THE UNIVERSITY *of* EDINBURGH

Kinetic Langevin Monte Carlo Methods

Peter Archibald Whalley

Doctor of Philosophy

The University of Edinburgh

2024

To Clare and Richard

Abstract

In this thesis, we study discretizations of kinetic Langevin dynamics within the context of Markov chain Monte Carlo. We compare the convergence properties for different choices of integrators, we provide asymptotic bias estimates and numerics to compare them. We also present an alternative to Metropolis-Hastings which corrects for the bias. This thesis consists of two parts.

In the first part, we provide a framework to analyze the convergence of discretized kinetic Langevin dynamics for M-∇Lipschitz, m-convex potentials with and without stochastic gradients. Our approach gives convergence rates of $\mathcal{O}(m/M)$, with explicit stepsize restrictions, which are of the same order as the stability threshold for Gaussian targets and are valid for a large interval of the friction parameter. We apply this methodology to various integration schemes which are popular in the molecular dynamics and machine learning communities. Further, we introduce the property "γ-limit convergence" (GLC) to characterize underdamped Langevin schemes that converge to overdamped dynamics in the high-friction limit and which have stepsize restrictions that are independent of the friction parameter; we show that this property is not generic by exhibiting methods from both the class and its complement. We present numerical experiments for a Bayesian logistic regression example, where BAOAB is shown to perform the best. Finally, we provide asymptotic bias estimates for the BAOAB scheme, which remain accurate in the high-friction limit by comparison to a modified stochastic dynamics which preserves the invariant measure.

In the second part, we present an unbiased method for Bayesian posterior means based on kinetic Langevin dynamics that combines advanced splitting methods with enhanced gradient approximations. Our approach avoids Metropolis correction by coupling Markov chains at different discretization levels in a multilevel Monte Carlo approach. Theoretical analysis demonstrates that our proposed estimator is unbiased, attains finite variance, and satisfies a central limit theorem. It can achieve accuracy $\epsilon > 0$ for estimating expectations of Lipschitz functions in d dimensions with $\mathcal{O}(d^{1/4}\epsilon^{-2})$ expected gradient evaluations, without assuming warm start. We exhibit similar bounds using both approximate and stochastic gradients, and our method's computational cost is shown to scale independently of the size of the dataset. The proposed method is tested using a multinomial regression problem on the MNIST dataset and a Poisson regression model for soccer scores. Experiments indicate that the number of gradient evaluations per effective sample is independent of dimension, even when using inexact gradients. For product distributions, we give dimension-independent variance bounds. Our results demonstrate that the unbiased algorithm we present can be much more efficient than the "gold-standard" randomized Hamiltonian Monte Carlo.

Lay Summary

This thesis is about studying the properties and introducing methodology for efficient methods for generating samples from probability distributions and estimating expected quantities of that probability distribution. This is particularly important in molecular dynamics, where one would want to calculate expected configurations for example for protein configurations, and in statistical computations such as Bayesian inference where one would want to estimate expected parameters of the posterior.

Typically these methods have computational cost that scale with the number of parameters in the model and properties of the model. In this thesis, we study how the cost scales with these quantities, and introduce more efficient methods for estimating these quantities.

Acknowledgements

Firstly, I would like to thank Prof. Ben Leimkuhler, Dr. Daniel Paulin and Dr. Neil Chada who have been fantastic supervisors. I have benefitted immensely from their diverse expertise and interests, ranging from numerical analysis to probability theory. I cannot thank them enough for all the time they put in and all their guidance throughout the numerous projects we have worked on, as well as all the opportunities they have given me. Their influence has been pivotal in shaping me as an applied mathematician.

I would like to thank Prof. Weizhu Bao, Prof. Gabriel Stoltz and Dr. Gilles Vilmart for hosting me during my PhD. I would also like to thank my collaborators at Rothamsted during my summers there, where I gained my first experience in research. In particular, Dr. Mikhail Semenov and Dr. Xiaoxian Zhang for providing me with lots of interesting mathematical problems. I would also like to thank my office mates Aidan, Andrew, Kat, Martin and Mike for making my time in JCMB fun.

Lastly, I would like to thank my family. Paul and Hannah, thank you for all your love and support. I would like to thank my parents to whom I dedicate this thesis for encouraging me to pursue my passions and providing me with the best childhood I could have ever asked for. Without all their time, effort and enthusiasm for my studies, I doubt I would have made it this far.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Peter Archibald Whalley

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

1.1 Markov chain Monte Carlo

Efficient sampling of high dimensional probability distributions is required for applications such as Bayesian inference and molecular dynamics (see for example [76] and [21, 100]). Efficient sampling is a challenge in many fields including pharmacology, economics, physics, political science and machine learning [154].

Bayesian inference requires one to sample from some d-dimensional target measure π defined on $\mathbb{R}^d$, with Lebesgue density proportional to $\exp(-U(x))$, where U is the negative log-density of π. We then wish to approximate expectations of a function $f : \mathbb{R}^d \to \mathbb{R}$, i.e.

$$\pi(f) := \int_{\mathbb{R}^d} f(x) d\pi(x), \tag{1.1.1}$$

where this could be expected parameters of a machine learning model under a posterior distribution or this could be expected potential energy of a molecular dynamics model [100, 154].

A popular approach to estimate expectations and generate samples for high-dimensional Bayesian inference [133] from our target measure is Markov chain Monte Carlo (MCMC). They enable the computation of posterior means and variances and other observable averages by replacing ensemble calculations with Monte Carlo sums over discrete Markov processes (1.1.1). More precisely, we sample from a Markov chain $X_{i+1} \sim p(\cdot \mid X_i)$ with invariant measure π. Then we wish to construct a chain such that

$$\pi(f) \approx \frac{1}{K} \sum_{i=1}^{K} f(X_i), \tag{1.1.2}$$

and a central limit theorem holds

$$\sqrt{K} \left(\frac{1}{K} \sum_{i=1}^{K} f(X_i) - \pi(f) \right) \to \mathcal{N}(0, \sigma^2), \tag{1.1.3}$$

with asymptotic variance σ^2.

MCMC relies on the construction of a Markov chain with the correct target measure or an approximation of the target measure as its invariant measure. A typical approach to constructing such a chain is finding a continuous time stochastic process whose unique invariant measure is the given measure. Then one can either simulate this process exactly or, more commonly, discretize the dynamics. Some

popular MCMC methods which rely on discretizing continuous dynamics which preserve the target measure are: overdamped Langevin dynamics [15, 136, 137], (randomized) Hamiltonian Monte Carlo [29, 62], kinetic/underdamped Langevin dynamics [34, 47, 55], adaptive Langevin dynamics [88] and generalized Langevin dynamics [2, 104].

There are also some recent approaches that rely on exact simulation of the continuous dynamics, which include the bouncy particle sampler [30, 125], the zig-zag process [18], and the Boomerang sampler [19]. However, these methods can also be discretized for use in MCMC procedures [13, 14].

A limitation to the broader uptake of Bayesian inference is the scaling of the computational cost of MCMC algorithms with model dimension and dataset size. Typical MCMC methods (Metropolis adjusted Langevin algorithm [16, 134], Hamiltonian Monte Carlo [62, 119]) employ Metropolis-Hastings correction steps to ensure convergence to the desired invariant distribution. In order to maintain a high acceptance rate, stepsizes must decrease as a function of the model dimension, which implies that convergence rates are also dependent on dimension [17, 45, 135]. This is due to the fact that the bias in the invariant measure due to discretization scales with dimension. To control the bias one typically needs to scale the stepsize with dimension. By contrast, optimization methods typically have convergence rates that are independent of the dimension and can make use of stochastic gradients based on a subset of the data instead of the entire dataset [87]. For these reasons, optimization algorithms are much more scalable than sampling methods, so practitioners often prefer machine-learning approaches. The relative inefficiency of sampling compared to optimization also limits the uptake of uncertainty quantification techniques (typically built on a Bayesian foundation) in high-dimensional machine learning applications. In the next sections, we will provide an informal overview of the relevant topics and literature to be made precise in the following chapters where relevant.

1.2 Overdamped Langevin dynamics

The first mentioned approach is the overdamped Langevin dynamics stochastic differential equation (SDE):

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dW_t, \tag{1.2.4}$$

where $t \geq 0$, $X_t \in \mathbb{R}^d$, U is a "potential energy" function and may be taken to be the negative log-density for a suitable target measure π and W_t is the driving d-dimensional Brownian motion. It is known that (1.2.4) has a unique strong solution under mild conditions on ∇U, for example, Lipschitz continuity (see [95, Sec. 5.2]). It can also be shown under mild assumptions on the potential U that the invariant measure of this process π has density proportional to $\exp(-U(x))$ [124]. Under the assumption of a Poincaré inequality, convergence rate guarantees can be established for the continuous dynamics [10].

Typical discretizations of these dynamics are the Euler-Maruyama discretization and the high-friction limit of the popular kinetic Langevin dynamics scheme BAOAB [98]. The simplest discretization of overdamped Langevin dynamics is using the Euler-Maruyama (EM) method which for a stepsize $h > 0$ is defined by the update rule

$$x_{n+1} = x_n - h\nabla U(x_n) + \sqrt{2h}\xi_{n+1}, \tag{1.2.5}$$

for some initial condition $x_0 \in \mathbb{R}^d$ where $(\xi_n)_{n\in\mathbb{N}}$ are independent d-dimensional standard Gaussian random variables, that is $\xi_n \sim \mathcal{N}(0_d, I_d)$ for all $n \in \mathbb{N}$, where I_d is the d-dimensional identity matrix.

An alternative method is the BAOAB limit method of Leimkuhler and Matthews (LM)([98], [102]) which is defined by the update rule

$$x_{n+1} = x_n - h\nabla U(x_n) + \sqrt{2h}\frac{\xi_{n+1} + \xi_n}{2}. \tag{1.2.6}$$

Under mild Lyapunov drift conditions, it can be shown that (1.2.4) and (1.2.6) have a unique invariant measure (see [63, 65]).

Example 1.2.1. *If we consider the discretizations for overdamped Langevin dynamics and implement them, then we can illustrate using histograms convergence of the invariant measure. In Figures 1.1 and 1.2 we can see the histograms of the points along the sample path converging towards the target density. Note that the Leimkuhler-Matthews method is exact for Gaussian targets and hence we can see no bias in the invariant measure.*

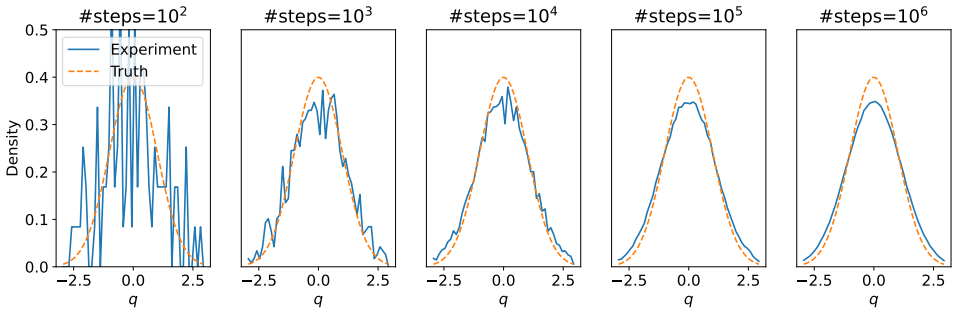


Figure 1.1: $h = 0.5$, one-dimensional standard Gaussian target using the Euler-Maruyama discretization (1.2.5).

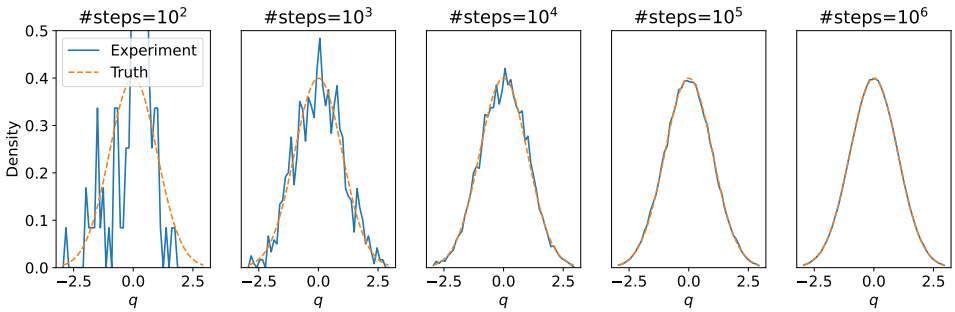


Figure 1.2: $h = 0.5$, one-dimensional standard Gaussian target using the Leimkuhler-Matthews discretization (1.2.6).

In Figures 1.3 and 1.4 we can see far more bias compared to the Gaussian case, with a smaller stepsize. There is also bias in the invariant measure for the Leimkuhler-Matthews method as it is only exact for Gaussian targets. However, the Leimkuhler Matthews method has significantly less bias than the Euler-Maruyama method.

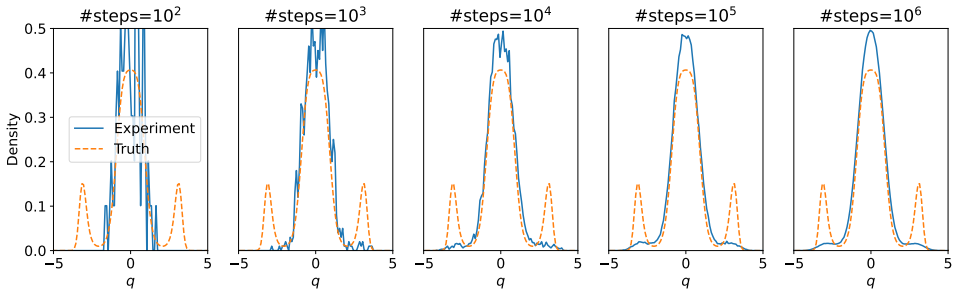


Figure 1.3: $h = 0.1$, one-dimensional multimodal distribution using the Euler-Maruyama discretization (1.2.5).

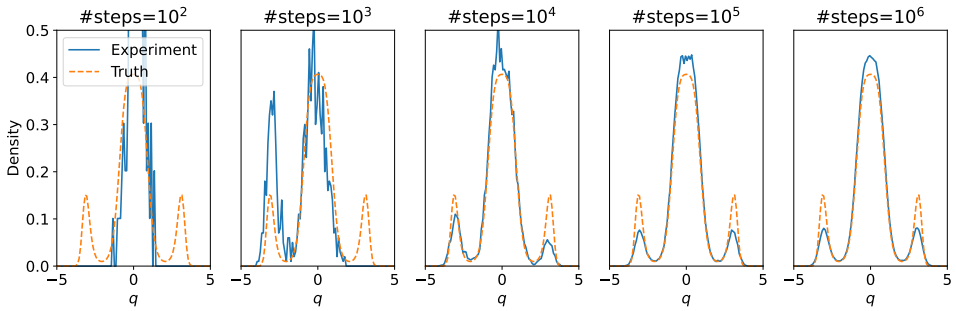


Figure 1.4: $h = 0.1$, one-dimensional multimodal distribution using the Leimkuhler-Matthews discretization (1.2.6).

For MCMC methods like Euler-Maruyama and Leimkuhler-Matthews it is important to quantify the bias in the invariant measure as well as the convergence rate towards the invariant measure, illustrated in the preceding figures. We can quantify these in terms of parameters such as dimension, parameters from assumptions on the potential, stepsize and any parameters which define the dynamics. The preceding figures illustrate the importance of the choice of your numerical integrator depending on the application, for example, the Leimkuhler-Matthews method has a much smaller bias in the invariant measure and there is no clear difference in convergence rate.

1.3 Kinetic Langevin dynamics

Another choice of stochastic dynamics and the main focus of this thesis is the kinetic Langevin dynamics which is the stochastic differential equation system defined by

$$\begin{aligned} dX_t &= V_t dt, \\ dV_t &= -\nabla U(X_t)dt - \gamma V_t dt + \sqrt{2\gamma} dW_t, \end{aligned} \tag{1.3.7}$$

where $X_t, V_t \in \mathbb{R}^d$, U is a "potential energy" function and may also be taken to be negative log-density for a suitable target measure π, $\gamma > 0$ is a friction parameter and W_t is the driving d-dimensional Brownian motion. Similarly to overdamped Langevin dynamics, one can show under mild conditions, for example, Lipschitz continuity of ∇U that (1.3.7) admits a unique strong solution [128, Sec. 9.2]. It can be shown under mild assumptions on the potential U that the invariant measure of this process is proportional to $\exp\left(-U(x) - \frac{1}{2}\|v\|^2\right)$ [124]. Normally, Langevin dynamics is developed in the physical setting with additional parameters representing temperature and mass. However, our primary aim in using (1.3.7) is, ultimately, the computation of statistical averages involving only the position X, and in such situations, both parameters can be neglected without any loss of generality.

Taking the limit as $\gamma \to \infty$ in (1.3.7), and introducing a suitable time-rescaling ($t' = \gamma t$) results in the overdamped Langevin dynamics given in (1.2.4)(see [Sec. 6.5][124]). In the case of kinetic Langevin dynamics, a more delicate argument is needed to establish exponential convergence than in the overdamped case, due to the hypoelliptic nature of the SDE (see [9, 11, 12, 38, 59, 60, 70, 141, 158]).

Due to the additional complexity of the kinetic Langevin dynamics compared to the overdamped dynamics, there have been far more integrators that have been proposed to discretize these dynamics. Many of these perform much better than the Euler-Maruyama discretization of the dynamics. A class of these rely on splitting methods [110, 144, 145, 149, 153], for example, a splitting method proposed in [98] is based on the following decomposition

$$\begin{pmatrix} dx \\ dv \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ -\nabla U(x)dt \end{pmatrix}}_{\mathcal{B}} + \underbrace{\begin{pmatrix} vdt \\ 0 \end{pmatrix}}_{\mathcal{A}} + \underbrace{\begin{pmatrix} 0 \\ -\gamma v dt + \sqrt{2\gamma} dW_t \end{pmatrix}}_{\mathcal{O}},$$

where the $\mathcal{B}$, $\mathcal{A}$ and $\mathcal{O}$ parts can be integrated exactly. Splitting methods (discretizations) can be made by composing different orders of these parts, each of which is integrated over a time interval of size $h > 0$. However, if a part is repeated, then one divides the timestep by the number of repeats, for example in the BAOAB integrator B and A are repeated twice so you integrate each of the parts over a time interval of size $h/2$, so that the total integration time of the B parts is of size h. Different combinations of these parts include the popular integrators BAOAB, OBABO and OABAO, noting that if the letters form a palindromic word, then this results in a method which has $\mathcal{O}(h^2)$ bias with respect to the stepsize in the invariant measure with respect to stepsize $h > 0$ [100].

Another splitting first considered in [148] and also studied in [3, 140] it requires only one gradient evaluation per step, yet has strong order two pathwise accuracy. It is based on splitting the SDE (1.3.7) into the following components

$$\begin{pmatrix} dx \\ dv \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ -\nabla U(x)dt \end{pmatrix}}_{\mathcal{B}} + \underbrace{\begin{pmatrix} vdt \\ -\gamma v dt + \sqrt{2\gamma} dW_t \end{pmatrix}}_{\mathcal{U}},$$

each of which can be integrated in the weak sense exactly over an interval of size $h > 0$. Similar to the overdamped case, under mild Lyapunov drift conditions, it can be shown that discretizations of kinetic Langevin dynamics have a unique invariant measure and geometric convergence towards the invariant measure (see [63]).

1.4 Non-asymptotic guarantees

The focus of this thesis is on Langevin dynamics in its kinetic and overdamped forms and using their discretizations for MCMC. The discretization with a step size $h > 0$ will have an invariant measure π_h on $\mathbb{R}^{2d}$. However, there is an inherent bias due to the finite difference numerical approximation ($\pi \neq \pi_h$) (this is illustrated in Example 1.2.1). This bias is usually addressed by choosing a sufficiently small stepsize, or by adding bias correction by methods like Metropolis-Hastings adjustment or unbiased estimation techniques. The choice of the discretization method has a significant effect on the quality of the samples and also on the computational cost of producing accurate samples, through stability properties, convergence rates, and asymptotic bias.

A metric that is typically used to quantify the performance of a sampling scheme is the number of steps required to reach a certain level of accuracy in Wasserstein distance. Non-asymptotic bounds in Wasserstein distance reflect computational complexity, convergence rate and accuracy. Achieving such bounds relies on two steps:

(1) determining explicit convergence rates of the process to its invariant measure

and

(2) proving asymptotic bias estimates for the invariant measure,

which are two important factors to take into consideration in algorithm design for MCMC and one wishes to optimise model parameters to maximise convergence rate and minimise discretization bias. Each of these quantities are illustrated in Example 1.2.1.

Kinetic Langevin dynamics has been shown to converge faster than its overdamped counterpart [38], it also allows for discretizations with higher orders of accuracy, for example, the UBU, BAOAB and OBABO schemes have a second-order bias in the invariant measure [105, 115, 140] compared to the Euler-Maruyama scheme for overdamped Langevin, which has first-order bias [65]. As a result, it has improved non-asymptotic guarantees [41, 47, 55, 105, 116, 140, 142]. It has also been the preferred method for sampling in molecular dynamics simulations for many years [100], due to its improved sampling performance in practice over its overdamped counterpart.

1.4.1 Wasserstein distance

Let $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a metric on $\mathbb{R}^d$, $\mathcal{P}(\mathbb{R}^d)$ be the set of all probability measures on $\mathbb{R}^d$ and $\Gamma(\mu, \nu)$ be the set of all couplings between $\mu, \nu \in \mathbb{R}^d$, i.e. the set of all probability measures in $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with first marginal μ and second marginal ν. Then the p-Wasserstein distance with respect to ρ is defined by

$$\mathcal{W}_{p,\rho}(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \rho(x, y)^p \gamma(dxdy) \right)^{1/p},$$

for all $\mu, \nu \in \mathcal{P}_{p,\rho}(\mathbb{R}^d)$, the set of all measures with finite p-th moment with respect to ρ. If ρ is omitted from $\mathcal{W}_{p,\rho}(\mu,\nu)$, i.e. $\mathcal{W}_p(\mu,\nu)$, then ρ is the Euclidean metric.

Considering the Markov kernel $(P_t)_{t\geq 0}$ of the continuous dynamics (1.2.4) or (1.3.7), or the transition kernel P_h of one step of a discretization with a stepsize $h > 0$, if one can design a coupling between two realizations of the continuous or discrete dynamics (for example shared Brownian motion or Gaussian increments) such that the expected distance between the realizations with respect to ρ converges to zero, then we have Wasserstein contraction. More precisely, for initial measures $\mu, \nu \in \mathcal{P}_{p,\rho}(\mathbb{R}^d)$ and a continuous Markov kernel $(P_t)_{t\geq 0}$ we define p-Wasserstein contraction with respect to ρ by

$$\mathcal{W}_{p,\rho}(\mu P_t, \nu P_t) \leq e^{-ct}\mathcal{W}_{p,\rho}(\mu,\nu),$$

with a rate $c > 0$ [58]. Equivalently for the transition kernel P_h of one step of a discretization with a stepsize $h > 0$ we define p-Wasserstein contraction with respect to ρ by

$$\mathcal{W}_{p,\rho}(\mu P_h^k, \nu P_h^k) \leq (1-ch)^k \mathcal{W}_{p,\rho}(\mu,\nu).$$

A direct consequence of Wasserstein contraction is the existence of a unique invariant measure and convergence towards it in Wasserstein distance. This can be proved by Banach's fixed point theorem [159, Corollary 5.22, Theorem 6.18].

Global strong order

In this thesis, we will be interested in the global strong order, that is pathwise accuracy estimates which are independent of time.

Definition 1.4.1. *Consider a discretization for* (1.2.4) *or* (1.3.7), $(X_t)_{t\geq 0}$ *to be the solution to the continuous dynamics and suppose they share Brownian motion. Then we say that a discretization method is globally strong order p, if there exists $h_0 > 0$ such that for all $h < h_0$ there exists $C > 0$ independent of k and h such that*

$$\left(\mathbb{E}\left|x_k - X_{kh}\right|^2\right)^{1/2} \leq Ch^p,$$

for all $k \in \mathbb{N}$, where $(x_k)_{k\in\mathbb{N}}$ is a discretization with stepsize $h > 0$ such that $x_0 = X_0$.

Remark 1.4.2. *Typically, strong order estimates are defined at a finite time [114], however, when we have convergence of the discretization we are able to get uniform-in-time strong order estimates and we define the global strong order in this way [140].*

Asymptotic bias

We will also be interested in asymptotic bias in the invariant measure, which is a weak order of accuracy. We will define this in terms of the Wasserstein two distance.

Definition 1.4.3. *Consider a discretization for* (1.2.4) *or* (1.3.7)*. Let* π_h *be the invariant measure of the discretization method with stepsize* $h > 0$*. Then we say that the discretization method has an asymptotic bias of order* q *in Wasserstein distance if there exists* $h_0 > 0$ *such that for all* $h < h_0$ *there exists* $C > 0$ *independent of* h *such that*

$$\mathcal{W}_2(\pi, \pi_h) \leq Ch^q.$$

Remark 1.4.4. *Typically in the literature (see [114]), the weak error of order* q *is defined by*

$$|\pi(f) - \pi_h(f)| \leq Ch^q$$

for smooth functions f *and the constant* $C > 0$ *depends on* f*. However, we will consider the stronger notion of weak error in Wasserstein distance.*

We remark that if a method is global strong order p, then it is at least order p in the asymptotic bias according to Definitions 1.4.1 and 1.4.3. We provide the strong order and asymptotic bias orders of some methods for kinetic Langevin dynamics in Table 1.1.

Algorithm	Strong Order	Asymptotic Bias Order	Reference
Euler-Maruyama	1	1	[142]
BAOAB	1	2	[105]
OBABO	1	2	[115]
UBU	2	2	[140]

Table 1.1: Global strong and asymptotic bias order for some kinetic Langevin dynamics methods.

For the objective of sampling from a measure π, if we use a discretization with transition kernel P_h with invariant measure π_h and initial measure μ_0, then we can consider the bias and convergence rate together. It is natural to consider the quantity

$$\begin{aligned}\mathcal{W}_2(\pi, \mu_0 P_h^k) &\leq \mathcal{W}_2(\pi_h, \mu_0 P_h^k) + \mathcal{W}_2(\pi, \pi_h) \\ &\leq \underbrace{(1 - ch)^k \mathcal{W}_2(\pi_h, \mu_0)}_{\text{convergence rate}} + \underbrace{\mathcal{W}_2(\pi, \pi_h)}_{\text{asymptotic bias}},\end{aligned}$$

then one wishes to minimise the number of steps k in terms of parameters in the discretization to achieve an accuracy $\epsilon > 0$ [67, 115, 140].

1.5 Metropolized and unbiased methods

Due to the fact that the discretized stochastic dynamics do not converge exactly to the correct target distribution, one often uses a Metropolis-Hasting accept/reject step to create Markov chains with the correct invariant measure. For example, the Metropolized version of the Euler-Maruyama scheme for overdamped Langevin dynamics (1.2.5) is the popular Metropolis-adjusted Langevin algorithm (MALA) (see [16, 134]). The MALA algorithm views the Euler scheme for overdamped Langevin as a proposal in a Metropolis-Hastings algorithm [112] which leaves the target measure invariant.

Other examples of Metropolis-adjusted discretizations for sampling include (randomized) Hamiltonian Monte Carlo (HMC) (see [29, 62, 119]) and generalized Hamiltonian Monte Carlo (GHMC) [80], which in a special case is a Metropolis-adjusted version of the OBABO algorithm [115]. We also introduce a (randomized) generalized Hamiltonian Monte Carlo in Algorithm 1, which we found to perform the best in practice as a comparison to the methods we introduce in this thesis. We introduce this in more detail in the following section.

1.5.1 Generalized Hamiltonian Monte Carlo

(Generalized) Hamiltonian Monte Carlo methods are based on Hamiltonian dynamics for the Hamiltonian $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by $H(x, v) = U(x) + \frac{1}{2}\|v\|^2$. More precisely the continuous solution to

$$\begin{aligned} dX_t &= V_t dt, \\ dV_t &= -\nabla U(X_t) dt, \end{aligned} \tag{1.5.8}$$

where $X_t, V_t \in \mathbb{R}^d$, and U is a "potential energy" function and can be taken to be the negative log-density of a suitable target measure. Then under mild assumptions, this has a unique invariant measure with density proportional to $\exp\left(-U(X) - \frac{1}{2}\|V\|^2\right)$. If we define $\Psi_H(x, v, t)$ to be the solution to the continuous dynamics (1.5.8) with initial condition $(x, v) \in \mathbb{R}^{2d}$ at time $T > 0$. Then for $T > 0$, we define the transition kernel of the Markov chain for GHMC by the update rule

$$(X_{k+1}, V_{k+1}) := \Psi_H(X_k, \alpha V_k + \sqrt{1 - \alpha^2}\xi_{k+1}),$$

where $\alpha \in [0, 1)$ is a partial velocity refreshment parameter, ξ_k are distributed according to $\mathcal{N}(0_d, I_d)$ for all $k \in \mathbb{N}$. The term generalized is used for the presence of the parameter α (see [80]), in traditional HMC $\alpha = 0$ (see [119]), however using partial velocity refreshment can lead to faster convergence. Additionally, the integration time does not need to be deterministic at each step, for example, $T \sim \text{Geom}\,(1/\lambda)$ for some rate $\lambda > 0$ in randomized Hamiltonian Monte Carlo, which has many advantages in terms of ergodicity and preventing periodic behaviour [29].

Except in special cases the Hamiltonian dynamics (1.5.8) cannot be solved exactly. Instead, a velocity Verlet integrator is used (or BAB in the splitting methods introduced for (1.3.7)). This is given by

$$\begin{aligned} v &\to v - \frac{h}{2}\nabla U(x), \\ x &\to x + hv, \\ v &\to v - \frac{h}{2}\nabla U(x), \end{aligned} \tag{1.5.9}$$

to integrate one step of size $h > 0$, then one performs $L = T/h \in \mathbb{N}$ steps. Then one can perform a full or partial velocity refreshment. When the dynamics are discretized instead of simulated exactly, there is an associated integration error which results in bias in the invariant measure. This is typically removed by using the dynamics as a proposal in a Metropolis-Hastings accept/reject step [112]. The acceptance ratio has a simplified form in terms of the difference between the Hamiltonian at the start and end of the Hamiltonian simulation governed by (1.5.9). In Algorithm 1 we detail precisely the algorithm for randomized Hamiltonian Monte Carlo with partial velocity refreshment, which, inspired by [36, 41, 80], we call generalized randomized Hamiltonian Monte Carlo (GRHMC). We also provide in Table 1.2 parameter selections for various popular methods in this general framework [80].

Algorithm 1 Randomized Hamiltonian Monte Carlo with Partial Refreshment (GRHMC)

Input:
- stepsize h.
- Initial distribution μ_0 on $\mathbb{R}^d \times \mathbb{R}^d$.
- Potential function $U : \mathbb{R}^d \to \mathbb{R}$ of target distribution.
- Number of samples parameter K.
- Expected number/Number of leapfrog steps parameter $E_L \geq 1$.
- Partial refreshment parameter α.

Initialise $(x_0, v_0) \sim \mu_0$.
for $i = 1, \ldots, K$ **do**
  Sample $L \sim \mathrm{Geom}(1/E_L)$ or set $L = E_L$.
  Perform L leapfrog steps.
  Set $(\tilde{x}_0, \tilde{v}_0) := (x_i, v_i)$.
  for $j = 0, \ldots, L-1$ **do**
    $\tilde{v}_{j+1/2} := \tilde{v}_j - \frac{h}{2}\nabla U(\tilde{x}_j)$
    $\tilde{x}_{j+1} := \tilde{x}_j + h\tilde{v}_{j+1/2}$
    $\tilde{v}_{j+1} := \tilde{v}_{j+1/2} - \frac{h}{2}\nabla U(\tilde{x}_{j+1})$
  end for
  Let $(x_i', v_i') = (\tilde{x}_L, \tilde{v}_L)$
  Compute Hamiltonian.
  $H(x_i, v_i) = U(x_i) + \frac{1}{2}\|v_i\|^2$, $H(x_i', v_i') = U(x_i') + \frac{1}{2}\|v_i'\|^2$.
  Perform Metropolis-Hastings accept/reject step.
  With probability $\min\left[1, \exp(H(x_i, v_i) - H(x_i', v_i'))\right]$, set $(x_{i+1}, v_{i+1}) = (x_i', v_i')$ (accept proposal).
  Otherwise, set $(x_{i+1}, v_{i+1}) = (x_i, -v_i)$ (reject proposal).
  Partial velocity refreshment.
  Sample $Z \sim \mathcal{N}(0_d, I_d)$ and update $v_{i+1} \to \alpha v_{i+1} + (1-\alpha^2)^{1/2} Z$.
end for
Output:
Samples $(x_1, v_1), \ldots, (x_K, v_K)$.

A disadvantage of Metropolis-adjustment is that dimension-dependent step size restrictions are needed to keep the acceptance rate from collapsing to 0. In particular, for MALA, it is shown in [48] that one requires $h \sim d^{-1/2}$, for HMC it is shown that one requires $h \sim d^{-1/2}$ [97], and an improvement is shown for RHMC for a Gaussian target under a warm start assumption $h \sim d^{-1/4}$ [45]. In many

Algorithm	L	α	Reference
MALA	$L = 1$	$\alpha = 0$	[16, 134]
MAKLA	$L = 1$	$\alpha = \exp(-\gamma h) \in (0, 1)$	[27, 80]
GHMC	$L > 1$	$\alpha = [0, 1)$	[80]
GRHMC	$L \sim \text{Geom}(1/E_L)$, $E_L \geq 1$	$\alpha = [0, 1)$	[29, 40]

Table 1.2: Parameters in Algorithm 1 for different Metropolis-adjusted Hamiltonian & Langevin based algorithms

applications of interest, for example, machine learning [154] and molecular dynamics [100] the target measures are typically very high-dimensional and the dimension-dependent step size restrictions result in a very high computational cost, for example, the computational cost for MALA with $h \sim d^{-1/2}$ would be $\mathcal{O}(d^{1/2})$.

We also mention that, in the area of molecular simulation, unadjusted numerical discretizations of kinetic Langevin dynamics have been employed for sampling from complex distributions for many years [34, 85, 98, 100]. Even though such discretizations introduce bias, this is often dominated by the Monte Carlo error–even at substantially larger stepsizes than would typically be used in Metropolized calculations [100]. On the other hand, the magnitude of the sampling bias due to finite stepsize is problem-dependent and can be difficult to quantify; thus, there are situations where the ability to ameliorate the discretization bias is crucial. Some methods have also been proposed for reducing the discretization bias by decreasing stepsize asymptotically [65, 164]. However, such a procedure can slow convergence or introduce heuristic schedules into the sampling apparatus.

1.5.2 Unbiased estimation

Ultimately, one is interested in producing an unbiased estimator for (1.1.1), and alternative methods to Metropolis-adjustment have recently been introduced to de-bias the estimator produced from MCMC methods based on discretizations. [79, 129] introduced methods to remove bias from (1.1.1) based on providing a sequence of measures $(\pi_k)_{k\in\mathbb{N}}$, such that $\lim_{k\to\infty}(\pi_k) = \pi$ and then to create an estimator of the form

$$\pi(f) = \pi_0(f) + \sum_{k=0}^{\infty} \pi_{k+1}(f) - \pi_k(f), \tag{1.5.10}$$

where one can create an unbiased estimator of the form

$$\widehat{\pi}(f) = \sum_{k=0}^{L} \frac{\xi_k}{\mathbb{P}_L(L \geq k)} \qquad \text{or} \qquad \widehat{\pi}(f) = \frac{\xi_L}{\mathbb{P}_L(L)}, \tag{1.5.11}$$

such that

$$\begin{aligned} \mathbb{E}[\xi_0] &= \pi_0(f), \\ \mathbb{E}[\xi_l] &= \pi_l(f) - \pi_{l-1}(f) \quad l \in \{1, 2, \dots\}, \end{aligned}$$

where L is a random variable with probability mass function $\mathbb{P}_L$ on $\mathbb{N}$ that is independent of the sequence $\{\xi_k\}_{k\in\mathbb{N}}$. For this to be a computationally implementable algorithm we require that $\pi_{k+1}(f) - \pi_k(f) \to 0$ at a fast rate. The sequence of measures considered in [78] for multilevel Monte Carlo to approximate the invariant measure was to consider π_k in (1.5.10) (using a truncated sum up to a level $K \in \mathbb{N}$) to be the invariant measure of the Euler scheme for overdamped Langevin dynamics with stepsize $h_k = h_0 2^{-k}$ at time T_k, where $T_k \to \infty$, hence reducing the cost to reach a desired accuracy for sampling the invariant measure. However, this approach had not been extended to produce unbiased estimates under the invariant measure until the recent work [40], which is the focus of Chapter 2, where we make use of kinetic Langevin dynamics discretization to increase the rate at which $\pi_{k+1}(f) - \pi_k(f) \to 0$ decreases. In principle, the framework introduced in [40] can be applied to any discretization-based MCMC methods, for example, unadjusted Hamiltonian Monte Carlo [26, 62], overdamped Langevin dynamics or discretized PDMPs [13, 14].

Unbiased Monte Carlo methods have been widely studied in the recent literature; see [86, Sec. 2.1] for an overview. The goal of the methods [49, 79, 82, 86, 129] is to remove burn-in bias via couplings. [90] proposed an alternative method for eliminating burn-in bias by considering a burn-in period of random length. The cited papers above all require that the stationary distribution of the Markov chain has no bias (hence, these methods typically involve Metropolization) and are not able to remove discretization bias in the numerical integration of SDEs such as (1.3.7). [113] extended unbiased methods to intractable likelihoods, and [61] created unbiased estimators of MCMC asymptotic variances. Recently, [138] introduced a method which allows sampling from the invariant measure without the use of Metropolis-Hastings accept/reject contrary to other work in the area, which offers an exciting new alternative to Metropolis-adjustment. However, their implementation is complicated (requiring the coupling of 4 Markov chains) and they have not demonstrated improvement over state-of-the-art Metropolis-adjusted methods only their unbiased versions (eliminating the initialization bias).

1.6 Stochastic gradients

Using full gradients at each iteration can be computationally expensive, as the cost of a gradient evaluation is high, as it requires an evaluation of the entire data set. The predominant approach used in machine learning optimization is to rely on a stochastic approximation of the gradient (see e.g. [132] for one of the first applications of such approaches). In many applications, this can dramatically reduce the computational cost, as the approximation will usually come at a fraction of the workload. In the context of sampling, there has been a great deal of interest to also use such ideas to improve the scalability of MCMC to large datasets, see e.g. [164], or the recent review paper [120]. This is typically done by considering a potential (the negative log-density) $U : \mathbb{R}^d \to \mathbb{R}$ of the form

$$U(x) = U_0(x) + \sum_{i=1}^{N_D} U_i(x), \tag{1.6.12}$$

where $x \in \mathbb{R}^d$, N_D is the size of your dataset and typically is large. After defining a posterior density, U_0 can be chosen to be the negative log density of the prior distribution. Then random variables of the form $\omega \in [N_D]^{N_b}$, which is a random selection of N_b indices, are to be selected uniformly on $[N_D] = \{1, \dots, N_D\}$, i.i.d. with replacement [8], then at each step of your MCMC method one uses an unbiased estimator of (1.6.12) instead of the full gradient evaluation, for example

$$\mathcal{G}(x, \omega|\hat{x}) = \nabla U_0(x) + \sum_{i=1}^{N_D} \nabla U_i(\hat{x}) + \frac{N_D}{N_b} \sum_{i \in \omega} [\nabla U_i(x) - \nabla U_i(\hat{x})], \tag{1.6.13}$$

such that $\mathbb{E}(\mathcal{G}(x, \omega)) = \nabla U(x)$.

When using stochastic gradients, the aim is to reduce the computational cost from $\mathcal{O}(N_D)$, when using full-gradients, to $\mathcal{O}(1)$ (in the regime where $N_b << N_D$). However, the use of stochastic gradients incurs an additional bias, for example, the BAOAB or UBU integrator with the control variate stochastic gradient estimator [8] (setting $\hat{x}$ to be the minimizer of U in (1.6.13), when sampling a log-concave measure) has order one bias in the stepsize compared to order two when full gradients are used [143]. When considering Metropolised methods, when using stochastic gradients, the cost of implementing bias correction scales linearly with dataset size, $\mathcal{O}(N_D)$, therefore mitigating the improved computational efficiency. It is a significant challenge to implement stochastic gradient algorithms in a computationally efficient way, whilst not significantly affecting the bias of your estimate.

We remark that one of the most efficient samplers in the big data regime is the Zig-Zag sampler [18] whose complexity is independent of the data size according to a limiting argument (although as stated in [18], some logarithmic factors were ignored). [50] is another recent paper that proposes a Metropolis-Hastings-type MCMC algorithm based on subsampling that only accesses $\mathcal{O}(1)$ or even $\mathcal{O}(1/\sqrt{N_D})$ data points per step. Although this method was shown to have state-of-the-art performance on a 10-dimensional logistic regression example, its efficiency on high-dimensional models has not yet been demonstrated. [40] introduce an estimation method which requires $\mathcal{O}(1)$ cost (in terms of N_D) per effective sample, which is the focus of Chapter 3, with state-of-the-art performance on a variety of high-dimensional models.

1.7 Notation

- Let $h > 0$ denote the stepsize of a discretization.
- Let $\gamma > 0$ denote the friction parameter in kinetic Langevin dynamics.
- Let $\eta := \exp(-\gamma h/2)$.
- Let m denote the strong convexity constant of a potential U.
- Let M denote the Lipschitz constant of the gradient of a potential ∇U.
- Let M_1 denote the Lipschitz constant of the Hessian of a potential $\nabla^2 U$.
- Let M_1^s denote the strongly-Hessian Lipschitz constant of a potential U.
- We let $z_{0:k} = (z_0, z_1, \dots, z_k)$ denote a sequence of variables.
- Let I_d denote the d-dimensional identity matrix.

