates, namely, $\boldsymbol{Z}_{k+1}$ and $\bar{\boldsymbol{Z}}_{k+1}$. This error can be written as

$$\boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1} = -\frac{h(1-e^{-2h})}{2M_2}\boldsymbol{g}_k(\theta_k), \qquad \boldsymbol{V}_{k+1} - \bar{\boldsymbol{V}}_{k+1} = -\frac{he^{-2h}}{M_2}\boldsymbol{g}_k(\theta_k), \tag{101}$$

where the deviation term $\boldsymbol{g}_k(\theta_k) \in \mathbb{R}^d$ is explicitly defined as

$$\boldsymbol{g}_k(\theta_k) = \nabla U_{\theta_k}(\boldsymbol{X}_k) - \nabla U_{\theta_k}(\boldsymbol{\Phi}_{\theta_k}) + \frac{1}{N}\sum_{i=1}^{N}\nabla U_i(\boldsymbol{\Phi}_i) - \frac{1}{N}\sum_{i=1}^{N}\nabla U_i(\boldsymbol{X}_k). \tag{102}$$

The analysis then proceeds by bounding the moments of this deviation term.

To bound the moments of $\boldsymbol{g}_k(\theta_k) \in \mathbb{R}^d$, we first note that $\boldsymbol{g}_k(\theta_k)$ is unbiased with respect to θ_k for a fixed history. Applying the triangle inequality gives

$$|\boldsymbol{g}_k(\theta_k)| \leqslant \left|\nabla U_{\theta_k}(\boldsymbol{X}_k) - \frac{1}{N}\sum_{i=1}^{N}\nabla U_i(\boldsymbol{X}_k)\right| + |\nabla U_{\theta_k}(\boldsymbol{\Phi}_{\theta_k})| + \frac{1}{N}\sum_{i=1}^{N}|\nabla U_i(\boldsymbol{\Phi}_i)|. \tag{103}$$

Taking the eighth moment and the full expectation over the numerical solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ and the random index $(\theta_k)_{k=0}^{\infty}$ in (103), we obtain

$$\begin{aligned}
\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^8 &\leqslant C\left(\mathbb{E}\left|\nabla U_{\theta_k}(\boldsymbol{X}_k) - \frac{1}{N}\sum_{i=1}^{N}\nabla U_i(\boldsymbol{X}_k)\right|^8 + \mathbb{E}|\nabla U_{\theta_k}(\boldsymbol{\Phi}_{\theta_k})|^8 + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}|\nabla U_i(\boldsymbol{\Phi}_i)|^8\right)\\
&\leqslant C\left((\sigma^2 d)^4 + \frac{2}{N}\sum_{i=1}^{N}\mathbb{E}|\nabla U_i(\boldsymbol{\Phi}_i)|^8\right) \leqslant C\left(d^4 + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big(|\boldsymbol{\Phi}_i|^2 + d\big)^4\right)\\
&\leqslant C\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}|\boldsymbol{\Phi}_i|^8 + d^4\right).
\end{aligned}$$

This chain of inequalities provides the key moment bound for the deviation term:

$$\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^8 \leqslant C\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}|\boldsymbol{\Phi}_i|^8 + d^4\right). \tag{104}$$

To establish the Lyapunov condition, we follow the analysis in the proof for Lemma 2. We first define the scaled stochastic gradient error at step k as $\boldsymbol{\Delta}_k = h^{-1}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}) \in \mathbb{R}^{2d}$. To manage the dependency on the algorithm's history, we also introduce a constant C_k that tracks the maximum normalized moment up to step k:

$$C_k = \frac{1}{d^4} \max_{0 \leqslant l \leqslant k} \mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_l)\big] = \frac{1}{d^4} \max_{0 \leqslant l \leqslant k} \mathbb{E}\big[(\boldsymbol{Z}_l^\top \boldsymbol{S} \boldsymbol{Z}_l)^4\big], \qquad k = 0, 1, \cdots.$$

The error term $\boldsymbol{\Delta}_k$ is unbiased. Since $\boldsymbol{X}_l$ is determined by $\theta_0, \cdots, \theta_{l-1}$ while $\mathbf{1}_{\{k_i = l\}}$ depends only on $\theta_l, \cdots, \theta_{k-1}$, the random variables $\boldsymbol{X}_l$ and $\mathbf{1}_{\{k_i = l\}}$ are independent, and

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}|\boldsymbol{\Phi}_i|^8 = \frac{1}{N} \sum_{i=1}^{N} \sum_{l=0}^{k-1} \mathbb{E}|\boldsymbol{X}_l|^8 \, \mathbb{P}(k_i = l) \leqslant \max_{0 \leqslant l \leqslant k-1} \mathbb{E}|\boldsymbol{X}_l|^8 \leqslant C d^4 (C_k + 1).$$

Consequently the moment bound on $\boldsymbol{g}_k(\theta_k)$ from (104) yields a bound on $\boldsymbol{\Delta}_k$ in terms of the historical maximum moment C_k:

$$\mathbb{E}|\boldsymbol{\Delta}_k|^8 \leqslant C d^4 (C_k + 1). \tag{105}$$

We now expand the eighth moment of $\boldsymbol{Z}_{k+1}$ using its decomposition $\boldsymbol{Z}_{k+1} = \bar{\boldsymbol{Z}}_{k+1} + h\boldsymbol{\Delta}_k$. This expansion, analogous to the one in the proof of Lemma 2, yields several terms:

$$\begin{aligned}
\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{Z}_{k+1}\big|^8 &= \mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1} + h\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_k\big|^8 \\
&\leqslant \mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|^8 + 28h^2\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|^6\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_k\big|^2\Big) \\
&\quad + 56h^3\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|^5\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_k\big|^3\Big) + 70h^4\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|^4\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_k\big|^4\Big) \\
&\quad + 56h^5\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|^3\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_k\big|^5\Big) + 28h^6\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|^2\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_k\big|^6\Big) \\
&\quad + 8h^7\mathbb{E}\Big(\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_k\big|^7\Big) + h^8\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{\Delta}_k\big|^8 \\
&\leqslant \mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|^8 + \frac{mh}{12M_2}\mathbb{E}\big|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}\big|^8 + \frac{Ch^5}{m^3}\mathbb{E}|\boldsymbol{\Delta}_k|^8,
\end{aligned}$$