- Let C denote an absolute constant (whose value may differ in each proposition or theorem).
- Let $C(\mathrm{var}_1, \ldots, \mathrm{var}_n)$ denote a constant that is a function of variables $\mathrm{var}_1, \ldots, \mathrm{var}_n$ (this function may differ in each proposition or theorem).
- Let G, SG and A denote an abbreviation for gradient, stochastic gradient and "approximate gradient".
- We let $l \in \mathbb{R}^+$ denote the level of discretization with respect to our discretized ULD, with stepsize h_l defined at each level.
- Let D_0 denote the empirical average of samples at level 0.
- Let $D_{l,l+1}$ denote the difference of empirical averages of samples at levels $l+1$ and l, which are generated jointly via a synchronous coupling.
- N denotes the number of samples taken at level 0.
- $N_{l,l+1}$ denotes the number of samples taken from the coupling of levels l and $l+1$.
- N_D is the size of the dataset (number of terms in potential $U(x) = U_0(x) + \sum_{i=1}^{N_D} U_i(x)$).
- N_b is the minibatch size.
- Let $z_k = (x_k, v_k)$ denotes step k in a numerical discretization of kinetic Langevin dynamics with time step h (specified each time this notation is used). Similarly, Z_t is the solution of (1.3.7) initialized at the invariant measure with synchronously coupled Brownian motion. $Z^k = Z_{kh}$ denotes the value of the continuous-time process at the same time as z_k.
- $\hat{x}_k$ denotes the point where the last gradient is evaluated for SVRG
- $\|\cdot\|_{L^2} := \left(\mathbb{E}\|\cdot\|^2\right)^{1/2}$ and $\|\cdot\|_{L^2,a,b} := \left(\mathbb{E}\|\cdot\|_{a,b}^2\right)^{1/2}$.
- $\psi_{\mathrm{method}}(x, v, h)$ is the one-step map of the "method" discretization with initial conditions $(x, v) \in \mathbb{R}^{2d}$ and stepsize $h > 0$, where "method" is any discretization considered in the thesis. $\psi_{\mathrm{method}}(x, v)$ may also be used when the stepsize is clear in the context.

1.8 Contributions and organisation of thesis

This thesis is organised as follows: Chapter 2 consists of the content of the papers:

Benedict J. Leimkuhler, Daniel Paulin and P.A.Whalley. Contraction and Convergence Rates for Discretized Kinetic Langevin Dynamics. SIAM Journal on Numerical Analysis, 62(3):1226–1258, 2024,

and

Benedict J. Leimkuhler, Daniel Paulin and Peter A. Whalley. Contraction rate estimates of stochastic gradient kinetic Langevin integrators. ESAIM: Mathematical Modelling and Numerical Analysis, 2024.

Chapter 3 consists of the preprint

Neil K. Chada, Benedict J. Leimkuhler, Daniel Paulin & Peter A. Whalley. Unbiased kinetic Langevin Monte Carlo methods. arXiv preprint arXiv:2311.05025, 2023.

In Chapter 2 (see [105] and [103]) we worked on proving Wasserstein convergence rates of numerical methods for kinetic Langevin dynamics, which hold for stepsizes all the way up to the stability threshold of the discretization. To achieve such convergence rates, it was necessary to construct stepsize dependent quadratic norms, as simply extending the norms available for the continuous

analysis was not possible in many cases. We further show a robustness property of certain schemes in the high friction limit. This robustness property is shown in Figure 2.2 for OBABO, BAOAB and rOABAO, where the convergence rate does not decrease like $\mathcal{O}(1/\gamma)$ like the other schemes. To support this we provide asymptotic bias estimates which remain accurate in the high friction limit for the BAOAB scheme by introducing modified stochastic dynamics which preserve the invariant measure and exhibit this robustness property. This can be combined with the convergence rate estimates to provide non-asymptotic guarantees.

In Chapter 3 (see [40]) we introduced the "UBUBU" algorithm, an alternative to Metropolis-Hastings accept-reject. It is an unbiased method for estimating Bayesian posterior means based on a high strong order numerical integrator for kinetic Langevin dynamics, which only requires one gradient evaluation per step. Our approach avoids Metropolis correction by coupling different discretization levels in a multilevel Monte Carlo approach and consequently removing bias [77] and making use of the theoretical results achieved in Chapter 2. We provide theoretical analysis to demonstrate in the setting, where the target measure is log-concave and under appropriate regularity assumptions on the potential, that it can achieve accuracy $\epsilon > 0$ for estimating expectations of Lipschitz functions in d dimensions with $\mathcal{O}(d^{1/4}\epsilon^{-2})$ expected gradient evaluations. This is state-of-the-art in terms of dimension dependence, as we do not assume a warm start. We successfully use this method to remove stochastic gradient bias within subsampling methods and remove bias within approximate gradient methods. We show that the computational complexity scales independently of the dataset size N_D, which is a significant improvement over $\mathcal{O}(N_D)$ in the full gradient setting, whilst being unbiased.

We test this method on real-world data for classification of the MNIST dataset and predicting premier league football scores. In these applications, we did not observe any dimension dependence and theoretically, we show dimension-independent bounds for sampling from product distributions. In our numerical experiments, we observe a major improvement of our method over the state-of-the-art Metropolis-Hastings-based methods with computational savings of the order of $\mathcal{O}(10^2)$. Figure 3.4 illustrates the complexity scaling for an independent sample of our method compared to randomized Hamiltonian Monte Carlo, the state-of-the-art method in terms of dimension dependence for Gaussian targets.

Contribution by the author of the thesis

The work in Chapter 2 was a collaboration with my supervisors Prof. Benedict Leimkuhler and Dr. Daniel Paulin. This project arose from trying to improve convergence bounds for unadjusted integrators within the context of unbiased estimation. The majority of the work in this chapter was conducted solely by me. However, Daniel performed the numerical experiments for the Bayesian logistic regression problem and Daniel came up with the idea to use an interpolation argument to improve the BAOAB bias bounds.

The work in Chapter 3 was a collaboration with my supervisors Dr. Neil Chada, Prof. Benedict Leimkuhler and Dr. Daniel Paulin. The original idea for the project was proposed by Daniel. My contributions were to suggest the use of the UBU integrator and provide the majority of the theoretical results in the Appendix, more specifically I contributed significantly to the results of Appendices C, D, E, F, G and H building the key Theorems in the main text. My theoretical insights from these results also led to improvements of the algorithm in the inexact and stochastic gradient setting to achieve complexity bounds which are independent of the size of the dataset.

Chapter 2

Wasserstein convergence and bias estimates of discretized kinetic Langevin dynamics

2.1 Introduction

In this chapter, based on [105] and [103] we focus our attention to proving convergence rates of numerical methods for kinetic Langevin sampling in Wasserstein distance (see [155]). We use coupling methods to establish these convergence rates (see [81]), more specifically synchronous coupling as in [47, 55, 80, 115, 140]. Proving contraction of a coupling has been a popular method for establishing convergence both in the continuous time setting and for the discretization for Langevin dynamics and Hamiltonian Monte Carlo ([22–24, 26, 57, 70, 131, 141]), since a consequence of such a contraction is convergence in Wasserstein distance (viewed as the infimum over all possible couplings with respect to some norm). The numerical methods we consider are the Euler-Maruyama discretisation (EM), splitting methods based on B,A and O splitting including BAOAB and OBABO [35, 98], the Brunger-Brooks Karplus discretisation (BBK) [34], the stochastic position and velocity Verlet (SPV, SVV) [111], a randomized method based on the Hamiltonian integrator of [26] (rOABAO) and the stochastic Euler scheme (SES/EB) [42, 71, 149].

When considering MCMC methods the performance of a sampling scheme is often assessed by measuring the number of steps needed to achieve a certain level of accuracy in the Wasserstein distance metric. By combining the results of this chapter with estimates of the stepsize-dependent bias of the numerical methods, it is possible to develop such non-asymptotic bounds in Wasserstein distance which can ultimately provide insight into the computational complexity, convergence rate, and accuracy of the sampling scheme. Bias estimates of some relevant numerical methods have been treated in [80, 115, 140]. Where available, bias analysis can be combined with our contraction results to provide non-asymptotic guarantees. In this chapter, we also provide new asymptotic bias estimates for a scheme (BAOAB) which remain finite in the high-friction limit, by comparing the scheme to an alternative scheme which exactly preserves the invariant measure. This scheme switches between exact Hamiltonian dynamics and Ornstein–Uhlenbeck process steps and we believe this technique can be extended to other integrators of this type.

Moreover, we discuss the use of stochastic gradients and how these proofs can be extended to that setting, which is particularly important in the context of machine learning. We demonstrate our results on an anisotropic Gaussian as well as a Bayesian logistic regression problem involving the MNIST dataset. We verify the convergence results for the anisotropic Gaussian example, by computing spectral gaps for the numerical methods.

It is also important to note that bias analysis of kinetic Langevin dynamics in the stochastic gradient setting has been considered in [143][Proposition 4] using the techniques of [1, 160]. Assuming smooth test functions that are compactly supported, they achieve order one bias estimates in the stepsize for the stochastic gradient OABAO scheme. This aligns with what is observed in practice. It remains an open problem to achieve order one in stepsize (for both overdamped Langevin and kinetic Langevin dynamics) in Wasserstein distance, where bias estimates of order $1/2$ have been shown in [43, 53, 80].

Algorithm	stepsize restriction	one-step contraction rate
EM	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(mh/\gamma)$
BAO,OBA,AOB	$\mathcal{O}((1-\eta^2)/\sqrt{M})$	$\mathcal{O}(mh^2/(1-\eta^2))$
OAB, ABO, BOA	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(\eta^2 mh^2/(1-\eta^2))$
BBK	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(mh/\gamma)$
SPV	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(mh/\gamma)$
SVV	$\mathcal{O}(1/\gamma)$	$\mathcal{O}(mh/\gamma)$
BAOAB	$\mathcal{O}((1-\eta^2)/\sqrt{M})$	$\mathcal{O}(mh^2/(1-\eta^2))$
OBABO	$\mathcal{O}((1-\eta^2)/\sqrt{M})$	$\mathcal{O}(mh^2/(1-\eta^2))$ [80]
rOABAO	$\mathcal{O}((1-\eta^2)/\sqrt{M})$	$\mathcal{O}(mh^2/(1-\eta^2))$
SES/EB	$\mathcal{O}(1/\gamma)$ [140]	$\mathcal{O}(mh/\gamma)$ [55, 140]

Table 2.1: The table provides our stepsize restrictions and optimal contraction rates of the discretized kinetic Langevin dynamics with stepsize h for an m-convex, M-∇ Lipschitz potential and previous results of [105] and other recent work for further integrators for comparison. We define $\eta = e^{-\gamma h/2}$.

There have been other recent work aimed at providing convergence rates for kinetic Langevin dynamics under explicit restrictions on the parameters ([47, 55, 80, 115]), but these guarantees are valid only with sharp restrictions on stepsize. There has also been the work of [140] which considers a slightly different version of the SDE (1.3.7), where time is rescaled depending on the smallest and largest eigenvalues of the Hessian to optimize contraction rates and bias. We have included their results in Table 2.1 after converting them into our framework using [55, Lemma 1]. The results of [140] rely on a stepsize restriction of $\mathcal{O}(1/\gamma)$, but their analysis does not provide the stepsize threshold [140, Example 9], and the class of schemes considered is different, with only the stochastic Euler scheme in common. Further, the techniques they use to quantify the asymptotic bias are not appropriate for many of the schemes considered in this chapter, as they are designed for high strong-order numerical integrators. For example, the BAOAB scheme is only strong order one but has asymptotic bias of order two, which cannot be estimated by their approach. Other works on contraction of kinetic Langevin and its discretization include [54, 73, 167].

In the current chapter, we apply direct convergence analysis to various popular integration methods and provide a general framework for establishing convergence rates of kinetic Langevin dynamics with tight explicit stepsize restrictions of $\mathcal{O}(1/\gamma)$ or $\mathcal{O}(1/\sqrt{M})$ (depending on the scheme). As a consequence, we improve the contraction rates significantly for many of the available algorithms (see Table 2.1). For a specific class of schemes, we establish explicit bounds on the convergence rate for stepsizes of $\mathcal{O}(1/\sqrt{M})$. In the limit of large friction, we distinguish two types of integrators – those that converge to overdamped dynamics ("γ-limit-convergent") and those that do not. We demonstrate with examples that this property is not universal: some seemingly reasonable methods have the property that the convergence rate falls to zero in the $\gamma \to \infty$ limit. This is verified numerically and analytically for an anisotropic Gaussian target. Further, our novel asymptotic bias estimates for the BAOAB scheme demonstrate accuracy in the high-friction limit.

Further, we generalize the contraction rates for all schemes in Table 2.1 to appropriate versions of these schemes using stochastic dynamics (see Table 2.3 for a summary of results). We allow for a flexible choice of unbiased gradient estimators (i.e they do not necessarily have to be based on subsampling) and control errors via expected variability in the Jacobian of the stochastic gradient versus the Hessian of the true potential. It turns out that for all schemes, there is some reduction in convergence rate as the gradient noise increases (we observed this in our numerical experiments when using sub-sampling with very small batch sizes). Nevertheless, for a fixed level of gradient noise, the relative reduction in the contraction rate due to stochastic gradients becomes negligible as the stepsize decreases.

The remainder of this chapter is structured as follows. We first introduce overdamped Langevin dynamics, the Euler-Maruyama (EM) and the high-friction limit of BAOAB (LM) and discuss their convergence guarantees. Next, we introduce kinetic Langevin dynamics and describe various popular discretizations, and give our results on convergence guarantees with mild stepsize assumptions. These schemes include first and second-order splittings and the stochastic Euler scheme (SES). Further, we compare the results of overdamped Langevin and kinetic Langevin and show how schemes like BAOAB, OBABO, rOABAO exhibit the positive qualities of both cases with the "γ-limit convergent" property, whereas schemes like EM, SES, BBK, SPV and SVV do not perform well for a large range of γ. Finally, we give asymptotic bias estimates for the BAOAB scheme which support the "γ-limit convergent" property.

2.2 Assumptions and definitions

2.2.1 Assumptions on U

We will make the following assumptions on the potential $U : \mathbb{R}^d \to \mathbb{R}$.

Assumption 2.2.1 (M-∇Lipschitz)**.** *U is twice continuously differentiable and there exists a $M > 0$ such that for all $X, Y \in \mathbb{R}^d$*

$$|\nabla U(X) - \nabla U(Y)| \leq M |X - Y| .$$

Assumption 2.2.2 (m-convexity)**.** *U is continuously differentiable and there exists a $m > 0$ such that for all $X, Y \in \mathbb{R}^d$*

$$\langle \nabla U(X) - \nabla U(Y), X - Y \rangle \geq m \left| X - Y \right|^2 .$$

Assumption 2.2.3 (M_1-Hessian Lipschitz)**.** *U is three times continuously differentiable and there exists a $M_1 > 0$ such that for all $X, Y \in \mathbb{R}^d$*

$$\left| \nabla^2 U\left(X\right) - \nabla^2 U\left(Y\right) \right| \leq M_1 \left| X - Y \right| .$$

The first two assumptions are popular conditions used to obtain explicit convergence rates, see [52, 55] for example. It is worth mentioning that these assumptions can also produce explicit convergence rates for gradient descent [32]. The final assumption is only used for proving higher order asymptotic bias estimates in Section 2.7.

2.2.2 Modified Euclidean norms

For kinetic Langevin dynamics, it is not possible to prove convergence with respect to the standard Euclidean norm due to the fact that the generator is hypoelliptic. We therefore work with a modified Euclidean norm as in [115]. For $z = (x, v) \in \mathbb{R}^{2d}$ we introduce the weighted Euclidean norm

$$||z||_{a,b}^2 = ||x||^2 + 2b \langle x, v \rangle + a \left| |v| \right|^2 ,$$

for $a, b > 0$, which is equivalent to the Euclidean norm on $\mathbb{R}^{2d}$ as long as $b^2 < a$. Under the condition $b^2 < a/4$, we have

$$\frac{1}{2}||z||_{a,0}^2 \leq ||z||_{a,b}^2 \leq \frac{3}{2}||z||_{a,0}^2.$$

2.2.3 Wasserstein distance

We define $\mathcal{P}_p\left(\mathbb{R}^{2d}\right)$ to be the set of probability measures which have finite p-th moment, then for $p \in [0, \infty)$ we define the p-Wasserstein distance on this space. Let μ and ν be two probability measures. We define the p-Wasserstein distance between μ and ν with respect to the norm $|| \cdot ||_{a,b}$ (introduced in Section 2.2.2) to be

$$\mathcal{W}_{p,a,b}\left(\nu, \mu\right) = \left(\inf_{\xi \in \Gamma(\nu,\mu)} \int_{\mathbb{R}^{2d}} ||z_1 - z_2||_{a,b}^p d\xi\left(z_1, z_2\right) \right)^{1/p},$$

where $\Gamma\left(\mu, \nu\right)$ is the set of measures with marginals μ and ν (the set of all couplings between μ and ν).

It is well known that the existence of couplings with a contractive property implies convergence in Wasserstein distance, which can be interpreted as the infimum over all such couplings. The simplest such coupling is the synchronous coupling, which considers simulations with common noise. If one can show contraction of two simulations that share noise increments with an explicit contraction rate, then one has convergence in Wasserstein distance at the same rate. Given all the constants and conditions derived for contraction in all schemes, we have convergence in Wasserstein distance by the following proposition:

Proposition 2.2.4. *Assume a numerical scheme for kinetic Langevin dynamics with a m-strongly convex M-∇Lipschitz potential U and transition kernel P_h. Let (x_n, v_n) and $(\tilde{x}_n, \tilde{v}_n)$ be two synchronously coupled chains of the numerical scheme that have the contraction property*

$$||(x_n - \tilde{x}_n, v_n - \tilde{v}_n)||_{a,b}^2 \leq C(1 - c\,(h))^n ||(x_0 - \tilde{x}_0, v_0 - \tilde{v}_0)||_{a,b}^2, \tag{2.2.1}$$

for $\gamma^2 \geq C_\gamma M$ and $h \leq C_h \left(\gamma, \sqrt{M}\right)$ for some $a, b > 0$ such that $b^2 < a/4$. Then we have that for all $\gamma^2 \geq C_\gamma M$, $h \leq C_h \left(\gamma, \sqrt{M}\right)$, $1 \leq p \leq \infty$ and all $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{2d})$, and all $n \in \mathbb{N}$,

$$\mathcal{W}_p^2 \left(\nu P_h^n, \mu P_h^n\right) \leq 3C \max\left\{a, \frac{1}{a}\right\} (1 - c\,(h))^n \, \mathcal{W}_p^2 \left(\nu, \mu\right).$$

Further to this, P_h has a unique invariant measure which depends on the stepsize, π_h, where $\pi_h \in \mathcal{P}_p(\mathbb{R}^{2d})$ for all $1 \leq p \leq \infty$.

Proof. The proof is given in [115, Corollary 20], which relies on [159, Corollary 5.22, Theorem 6.18]. □

The focus of this chapter is to prove contractions of the form (2.2.1), and hence to achieve Wasserstein convergence rates by Proposition 2.2.4. With convergence to the invariant measure of the discretizations of kinetic Langevin dynamics considered here it will be possible to combine our results with estimates of the bias of each scheme as in [55], [115], [140] and [47] to obtain non-asymptotic estimates. However, with the bias bounds provided in Section 2.7 we provide non-asymptotic estimates for BAOAB.

2.3 Overdamped Langevin discretizations and contraction

We first consider two discretizations of the SDE (1.2.4), namely the Euler-Maruyama discretization and the high-friction limit of the popular kinetic Langevin dynamics scheme BAOAB [98]. The simplest discretization of overdamped Langevin dynamics is using the Euler-Maruyama (EM) method which is defined by the update rule

$$x_{n+1} = x_n - h\nabla U\left(x_n\right) + \sqrt{2h}\xi_{n+1}, \tag{2.3.2}$$

where $(\xi_n)_{n\in\mathbb{N}}$ are independent d-dimensional standard Gaussian random variables, that is $\xi_n \sim \mathcal{N}(0_d, I_d)$ for all $n \in \mathbb{N}$, where I_d is the d-dimensional identity matrix. This scheme is combined with Metropolization in the popular MALA algorithm.

An alternative method is the BAOAB limit method of Leimkuhler and Matthews (LM)([98], [102]) which is defined by the update rule

$$x_{n+1} = x_n - h\nabla U\left(x_n\right) + \sqrt{2h}\frac{\xi_{n+1} + \xi_n}{2}. \tag{2.3.3}$$

The advantage of this method is that it gains a weak order of accuracy asymptotically.

2.3.1 Convergence guarantees

The convergence guarantees of overdamped Langevin dynamics and its discretizations have been extensively studied under the assumptions presented (see [46, 51, 52, 64–66, 69]). We use synchronous coupling as a proof strategy to obtain convergence rates as in [52]. We first consider two chains x_n and y_n with shared noise such that

$$x_{n+1} = x_n - h\nabla U(x_n) + \sqrt{2h}\xi_{n+1}, \quad y_{n+1} = y_n - h\nabla U(y_n) + \sqrt{2h}\xi_{n+1}.$$

Then we have that

$$\begin{aligned}
&||x_{n+1} - y_{n+1}||^2 = ||x_n - y_n + h(-\nabla U(x_n) - (-\nabla U(y_n))||^2 \\
&= ||x_n - y_n||^2 - 2h\langle \nabla U(x_n) - \nabla U(y_n), x_n - y_n\rangle + h^2||\nabla U(x_n) - \nabla U(y_n)||^2 \\
&= ||x_n - y_n||^2 - 2h\langle x_n - y_n, Q(x_n - y_n)\rangle + h^2\langle x_n - y_n, Q^2(x_n - y_n)\rangle,
\end{aligned}$$

where $Q = \int_{t=0}^{1} \nabla^2 U(x_n + t(y_n - x_n))dt$. Q has eigenvalues which are bounded between m and M, so $Q^2 \preceq MQ$, and hence

$$h^2\langle x_n - y_n, Q^2(x_n - y_n)\rangle \leq h^2 M\langle x_n - y_n, Q(x_n - y_n)\rangle.$$

Therefore $||x_{n+1} - y_{n+1}||^2 \leq ||x_n - y_n||^2(1 - hm(2 - hM))$, assuming that $h \leq \frac{2}{M}$ we have contraction and

$$||x_n - y_n|| \leq (1 - hm(2 - hM))^{n/2}||x_0 - y_0||. \tag{2.3.4}$$

A consequence of this contraction result is that we have convergence in Wasserstein distance to the invariant measure with rate $hm\,(2 - hM)$, under the imposed assumptions on h (as discussed in Section 2.2.3)[115, 159].

This argument is similar to the LM discretization (2.3.3) of overdamped Langevin dynamics. Note that $(x_n)_{n\geq 0}$ by itself does not define a Markov chain for this discretization, but by extending the state space to include the noise, $(x_n, \xi_n)_{n\geq 0}$ does. For two initial points (x_0, ξ_0) and (y_0, ξ_0'), by using shared noise $\xi_n = \xi_n'$ for $n \geq 1$, it follows by the same argument that

$$||x_n - y_n|| \leq (1 - hm(2 - hM))^{n/2}||x_1 - y_1||, \tag{2.3.5}$$

for $n \geq 2$, since two processes have shared noise after $n = 0$.

The stepsize assumption for convergence of overdamped Langevin dynamics in this setting is weak and is the same assumption as is needed to guarantee convergence of gradient descent in optimization [32].

2.4 Kinetic Langevin dynamics

We now consider several discretization methods for the SDE (1.3.7). The aim is to construct an alternative Euclidean norm in which we can prove contraction (it is not possible to prove contraction in the standard Euclidean norm). Essentially, we convert the problem of proving contraction to the problem of showing that certain matrices are positive definite. First, in Section 2.4.1 we introduce the discretization methods we consider; then in Section 2.4.2 we provide a framework for proving contraction for these discretizations, and in Section 2.4.3 we detail the contraction results for each of the schemes.

2.4.1 Discretization schemes

We will consider several popular numerical integrators for kinetic Langevin dynamics, arising in the literatures of molecular dynamics and machine learning. The numerical methods are generally defined by $(x_k, v_k) \in \mathbb{R}^d \times \mathbb{R}^d$ for $k \in \mathbb{N}$ with initial conditions $(x_0, v_0) \in \mathbb{R}^d \times \mathbb{R}^d$ and given noise sequences.

Choices for the algorithm include:

- the Euler-Maruyama discretization (EM)
- splitting methods based on breaking the dynamics into parts which can be solved analytically (in the weak sense) [35, 98, 101]
- the stochastic Euler scheme (SES/EB) (see [42, 71, 149]), which is popular in the machine learning literature (see [47, 55, 140]) and is based on keeping the force constant and integrating exactly over the interval
- the Brunger-Brooks Karplus (BBK) scheme which uses a leapfrog-like approach to propagate position and velocity components, combined with implicit and explicit Euler steps in velocity [34, 72]
- the stochastic position and velocity Verlet schemes (SPV,SVV) based on integrating the force and the OU process together in a splitting scheme introduced in [111]
- a new randomized midpoint method based on a Hamiltonian integrator from [26]

We recommend [63] and [72] for an introduction to many of these schemes. We next describe these algorithms by giving their respective update rules.

The Euler-Maruyama method

First, we consider the simplest discretization, the Euler-Maruyama method. For the initial condition $(x_0, v_0) \in \mathbb{R}^{2d}$, the iterates (x_n, v_n, ξ_n) for $n \in \mathbb{N}$ are defined by:

$$x_{n+1} = x_n + hv_n, \tag{2.4.6}$$

$$v_{n+1} = v_n - h\nabla U(x_n) - h\gamma v_n + \sqrt{2h\gamma}\xi_{n+1}, \tag{2.4.7}$$

where $(\xi_n)_{n\in\mathbb{N}}$ are independent $\mathcal{N}(0, I_d)$ draws.

Splitting Methods

More advanced integrators than Euler-Maruyama can be constructed based on splitting. These rely on an additive decomposition of the SDE into various terms which can be easily (often exactly) integrated. A useful class of schemes relies on the exact integration of linear positional drift, impulse due to the force and a dissipative-stochastic term corresponding to an Ornstein-Uhlenbeck equation [35]. The solution maps corresponding to these parts may be denoted by $\mathcal{B}$, $\mathcal{A}$, and $\mathcal{O}$ with update rules given by

$$\begin{aligned} \mathcal{B} &: (x, v) \to (x, v - h\nabla U(x)), \\ \mathcal{A} &: (x, v) \to (x + hv, v), \\ \mathcal{O} &: (x, v) \to \left(x, \eta^2 v + \sqrt{1 - \eta^4}\xi\right), \end{aligned} \tag{2.4.8}$$

where $\xi \sim \mathcal{N}(0_d, I_d)$ and

$$\eta := \exp(-\gamma h/2).$$

The infinitesimal generator of the SDE dynamics (1.3.7) can be split as $\mathcal{L} = \mathcal{L}_\mathcal{A} + \mathcal{L}_\mathcal{B} + \gamma\mathcal{L}_\mathcal{O}$, where

$$\mathcal{L}_\mathcal{A} = \langle v, \nabla_x \rangle, \qquad \mathcal{L}_\mathcal{B} = -\langle \nabla U(x), \nabla_v \rangle, \qquad \mathcal{L}_\mathcal{O} = -\langle v, \nabla_v \rangle + \triangle_v, \tag{2.4.9}$$

where $\mathcal{L}_\mathcal{A}$ and $\mathcal{L}_\mathcal{B}$ are the deterministic dynamics related to the both $\mathcal{A}$ and $\mathcal{B}$, while the dynamics of $\gamma\mathcal{L}_\mathcal{O}$ corresponds to the dynamics of an Ornstein-Uhlenbeck process. The idea of splitting schemes is that we compose the maps $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{O}$ as functions in some way. Each of the deterministic terms can be solved analytically, whereas $\exp(h\mathcal{L}_\mathcal{O})$ is realized by a weakly exact process. There are several possible options for the ordering of the different parts which are denoted by different strings. For example, the scheme ABO would apply, in sequence, the A, B and O propagators. Other alternative options include BAO, BOA, AOB, OAB, and OBA. Such schemes are called first order because they have a weak order of one.

Alternatively, we can go beyond first-order methods to attain higher weak orders than one by using symmetric Strang splittings. These are defined by palindromic letter sequences such as ABOBA, OBABO, etc. The interpretation of a string such as ABOBA is the following: we apply A for half a timestep (drift), then B for half a timestep (kick), generate a stochastic path corresponding to the Ornstein-Uhlenbeck equation in momentum, then follow with a half-step kick and finally a half-step drift. Such symmetric schemes can typically be applied using only one new force evaluation at each timestep (with the second force evaluation re-used at the start of the following step). Despite this, the symmetry implies that they have second (weak) order of accuracy (see [98, 101]), meaning that the bias in long run averages is $O(h^2)$ for stepsize h as $h \to 0$. Thus they are efficient in providing high accuracy at little additional cost compared to first-order methods. Moreover, as shown in [99], a particular choice of splitting, namely the BAOAB method, has no bias at all for Gaussian targets.

We will use the notation $\psi_{\mathrm{BAOAB}}(x, v, h)$ throughout to denote the one-step map of the BAOAB discretization with initial conditions $(x, v) \in \mathbb{R}^{2d}$ and stepsize $h > 0$, and similarly for other discretizations.

The stochastic exponential Euler method

See [63] for an introduction to the stochastic exponential Euler scheme and a derivation. This scheme is based on keeping the force constant and analytically integrating the whole process over a time interval. This scheme is the one considered in [47, 55] and has gained a lot of attention in the machine learning community and we can apply our methods to this scheme. Similar schemes have also been considered in [42, 71, 149] and it has been analyzed in [63, 147]. In the notation we have used, it is defined by the updates

$$\begin{aligned} x_{n+1} &= x_n + \frac{1-\eta^2}{\gamma} v_n - \frac{\gamma h + \eta^2 - 1}{\gamma^2} \nabla U(x_n) + \zeta_{n+1}, \\ v_{n+1} &= \eta^2 v_n - \frac{1-\eta^2}{\gamma} \nabla U(x_n) + \omega_{n+1}, \end{aligned} \tag{2.4.10}$$

where

$$\zeta_{n+1} = \sqrt{2\gamma} \int_0^h e^{-\gamma(h-s)} dW_{h\gamma+s}, \qquad \omega_{n+1} = \sqrt{2\gamma} \int_0^h \frac{1 - e^{-\gamma(h-s)}}{\gamma} dW_{h\gamma+s}. \tag{2.4.11}$$

$(\zeta_n, \omega_n)_{n\in\mathbb{N}}$ are i.i.d Gaussian random vectors with covariances matrix $\Sigma \otimes I_d$ with Σ given by

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_3 \end{pmatrix}, \tag{2.4.12}$$

where

$$\begin{aligned} \Sigma_1 &= \frac{1}{\gamma} \left(2h - \frac{3 - 4\eta^2 + \eta^4}{\gamma} \right), \\ \Sigma_2 &= \frac{1}{\gamma} \left(1 - \eta^2\right)^2, \\ \Sigma_3 &= 1 - \eta^4, \end{aligned}$$

as defined in [63]. We can couple two trajectories which have common noise $(\zeta_n, \omega_n)_{n\in\mathbb{N}}$ to obtain contraction rates by the previously introduced methods.

BBK

For initial conditions $(x_0, v_0) \in \mathbb{R}^{2d}$, the iterations $(x_k, v_k) \in \mathbb{R}^{2d}$ for $k \in \mathbb{N}$ of the BBK method of [34] are defined by the update rule

$$\begin{aligned} x_{k+1} &= x_k + h(1 - \frac{\gamma h}{2}) v_k - \frac{h^2}{2} \nabla U(x_k) + \sqrt{2\gamma} \frac{h^{3/2}}{2} \xi_k, \\ v_{k+1} &= \frac{1 - \gamma h/2}{1 + \gamma h/2} v_k - \frac{h}{2(1+\gamma h/2)} \left(\nabla U(x_k) + \nabla U(x_{k+1}) \right) + \frac{\sqrt{2\gamma h}}{2(1+\gamma h/2)} (\xi_{k+1} + \xi_k), \end{aligned}$$

which can be rewritten as [72]

$$
\begin{aligned}
(B_1) \qquad & v_{k+1/2} = v_k + \frac{h}{2}\left(-\nabla U(x_k) - \gamma v_k + \frac{\sqrt{2\gamma}}{\sqrt{h}}\xi_k\right), \\
(A) \qquad & x_{k+1} = x_k + h v_{k+1/2}, \\
(B_2) \qquad & v_{k+1} = v_{k+1/2} + \frac{h}{2}\left(-\nabla U(x_{k+1}) - \gamma v_{k+1} + \frac{\sqrt{2\gamma}}{\sqrt{h}}\xi_{k+1}\right).
\end{aligned}
$$

This can be viewed as an explicit Euler step followed by a position update followed by an implicit Euler step. We denote the explicit Euler step by B_1 and the implicit Euler step by B_2

Stochastic position and velocity Verlet

The stochastic position and velocity Verlet schemes are defined through an alternative splitting of the dynamics based on keeping the B and O steps together in an exact integration. We define the operators involved in the update rule by

$$
\begin{aligned}
&\mathcal{V}(h) : v \to \eta^2 v - \frac{1-\eta^2}{\gamma}\nabla U(x) + \sqrt{1-\eta^4}\xi, \\
&\mathcal{A}(h) : x \to x + hv,
\end{aligned}
$$

where $\eta = e^{-\gamma h/2}$. Then the stochastic position Verlet is defined by $\mathcal{A}(h/2)\mathcal{V}(h)\mathcal{A}(h/2)$ and the stochastic velocity Verlet is defined by $\mathcal{V}(h/2)\mathcal{A}(h)\mathcal{V}(h/2)$.

Randomized midpoint method

Other algorithms for kinetic Langevin dynamics include the randomized midpoint methods considered in [146] and analyzed in [37], which have improved dimension dependence in non-asymptotic estimates. However, they involve multiple gradient evaluations at each step and cannot be analyzed in our framework; this problem has been discussed in [140]. For contractivity of algorithms involving several gradient evaluations we refer the reader to [139].

In the recent paper [26] the authors consider such a method for Hamiltonian Monte Carlo, whose discretization is closely related to the OBABO or OABAO discretization in the $\gamma \to \infty$ limit [80]. More precisely one could consider the following procedure.

Fix a stepsize h then sample $u \sim [0, h]$ and compute

$$
\begin{aligned}
&\mathcal{A} : x \to x + uv, \\
&\mathcal{B} : v \to v - h\nabla U(x), \\
&\mathcal{A} : x \to x + (h-u)v,
\end{aligned}
$$

which is the following update

$$
\begin{aligned}
x_{k+1} &= x_k + hv_k - h(h-u)\nabla U(x_k + uv_k), \\
v_{k+1} &= v_k - h\nabla U(x_k + uv_k),
\end{aligned}
$$

then only considering the randomness in the gradient evaluation we arrive at the Verlet scheme considered in [26]. We define $r\mathcal{ABA}$ to be the update

$$x_{k+1} = x_k + hv_k - \frac{h^2}{2}\nabla U(x_k + uv_k),$$
$$v_{k+1} = v_k - h\nabla U(x_k + uv_k),$$

where $u \sim \mathcal{U}(0,h)$ as introduced in [26]. The key difference being that the gradient is evaluated at a random midpoint in the interval of numerical integration. We define the kinetic Langevin dynamics integrator rOABAO to be $\mathcal{O}(r\mathcal{ABA})\mathcal{O}$.

We remark that we can achieve contraction rates by coupling two trajectories which have common noise (in Brownian increment and randomized midpoint) $(\zeta_k, u_k)_{k\in\mathbb{N}}$ with the previously introduced methods. The convergence rates will be established in Section A.1.

2.4.2 Proof strategy

To prove contraction and Wasserstein convergence of the kinetic Langevin integrators we will consider a modified Euclidean norm as defined in Section 2.2.2 for some choice of a and b. We aim to construct an equivalent Euclidean norm such that contraction occurs for two Markov chains simulated by the same discretization scheme $z_n = (x_n, v_n) \in \mathbb{R}^{2d}$ and $\tilde{z}_n = (\tilde{x}_n, \tilde{v}_n) \in \mathbb{R}^{2d}$ that are synchronously coupled. That is, for some choice of a and b such that $a, b > 0$ and $b^2 < a/4$

$$||\tilde{z}_{n+1} - z_{n+1}||_{a,b}^2 < (1 - c(h))\,||\tilde{z}_n - z_n||_{a,b}^2, \tag{2.4.13}$$

where a and b are chosen to provide reasonable explicit assumptions on the stepsize h and friction parameter γ. Our initial choices of a and b for simple schemes are motivated by [115], and are derived by considering contraction of the continuous dynamics. Let $\overline{z}_j = \tilde{z}_j - z_j$ for $j \in \mathbb{N}$, then (2.4.13) is equivalent to showing that

$$\overline{z}_n^T\left((1 - c(h))\,G - P^T G P\right)\overline{z}_n > 0, \quad \text{where} \quad G = \begin{pmatrix} I_d & bI_d \\ bI_d & aI_d \end{pmatrix}, \tag{2.4.14}$$

and $\overline{z}_{n+1} = P\overline{z}_n$ (P depends on z_n and $\tilde{z}_n$, but we omit this in the notation).

Example 2.4.1. *As an example, we have for the Euler-Maruyama method the update rule for* $\overline{z}_n$

$$\overline{x}_{n+1} = \overline{x}_n + h\overline{v}_n, \qquad \overline{v}_{n+1} = \overline{v}_n - \gamma h\overline{v}_n - hQ\overline{x}_n,$$

where by the mean value theorem we can define $Q = \int_{t=0}^1 \nabla^2 U(\tilde{x}_n + t(x_n - \tilde{x}_n))dt$, *then* $\nabla U(\tilde{x}_n) - \nabla U(x_n) = Q\overline{x}$. *One can show that in the notation of equation* (2.4.14) *we have*

$$P = \begin{pmatrix} I_d & hI_d \\ -hQ & (1 - \gamma h)\,I_d \end{pmatrix}. \tag{2.4.15}$$

Proving contraction for a general scheme is equivalent to showing that the matrix $\mathcal{H} := (1 - c(h))\, G - P^T G P \succ 0$ is positive definite. The matrix $\mathcal{H}$ is symmetric and hence of the form

$$\mathcal{H} = \begin{pmatrix} A & B \\ B & C \end{pmatrix}, \tag{2.4.16}$$

we can show that $\mathcal{H}$ is positive definite by applying the following Proposition 2.4.2.

Proposition 2.4.2. *Let $\mathcal{H}$ be a symmetric matrix of the form (2.4.16), then $\mathcal{H}$ is positive definite if and only if $A \succ 0$ and $C - BA^{-1}B \succ 0$. Further if A, B and C commute then $\mathcal{H}$ is positive definite if and only if $A \succ 0$ and $AC - B^2 \succ 0$.*

Remark 2.4.3. *If A, C commute and are symmetric (which is true when $\mathcal{H}$ is symmetric), then $AC = CA$ is symmetric.*

Proof. The proof of the first result is given in [84]. To establish the second statement, observe from [83] that if two matrices are positive definite and they commute then the product is positive definite. Also if $A \succ 0$ then $A^{-1} \succ 0$ (as A is symmetric positive definite). Further A, B and C commute and hence B, C and A^{-1} commute. Therefore by applying the first result, we have that $A \succ 0$ and

$$A^{-1}\left(AC - B^2\right) = C - BA^{-1}B \succ 0,$$

hence $\mathcal{H}$ is positive definite. If $\mathcal{H}$ is positive definite then $A \succ 0$ and $C - BA^{-1}B \succ 0$ by the first result. Thus as A, B and C commute we have $AC - B^2 \succ 0$. □

Remark 2.4.4. *An equivalent condition for a symmetric matrix $\mathcal{H}$ of the form (2.4.16) to be positive definite is $C \succ 0$ and $AC - B^2 \succ 0$ when A, B and C commute. One could equivalently prove that $C \succ 0$ instead of $A \succ 0$ if it is more convenient.*

Our general approach to prove contraction of kinetic Langevin dynamics schemes is to prove the conditions of Proposition 2.4.2 are satisfied to establish contraction. We will use the notation laid out in this section in the proofs given in the appendix.

2.4.3 Convergence results

We now detail contraction results of all the schemes introduced in Section 2.4.1. These results use the proof strategy of Section 2.4.2.

Theorem 2.4.5. *For the numerical schemes for kinetic Langevin dynamics given in Table 2.2 with an m-strongly convex, M-∇Lipschitz potential U we consider any sequence of synchronously coupled random variables with initial conditions $(x_0, v_0) \in \mathbb{R}^{2d}$ and $(\tilde{x}_0, \tilde{v}_0) \in \mathbb{R}^{2d}$.*

Under stepsize restrictions $h < h_0$ and $\gamma \geq \gamma_0$ for constants given in Table 2.2 we have the contraction

$$\|(x_k - \tilde{x}_k, v_k - \tilde{v}_k)\|_{a,b} \leq C(1 - c(h))^{s/2} \|(x_0 - \tilde{x}_0, v_0 - \tilde{v}_0)\|_{a,b},$$

with norm given with constants $a = 1/M$ and b, where b, the contraction rate $c(h)$, the preconstant C and the number of steps s are given in Table 2.2 and are specific to each scheme.

Algorithm	h_0	γ_0	b	c(h)	C	s
BAO	$(1-\eta^2)/\sqrt{6M}$	implicit	$h/(1-\eta^2)$	$h^2m/4(1-\eta^2)$	1	k
OAB	$\min\left\{1/4\gamma, (1-\eta^2)/\sqrt{6M}\right\}$	implicit	$\eta^2h/(1-\eta^2)$	$\eta^2h^2m/(1-\eta^2)$	1	k
EM	$1/2\gamma$	$2\sqrt{M}$	$1/\gamma$	$mh/2\gamma$	1	k
BBK	$1/4\gamma$	$\sqrt{12M}$	$h/2+1/\gamma$	$mh/4\gamma$	7	$k-1$
SPV	$1/2\gamma$	$\sqrt{11M}$	$h/(1-\eta^2)$	$mh/4\gamma$	7	$k-1$
SVV	$1/2\gamma$	$\sqrt{11M}$	$h/(1-\eta^2)$	$mh/4\gamma$	7	$k-1$
BAOAB	$(1-\eta^2)/2\sqrt{M}$	implicit	$h/(1-\eta^2)$	$h^2m/4(1-\eta^2)$	7	$k-1$
OBABO	$(1-\eta^2)/4\sqrt{M}$	implicit	$h/(1-\eta^2)$	$h^2m/4(1-\eta^2)$	7	$k-1$
rOABAO	$(1-\eta^2)/2\sqrt{M}$	implicit	$h/(1-\eta^2)$	$h^2m/4(1-\eta^2)$	7	$k-1$
SES/EB	$1/2\gamma$	$5\sqrt{M}$	$1/\gamma$	$mh/4\gamma$	1	k

Table 2.2: Constants for contraction of each scheme. Implicit refers to the implicit assumption on γ through the stepsize restriction h_0, the value of γ^2 must be greater than a certain constant multiple of M. For example, the condition $h_0 = (1-\eta^2)/\alpha\sqrt{M}$ is satisfied when $\gamma \geq 2\alpha\sqrt{M}$ and $h < 1/(2\gamma)$ and for $h \geq 1/(2\gamma)$ we have $h_0 \geq 1/(6\alpha\sqrt{M})$.

Referring to Table 2.2 we have that the convergence rate $c(h)$ is proportional to m/γ for small h, which is shown to match the convergence rate of the continuous dynamics for large γ (see for example [55]). We have that the convergence rate is hm/γ for all the schemes apart from BAOAB, OBABO and rOABAO, which have convergence rates which are faster than the continuous dynamics for large values of γ and h. This is due to the fact that the $\mathcal{O}$ step is integrated exactly separately and one can take the high friction limit. Since the $\mathcal{O}$ step also leaves the measure invariant the bias in these types of schemes comes from the discretization error of the Hamiltonian integrator and hence retains high order asymptotic bias [101]. However, these splitting schemes are only strong order 1 and this can be seen particularly for large values of friction when the convergence rates are higher than for the continuous dynamics. (It fails to approximate the continuous dynamics, but it is accurate in the sampling context as illustrated by the BAOAB asymptotic bias bounds in Section 2.7).

Example 2.4.6. *An example to illustrate the tightness of the restrictions on the stepsize h and the restriction on the friction parameter γ. We consider the anisotropic Gaussian distribution on $\mathbb{R}^2$ with potential $U : \mathbb{R}^2 \mapsto \mathbb{R}$ given by $U(x,y) = \frac{1}{2}mx^2 + \frac{1}{2}My^2$. This potential satisfies Assumptions 2.2.1 and 2.2.2 with constants M and m respectively. By computing the eigenvalues of the transition matrix P (for contraction) we can see for what values of h contraction occurs. For a Gaussian target, stability and asymptotic convergence speed are completely determined by the eigenvalues of P (which is constant). For EM we have that*

$$P = \begin{pmatrix} I_2 & hI_2 \\ -hQ & (1-\gamma h)\, I_2 \end{pmatrix}, \text{ where } Q = \begin{pmatrix} m & 0 \\ 0 & M \end{pmatrix},$$

with eigenvalues $\frac{1}{2}\left(2-\gamma h \pm h\sqrt{\gamma^2 - 4\lambda}\right)$, for $\lambda = m, M$. For stability and contraction, we require that

$$\lambda_{\max} := \max_{\lambda \in \{m,M\}} \left| \frac{1}{2}\left(2-\gamma h \pm h\sqrt{\gamma^2-4\lambda}\right)\right| < 1. \tag{2.4.17}$$

By Gelfand's formula, the asymptotic contraction rate exactly equals $1-\lambda_{\max}$. *For* $\gamma \geq 2\sqrt{M}$, *all 4 eigenvalues are real, and this condition requires that* $h < 4/(\gamma+\sqrt{\gamma^2-4\lambda}) \approx 2/\gamma$. *Up to a constant, this is consistent with the stepsize restriction in our contraction rate results.* $\lambda_{\max}$ *in this range of* γ *equals* $1-\frac{1}{2}h\gamma\left(1-\sqrt{1-\frac{4m}{\gamma^2}}\right)$, *so the best possible contraction rate is* $O(m/M)$ *(for the choice* $\gamma = 2\sqrt{M}$*), which is also consistent with our results.*

For $\gamma \in [2\sqrt{m}, 2\sqrt{M})$, *the absolute value of the eigenvalues are* $\sqrt{1-\gamma h+Mh^2}$, *and* $\frac{1}{2}|2-\gamma h \pm h\sqrt{\gamma^2-4m}|$, *where we need all of these to be less than 1. The first condition* $1-\gamma h+Mh^2 \leq 1$ *requires that* $h \leq \frac{\gamma}{M}$. *The second condition* $\frac{1}{2}\left|2-\gamma h-h\sqrt{\gamma^2-4m}\right| \leq 1$ *requires that* $h \leq \frac{4}{\gamma+\sqrt{\gamma^2-4m}}$. *Using our assumption that* $\gamma \in [2\sqrt{m}, 2\sqrt{M})$, $\gamma^2 \leq 4M$, *so* $\frac{4}{\gamma+\sqrt{\gamma^2-4m}} \geq \frac{2\gamma}{\gamma^2} \geq \frac{\gamma}{2M}$, *and the second condition holds whenever* $h \leq \frac{\gamma}{2M}$. *The third condition is satisfied whenever the second holds. Hence the stepsize restriction in this regime is* $\frac{\gamma}{2M}$. *One can show that the best possible convergence rate is still* $O(m/M)$.

Finally, when $\gamma < 2\sqrt{m}$, (2.4.17) *becomes equivalent to* $\sqrt{1-\gamma h+Mh^2} < 1$ *and* $\sqrt{1-\gamma h+mh^2} < 1$, *which results in the stepsize restriction* $h \leq \frac{\gamma}{M}$. *The best possible convergence rate in this regime is still* $O(m/M)$.

Example 2.4.7. *An example to illustrate the tightness of the restrictions on the stepsize* h *and the restriction on the friction parameter* γ. *We consider the anisotropic Gaussian distribution on* $\mathbb{R}^2$ *with potential* $U : \mathbb{R}^2 \mapsto \mathbb{R}$ *given by* $U(x,y) = \frac{1}{2}mx^2 + \frac{1}{2}My^2$. *By computing the eigenvalues of the transition matrix* P *(for contraction) we can see for what values of* h *contraction occurs. For BAO we have that*

$$P = \begin{pmatrix} I_2 - h^2 Q & hI_2 \\ -h\eta^2 Q & \eta^2 I_2 \end{pmatrix}, \text{ where } Q = \begin{pmatrix} m & 0 \\ 0 & M \end{pmatrix},$$

with eigenvalues $\frac{1}{2}\left(1+\eta^2-h^2\lambda \pm \sqrt{-4\eta^2+\left(-1-\eta^2+h^2\lambda\right)^2}\right)$ *for* $\lambda = m, M$, *where* $\eta = \exp\{-\gamma h/2\}$. *For stability and contraction, it is necessary and sufficient that*

$$\lambda_{\max} := \max_{\lambda \in \{m,M\}} \left|\frac{1}{2}\left(1+\eta^2-h^2\lambda \pm \sqrt{-4\eta^2+\left(-1-\eta^2+h^2\lambda\right)^2}\right)\right| < 1.$$

Due to the convexity of the absolute value function, it is clear that $|\frac{1}{2}(1+\eta^2-h^2M)| < 1$ *is necessary for the stability condition to hold, so for any value of* γ *and* m, *we need that* $h \leq \frac{2}{\sqrt{M}}$. *Theorem 2.4.5 implies that the stepsize restriction* $h \leq \frac{1}{8\sqrt{M}}$ *suffices for stability for any* $\gamma \geq 5\sqrt{M}$. *Further when* $\gamma \geq 5\sqrt{M}$ *we have that the asymptotic contraction rate for this Gaussian target simplifies to* $c_{\mathcal{N}} = \frac{1}{2}\left(1-\eta^2+h^2m-\sqrt{\left(1-\eta^2+h^2m\right)^2-4h^2m}\right)$. *It can be shown that* $4c(h) > c_{\mathcal{N}}$ *for* $\gamma \geq 5\sqrt{M}, h \leq \frac{1}{8\sqrt{M}}$. *It is shown in [116, Proposition 4] that for the continuous dynamics the condition* $\gamma > c\sqrt{M}$ *for some constant* $c > 0$ *is necessary to show contraction using a quadratic form argument similar to ours.*

Despite this fact, for Gaussian targets, faster convergence rates can be achieved for BAO in the low-friction regime ($\gamma < 5\sqrt{M}$*). In particular, when we set* $\gamma = 2\sqrt{m}$, *and use stepsize* $h = \frac{1}{\sqrt{M}}$, *with*

the notation $\rho = \frac{\sqrt{m}}{\sqrt{M}}$*, the maximum norm becomes*

$$\lambda_{\max} := \\ \frac{1}{2}\max\left(\left|e^{-2\rho} \pm \sqrt{e^{-4\rho} - 4e^{-2\rho}}\right|, \left|1 - \rho^2 + e^{-2\rho} \pm \sqrt{(1 + e^{-2\rho} - \rho^2)^2 - 4e^{-2\rho}}\right|\right).$$

It is not difficult to show with symbolic computing that for this choice of γ *and* h*, for any* $0 \leq \rho \leq 1$*, we have*

$$c_{\mathcal{N}} = 1 - \lambda_{\max} \geq \frac{3}{5}\rho = \frac{3}{5}\frac{\sqrt{m}}{\sqrt{M}}. \tag{2.4.18}$$

This is an accelerated convergence rate that is faster than what we could prove for general strongly convex and smooth potentials in Theorem 2.4.5.

Considering other splittings one could use the same techniques as above or we can use the contraction results of BAO and OAB to achieve a contraction result for the remaining permutations by writing $(\mathcal{ABO})^n = \mathcal{AB}(\mathcal{OAB})^{n-1}\mathcal{O}, (\mathcal{BOA})^n = \mathcal{B}(\mathcal{OAB})^{n-1}\mathcal{OA}, (\mathcal{OBA})^n = \mathcal{O}(\mathcal{BAO})^{n-1}\mathcal{BA}$, and $(\mathcal{AOB})^n = \mathcal{AO}(\mathcal{BAO})^{n-1}\mathcal{B}$. However, by applying direct arguments as done for OAB and BAO one would achieve better preconstants. Let $(\tilde{x}_0, \tilde{v}_0) \in \mathbb{R}^{2d}$ and $(x_0, v_0) \in \mathbb{R}^{2d}$ be two initial conditions for a synchronous coupling of sample paths of the ABO splitting and $\overline{x}_0 := \tilde{x}_0 - x_0$, $\overline{v}_0 := \tilde{v}_0 - v_0$. We use the notation ψ_{ABO} to denote the one-step map of the ABO discretization with stepsize $h > 0$, and equivalently for other operators (omitting the stepsize in the argument, which is $h > 0$ for all the one-step maps considered). We have that for $h < \min\{\frac{1}{4\gamma}, \frac{1-\eta^2}{\sqrt{6M}}\}$

$$\begin{aligned}
&||\left(\tilde{x}_n, \tilde{v}_n\right) - \left(x_n, v_n\right)||_{a,b}^2 = ||\left(\psi_{\mathrm{ABO}}\right)^n\left(\tilde{x}_0, \tilde{v}_0\right) - \left(\psi_{\mathrm{ABO}}\right)^n\left(x_0, v_0\right)||_{a,b}^2 \\
&= ||\psi_{\mathrm{O}} \circ \left(\psi_{\mathrm{OAB}}\right)^{n-1} \circ \psi_{\mathrm{AB}}\left(\tilde{x}_0, \tilde{v}_0\right) - \psi_{\mathrm{O}} \circ \left(\psi_{\mathrm{OAB}}\right)^{n-1} \circ \psi_{\mathrm{AB}}\left(x_0, v_0\right)||_{a,b}^2 \\
&\leq 3\left(1 - c\left(h\right)\right)^{n-1}||\psi_{\mathrm{AB}}\left(\tilde{x}_0, \tilde{v}_0\right) - \psi_{\mathrm{AB}}\left(x_0, v_0\right)||_{a,b}^2 \\
&\leq 9\left(1 - c\left(h\right)\right)^{n-1}\left(\left(1 + 2h^2M^2a\right)||\overline{x}_0||^2 + \left(h^2 + a + 2h^4M^2a\right)||\overline{v}_0||^2\right) \\
&\leq 27\left(1 - c\left(h\right)\right)^{n-1}||\left(\overline{x}_0, \overline{v}_0\right)||_{a,b}^2,
\end{aligned}$$

where we have used the norm equivalence introduced in Section 2.2.2. The same method of argument can be used for the other first-order splittings.

2.5 Overdamped limit

We will now compare and analyze how the different schemes behave in the high-friction limit, where we first start with the first-order schemes. It is a desirable property that the high-friction limit is a discretization of the overdamped dynamics, therefore if a user of such a scheme sets the friction parameter γ large, they will not suffer from the $\mathcal{O}(1/\gamma)$ scaling of the convergence rate. We will call schemes with this desirable property γ-limit convergent (GLC), out of the schemes we have analysed it is only BAOAB and OBABO which are GLC.

BAO

If we consider the update rule of the BAO scheme

$$x_{n+1} = x_n + h\left(v_n - h\nabla U(x_n)\right), \quad v_{n+1} = \eta^2 v_n - h\eta^2 \nabla U(x_n) + \sqrt{1-\eta^4}\xi_{n+1},$$

and take the limit as $\gamma \to \infty$ we obtain

$$x_{n+1} = x_n - h^2 \nabla U(x_n) + h\xi_n,$$

which is simply the Euler-Maruyama scheme with stepsize $h^2/2$ for potential $\tilde{U} := 2U$, which imposes stepsize restrictions which are consistent with our analysis. Further, if we take the limit of the contraction rate and the modified Euclidean norm we have

$$\lim_{\gamma\to\infty} c(h) = \frac{h^2 m}{4}, \qquad \lim_{\gamma\to\infty} ||x||^2 + 2b\langle x, v\rangle + a||v||^2 = ||x||^2 + 2h\langle x, v\rangle + \frac{1}{M}||v||^2,$$

which is again consistent with the convergence rates achieved in Section 2.3.1 and the norm is essentially the Euclidean norm when considered on the overdamped process as $\overline{v} = 0$. Due to the fact that the potential is rescaled in the limit, this is not a discretization of the correct overdamped dynamics.

OAB

If we consider the update rule of the OAB scheme

$$\begin{aligned} x_{n+1} &= x_n + h\eta^2 v_n + h\sqrt{1-\eta^4}\xi_{n+1}, \\ v_{n+1} &= \eta v_n \sqrt{1-\eta^4}\xi_{n+1} - h\eta^2 \nabla U(x_n + h\eta^2 v_n + h\sqrt{1-\eta^4}\xi_{n+1}), \end{aligned}$$

and take the limit as $\gamma \to \infty$ we obtain the update rule $x_{n+1} = x_n + h\xi_{n+1}$, therefore the overdamped limit is not inherited by the scheme and further we do not expect contraction. This is consistent with our analysis of OAB and our contraction rate which vanishes in the high-friction limit.

BAOAB

If we consider the update rule of the BAOAB scheme

$$\begin{aligned} x_{n+1} &= x_n + \frac{h}{2}\left(1+\eta^2\right) v_n - \frac{h^2}{4}\left(1+\eta^2\right)\nabla U(x_n) + \frac{h}{2}\sqrt{1-\eta^4}\xi_{n+1}, \\ v_{n+1} &= \eta^2\left(v_n - \frac{h}{2}\nabla U(x_n)\right) + \sqrt{1-\eta^4}\xi_{n+1} - \frac{h}{2}\nabla U(x_{n+1}), \end{aligned}$$

and take the limit as $\gamma \to \infty$ we obtain

$$x_{n+1} = x_n - \frac{h^2}{2}\nabla U(x_n) + \frac{h}{2}\left(\xi_n + \xi_{n+1}\right),$$

which is simply the LM scheme with stepsize $h^2/2$ (as originally noted in [98]), which imposes stepsize restrictions $h^2 \leq 2/M$ and hence consistent with our analysis. Further, if we take the limit of the contraction rate and the modified Euclidean norm we have

$$\lim_{\gamma\to\infty} c\left(h\right) = \frac{h^2 m}{4}, \qquad \lim_{\gamma\to\infty} ||x||^2 + 2b\langle x, v\rangle + a||v||^2 = ||x||^2 + 2h\langle x, v\rangle + \frac{1}{M}||v||^2,$$

which is again consistent with the convergence rates achieved in Section 2.3.1 and the modified Euclidean norm is essentially the Euclidean norm when considered on the overdamped process as $\overline{v} = 0$.

OBABO

If we consider the update rule of the OBABO scheme

$$\begin{aligned} x_{n+1} &= x_n + h\eta v_n + h\sqrt{1-\eta^2}\xi_{1,n+1} - \frac{h^2}{2}\nabla U(x_n), \\ v_{n+1} &= \eta\left(\eta v + \sqrt{1-\eta^2}\xi_{1,n+1} - \frac{h}{2}\nabla U(x_n) - \frac{h}{2}\nabla U(x_{n+1})\right) + \sqrt{1-\eta^2}\xi_{2,n+1}, \end{aligned}$$

where ($\eta = \exp\left(-\gamma h/2\right)$) and for ease of notation in the above scheme and we have labelled the two noises of one step ξ_1 and ξ_2. Now we take the limit as $\gamma \to \infty$ we obtain

$$x_{n+1} = x_n - \frac{h^2}{2}\nabla U(x_n) + h\xi_{n+1},$$

which is the Euler-Maruyama scheme for overdamped Langevin with stepsize $h^2/2$, which has convergence rate $\mathcal{O}\left(h^2 m\right)$. Hence consistent with our analysis of OBABO and our contraction rate which tends towards $h^2 m/4$ in the high-friction limit.

SES

If we consider the limit as $\gamma \to \infty$ of the scheme (2.4.10) we obtain the update rule $x_{n+1} = x_n$ and therefore the overdamped limit is not inherited by the scheme and further we do not expect contraction. Hence consistent with our analysis of the stochastic Euler scheme as the contraction rate tends to zero in the high-friction limit.

BBK Integrator

Taking the limit as $\gamma \to \infty$ and by considering two consecutive iterations ($v_{k+1} = -v_k$ in this limit) one arrives at the following update rule

$$x_{k+2} = x_k - \frac{h^2}{2}\left(\nabla U(x_{k+1}) + \nabla U(x_k)\right) + \frac{\sqrt{2\gamma}h^{3/2}}{2}\left(\xi_k + \xi_{k+1}\right),$$

and hence the method is not GLC as this does not converge to overdamped dynamics as the stepsize is taken to zero.

Stochastic position Verlet and stochastic velocity Verlet

If one takes the limit as $\gamma \to \infty$ for the stochastic position and velocity Verlet then we get the operators

$$\mathcal{V}(h) : v \to \xi,$$
$$\mathcal{A}(h) : x \to x + hv,$$

hence these schemes do not converge to the overdamped dynamics as one takes the stepsize to zero.

rOABAO

The rOABAO scheme is GLC and, interestingly, by taking the high friction limit one arrives at the scheme

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k + u\xi_k) + h\xi_k,$$

where $u \sim \mathcal{U}(0, h)$, which has the correct invariant measure and is a randomized midpoint version of the Euler-Maruyama scheme for overdamped Langevin dynamics and the one-step HMC scheme of [26].

2.6 Stochastic gradients

An analysis of convergence rates of the discretizations with stochastic gradients is performed in [80].

Definition 2.6.1. *A stochastic gradient approximation of a potential U is defined by a function $\mathcal{G} : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ and a probability distribution ρ on a Polish space Ω, satisfying that $\mathcal{G}$ is measurable on $(\Omega, \mathcal{F})$, and that for every $x \in \mathbb{R}^d$, for $W \sim \rho$,*

$$\mathbb{E}(\mathcal{G}(x, W)) = \nabla U(x).$$

The function $\mathcal{G}$ and the distribution ρ together define the stochastic gradient, which we denote as $(\mathcal{G}, \rho)$.

The numerical schemes considered in this chapter are roughly one gradient evaluation per sample, roughly meaning when negating the extra gradient evaluations at the head and tail of the simulation of the algorithm (for the first and last sample). This is done by using the same gradient evaluation in consecutive velocity updates when the position has not been updated, for an increase in computational efficiency. We treat this case when it also comes to stochastic gradients to improve computational efficiency, for example in the BAOAB scheme the last B and first B of each iteration will share an estimate of the force (using the same stochastic gradient evaluation). For clarity, a stochastic gradient version of each algorithm is provided in Appendix A.3.

In our convergence rate estimates, we impose the assumption that the variance of the Jacobian of the stochastic gradient is bounded.

Assumption 2.6.2. *We assume that the Jacobian of the stochastic gradient $\mathcal{G}$, $D_x\mathcal{G}(x,W)$ exists and it is measurable on $(\Omega,\mathcal{F})$. We also assume there exists $C_G>0$ such that for $W\sim\rho$,*

$$\sup_{x\in\mathbb{R}^d}\mathbb{E}\|D_x\mathcal{G}(x,W)-\nabla^2U(x)\|^2\leq C_G.$$

Our results extend to the stochastic gradient setting by including a coupling in the mini-batches or the stochastic gradients in the same way as OABAO in [80]. Further, the results for the other schemes in this chapter are generalized to the case of stochastic gradients when the same stochastic gradient is chosen as in Appendix A.3 for each algorithm. In this way, there is still one gradient evaluation per step.

We remark that these assumptions hold when $\mathcal{G}$ is of the form $\mathcal{G}(x,W)=\sum_{i\in W}\nabla U_i(x)$, where $W\in\Omega\subset[N_D]^{N_b}$, N_b is the batch size, and $(U_i)_{i\in[N_D]}$ are strongly convex and gradient-Lipschitz, this is the setting of minibatching in many Bayesian learning problems. The contraction results of Theorem 2.4.5 are extended to the stochastic gradient setting in Theorem 2.6.3. We remark that our assumptions are more flexible than the assumptions imposed in [80], where they assume that the stochastic gradient is universally gradient Lipschitz and strongly convex over the entire state space Ω.

Theorem 2.6.3. *Consider the numerical schemes for stochastic gradient kinetic Langevin dynamics given in Appendix A.3 and Table 2.3, where the potential U is m-strongly convex and M-∇Lipschitz. Assume a stochastic gradient approximation defined by $(\mathcal{G},\rho)$ (see Definition 2.6.1) satisfying Assumption 2.6.2 with constant C_G. We consider any sequence of synchronously coupled random variables (in Brownian increment and stochastic gradient) with initial conditions $(x_0,v_0)\in\mathbb{R}^{2d}$ and $(\tilde{x}_0,\tilde{v}_0)\in\mathbb{R}^{2d}$.*

Algorithm	c(h)	C(h)
EM	$mh/2\gamma-2h^2C_G/M$	1
BBK	$mh/4\gamma-4h^2C_G/M$	$7+3h^2C_G/M$
SPV	$mh/4\gamma-4h^2C_G/M$	$7+12h^2C_G/M$
SVV	$mh/4\gamma-4h^2C_G/M$	$7+6h^2C_G/M$
BAOAB	$h^2m/4(1-\eta^2)-5h^2C_G\left(\eta^2/M+\frac{1}{4}h^2\right)$	$7+3h^2C_G/M$
OBABO	$h^2m/4(1-\eta^2)-4h^2C_G/M$	$8+3h^2C_G/M$
rOABAO	$h^2m/4(1-\eta^2)-5h^2C_G\left(\eta^2/M+\frac{1}{4}h^2\right)$	$8+8h^2C_G/M$
SES/EB	$mh/4\gamma-4h^2C_G/M$	1

Table 2.3: Contraction rates $c(h)$ and preconstants $C(h)$ in (2.6.19).

Under stepsize restrictions $h<h_0$ and $\gamma\geq\gamma_0$, where h_0 and γ_0 are given in Table 2.2, and given initial conditions $(x_0,v_0)\in\mathbb{R}^{2d}$ and $(\tilde{x}_0,\tilde{v}_0)\in\mathbb{R}^{2d}$ we have the expected contraction

$$\left(\mathbb{E}\|(x_k-\tilde{x}_k,v_k-\tilde{v}_k)\|_{a,b}^2\right)^{1/2}\leq C(h)(1-c(h))^{s/2}\|(x_0-\tilde{x}_0,v_0-\tilde{v}_0)\|_{a,b},\tag{2.6.19}$$

with norm given with constants $a=1/M$ and b, where the contraction rate $c(h)$ and the preconstant $C(h)$ are given in Table 2.3 and b and the number of steps s are given in Table 2.2 with all parameters specific to each scheme.

Remark 2.6.4. *Compared to Theorem 2.4.5 with deterministic gradients, Theorem 2.6.3 demonstrates expected contraction, because the randomness from the stochastic gradients can be integrated out. This allows us to make Assumption 2.6.2 less restrictive than it would need to be otherwise. Rather than deterministic contraction we have contraction in expectation.*

Remark 2.6.5. *Our analysis suggests a reduction in the convergence rate for large gradient noises, which we have observed in numerical experiments when using sub-sampling and very small batches. For large gradient noise C_G and stepsize h it is possible that these bounds become vacuous and the loss of convergence was also confirmed in our experiments.*

Remark 2.6.6. *The implementation of the BAOAB algorithm and other algorithms considered in Section A.3 is non-Markovian, because the last B step of each iteration and the first B step of the next iteration share the same stochastic gradient sample. This is not an issue in our convergence rate framework as we consider convergence of a different operator, which is Markovian, for example $\mathcal{ABAO}$ for BAOAB, which does not share stochastic gradients with consecutive iterations. We simplify the problem into proving convergence of an operator which only has a single gradient evaluation and hence is Markovian in the stochastic gradient setting.*

Proof of Theorem 2.6.3. For stochastic gradients, we synchronously couple Brownian increments as well as the stochastic gradients. We wish to instead consider expected contraction of the update rule we used to prove contraction in the full gradient setting, i.e. for synchronously coupled (in stochastic gradient and Brownian increment) iterates $(x_l, v_l), (\tilde{x}_l, \tilde{v}_l) \in \mathbb{R}^{2d}$ for $l \in \mathbb{N}$ and $(\overline{x}_l, \overline{v}_l) = (\tilde{x}_l, \tilde{v}_l) - (x_l, v_l)$ and for $k \in \mathbb{N}$

$$\mathbb{E}\|(\overline{x}_{k+1}, \overline{v}_{k+1})\|_{a,b}^2 \leq (1 - c(h)) \, \|(\overline{x}_k, \overline{v}_k)\|_{a,b}^2$$

then we have

$$\mathbb{E}\left(\overline{z}_k^T P^T M P \overline{z}_k\right) \leq (1 - c(h)) \, \overline{z}_k^T M \overline{z}_k.$$

Now if $\tilde{Q}$ is defined through the mean value theorem of $D_x\mathcal{G}$ (the Jacobian of $\mathcal{G}$) and is a random variable in W, such that $\mathbb{E}(\tilde{Q}) = Q$, then $P^T M P$ is of the form

$$\mathcal{P}(\tilde{Q}) = \begin{pmatrix} P_1(\tilde{Q}) & P_2(\tilde{Q}) \\ P_2(\tilde{Q}) & P_3(\tilde{Q}) \end{pmatrix},$$

where P_1, P_2 and P_3 are quadratics in $\tilde{Q}$ of the form

$$\begin{aligned} P_1(\tilde{Q}) &= a_0 + a_1\tilde{Q} + a_2\tilde{Q}^2, \\ P_2(\tilde{Q}) &= b_0 + b_1\tilde{Q} + b_2\tilde{Q}^2, \\ P_3(\tilde{Q}) &= c_0 + c_1\tilde{Q} + c_2\tilde{Q}^2. \end{aligned}$$

Then we have

$$\mathbb{E}(P^T M P) = \mathcal{P}(Q) + \begin{pmatrix} a_2\mathbb{E}(\tilde{Q} - Q)^2 & b_2\mathbb{E}(\tilde{Q} - Q)^2 \\ b_2\mathbb{E}(\tilde{Q} - Q)^2 & c_2\mathbb{E}(\tilde{Q} - Q)^2 \end{pmatrix},$$

in combination with the Theorem 2.4.5 result we have that

$$\begin{aligned}\mathbb{E}\|(x_{k+1},v_{k+1})\|_{a,b}^2 &\leq (1-c(h))\,\|(x_k,v_k)\|_{a,b}^2 + z_k^T\begin{pmatrix} a_2\mathbb{E}(\tilde{Q}-Q)^2 & b_2\mathbb{E}(\tilde{Q}-Q)^2 \\ b_2\mathbb{E}(\tilde{Q}-Q)^2 & c_2\mathbb{E}(\tilde{Q}-Q)^2\end{pmatrix} z_k \\ &= (1-c(h))\,\|(x_k,v_k)\|_{a,b}^2 + z_k^T\mathcal{R}(\tilde{Q})z_k,\end{aligned}$$

where we use the notation

$$\mathcal{R}(\tilde{Q}) := \begin{pmatrix} a_2\mathbb{E}(\tilde{Q}-Q)^2 & b_2\mathbb{E}(\tilde{Q}-Q)^2 \\ b_2\mathbb{E}(\tilde{Q}-Q)^2 & c_2\mathbb{E}(\tilde{Q}-Q)^2\end{pmatrix}.$$

Then we will bound the remainder term $z^T\mathcal{R}(\tilde{Q})z$ separately for each scheme and we refer the reader to the contraction estimate proofs in Appendix A.1 and [105] for the coefficients a_2, b_2 and c_2 for Euler-Maruyama, BAOAB, OBABO and SES and to Appendix A.1 and [103] for the schemes analyzed in this chapter. We remark that we analyze the update rules for which we proved contraction for all the schemes, which aren't necessarily the same as the scheme for example we analyze $\mathcal{ABAO}$ for BAOAB. Throughout these estimates we use the equivalence of norms in Section 2.2.2 and the stepsize and parameter restrictions imposed in the contraction estimates of the respective schemes. We define $\mathsf{Var}(\tilde{Q}) := \mathbb{E}(\tilde{Q}-Q)^2$, to be the variance of $\tilde{Q}$.

1. For the Euler-Maruyama $a_2 > 0$ and $b_2 = c_2 = 0$, therefore we have for $z = (x,v) \in \mathbb{R}^{2d}$

$$\begin{aligned}z^T\mathcal{R}(\tilde{Q})z &\leq h^2 a C_G\|x\|^2 \\ &\leq 2h^2 a C_G\|z\|_{a,b}^2.\end{aligned}$$

2. For BAOAB we have for $\mathcal{ABAO}$

$$\begin{aligned}z^T\mathcal{R}(\tilde{Q})z &= ah^2\left(\eta^4 + b\eta^2 hM + \frac{h^2}{4}M\right)\left(x+\frac{h}{2}v\right)^T\mathsf{Var}(\tilde{Q})\left(x+\frac{h}{2}v\right) \\ &\leq 4ah^2C_G\left(\eta^4 + b\eta^2 hM + \frac{h^2}{4}M\right)\|(x,v)\|_{a,b}^2 \\ &\leq 5ah^2C_G\left(\eta^2 + \frac{h^2}{4}M\right)\|(x,v)\|_{a,b}^2.\end{aligned}$$

For the other schemes they follow similarly and can be found in [103]. We have all desired preconstants and penalty terms for the contraction rate when the gradient is a stochastic estimate.

□

Proposition 2.6.7. *Consider the numerical schemes for stochastic gradient kinetic Langevin dynamics given in Appendix A.3 and Table 2.3, where the potential U is m-strongly convex and M-∇Lipschitz. Assume a stochastic gradient approximation defined by $(\mathcal{G},\rho)$ (see Definition 2.6.1) satisfying Assumption 2.6.2 with constant C_G. We use P_h to denote the marginal transition kernel of the numerical schemes. For the constants given in Table 2.2 and 2.3 we have for any two synchronously coupled chains, (x_k,v_k) and $(\tilde{x}_k,\tilde{v}_k)$ under the assumptions specific to the schemes imposed in Theorem 2.6.3*

we have for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^{2d})$*, and all* $k \in \mathbb{N}$*,*

$$\mathcal{W}_2^2\left(\nu P_h^k, \mu P_h^k\right) \leq 3C(h)\max\left\{M, \frac{1}{M}\right\}(1-c\,(h))^k\,\mathcal{W}_2^2\,(\nu,\mu)\,.$$

Proof. We remark that the stochastic gradients are independent from position and hence can be marginalized out in the following estimates over the extended state space. We first denote $\hat{P}_h$ to be the transition kernel for which contraction is proved in Theorem 2.4.5 or [105]. For example stochastic gradient $\mathcal{ABAO}$ for BAOAB. From Theorem 2.6.3 and following [115, Corollary 20] we know that for $z_k = (x_k, v_k)$, $\tilde{z}_k = (\tilde{x}_k, \tilde{v}_k)$ such that $z_0 = (x_0, v_0) \sim \mu$ and $\tilde{z}_0 = (\tilde{x}_0, \tilde{v}_0) \sim \nu$ and $(z_0, \tilde{z}_0)$ is a $\mathcal{W}_2$ optimal coupling of μ and ν then under $\hat{P}_h$

$$\mathcal{W}_{2,a,b}^2\left(\mu\hat{P}_h^k, \nu\hat{P}_h^k\right) \leq \mathbb{E}\|z_k - \tilde{z}_k\|_{a,b}^2 \leq (1-c(h))^k\mathcal{W}_{2,a,b}^2\,(\mu,\nu)\,,$$

then we can use the equivalence of norms in Section 2.2.2 and the preconstant estimates of Table 2.3 to achieve the desired result for P_h. □

Remark 2.6.8. *We remark that the contraction rate of BAOAB and rOABAO can be upper bounded by a simpler form, for example,* $\mathcal{O}\left(mh^2/(1-\eta^2) - h^2C_G/M\right)$*, but we have included the more detailed estimate because it has the property that as you take the friction parameter* $\gamma \to \infty$ *then the contraction rate is of the same order as the overdamped Langevin dynamics scheme as discussed in [103].*

If we take the limit as $\gamma \to \infty$ *for the BAOAB and rOABAO scheme we get a contribution from the stochastic gradient of* $\mathcal{O}\left(h^4C_G\right)$ *in the convergence rate estimate, and for the overdamped analysis in [103] we have a contribution of* $\mathcal{O}\left(h^4C_G\right)$ *in the high friction limit of BAOAB and rOABAO. However, for OBABO we get a contribution of* $\mathcal{O}\left(h^2C_G/M\right)$*, which agrees with the overdamped Langevin analysis for the largest choice of stepsize.*

2.7 Asymptotic bias of BAOAB

There are results for the asymptotic bias of OBABO available in [80, 115] and for the SES in [140] which can easily be combined with our results. However there are no results for BAOAB available in the literature, we will provide asymptotic bias estimates for this scheme. We will do this with the aim of achieving bias estimates which remain finite in the high-friction limit to show that BAOAB and GLC schemes remain useful for sampling even though they deviate from the continuous dynamics.

We define the solution map $\mathcal{H}$ to have update rule

$$\mathcal{H} : (x, v) \to \phi_h(x, v), \tag{2.7.20}$$

where $\phi_h(x, v)$ is the solution to the ODE

$$dX_t = V_t dt, \qquad dV_t = -\nabla U(X_t)dt,$$

initialized at $(X_0, V_0) := (x, v) \in \mathbb{R}^{2d}$ at time $h > 0$. Since the BAOAB scheme converges faster than the continuous dynamics, if we compare BAOAB to the underdamped Langevin dynamics in the high-friction regime, we will get discretization bounds which diverge as $\gamma \to \infty$ and the stepsize is kept constant. We instead compare BAOAB to a scheme which performs exact Hamiltonian dynamics for half a step, followed by an OU-process for a full step, followed by Hamiltonian dynamics for half a step. We call this scheme the HOH scheme with the update rule given by $\mathcal{HOH}$ with $\mathcal{H}$ defined in (2.7.20). This process exactly preserves the invariant measure and is a more accurate approximation of the BAOAB scheme than (1.3.7).

Remark 2.7.1. *In [115] and [80], the authors provide bias estimates for the OBABO and OABAO schemes. In their analysis of these schemes, they use the fact that the "O" step preserves the invariant measure and doesn't increase Wasserstein distance. They are then able to exploit L^2-accuracy results of the embedded Verlet integrators BAB and ABA in their analysis.*

Proposition 2.7.2. *Consider an HOH scheme initialized at $(x, v) \in \mathbb{R}^{2d}$ and a BAOAB scheme initialized at $(x', v') \in \mathbb{R}^{2d}$ with synchronously coupled Gaussian increments and stepsize $h > 0$, then we define $(\triangle_x, \triangle_v) := \psi_{\mathsf{HOH}}(x, v, h) - \psi_{\mathsf{BAOAB}}(x', v', h)$ with shared noise $\xi \sim \mathcal{N}(0_d, I_d)$. We assume that $h < \frac{1-\eta^2}{2\sqrt{M}}$ and Assumptions 2.2.1 - 2.2.2 on the potential, then we have that*

$$\|\triangle_x\|_{L^2} \leq \left(1 + (1+\eta^2)\frac{h^2}{4}M\right)\|x - x'\|_{L^2} + \frac{h}{2}(1+\eta^2)\|v - v'\|_{L^2} + \frac{3h^3 M\sqrt{d}}{8},$$
$$\triangle_v = \left(\eta^2 I_d - \frac{h^2(1+\eta^2)}{4}Q_2\right)(v - v') + \left(-\frac{h\eta^2}{2}Q_1 - \frac{h}{2}Q_2 + \frac{h^3(1+\eta^2)}{8}Q_2Q_1\right)(x - x')$$
$$+ \epsilon_v,$$

where $\|\epsilon_v\|_{L^2} \leq 2h^2 M\sqrt{d}$. Further if we assume Assumption 2.2.3 we have

$$\triangle_v = \left(\eta^2 I_d - \frac{h^2(1+\eta^2)}{4}Q_2\right)(v - v') + \left(-\frac{h\eta^2}{2}Q_1 - \frac{h}{2}Q_2 + \frac{h^3(1+\eta^2)}{8}Q_2Q_1\right)(x - x')$$
$$+ \epsilon_v + h^2\sqrt{1-\eta^4}A(x, x')\xi,$$

where $mI_d \prec Q_1, Q_2 \prec MI_d$, $\epsilon_v \in \mathbb{R}^{2d}$, $A(x, x') \in \mathbb{R}^{d\times d}$ and $\|\cdot\|_{L^2} := (\mathbb{E}\|\cdot\|^2)^{1/2}$, with $\|\epsilon_v\|_{L^2} \leq 2h^3\sqrt{d}(M^{3/2} + M_1\sqrt{d})$. Further we have that $\|A(x, x')\xi\|_{L^2} \leq \frac{3}{8}M\sqrt{d}$.