where we have utilized the binomial theorem. The term linear in $\boldsymbol{\Delta}_k$ is dropped from the binomial expansion because $\mathbb{E}[\boldsymbol{\Delta}_k|\mathcal{F}_k] = \mathbf{0}$, and the remaining terms are absorbed into the first and last ones by Young's inequality. Applying the Lyapunov condition for the FG-BU integrator in Lemma 20 to the $\mathbb{E}|\sqrt{\boldsymbol{S}}\bar{\boldsymbol{Z}}_{k+1}|^8$ terms and substituting the bound (105) for the moment of $\boldsymbol{\Delta}_k$, we get

$$\begin{aligned}
\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{Z}_{k+1}\big|^8 &\leqslant \left(1 + \frac{mh}{12M_2}\right)\left(\left(1 - \frac{mh}{2M_2}\right)\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{Z}_k\big|^8 + \frac{Cd^4h}{m^3}\right) + \frac{Cd^4h^5}{m^3}(C_k + 1) \\
&\leqslant \left(1 - \frac{mh}{3M_2}\right)\mathbb{E}\big|\sqrt{\boldsymbol{S}}\boldsymbol{Z}_k\big|^8 + \frac{Cd^4h}{m^3} + \frac{Cd^4h^5}{m^3}C_k.
\end{aligned}$$

Expressing this inequality in terms of the Lyapunov function $\mathcal{V}(\boldsymbol{x}, \boldsymbol{v})$, we arrive at the desired recursive relationship

$$\mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_{k+1})\big] \leqslant \left(1 - \frac{mh}{3M_2}\right)\mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_k)\big] + \frac{Cd^4h}{m^3} + \frac{Ch^5}{m^3} \max_{0 \leqslant l \leqslant k} \mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_l)\big].$$

This completes the proof of the Lyapunov condition for the SAGA-BU integrator. ■

The recursive Lyapunov condition for SAGA-BU from Lemma 29 is sufficient to establish a uniform-in-time moment bound for the full SAGA-UBU integrator. This result requires a sufficiently small step size h to ensure that the contractive dynamics dominate the historical dependency inherent in the SAGA method.

Theorem 30 *Under Assumptions* (P) *and* (SgN^+)*, let the step size* $h \leqslant c_0 m$*, and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the SAGA-UBU solution $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ *satisfies*

$$\sup_{k\geqslant 0} \mathbb{E}\big[|\boldsymbol{Z}_k|^8\big] \leqslant \frac{Cd^4}{m^4},$$

where the constants c_0 *and* C *depend only on* M_1, M_2*.*

Proof To prove the uniform-in-time moment bound for the SAGA-BU integrator, we analyze the recursive relationship for the maximum normalized moment, defined as

$$C_k = \frac{1}{d^4} \max_{0\leqslant l\leqslant k} \mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_l)\big], \qquad k = 0, 1, \cdots.$$

Writing $a_k = \frac{1}{d^4}\mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_k)\big]$, so that $C_k = \max_{0\leqslant l\leqslant k} a_l$, the Lyapunov condition from Lemma 29 divided by d^4 provides the following inequality:

$$a_{k+1} \leqslant \left(1 - \frac{mh}{3M_2}\right)a_k + \frac{Ch}{m^3} + \frac{Ch^5}{m^3}C_k \leqslant \left(1 - \frac{mh}{3M_2}\right)C_k + \frac{Ch}{m^3} + \frac{Ch^5}{m^3}C_k. \tag{106}$$

The challenge lies in the final term, $\frac{Ch^5}{m^3}C_k$, which reflects the integrator's dependency on its entire history. To ensure stability, we must choose a step size h small enough for the contractive term to dominate this historical dependency. By setting a threshold

$$c_0 = \min\big\{(6CM_2)^{-1/4}, (4M_2)^{-1}\big\},$$

a step size $h \leqslant c_0 m$ guarantees that $\frac{Ch^5}{m^3} \leqslant \frac{mh}{6M_2}$, and also that $h \leqslant \frac{m}{4M_2} \leqslant \frac{1}{4}$, as required by Lemma 29. Writing $\gamma = \frac{m}{3M_2}$ and $\beta = \frac{C}{m^3}$, the inequality (106) simplifies to

$$a_{k+1} \leqslant \left(1 - \frac{\gamma h}{2}\right)C_k + \beta h.$$

Next we prove $C_k \leqslant \max\big\{C_0, \frac{2\beta}{\gamma} + \beta h\big\}$ by induction on k. If $C_k \geqslant \frac{2\beta}{\gamma}$, then

$$a_{k+1} \leqslant C_k - \frac{\gamma h}{2}C_k + \beta h \leqslant C_k - \beta h + \beta h = C_k,$$

so that $C_{k+1} = \max\{C_k, a_{k+1}\} = C_k$ and the bound is inherited. If instead $C_k < \frac{2\beta}{\gamma}$, then $a_{k+1} \leqslant \frac{2\beta}{\gamma} + \beta h$, and again $C_{k+1} \leqslant \frac{2\beta}{\gamma} + \beta h$. Since $\frac{\beta}{\gamma} = \frac{3CM_2}{m^4}$ and $\beta h \leqslant \frac{\beta}{\gamma}$, the sequence $(C_k)_{k=0}^{\infty}$ is uniformly bounded:

$$\sup_{k\geqslant 0} C_k \leqslant \max\left\{C_0, \frac{C}{m^4}\right\} \leqslant \frac{C}{m^4},$$

and therefore $\sup_{k\geqslant 0}\mathbb{E}\big[\mathcal{V}(\boldsymbol{Z}_k)\big] \leqslant d^4 \sup_{k\geqslant 0} C_k \leqslant \frac{Cd^4}{m^4}$.

This establishes the desired uniform-in-time moment bound for the SAGA-BU integrator, and the same argument extends to SAGA-UBU by conjugation. Writing $\tilde{\boldsymbol{Z}}_k = \Phi^{\mathfrak{U}}_{h/2}\boldsymbol{Z}_k$ for the SAGA-UBU solution (97), the semigroup property $\Phi^{\mathfrak{U}}_{h/2} \circ \Phi^{\mathfrak{U}}_{h/2} = \Phi^{\mathfrak{U}}_h$ gives

$$\tilde{\boldsymbol{Z}}_{k+1} = \Phi^{\mathfrak{U}}_{h/2}\boldsymbol{Z}_{k+1} = \big(\Phi^{\mathfrak{U}}_h \circ \Phi^{\mathfrak{B}}_h(\{\boldsymbol{\Phi}_i\}_{i=1}^N, \theta_k)\big)\tilde{\boldsymbol{Z}}_k,$$