Proof. The proof is found in Appendix A.2. □

Proposition 2.7.3. *Consider an HOH scheme, $(x_i, v_i)_{i\in\mathbb{N}}$ and a BAOAB scheme $(x'_i, v'_i)_{i\in\mathbb{N}}$ initialized at $(x_0, v_0) = (x'_0, v'_0) = (x, v) \sim \pi$ in $\mathbb{R}^{2d}$ with synchronously coupled Gaussian increments and stepsize $h < \frac{1-\eta^2}{2\sqrt{M}}$, for $l \in \mathbb{N}$ we define $(\triangle_x^l, \triangle_v^l) := (x_l - x'_l, v_l - v'_l)$. For $a = 1/M$ and $b = h/(1-\eta^2)$ we have that under Assumptions 2.2.1-2.2.2, for any $l \geq 1$,*

$$\|(\triangle_x^l, \triangle_v^l)\|_{L^2,a,b} \leq 4400\frac{M}{m}\sqrt{d}h.$$

Additionally, if Assumption 2.2.3 is satisfied, we have for any $l \geq 1$,

$$\|(\triangle_x^l, \triangle_v^l)\|_{L^2,a,b} \leq 1500\frac{\sqrt{M}}{m}\left(4\sqrt{Md} + 3\frac{M_1}{M}d\right)h(1-\eta^2).$$

Proof. The proof is found in Appendix A.2. □

Remark 2.7.4. *In Proposition 2.7.3 we provide L^2 error estimates over $l \in \mathbb{N}$ steps, this allows one to go from order $h^{5/2}$ local error to order h^2 global error, as the $h^{5/2}$ term in the local error estimate is due to independent Gaussian increments. Similarly, the strong convergence of numerical solutions of SDEs only loses an order of $1/2$ accuracy (see [114, Theorem 1.1.1]), for example, the Euler-Maruyama scheme or the UBU scheme in [140].*

Theorem 2.7.5. *Consider a BAOAB scheme with stepsize $h < \frac{1-\eta^2}{2\sqrt{M}}$ and invariant measure π_h on $\mathbb{R}^{2d}$, with target measure π on $\mathbb{R}^{2d}$. For $a = 1/M$ and $b = h/(1-\eta^2)$ we have that under Assumptions 2.2.1-2.2.2*

$$\mathcal{W}_{2,a,b}(\pi, \pi_h) \leq 4400 \frac{M}{m} \sqrt{d} h.$$

Additionally, if Assumption 2.2.3 is satisfied we have

$$\mathcal{W}_{2,a,b}(\pi, \pi_h) \leq 1500 \frac{\sqrt{M}}{m} \left(4\sqrt{Md} + 3\frac{M_1}{M} d \right) h(1-\eta^2).$$

Proof. Let P_h denote the transition kernel of the BAOAB scheme, which has invariant measure π_h. Then we have that for $l \in \mathbb{N}$

$$\mathcal{W}_{2,a,b}(\pi, \pi_h) \leq \mathcal{W}_{2,a,b}(\pi P_h^l, \pi_h) + \mathcal{W}_{2,a,b}(\pi P_h^l, \pi).$$

We now estimate each of the terms on the right-hand side separately. We have by Theorem 2.4.5

$$\mathcal{W}_{2,a,b}(\pi P_h^l, \pi_h) = \mathcal{W}_{2,a,b}(\pi P_h^l, \pi_h P_h^l) \leq 7(1 - c(h))^{(l-1)/2} \mathcal{W}_{2,a,b}(\pi, \pi_h).$$

This term tends to zero as l tends to infinity. By Proposition 2.7.3, and the definition of the Wasserstein distance, we can bound the term $\mathcal{W}_{2,a,b}(\pi P_h^l, \pi)$ uniformly for any l, hence our claims follow. □

Note that the constraints on the stepsize and friction parameter are the same as for our contraction result i.e. close to the stability threshold. A consequence of the asymptotic bias results together with the contraction results is that when the potential satisfies Assumptions 2.2.1-2.2.2 the BAOAB scheme requires $\mathcal{O}(d^{1/2}/\epsilon)$ steps to reach an accuracy $\epsilon > 0$ in Wasserstein distance from the target. If the potential further satisfies Assumption 2.2.3 this can be improved to $\mathcal{O}(d^{1/2}/\epsilon^{1/2})$ steps to reach an accuracy $\epsilon > 0$ in Wasserstein distance from the target. Perhaps under stronger assumptions like the strongly Hessian Lipschitz assumption of [45], the dimension dependency can be improved. We remark that the strongly Hessian Lipschitz assumption of [45] implies Assumption 2.2.3.

2.8 Numerical experiments

To quantify and validate our convergence results and contraction rates we approximate the spectral gap of the numerical scheme $c(h)$ for an Anisotropic Gaussian example. We then compare this to the continuous dynamics via

$$\frac{1-c(h)}{h},$$

which converges to the spectral gap of the continuous dynamics as $h \to 0$, and is normalized by stepsize. We also compare the bias of the numerical integrators in a Bayesian classification application.

2.8.1 Anisotropic Gaussian

We first consider a simple low-dimensional example to compare the convergence rates, the anisotropic Gaussian distribution on $\mathbb{R}^2$ with potential $U : \mathbb{R}^2 \mapsto \mathbb{R}$ given by $U(x,y) = \frac{1}{2}mx^2 + \frac{1}{2}My^2$. This potential satisfies Assumption 2.2.1 with constants M and m respectively. For this example, we can analytically solve for the contraction rates, which coincide with the convergence rates of $\mathbb{E}(X_n)$. We can do this by computing the spectral gap of the transition matrix P, by $1 - |\lambda_{\max}|$, where $\lambda_{\max}$ is the largest eigenvalue of the matrix P due to Gelfand's formula. This converges to the spectral gap of the continuous dynamics as $h \to 0$.

The dependence of the convergence rate on the friction parameter γ is given in Figure 2.1. We will study how this changes for the discretisations with contour plots of stepsize versus contraction rate for all the numerical methods we consider. If we take a slice of our contour plots for small stepsizes then this will coincide with Figure 2.1. This is given in Figure 2.2.

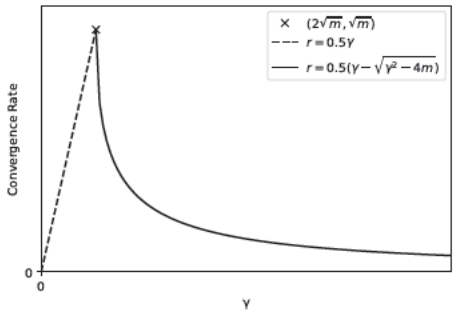
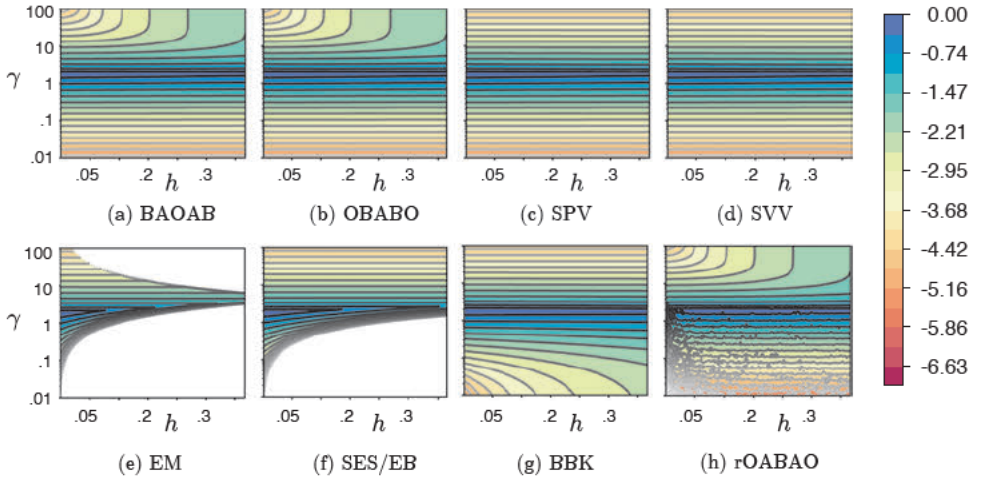


Figure 2.1: Contraction rate of continuous kinetic Langevin dynamics for an anisotropic Gaussian with parameters m and M.

Due to the fact that each update matrix P for the anisotropic Gaussian using the rOABAO scheme is in fact a random matrix, we estimate the contraction rate using [75, 91], where

$$\lim_{N\to\infty} \log \|P_1 P_2 ... P_N\|/N \to \log(1 - c(h)),$$

where P_i for $i \in \mathbb{N}$ is the transition matrix of the i^{th} iteration. We approximate this limit by Monte Carlo simulations with a random $u \sim [0, h]$ from the randomized midpoint at each stage to approximate the spectral radius.



(a) BAOAB (b) OBABO (c) SPV (d) SVV

(e) EM (f) SES/EB (g) BBK (h) rOABAO

Figure 2.2: Contour plots of $\ln\left(\frac{1-c(h)}{h}\right)$ for various schemes in the case of an anisotropic Gaussian with parameters $m = 1$ and $M = 10$. Regions of white indicate instability. The rOABAO contour plot is approximate and all other plots are exact (analytic).

Figure 2.2 illustrates the exact synchronously coupled contraction rates for all the numerical integrators we consider (apart from for rOABAO, which is an approximate Monte Carlo estimation) for a range of stepsizes h and friction parameters γ. BAOAB, OBABO, rOABAO fail to approximate the true kinetic Langevin dynamics for large stepsizes, but still have low bias in the invariant measure as they act like overdamped Langevin dynamics. The γ-limit convergent property is reflected in Figure 2.2 for large γ as BAOAB, OBABO and rOABAO have large contraction rates for large values of the stepsize and no longer scale with $1/\gamma$, like the other schemes. SVV, SPV and BBK remain stable, but have convergence rates which scale with $1/\gamma$, indicated by the parallel contour lines in Figure 2.2. The SES and EM methods have large regions of instability, SES being unstable for small values of the friction parameter when h scales larger than γ.

We have only illustrated convergence results towards the invariant measure, there has been work which provides Wasserstein bias estimates for a few of the numerical methods explored (see [80, 115, 140]). Although the focus of this chapter is to provide convergence rate estimates, we will provide a comparative numerical study of the bias of each of these numerical methods for some choices of the friction parameter for an application in the following section.

2.8.2 Bayesian logistic regression on MNIST

We next consider a more involved example, which has a ∇-Lipschitz and convex potential. This is a Bayesian posterior sampling application in multinomial logistic regression using the MNIST machine learning data set [96]. The data set contains $60,000$ training data points and $10,000$ test data points. The images are of size 28 by 28 pixels and hence can be represented in $\mathbb{R}^{784}$. However, we will consider the reduced problem of classifying digits 3 and 5. Sample images are shown in Figure 3.7.

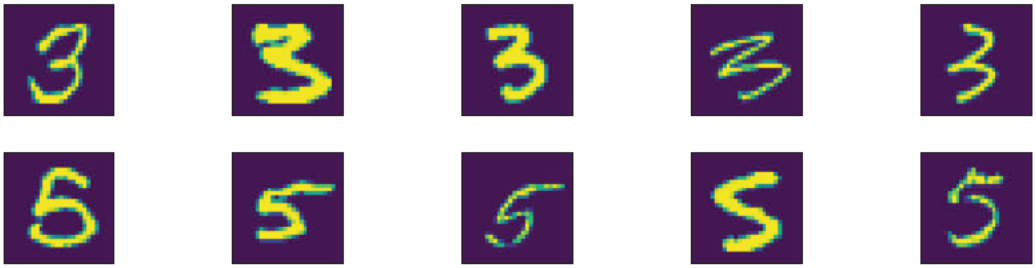


Figure 2.3: MNIST 3 and 5 digits.

We use a i.i.d. Gaussian prior p_0 with mean 0 and variance $\sigma^2 = 0.001$. The likelihood function for logistic regression is

$$p(y^j \mid x^j, \mathbf{q}) = \frac{\exp\left(y^j \left\langle x^j, \mathbf{q}\right\rangle\right)}{1 + \exp\left(y^j \left\langle x^j, \mathbf{q}\right\rangle\right)},$$

where there are 2 classes (i.e. y^j can take values 0 and 1, with 1 corresponding to digit 5, and 0 corresponding to digit 3) and $(x^j, y^j)_{j=1}^{N_D}$ are the respective training points and labels for a data set of size N_D (there are $N_D = 11552$ training images of 3 or 5). We then define the posterior potential by

$$U(\mathbf{q}) = -\log\left(p_0(\mathbf{q})\right) - \sum_{i=1}^{N_D} \log\left(p\left(y^j \mid x^j, \mathbf{q}\right)\right). \tag{2.8.21}$$

A commonly used method in machine learning and other fields relies on a stochastic gradient approximation, an unbiased estimator of the gradient of the potential defined in (2.8.21). This is typically obtained based on a sub-sample of size N_b of a data set of size N_D, where $N_b << N_D$, i.e. for a random selection $I_{N_b} \subset [N_D] := \{1, ..., N_D\}$ one would consider the gradient of

$$\widehat{\nabla_{\mathbf{q}} U(\mathbf{q})}_{SG} = -\nabla_{\mathbf{q}} \log\left(p_0(\mathbf{q})\right) - \frac{N_D}{N_b} \sum_{i \in I_{N_b}} \nabla_{\mathbf{q}} \log\left(p\left(y^j \mid x^j, \mathbf{q}\right)\right), \tag{2.8.22}$$

where the sub-samples are chosen i.i.d at each gradient evaluation (or iteration) of the algorithm. Let $W = I_{N_b}$ as defined above, and $\mathcal{G}(\mathbf{q}, W) = \widehat{\nabla_{\mathbf{q}} U(\mathbf{q})}_{SG}$, then it is easy to see that the conditions of Definition 2.6.1 hold.

One more accurate estimator for the gradient is the variance-reduced stochastic gradient ([87]), also called the control variate method in the context of MCMC (see [8, 127]). This estimator uses the minimizer (or an approximation) $\mathbf{q}_{\min}$, and estimates the gradient as

$$\begin{aligned}\widehat{\nabla_{\mathbf{q}} U(\mathbf{q})}_{VRSG} = &-\nabla_{\mathbf{q}} \log\left(p_0(\mathbf{q})\right) - \nabla_{\mathbf{q}_{\min}} \sum_{i=1}^{N_D} \log\left(p\left(y^j \mid x^j, \mathbf{q}_{\min}\right)\right) \\ &- \frac{N_D}{N_b} \sum_{i \in I_{N_b}} \left[\nabla_{\mathbf{q}} \log\left(p\left(y^j \mid x^j, \mathbf{q}\right)\right) - \nabla_{\mathbf{q}_{\min}} \log\left(p\left(y^j \mid x^j, \mathbf{q}_{\min}\right)\right)\right].\end{aligned} \tag{2.8.23}$$

This can be also shown to satisfy Definition 2.6.1 with $W = I_{N_b}$ and $\mathcal{G}(\mathbf{q}, W) = \widehat{\nabla_{\mathbf{q}} U(\mathbf{q})}_{VRSG}$. Both (2.8.22) and (2.8.23) are unbiased estimators of the gradient. In situations where the distribution is concentrated near the minimizer (as the sample size is large compared to the number of parameters, or the prior is sufficiently strong), the (2.8.23) approximation has a much smaller variance, and we found that this reduces the bias of sampling algorithms. In the following numerics, we first consider full gradients for each scheme. We also implemented variance-reduced stochastic gradients for BAOAB, based on (2.8.23) with batch size $N_b = 100$.

We minimized the potential based on the BFGS algorithm and computed the smallest and largest eigenvalues of the Hessian at the minimizer, which were $m = 10^3$ and $M = 1.7342 \cdot 10^5$. Note that computing the upper and lower bounds on the Hessian globally is not easy for this problem, so we used these eigenvalues at the minimizer instead for setting the parameters in our simulations. We tried two different friction parameters: $\gamma = \sqrt{M}$ (the lowest value of γ for which our theory works) and $\gamma = \sqrt{m}$ (a good choice based on the contraction rates for Gaussians shown on Figure 2.2). In terms of stepsize, we tried $h \in \{2/\sqrt{M}, 1/\sqrt{M}, 1/(2\sqrt{M}), 1/(4\sqrt{M})\}$. The stepsize $h = 2/\sqrt{M}$ is near the anticipated stability threshold of these methods, this is confirmed by the fact that a larger stepsize ($h = 4/\sqrt{M}$) resulted in unstable behaviour and biases above 10^3 for all methods.

We used the potential U as a test function, which is often a good choice for examining convergence of Markov chains. The ground truth posterior mean of U was established based on running a well-tuned HMC with accept/reject steps (400 parallel runs, 440 million gradient evaluations in total, with 10% burn-in), this had a standard deviation of 0.023. The posterior standard deviation of U was also estimated based on these samples, it was found to be 19.82.

All tested methods were run in parallel 80 times for 120000 iterations per run (20000 burn-in, 100000 samples), initiated from the minimum of the potential. We computed effective sample sizes based on the approach of [156], using the Matlab package `https://github.com/lacerbi/multiESS`. All methods were implemented in Matlab on a desktop computer using GPU acceleration.

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{M}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{M}$
EM	$4.2(\pm 0.089)$	$1.5(\pm 0.13)$	$0.79(\pm 0.18)$	$0.28(\pm 0.23)$
BBK	$2.7(\pm 0.061)$	$0.67(\pm 0.099)$	$0.016(\pm 0.14)$	$-0.18(\pm 0.2)$
SPV	$123(\pm 0.079)$	$32.1(\pm 0.091)$	$8.19(\pm 0.13)$	$2.07(\pm 0.18)$
SVV	$126(\pm 0.097)$	$32.8(\pm 0.091)$	$8.17(\pm 0.13)$	$2.03(\pm 0.17)$
BAOAB	$-0.043(\pm 0.049)$	$-0.002(\pm 0.058)$	$0.13(\pm 0.086)$	$-0.055(\pm 0.12)$
BAOAB VRSG	$0.47(\pm 0.043)$	$0.23(\pm 0.066)$	$0.035(\pm 0.087)$	$0.036(\pm 0.12)$
OBABO	$2.7(\pm 0.056)$	$0.67(\pm 0.076)$	$0.22(\pm 0.13)$	$0.17(\pm 0.19)$
rOABAO	$-2.6(\pm 0.062)$	$-0.61(\pm 0.094)$	$0.025(\pm 0.13)$	$-0.16(\pm 0.19)$
SES/EB	$2.6(\pm 0.072)$	$1.2(\pm 0.094)$	$0.71(\pm 0.11)$	$0.2(\pm 0.18)$

Table 2.4: Bias for potential function, $\gamma = \sqrt{M}$

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{m}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{m}$
EM	$6.4 \cdot 10^4(\pm 0.82)$	$1.5 \cdot 10^4(\pm 0.72)$	$1.1 \cdot 10^3(\pm 0.73)$	$4.9(\pm 0.11)$
BBK	$2.8(\pm 0.034)$	$0.68(\pm 0.041)$	$0.1(\pm 0.05)$	$0.0038(\pm 0.066)$
SPV	$0.72(\pm 0.036)$	$0.14(\pm 0.043)$	$0.06(\pm 0.054)$	$-0.014(\pm 0.073)$
SVV	$3.5(\pm 0.036)$	$0.81(\pm 0.043)$	$0.26(\pm 0.061)$	$0.05(\pm 0.089)$
BAOAB	$0.03(\pm 0.038)$	$-0.011(\pm 0.049)$	$-0.046(\pm 0.062)$	$0.043(\pm 0.074)$
BAOAB VRSG	$6.4(\pm 0.04)$	$2.4(\pm 0.051)$	$1.1(\pm 0.063)$	$0.55(\pm 0.075)$
OBABO	$2.7(\pm 0.032)$	$0.65(\pm 0.041)$	$0.22(\pm 0.052)$	$0.11(\pm 0.071)$
rOABAO	$-1.7(\pm 0.041)$	$-0.55(\pm 0.041)$	$-0.2(\pm 0.054)$	$-0.033(\pm 0.081)$
SES/EB	$6.0 \cdot 10^4(\pm 0.61)$	$1.5 \cdot 10^4(\pm 0.48)$	$1.1 \cdot 10^3(\pm 0.59)$	$4.7(\pm 0.068)$

Table 2.5: Bias for potential function, $\gamma = \sqrt{m}$

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{M}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{M}$
EM	$146(\pm 0.7)$	$221(\pm 0.998)$	$282(\pm 0.822)$	$327(\pm 0.581)$
BBK	$85(\pm 0.535)$	$148(\pm 0.726)$	$221(\pm 0.969)$	$285(\pm 0.933)$
SPV	$86.7(\pm 0.554)$	$148(\pm 0.775)$	$221(\pm 0.887)$	$284(\pm 0.992)$
SVV	$86.5(\pm 0.645)$	$147(\pm 0.801)$	$222(\pm 0.916)$	$283(\pm 0.825)$
BAOAB	$44.3(\pm 0.304)$	$88.7(\pm 0.585)$	$152(\pm 0.812)$	$228(\pm 0.822)$
BAOAB VRSG	$44.6(\pm 0.332)$	$86.8(\pm 0.578)$	$152(\pm 0.915)$	$226(\pm 0.934)$
OBABO	$68.6(\pm 0.491)$	$140(\pm 0.84)$	$218(\pm 0.942)$	$282(\pm 0.809)$
rOABAO	$68.5(\pm 0.507)$	$140(\pm 0.692)$	$219(\pm 0.781)$	$283(\pm 0.862)$
SES/EB	$87.4(\pm 0.593)$	$149(\pm 0.663)$	$220(\pm 0.831)$	$284(\pm 0.809)$

Table 2.6: Gradient evaluations / ESS (potential function), $\gamma = \sqrt{M}$

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{m}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{m}$
EM	N.A.	N.A.	N.A.	$189(\pm 0.955)$
BBK	$15(\pm 0.124)$	$30.1(\pm 0.233)$	$57.5(\pm 0.352)$	$108(\pm 0.717)$
SPV	$15.1(\pm 0.106)$	$29.7(\pm 0.209)$	$57.4(\pm 0.408)$	$109(\pm 0.725)$
SVV	$15(\pm 0.121)$	$29.9(\pm 0.222)$	$57.5(\pm 0.341)$	$108(\pm 0.628)$
BAOAB	$18.8(\pm 0.128)$	$36.4(\pm 0.288)$	$66.4(\pm 0.461)$	$116(\pm 0.849)$
BAOAB VRSG	$19.7(\pm 0.169)$	$36.4(\pm 0.242)$	$67.8(\pm 0.447)$	$114(\pm 0.662)$
OBABO	$15(\pm 0.118)$	$30(\pm 0.204)$	$57.5(\pm 0.471)$	$108(\pm 0.711)$
rOABAO	$16.5(\pm 0.236)$	$29.7(\pm 0.218)$	$58.2(\pm 0.356)$	$109(\pm 0.669)$
SES/EB	N.A.	N.A.	N.A.	$108(\pm 0.652)$

Table 2.7: Gradient evaluations / ESS (potential function), $\gamma = \sqrt{m}$. N.A. indicates that the method did not converge for the given stepsize.

Firstly, when changing from $\gamma = \sqrt{M}$ to $\gamma = \sqrt{m}$, we can see that the changes in bias are not significant for BAOAB, OBABO, rOABAO, and BBK, the bias increases significantly for EM and SES (instability issues) and somewhat for BAOAB VRSG, and the bias decreases significantly for SPV and SVV. In terms of gradient evaluations per ESS, the choice $\gamma = \sqrt{m}$ is more efficient by a factor of 2-6 for all methods except EM and SES. This is in line with the recent research in accelerated convergence rates for underdamped Langevin dynamics ([38, 167]).

We can see that BAOAB has an impressively low bias for the potential test function even at the largest stepsize $2/\sqrt{M}$, and it also has a competitive computational cost in terms of gradient evaluations / effective sample size (ESS). The VRSG variant of BAOAB has a somewhat larger bias (especially at lower frictions), but it requires a similar number of iterations per ESS, with much lower computational cost per iteration compared to using full gradients. The rOABAO scheme based on randomized midpoints has a relatively low bias at all stepsizes and requires a rather small number of gradient evaluations per iteration. It is beyond the scope of this chapter, but we think that more significant differences could arise between these schemes for less smooth potentials.

Chapter 3

Unbiased kinetic Langevin Monte Carlo

3.1 Introduction

3.1.1 Unbiased estimation without accept/reject steps

This chapter describes a technique for performing Bayesian inference based on unbiased unadjusted Markov chain Monte Carlo that does not rely on Metropolis-Hastings accept/reject steps. Our algorithm is based on a multilevel scheme [77] that combines several different unadjusted MCMC chains to eliminate bias efficiently. Our approach is related to a recent paper [138] that introduced an unbiased unadjusted MCMC method, however we employ state-of-the-art integrators, and we extend the method with modifications for handling incomplete (or approximate) gradients, thus obtaining a procedure with improved scalability and competitiveness compared to state-of-the-art algorithms such as randomized Hamiltonian Monte Carlo (RHMC) [29, 45].

There have been several proposals for creating computationally efficient estimators for functions of SDE paths based on numerical discretization using multilevel Monte Carlo variance reduction techniques. Our scheme relates to the method of Muller et al [118] for approximating functions of whole paths of kinetic Langevin dynamics using integrators based on splitting. Unlike our approach, that work did not address the stationary distribution; moreover, the burn-in bias was not eliminated, and they did not consider the incorporation of approximate or stochastic gradients. More recently, Giles et al [78] introduced a general framework for multilevel approximation of expectations with respect to the stationary distribution of overdamped Langevin dynamics and also considered stochastic gradients. However, their approach does not produce unbiased samples, and overdamped Langevin dynamics generally appear less efficient at exploring distributions with high condition numbers than well-tuned kinetic Langevin dynamics [124], as considered here. Until this work, multilevel approaches have not been shown to be competitive with Hamiltonian Monte Carlo methods for high-dimensional sampling.

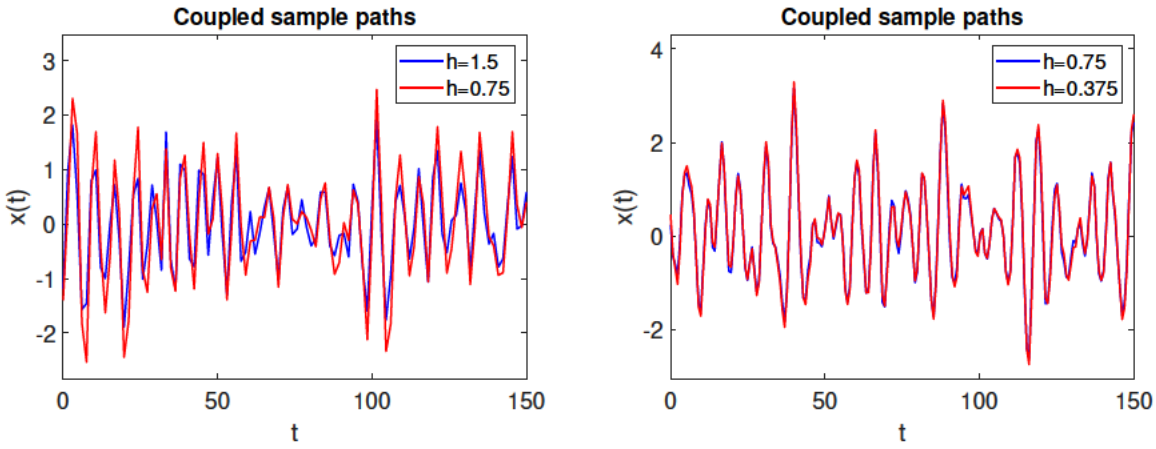


Figure 3.1: Coupled sample paths based on synchronous coupling from UBU (Section 3.2) discretization scheme of kinetic Langevin diffusion for a Gaussian target at stepsizes $h = 1.5, 0.75$ and $h = 0.75, 0.375$. UBU is strong order 2, so the typical distance between coupled paths is $\mathcal{O}(h^2)$.

3.1.2 Proposed methodology

We consider kinetic Langevin dynamics (also referred to as underdamped Langevin dynamics [47, 55]):

$$\begin{aligned} dX_t &= V_t dt, \\ dV_t &= -\nabla U(X_t)dt - \gamma V_t dt + \sqrt{2\gamma} dW_t, \end{aligned} \tag{3.1.1}$$

where $U : \mathbb{R}^d \to \mathbb{R}$ is a potential energy function, $\{W_t\}_{t\geq 0}$ is a standard $d-$dimensional Brownian motion, and $\gamma > 0$ is a friction coefficient. Under fairly weak assumptions, the unique invariant measure of the process $\{X_t, V_t\}_{t\geq 0}$ is of the form

$$\pi(dxdv) \propto \exp\left(-U(x) - \frac{\|v\|^2}{2}\right) dxdv. \tag{3.1.2}$$

This dynamics forms the basis of many sampling methods [34, 105], and it has a dimension-independent convergence rate for a large class of distributions [38]. In this chapter, we expand on the work of [138] and develop a comprehensive and practical framework for unbiased estimation. Specifically, we consider using a splitting integrator called UBU [140], which is strongly second-order accurate, where the unbiased estimator we introduce is referred to as UBUBU (Unbiased-UBU). Figure 3.1 illustrates the synchronously coupled paths of UBU discretizations of kinetic Langevin dynamics.

In Figure 3.1 we see that UBU discretization can be pathwise accurate even at large stepsize. Nevertheless, there is always some residual bias, and the stationary distribution of the discretization with stepsize h, π_h, differs from the target distribution π. The idea of unbiased estimation as proposed in [138] was to consider a sequence of discretization levels $h_l = 2^{-l}h_0$ for $l = 0, 1, 2, \ldots$ to create an estimator of the form

$$\hat{\pi}(f) = \hat{\pi}_{h_0}(f) + \sum_{l=0}^{\infty} \hat{\pi}_{h_{l+1},h_l}(f), \tag{3.1.3}$$

where f is some arbitrary quantity of interest, $\hat{\pi}_{h_0}(f)$ is an unbiased estimator of $\pi_{h_0}(f)$, and $\hat{\pi}_{h_{l+1},h_l}(f)$ is an unbiased estimator of $\pi_{h_{l+1}}(f) - \pi_{h_l}(f)$. A sophisticated coupling construction was used for defining $\hat{\pi}_{h_{l+1},h_l}(f)$ based on four Markov chains using Euler–Maruyama discretization of (3.1.1). Under certain weak assumptions, the estimator (3.1.3) was shown to have no bias, finite variance and finite expected computational cost.

In our algorithm:

(i) The burn-in bias is eliminated differently, resulting in simpler couplings. Our estimator is still of the form (3.1.3). However, instead of estimating $\pi_{h_0}(f)$ and $\pi_{h_{l+1}}(f) - \pi_{h_l}(f)$, which requires eliminating the burn-in bias for both discretization levels, we let $\hat{\pi}_{h_0}(f)$ be an unbiased estimator of $\tilde{\pi}_{h_0}(f)$, and $\hat{\pi}_{h_{l+1},h_l}(f)$ be an unbiased estimator $\tilde{\pi}_{h_{l+1}}(f) - \tilde{\pi}_{h_l}(f)$. Here $\tilde{\pi}_{h_l}(f)$ denotes the expected value of f according to the empirical distribution of a Markov chain using discretization stepsize h_l, thinning 2^l, and burn-in period of length $(B_0 + l \cdot B)/h_l$, for some constants $B_0, B > 0$. See Figure 3.2 for an illustration. Due to the increasing burn-in periods at smaller stepsizes, the bias of $\tilde{\pi}_{h_l}(f)$ shrinks to zero as $l \to \infty$. With this approach, we only need to couple two chains for creating unbiased estimators of $\tilde{\pi}_{h_{l+1}}(f) - \tilde{\pi}_{h_l}(f)$, and simple synchronous couplings can be used.

(ii) We use UBU discretization instead of Euler-Maruyama. The higher accuracy of UBU means that the differences between consecutive discretization levels h_l and h_{l+1} are smaller, and as a result, our estimator has a lower variance. We show that under certain assumptions, it is unbiased, has finite variance and finite expected computational cost.

(iii) In our method, the number of samples per level is deterministic (except at very small stepsize), and we can use Richardson extrapolation [130] to further lower the variance.

(iv) We show unbiasedness and finite variance even when using approximate or stochastic gradients. This dramatically improves the scalability of our method to large datasets.

(v) The usual unbiased estimator of Rhee and Glynn takes the form

$$\widehat{\pi}(f) = \frac{\xi_L}{\mathbb{P}_L(L)}, \tag{3.1.4}$$

such that

$$\begin{aligned} \mathbb{E}[\xi_{l_*}] &= \pi_{l_*}(f), \\ \mathbb{E}[\xi_l] &= \pi_l(f) - \pi_{l-1}(f) \quad l \in \{l_* + 1, l_* + 2, \ldots\}, \end{aligned}$$

where L is a random variable with probability mass function $\mathbb{P}_L$ on $\mathbb{N}_{l^*} := \{l_*, l_* + 1, \ldots\}$ that is independent of the sequence $\{\xi_l\}_{l \in \mathbb{N}_{l_*}}$. This approach was also used in [138]. Various alternative schemes with lower variance were proposed in [157].

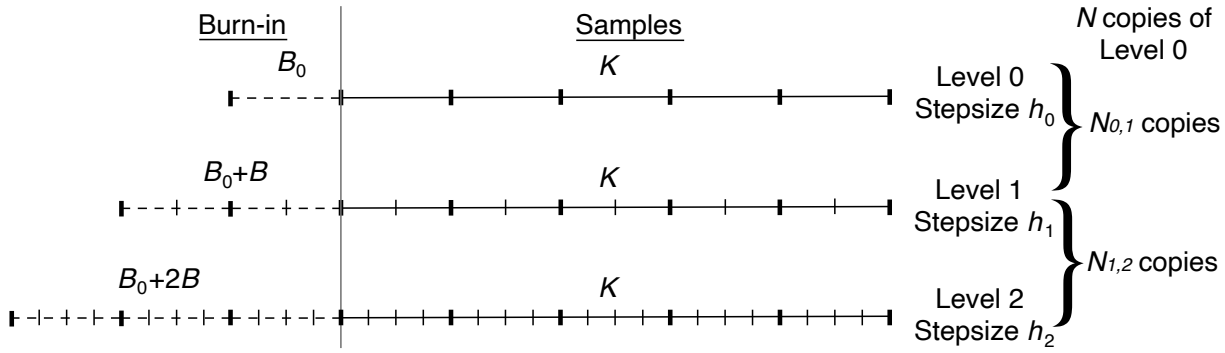


Figure 3.2: Elimination of bias by increasing burn-in lengths at higher discretization levels.

3.1.3 Organization

This chapter is organized as follows. In Section 3.2, we provide the necessary background material related to this work, including a discussion of splitting methods for kinetic Langevin dynamics, in particular the UBU discretization, as well as others such as BAOAB. We then discuss variants of our algorithm based on UBU, which includes an extension to stochastic gradients.
Section 3.3 is devoted to introducing our unbiased algorithms. We first provide some simple conditions for creating unbiased estimators with finite variance based on telescopic sums, together with a central limit theorem for such estimators. We then present our method using exact gradients and discuss necessary assumptions for unbiasedness and finite variance including showing that the variance of the estimator is finite. In addition to exact gradients, we also state versions of our method using stochastic and approximate gradients, with theoretical analysis.
Numerical experiments are provided in Section 3.4 on a range of model problems, including a simple Gaussian problem, an MNIST multinomial regression problem and a Poisson regression model applied to soccer game outcome prediction. Our unbiased methods are compared to RHMC, and demonstrate gains in terms of accuracy, and computational efficiency for high-dimensional problems, while eliminating bias.
Finally, we provide detailed proofs of all theorems in the appendices, as summarized in B.1.

Table 3.1 compares various Metropolized methods of the literature with our approach. [45] states that the warm start assumption cannot be removed as [97] has shown a lower bound of $\mathcal{O}(d^{1/2})$ without it. [4] proposes an algorithmic warm start using unadjusted kinetic Langevin dynamics at $\mathcal{O}(d^{1/2})$ gradient evaluations. For Gaussian targets, [5] has shown that it is possible to achieve a warm start using $\mathcal{O}(d^{1/4})$ gradient evaluations.

Algorithm	Gradient Evaluations	Conditions	Reference
MALA	$\mathcal{O}(d^{3/7})$	$h = \mathcal{O}(d^{-3/7})$, warm start	[45]
HMC	$\mathcal{O}(d^{1/4})$	$h = \mathcal{O}(d^{-1/4})$, warm start	[45]
RHMC	$\mathcal{O}(d^{1/4})$	$h = \mathcal{O}(d^{-1/4})$, warm start, Gaussian target	[5]
UBUBU	$\mathcal{O}(d^{1/4})$	$h_0 = \mathcal{O}(d^{-1/4})$	this work

Table 3.1: Dimension dependency of gradient evaluations per effective sample for different algorithms for m-strongly convex, M-∇Lipschitz, M_1^s-strongly Hessian Lipschitz potentials, in comparison to UBUBU.

3.2 Background & preliminary material

In this section, we provide the essential background material on kinetic (underdamped) Langevin dynamics and a splitting-type scheme called UBU. We then discuss the extension to stochastic gradients and state assumptions required in the remainder of the chapter.

For this work, we consider Langevin dynamics as defined by Equation (3.1.1) under temporal discretization. The simplest discretization is the Euler-Maruyama scheme. For a given stepsize $h > 0$, this proceeds, after initialization of x_0, v_0, with the following recursion:

$$\begin{aligned} x_{k+1} &= x_k + hv_k, \\ v_{k+1} &= v_k - h\nabla U(x_k) - h\gamma v_k + \sqrt{2\gamma h}\xi_{k+1}, \end{aligned} \tag{3.2.5}$$

where $(\xi_k)_{k\in\mathbb{N}}$ are i.i.d. $\mathcal{N}(0_d, I_d)$ random variables. Under suitable assumptions on the potential U, for h small enough, the discrete-time Markov chain expressed as $\{x_k, v_k\}_{k\in\mathbb{N}}$ admits a unique invariant measure π_h and moreover converges geometrically, meaning that for suitable classes of functions $f : \mathbb{R}^{2d} \to \mathbb{R}$,

$$\mathbb{E}\left[\left(K^{-1}\sum_{k=1}^{K} f(x_k, v_k) - \int f(x, v)\pi_h(dxdv)\right)^2\right] = \mathcal{O}(K^{-1}),$$

see [63]. In addition to this, π_h converges to π in distribution, as $h \to 0$. These types of results regarding convergence and accuracy can be extended to other numerical discretizations for the underdamped system, which we next discuss.

3.2.1 Splitting methods

Improved discretization methods with a high order of accuracy in both the weak and strong senses can be constructed by *splitting* [21, 98, 149], in which the SDE is broken into parts that can be either be solved analytically or which are in some way easier to handle numerically.

An accurate splitting method was introduced in [169] and was also studied in [140]. This splitting method only requires one gradient evaluation per iteration but has strong order two. The method is based on splitting the SDE (3.1.1) as follows

$$\begin{pmatrix} dx \\ dv \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ -\nabla U(x)dt \end{pmatrix}}_{\mathcal{B}} + \underbrace{\begin{pmatrix} vdt \\ -\gamma v dt + \sqrt{2\gamma}dW_t \end{pmatrix}}_{\mathcal{U}},$$

which can be integrated exactly over a step of size h. Given $\gamma > 0$, let $\eta = \exp(-\gamma h/2)$, and for ease of notation, we define the following operators

$$\mathcal{B}(x, v, h) = (x, v - h\nabla U(x)), \tag{3.2.6}$$

and

$$\begin{aligned}\mathcal{U}(x,v,h/2,\xi^{(1)},\xi^{(2)}) = \Big(x + \frac{1-\eta}{\gamma}v + \sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(h/2,\xi^{(1)}\right) - \mathcal{Z}^{(2)}\left(h/2,\xi^{(1)},\xi^{(2)}\right)\right),\\ \eta v + \sqrt{2\gamma}\mathcal{Z}^{(2)}\left(h/2,\xi^{(1)},\xi^{(2)}\right)\Big),\end{aligned} \tag{3.2.7}$$

where

$$\begin{aligned}\mathcal{Z}^{(1)}\left(h/2,\xi^{(1)}\right) &= \sqrt{\frac{h}{2}}\xi^{(1)},\\ \mathcal{Z}^{(2)}\left(h/2,\xi^{(1)},\xi^{(2)}\right) &= \sqrt{\frac{1-\eta^2}{2\gamma}}\left(\sqrt{\frac{1-\eta}{1+\eta}\cdot\frac{4}{\gamma h}}\xi^{(1)} + \sqrt{1-\frac{1-\eta}{1+\eta}\cdot\frac{4}{\gamma h}}\xi^{(2)}\right).\end{aligned} \tag{3.2.8}$$

The $\mathcal{B}$ operator indicated here is as given previously, whereas $\mathcal{U}$ as defined above is the exact solution in the weak sense of the remainder of the dynamics when $\xi^{(1)},\xi^{(2)} \sim \mathcal{N}(0,I_d)$ are independent random vectors. Different orders of composition of $\mathcal{B}$ and $\mathcal{U}$ can be taken to define different numerical integrators of kinetic Langevin dynamics, two such methods considered in [169] are BUB, a half step in $\mathcal{B}$, followed by a full step in $\mathcal{U}$ and a further half step in $\mathcal{B}$ and UBU, a half step in $\mathcal{U}$ followed by a full $\mathcal{B}$ step, followed by a half $\mathcal{U}$ step.

The Markov kernel for an UBU step with stepsize h will be denoted by P_h, which can be described by (3.2.9) as follows.

$$\begin{aligned}&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{4}, \quad \xi_{k+1}^{(i)} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1,...,4.\\ &(x_{k+1},v_{k+1}) = \mathcal{U}\mathcal{B}\mathcal{U}\left(x_k,v_k,h,\xi_{k+1}^{(1)},\xi_{k+1}^{(2)},\xi_{k+1}^{(3)},\xi_{k+1}^{(4)}\right)\\ &\qquad = \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(x_k,v_k,h/2,\xi_{k+1}^{(1)},\xi_{k+1}^{(2)}\right),h\right),h/2,\xi_{k+1}^{(3)},\xi_{k+1}^{(4)}\right).\end{aligned} \tag{3.2.9}$$

We have found that the strong second-order property and generally high accuracy of UBU makes it suitable for unbiased estimation, as described in Section 3.3.

The BAOAB method is an alternative splitting scheme that is known to be second-order weakly accurate and has small bias (see [28, 98, 99, 101]). BAOAB is exact for Gaussian targets and has a robustness property for large values of the friction parameter γ (see [105]), but its strong order is one. Theorem 3.3 of [152] claims that the stochastic velocity Verlet (SVV) method is, like UBU, also strongly second-order accurate. Despite their strengths as raw sampling schemes, both BAOAB and SVV exhibited worse performance than UBU in our preliminary numerical experiments in the setting of unbiased estimation. For this reason, we focus on UBU in this chapter. Nevertheless, it is important to note that the unbiased estimation approach of this chapter is by no means limited to the UBU integrator, and its performance could be further improved by more accurate integrators developed in the future.

3.2.2 Extension to stochastic gradients

In this subsection, we consider extending splitting methods with the use of stochastic gradients. We use the following definition from [103].

Definition 3.2.1. *A stochastic gradient approximation of a potential U is defined by a function $\mathcal{G} : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ and a probability distribution ρ on a Polish space Ω, such that for every $x \in \mathbb{R}^d$, $\mathcal{G}(x, \cdot)$ is measurable on $(\Omega, \mathcal{F})$, and for $\omega \sim \rho$,*

$$\mathbb{E}(\mathcal{G}(x, \omega)) = \nabla U(x).$$

The function $\mathcal{G}$ and the distribution ρ together define the stochastic gradient, which we denote as $(\mathcal{G}, \rho)$.

The following assumption is useful for controlling the accuracy of the stochastic gradient approximations.

Assumption 3.2.2. *We assume that the Jacobian of the stochastic gradient $\mathcal{G}$, $D_x\mathcal{G}(x, \omega)$ exists and it is measurable on $(\Omega, \mathcal{F})$. We also assume there exists $C_G > 0$ such that for $\omega \sim \rho$,*

$$\sup_{x \in \mathbb{R}^d} \mathbb{E}\|D_x\mathcal{G}(x, \omega) - \nabla^2 U(x)\|^2 \leq C_G.$$

Replacing the exact gradients with such stochastic gradients in the $\mathcal{B}$ step yields

$$\mathcal{B}_{\mathcal{G}}(x, v, h, \omega) = (x, v - h\mathcal{G}(x, \omega)), \tag{3.2.10}$$

and we can use this inside BAOAB and UBU to obtain stochastic gradient variants.

[103] has proven convergence bounds for BAOAB with stochastic gradients in Wasserstein distance and also shown that some widely used stochastic gradient schemes (random sampling with replacement, control variate gradient estimator) satisfy the conditions of Definition 3.2.1 and Assumption 3.2.2.

3.3 Unbiased multilevel Monte Carlo methods

In this section, we introduce and motivate our proposed algorithm, which we refer to as Unbiased UBU (UBUBU). We first describe the basic unbiased Monte Carlo scheme and introduce some essential assumptions. We then give relevant results which help to motivate our estimator, including a central limit theorem, a non-asymptotic bound on the variance with exact gradients, and other related results. Finally, we state our algorithm.

Suppose that for each $h \in (0, h_{\max}]$ (stepsize parameter), Q_h is a Markov kernel on some Polish state space Λ with stationary distribution μ_h such that μ_h converges to μ in distribution as $h \to 0$ (for example, these might be discretizations of a diffusion with different time stepsizes). Assume that we are interested in computing the expectation $\mu(f)$ of a function f satisfying $\pi_h(f^2) < \infty$ for every $h \in (0, h_{\max}]$ and $\mu(f^2) < \infty$. [138] suggested a multilevel estimation method based on stepsizes

$$h_0 \in (0, h_{\max}] \text{ and } h_l = h_0 \cdot 2^{-l} \text{ for } l = 1, 2, \ldots, \tag{3.3.11}$$

using a telescopic sum of the form

$$\mu(f) = \mu_{h_0}(f) + \sum_{j=1}^{\infty}(\mu_{h_j}(f) - \mu_{h_{j-1}}(f)).$$

Unbiased estimators of each term in the sum can be constructed via coupling. A challenge with this approach is that obtaining an unbiased estimator for $\mu_{h_0}(f)$ already requires two chains to be coupled based on the approach proposed in the papers [39, 79, 82, 86]. Estimating the expectations $\mu_{h_j}(f) - \mu_{h_{j-1}}(f)$ is even more challenging, requiring the coupling of four chains. The nature of the couplings means that it is not straightforward to use splitting methods such as UBU or BAOAB (as Markov kernels from different starting points need to be coupled closely in total variation distance, and this is difficult unless the distributions are Gaussian).

To overcome such issues, we propose a different telescoping sum for estimating $\mu(f)$,

$$\mu(f) = \tilde{\mu}_{h_0}(f) + \sum_{l=0}^{\infty}(\tilde{\mu}_{h_{l+1}}(f) - \tilde{\mu}_{h_l}(f)). \tag{3.3.12}$$

Here $\tilde{\mu}_{h_l}$ are created using some empirical averages, which will be defined in the rest of this section for exact, stochastic, and approximate gradients.

Suppose that D_0 is a random variable satisfying that $\mathbb{E}(D_0) = \tilde{\mu}_0(f)$. Let $\{D_0^{(r)}\}_{r=1}^N$ be N i.i.d. copies of D_0, and we define

$$S_0 = \frac{1}{N}\sum_{r=1}^{N} D_0^{(r)}. \tag{3.3.13}$$

Then it is clear that $\mathbb{E}(S_0) = \mathbb{E}(D_0) = \tilde{\mu}_{h_0}(f)$.

Let $D_{l,l+1}$ be a random variable such that

$$\mathbb{E}D_{l,l+1} = \tilde{\mu}_{h_{l+1}}(f) - \tilde{\mu}_{h_l}(f).$$

Let $c_{0,1}, c_{1,2}, \ldots$ be positive constants such that $c_{l,l+1} \to 0$ as $l \to \infty$, and let

$$\begin{aligned} &L(N) = \max\{l \in \mathbb{N} : c_{l,l+1}N \geq 0.5\}, \\ &N_{l,l+1} = \lceil c_{l,l+1}N \rceil \text{ for } l \leq L(N), \\ &N_{l,l+1} \sim \text{Bernoulli}(c_{l,l+1}N) \text{ for } l > L(N). \end{aligned} \tag{3.3.14}$$

For each $l \geq 1$, let $\{D_{l,l+1}^{(r)}\}_{r=1}^{N_{l,l+1}}$ be $N_{l,l+1}$ i.i.d. copies of $D_{l,l+1}$, and

$$S_{l,l+1} = \frac{1}{\mathbb{E}(N_{l,l+1})}\sum_{r=1}^{N_{l,l+1}} D_{l,l+1}^{(r)}. \tag{3.3.15}$$

It is clear from the definitions and Wald's equation that

$$\mathbb{E}S_{l,l+1} = \mathbb{E}D_{l,l+1} = \tilde{\mu}_{h_{l+1}}(f) - \tilde{\mu}_{h_l}(f).$$

Our first estimator is defined as

$$S = S_0 + \sum_{l=0}^{\infty} S_{l,l+1}, \tag{3.3.16}$$

where the terms $S_0, S_{0,1}, S_{1,2}, \ldots$ are independent.

The random $D_{l,l+1}$ variable will play a key role in our approach, as it is going to link two different discretization levels with stepsizes h_l and h_{l+1}. $\mathrm{Var}(S)$ depends on $\mathrm{Var}(D_{l,l+1})$, which is determined by how closely we couple the two discretizations. This is closely related to the strong order of the discretizations, determining how close they are to the underlying diffusion.

It is possible to improve estimator (3.3.16) slightly by the use of Richardson extrapolation [130]. The idea is that when h is sufficiently small, for Q_h defined in terms of an SDE discretization, the differences $\mu_h(f) - \mu(f)$ tend to follow a certain asymptotic behaviour in h, which can be characterized by an asymptotic expansion [98, 101]. For symmetric splittings like BAOAB it is known that $\mu_h(f) - \mu(f) = c_{f,\mu} h^2 (1 + \mathcal{O}(h))$ for some constant $c_{f,\mu}$ depending on f and μ. The same property can be established for UBU, using similar arguments. Based on this observation, and taking into account that such behaviour may only be valid at small stepsizes, our refined estimator is defined as

$$S(c_R) = S_0 + \sum_{l=0}^{L(N)-1} S_{l,l+1} + \frac{S_{L(N),L(N)+1}}{1 - c_R} + \sum_{l=L(N)+1}^{\infty} \overline{S}_{l,l+1}, \tag{3.3.17}$$

$$\overline{S}_{l,l+1} = \frac{1}{\mathbb{E}(N_{l,l+1})} \sum_{r=1}^{N_{l,l+1}} \left[D_{l,l+1}^{(r)} - S_{L(N),L(N)+1} \cdot c_R^{l-L(N)} \right],$$

where $c_R \in [0, 1)$ can be any number (we state the recommended choice of this in our algorithms). Our first estimator S is a special case since $S(0) = S$.

The key assumptions we make on the variances are as follows:

Assumption 3.3.1. *$f : \Lambda \to \mathbb{R}$ is a measurable function. $(\tilde{\mu}_{h_l})_{l \geq 0}$ is a sequence of distributions satisfying that $\tilde{\mu}_{h_l}(f) \to \mu(f)$ as $l \to \infty$. The random variable D_0 satisfies that $\mathbb{E}(D_0) = \mu_{h_0}(f)$, $\mathrm{Var}(D_0) < \infty$, for every $l \geq 0$, the random variable $D_{l,l+1}$ satisfies that $\mathbb{E}(D_{l,l+1}) = \tilde{\mu}_{l+1}(f) - \tilde{\mu}_{l+1}(f)$ and $|\mathbb{E}(D_{l,l+1}^2)| \leq V_D \phi_D^{-l}$ for some finite constants $V_D > 0$, $\phi_D > 2$.*

Assumption 3.3.2. *The constants $c_{l,l+1}$ controlling $N_{l,l+1}$ satisfy*

$$\underline{c}_N \phi_N^{-l} \leq c_{l,l+1} \leq \overline{c}_N \phi_N^{-l},$$

for some finite constants $0 < \underline{c}_N \leq \overline{c}_N$, $\phi_N > 2$.

Assumption 3.3.3. *The computational cost of generating a sample from $D_{l,l+1}$ is $\mathcal{O}(2^l(K + lB + B_0))$ for some finite constants B, B_0, and generating a sample from D_0 has a finite computational cost.*

Proposition 3.3.4. *Suppose that Assumptions 3.3.1, 3.3.2 and 3.3.3 hold, and that $2 < \phi_N < \phi_D$. Then S as defined in* (3.3.16) *is an unbiased estimator of $\mu(f)$ that has finite variance*

$$\mathrm{Var}(S) \leq \frac{\mathrm{Var}(D_0)}{N} + \frac{V_D}{N \underline{c}_N \left(1 - \frac{\phi_N}{\phi_D}\right)},$$

and finite expected computational cost.

Similarly, for any $c_R \in [0,1)$, $S(c_R)$ as defined in (3.3.17) *is also an unbiased estimator of $\mu(f)$ with finite variance*

$$\mathrm{Var}(S(c_R)) \leq \\ \frac{\mathrm{Var}(D_0)}{N} + \frac{V_D}{N\underline{c}_N\left(1-\frac{\phi_N}{\phi_D}\right)} + \frac{V_D}{N\underline{c}_N\left(1-\frac{\phi_N}{\phi_D}\right)}\frac{2}{(1-c_R)^2}\left(\frac{\phi_N}{\phi_D}\right)^{\log(2\underline{c}_N N/\phi_N)/\log(\phi_N)},$$

and finite expected computational cost.

Proof. See Section B.2 of the Appendix. □

We show below that a Central Limit Theorem (CLT) holds for these estimators.

Theorem 3.3.5. *Under the assumptions of Proposition 3.3.4, we have that, as $N \to \infty$,*

$$\sqrt{N}(S-\mu(f)) \Rightarrow \mathcal{N}(0,\sigma_S^2) \quad \textit{and} \quad \sqrt{N}(S(c_R)-\mu(f)) \Rightarrow \mathcal{N}(0,\sigma_S^2),$$

where

$$\sigma_S^2 := \mathrm{Var}(D_0) + \sum_{l=0}^{\infty}\frac{\mathrm{Var}(D_{l,l+1})}{c_{l,l+1}}. \tag{3.3.18}$$

Proof. See Appendix B.2. □

3.3.1 UBUBU with exact gradients

Now, we will specify the way D_0 and $D_{l,l+1}$ are defined based on UBU discretization of (3.1.1) with exact gradients, as defined in (3.2.9). Let μ_0 be an initial distribution on Λ that we can readily sample from, for example, a Dirac-δ at the maximum-a-posteriori (MAP) estimator. Let

$$R_0 = P_{h_0} \text{ and } R_l = P_{h_l}^{2^l} \text{ for } l = 1,2,\ldots. \tag{3.3.19}$$

These Markov kernels correspond to the same amount of time h_0 in the timescale of the limiting diffusion (and clearly, R_l still has μ_{h_l} as its stationary distribution). Consider B_0 burn-in steps with kernel R_0 at level 0, and $B_l = B_0 + lB$ steps with kernel R_l at level l. Define the approximate versions of μ_{h_l} as

$$\tilde{\mu}_{h_l} = \frac{1}{K}\sum_{i=1}^{K}\mu_0 R_l^{B_l+i}. \tag{3.3.20}$$

Estimates with respect to this can be computed by taking B_l burn-in steps according to R_l (equivalently $2^l B_l$ burn-in steps according to P_{h_l}), and then K additional steps that are used for computing an empirical average. In this way, we can compute expectations with respect to $\tilde{\mu}_{h_l}$ without the use of couplings. Moreover, given that at the diffusion time scale, the burn-in time tends to infinity as l grows, it is reasonable to expect that under suitable assumptions, $\tilde{\mu}_{h_l}$ converges to μ as $l \to \infty$.

Let D_0 be the empirical average of a function f based on K samples from Markov chain with kernel R_0 with burn-in B_0 initiated from μ_0, i.e. for the Markov chain $z_{-B_0}^{(0)} \sim \mu_0$, $z_{-B_0+1}^{(0)} \sim R_0(z_{-B_0}^{(0)}, \cdot), \ldots, z_K^{(0)} \sim R_0(z_{K-1}^{(0)}, \cdot)$. Let ν_0 denote the joint distribution of $z_{-B_0}^{(0)}, \ldots, z_K^{(0)}$, and define

$$D_0 = \frac{1}{K} \sum_{i=1}^{K} f(z_i^{(0)}). \tag{3.3.21}$$

Let $\{D_0^{(r)}\}_{r=1}^N$ be N i.i.d. copies of D_0, and define

$$S_0 = \frac{1}{N} \sum_{r=1}^{N} D_0^{(r)}. \tag{3.3.22}$$

Then it is clear that $\mathbb{E}(S_0) = \mathbb{E}(D_0) = \tilde{\mu}_{h_0}(f)$. For $l \geq 0$, let $z_{-B_l}^{(l,l+1)}, \ldots, z_K^{(l,l+1)}, z'^{(l,l+1)}_{-B_{l+1}}, \ldots, z'^{(l,l+1)}_K$ be Λ valued random variables defined on the same probability space (i.e. coupled) such that

- $z_{-B_l}^{(l,l+1)}, \ldots, z_K^{(l,l+1)}$ is a Markov chain with kernel R_l initiated as $z_{-B_l}^{(l,l+1)} \sim \mu_0$, and
- $z'^{(l,l+1)}_{-B_{l+1}}, \ldots, z'^{(l,l+1)}_K$ is a Markov chain with kernel R_{l+1} initiated $z'^{(l,l+1)}_{-B_{l+1}} \sim \mu_0$.

Let

$$D_{l,l+1} = \frac{1}{K} \sum_{i=1}^{K} [f(z'^{(l,l+1)}_i) - f(z_i^{(l,l+1)})]. \tag{3.3.23}$$

From the definitions, it follows that

$$\mathbb{E}D_{l,l+1} = \tilde{\mu}_{h_{l+1}}(f) - \tilde{\mu}_{h_l}(f),$$

hence $D_{l,l+1}$ is an unbiased estimator of the difference $\tilde{\mu}_{h_{l+1}}(f) - \tilde{\mu}_{h_l}(f)$.

When these Markov chains are discretizations of the same diffusion, it is natural to create synchronous couplings by using the same Brownian noise to generate the Gaussian random variables used during the periods $z_{-B_l}, \ldots, z_K$ and $z'_{-B_l}, \ldots, z'_K$. Such couplings can significantly reduce the variance of $D_{l,l+1}$. Let $\mathcal{B}$ and $\mathcal{U}$ be as in (3.2.6-3.2.7). Further we define $\mathcal{U}^2$ to be

$$\mathcal{U}^2(x, v, h, \xi^{(1)}, \xi^{(2)}, \xi^{(3)}, \xi^{(4)}) = \mathcal{U}\left(\mathcal{U}\left(x, v, h/2, \xi^{(1)}, \xi^{(2)}\right), h/2, \xi^{(3)}, \xi^{(4)}\right). \tag{3.3.24}$$

As $\mathcal{U}$ is an exact solution in the weak sense to its respective component in the splitting, this is an exact solution in the weak sense which uses Brownian increments $\left(\xi^{(1)}, \xi^{(2)}\right)$ in the first half step $h/2$ and $\left(\xi^{(3)}, \xi^{(4)}\right)$ in the second half step $h/2$. The $\mathcal{U}^2$ operator is an exact solution over stepsize h.

A coupling can be constructed between discretization levels so that the two discretization levels share Brownian motion in the exact integration of the $\mathcal{U}$ steps. This is done by using the Brownian increments from two respective $\mathcal{U}$ solutions at the higher level and concatenating them using the $\mathcal{U}^2$ operator at the lower level. Next, the stochastic integrals in the two levels are coupled by sharing the same Brownian noise. The Markov kernel $P_{h,h/2}$ for the two discretization levels $h, h/2$ is defined as

follows.

$$
\begin{aligned}
&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{8} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1, ..., 8. \\
&\left(x'_{k+1/2}, v'_{k+1/2}\right) = \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(x'_k, v'_k, h/4, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right), h/2\right), h/4, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right) \\
&\left(x'_{k+1}, v'_{k+1}\right) = \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(x'_{k+1/2}, v'_{k+1/2}, h/4, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}\right), h/2\right), h/4, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right) \\
&(x_{k+1}, v_{k+1}) = \\
&\mathcal{U}^2\left(\mathcal{B}\left(\mathcal{U}^2\left(x_k, v_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right), h\right), h/2, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right).
\end{aligned}
\tag{3.3.25}
$$

This Markov chain acts on the state space $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, moving from (x_k, v_k, x'_k, v'_k) to $(x_{k+1}, v_{k+1}, x'_{k+1}, v'_{k+1})$ via the steps in (3.3.25). When looking at the individual components, $(x_k, v_k) \to (x_{k+1}, v_{k+1})$ corresponds to one UBU step at stepsize h, while $(x'_k, v'_k) \to (x'_{k+1}, v'_{k+1})$ corresponds to two UBU steps at stepsize $h/2$. A key property here is that the stochastic integrals between two steps are synchronously coupled, which ensures that these two chains approximate the same underlying diffusion (in the strong sense). Hence, they are expected to remain close, which was observed in our numerical simulations.

We now create a coupling between levels l and $l+1$, denoted by $\nu_{l,l+1}$.

$\nu_{l,l+1}$ coupling

1: For given initial distribution μ_0 on Λ, we define $z_{-B_l}^{(l,l+1)} \sim \mu_0$ and $z_{-B_{l+1}}^{\prime(l,l+1)} \sim \mu_0$ as independent random variables.
2: We let $z_{-B_{l+1}}^{\prime(l,l+1)}, \ldots, z_{-B_l}^{\prime(l,l+1)}$ be a Markov chain evolving according to $R_{l+1} = (P_{h_{l+1}})^{2^{l+1}}$.
3: Let $(z_{-B_l}^{(l,l+1)}, z_{-B_l}^{\prime(l,l+1)}), (z_{-B_l+1}^{(l,l+1)}, z_{-B_l+1}^{\prime(l,l+1)}), \ldots, (z_K^{(l,l+1)}, z_K^{\prime(l,l+1)})$ be a Markov chain evolving according to $R_{l,l+1} = (P_{h_l,h_{l+1}})^{2^l}$.
4: Let $\nu_{l,l+1}$ denote the joint distribution of $z_{-B_l}^{(l,l+1)}, \ldots, z_K^{(l,l+1)}, z_{-B_{l+1}}^{\prime(l,l+1)}, \ldots, z_K^{\prime(l,l+1)}$.

The motivation for this $\nu_{l,l+1}$ coupling is that if two coupled chains are driven by the same noise and approximate the same diffusion, they are expected to be close most of the time. Given a sufficiently long burn-in, they will likely stay close during the iterations $1, 2, \ldots, K$ used for computing the differences in their empirical averages, reducing the variance of $D_{l,l+1}$. Let $c_N > 0$ and $\phi_N > 2$ be constants, and let

$$
c_{l,l+1} = c_N \phi_N^{-l} \text{ for } l \in \mathbb{N}. \tag{3.3.26}
$$

We call the overall estimator $S(c_R)$ based on formula (3.3.17) with $D_{l,l+1}$ defined based on coupling construction $\nu_{l,l+1}$ as *Unbiased UBU* (or UBUBU, for short). The steps for constructing this estimator are summarized in Algorithm 2.

Now, we will state our theoretical results for this algorithm. To prove unbiasedness and finite variance for our estimator $S(c_R)$, we require several assumptions, which we state below. These include assumptions on the smoothness and strong convexity of our potential, as well as restrictions on various parameters of the algorithm.

Algorithm 2 Unbiased-UBU (UBUBU)

Input:
- Maximum stepsize h_0.
- Friction parameter $\gamma > 0$.
- Initial distribution μ_0 on $\mathbb{R}^d \times \mathbb{R}^d$ for $l \geq 0$.
- Potential function $U : \mathbb{R}^d \to \mathbb{R}$ of target distribution.
- Burn-in length parameters B_0 and B.
- Number of samples parameter K.
- Number of parallel chains parameters N, c_N and ϕ_N.
- Richardson extrapolation parameter $c_R \in [0, 1)$ (default value $c_R = \frac{1}{4}$).
- Test function f.

Averages from level 0**:**
for $r = 1, \dots, N$ **do**
  Sample $z_{-B_0}^{(0,r)}, \dots, z_K^{(0,r)}$ from ν_0.
  Compute $D_0^{(r)}$ based on (3.3.21) using the samples $z_1^{(0,r)}, \dots, z_K^{(0,r)}$.
end for
Compute S_0 using (3.3.22).
Generate number of chains:
Sample $N_{l,l+1}$ according to (3.3.14), let $l_{\max} = \max\{l : N_{l,l+1} > 0\}$.
Averages of differences $D_{l,l+1}$ **from** $l = 0, \dots, l_{\max}$**:**
for $l = 0, \dots, l_{\max}$ **do**
  for $r = 1, \dots, N_{l,l+1}$ **do**
    Sample $z_{-B_l}^{(l,l+1,r)}, \dots, z_K^{(l,l+1,r)}, z'^{(l,l+1,r)}_{-B_{l+1}}, \dots, z'^{(l,l+1,r)}_K$ according to $\nu_{l,l+1}$.
    Compute $D_{l,l+1}^{(r)}$ based on (3.3.23) using $z_1^{(r,l,l+1)}, \dots, z_K^{(r,l,l+1)}, z_1'^{(r,l,l+1)}, \dots, z_K'^{(r,l,l+1)}$.
  end for
  Compute $S_{l,l+1}$ using (3.3.15).
end for
Compute $S(c_R)$ using (3.3.17).
Output:
Unbiased estimator $S(c_R)$,
Samples $z_1^{(0,r)}, \dots, z_K^{(0,r)}$ for parallel chains $1 \leq r \leq N$,
Samples $z_1^{(l,l+1,r)}, \dots, z_K^{(l,l+1,r)}, z_1'^{(r,l,l+1)}, \dots, z_K'^{(r,l,l+1)}$ for $0 \leq l \leq l_{\max}$, chains $1 \leq r \leq N_{l,l+1}$.

Assumption 3.3.6 (M-∇ Lipschitz)**.** *$U : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable and there exists $M > 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\|\nabla U(x) - \nabla U(y)\| \leq M\|x - y\|.$$

Assumption 3.3.7 (m-strong convexity)**.** *$U : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and there exists $m > 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m|x - y|^2.$$

The strongly Hessian Lipschitz property relies on the following tensor norm from [45], which we require in our setup.

Definition 3.3.8. *For $A \in \mathbb{R}^{d \times d \times d}$, let*

$$\|A\|_{\{1,2\}\{3\}} = \sup_{x \in \mathbb{R}^{d \times d}, y \in \mathbb{R}^d} \left\{ \sum_{i,j,k=1}^{d} A_{ijk} x_{ij} y_k \,\middle|\, \sum_{i,j=1}^{d} x_{ij}^2 \leq 1, \sum_{k=1}^{d} y_k^2 \leq 1 \right\}.$$

Remark 3.3.9. *The $\|A\|_{\{1,2\}\{3\}}$ norm in Definition 3.3.8 can be equivalently written as*

$$\|A\|_{\{12\}\{3\}} = \left\| \sum_{i_1} A_{i_1,\cdot,\cdot}^T \cdot A_{i_1,\cdot,\cdot} \right\|^{1/2}, \tag{3.3.27}$$

where $A_{i_1,\cdot,\cdot} = (A_{i_1,i_2,i_3})_{1\le i_2\le d, 1\le i_3\le d}$ is a $d \times d$ matrix, see the proof of Lemma 7 of [123].

Assumption 3.3.10 (M_1^s-strongly Hessian Lipschitz)**.** *$U : \mathbb{R}^d \to \mathbb{R}$ is three times continuously differentiable and M_1^s-strongly Hessian Lipschitz if there exists $M_1^s > 0$ such that*

$$\|\nabla^3 U(x)\|_{\{1,2\}\{3\}} \le M_1^s$$

for all $x \in \mathbb{R}^d$.

Remark 3.3.11. *In Lemma B.8.6 in the Appendix, we show that Bayesian multinomial regression satisfies this assumption.*

Assumption 3.3.12 (1-Lipschitzness of f)**.** *f is a 1-Lipschitz function with respect to the Euclidean distance on $\mathbb{R}^{2d}$, that only depends on x, not v (i.e. $f(x,v) = f(x,v')$ for any $x, v, v' \in \mathbb{R}^d$).*

Assumption 3.3.13 (Distance of initial distribution from target)**.** *The initial distribution on $\Lambda = \mathbb{R}^{2d}$ satisfy that $\mathcal{W}_2(\pi, \mu_0) \le c_{\mu_0}\sqrt{\frac{d}{m}}$, for some $c_{\mu_0} > 0$.*

Remark 3.3.14. *It is easy to show that under Assumption* (3.3.7)*, for $\mu_0 = \delta_{x^*} \times \mathcal{N}(0_d, I_d)$, and for $\mu_0 = \mathcal{N}(x^*, (\nabla^2 U(x^*))^{-1}) \times \mathcal{N}(0_d, I_d)$ (Gaussian approximation), this condition holds with $c_{\mu_0} = 2$.*

The heart of the work is related to demonstrating unbiasedness and finite variance as a result of the multilevel scheme presented in Figure 3.2. We now state our first main result, which is a non-asymptotic bound on the variance of our estimator (3.3.17).

Theorem 3.3.15. *Suppose that Assumptions 3.3.6, 3.3.7, 3.3.10, 3.3.12, 3.3.13 hold, and in addition,*

$$\gamma \ge \sqrt{8M}, \quad h_0 \le \frac{1}{\gamma} \cdot \frac{m}{264M}, \quad B \ge \frac{16\log(4)\gamma}{mh_0}, \quad B_0 \ge \frac{16\gamma}{mh_0} \log\left(\frac{c_{\mu_0}+1}{\sqrt{M}\gamma h_0^2}\right).$$

Suppose that $c_R \in [0,1)$, and $2 < \phi_N < 16$. Then for any $N \ge 1$, the UBUBU estimator $S(c_R)$ has finite expected computational cost, $\mathbb{E}S(c_R) = \pi(f)$, and it has finite variance. Moreover, it satisfies a CLT as $N \to \infty$, and the asymptotic variance σ_S^2 defined in (3.3.18) *can be bounded as*

$$\sigma_S^2 \le \frac{C(m, M, M_1^s, \gamma, c_N, \phi_N)}{Kh_0}\left(1 + \frac{1}{h_0 K} + dh_0^4\right).$$

Proof. See Section B.4.3 in the Appendix. □

Remark 3.3.16. *In particular, when setting $h_0 = \mathcal{O}(d^{-1/4})$, and $K > 1/h_0$, the bound simplifies to $\mathrm{Var}(S(c_R)) \le \frac{C(\gamma,m,M,M_1)}{NKh_0}$. This indicates that the overall number of gradient evaluations per effective sample in this setting is $\mathcal{O}(d^{1/4})$, which matches the best available bounds for HMC in [45], without the warm start assumption required in that paper.*

The following proposition shows dimension-free bounds for product distributions. We are going to use an assumption on the initial distribution μ_0.

Assumption 3.3.17. *Suppose that μ_0 and the target distribution π are of product form*

$$\mu_0(dx, dv) = \prod_{i=1}^{d} \mu_{0,i}(dx_i, dv_i) \quad \text{for all } l \geq 0, \quad \pi(dx, dv) = \prod_{i=1}^{d} \tilde{\pi}_i(dx_i) \frac{e^{-v_i^2/2} dv_i}{\sqrt{2\pi}},$$

for $x = (x_1, ..., x_d) \in \mathbb{R}^d$, $v = (v_1, ..., v_d) \in \mathbb{R}^d$, and that

$$\max_{1 \leq i \leq d} \mathcal{W}_2(\pi_i, \mu_{0,i}) \leq c_{\mu_0} \sqrt{\frac{1}{m}},$$

for some finite constant c_{μ_0}, where $\pi_i(dx_i, dv_i) = \tilde{\pi}_i(dx_i) \frac{e^{-v_i^2/2}}{\sqrt{2\pi}} dv_i$ is the joint distribution of (x_i, v_i) according to the target π.

Proposition 3.3.18. *Suppose that Assumption 3.3.17 holds, and denote the potential U as $U(x) = \sum_{i=1}^{d} U_i(x_i)$. Suppose that Assumptions 3.3.6, 3.3.7, and 3.3.10 hold for each component $(U_i)_{1 \leq i \leq d}$, and that*

$$\gamma \geq \sqrt{8M}, \quad h_0 \leq \frac{1}{\gamma} \cdot \frac{m}{264M}, \quad B \geq \frac{16 \log(4) \gamma}{m h_0}, \quad B_0 \geq \frac{16\gamma}{m h_0} \log\left(\frac{c_{\mu_0} + 1}{\sqrt{M} \gamma h_0^2} \right).$$

Suppose that f is of the form

$$f(x, v) = g(\langle w^{(1)}, x \rangle, \dots \langle w^{(r)}, x \rangle), \tag{3.3.28}$$

where $g : \mathbb{R}^r \to \mathbb{R}$ is 1-Lipschitz, and $w^{(1)}, \dots, w^{(r)} \in \mathbb{R}^d$. Suppose that $c_R \in [0, 1)$ and $2 < \phi_N < 16$. Then for any $N \geq 1$, the UBUBU estimator $S(c_R)$ has finite expected computational cost, $\mathbb{E}S(c_R) = \pi(f)$, and it has finite variance. Moreover, it satisfies a CLT as $N \to \infty$, and the asymptotic variance can be bounded as

$$\sigma_S^2 \leq \frac{C(m, M, M_1, \gamma, r, c_N, \phi_N)}{K h_0} \sum_{1 \leq i \leq r} \|w^{(i)}\|^2.$$

Proof. See Section B.4.3 in the Appendix. □

Remark 3.3.19. *These bounds are independent of the dimension d. This is not surprising as the different components evolve independently according to the kinetic Langevin diffusion* (3.1.1)*, and we do not introduce any dependencies in the UBUBU algorithm. This is in contrast with Metropolized methods, where the accept/reject steps introduce dependencies in the evolution of the components. The results could be generalized to potentials which are separable into independent groups of coordinates, i.e. $U(x) = \sum_{i=1}^{s} U_i(x_{G_i})$, where $G_1, \dots, G_s$ is a partition of $[d]$, and the size of each group $|G_i|$ is small.*

3.3.2 UBUBU with stochastic gradients

In this section, we extend the unbiased estimation methods of the previous section to the setting where we have instead stochastic or approximate gradient evaluations, combined with a control variate approach that occasionally computes full gradients for variance reduction, as in [87, 168].

In many applications, particularly in data science and machine learning, gradient computations are computationally expensive due to large datasets and the need to iterate through the entire dataset at each gradient evaluation. A common approach for reducing the cost of the gradient-based methods is to use stochastic gradient approximations based on subsampling the dataset to compute unbiased estimates (see [8, 33, 87, 127, 151, 160]).

In these applications the potential $U : \mathbb{R}^d \to \mathbb{R}$ is typically of the form

$$U(x) = U_0(x) + \sum_{i=1}^{N_D} U_i(x), \tag{3.3.29}$$

where $x \in \mathbb{R}^d$, the dataset is of size $N_D \in \mathbb{N}$. U_0 can be chosen as the negative log density of the prior distribution or some other term that does not require accessing the data. In our examples, U_0 can be taken to be a quadratic function, for example, a quadratic matching the Hessian at the minimizer (which can be computed before sampling).

We remark that one of the most efficient samplers in the big data regime is the Zig-Zag sampler [18] whose complexity is independent of the data size according to a limiting argument (although as stated in [18], some logarithmic factors were ignored). [50] is another recent paper that proposes a Metropolis-Hastings-type MCMC algorithm based on subsampling that only accesses $\mathcal{O}(1)$ or even $\mathcal{O}(1/\sqrt{N_D})$ data points per step. Although this method was shown to have state-of-the-art performance on a 10-dimensional logistic regression example, its efficiency on high-dimensional models has not yet been demonstrated.

In this section, we will develop a version of UBUBU using stochastic gradients. We are going to use random variables of the form $\omega \in [N_D]^{N_b}$, which is a random selection of N_b indices to be selected uniformly on $[N_D] = \{1, \dots, N_D\}$, i.i.d. with replacement [8]. We denote the distribution of ω here as $\mathcal{SWR}(N_D, N_b)$.

Definition 3.3.20. *The sub-sampled stochastic gradient of U at x with respect to $\hat{x}$ is*

$$\mathcal{G}(x, \omega|\hat{x}) = \nabla U_0(x) + \sum_{i=1}^{N_D} \nabla U_i(\hat{x}) + \frac{N_D}{N_b} \sum_{i \in \omega} [\nabla U_i(x) - \nabla U_i(\hat{x})], \tag{3.3.30}$$

where $\omega \sim \mathcal{SWR}(N_D, N_b)$.

$\mathcal{G}(\cdot|\hat{x}) : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ is an unbiased estimator of $\nabla U(x)$ in the sense of Definition 3.2.1. We can use this estimator in UBU by replacing the $\mathcal{B}$ step with

$$\mathcal{B}_{\mathcal{G}}(x, v, h, \omega|\hat{x}) = (x, v - h\mathcal{G}(x, \omega|\hat{x})). \tag{3.3.31}$$

Let $x^* \in \mathbb{R}^d$ be the minimizer of the potential U, then the selection $\hat{x} = x^*$ at each step corresponds to the control variate gradient estimator, see [8]. When approximating the step $\mathcal{B}$ in UBU using this control variate approach, we can only achieve strong order $1/2$.

Another possibility is to update $\hat{x}$ every $\tau = \lceil N_D/N_b \rceil$ iterations with the latest position where the gradient was evaluated (this is not x_k for UBU as the gradients are evaluated after moving forward by a $\mathcal{U}$ step with stepsize $h/2$). We refer to this as the stochastic variance reduced gradient (SVRG) approach (see [87, 168]). The overall computational cost of this approach is approximately twice that of the control variate approach (due to the need for a full gradient evaluation). Since the gradient is reevaluated every τ iterations, when h is small, the position $\hat{x}$ becomes closer to the positions x that are considered, and the approximate dynamics provide a better approximation of the underlying diffusion (3.1.1). We will show that the SVRG discretization has strong order $3/2$, which is better than the control variate estimator, hence we will use it within our unbiasing scheme. The evolution of SVRG steps can be written as follows,

$$\begin{aligned}
&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{4} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1, ..., 4. \\
&\omega_{k+1} \sim \mathcal{SWR}(N_D, N_b) \\
&(\overline{x}_k, \overline{v}_k) = \mathcal{U}\left(x_k, v_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right) \\
&\hat{x}_k = \overline{x}_{\lfloor k/\tau \rfloor \tau} \\
&(x_{k+1}, v_{k+1}) = \mathcal{U}\left(\mathcal{B}_{\mathcal{G}}\left(\overline{x}_k, \overline{v}_k, h, \omega_{k+1} | \hat{x}_k\right), h/2, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right).
\end{aligned} \tag{3.3.32}$$

Let P_h^{SVRG} denote the time inhomogenous Markov kernel describing the evolution of $(x_k, \hat{x}_k, v_k)$ according to the SVRG steps (3.3.32).

We use a Gaussian approximation of the target. Let

$$\mu_G = \mathcal{N}(x^*, (H^*)^{-1}) \times \mathcal{N}(0_d, I_d) \quad \text{with} \quad H^* = \nabla^2 U(x^*), \tag{3.3.33}$$

and define

$$\mathcal{H}_*(x, v, h) = \begin{pmatrix} x^* \\ 0_d \end{pmatrix} + \exp\left(h \begin{pmatrix} 0_d & I_d \\ -H_* & 0_d \end{pmatrix}\right) \begin{pmatrix} x - x^* \\ v \end{pmatrix}, \tag{3.3.34}$$

$$\mathcal{O}(x, v, h/2, \xi^{(1)}, \xi^{(2)}) = \left(x, \eta v + \sqrt{2\gamma}\mathcal{Z}^{(2)}\left(h/2, \xi^{(1)}, \xi^{(2)}\right)\right), \tag{3.3.35}$$

$$\mathcal{O}^2(x, v, h, \xi^{(1)}, \xi^{(2)}, \xi^{(3)}, \xi^{(4)}) = \mathcal{O}\left(\mathcal{O}\left(x, v, h/2, \xi^{(1)}, \xi^{(2)}\right), h/2, \xi^{(3)}, \xi^{(4)}\right). \tag{3.3.36}$$

with $\mathcal{H}_*(x,v,h)$ corresponding the solution of the Hamiltonian dynamics on target $\mu_G \times \mathcal{N}(0_d, I_d)$ initiated in (x,v) at time h. It follows from (3.2.8) that $\sqrt{2\gamma}\mathcal{Z}^{(2)}\left(h/2, \xi^{(1)}, \xi^{(2)}\right) \sim \mathcal{N}(0_d, (1-\eta^2)I_d)$, so this $\mathcal{O}$ steps keeps the target invariant. We are going to use the OHO scheme as part of our algorithm.

$$\begin{aligned}
&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{4} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1, ..., 4. \\
&(\overline{x}_k, \overline{v}_k) = \mathcal{O}\left(x_k, v_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right) \\
&(x_{k+1}, v_{k+1}) = \mathcal{O}\left(\mathcal{H}_*\left(\overline{x}_k, \overline{v}_k, h\right), h/2, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right).
\end{aligned} \tag{3.3.37}$$

Let P_h^{OHO} denote the time homogeneous Markov kernel describing the evolution of (x_k, v_k) according to the OHO steps (3.3.37).

Two chains evolving according to SVRG with step size h and SVRG with step size $h/2$ can be coupled as follows.

$$\begin{aligned}
&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{8} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1, ..., 8. \\
&\omega'_{k+1/2}, \omega'_{k+1} \sim \mathcal{SWR}(N_D, N_b), \\
&(\overline{x}_k, \overline{v}_k) = \mathcal{U}^2\left(x_k, v_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right) \\
&\hat{x}_k = \overline{x}_{\lfloor k/\tau \rfloor \tau} \\
&(x_{k+1}, v_{k+1}) = \mathcal{U}^2\left(\mathcal{B}_{\mathcal{G}}\left(\overline{x}_k, \overline{v}_k, h, \omega_{k+1} \middle| \hat{x}_k\right), h/2, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right), \\
&\left(\overline{x}'_k, \overline{v}'_k\right) = \mathcal{U}\left(x'_k, v'_k, h/4, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right) \\
&\hat{x}'_k = \overline{x}'_{\lfloor 2k/\tau \rfloor \tau/2} \\
&\left(x'_{k+1/2}, v'_{k+1/2}\right) = \mathcal{U}\left(\mathcal{B}_{\mathcal{G}}\left(\overline{x}'_k, \overline{v}'_k, h/2, \omega'_{k+1/2}, v \middle| \hat{x}'_k\right), h/4, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right) \\
&\left(\overline{x}'_{k+1/2}, \overline{v}'_{k+1/2}\right) = \mathcal{U}\left(x'_{k+1/2}, v'_{k+1/2}, h/4, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}\right) \\
&\hat{x}'_{k+1/2} = \overline{x}'_{\lfloor (2k+1)/\tau \rfloor \tau/2} \\
&\left(x'_{k+1}, v'_{k+1}\right) = \mathcal{U}\left(\mathcal{B}_{\mathcal{G}}\left(\overline{x}'_{k+1/2}, \overline{v}'_{k+1/2}, h/2, \omega'_{k+1} \middle| \hat{x}'_{k+1/2}\right), h/4, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right).
\end{aligned} \tag{3.3.38}$$

Let $P_{h,h/2}^{SVRG}$ denote the time inhomogenous Markov kernel describing the evolution of $(x_k, \hat{x}_k, v_k, x'_k, \hat{x}'_k, v'_k)$ according to the SVRG steps (3.3.38).

Finally, we will also need to couple one chain with step size h running OHO on the Gaussian approximation μ_G, and another chain running SVRG on the target with step size $h/2$.

$$
\begin{aligned}
&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{8} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1, ..., 8. \\
&\omega'_{k+1/2}, \omega'_{k+1} \sim \mathcal{SWR}(N_D, N_b), \\
&(\overline{x}_k, \overline{v}_k) = \mathcal{O}^2\left(x_k, v_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right) \\
&(x_{k+1}, v_{k+1}) = \mathcal{O}^2\left(\mathcal{H}_*\left(\overline{x}_k, \overline{v}_k, h\right), h/2, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right), \\
&\left(\overline{x}'_k, \overline{v}'_k\right) = \mathcal{U}\left(x'_k, v'_k, h/4, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right) \\
&\hat{x}'_k = \overline{x}'_{\lfloor 2k/\tau \rfloor \tau/2} \\
&\left(x'_{k+1/2}, v'_{k+1/2}\right) = \mathcal{U}\left(\mathcal{B}_{\mathcal{G}}\left(\overline{x}'_k, \overline{v}'_k, h/2, \omega'_{k+1/2}, v \middle| \hat{x}'_k\right), h/4, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right) \\
&\left(\overline{x}'_{k+1/2}, \overline{v}'_{k+1/2}\right) = \mathcal{U}\left(x'_{k+1/2}, v'_{k+1/2}, h/4, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}\right) \\
&\hat{x}'_{k+1/2} = \overline{x}'_{\lfloor (2k+1)/\tau \rfloor \tau/2} \\
&\left(x'_{k+1}, v'_{k+1}\right) = \mathcal{U}\left(\mathcal{B}_{\mathcal{G}}\left(\overline{x}'_{k+1/2}, \overline{v}'_{k+1/2}, h/2, \omega'_{k+1} \middle| \hat{x}'_{k+1/2}\right), h/4, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right).
\end{aligned}
\tag{3.3.39}
$$

Let $P_{h,h/2}^{OHO/SVRG}$ denote the time inhomogenous Markov kernel describing the evolution of $(x_k, v_k, x'_k, \hat{x}'_k, v'_k)$ according to the steps (3.3.39).