while the position component of $\tilde{\boldsymbol{Z}}_k$ is exactly $\boldsymbol{Y}_k$, which is the state stored in the table by (97). Hence $(\tilde{\boldsymbol{Z}}_k)_{k=0}^{\infty}$ is a SAGA-BU solution in the sense of (99), and the bound just proved applies to it. Since $\boldsymbol{Z}_{k+1} = \Phi^{\mathfrak{U}}_{h/2}\big(\Phi^{\mathfrak{B}}_h(\{\boldsymbol{\Phi}_i\}_{i=1}^N, \theta_k)\tilde{\boldsymbol{Z}}_k\big)$ is obtained from $\tilde{\boldsymbol{Z}}_k$ by one velocity update $-\frac{h}{M_2}\boldsymbol{b}_k$ followed by the integrator flow $\Phi^{\mathfrak{U}}_{h/2}$, and $\mathbb{E}|\boldsymbol{b}_k|^8$ is controlled by the same factorization $\mathbb{E}|\boldsymbol{\Phi}_i|^8 = \sum_l \mathbb{E}|\boldsymbol{Y}_l|^8\,\mathbb{P}(k_i = l) \leqslant \max_l \mathbb{E}|\boldsymbol{Y}_l|^8$ used above, we conclude

$$\sup_{k\geqslant 0}\mathbb{E}|\boldsymbol{Z}_k|^8 \leqslant C\big(\sup_{k\geqslant 0}\mathbb{E}|\tilde{\boldsymbol{Z}}_k|^8 + d^4\big) \leqslant \frac{Cd^4}{m^4},$$

completing the proof of Theorem 30. ■

Due to the non-Markovian nature of the SAGA-UBU solution, a standard invariant distribution cannot be defined. Instead, the numerical bias is analyzed with respect to $\tilde{\pi}_h(\boldsymbol{x}, \boldsymbol{v})$, which represents the subsequential weak limit of the time-averaged distributions along the trajectory, as established in the following lemma.

Lemma 31 *Under Assumptions* (P) *and* (SgN^+)*, let the step size $h \leqslant c_0 m$; define $\tilde{\pi}_h^K$ as the averaged distribution of SAGA-UBU over the first K steps:*

$$\tilde{\pi}_h^K(A) = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{P}\big(\boldsymbol{Z}_k \in A \mid \boldsymbol{X}_0 = \boldsymbol{0},\ \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d)\big).$$

Then $\{\tilde{\pi}_h^K\}_{K=1}^{\infty}$ admits a weakly convergent subsequence with limit $\tilde{\pi}_h$ satisfying

$$\int_{\mathbb{R}^{2d}} |\boldsymbol{z}|^8 \mathrm{d}\tilde{\pi}_h(\boldsymbol{z}) \leqslant \frac{Cd^4}{m^4},$$

where the constants c_0, C depend only on M_1, M_2.

Moreover, for any Lipschitz continuous test function $g(\boldsymbol{x}, \boldsymbol{v})$ on $\mathbb{R}^{2d}$, $\tilde{\pi}_h(g)$ is the subsequential limit of $\{\tilde{\pi}_h^K(g)\}_{K=1}^{\infty}$ in $\mathbb{R}$.

Proof To prove the existence of a subsequential weak limit $\tilde{\pi}_h$, we introduce a sequence of weighted measures. For each $K \in \mathbb{N}$, define the positive measure $\tilde{\mu}_h^K$ on $\mathbb{R}^{2d}$ via the Radon–Nikodym derivative with respect to the averaged distribution $\tilde{\pi}_h^K$:

$$\frac{\mathrm{d}\tilde{\mu}_h^K}{\mathrm{d}\tilde{\pi}_h^K}(\boldsymbol{x},\boldsymbol{v}) = |\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d.$$

The uniform-in-time moment of SAGA-UBU in Theorem 30 implies uniform bounds on the total mass and higher moments of these weighted measures:

$$\int_{\mathbb{R}^{2d}} \frac{1}{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d} \mathrm{d}\tilde{\mu}_h^K(\boldsymbol{x},\boldsymbol{v}) = \int_{\mathbb{R}^{2d}} \mathrm{d}\tilde{\pi}_h^K(\boldsymbol{x},\boldsymbol{v}) = 1,$$

and

$$\int_{\mathbb{R}^{2d}} (|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d)^3 \mathrm{d}\tilde{\mu}_h^K(\boldsymbol{x},\boldsymbol{v}) = \int_{\mathbb{R}^{2d}} (|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d)^4 \mathrm{d}\tilde{\pi}_h^K(\boldsymbol{x},\boldsymbol{v}) \leqslant \frac{Cd^4}{m^4}.$$

These uniform bounds ensure that the sequence of measures $\{\tilde{\mu}_h^K\}_{K=1}^{\infty}$ is tight. By Prokhorov's theorem, there exists a subsequence $\{\tilde{\mu}_h^{K_j}\}_{j=1}^{\infty}$ that converges weakly to a limit measure $\tilde{\mu}_h$. This limit measure inherits the same moment bounds:

$$\int_{\mathbb{R}^{2d}} \frac{1}{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d} \mathrm{d}\tilde{\mu}_h(\boldsymbol{x},\boldsymbol{v}) = 1, \quad \int_{\mathbb{R}^{2d}} (|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d)^3 \mathrm{d}\tilde{\mu}_h(\boldsymbol{x},\boldsymbol{v}) \leqslant \frac{Cd^4}{m^4}.$$

We define the probability measure $\tilde{\pi}_h$ via its Radon–Nikodym derivative with respect to $\tilde{\mu}_h$:

$$\frac{\mathrm{d}\tilde{\pi}_h}{\mathrm{d}\tilde{\mu}_h}(\boldsymbol{x},\boldsymbol{v}) = \frac{1}{|\boldsymbol{x}|^2 + |\boldsymbol{v}|^2 + d}.$$

By construction, $\tilde{\pi}_h$ is the weak limit of the subsequence $\{\tilde{\pi}_h^{K_j}\}$. Therefore, for any test function $g(\boldsymbol{x},\boldsymbol{v})$ with linear growth, the following holds

$$\lim_{j\to\infty} \int_{\mathbb{R}^{2d}} g(\boldsymbol{x},\boldsymbol{v}) \mathrm{d}\tilde{\pi}_h^{K_j}(\boldsymbol{x},\boldsymbol{v}) = \int_{\mathbb{R}^{2d}} g(\boldsymbol{x},\boldsymbol{v}) \mathrm{d}\tilde{\pi}_h(\boldsymbol{x},\boldsymbol{v}).$$

This establishes $\tilde{\pi}_h$ as a subsequential weak limit possessing the stated moment bound and convergence property, completing the proof. ■

F.2 Local error analysis of SAGA-UBU

A key step in the local error analysis for the SAGA-UBU integrator is to bound the distance between the current intermediate position $\boldsymbol{Y}_k$ and the historical positions $\{\boldsymbol{\Phi}_i\}_{i=1}^N$. The magnitude of this distance directly determines the variance, and thus the accuracy, of the SAGA gradient estimator. The following lemma establishes the required moment bound.