We now create a coupling between levels 0 and 1, denoted by $\nu_{0,1}^{SG}$.

$\nu_{0,1}^{SG}$ coupling

1: Define $z_{-B_1}^{(0,1)} \sim \mu_G$ and let $z_{-B_1}^{\prime(0,1)} = z_{-B_1}^{(0,1)}$.
2: Let $(z_{-B_1}^{(0,1)}, z_{-B_1}^{\prime(0,1)}, \hat{x}_{-B_1}^{\prime(0,1)}), (z_{-B_1+1}^{(0,1)}, z_{-B_1+1}^{\prime(0,1)}, \hat{x}_{-B_1+1}^{\prime(0,1)}), \ldots, (z_K^{(0,1)}, z_K^{\prime(0,1)}, \hat{x}_K^{\prime(0,1)})$ be a Markov chain with kernel $R_{0,1}^{OHO/SVRG} = P_{h_0,h_1}^{OHO/SVRG}$ (satisfying that $z_k^{(0,1)} \sim \mu_G$ for all k).
3: Let $\nu_{0,1}$ denote the joint distribution of $z_{-B_0}^{(0,1)}, \ldots, z_K^{(0,1)}, z_{-B_1}^{\prime(0,1)}, \ldots, z_K^{\prime(0,1)}$.

We now create a coupling between levels l and $l+1$ for $l \geq 1$, denoted by $\nu_{l,l+1}^{SG}$.

$\nu_{l,l+1}^{SG}$ coupling

1: Define $z_{-B_{l+1}}^{(l,l+1)} \sim \mu_G$ and let $z_{-B_{l+1}}^{\prime(l,l+1)} = z_{-B_{l+1}}^{(l,l+1)}$.
2: Let $\left(z_{-B_{l+1}}^{(l,l+1)}, z_{-B_{l+1}}^{\prime(l,l+1)}, \hat{x}_{-B_{l+1}}^{\prime(l,l+1)}\right), \ldots, \left(z_{-B_l}^{(l,l+1)}, z_{-B_l}^{\prime(l,l+1)}, \hat{x}_{-B_l}^{\prime(l,l+1)}\right)$ be a Markov chain with kernel $R_{l,l+1}^{OHO/SVRG} = \left(P_{h_l,h_{l+1}}^{OHO/SVRG}\right)^{2^l}$ (satisfying that $z_{-B_l}^{(l,l+1)} \sim \mu_G$).
3: Let $\left(z_{-B_l}^{(l,l+1)}, z_{-B_l}^{\prime(l,l+1)}, \hat{x}_{-B_l}^{\prime(l,l+1)}\right), \left(z_{-B_l+1}^{(l,l+1)}, z_{-B_l+1}^{\prime(l,l+1)}, \hat{x}_{-B_l+1}^{\prime(l,l+1)}\right), \ldots, \left(z_K^{(l,l+1)}, z_K^{\prime(l,l+1)}, \hat{x}_K^{\prime(l,l+1)}\right)$ be a Markov chain evolving according to $R_{l,l+1}^{SVRG} = (P_{h_l,h_{l+1}}^{SVRG})^{2^l}$.
4: Let $\nu_{l,l+1}^{SG}$ denote the joint distribution of $z_{-B_l}^{(l,l+1)}, \ldots, z_K^{(l,l+1)}, z_{-B_{l+1}}^{\prime(l,l+1)}, \ldots, z_K^{\prime(l,l+1)}$.

Figure 3.3 illustrates our couplings between different levels using OHO/UBU discretizations. Given some constants $c_N > 0$, $\phi_N > 2$, we let

$$
c_{l,l+1} = c_N \phi_N^{-l} \text{ for } l \in \mathbb{N}. \tag{3.3.40}
$$

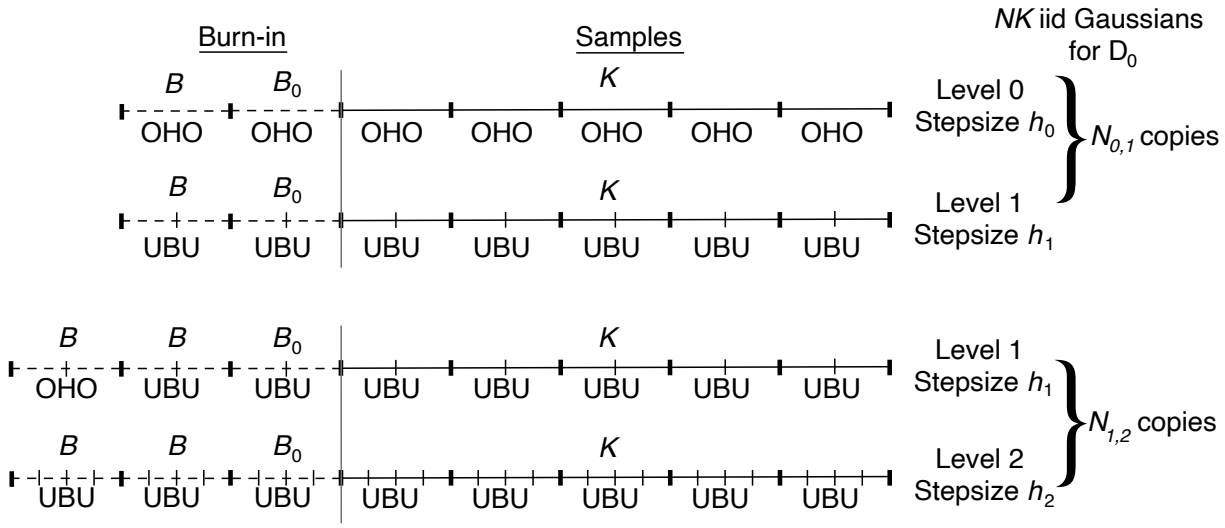


Figure 3.3: Coupling scheme for UBUBU-SG.

Our stochastic gradient-based method (UBUBU-SG) proceeds as stated in Algorithm 3. We recommend setting the Richardson extrapolation parameter $c_R = \frac{1}{2\sqrt{2}}$ in this case (as SVRG has strong order 3/2).

In order to show variance bounds for this algorithm, we make the following assumptions.

Assumption 3.3.21 (∇Lipschitz property)**.** *For every $1 \leq i \leq N_D$, $U_i : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable and there exists a $\tilde{M} > 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla U_i(x) - \nabla U_i(y)\| \leq \tilde{M}\|x - y\|,$$

for every $1 \leq i \leq N_D$ and moreover,

$$\|\nabla U(x) - \nabla U(y)\| \leq M\|x - y\| \quad \textit{for} \quad M = N_D\tilde{M}.$$

Assumption 3.3.22 ($N_D\tilde{m}$-strong convexity)**.** *There exists a $\tilde{m} > 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\langle \nabla U(x) - \nabla U(y), x - y\rangle \geq m\|x - y\|^2 \quad \textit{for} \quad m = N_D\tilde{m}.$$

Assumption 3.3.23 (strongly Hessian Lipschitz property)**.** *$U : \mathbb{R}^d \to \mathbb{R}$ is three times continuously differentiable and M_1^s-strongly Hessian Lipschitz if there exists $M_1^s > 0$ such that*

$$\|\nabla^3 U(x)\|_{\{1,2\}\{3\}} \leq M_1^s \quad \textit{for} \quad M_1^s = N_D\tilde{M}_1^s,$$

for all $x \in \mathbb{R}^d$.

The next theorem states our bounds on the asymptotic variance for this algorithm.

Algorithm 3 Unbiased-UBU with stochastic gradients (UBUBU-SG)

Input:
- Maximum stepsize h_0.
- Friction parameter $\gamma > 0$.
- Individual potential terms $(U_i)_{0 \le i \le N_D}$.
- Minimizer x^* of Potential $U(x)$ and its Hessian $H^* = \nabla^2 U(x^*)$.
- Batch size parameter N_b (related to $\tau = \lceil N_D/N_b \rceil$).
- Burn-in length parameters B_0 and B.
- Number of samples parameter K.
- Number of parallel chains parameters N, c_N and ϕ_N.
- Richardson extrapolation parameter $c_R \in [0, 1)$ (default value $c_R = \frac{1}{2\sqrt{2}}$).
- Test function f.

Samples from Gaussian approximation at level 0**:**
Sample NK i.i.d. samples $z_1^{(0)}, \dots, z_{NK}^{(0)}$ from μ_G.
Compute $S_0 := \frac{1}{NK} \sum_{i=1}^{NK} f(z_i^{(0)})$.
Generate number of chains:
Sample $(N_{l,l+1})_{l \ge 0}$ according to (3.3.14) and (3.3.40), let $l_{\max} = \max\{l : N_{l,l+1} > 0\}$.
Averages of differences $D_{l,l+1}$ **from** $l = 0, \dots, l_{\max}$**:**
for $l = 0, \dots, l_{\max}$ **do**
    for $r = 1, \dots, N_{l,l+1}$ **do**
        Sample $z_{-B_l}^{(l,l+1,r)}, \dots, z_K^{(l,l+1,r)}, z_{-B_{l+1}}^{\prime(l,l+1,r)}, \dots, z_K^{\prime(l,l+1,r)}$ from $\nu_{l,l+1}^{SG}$.
        Compute $D_{l,l+1}^{(r)}$ based on (3.3.23) using $z_1^{(r,l,l+1)}, \dots, z_K^{(r,l,l+1)}, z_1^{\prime(r,l,l+1)}, \dots, z_K^{\prime(r,l,l+1)}$.
    end for
    Compute $S_{l,l+1}$ using (3.3.15).
end for
Compute $S(c_R)$ using (3.3.17).
Output:
Unbiased estimator $S(c_R)$,
Samples $z_1^{(0)}, \dots, z_{NK}^{(0)}$,
Samples $z_1^{(l,l+1,r)}, \dots, z_K^{(l,l+1,r)}, z_1^{\prime(r,l,l+1)}, \dots, z_K^{\prime(r,l,l+1)}$ for levels $0 \le l \le l_{\max}$, parallel chains $1 \le r \le N_{l,l+1}$.

Theorem 3.3.24. *Considering UBUBU with stochastic gradients, suppose that Assumptions 3.3.12, 3.3.21, 3.3.22, 3.3.23 hold, and in addition* $\gamma \geq \sqrt{8M}$,

$$h_0 \leq \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tau/N_D, N_b)}{N_D^{3/2}}, \quad B \geq \frac{16\log(2^{3/2})\tilde{\gamma}}{\tilde{m}h_0 N_D^{1/2}}, \quad B_0 \geq \frac{16\tilde{\gamma}}{\tilde{m}h_0 N_D^{1/2}} \log\left(\frac{1}{N_D^{9/4}h_0^{3/2}}\right).$$

Suppose that $c_R \in [0,1)$ *and* $2 < \phi_N < 8$. *Then for any* $N \geq 1$, *the UBUBU estimator* $S(c_R)$ *has finite expected computational cost,* $\mathbb{E}S(c_R) = \pi(f)$, *and it has finite variance. Moreover, it satisfies a CLT as* $N \to \infty$, *and the asymptotic variance* σ_S^2 *defined in* (3.3.18) *can be bounded as*

$$\sigma_S^2 \leq \frac{1}{\tilde{m}N_D K} + C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, N_b, \phi_N)\frac{d^2}{c_N N_D^2}.$$

Proof. See Section B.6.2 in the Appendix. □

Remark 3.3.25. *With the choice* $c_N = \mathcal{O}\left(\frac{1}{N_D}\right)$ *and* $K = \mathcal{O}(1)$, *we get a bound* $\sigma_S^2 \leq \mathcal{O}\left(\frac{d^2}{\tilde{m}N_D}\right)$, *which, except for the dimension dependence, is similar to the variance of a 1-Lipschitz function according to the target. Hence obtaining an effective sample only requires evaluating a full gradient only once per* $\mathcal{O}(N_D)$ *iteration, so there is no increase in computational cost as the dataset size* N_D *increases. The dimension dependency* $\mathcal{O}(d^2)$ *in our bound is likely not sharp as we have not observed any dimension dependency in our simulations.*

Remark 3.3.26. *Although we have used Gaussian approximation at level* 0 *in Theorem 3.3.24 as this allows us to obtain better computational complexity in terms of* N_D, *one could also consider using UBU discretizations with SVRG gradients starting from level* 0. *This might be advantageous when the Gaussian approximation is not yet accurate. One could also consider different initial distributions. It is straightforward to adapt the proofs of Theorem 3.3.24 to show that even in such situations, under appropriate assumptions on the burn-in times, the UBUBU-SG method produces unbiased estimators with finite variance.*

3.3.3 UBUBU with approximate gradients

Stochastic gradients are not the only possible approach for computing accurate approximations of the gradient. In case the potential is close to a Gaussian (which is typical in the big data regime due to the Bernstein-von-Mises theorem), the following approximation can be quite accurate.

Definition 3.3.27. *The quadratic approximate gradient of* U *at* x *with respect to* $\hat{x}$ *is defined by*

$$\mathcal{Q}(x|\hat{x}) = \nabla U(\hat{x}) + \nabla^2 U(x^*)(x - \hat{x}), \tag{3.3.41}$$

where x^* *is the minimizer of* U.

When using this approximation for the gradient, the $\mathcal{B}$ step becomes

$$\mathcal{B}_{\mathcal{Q}}(x, v, h|\hat{x}) = (x, v - h\mathcal{Q}(x|\hat{x})). \tag{3.3.42}$$

The UBU iterations in this case become

$$\begin{aligned}
&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{4} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1, ..., 4. \\
&(\overline{x}_k, \overline{v}_k) = \mathcal{U}\left(x_k, v_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right), \\
&\hat{x}_k = \overline{x}_{\lfloor k/\tau \rfloor \tau} \\
&(x_{k+1}, v_{k+1}) = \mathcal{U}\left(\mathcal{B}_{\mathcal{Q}}\left(\overline{x}_k, \overline{v}_k, h \middle| \hat{x}_k\right), h/2, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right).
\end{aligned} \tag{3.3.43}$$

Let P_h^A denote the time inhomogenous Markov kernel describing the evolution of $(x_k, \hat{x}_k, v_k)$ according to the approximate gradient steps (3.3.43).

The reference point $\hat{x}$ is updated after every τ iterations for some $\tau \geq 1$. We only need to evaluate the full gradient once per τ iterations, and use an approximation based on the Hessian at the minimizer otherwise. Since the Hessian $H^* = \nabla^2 U(x^*)$ only has to be computed once, this does not affect overall efficiency when the number of samples N is sufficiently high. For many potentials of interest, the approximation steps in (3.3.41) can be computed at a much smaller cost than the gradient of U. Moreover, when thinning is used (such at levels $l = 1$ and higher), multiple steps according to (3.3.43) can be combined into one using the fact that this is a linear system, further reducing the number of matrix-vector products required.

We follow a similar strategy as in the UBUBU-SG case (see Figure 3.3). We use Gaussian samples at level 0, and couplings involving both OHO and UBU discretizations.

Two chains evolving according to approximate gradients with step sizes h and $h/2$ can be coupled as follows.

$$\begin{aligned}
&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{8} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1, ..., 8. \\
&(\overline{x}_k, \overline{v}_k) = \mathcal{U}^2\left(x_k, v_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right), \\
&\hat{x}_k = \overline{x}_{\lfloor k/\tau \rfloor \tau} \\
&(x_{k+1}, v_{k+1}) = \mathcal{U}^2\left(\mathcal{B}_{\mathcal{Q}}\left(\overline{x}_k, \overline{v}_k, h \middle| \hat{x}_k\right), h/2, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right). \\
&\left(\overline{x}'_k, \overline{v}'_k\right) = \mathcal{U}\left(x'_k, v'_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right), \\
&\hat{x}'_k = \overline{x}'_{\lfloor 2k/\tau \rfloor \tau/2} \\
&\left(x'_{k+1/2}, v'_{k+1/2}\right) = \mathcal{U}\left(\mathcal{B}_{\mathcal{Q}}\left(\overline{x}'_k, \overline{v}'_k, h/2, v \middle| \hat{x}'_k\right), h/4, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right), \\
&\left(\overline{x}'_{k+1/2}, \overline{v}'_{k+1/2}\right) = \mathcal{U}\left(x'_{k+1/2}, v'_{k+1/2}, h/2, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}\right), \\
&\hat{x}'_{k+1/2} = \overline{x}'_{\lfloor (2k+1)/\tau \rfloor \tau/2} \\
&\left(x'_{k+1}, v'_{k+1}\right) = \mathcal{U}\left(\mathcal{B}_{\mathcal{Q}}\left(\overline{x}'_{k+1/2}, \overline{v}'_{k+1/2}, h/2 \middle| \hat{x}'_{k+1/2}\right), h/4, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right),
\end{aligned} \tag{3.3.44}$$

Let $P_{h,h/2}^A$ denote the time inhomogenous Markov kernel describing the evolution of $(x_k, \hat{x}_k, v_k, x'_k, \hat{x}'_k, v'_k)$ according to the coupled approximate gradient steps (3.3.44).

As with stochastic gradients, will also need to couple one chain with step size h running OHO on the Gaussian approximation μ_G, and another chain based on UBU with approximate gradients on the target with step size $h/2$.

$$\begin{aligned}
&\left(\xi_{k+1}^{(i)}\right)_{i=1}^{8} \sim \mathcal{N}(0_d, I_d) \text{ for all } i = 1, ..., 8. \\
&(\overline{x}_k, \overline{v}_k) = \mathcal{O}^2\left(x_k, v_k, h/2, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right) \\
&(x_{k+1}, v_{k+1}) = \mathcal{O}^2\left(\mathcal{H}_*\left(\overline{x}_k, \overline{v}_k, h\right), h/2, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right), \\
&\left(\overline{x}'_k, \overline{v}'_k\right) = \mathcal{U}\left(x'_k, v'_k, h/4, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right) \\
&\hat{x}'_k = \overline{x}'_{\lfloor 2k/\tau \rfloor \tau/2} \\
&\left(x'_{k+1/2}, v'_{k+1/2}\right) = \mathcal{U}\left(\mathcal{B}_{\mathcal{Q}}\left(\overline{x}'_k, \overline{v}'_k, h/2 \middle| \hat{x}'_k\right), h/4, \xi_{k+1}^{(3)}, \xi_{k+1}^{(4)}\right) \\
&\left(\overline{x}'_{k+1/2}, \overline{v}'_{k+1/2}\right) = \mathcal{U}\left(x'_{k+1/2}, v'_{k+1/2}, h/4, \xi_{k+1}^{(5)}, \xi_{k+1}^{(6)}\right) \\
&\hat{x}'_{k+1/2} = \overline{x}'_{\lfloor (2k+1)/\tau \rfloor \tau/2} \\
&\left(x'_{k+1}, v'_{k+1}\right) = \mathcal{U}\left(\mathcal{B}_{\mathcal{Q}}(\overline{x}'_{k+1/2}, \overline{v}'_{k+1/2}, h/2 | \hat{x}'_{k+1/2}), h/4, \xi_{k+1}^{(7)}, \xi_{k+1}^{(8)}\right)
\end{aligned} \tag{3.3.45}$$

Let $P_{h,h/2}^{OHO/A}$ denote the time inhomogenous Markov kernel describing the evolution of $(x_k, v_k, x'_k, \hat{x}'_k, v'_k)$ according to the steps (3.3.45).

We define $\nu_{0,1}^A$ as joint distribution of $z_{-B_0}^{(0,1)}, \ldots, z_K^{(0,1)}, z_{-B_1}^{\prime(0,1)}, \ldots, z_K^{\prime(0,1)}$ that is similar to the $\nu_{0,1}^{SG}$ coupling for UBUBU-SG, but using inhomogenous Markov kernels $P_{h,h/2}^{OHO/A}$ instead of $P_{h,h/2}^{OHO/SVRG}$. Similarly, we let $\nu_{l,l+1}^A$ denote the joint distribution of $z_{-B_l}^{(l,l+1)}, \ldots, z_K^{(l,l+1)}, z_{-B_{l+1}}^{\prime(l,l+1)}, \ldots, z_K^{\prime(l,l+1)}$, defined analogously to $\nu_{l,l+1}^{SG}$ for UBUBU-SG, but using $P_{h,h/2}^{OHO/A}$ and $P_{h,h/2}^{A}$ in place of $P_{h,h/2}^{OHO/SVRG}$ and $P_{h,h/2}^{SVRG}$. We choose $c_{l,l+1}$ as

$$c_{l,l+1} = c_N \phi_N^{-l} \text{ for } l \in \mathbb{N}. \tag{3.3.46}$$

The UBUBU-Approx method follows similar steps as in Algorithm 3, but it uses the couplings ν_0^A and $\nu_{l,l+1}^A$ instead of ν_0^{SG}, and $\nu_{l,l+1}^{SG}$. In terms of input, unlike in Algorithm 3, we do not use individual potential terms $U_i(x)$ and batch size N_b, but require gradient calculation frequency τ. We recommend setting the Richardson extrapolation parameter $c_R = \frac{1}{2}$ in this case (as this approximate gradient scheme has strong order 1). A slightly weaker form of Assumption 3.3.21 will suffice here.

Assumption 3.3.28 (∇Lipschitz property)**.** *There is a $\tilde{M} > 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla U(x) - \nabla U(y)\| \leq M\|x - y\| \quad \textit{for} \quad M = N_D \tilde{M}.$$

Our results for this scheme are stated in Theorem 3.3.29.

Theorem 3.3.29. *Considering UBUBU-Approx method, suppose that Assumptions 3.3.12, 3.3.22, 3.3.23, 3.3.28 hold, and in addition $\gamma \geq \sqrt{8M}$,*

$$h_0 \leq \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tau/N_D)}{N_D^{3/2}}, \quad B \geq \frac{16\log(2)\tilde{\gamma}}{\tilde{m} h_0 N_D^{1/2}}, \quad B_0 \geq \frac{16\tilde{\gamma}}{\tilde{m} N_D^{1/2} h_0} \log\left(\frac{1}{N_D^3 h_0^2}\right).$$

Suppose that $c_R \in [0,1)$ and $2 < \phi_N < 4$. Then for any $N \geq 1$, $S(c_R)$ has finite expected computational cost, $\mathbb{E}S(c_R) = \pi(f)$, and it has finite variance. Moreover, it satisfies a CLT as $N \to \infty$, and the asymptotic variance σ_S^2 defined in (3.3.18) *can be bounded as*

$$\sigma_S^2 \leq \frac{1}{\tilde{m} N_D K} + \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, \phi_N) d^2}{c_N N_D^2}.$$

Proof. See Section B.7.3 in the Appendix. □

Remark 3.3.30. *To control the asymptotic variance of Theorem 3.3.24 and Theorem 3.3.29 for large d we would need to set $h_0 < \mathcal{O}(d^{-2})$; the dimension dependency in this bound might not be sharp, and we did not observe such limitations in our simulations. UBU iterations with AG and SVRG gradient approximations no longer form a time homogeneous Markov chain (unless the state space is extended), so it is challenging to establish $\mathcal{O}(1/K)$ scaling in the bound on σ_S^2, like in Theorem 3.3.15. If we select $h_0 \sim \mathcal{O}(1/N_D^{3/2})$, then for large N_D, the total computational cost of the approximate and stochastic gradient methods scales like $\mathcal{O}(N)$ due to Proposition 3.3.4. This is a significant improvement over UBUBU with exact gradients, which has a computational cost of $\mathcal{O}(N_D N)$.*

Algorithm	Computational Cost
UBUBU (Exact gradients)	$\mathcal{O}(N_D N)$
UBUBU (stochastic gradients)	$\mathcal{O}(N)$
UBUBU (approximate gradients)	$\mathcal{O}(N)$

Table 3.2: Comparison of the computational cost of the various UBUBU methods in terms of N and N_D.

Remark 3.3.31. *Although we have used Gaussian approximation at level 0 in Theorem 3.3.29 to understand the complexity in terms of N_D, one could also consider using UBU discretizations with approximate gradients starting from level 0. We could also use different initial distributions. It is not difficult to adapt the proof to show that even in such situations, under appropriate assumptions on the burn-in times, the UBUBU-Approx method produces unbiased estimators with finite variance.*

3.4 Numerical results

In this section, we provide numerical examples to demonstrate the effectiveness of our unbiased estimator UBUBU with exact, approximate and stochastic gradients. We test this on a range of problems, including a Gaussian example, a multinomial regression problem on the MNIST dataset, and a Poisson regression model for soccer scores; these computations serve to highlight the comparisons of our method with RHMC, which we view as the gold standard. We briefly describe the latter in Algorithm 1, stated in the introduction.

For RHMC, we have used a partial refreshment parameter of $\alpha = 0.7$, which typically performed $50\% - 70\%$ better than doing full velocity refreshment ($\alpha = 0$). We choose parameters E_L (expected number of leapfrog steps) and h (stepsize) such that the acceptance rate is approximately 0.65 (as recommended in [17]), and that $E_L h \approx \frac{1}{\sqrt{m}}$ (m is the minimal eigenvalue of the Hessian at the

mode), in line with the theoretical results for optimal convergence of the continuous time RHMC process [106]. We found that the effective sample sizes obtained in all of our experiments are in line with the continuous convergence rates of [106] scaled by the stepsize h, so we do not think that other parameter choices can significantly improve the performance of RHMC.

Our numerical experiments with unbiased estimators are specific to the UBU splitting method, as was the analysis. We also ran some preliminary numerical experiments with an unbiased version of BAOAB, but found that UBUBU was more efficient in all cases.

We estimated the ESS values based on at least 16 parallel runs in each simulation. For UBUBU, the number of parallel chains N was chosen in the range $N \in [64, 256]$, and we set $c_N = 1/16$, $\phi_N = 2\sqrt{2}$ in each case.

We will post the Matlab code of our simulations at `https://github.com/paulindani`.

3.4.1 Gaussian target

Here we consider a Gaussian target in d dimensions whose precision matrix has eigenvalues

$$1, 1 + \frac{\kappa - 1}{d - 1}, 1 + \frac{2(\kappa - 1)}{d - 1}, \ldots, \kappa.$$

Theorem 4 of [97] has shown that for some Gaussian targets with condition number κ, the inverse spectral gap of HMC taking K leapfrog steps per iteration was shown to be at least $\mathcal{O}(K\kappa\sqrt{d}/\log(d))$. More recently, it has been shown that randomizing the integration time can substantially improve the performance of HMC [29]. In continuous time, sharp convergence results have been obtained for RHMC in [106]. Moreover, for Gaussians with condition number κ, RHMC can approximate the target distribution with $\mathcal{O}(\sqrt{\kappa}d^{1/4})$ queries under a warm-start assumption [5]. In our preliminary experiments, RHMC significantly outperformed HMC on high-dimensional problems, so we only consider RHMC here.

We provided RHMC with the advantage of being initialized from the Gaussian target distribution, while UBUBU was initialized in $\mu_0(x, v) = \delta_0(x)\mathcal{N}(v, I_d)$.Our numerical simulations are presented in Figures 3.4-3.6.

Figure 3.4 shows the maximum number gradient evaluations per effective sample (ESS) among all components $f(x) = x_i$ for $1 \leq i \leq d$ as a function of the dimension $d = 10, 10^2, \ldots, 10^5$, for condition number $\kappa \in \{4, 100\}$. Figure 3.5 shows the number of gradient evaluations per ESS for the norm test function $f(x) = \|x\|$ as a function of the dimension d. As we can see, in both scenarios, UBUBU does not show any dimension dependence, while the number of gradient evaluations per ESS scales as $\mathcal{O}(d^{1/4})$ for RHMC. In our experiments, UBUBU is 20-40 times more efficient than RHMC for $d = 100000$.

Figure 3.6 presents the histograms of the number of gradient evaluations per effective sample size (ESS) amongst test functions $f(x) = x_1, \ldots, f(x) = x_d$, when comparing UBUBU with RHMC. This experiment is for a specific dimensions size of $d = 10^5$ and condition numbers $\kappa \in \{4, 100\}$. As we can observe, UBUBU outperforms RHMC in terms of gradient evaluations per ESS.

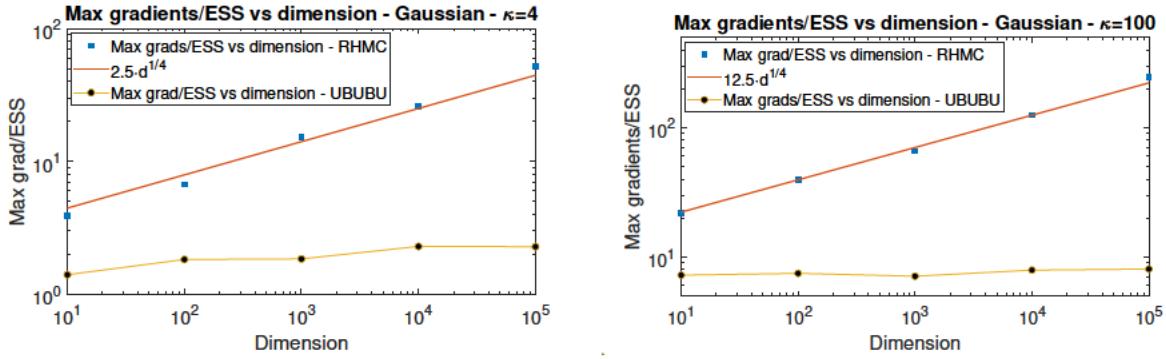


Figure 3.4: Dimensional dependence of gradients/ESS over all components for Gaussian targets.

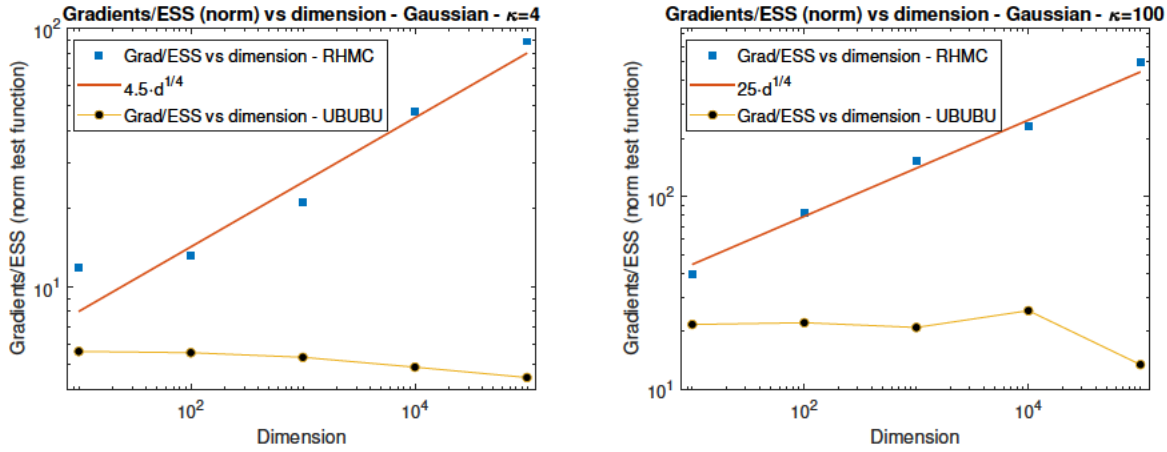


Figure 3.5: Dimension dependence of gradients/ESS for test function $\|x\|$ for Gaussian targets.

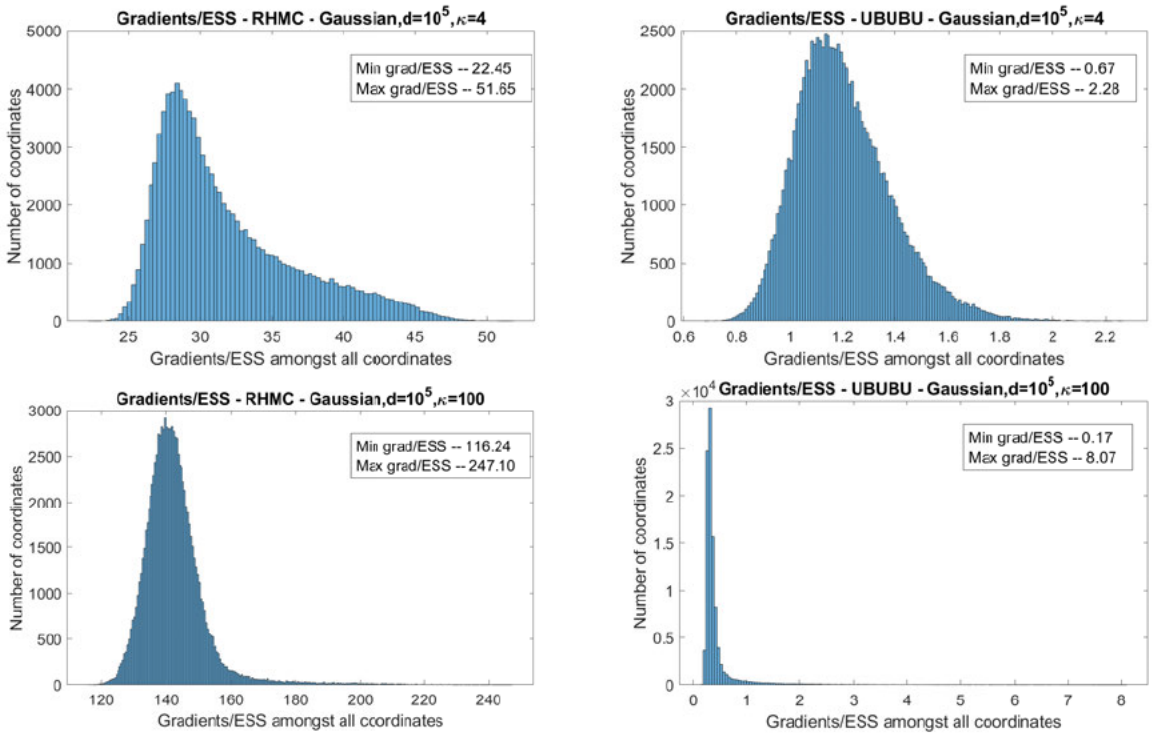


Figure 3.6: Gradient/ESS over all components for the Gaussian target example.

An important question related to this example is the dimension dependence of the original unbiased kinetic Langevin scheme based on Euler–Maruyama discretization presented in [138]. Due to the different estimator proposed there, the number of samples $N_{l,l+1}$ is random for every l, and the variance of the term equivalent to $S_{l,l+1} = \frac{1}{N_{l,l+1}} \sum_{i=1}^{N_{l,l+1}} D_{l,l+1}^{(r)}$ will be proportional to $\mathbb{E}(D_{l,l+1}^2)$, not $\mathrm{Var}(D_{l,l+1})$ like in our case. For functions like the norm $f(x,v) = \|x\|$, in general, using the strong order one property of the Euler–Maruyama scheme ([140]), $\mathbb{E}(D_{l,l+1}) = \mathcal{O}(\sqrt{d}h_l)$ and $\mathbb{E}(D_{l,l+1}^2) = \mathcal{O}(dh_l^2)$. So the asymptotic variance of the final estimator is $\mathcal{O}(1 + dh_0^2)$, and by choosing $h_0 = \mathcal{O}(d^{-1/2})$, we expect that this will require $\mathcal{O}(d^{1/2})$ gradient evaluations per effective sample.

3.4.2 Bayesian multinomial regression

Our second numerical example is to consider a Bayesian multinomial regression (BMR) problem. BMR is a generalized linear regression model which estimates probabilities for r different categories of dependent variable y using a set of explanatory variables x. Here, provided m classes, we let $q = (q^1, \dots q^m) \in \mathbb{R}^d$ with $d = md_o$ and $q^i \in \mathbb{R}^{d_o}$. The likelihood associated with the problem is given as

$$p(y^j|q) = \frac{\exp(\langle x^j, q^{y^j}\rangle)}{\sum_{1\le k\le m} \exp(\langle x^j, q^k\rangle)}. \tag{3.4.47}$$

Our focus is on estimating a posterior distribution, where the posterior potential is given as

$$U(q) = -\log(p_0(q)) - \sum_{k=1}^{N_D} \log\left(p(y^j|q)\right). \tag{3.4.48}$$

Here we chose p_0 as a Gaussian prior $p_0(q) = \frac{\exp(-\|q\|^2/(2\sigma_0^2))}{(\pi\sigma_0^2)^{d/2}}$. In Lemma B.8.6 in Appendix B.8, we show that the gradient-Lipschitz and strongly Hessian Lipschitz conditions (Assumptions 3.3.6 and 3.3.10) hold for this example.

We are interested in applying our BMR model to the MINST dataset [96] about classifying handwritten digits from 0 to 9, which are shown as examples in Figure 3.7. The dataset contains 60,000 training data points and 10,000 test data points where the images are of size 28 by 28 pixels. The covariate vectors x^j are obtained by flattening the images into vectors taking values on the interval $[0,1]$, and adding a 1 in the end for the intercept term. Hence $d_0 = 28^2+1 = 785$, $m = 10$, and $d = d_0 m = 7850$. We set the prior variance $\sigma_0^2 = 0.1$ (this was tested to provide good prediction performance).

Figure 3.7: MNIST datasets containing images of handwritten digits from 0 to 9.

For our numerical simulations, we will present two different scenarios: one without preconditioning (Figure 3.9) and one with preconditioning (Figure 3.10). In both figures, we evaluated the efficiency of the methods in terms of gradient evaluations per ESS for the coordinate test functions $f(x) = x_1, \ldots, f(x) = x_d$.

To compare the posterior distribution with a Gaussian approximation, we have selected a component with a relatively large third derivative. Figure 3.8 illustrates the potential function and the Gaussian approximation with precision $\nabla^2 U(x^*)$ along the line $x^* + te_i$. Here $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the unit vector of the chosen component ($i = 7491$ in our implementation), and t is chosen to cover up to 3 times the standard deviation difference from x_i^*. As we can see, the distribution of this component has a significant skewness, and the density values can differ by up 40% even in the bulk of the distribution.

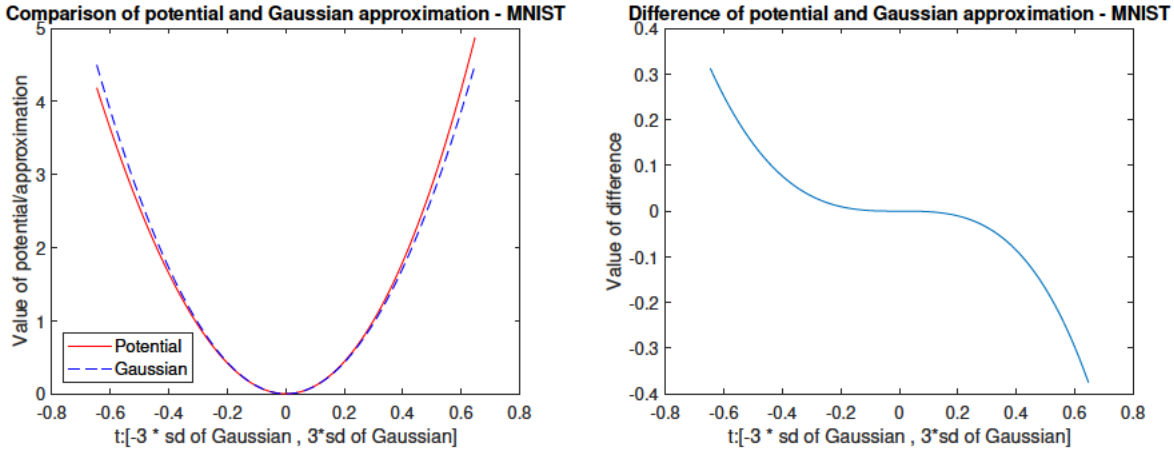


Figure 3.8: MNIST example. Left: Comparison between potential and quadratic approximation. Right: Difference between the potential and quadratic approximation.

In the first scenario (no preconditioning), the condition number of the Hessian at the mode $\nabla^2 U(x^*)$ is $\kappa \approx 7.2 \times 10^3$. We included simulation results with RHMC, UBUBU, and UBUBU-SG. For UBUBU-SG, we used a 10% batch ($N_b = 6000$, $N_D = 60000$), and set the maximum level with control variate stochastic gradient approximation as $s_{\max} = 2$. As we can see, UBUBU improves upon RHMC, and this is further improved by UBUBU-SG.

By preconditioning, we mean that we obtain samples from a transformed potential $U(Ax)$ for some matrix A, which may have a better condition number than the original potential. It is easy to see that if X follows a distribution with density proportional to $\exp(-U(x))$, then $X' = A^{-1}X$ has a density proportional to $\exp(-U(Ax))$.

In the case of RHMC, the best performance was obtained by preconditioning using the matrix square root of the Hessian at the mode, $A = (\nabla^2 U(x^*))^{-1/2}$. For UBUBU, this same approach worked reasonably well, but the best performance was obtained by only preconditioning in the eigenvectors corresponding to the largest 1000 eigenvalues of $\nabla^2 U(x^*)$ (i.e. shrinking them to the same size as the 1000th largest eigenvalue), and keeping the other directions unchanged. This resulted in a condition number of $\kappa \approx 4.5$ for the Hessian of the transformed potential at its mode.

We also included the implementation of the approximate gradient version UBUBU-Approx with the same preconditioning as for UBUBU and set the frequency of full gradient evaluations as $\tau = 15$. This has drastically reduced the number of gradient evaluations without hurting performance, and it shows approximately 100 times improvement over RHMC.

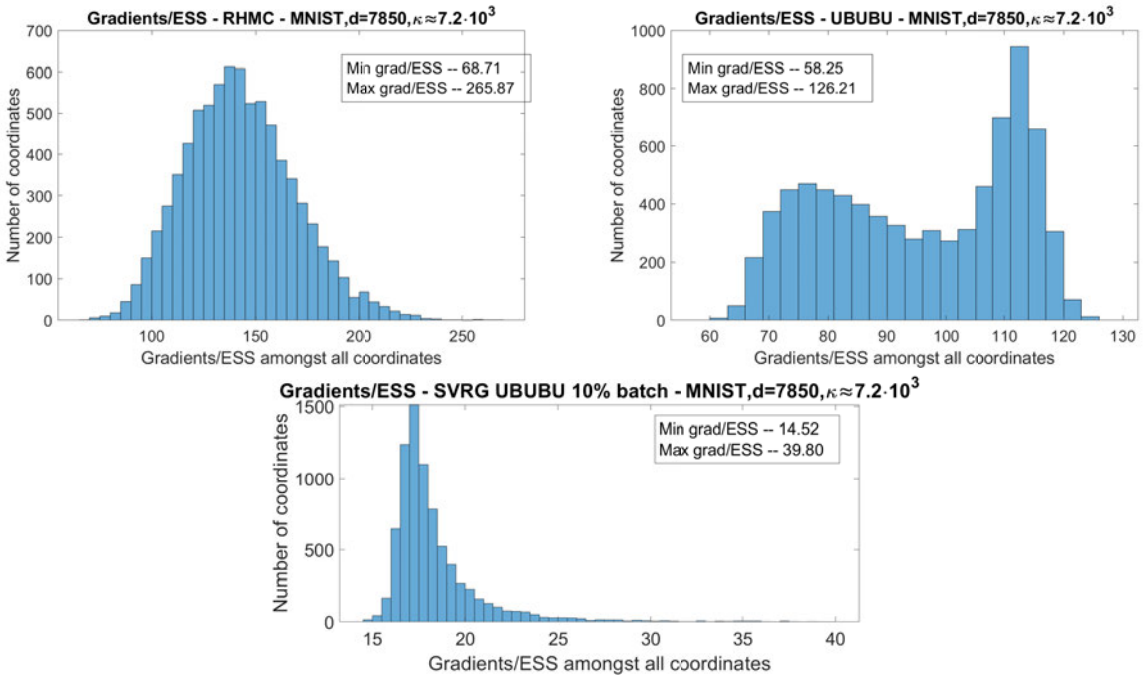


Figure 3.9: Gradient/ESS over all components for MNIST dataset without preconditioning.

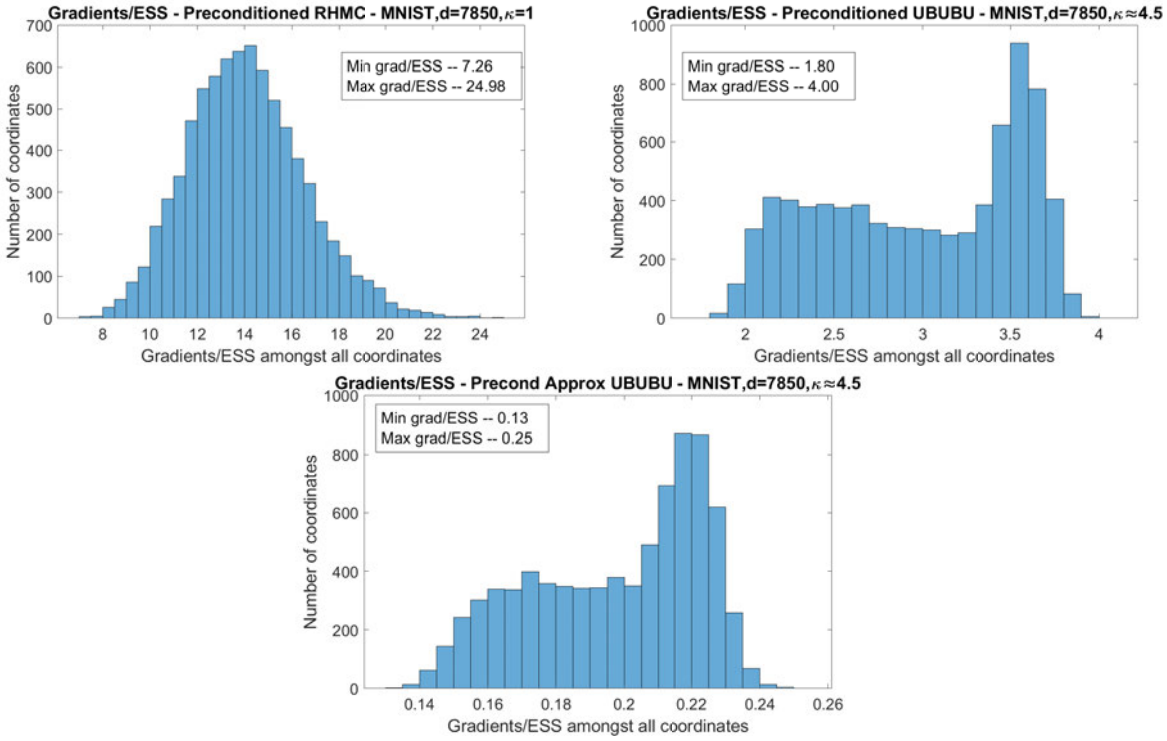


Figure 3.10: Gradient/ESS over all components for MNIST dataset with preconditioning.

In addition to the coordinate test functions, we have also evaluated the efficiency of these methods for the posterior predictive probability of digits $0, 1 \ldots, 9$ on the test dataset (10000 images, 100000 test functions in total). Figure 3.11 presents experiments comparing RHMC and UBUBU-Approx on these test functions. The experiments show approximately 60 times improvement in efficiency for UBUBU-Approx compared to RHMC, which is in line with our theory proving that UBUBU does not exhibit dimension dependency (Proposition 3.3.18).

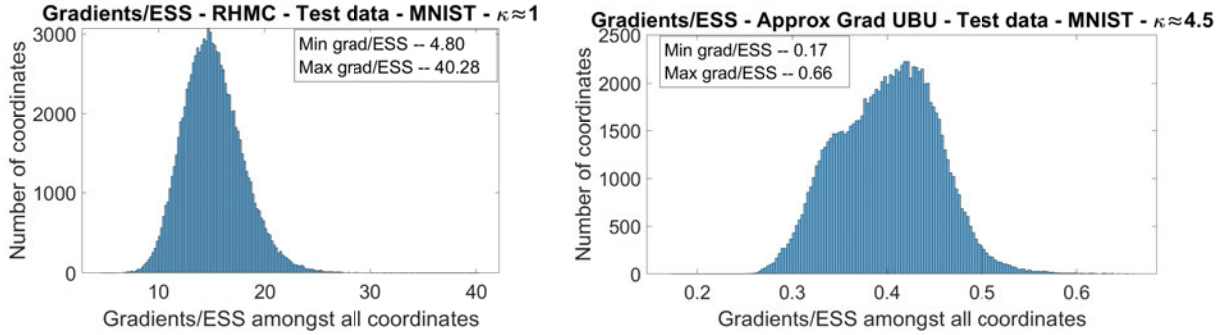


Figure 3.11: Gradient/ESS for probabilities of all 10 digits over 10000 test images for MNIST dataset with preconditioning.

3.4.3 Poisson regression model

Our final example is a Poisson regression model for predicting soccer scores taken from [94].

Let $g = 1, \ldots, G$ be the index of games. Let S_g^H denote the number of goals scored by the home team at game g, and let S_g^A denote the number of goals scored by the away team. The independent Poisson model [107] assumes that these scores are distributed as

$$S_g^H \sim \text{Poisson}(\lambda_g^H), \quad S_g^A \sim \text{Poisson}(\lambda_g^A),$$

conditionally independently given the rates λ_g^H and λ_g^A.

In our implementation, the rates are connected to the linear predictors η_g^H and η_g^A using the function $\text{softplus}(x) = \log(1 + \exp(x))$ (see Figure 3.12), i.e.

$$\lambda_g^A = \text{softplus}(\eta_g^A), \quad \lambda_g^H = \text{softplus}(\eta_g^H). \tag{3.4.49}$$

This function is Lipschitz and also gradient Lipschitz, which is desirable given our theory. Although this is less frequently used in the literature than the log link function, it was shown to be more robust and less sensitive to outliers [163, 165]. The linear predictors are modelled based on a random effect model with time-dependent attacking and depending strengths for each team. Let $w(g)$ denote the week of game g, then we set

$$\eta_g^H = a_{\text{home.team(g)},w(g)} + d_{\text{away.team(g)},w(g)}, \quad \eta_g^A = a_{\text{away.team(g)},w(g)} + d_{\text{home.team(g)},w(g)} \tag{3.4.50}$$

Let $\mathbf{a}$ be all attacking strengths of all teams over the whole period, and $\mathbf{d}$ denote all defending

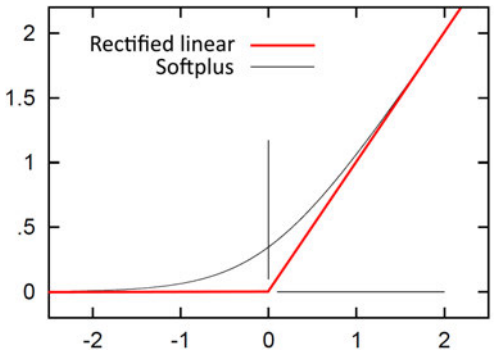


Figure 3.12: Softplus and ReLU activation functions.

strengths. Then the log-likelihood is of the form

$$\log(p(\mathbf{a},\mathbf{d})) = \\ C(S_1^H,\dots,S_G^H,S_1^A,\dots,S_G^H) + \sum_{g=1}^{G}\left(-\lambda_g^H + S_g^H\log(\lambda_g^H) - \lambda_g^A + S_g^H\log(\lambda_g^H)\right),$$

which can be written as a function of $\mathbf{a}$ and $\mathbf{d}$ using (3.4.49) and (3.4.50).

We used a Gaussian random walk prior for the attacking/defending strengths $a_{\text{team},w}$ and $d_{\text{team},w}$, together with a weak Gaussian prior on every attacking and defending strength. Let $\mathcal{T}$ denote the set of teams during the whole period considered (teams change from season to season due to relegation/promotion), then the overall log prior is of the form

$$\log p_0(\mathbf{a},\mathbf{d}) = C(\sigma,\sigma_0) - \sum_{\text{team}\in\mathcal{T}}\left(\sum_{w=w(1)}^{w(G)}\frac{a_{\text{team},w}^2}{2\sigma_0^2} + \sum_{w=w(1)}^{w(G)-1}\frac{(a_{\text{team},w+1}-a_{\text{team},w})^2}{2\sigma^2}\right) \\ - \sum_{\text{team}\in\mathcal{T}}\left(\sum_{w=w(1)}^{w(G)}\frac{d_{\text{team},w}^2}{2\sigma_0^2} + \sum_{w=w(1)}^{w(G)-1}\frac{(d_{\text{team},w+1}-d_{\text{team},w})^2}{2\sigma^2}\right),$$

We set $\sigma^2 = 0.01$ (this means a strong correlation for about two years), and $\sigma_0^2 = 10$ (weakly informative prior).

We considered 20 years of Premier League data (7600 games) from 19/08/2000 until 26/07/2020. Our model has $d = 89526$ parameters, and the condition number of the Hessian at the mode is $\kappa \approx 4\cdot 10^3$.

We have implemented RHMC, UBUBU and UBUBU-Approx with $\tau = 20$ for this model. In the UBUBU-Approx algorithm, the target at level 0 was chosen as the Gaussian approximation (with mean x^*, and precision matrix $\nabla^2 U(x^*)$), meaning that gradient evaluations were only used from level 1 onwards. The test functions were chosen as $f(x) = x_1,\dots,f(x) = x_d$. Our numerical simulations are presented in Figure 3.13. As we can see, UBUBU uses approximately 30 times fewer gradient evaluations per effective sample than RHMC, and UBUBU-Approx uses 2000 times fewer gradient evaluations than RHMC.

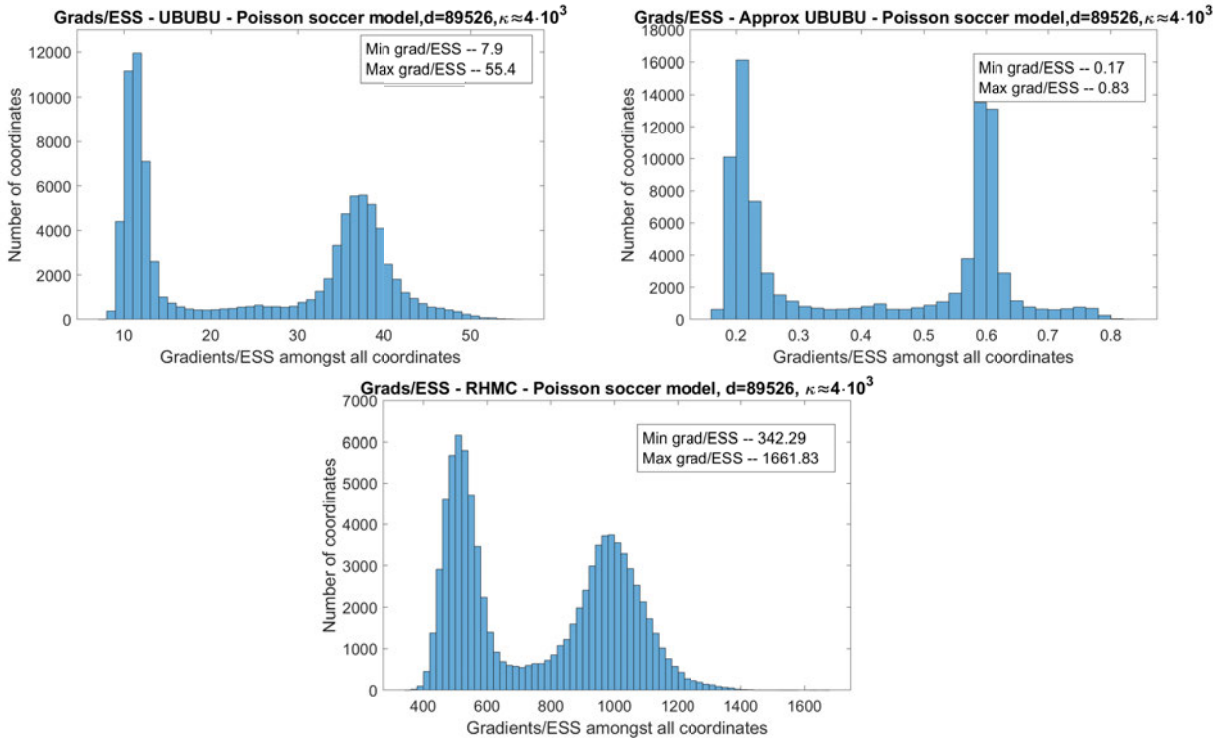


Figure 3.13: Gradient/ESS over all components of a Poisson regression model for soccer scores.

Chapter 4

Conclusion and future directions

To conclude Chapter 2, in [55] it is shown that the optimal convergence rate for the continuous time dynamics is $\mathcal{O}(m/\gamma)$, therefore our contraction rates are consistent up to a constant for the discretizations, however for some of the schemes considered for example BAOAB and OBABO we have that the scheme inherits convergence to the overdamped Langevin dynamics (without time rescaling) and this is reflected in our convergence rate estimates. For MCMC applications, our estimates of convergence rate are independent of γ. In particular, in the case of γ-limit-convergence methods, our convergence guarantees are valid for large friction. BAOAB and OBABO do not suffer from slow convergence in the limit and we do not expect a large bias because they converge to a consistent overdamped Langevin dynamics numerical scheme. This is shown by our provided asymptotic bias estimates for the BAOAB scheme which remain finite in the high-friction limit and provide non-asymptotic guarantees for this scheme. We recommend these schemes in the context of sampling due to their robustness with respect to the choice of friction parameter and increased stability.

The constants in our arguments can be improved by sharper bounds and a more careful analysis, but the restriction on γ is consistent with other works on synchronous coupling for the continuous time Langevin diffusions [20, 47, 55, 57, 166]. Further it is shown in [116, Proposition 4] that the continuous time process yields Wasserstein contraction by synchronous coupling for all M-∇Lipschitz and m-strongly convex potentials U if and only if $M - m < \gamma(\sqrt{M} + \sqrt{m})$ for the norms that we considered. This condition when M is much larger than m is $\mathcal{O}(\sqrt{M})$. It may be possible to achieve convergence rates for small γ, by using a more sophisticated argument like that of [70]. Using a different Lyapunov function or techniques may lead to being able to extend these results to all $\gamma > 0$ [63, 126], but this is beyond the scope of this paper.

The constants for the discretization analysis for BAOAB can be improved. However, the dimension-dependence of the non-asymptotic guarantees is in accordance with other numerical integrators for kinetic Langevin dynamics without the use of randomized midpoint methods (see [146]).

We have shown BAO, OBA, AOB, BAOAB, OBABO and rOABAO have convergence guarantees for stepsizes $\mathcal{O}(1/\sqrt{M})$ and BAOAB, OBABO and rOABAO have the desirable GLC property which is not common amongst the schemes we studied, for example EM, SES, BBK, SPV and SVV. For the choice of parameters which achieve optimal contraction rate, we derive $\mathcal{O}(m/M)$ rates of contraction, which are sharp up to a constant and we achieve this for every scheme that we studied.

We further considered the case of stochastic gradients, where we allow a flexible choice of unbiased gradient estimator under the assumption that the expected variance of the Jacobian of the estimator is bounded. We show that this results in a reduced convergence rate based on the variance of the Jacobian of the estimator, which coincides with what we have observed numerically for small batch sizes in a subsampled stochastic gradient. Most previous results in the literature (see e.g. [53]) require the mean square error of the stochastic gradients $\mathbb{E}(\|G_k - \nabla U(x_k)\|^2)$ to be uniformly bounded, which can be easily violated even for strongly convex and gradient-Lipschitz potentials U_i when using standard subsampling schemes. We do not need such a stringent requirement; our conditions on the stochastic gradients stated in Assumption 2.6.2 are applicable for subsampling-based estimators as long as each U_i is gradient-Lipschitz.

We have provided numerical results comparing the bias of each of the numerical methods based on choices of the friction parameter which are optimal according to our theory or the optimal choice for the Gaussian distribution, where we solved for the convergence rates exactly. We compared the errors of the integrators in a Bayesian logistic regression application and have seen that some of the integrators performed well with large stepsizes, even in the presence of stochastic gradients. Our theoretical and numerical results indicate that using stochastic gradients with advanced numerical integrators can perform well and have significant computational advantages compared to full-gradient methods. In the case of sufficiently large batch sizes, there is little change in terms of the stability threshold and convergence rate compared to the full-gradient version.

In this thesis, in Chapter 2 we have developed theory to show convergence of discretizations of kinetic Langevin dynamics all the way up to the stability threshold (with and without the use of stochastic gradients). This theory was particularly useful in the analysis of the UBUBU method we introduced in Chapter 3.

In Chapter 3, we presented a new unbiased estimator which can exploit high strong-order numerical integrators for underdamped Langevin dynamics. We refer to our estimator as UBUBU which does not rely on the Metropolis acceptance/reject step. Our estimator is influenced by the work of [138], and instead is constructed using a telescoping sum for different discretization levels [79, 129]. We were able to show various theoretical insights, which include showing unbiasedness and finite variance, a central limit theorem, and asymptotic and non-asymptotic bounds on the variance for three algorithms, based on exact, stochastic, and approximate gradients. We have studied the behaviour of our algorithm for product target distributions and shown that for a large class of test functions, it has dimension-independent computational complexity. For stochastic gradients, we also considered the dependency on the size of the data in the big data limit and shown that our method is very efficient in such situations. The proof of these results relies on Wasserstein contraction results for the UBU dynamics. We provided numerical experiments verifying our theory and demonstrating the performance gains over other well-known methods such as randomized HMC. We have considered a range of model problems including an MNIST multinomial regression problem, and a Poisson regression model tested on a real-world dataset. Our comparisons are based on gradient evaluations per effective sample size.

In terms of future work, there are various directions which could be taken up. One of them is related to exploiting higher-order schemes, which were provided in [73, 74]. Numerically, strong orders of up to 4 have been observed. [73] have proven strong order $3/2, 5/2$ and 3 under gradient Lipschitz, Hessian Lipschitz and third-order Lipschitz assumptions, respectively. However, the dimensional dependence obtained under each of these assumptions has not been shown to improve on the UBU scheme in [140]. Furthermore, such splitting schemes typically require more than one gradient evaluation per step, unlike our strategy. In a different direction, one could consider integrators adapted to potentials that do not have the gradient-Lipschitz property (such as in the case of sparsity-inducing priors [122] or log link functions). Other potential directions are nested expectations [162] and static parameter estimation [6, 56]. Finally, one could consider the setting where one does not assume convexity [41, 70, 108, 142], where one could use an alternative coupling to synchronous coupling to achieve theoretical guarantees in the setting where the potential is non-convex with potential applications in sampling Bayesian neural networks, as an alternative to Hamiltonian Monte Carlo.

Appendix A

Appendix of Wasserstein convergence and bias estimates of discretized kinetic Langevin dynamics

A.1 Convergence rates

In this section, we detail proofs for the convergence rate for Euler-Maruyama (the simplest scheme), BAOAB, OBABO and the stochastic Euler scheme. Proofs for the first-order splitting methods, BBK, SVV, SPV and rOABAO can be found in [103, 105].

Proof for Euler-Maruyama. We will denote two synchronous realisations of EM as (x_j, v_j) and $(\tilde{x}_j, \tilde{v}_j)$ for $j \in \mathbb{N}$. Now we will denote $\overline{x}_j := (\tilde{x}_j - x_j)$, $\overline{v}_j = (\tilde{v}_j - v_j)$ and $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$, where $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$ for $j = n, n+1$ for $n \in \mathbb{N}$. We have the following update rule for $\overline{z}_n$

$$\overline{x}_{n+1} = \overline{x}_n + h\overline{v}_n, \qquad \overline{v}_{n+1} = \overline{v}_n - \gamma h \overline{v}_n - hQ\overline{x}_n,$$

where by mean value theorem we define $Q = \int_{t=0}^{1} \nabla^2 U(\tilde{x}_n + t(x_n - \tilde{x}_n))dt$, then $\nabla U(\tilde{x}_n) - \nabla U(x_n) = Q\overline{x}_n$. One can show that in the notation of equation (2.4.14) we have

$$P = \begin{pmatrix} I_d & hI_d \\ -hQ & (1-\gamma h)\, I_d \end{pmatrix}, \tag{A.1}$$

and therefore for Euler-Maruyama (using the notation of equation (2.4.16))

$$\begin{aligned} A &= -c\,(h)\, I_d + 2bhQ - h^2 a Q^2, \\ B &= -bc\,(h)\, I_d + h\left((b\gamma - 1)I_d + (a + h(b - a\gamma))Q\right), \\ C &= \left(-c\,(h)\, a + h\left(2a\gamma - 2b - h(1 - 2b\gamma + a\gamma^2)\right)\right) I_d. \end{aligned}$$

We now invoke Proposition 2.4.2 as A, B and C commute as they are all polynomials in Q. A is positive definite if and only if all its eigenvalues are positive. We note that the eigenvalues of A are precisely $P_A(\lambda) := -c\,(h) + h\left(2b\lambda - ha\lambda^2\right)$, where λ are the eigenvalues of Q, where $m \leq \lambda \leq M$. We wish to show that $P_A(\lambda) > 0$ for all $\lambda \in [m, M]$. This is equivalent to

$$\frac{P_A\,(\lambda)}{h} = -\frac{m}{2\gamma} + \frac{2\lambda}{\gamma} - \frac{h\lambda^2}{M} \geq \lambda\left(-\frac{1}{2\gamma} + \frac{2}{\gamma} - h\right) > 0,$$

which is satisfied when $h < 1/\gamma$. Hence we have that $A \succ 0$. Now it remains to prove that $AC - B^2 \succ 0$, where $AC - B^2$ is a polynomial of Q, which we denote $P_{AC-B^2}(Q)$. Hence it has eigenvalues dictated by the eigenvalues λ of Q. That is the eigenvalues of $AC - B^2$ are $P_{AC-B^2}(\lambda)$ for λ an eigenvalue of Q. Considering P_{AC-B^2} we have

$$\frac{P_{AC-B^2}(\lambda)}{h^2\lambda} = \frac{4}{M} - \frac{4}{\gamma^2} + \frac{m^2}{4\gamma^2 M\lambda} - \frac{m^2}{4\gamma^4\lambda} + \frac{m}{\gamma^2\lambda} - \frac{m}{M\lambda} - \frac{\lambda}{M^2} + h^2\left(\frac{\lambda}{M} - \frac{\lambda}{\gamma^2}\right)$$
$$+ h\left(\frac{2}{\gamma} - \frac{m}{\gamma M} - \frac{m}{2\gamma\lambda} + \frac{m}{\gamma^3} + \frac{m\lambda}{2\gamma M^2} + \frac{\gamma m}{2M\lambda} - \frac{2\gamma}{M}\right) > \frac{1}{M} - h\frac{2\gamma}{M} > 0,$$

where we have used the fact that $\gamma \geq 2\sqrt{M}$ and $h < \frac{1}{2\gamma}$ and hence $AC - B^2 \succ 0$. □

Proof for BAOAB. We first note that $(\mathcal{BAOAB})^n = \mathcal{BAO}(\mathcal{ABAO})^{n-1}\mathcal{AB}$. We will now focus our attention on proving contraction of $\mathcal{ABAO}$, by doing this we only have to deal with a single evaluation of the Hessian at each step. We will denote two synchronous realizations of ABAO as (x_j, v_j) and $(\tilde{x}_j, \tilde{v}_j)$ for $j \in \mathbb{N}$. Now we will denote $\overline{x}_j := (\tilde{x}_j - x_j)$, $\overline{v}_j = (\tilde{v}_j - v_j)$ and $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$, where $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$ for $j = n, n+1$ for $n \in \mathbb{N}$. We have the following update rule for $\overline{z}_n$

$$\overline{x}_{n+1} = \overline{x}_n + h\overline{v}_n - \frac{h^2}{2}Q\left(\overline{x} + \frac{h}{2}\overline{v}\right), \quad \overline{v}_{n+1} = \eta^2\overline{v}_n - h\eta^2 Q\left(\overline{x} + \frac{h}{2}\overline{v}\right),$$

where by mean value theorem we define $Q = \int_{t=0}^{1} \nabla^2 U(\tilde{x}_n + \frac{h}{2}\tilde{v}_n + t(x_n - \tilde{x}_n + \frac{h}{2}(v_n - \tilde{v}_n)))dt$, then $\nabla U\left(\tilde{x}_n + \frac{h}{2}\tilde{v}_n\right) - \nabla U\left(x_n + \frac{h}{2}v_n\right) = Q\left(\overline{x} + \frac{h}{2}\overline{v}\right)$. In the notation of (2.4.16) we have that for this scheme

$$A = -c(h)I_d + \left(2b\eta^2 h + h^2\right)Q + \left(-a\eta^4 h^2 - b\eta^2 h^3 - \frac{1}{4}h^4\right)Q^2,$$
$$B = \left(b\left(1-\eta^2\right) - h - bc(h)\right)I_d + \left(a\eta^4 h + 2b\eta^2 h^2 + \frac{3}{4}h^3\right)Q$$
$$+ \left(-\frac{1}{2}a\eta^4 h^3 - \frac{1}{2}b\eta^2 h^4 - \frac{1}{8}h^5\right)Q^2,$$
$$C = \left(a\left(1-\eta^4\right) - 2b\eta^2 h - h^2 - ac(h)\right)I_d + \left(a\eta^4 h^2 + \frac{3}{2}b\eta^2 h^3 + \frac{1}{2}h^4\right)Q$$
$$+ \left(-\frac{1}{4}a\eta^4 h^4 - \frac{1}{4}b\eta^2 h^5 - \frac{1}{16}h^6\right)Q^2,$$

where $\eta = \exp\{-\gamma h/2\}$. For our choice of a and b, B simplifies to $B = -bc(h) + \left(a\eta^4 h + 2b\eta^2 h^2 + \frac{3}{4}h^3\right)Q + \left(-\frac{1}{2}a\eta^4 h^3 - \frac{1}{2}b\eta^2 h^4 - \frac{1}{8}h^5\right)Q^2$. Now it is sufficient to prove that $A \succ 0$ and that $C - BA^{-1}B \succ 0$, noting that A, B and C commute as they are all polynomials in Q; it is sufficient to prove that $A \succ 0$ and $AC - B^2 \succ 0$. First considering A we have that A is symmetric and hence it is positive definite if and only if all its eigenvalues are positive. We note that the eigenvalues of A are precisely

$$P_A(\lambda) = h\left(-\frac{c(h)}{h} + (2b\eta^2 + h)\lambda + (-a\eta^4 h - b\eta^2 h^2 - \frac{1}{4}h^3)\lambda^2\right)$$
$$\geq h\lambda\left(\frac{7b\eta^2}{4} + \frac{3h}{4} - \eta^4 h - b\eta^2 h^2 M - \frac{h^3 M}{4}\right) \geq h\lambda\left(\frac{b\eta^2}{2} + \frac{3h}{4} - \frac{h}{16}\right) > 0,$$

where λ are the eigenvalues of Q, where $m \leq \lambda \leq M$ and we have used the fact that $h < \frac{1-\eta^2}{\sqrt{4M}}$ and $b \geq h$. Hence we have that $A \succ 0$. Now it remains to prove that $AC - B^2 \succ 0$, now we have that $AC - B^2$ is a polynomial of Q, which we denote $P_{AC-B^2}(Q)$ and hence has eigenvalues dictated by the eigenvalues λ of Q. Now considering

$$\begin{aligned}
&\frac{P_{AC-B^2}(\lambda)}{h\lambda} = \frac{\left(\eta^4-1\right)c(h)}{hM\lambda} + h\left(\frac{1-\eta^4}{M} - \frac{\eta^4\lambda}{M^2} + \frac{2\eta^2\left(1-\eta^4\right)}{\left(1-\eta^2\right)M} + \frac{c(h)^2}{h^2M\lambda}\right)\\
&+ h^2\frac{c(h)}{h}\left(-\frac{\left(1+\eta^2\right)^2}{M} + \frac{1+\eta^2}{\left(1-\eta^2\right)\lambda} + \frac{\eta^4\lambda}{M^2}\right)\\
&+ h^3\left(-1 - \frac{4\eta^2}{\left(1-\eta^2\right)^2} - \frac{c(h)^2}{h^2\left(1-\eta^2\right)^2\lambda} - \frac{\lambda}{4M} + \frac{3\eta^4\lambda}{4M} - \frac{\eta^2\lambda\left(1-\eta^4\right)}{\left(1-\eta^2\right)M}\right)\\
&+ h^4\frac{c(h)}{h}\left(1 + \frac{4\eta^2}{\left(1-\eta^2\right)^2} + \frac{\lambda\left(1+\eta^4\right)}{4M} + \frac{\eta^2\lambda}{M}\right) + h^5\lambda\left(\frac{3}{16} + \frac{\eta^2}{\left(1-\eta^2\right)^2}\right)\\
&+ h^6\frac{c(h)}{h}\left(-\frac{3\lambda}{16} - \frac{\eta^2\lambda}{\left(1-\eta^2\right)^2}\right) \geq -\frac{\left(1+\eta^2\right)h}{4M} + h\left(\frac{1+2\eta^2}{M}\right)\\
&+ h^3\left(-\frac{5}{4} - \frac{4\eta^2}{\left(1-\eta^2\right)^2} - \frac{1}{64\left(1-\eta^2\right)^2} - \eta^2\left(1+\eta^2\right)\right) - \frac{h^3}{32} > 0,
\end{aligned}$$

where we have used the fact that $h < \frac{1-\eta^2}{\sqrt{4M}}$. Hence $AC-B^2 \succ 0$ and our contraction results hold. All computations can be checked using symbolic computing. We can bound the $\mathcal{AB}$ operator on $||\cdot||_{a,b}$ by

$$\begin{aligned}
&||\psi_{\mathrm{AB}}\left(\tilde{x}_n, \tilde{v}_n, h/2\right) - \psi_{\mathrm{AB}}\left(x_n, v_n, h/2\right)||_{a,b}^2 \leq\\
&3\left(\left(1 + \frac{ah^2M^2}{2}\right)||\overline{x}_n||^2 + \left(a + \frac{h^2}{4} + \frac{ah^4M^2}{8}\right)||\overline{v}_n||^2\right) \leq 7||\overline{x}_n, \overline{v}_n||_{a,b}^2,
\end{aligned}$$

where we have used the norm equivalence in Section 2.2.2. We can also bound

$$\begin{aligned}
&||\psi_{\mathrm{O}}\left(\psi_{\mathrm{BA}}(\tilde{x}_n, \tilde{v}_n, h/2), h\right) - \psi_{\mathrm{O}}\left(\psi_{\mathrm{BA}}(x_n, v_n, h/2), h\right)||_{a,b}^2\\
&\leq 3\left(\left(1 + \frac{ah^2M^2}{4} + \frac{h^4M^2}{8}\right)||\overline{x}_n||^2 + \left(\frac{h^2}{2} + a\right)||\overline{v}_n||^2\right) \leq 7||\overline{x}_n, \overline{v}_n||_{a,b}^2.
\end{aligned}$$

Combining these estimates we have the required result. □

Proof for OBABO. We first note that $(\mathcal{OBABO})^n = \mathcal{OB}\left(\mathcal{ABOB}\right)^{n-1}\mathcal{ABO}$. We will now focus our attention on proving contraction of $\mathcal{ABOB}$. Note we only have to deal with a single evaluation of the Hessian at each step as the position variable is not updated between gradient evaluations. We will denote two synchronous realisations of ABOB as (x_j, v_j) and $(\tilde{x}_j, \tilde{v}_j)$ for $j \in \mathbb{N}$. Now we will denote $\overline{x}_j := (\tilde{x}_j - x_j)$, $\overline{v}_j = (\tilde{v}_j - v_j)$ and $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$, where $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$ for $j = n, n+1$ for $n \in \mathbb{N}$. We have the following update rule for $\overline{z}_n$

$$\overline{x}_{n+1} = \overline{x}_n + h\overline{v}_n, \quad \overline{v}_{n+1} = \eta^2\overline{v}_n - \frac{h}{2}(\eta^2+1)Q\left(\overline{x} + h\overline{v}\right),$$

where by mean value theorem we define $Q = \int_{t=0}^{1} \nabla^2 U(\tilde{x}_n + h\tilde{v}_n + t(x_n - \tilde{x}_n + h(v_n - \tilde{v}_n)))dt$, then $\nabla U\left(\tilde{x}_n + h\tilde{v}_n\right) - \nabla U\left(x_n + hv_n\right) = Q\left(\overline{x} + h\overline{v}\right)$. In the notation of (2.4.16) we have that for this scheme

$$
\begin{aligned}
A &= -c\left(h\right) I_d + bh\left(1+\eta^2\right) Q - (1+\eta^2)^2 \frac{ah^2Q^2}{4}, \\
B &= \left(b\left(1-\eta^2\right) - h - bc\left(h\right)\right) I_d + \left(\frac{1}{2}a\eta^2 + bh\right)\left(\eta^2+1\right) hQ - \left(\eta^2+1\right)^2 \frac{ah^3}{4} Q^2, \\
C &= \left(a\left(1-\eta^4\right) - 2b\eta^2 h - h^2 - ac\left(h\right)\right) I_d + \left(a\eta^2 + bh\right)\left(\eta^2+1\right) h^2 Q - a\left(\eta^2+1\right)^2 \frac{h^4}{4} Q^2,
\end{aligned}
$$

where $\eta = \exp\{-\gamma h/2\}$. This form motivates the choice $b = \frac{h}{1-\eta^2}$ and $a = \frac{1}{M}$ inspired by the continuous dynamics. For our choice of a and b, B simplifies to $B = -bc\left(h\right) + (\frac{1}{2}a\eta^2 + bh)(\eta^2 + 1)hQ - (\eta^2+1)^2\frac{ah^3}{4}Q^2$. We will now apply Proposition 2.4.2, first considering A we have that the eigenvalues are precisely

$$
P_A\left(\lambda\right) = -c\left(h\right) + bh(1+\eta^2)\lambda - (1+\eta^2)^2\frac{ah^2\lambda^2}{4} \geq h\lambda\left(\frac{3b}{4} + b\eta^2 - \frac{h}{4} - \frac{3b\eta^2}{4}\right) > 0,
$$

where λ are the eigenvalues of Q, where $m \leq \lambda \leq M$ and we have used the fact that $b \geq h$. Hence we have that $A \succ 0$. Now it remains to prove that $AC - B^2 \succ 0$, now we have that $AC - B^2$ is a polynomial of Q, which we denote $P_{AC-B^2}(Q)$ and hence has eigenvalues dictated by the eigenvalues λ of Q. Now considering

$$
\begin{aligned}
&\frac{P_{AC-B^2}\left(\lambda\right)}{h\lambda} = \frac{\left(\eta^4-1\right)c\left(h\right)}{hM\lambda} + h\left(\frac{\left(1+\eta^2\right)^2}{M} + \frac{c\left(h\right)^2}{h^2M\lambda} - \frac{\left(1+\eta^2\right)^2\lambda}{4M^2}\right) \\
&+ h^2\frac{c\left(h\right)}{h}\left(-\frac{\left(1+\eta^2\right)^2}{M} - \frac{1}{\lambda} + \frac{2}{\left(1-\eta^2\right)\lambda} + \frac{\lambda\left(1+\eta^2\right)^2}{4M^2}\right) \\
&+ h^3\left(-\frac{\left(1+\eta^2\right)^2}{\left(1-\eta^2\right)^2} - \frac{c\left(h\right)^2}{\left(1-\eta^2\right)^2 h^2\lambda} - \frac{\lambda\left(1+\eta^2\right)^2}{4M} + \frac{\lambda\left(1+\eta^2\right)^2}{2\left(1-\eta^2\right)M}\right) \\
&+ h^4\frac{c\left(h\right)}{h}\left(\frac{\left(1+\eta^2\right)^2}{\left(1-\eta^2\right)^2} + \frac{\lambda\left(1+\eta^2\right)^2}{4M} - \frac{\lambda\left(1+\eta^2\right)^2}{2M\left(1-\eta^2\right)}\right) \\
&> -\frac{h\left(1+\eta^2\right)}{4M} + h\left(\frac{3\left(1+\eta^2\right)^2}{4M}\right) - h\left(\frac{3\left(1+\eta^2\right)^2}{64M}\right) + h\left(-\frac{\left(1+\eta^2\right)^2}{4M} - \frac{1}{64M}\right) > 0,
\end{aligned}
$$

where we have used the fact that $h < \frac{1-\eta^2}{\sqrt{4M}}$. Hence $AC - B^2 \succ 0$ and our contraction results hold. All computations can be checked using symbolic computing. We can bound the $\mathcal{ABO}$ operator on $||\cdot||_{a,b}$ by

$$
||\psi_{\mathrm{BO}}\left(\psi_{\mathrm{A}}\left(\tilde{x}_n, \tilde{v}_n, h\right), h/2\right) - \psi_{\mathrm{BO}}\left(\psi_{\mathrm{A}}\left(x_n, v_n, h\right), h/2\right)||_{a,b}^2 \leq 8||\overline{x}_n, \overline{v}_n||_{a,b}^2,
$$

where we have used the norm equivalence in Section 2.2.2. Similarly, we can also bound

$$
||\psi_{\mathrm{OB}}\left(\tilde{x}_n, \tilde{v}_n, h/2\right) - \psi_{\mathrm{OB}}\left(x_n, v_n, h/2\right)||_{a,b}^2 \leq 6||\overline{x}_n, \overline{v}_n||_{a,b}^2.
$$