Lemma 32 *Under Assumptions* (P) *and* (SgN^+)*, let the step size* $h \leqslant c_0 m$*, and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \boldsymbol{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, M_2^{-1}\boldsymbol{I}_d);$$

for each $k \geqslant 1$, let $\boldsymbol{Y}_k$ be the position component of $\Phi_{h/2}^{\mathfrak{U}}\boldsymbol{Z}_k$ defined in (9)*, and $\{\boldsymbol{\Phi}_i\}_{i=1}^N$ be the historical positions of SAGA-UBU in* (98)*, then*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big[|\boldsymbol{Y}_k-\boldsymbol{\Phi}_i|^4\big]\leqslant\frac{CN^4d^2h^4}{m^2},$$

where the constants c_0 and C depend only on M_1, M_2.

Proof The proof begins by expressing the average fourth moment distance between the current position $\boldsymbol{Y}_k$ and the historical positions $\{\boldsymbol{\Phi}_i\}_{i=1}^N$. For each component i, the historical position is $\boldsymbol{\Phi}_i=\boldsymbol{Y}_{k_i}$, where k_i is the last step at which the gradient ∇U_i was evaluated. By repeatedly applying Hölder's inequality, we can bound this distance by

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}|\boldsymbol{Y}_k-\boldsymbol{\Phi}_i|^4&=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}|\boldsymbol{Y}_k-\boldsymbol{Y}_{k_i}|^4\\&\leqslant\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\bigg(\sum_{l=k_i}^{k-1}|\boldsymbol{Y}_{l+1}-\boldsymbol{Y}_l|\bigg)^4\\&\leqslant\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\bigg[(k-k_i)^3\sum_{l=k_i}^{k-1}|\boldsymbol{Y}_{l+1}-\boldsymbol{Y}_l|^4\bigg]\\&=\frac{1}{N}\sum_{i=1}^{N}\sum_{l=0}^{k-1}\mathbb{E}\Big[(k-k_i)^3\mathbf{1}_{\{k_i\leqslant l\}}|\boldsymbol{Y}_{l+1}-\boldsymbol{Y}_l|^4\Big].\end{aligned}$$

Unlike the estimate in the proof of Lemma 33, the summand here involves $\boldsymbol{Y}_{l+1}-\boldsymbol{Y}_l$ for $l\geqslant k_i$, which is not independent of $\mathbf{1}_{\{k_i\leqslant l\}}$; the expectation therefore does not factor, and the following split cannot be avoided. Applying Cauchy's inequality separates the expectation, isolating the term related to the random indices from the term involving the dynamics:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}|\boldsymbol{Y}_k-\boldsymbol{\Phi}_i|^4\leqslant\frac{1}{N}\sum_{i=1}^{N}\sum_{l=0}^{k-1}\sqrt{\mathbb{E}\big[(k-k_i)^6\mathbf{1}_{\{k_i\leqslant l\}}\big]}\sqrt{\mathbb{E}|\boldsymbol{Y}_{l+1}-\boldsymbol{Y}_l|^8}.\tag{107}$$

The eighth moment of the single-step deviation, $\boldsymbol{Y}_{l+1}-\boldsymbol{Y}_l$, can be bounded using an argument similar to that in the proof of Lemma 21, yielding

$$\sqrt{\mathbb{E}|\boldsymbol{Y}_{l+1}-\boldsymbol{Y}_l|^8}\leqslant\frac{Cd^2h^4}{m^2},\qquad l=0,1,\cdots.$$

Substituting this bound into (107) simplifies the expression. Since the random indices $\{k_i\}_{i=1}^N$ are identically distributed, the average over i can be removed:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}|\boldsymbol{Y}_k-\boldsymbol{\Phi}_i|^4&\leqslant\frac{Cd^2h^4}{Nm^2}\sum_{i=1}^{N}\sum_{l=0}^{k-1}\sqrt{\mathbb{E}\big[(k-k_i)^6\mathbf{1}_{\{k_i\leqslant l\}}\big]}\\&=\frac{Cd^2h^4}{m^2}\sum_{l=0}^{k-1}\sqrt{\mathbb{E}\big[(k-k_1)^6\mathbf{1}_{\{k_1\leqslant l\}}\big]}.\end{aligned}$$

Thus, the proof of Lemma 32 is reduced to establishing the following purely probabilistic inequality concerning the distribution of the random index k_1:

$$\sum_{l=0}^{k-1} \sqrt{\mathbb{E}\big[(k-k_1)^6 \mathbf{1}_{\{k_1 \leqslant l\}}\big]} \leqslant CN^4. \tag{108}$$

To establish the probabilistic inequality (108), we formally define the random variable k_1. It signifies the index of the last iteration before step k where the gradient $\nabla U_1(\boldsymbol{x})$ was computed. Given the sequence of randomly chosen indices $\theta_1, \ldots, \theta_{k-1}$, k_1 is defined as

$$k_1 = \begin{cases} \max\{l : \theta_l = 1\}, & \text{if } \exists\, l \in \{1, \cdots, k-1\} \text{ with } \theta_l = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Due to the i.i.d. nature of the index selection, a time-reversal symmetry argument shows that the distribution of the "age" of the gradient, $k - k_1$, is identical to the distribution of the "waiting time" for the first selection of index 1, which we denote by k_1':

$$k_1' = \begin{cases} \min\{l : \theta_l = 1\}, & \text{if } \exists\, l \in \{1, \cdots, k-1\} \text{ with } \theta_l = 1, \\ k, & \text{otherwise.} \end{cases}$$

The map $l \mapsto k - l$ is a bijection of $\{0, \cdots, k-1\}$ onto $\{1, \cdots, k\}$, and $\{k_1 \leqslant l\} = \{k - k_1 \geqslant k - l\}$. Combining this reindexing with the identity in law $k - k_1 \overset{\mathrm{d}}{=} k_1'$ transforms (108) into the equivalent statement

$$\sum_{l=1}^{k} \sqrt{\mathbb{E}\big[(k_1')^6 \mathbf{1}_{\{k_1' \geqslant l\}}\big]} \leqslant CN^4.$$

Expanding the expectation inside the square root, we only need to prove

$$\sum_{l=1}^{k} \sqrt{\sum_{s=l}^{k} s^6 \cdot \mathbb{P}(k_1' = s)} \leqslant CN^4. \tag{109}$$

The random variable k_1' follows a truncated geometric distribution. As a consequence, the probability $\mathbb{P}(k_1' = s)$ is given by the standard geometric probability:

$$\mathbb{P}(k_1' = s) = \left(1 - \frac{1}{N}\right)^{s-1} \frac{1}{N}, \qquad s = 1, \cdots, k-1.$$

The event $k_1' = k$ occurs if index 1 is not selected in any of the first $k-1$ steps:

$$\mathbb{P}(k_1' = k) = \left(1 - \frac{1}{N}\right)^{k-1}.$$

Conveniently, this probability can be expressed as the tail sum of the corresponding infinite geometric distribution: $\mathbb{P}(k_1' = k) = \sum_{s=k}^{\infty} (1 - \frac{1}{N})^{s-1} \frac{1}{N}$.

Substituting this distribution into (109) allows us to replace the finite sum up to k with an infinite sum, effectively removing the dependency on k and reducing the problem to proving the following algebraic inequality:

$$\sum_{l=1}^{\infty}\sqrt{\sum_{s=l}^{\infty}s^6\Big(1-\frac{1}{N}\Big)^{s-1}\frac{1}{N}}\leqslant CN^4. \tag{110}$$

To establish (110), we bound the terms in the inner sum. Assume $N\geqslant 2$; the case $N=1$ is trivial, since then $k_1=k-1$ almost surely and the left-hand side of (108) equals 1. Let $j=\lfloor s/N\rfloor$. For s such that $jN\leqslant s<(j+1)N$, we have $s^6\leqslant(j+1)^6N^6$, and combining $(1-1/N)^{jN}\leqslant e^{-j}$ with $(1-1/N)^{-1}=\frac{N}{N-1}\leqslant 2$ gives

$$s^6\Big(1-\frac{1}{N}\Big)^{s-1}\leqslant(j+1)^6N^6\Big(1-\frac{1}{N}\Big)^{jN-1}\leqslant 2(j+1)^6N^6e^{-j}.$$

Each block $\{s: jN\leqslant s<(j+1)N\}$ contains N values of s, so the factor $\frac{1}{N}$ carried by $\mathbb{P}(k_1'=s)$ cancels against the block size:

$$\sum_{s=l}^{\infty}s^6\Big(1-\frac{1}{N}\Big)^{s-1}\frac{1}{N}\leqslant\frac{1}{N}\sum_{j\geqslant\lfloor l/N\rfloor}N\cdot 2(j+1)^6N^6e^{-j}=2N^6\sum_{j\geqslant\lfloor l/N\rfloor}(j+1)^6e^{-j}.$$

The inequality (110) then simplifies to

$$\sum_{l=1}^{\infty}\sqrt{2N^6\sum_{j=\lfloor l/N\rfloor}^{\infty}(j+1)^6e^{-j}}\leqslant CN^4. \tag{111}$$

Factoring N^3 out of the square root, and grouping the outer sum into blocks of N consecutive values of l sharing the same $i=\lfloor l/N\rfloor$, contributes a further factor N; since $N^3\cdot N=N^4$, we arrive at the final simplified inequality, which is independent of N:

$$\sum_{i=0}^{\infty}\sqrt{\sum_{j=i}^{\infty}(j+1)^6e^{-j}}\leqslant C. \tag{112}$$

This final inequality (112) holds because the exponential term e^{-j} decays faster than any polynomial in j. Specifically, using the bound $e^{-j}\leqslant C(j+1)^{-10}$, we have:

$$\sum_{j=i}^{\infty}(j+1)^6e^{-j}\leqslant C\sum_{j=i}^{\infty}\frac{(j+1)^6}{(j+1)^{10}}=C\sum_{j=i}^{\infty}\frac{1}{(j+1)^4}\leqslant\frac{C}{(i+1)^3}.$$

Therefore, the outer sum converges:

$$\sum_{i=0}^{\infty}\sqrt{\sum_{j=i}^{\infty}(j+1)^6e^{-j}}\leqslant C\sum_{i=0}^{\infty}\frac{1}{(i+1)^{3/2}}<+\infty.$$

This confirms (112), which in turn proves (108) and thus establishes Lemma 32. ■

Building on the previous results, we now establish bounds for the stochastic gradient error of the SAGA-UBU integrator.

Lemma 33 *Under Assumptions* (P) *and* (SgN^+), *let the step size* $h \leqslant c_0 m$, *and*

$$\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{V}_0) \text{ with } \boldsymbol{X}_0 = \mathbf{0} \text{ and } \boldsymbol{V}_0 \sim \mathcal{N}(\mathbf{0}, M_2^{-1}\boldsymbol{I}_d);$$

then the stochastic gradient error of SAGA-UBU with the batch size 1 *satisfies*

$$\sup_{k\geqslant 0} \sqrt{\mathbb{E}\big|\boldsymbol{X}_{k+1} - \bar{\boldsymbol{X}}_{k+1}\big|^4} \leqslant \frac{Cdh^4}{m} \min\big\{1, N^2h^2\big\},$$
$$\sup_{k\geqslant 0} \sqrt{\mathbb{E}\big|\boldsymbol{V}_{k+1} - \bar{\boldsymbol{V}}_{k+1}\big|^4} \leqslant \frac{Cdh^2}{m} \min\big\{1, N^2h^2\big\},$$

where c_0 *depends only on* $(M_i)_{i=1}^2$, *and* C *depends only on* $(M_i)_{i=1}^3$.

Proof The SAGA stochastic gradient deviation at step k, denoted by $\boldsymbol{g}_k(\theta_k)$, is defined as

$$\boldsymbol{g}_k(\theta_k) = \nabla U_{\theta_k}(\boldsymbol{Y}_k) - \nabla U_{\theta_k}(\boldsymbol{\Phi}_{\theta_k}) + \frac{1}{N}\sum_{i=1}^N \nabla U_i(\boldsymbol{\Phi}_i) - \frac{1}{N}\sum_{i=1}^N \nabla U_i(\boldsymbol{Y}_k). \tag{113}$$

As established in (84), the stochastic gradient error $\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}$ is directly proportional to this deviation term $\boldsymbol{g}_k(\theta_k)$. Therefore, bounding the moments of $\boldsymbol{g}_k(\theta_k)$ is sufficient to control the stochastic gradient error. The specific target bound required is

$$\sqrt{\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^4} \leqslant \frac{Cd}{m} \min\big\{1, N^2h^2\big\}. \tag{114}$$

We establish the required bound (114) by deriving two separate estimates for $\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^4$ and then taking the minimum.

First, a uniform bound independent of h can be obtained using an argument similar to the one leading to (104). We start with the triangle inequality (103):

$$|\boldsymbol{g}_k(\theta_k)| \leqslant \left|\nabla U_{\theta_k}(\boldsymbol{Y}_k) - \frac{1}{N}\sum_{i=1}^N \nabla U_i(\boldsymbol{Y}_k)\right| + |\nabla U_{\theta_k}(\boldsymbol{\Phi}_{\theta_k})| + \frac{1}{N}\sum_{i=1}^N |\nabla U_i(\boldsymbol{\Phi}_i)|.$$

Using Assumption (SgN^+) and Theorem 30 (since $\boldsymbol{\Phi}_i = \boldsymbol{Y}_{k_i}$ are previous states), we obtain

$$\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^4 \leqslant C\bigg(d^2 + \frac{1}{N}\sum_{i=1}^N \mathbb{E}\big(|\boldsymbol{\Phi}_i|^2 + d\big)^2\bigg) \leqslant C\bigg(d^2 + \frac{1}{N}\sum_{i=1}^N \mathbb{E}|\boldsymbol{Y}_{k_i}|^4\bigg).$$

The historical index k_i takes values in $\{0, \cdots, k-1\}$, and the events $\{k_i = l\}_{l=0}^{k-1}$ partition the sample space, thus

$$\begin{aligned}\mathbb{E}\big|\boldsymbol{Y}_{k_i}\big|^4 &= \sum_{l=0}^{k-1} \mathbb{E}\big(|\boldsymbol{Y}_l|^4\, \mathbf{1}_{\{k_i=l\}}\big) = \sum_{l=0}^{k-1} \mathbb{E}|\boldsymbol{Y}_l|^4\, \mathbb{P}(k_i = l) \\ &\leqslant \Big(\max_{0\leqslant l\leqslant k-1} \mathbb{E}|\boldsymbol{Y}_l|^4\Big) \sum_{l=0}^{k-1} \mathbb{P}(k_i = l) = \max_{0\leqslant l\leqslant k-1} \mathbb{E}|\boldsymbol{Y}_l|^4.\end{aligned}$$

The factorization uses the independence of $\boldsymbol{Y}_l$ and $\mathbf{1}_{\{k_i=l\}}$: by (98),

$$\{k_i = l\} = \{\theta_l = i\} \cap \bigcap_{m=l+1}^{k-1} \{\theta_m \neq i\},$$

so $\mathbf{1}_{\{k_i=l\}}$ depends only on $\theta_l, \cdots, \theta_{k-1}$, whereas Algorithm 4 evaluates $\boldsymbol{Y}_l$ before θ_l is drawn. The random indices are independent of each other and of the Brownian motion.

Theorem 30 controls that maximum: since $\mathcal{V}(\boldsymbol{z}) = (\boldsymbol{z}^\top \boldsymbol{S}\boldsymbol{z})^4$ with $\boldsymbol{S}$ positive definite, we have $\mathbb{E}|\boldsymbol{Y}_l|^8 \leqslant C\,\mathbb{E}[\mathcal{V}(\boldsymbol{Z}_l)] \leqslant \frac{Cd^4}{m^4}$ uniformly in l, and Jensen's inequality gives $\mathbb{E}|\boldsymbol{Y}_l|^4 \leqslant \left(\mathbb{E}|\boldsymbol{Y}_l|^8\right)^{\frac{1}{2}} \leqslant \frac{Cd^2}{m^2}$. Therefore

$$\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^4 \leqslant C\left(d^2 + \frac{d^2}{m^2}\right) \leqslant \frac{Cd^2}{m^2},$$

providing the first part of the minimum bound

$$\sqrt{\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^4} \leqslant \frac{Cd}{m}. \tag{115}$$

Starting from the definition of $\boldsymbol{g}_k(\theta_k)$, we apply the triangle inequality to obtain

$$\begin{aligned}
|\boldsymbol{g}_k(\theta_k)| &\leqslant \left|\nabla U_{\theta_k}(\boldsymbol{Y}_k) - \nabla U_{\theta_k}(\boldsymbol{\Phi}_{\theta_k})\right| + \frac{1}{N}\sum_{i=1}^{N}\left|\nabla U_i(\boldsymbol{Y}_k) - \nabla U_i(\boldsymbol{\Phi}_i)\right| \\
&\leqslant C|\boldsymbol{Y}_k - \boldsymbol{\Phi}_{\theta_k}| + \frac{C}{N}\sum_{i=1}^{N}|\boldsymbol{Y}_k - \boldsymbol{\Phi}_i|.
\end{aligned}$$

Taking the fourth moment and the full expectation over the trajectory and θ_k gives

$$\begin{aligned}
\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^4 &\leqslant C\,\mathbb{E}|\boldsymbol{Y}_k - \boldsymbol{\Phi}_{\theta_k}|^4 + C\,\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}|\boldsymbol{Y}_k - \boldsymbol{\Phi}_i|\right)^4\right] \\
&\leqslant C\,\mathbb{E}|\boldsymbol{Y}_k - \boldsymbol{\Phi}_{\theta_k}|^4 + \frac{C}{N}\sum_{i=1}^{N}\mathbb{E}|\boldsymbol{Y}_k - \boldsymbol{\Phi}_i|^4 = \frac{C}{N}\sum_{i=1}^{N}\mathbb{E}|\boldsymbol{Y}_k - \boldsymbol{\Phi}_i|^4.
\end{aligned}$$

Applying the bound from Lemma 32 yields the h-dependent estimate:

$$\sqrt{\mathbb{E}|\boldsymbol{g}_k(\theta_k)|^4} \leqslant \frac{CN^2dh^2}{m}. \tag{116}$$

Combining the uniform bound implied by (115) and the h-dependent bound from (116), we obtain the desired composite bound (114). ∎

F.3 MSE of SAGA-UBU

The MSE bound for SAGA-UBU stated in Theorem 13 is derived using the same analytical framework employed for SG-UBU and SVRG-UBU. The proof combines the uniform-in-time moment bound from Theorem 30 and the local error estimates from Lemma 33.

Proof of Theorem 13 For SAGA-UBU, we note that the stochastic gradient error bound from Lemma 33 exhibits a stronger dependence on the parameter m. Therefore, the contribution of this error to the total MSE must be bounded carefully.

Following the proof strategy for SG-UBU (Theorem 7), we decompose the stochastic gradient error term R_k as:

$$
\begin{aligned}
R_{k,1} &= \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \nabla\phi_h(\bar{\boldsymbol{Z}}_{k+1}),\\
R_{k,2} &= \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \int_0^1 \Big(\nabla\phi_h\big(\lambda\boldsymbol{Z}_{k+1} + (1-\lambda)\bar{\boldsymbol{Z}}_{k+1}\big) - \nabla\phi_h(\bar{\boldsymbol{Z}}_{k+1})\Big)\mathrm{d}\lambda.
\end{aligned}
$$

This decomposition leads to the following moment bounds:

$$
\begin{aligned}
\mathbb{E}[R_{k,1}^2] &\leqslant \frac{C}{m^2h^2}\mathbb{E}|\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}|^2 \leqslant \frac{Cd}{m^3}\min\big\{1, N^2h^2\big\} \leqslant \frac{Cd}{m^3},\\
\mathbb{E}[R_{k,2}^2] &\leqslant \frac{C}{m^4h^2}\mathbb{E}|\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}|^4 \leqslant \frac{Cd^2h^2}{m^6}\min\big\{1, N^4h^4\big\}.
\end{aligned}
$$

Then the contribution of the stochastic gradient error $\sum_{k=0}^{K-1} R_k$ to the MSE is bounded by

$$
\begin{aligned}
\frac{1}{K^2}\mathbb{E}\left[\Big(\sum_{k=0}^{K-1} R_k\Big)^2\right] &\leqslant \frac{C}{K^2}\sum_{k=0}^{K-1}\mathbb{E}[R_{k,1}^2] + \frac{C}{K}\sum_{k=0}^{K-1}\mathbb{E}[R_{k,2}^2]\\
&\leqslant C\bigg(\frac{d}{m^3K} + \frac{d^2h^2}{m^6}\min\big\{1, N^4h^4\big\}\bigg).
\end{aligned}
$$

The analysis of the discretization error and martingale components remains unchanged. Therefore, the total MSE is bounded by

$$
\mathbb{E}\left[\bigg(\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f)\bigg)^2\right] \leqslant C\bigg(\frac{d}{m^3Kh} + \frac{d^2h^2}{m^6}\min\big\{1, N^4h^4\big\} + \frac{d^2h^4}{m^5}\Big(1 + \frac{dh}{m^2}\Big)\bigg),
$$

completing the proof of Theorem 13. ■

F.4 Numerical bias of SAGA-UBU

Proof of Theorem 14 We begin with the time average error decomposition in (18):

$$
\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k) - \pi(f) = \frac{\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_K)}{Kh} + \frac{1}{K}\sum_{k=0}^{K-1}(R_k + S_k + T_k),
$$

where R_k, S_k and T_k represent the stochastic gradient error, discretization error, and martingale difference term, respectively. To analyze the numerical bias, we take the expectation of this expression. Since the martingale term satisfies $\mathbb{E}[T_k] = 0$, the expectation becomes

$$\mathbb{E}\Big[\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k)\Big] - \pi(f) = \frac{\mathbb{E}[\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_K)]}{Kh} + \frac{1}{K}\sum_{k=0}^{K-1}\big(\mathbb{E}[R_k] + \mathbb{E}[S_k]\big). \tag{117}$$

The subsequent analysis focuses on bounding the expectations of the stochastic error term $\mathbb{E}[R_k]$ and the discretization error term $\mathbb{E}[S_k]$.

For the stochastic gradient error R_k, we use its integral representation

$$R_k = \frac{\phi_h(\boldsymbol{Z}_{k+1}) - \phi_h(\bar{\boldsymbol{Z}}_{k+1})}{h} = \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \int_0^1 \nabla\phi_h\big(\lambda\boldsymbol{Z}_{k+1} + (1-\lambda)\bar{\boldsymbol{Z}}_{k+1}\big)\mathrm{d}\lambda.$$

Following the decomposition strategy used in the proof of Theorem 7, we split R_k into a martingale difference term R_k^1 and a higher-order remainder R_k^2:

$$\begin{aligned}
R_k^1 &= \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \nabla\phi_h(\bar{\boldsymbol{Z}}_{k+1}),\\
R_k^2 &= \frac{1}{h}(\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1})^\top \int_0^1 \Big(\nabla\phi_h\big(\lambda\boldsymbol{Z}_{k+1} + (1-\lambda)\bar{\boldsymbol{Z}}_{k+1}\big) - \nabla\phi_h(\bar{\boldsymbol{Z}}_{k+1})\Big)\mathrm{d}\lambda.
\end{aligned}$$

The martingale property ensures $\mathbb{E}[R_k^1] = 0$. For the remainder R_k^2, we apply the Hessian bound from Theorem 1 and the stochastic gradient error bounds from Lemma 33 to obtain

$$\mathbb{E}|R_k^2| \leqslant \frac{C}{m^2 h}\mathbb{E}\big|\boldsymbol{Z}_{k+1} - \bar{\boldsymbol{Z}}_{k+1}\big|^2 \leqslant \frac{Cdh}{m^3}\min\big\{1, N^2h^2\big\}.$$

Since $\mathbb{E}[R_k] = \mathbb{E}[R_k^2]$, the average expected stochastic gradient error is bounded by:

$$\Big|\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[R_k]\Big| \leqslant \frac{Cdh}{m^3}\min\big\{1, N^2h^2\big\}. \tag{118}$$

For the discretization error S_k, the estimate (75) established in the proof of Theorem 9 applies verbatim: it involves $\boldsymbol{Z}_k$ only through the uniform-in-time moment bounds, which Theorem 30 supplies at the same order for SAGA-UBU. Hence

$$\Big|\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[S_k]\Big| \leqslant \frac{Cdh^2}{m^3}. \tag{119}$$

Finally, we need to bound the boundary term $\mathbb{E}[\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_K)]/(Kh)$ in (117). Using the Lipschitz property of ϕ_h from Theorem 1 and an argument similar to Lemma 21, we can estimate the displacement moment:

$$\mathbb{E}\big|\phi_h(\boldsymbol{Z}_0) - \phi_h(\boldsymbol{Z}_K)\big| \leqslant \frac{C}{m}\mathbb{E}|\boldsymbol{Z}_0 - \boldsymbol{Z}_K| \leqslant C\sqrt{\frac{dKh}{m^3}}. \tag{120}$$

Substituting the bounds for the expected stochastic error (118), the discretization error (119), and the boundary term (120) into the expression (117) yields an estimate for the time average bias over K steps:

$$\left|\mathbb{E}\Big[\frac{1}{K}\sum_{k=0}^{K-1} f(\boldsymbol{X}_k)\Big] - \pi(f)\right| \leqslant C\left(\sqrt{\frac{d}{m^3 K h}} + \frac{dh}{m^3}\min\left\{1, N^2h^2\right\} + \frac{dh^2}{m^3}\right).$$

As established in Lemma 31, the time-averaged distribution $\tilde{\pi}_h^K$ converges weakly along a subsequence as $K \to \infty$ to the limit distribution $\tilde{\pi}_h$. Taking the subsequential limit $K \to \infty$, the boundary term vanishes, resulting in the final bound for the numerical bias:

$$|\tilde{\pi}_h(f) - \pi(f)| \leqslant C\left(\frac{dh}{m^3}\min\left\{1, N^2h^2\right\} + \frac{dh^2}{m^3}\right).$$

This bound exactly matches the result stated in Theorem 14. ■

Appendix G. Numerical calculation of leading-order coefficient

In this section, we present the computational method for the leading-order coefficient (25) in the numerical bias of SG-UBU. This calculation requires computing the quantity

$$C_0 = \mathbb{E}^{\pi}\Big[\big(b(\boldsymbol{x}_0,\theta) - \nabla U(\boldsymbol{x}_0)\big)^{\top}\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0,\boldsymbol{v}_0)\big(b(\boldsymbol{x}_0,\theta) - \nabla U(\boldsymbol{x}_0)\big)\Big].$$

We define the noise outer-product matrix $\boldsymbol{E}(\boldsymbol{x}_0) \in \mathbb{R}^{d\times d}$ as

$$\boldsymbol{E}(\boldsymbol{x}_0) = \mathbb{E}_{\theta}\Big[\big(b(\boldsymbol{x}_0,\theta) - \nabla U(\boldsymbol{x}_0)\big)\big(b(\boldsymbol{x}_0,\theta) - \nabla U(\boldsymbol{x}_0)\big)^{\top}\Big],$$

which allows C_0 to be written equivalently using the trace:

$$C_0 = \mathbb{E}^{\pi}\Big[\operatorname{Tr}\big(\boldsymbol{E}(\boldsymbol{x}_0)\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0,\boldsymbol{v}_0)\big)\Big].$$

Since the continuous Poisson solution $\phi_0(\boldsymbol{x},\boldsymbol{v})$ is independent of the specific stochastic gradient model, we can compute the Hessian matrix $\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x},\boldsymbol{v})$ and apply this result to both the Gaussian noise and finite-sum cases.

Using the Kolmogorov solution $u(\boldsymbol{x},\boldsymbol{v},t)$ from (11), the Hessian matrix $\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x},\boldsymbol{v})$ is

$$\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0,\boldsymbol{v}_0) = \int_0^{\infty}\mathbb{E}^{(\boldsymbol{x}_0,\boldsymbol{v}_0)}\Big[(D_{\boldsymbol{vv}}\boldsymbol{x}_t)^{\top}\nabla f(\boldsymbol{x}_t) + (D_{\boldsymbol{v}}\boldsymbol{x}_t)^{\top}\nabla^2 f(\boldsymbol{x}_t)D_{\boldsymbol{v}}\boldsymbol{x}_t\Big],$$

where $(\boldsymbol{x}_t,\boldsymbol{v}_t)_{t\geqslant 0}$ solves the exact underdamped Langevin dynamics (6) with the initial state $(\boldsymbol{x}_0,\boldsymbol{v}_0)$, and $D_{\boldsymbol{v}}\boldsymbol{x}_t$ and $D_{\boldsymbol{vv}}\boldsymbol{x}_t$ are the first and second variation processes, respectively.

Here, the second variation process can be derived by taking derivatives with respect to $\boldsymbol{v}$ in the first variation process (36), and it is governed by the coupled ODE system:

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}t}D_{\boldsymbol{vv}}\boldsymbol{x}_t = D_{\boldsymbol{vv}}\boldsymbol{v}_t,\\ \dfrac{\mathrm{d}}{\mathrm{d}t}D_{\boldsymbol{vv}}\boldsymbol{v}_t = -\dfrac{1}{M_2}\Big(\nabla^3 U(\boldsymbol{x}_t)\langle D_{\boldsymbol{v}}\boldsymbol{x}_t, D_{\boldsymbol{v}}\boldsymbol{x}_t\rangle + \nabla^2 U(\boldsymbol{x}_t)D_{\boldsymbol{vv}}\boldsymbol{x}_t\Big) - 2D_{\boldsymbol{vv}}\boldsymbol{v}_t,\end{cases} \tag{121}$$

with the initial configuration $D_{\boldsymbol{vv}}\boldsymbol{x}_0 = D_{\boldsymbol{vv}}\boldsymbol{v}_0 = \boldsymbol{O}_{d\times d\times d}$, which are zero third-order tensors. The tensor quadratic form $\boldsymbol{B}\langle \boldsymbol{A}, \boldsymbol{A}\rangle \in \mathbb{R}^{d\times d\times d}$ is defined for $\boldsymbol{B} \in \mathbb{R}^{d\times d\times d}$ and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ by

$$\big(\boldsymbol{B}\langle \boldsymbol{A}, \boldsymbol{A}\rangle\big)_{kmn} := \sum_{i,j=1}^{d} B_{ijk}A_{im}A_{jn}, \qquad k, m, n = 1, \cdots, d.$$

For the discrete-time setting, let $(\boldsymbol{X}_k, \boldsymbol{V}_k)_{k=0}^{\infty}$ be the FG-UBU solution with step size $h > 0$ starting from $(\boldsymbol{x}_0, \boldsymbol{v}_0)$. Let $(D_{\boldsymbol{v}}\boldsymbol{X}_k)_{k=0}^{\infty}$ and $(D_{\boldsymbol{vv}}\boldsymbol{X}_k)_{k=0}^{\infty}$ be the corresponding first and second variation processes, both solved using an exponential integrator. The Hessian matrix $\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0, \boldsymbol{v}_0)$ can then be approximated by the sum

$$\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0, \boldsymbol{v}_0) \approx h\sum_{k=0}^{\infty} \mathbb{E}^{(\boldsymbol{x}_0,\boldsymbol{v}_0)}\Big[(D_{\boldsymbol{vv}}\boldsymbol{X}_k)^\top \nabla f(\boldsymbol{X}_k) + (D_{\boldsymbol{v}}\boldsymbol{X}_k)^\top \nabla^2 f(\boldsymbol{X}_k) D_{\boldsymbol{v}}\boldsymbol{X}_k\Big].$$

Note that the second variation $D_{\boldsymbol{vv}}\boldsymbol{x}_t$ is only guaranteed to decay exponentially almost surely when the potential function $U(\boldsymbol{x})$ is globally convex. For non-convex potentials, this computation of $\nabla^2_{\boldsymbol{vv}}\phi_0(\boldsymbol{x}_0, \boldsymbol{v}_0)$ may be unstable.