Combining these estimates we have the required result. □

Proof for the Stochastic exponential Euler scheme. We remark that synchronous coupling between two realisations of the stochastic Euler scheme results in a synchronous coupling of $(\zeta_n, \omega_n)_{n\in\mathbb{N}}$. Now we will denote $\overline{x}_j := (\tilde{x}_j - x_j)$, $\overline{v}_j = (\tilde{v}_j - v_j)$ and $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$, where $\overline{z}_j = (\overline{x}_j, \overline{v}_j)$ for $j = n, n+1$ for $n \in \mathbb{N}$. We have the following update rule for $\overline{z}_n$

$$\overline{x}_{n+1} = \overline{x}_n + \frac{1-\eta^2}{\gamma}\overline{v}_n - \frac{\gamma h + \eta^2 - 1}{\gamma^2}Q\overline{x}_n, \quad \overline{v}_{n+1} = \eta^2\overline{v}_n - \frac{1-\eta^2}{\gamma}Q\overline{x}_n,$$

where by mean value theorem we define $Q = \int_{t=0}^{1} \nabla^2 U(\tilde{x}_n + t(x_n - \tilde{x}_n))dt$, then $\nabla U(\tilde{x}_n) - \nabla U(x_n) = Q(\overline{x} + h\overline{v})$. In the notation of (2.4.16) we have that for this scheme

$$\begin{aligned}
A &= -c(h) I_d + 2\left(\frac{b\left(1-\eta^2\right)}{\gamma} + \frac{\eta^2 - 1 + \gamma h}{\gamma^2}\right) Q \\
&\quad - \left(\frac{a\left(1-\eta^2\right)^2}{\gamma^2} + \frac{2b\left(1-\eta^2\right)\left(-1+\eta^2+\gamma h\right)}{\gamma^3} + \frac{\left(-1+\eta^2+\gamma h\right)^2}{\gamma^4}\right) Q^2, \\
B &= \left(b\left(1-\eta^2\right) - \frac{\left(1-\eta^2\right)}{\gamma} - bc(h)\right) I_d \\
&\quad + \left(\frac{a\eta^2\left(1-\eta^2\right)}{\gamma} + \frac{b\left(1-\eta^2\right)^2}{\gamma^2} + \frac{b\eta^2\left(-1+\eta^2+\gamma h\right)}{\gamma^2} + \frac{\left(1-\eta^2\right)\left(-1+\eta^2+\gamma h\right)}{\gamma^3}\right) Q, \\
C &= \left(a(1-\eta^4) - ac(h) - \frac{2b\eta^2\left(1-\eta^2\right)}{\gamma} - \frac{\left(1-\eta^2\right)^2}{\gamma^2}\right) I_d,
\end{aligned}$$

where $\eta = \exp\{-\gamma h/2\}$. This form motivates the choice $b = \frac{1}{\gamma}$ and $a = \frac{1}{M}$ inspired by the continuous dynamics. For our choice of a and b, B simplifies to $B = -bc(h) + \mathcal{O}(Q)$. We will now apply Proposition 2.4.2, first considering A we wish to show that all its eigenvalues are positive which are precisely

$$\begin{aligned}
P_A(\lambda) &:= -c(h) + 2\left(\frac{b\left(1-\eta^2\right)}{\gamma} + \frac{\eta^2 - 1 + \gamma h}{\gamma^2}\right)\lambda \\
&\quad - \left(\frac{a\left(1-\eta^2\right)^2}{\gamma^2} + \frac{2b\left(1-\eta^2\right)\left(-1+\eta^2+\gamma h\right)}{\gamma^3} + \frac{\left(-1+\eta^2+\gamma h\right)^2}{\gamma^4}\right)\lambda^2, \\
&\geq h\lambda\left(\frac{7}{4\gamma} - \left(h + \frac{h^2 M}{\gamma} + \frac{h^3 M}{4}\right)\right) > 0,
\end{aligned}$$

where λ are the eigenvalues of Q, where $m \leq \lambda \leq M$ and using the fact that, $\gamma^2 \geq 4M$, $1-\eta^2 \leq h\gamma$, $h\gamma + \eta^2 - 1 \leq \frac{(h\gamma)^2}{2}$ and $h < \frac{1}{2\gamma}$. Hence we have that $A \succ 0$. Now it remains to prove that $AC - B^2 \succ 0$, now we have that $AC - B^2$ is a polynomial of Q, which we denote $P_{AC-B^2}(Q)$ and hence has eigenvalues dictated by the eigenvalues λ of Q. Since the terms are more complicated than the previous discretizations we choose a convenient way of expanding the expression which can obtain positive definiteness. That is to expand the expression in terms of a. By using symbolic computing

one can show that $P_{AC-B^2}(\lambda) = c_0 + c_1 a + c_2 a^2$, where

$$
\begin{aligned}
c_1 + c_2 a &= \left(\eta^4 - 1\right) c\left(h\right) + c\left(h\right)^2 + 2\left(1-\eta^4\right)\left(\frac{b\left(1-\eta^2\right)}{\gamma} + \frac{-1+\eta^2+\gamma h}{\gamma^2}\right)\lambda \\
&+ \frac{2b\left(1-\eta^2\right)\eta^2 c\left(h\right)\lambda}{\gamma} - 2\left(\frac{b\left(1-\eta^2\right)}{\gamma} + \frac{-1+\eta^2+\gamma h}{\gamma^2}\right) c\left(h\right)\lambda + \frac{\left(1-\eta^2\right)^4\lambda^2}{\gamma^4} \\
&- \frac{2\eta^2\left(1-\eta^2\right)^2\left(-1+\eta^2+\gamma h\right)\lambda^2}{\gamma^4} - \frac{2b\left(1-\eta^2\right)\left(-1+\eta^2+\gamma h\right)\lambda^2}{\gamma^3} \\
&- \frac{\left(1-\eta^4\right)\left(-1+\eta^2+\gamma h\right)^2\lambda^2}{\gamma^4} + \frac{2b\left(1-\eta^2\right)\left(-1+\eta^2+\gamma h\right) c\left(h\right)\lambda^2}{\gamma^3} \\
&+ \frac{\left(-1+\eta^2+\gamma h\right)^2 c\left(h\right)\lambda^2}{\gamma^4} + a\left(-\frac{\left(1-\eta^2\right)^2\lambda^2}{\gamma^2} + \frac{\left(1-\eta^2\right)^2 c\left(h\right)\lambda^2}{\gamma^2}\right) \\
&\geq c_1 - \frac{\left(1-\eta^2\right)^2\lambda}{\gamma^2} + \frac{\left(1-\eta^2\right)^2 c\left(h\right)\lambda}{\gamma^2} \\
&\geq \left(\eta^4 - 1\right) c\left(h\right) + \left(1-\eta^4\right)\left(\frac{2h}{\gamma} - \frac{1-\eta^2}{\left(1+\eta^2\right)\gamma^2}\right)\lambda + ... \\
&\geq \lambda\left(\left(-\frac{h^2}{2}\right) + h^2\left(2 - \frac{1}{1+\eta^2}\right) + ...\right) \\
&> \lambda\left(\frac{h^2}{2} - \frac{h^2}{16} - \frac{h^3\gamma}{16} - \frac{h^3\gamma}{8} - \frac{h^3\gamma}{16}\right) \geq \lambda\left(\frac{7h^2}{16} - \frac{h^3\gamma}{4}\right),
\end{aligned}
$$

where $h < \frac{1}{2\gamma}$, $\gamma^2 \geq 8M \geq 8m$, $\frac{h\gamma}{2} \leq 1 - \eta^2 \leq h\gamma$ and $h\gamma + \eta^2 - 1 \leq \frac{(h\gamma)^2}{2}$. Further, we have that

$$
\begin{aligned}
c_0 &= \frac{\left(1-\eta^2\right)^2 c\left(h\right)}{\gamma^2} + \frac{2b\left(1-\eta^2\right)\eta^2 c\left(h\right)}{\gamma} - b^2 c\left(h\right)^2 - \frac{2b\left(1-\eta^2\right)^3\lambda}{\gamma^3} \\
&- \frac{2\left(1-\eta^2\right)^2\left(-1+\eta^2+\gamma h\right)\lambda}{\gamma^4} - \frac{4b^2\eta^2\left(1-\eta^2\right)^2\lambda}{\gamma^2} - \frac{4b\eta^2\left(1-\eta^2\right)\left(-1+\eta^2+\gamma h\right)\lambda}{\gamma^3} \\
&- \frac{b^2\eta^4\lambda^2(\eta^2+\gamma h-1)^2}{\gamma^4} + \frac{2b^2(1-\eta^2)^2 c\left(h\right)\lambda}{\gamma^2} + \frac{2b^2\eta^2 c\left(h\right)\lambda(\eta^2+\gamma h-1)}{\gamma^2} \\
&+ \frac{2b^2\eta^2(1-\eta^2)^2\lambda^2(\eta^2+\gamma h-1)}{\gamma^4} - \frac{b^2(1-\eta^2)^4\lambda^2}{\gamma^4} + \frac{2b(1-\eta^2)c\left(h\right)\lambda(\eta^2+\gamma h-1)}{\gamma^3} \\
&> \lambda\left(-\frac{2\left(h\gamma\right)^3}{\gamma^4} - \frac{\left(h\gamma\right)^4}{\gamma^4} - \frac{4\left(h\gamma\right)^2}{\gamma^4} - \frac{2\left(h\gamma\right)^3}{\gamma^4} - \frac{\left(h\gamma\right)^4\lambda}{4\gamma^6} - \frac{\left(h\gamma\right)^4\lambda}{\gamma^6}\right) > \lambda\left(-\frac{7h^2}{\gamma^2}\right),
\end{aligned}
$$

now we can combine this with the previous estimate and we have

$$
P_{AC-B^2}(\lambda) > \lambda\left(\frac{7h^2}{16M} - \frac{h^3\gamma}{4M} - \frac{7h^2}{\gamma^2}\right) > h^2\lambda\left(\frac{5}{16M} - \frac{7}{\gamma^2}\right) \geq 0,
$$

which is true when $\gamma \geq 5\sqrt{M}$. Hence $AC - B^2 \succ 0$ and our contraction results hold. All computations can be checked using symbolic computing. □

A.2 Asymptotic bias of BAOAB

Proof of Proposition 2.7.2. We start by estimating $\triangle_x$. We use the following Taylor expansion for the Hamiltonian dynamics

$$\psi_{\mathsf{H}}(z,h) = \left(x + hv - \int_0^h \nabla U(x(t))(h-t)dt, v - h\nabla U(x) - \int_0^h \nabla^2 U(x(t))v(t)(h-t)dt\right),$$

we then define $(\overline{x}(t), \overline{v}(t))$ to be Hamiltonian dynamics initialised at $\psi_{\mathsf{O}}(\psi_{\mathsf{H}}(z, h/2), h)$ at time $t > 0$ and $(\overline{x}, \overline{v}) := (\overline{x}(0), \overline{v}(0))$. Then the x-component of the HOH scheme is

$$\begin{aligned} x_{\mathsf{HOH}} := x + \frac{h}{2}v(1+\eta^2) - \int_0^{h/2} \nabla U(x(t))\left(\frac{h}{2} - t\right)dt - \int_0^{h/2} \nabla U(\overline{x}(t))\left(\frac{h}{2} - t\right)dt \\ + \frac{h}{2}\left(\eta^2\left(-\frac{h}{2}\nabla U(x) - \int_0^{h/2} \nabla^2 U(x(t))v(t)\left(\frac{h}{2} - t\right)dt\right) + \sqrt{1-\eta^4}\xi\right) \end{aligned}$$

and the x-component of the BAOAB scheme is

$$x_{\mathsf{BAOAB}} := x' + \frac{h}{2}v'(1+\eta^2) - \frac{h^2}{4}(1+\eta^2)\nabla U(x') + \frac{h}{2}\sqrt{1-\eta^4}\xi.$$

Therefore the difference $\triangle_x = x_{\mathsf{HOH}} - x_{\mathsf{BAOAB}}$ in x satisfies

$$\begin{aligned} \|\triangle_x\|_{L^2} \leq \left(1 + \eta^2\frac{h^2}{4}M\right)\|x - x'\|_{L^2} + \frac{h}{2}(1+\eta^2)\|v - v'\|_{L^2} \\ + \left\| \int_0^{h/2} \left(\nabla U(x(t)) - \nabla U(x')\right)\left(\frac{h}{2} - t\right)dt - \eta^2\frac{h}{2}\int_0^{h/2} \nabla^2 U(x(t))v(t)\left(\frac{h}{2} - t\right)dt \right. \\ \left. + \int_0^{h/2} \left(\nabla U(\overline{x}(t)) - \nabla U(x')\right)\left(\frac{h}{2} - t\right)dt \right\|_{L^2}. \end{aligned}$$

We now bound the final expression by

$$\begin{aligned} M\int_0^{h/2} \left\|x(t) - x'\right\|_{L^2}\left(\frac{h}{2} - t\right)dt + \frac{h\eta^2 M}{2}\int_0^{h/2} \|v(t)\|_{L^2}\left(\frac{h}{2} - t\right)dt \\ + M\int_0^{h/2} \left\|\overline{x}(t) - x'\right\|_{L^2}\left(\frac{h}{2} - t\right)dt, \end{aligned}$$

where we have that the second term is bounded by $\frac{h^3\eta^2 M}{16}\sqrt{d}$ and considering the first term we have for $t \in [0, h/2]$

$$\|x(t) - x'\|_{L^2} \leq \|x - x'\|_{L^2} + \|tv - \int_0^t \nabla U(x(s))(t-s)ds\|_{L^2} \leq \|x - x'\|_{L^2} + \frac{3h\sqrt{d}}{4},$$

similarly we can bound $M\int_0^{h/2} \|\overline{x}(t) - x'\|_{L^2}\left(\frac{h}{2} - t\right)dt$. Therefore we have by summing the estimates

$$\|\triangle_x\|_{L^2} \leq \left(1 + (1+\eta^2)\frac{h^2}{4}M\right)\|x - x'\|_{L^2} + \frac{h}{2}(1+\eta^2)\|v - v'\|_{L^2} + \frac{3h^3 M\sqrt{d}}{8},$$

where we have used the fact that $h < \frac{1}{2\sqrt{M}}$. Now we estimate $\triangle_v$, considering the velocity components we have that the v-component of the HOH scheme is

$$v_{\mathsf{HOH}} := \eta^2 \left(v - \frac{h}{2}\nabla U(x) - \int_0^{h/2} \nabla^2 U(x(t))v(t) \left(\frac{h}{2} - t\right) dt \right) + \sqrt{1-\eta^4}\xi$$
$$- \frac{h}{2}\nabla U(\overline{x}) - \int_0^{h/2} \nabla^2 U(\overline{x}(t))\overline{v}(t) \left(\frac{h}{2} - t\right) dt,$$

where $\overline{x} := x + \frac{h}{2}v - \int_0^{h/2} \nabla U(x(t)) \left(\frac{h}{2} - t\right) dt$. The v-component of BAOAB is

$$v_{\mathsf{BAOAB}} := \eta^2(v' - \frac{h}{2}\nabla U(x')) + \sqrt{1-\eta^4}\xi - \frac{h}{2}\nabla U(\hat{x}),$$

where $\hat{x} := x' + \frac{h}{2}(1+\eta^2)v' - \frac{h^2}{4}(1+\eta^2)\nabla U(x') + \frac{h}{2}\sqrt{1-\eta^4}\xi$. For $\triangle_v = v_{\mathsf{HOH}} - v_{\mathsf{BAOAB}}$ we have that

$$\triangle_v = \eta^2(v - v') - \frac{h\eta^2}{2}\left(\nabla U(x) - \nabla U(x')\right) - \eta^2 \int_0^{h/2} \nabla^2 U(x(t))v(t) \left(\frac{h}{2} - t\right) dt$$
$$- \frac{h}{2}\left(\nabla U(\overline{x}) - \nabla U(\hat{x})\right) - \int_0^{h/2} \nabla^2 U(\overline{x}(t))\overline{v}(t) \left(\frac{h}{2} - t\right) dt,$$

we now consider the Taylor expansion

$$\nabla U(\hat{x}) = \nabla U(\hat{x}_c)$$
$$- \nabla^2 U([\hat{x}_c, \hat{x}]) \left(x - x' + \frac{h}{2}(1+\eta^2)(v - v') - \frac{h^2}{4}(1+\eta^2)(\nabla U(x) - \nabla U(x')) \right),$$

where we define

$$\hat{x}_c := x + \frac{h}{2}(1+\eta^2)v - \frac{h^2}{4}(1+\eta^2)\nabla U(x) + \frac{h}{2}\sqrt{1-\eta^4}\xi,$$
$$\nabla^2 U([v_1, v_2]) := \int_0^1 \nabla^2 U(v_1 + s(v_2 - v_1))ds,$$

for any $v_1, v_2 \in \mathbb{R}^d$. We then define

$$\alpha_v := \eta^2(v - v') - \frac{h\eta^2}{2}\left(\nabla U(x) - \nabla U(x')\right)$$
$$- \frac{h}{2}\nabla^2 U([\hat{x}_c, \hat{x}]) \left(x - x' + \frac{h}{2}(1+\eta^2)(v - v') - \frac{h^2}{4}(1+\eta^2)\left(\nabla U(x) - \nabla U(x')\right) \right)$$
$$= (\eta^2 I_d - \frac{h^2(1+\eta^2)}{4}Q_2)(v - v') + \left(-\frac{h\eta^2}{2}Q_1 - \frac{h}{2}Q_2 + \frac{h^3(1+\eta^2)}{8}Q_2Q_1 \right)(x - x'),$$

where $Q_1 = \nabla^2 U([x, x'])$ and $Q_2 = \nabla^2 U([\hat{x}, \hat{x}_c])$. Consider $\triangle_v - \alpha_v$, which can be written as

$$-\eta^2 \int_0^{h/2} \nabla^2 U(x(t))v(t) \left(\frac{h}{2} - t\right) dt + \frac{h}{2}\nabla^2 U([\hat{x}_c, \overline{x}])(\hat{x}_c - \overline{x}) - \int_0^{h/2} \nabla^2 U(\overline{x}(t))\overline{v}(t) \left(\frac{h}{2} - t\right) dt,$$

which only contains terms from the continuous dynamics. Removing some third order terms which we can bound in L^2 by $h^3M^{3/2}\sqrt{d}$, where we have used that $h < \frac{1}{2\sqrt{M}}$, we have the additional terms are given by

$$
\begin{aligned}
&-\eta^2\int_0^{h/2}\nabla^2U(x(t))v\left(\frac{h}{2}-t\right)dt+\frac{h}{2}\nabla^2U([\hat{x}_c,\overline{x}])\left(\frac{h}{2}\eta^2v+\frac{h}{2}\sqrt{1-\eta^4}\xi\right)\\
&-\int_0^{h/2}\nabla^2U(\overline{x}(t))\left(\eta^2\tilde{v}+\sqrt{1-\eta^4}\xi\right)\left(\frac{h}{2}-t\right)dt,
\end{aligned}
$$

where $\tilde{v} := v-\frac{h}{2}\nabla U(x)-\int_0^{h/2}\nabla^2U(x(t))v(t)\left(\frac{h}{2}-t\right)dt$ and we can bound in L^2

$$
\begin{aligned}
&\left\|\eta^2\int_0^{h/2}\nabla^2U(x(t))v\left(\frac{h}{2}-t\right)dt-\eta^2\frac{h^2}{4}\nabla^2U([\hat{x}_c,\overline{x}])v+\eta^2\int_0^{h/2}\nabla^2U(\overline{x}(t))\tilde{v}\left(\frac{h}{2}-t\right)dt\right\|_{L^2}\\
&\leq\frac{\eta^2h^3M^{3/2}\sqrt{d}}{8}+\eta^2\left(\left\|\int_0^{h/2}\left(\nabla^2U(x(t))-\nabla^2U([\hat{x}_c,\overline{x}])\right)v\left(\frac{h}{2}-t\right)dt\right\|_{L^2}\right.\\
&\left.+\left\|\int_0^{h/2}\left(\nabla^2U(\overline{x}(t))-\nabla^2U([\hat{x}_c,\overline{x}])\right)v\left(\frac{h}{2}-t\right)dt\right\|_{L^2}\right),
\end{aligned}
$$

and you can bound the second term under Assumption 2.2.3 by

$$
\begin{aligned}
&\eta^2M_1\int_0^{h/2}\int_0^1\left\|\|x(t)-\hat{x}_c-s(\overline{x}-\hat{x}_c)\|\,\|v\|\right\|_{L^2}\left(\frac{h}{2}-t\right)dsdt\\
&\leq\eta^2M_1\int_0^{h/2}\left\|\left(2h\|v\|+\frac{3}{2}h^2\sqrt{Md}+h\sqrt{1-\eta^4}\|\xi\|\right)\|v\|\right\|_{L^2}\left(\frac{h}{2}-t\right)dt\leq\frac{\eta^2h^3M_1d}{2}
\end{aligned}
$$

and similarly for the third term. Without Assumption 2.2.3, we can bound each of these terms by $\eta^2\frac{h^2M\sqrt{d}}{4}$.

The remaining terms we have are

$$
\sqrt{1-\eta^4}\frac{h^2}{4}\nabla^2U([\hat{x}_c,\overline{x}])\xi-\sqrt{1-\eta^4}\int_0^{h/2}\nabla^2U(\overline{x}(t))\left(\frac{h}{2}-t\right)\xi dt, \tag{A.2}
$$

which can be bounded by $3h^2M\sqrt{d}/8$ in L^2. For higher order estimates we have that the ∇^2U terms contain noise and hence we use the Taylor expansions

$$
\begin{aligned}
&\nabla^2U(\overline{x}(t))=\nabla^2U(\overline{x})+\nabla^3U([\overline{x}(t),\overline{x}])\left(t\overline{v}-\int_0^t\nabla U(\overline{x}(s))(t-s)ds\right),\\
&\nabla^2U([\hat{x}_c,\overline{x}])=\int_0^1\nabla^2U(\hat{x}+s(\overline{x}-\hat{x})-(1-s)\frac{h}{2}\sqrt{1-\eta^4}\xi)ds\\
&+\frac{h}{2}\sqrt{1-\eta^4}\int_0^1\nabla^3U([\hat{x}+s(\overline{x}-\hat{x}),\hat{x}+s(\overline{x}-\hat{x})-(1-s)\frac{h}{2}\sqrt{1-\eta^4}\xi])(1-s)\xi ds.
\end{aligned}
$$

Therefore we have (A.2) is of the form $\sqrt{1-\eta^4}h^2A(x,x')\xi+r_h$, where

$$
A(x,x')=\frac{1}{8}\left(2\int_0^1\nabla^2U(\hat{x}+s(\overline{x}-\hat{x})-(1-s)\frac{h}{2}\sqrt{1-\eta^4}\xi)ds-\nabla^2U(\overline{x})\right)
$$

is independent of ξ, $\|A(x,x')\xi\|_{L^2} \leq \frac{3}{8}M\sqrt{d}$, and $\|r_h\|_{L^2} \leq h^3 M_1 d$. Combining all the estimates we have

$$\triangle_v = \left(\eta^2 I_d - \frac{h^2(1+\eta^2)}{4}Q_2\right)(v-v') + \left(-\frac{h\eta^2}{2}Q_1 - \frac{h}{2}Q_2 + \frac{h^3(1+\eta^2)}{8}Q_2Q_1\right)(x-x') + \epsilon_v + h^2\sqrt{1-\eta^4}A(x,x')\xi,$$

where $\|\epsilon_v\|_{L^2} \leq 2h^3\sqrt{d}(M^{3/2} + M_1\sqrt{d})$ under Assumptions 2.2.1-2.2.3 and $\|\epsilon_v\|_{L^2} \leq 2h^2 M\sqrt{d}$ with $A(x,x') = 0$ under Assumptions 2.2.1-3.3.7. □

We will use the following proposition to control the evolution of the error between BAOAB and HOH steps.

Proposition A.2.1. *Consider an HOH scheme, $(x_i, v_i)_{i\in\mathbb{N}}$ and a BAOAB scheme $(x'_i, v'_i)_{i\in\mathbb{N}}$ initialized at $(x_0, v_0) = (x'_0, v'_0) = (x, v) \sim \pi$ in $\mathbb{R}^{2d}$ with synchronously coupled Gaussian increments and stepsize $h < \frac{1-\eta^2}{2\sqrt{M}}$, for $l \in \mathbb{N}$ we define $(\triangle_x^l, \triangle_v^l) := (x_l - x'_l, v_l - v'_l)$. For $a = 1/M$ and $b = h/(1-\eta^2)$ we have that under Assumptions 2.2.1-3.3.7*

$$\|(\triangle_x^l, \triangle_v^l)\|_{L^2,a,b} \leq \sqrt{3}e^{(l-1)5h\sqrt{M}}C_l,$$

where $C_l = \frac{3lh^3M\sqrt{d}}{8} + \frac{2h^2\sqrt{Md}}{1-\eta^2}$. Additionally, if Assumption 2.2.3 is satisfied we have

$$\|(\triangle_x^l, \triangle_v^l)\|_{L^2,a,b} \leq \sqrt{3}e^{(l-1)4h\sqrt{M}}C_l,$$

where $C_l := 3h^3\sqrt{d}\left(M + \frac{M_1}{\sqrt{M}}\sqrt{d}\right)l + \frac{3h^2\sqrt{Md}}{8}$.

Proof of Proposition A.2.1. Let $(x'_0, v'_0) = (x_0, v_0) \sim \pi$ and $(x_i, v_i)_{i=1}^l$ be defined by the HOH scheme with stepsize h and friction parameter γ and $(x'_i, v'_i)_{i=1}^l$ be defined by the BAOAB scheme with the same stepsize and friction parameter, we define these such that they have synchronously coupled Gaussian increments in the O steps. Then let us define $\triangle_x^i := x_i - x'_i$ and $\triangle_v^i := v_i - v'_i$ for $i \in \mathbb{N}$, we have by Proposition 2.7.2 that

$$\begin{aligned}\|\triangle_x^l\|_{L^2} &\leq \|\triangle_x^{l-1}\|_{L^2} + h\left(2hM\|\triangle_x^{l-1}\|_{L^2} + \|\triangle_v^{l-1}\|_{L^2}\right) + \frac{3h^3M\sqrt{d}}{8} \\ &\leq \sum_{i=1}^{l-1} h\left(2hM\|\triangle_x^i\|_{L^2} + \|\triangle_v^i\|_{L^2}\right) + l\frac{3h^3M\sqrt{d}}{8}.\end{aligned}$$

Without the additional Assumption 2.2.3 we have

$$\begin{aligned}\|\triangle_v^l\|_{L^2} &\leq (\eta^2 + \frac{h^2M}{2})\|\triangle_v^{l-1}\|_{L^2} + hM\left(2 + \frac{h^2M}{4}\right)\|\triangle_x^{l-1}\|_{L^2} + 2h^2M\sqrt{d} \\ &\leq \sum_{i=1}^{l-1}\eta^{2(l-1-i)}\frac{h}{2}\left(5M\|\triangle_x^i\|_{L^2} + hM\|\triangle_v^i\|_{L^2}\right) + \sum_{i=1}^{l}\eta^{2(l-i)}2h^2M\sqrt{d},\end{aligned}$$

and therefore we have $\|(\triangle_x^l, \triangle_v^l)\|_{L^2,a,0} \leq \sum_{i=1}^{l-1} 5h\sqrt{M}\|(\triangle_x^i, \triangle_v^i)\|_{L^2,a,0} + C_l$, where $C_l = \frac{3lh^3M\sqrt{d}}{8} + \frac{2h^2\sqrt{Md}}{1-\eta^2}$ and hence

$$\|(\triangle_x^l, \triangle_v^l)\|_{L^2,a,0} \leq e^{(l-1)5h\sqrt{M}} C_l.$$

With Assumption 2.2.3 we have

$$\triangle_v^l = \sum_{i=1}^{l-1} \eta^{2(l-1-i)} \left(A_v^i \triangle_v^i + A_x^i \triangle_x^i\right) + \sum_{i=1}^{l} \eta^{2(l-i)} \epsilon_v^i + \sum_{i=1}^{l} \eta^{2(l-i)} h^2 \sqrt{1-\eta^4} A_i(x_i, x_i')\xi_i,$$

where ξ_i is the noise increment from the iteration i of BAOAB, A_i and ϵ_v^i are defined by Proposition 2.7.2 for each iteration and $A_x^i := -\frac{h\eta^2}{2} Q_1^i - \frac{h}{2} Q_2^i + \frac{h^3(1+\eta^2)}{8} Q_2^i Q_1^i$, $A_v^i := -\frac{h^2(1+\eta^2)}{4} Q_2^i$, where Q_1^i and Q_2^i are defined by Proposition 2.7.2 for each i. Therefore

$$\|\triangle_v^l\|_{L^2} \leq \sum_{i=1}^{l-1} \frac{h^2 M}{2} \|\triangle_v^i\|_{L^2} + 2hM\|\triangle_x^i\|_{L^2} + 2h^3\sqrt{d}\left(M^{3/2} + M_1\sqrt{d}\right) l + h^2 \frac{3M\sqrt{d}}{8},$$

where we only lose an order of $1/2$ in the last term due to the independence of the Gaussian increments, more precisely for $i \neq j$ and without loss of generality assume $i > j$ we have $\mathbb{E}\langle A_i\xi_i, A_j\xi_j\rangle = \mathbb{E}_{\xi_j}\mathbb{E}\left[\langle A_i\xi_i, A_j\xi_j\rangle \mid \xi_j\right] = 0$. We therefore have that

$$\|(\triangle_x^l, \triangle_v^l)\|_{L^2,a,0} \leq \sum_{i=1}^{l-1} 4h\sqrt{M}\|(\triangle_x^i, \triangle_v^i)\|_{L^2,a,0} + C_l \leq e^{(l-1)4h\sqrt{M}} C_l,$$

where $C_l := 3h^3\sqrt{d}\left(M + \frac{M_1}{\sqrt{M}}\sqrt{d}\right) l + \frac{3h^2\sqrt{Md}}{8}$. □

The previous error estimate can be refined using our contraction results, hence we can avoid blowing up exponentially in l.

Proof of Proposition 2.7.3. We have $(\triangle_x^l, \triangle_v^l) := (x_l - x_l', v_l - v_l')$, where (x_l, v_l) corresponds to l steps of HOH initiated in $(x_0, v_0) = (x, v) \sim \pi$, and (x_l', v_l') corresponds to l steps of BAOAB initiated in $(x_0', v_0') = (x, v)$, driven by the same noise.

We define a sequence of interpolating variants $(x_l^{(k)}, v_l^{(k)})$ for every $k = 0, 1, \ldots, l$ as follows. First, $(x_0^{(k)}, v_0^{(k)}) = (x, v)$. We then define $(x_i^{(k)}, v_i^{(k)})_{i=1}^{k}$ as HOH steps, followed by $(x_i^{(k)}, v_i^{(k)})_{i=k+1}^{l}$ as BAOAB steps. So we take k HOH steps, followed by $l-k$ BAOAB steps. From the definition, it follows that $(x_l^{(l)}, v_l^{(l)}) = (x_l, v_l)$ and $(x_l'^{(0)}, v_l'^{(0)}) = (x_l', v_l')$. We break l steps into blocks of size $\tilde{l} = \left\lceil \frac{1}{2\sqrt{M}h} \right\rceil$, as follows,

$$\begin{aligned}
\|(\triangle_x^l, \triangle_v^l)\|_{L^2,a,b} &= \left\| \left(x_l^{(0)} - x_l^{(l)}, v_l^{(0)} - v_l^{(l)}\right) \right\|_{L^2,a,b} \\
&\leq \sum_{j=0}^{\lfloor l/\tilde{l} \rfloor - 1} \left\| \left(x_l^{(j\tilde{l})} - x_l^{((j+1)\tilde{l})}, v_l^{(j\tilde{l})} - v_l^{((j+1)\tilde{l})}\right) \right\|_{L^2,a,b} \\
&+ \left\| \left(x_l^{(\lfloor l/\tilde{l} \rfloor \tilde{l})} - x_l^{(l)}, v_l^{(\lfloor l/\tilde{l} \rfloor \tilde{l})} - v_l^{(l)}\right) \right\|_{L^2,a,b}.
\end{aligned}$$

When we bound the terms $\left\|\left(x_l^{(j\tilde{l})} - x_l^{((j+1)\tilde{l})}, v_l^{(j\tilde{l})} - v_l^{((j+1)\tilde{l})}\right)\right\|_{L^2,a,b}$, we can use the fact that the first $j\tilde{l}$ according to HOH keep the stationary distribution invariant, and the two chains deviate in the following $\tilde{l}$ steps (since one of them is doing HOH, and the other one is doing BAOAB steps). Still, the remaining steps are doing BAOAB for both chains (hence, there is a contraction). Using Proposition A.2.1 with l chosen as $\tilde{l}$, and Theorem 2.4.5, we have

$$\left\|\left(x_l^{(j\tilde{l})} - x_l^{((j+1)\tilde{l})}, v_l^{(j\tilde{l})} - v_l^{((j+1)\tilde{l})}\right)\right\|_{L^2,a,b} \leq \sqrt{3}e^{5/2}C_{\tilde{l}} \cdot 7(1-c(h))^{\frac{l-1-(j+1)\tilde{l}}{2}}.$$

Under Assumptions 2.2.1-3.3.7, $C_{\tilde{l}} = \frac{3\tilde{l}h^3M\sqrt{d}}{8} + \frac{2h^2\sqrt{Md}}{1-\eta^2} \leq \frac{3h^2\sqrt{Md}}{1-\eta^2}$. If we also include Assumption 2.2.3, $C_{\tilde{l}} := 3h^3\sqrt{d}\left(M + \frac{M_1}{\sqrt{M}}\sqrt{d}\right)\tilde{l} + \frac{3h^2\sqrt{Md}}{8} \leq h^2(4\sqrt{Md} + 3\frac{M_1}{M}d)$. By some simple algebra, we have that

$$\|\triangle_x^l, \triangle_v^l\|_{L^2,a,b} \leq \sqrt{3}e^{5/2}C_{\tilde{l}}\left(1 + \frac{7}{1-(1-c(h))^{\tilde{l}/2}}\right) \leq 170C_{\tilde{l}} \cdot \frac{1}{1-e^{-c(h)\tilde{l}/2}}$$
$$= 170C_{\tilde{l}} \cdot \frac{1}{1-e^{-\frac{h^2m}{4(1-\eta^2)}\tilde{l}/2}} \leq 170C_{\tilde{l}} \cdot \frac{1}{1-e^{-\frac{hm}{8\sqrt{M}(1-\eta^2)}}} \leq 170C_{\tilde{l}} \cdot \left(1 + \frac{8\sqrt{M}\left(1-\eta^2\right)}{hm}\right),$$

where we have used the fact that $1/(1-e^{-x}) \leq 1 + 1/x$ for $x > 0$. The claims now follow by rearrangement. □

A.3 Stochastic gradient kinetic Langevin dynamics integrators

For the Euler-Maruyama, stochastic Euler scheme, rOABAO, stochastic position Verlet only one force evaluation is used in each iteration, so every gradient evaluation is taken as a stochastic gradient estimate. The complete algorithms are stated below in Algorithms 4,5, 6 and 7.

Algorithm 4 Stochastic Gradient Euler-Maruyama (EM)

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- for $k = 1, 2, ..., K$ do
  - Sample $W_k \sim \rho$
  - $G_{k-1} \to \mathcal{G}(x_{k-1}, W_k)$
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - $x_k \to x_{k-1} + hv_{k-1}$
  - $v_k \to v_{k-1} - hG_{k-1} - h\gamma v_{k-1} + \sqrt{2\gamma h}\xi_k$
- Output: Samples $(x_k)_{k=0}^K$.

Algorithm 5 Stochastic Gradient Stochastic Euler Scheme (SES/EB)

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- for $k = 1, 2, ..., K$ do
  - Sample $(\zeta_k, \omega_k) \sim \mathcal{N}(0_{2d}, \Sigma)$, where Σ is given in (2.4.12).
  - Sample $W_k \sim \rho$
  - $G_{k-1} \to \mathcal{G}(x_{k-1}, W_k)$
  - $x_k \to x_{k-1} + \frac{1-\eta^2}{\gamma}v_{k-1} - \frac{\gamma h+\eta^2-1}{\gamma^2}G_{k-1} + \zeta_k$
  - $v_k \to \eta^2 v_{k-1} - \frac{1-\eta^2}{\gamma}G + \omega_k$
- Output: Samples $(x_k)_{k=0}^K$.

Algorithm 6 Stochastic Gradient rOABAO

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- for $k = 1, 2, ..., K$ do
  - Sample $u_k \sim \mathcal{U}(0, h)$
  - Sample $W_k \sim \rho$
  - $G_{k-1} \to \mathcal{G}(x_{k-1} + u_k v_{k-1}, W_k)$
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v \to \eta v_{k-1} + \sqrt{1-\eta^2}\xi_k$
  - $x_k \to x_{k-1} + h v_{k-1} - \frac{h^2}{2} G_{k-1}$
  - $v_k \to v_{k-1} - h G_{k-1}$
  - Sample $\xi'_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v_k \to \eta v + \sqrt{1-\eta^2}\xi'_k$
- Output: Samples $(x_k)_{k=0}^K$.

Algorithm 7 Stochastic Gradient Stochastic Velocity Verlet (SVV)

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- for $k = 1, 2, ..., K$ do
  - (A) $x \to x_{k-1} + \frac{h}{2} v_{k-1}$
  - Sample $W_k \sim \rho$
  - $G_{k-1} \to \mathcal{G}(x, W_k)$
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - ($\mathcal{V}$) $v_k \to \eta^2 v_{k-1} - \frac{1-\eta^2}{\gamma} G_{k-1} + \sqrt{1-\eta^4}\xi_k$
  - (A) $x_k \to x + \frac{h}{2} v_k$
- Output: Samples $(x_k)_{k=0}^K$.

For BAOAB the first and last $\mathcal{B}$ of each iteration share a stochastic gradient evaluation to make the algorithm roughly one gradient evaluation per step. The complete algorithm is given in Algorithm 8.

Algorithm 8 Stochastic Gradient BAOAB

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- Sample $W_1 \sim \rho$
- $G_0 \to \mathcal{G}(x_0, W_1)$
- for $k = 1, 2, ..., K$ do
  - (B) $v \to v_{k-1} - \frac{h}{2} G_{k-1}$
  - (A) $x \to x_{k-1} + \frac{h}{2} v$
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v \to \eta^2 v + \sqrt{1-\eta^4}\xi_k$
  - (A) $x_k \to x + \frac{h}{2} v$
  - Sample $W_{k+1} \sim \rho$
  - $G_k \to \mathcal{G}(x_k, W_{k+1})$
  - (B) $v_k \to v - \frac{h}{2} G_k$
- Output: Samples $(x_k)_{k=0}^K$.

Similarly for OBABO the first and last $\mathcal{B}$ of each iteration share a stochastic gradient evaluation to make the algorithm roughly one gradient evaluation per step. The complete algorithm is given in Algorithm 9.

Algorithm 9 Stochastic Gradient OBABO

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- Sample $W_1 \sim \rho$
- $G_0 \to \mathcal{G}(x_0, W_1)$
- for $k = 1, 2, ..., K$ do
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v \to \eta v_{k-1} + \sqrt{1-\eta^2}\xi_k$
  - (B) $x \to x_{k-1} - \frac{h}{2} G_{k-1}$
  - (A) $x \to x + hv$
  - Sample $W_{k+1} \sim \rho$
  - $G_k \to \mathcal{G}(x, W_{k+1})$
  - (B) $x_k \to x - \frac{h}{2} G_k$
  - Sample $\xi'_k \sim \mathcal{N}(0_d, I_d)$
  - (O) $v_k \to \eta v + \sqrt{1-\eta^2}\xi'_k$
- Output: Samples $(x_k)_{k=0}^K$.

If we express each iteration of the BBK methods as $\Phi_{\mathrm{B}_2} \circ \Phi_{\mathrm{A}} \circ \Phi_{\mathrm{B}_1}$, then the last Φ_{B_2} step of each iteration and the first Φ_{B_1} of the next iteration share the same stochastic gradient evaluation to make the algorithm roughly one gradient evaluation per step. The complete algorithm is given in Algorithm 10.

Algorithm 10 Stochastic Gradient Brunger-Brooks Karplus (BBK)

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- Sample $W_1 \sim \rho$
- $G_0 \to \mathcal{G}(x_0, W_1)$
- for $k = 1, 2, ..., K$ do
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - (B_1) $v \to v_{k-1} + \frac{h}{2}\left(-G_{k-1} - \gamma v_{k-1} + \frac{\sqrt{2\gamma}}{\sqrt{h}}\xi_k\right)$
  - (A) $x_k \to x_{k-1} + hv$
  - Sample $W_{k+1} \sim \rho$
  - $G_k \to \mathcal{G}(x_k, W_{k+1})$
  - Sample $\xi'_k \sim \mathcal{N}(0_d, I_d)$
  - (B_2) $v_k \to \left(1 + \frac{h}{2}\gamma\right)^{-1}\left(v - \frac{h}{2}G_k + \frac{\sqrt{2\gamma}}{\sqrt{h}}\xi'_k\right)$
- Output: Samples $(x_k)_{k=0}^K$.

Finally for SVV the last $\mathcal{V}$ step of each iteration and the first $\mathcal{V}$ of the next iteration share the same stochastic gradient evaluation. The complete algorithm is given in Algorithm 11.

Algorithm 11 Stochastic Gradient Stochastic Velocity Verlet (SVV)

- Initialize $(x_0, v_0) \in \mathbb{R}^{2d}$, stepsize $h > 0$ and friction parameter $\gamma > 0$.
- Sample $W_1 \sim \rho$
- $G_0 \to \mathcal{G}(x_0, W_1)$
- for $k = 1, 2, ..., K$ do
  - Sample $\xi_k \sim \mathcal{N}(0_d, I_d)$
  - ($\mathcal{V}$) $v \to \eta v_{k-1} - \frac{1-\eta}{\gamma} G_{k-1} + \sqrt{1-\eta^2}\xi_k$
  - (A) $x_k \to x_{k-1} + hv$
  - $G_k \to \mathcal{G}(x_k, W_{k+1})$
  - Sample $\xi'_k \sim \mathcal{N}(0_d, I_d)$
  - ($\mathcal{V}$) $v_k \to \eta v - \frac{1-\eta}{\gamma} G_k + \sqrt{1-\eta^2}\xi'_k$
- Output: Samples $(x_k)_{k=0}^K$.

Appendix B

Appendix of Unbiased kinetic Langevin Monte Carlo

B.1 Discussion and outline of results

The beginning of this appendix is devoted to providing a road-map for our results. In Appendix B.2, which follows, we provide variance estimates of the full gradient multilevel UBUBU method. The approach we use is to bound $\mathrm{Var}(D_0)$ using Theorem 2 of [89] and to use the strong error estimates of [140] for UBU to estimate $\mathrm{Var}(D_{l,l+1})$. [89] requires Ricci curvature of the UBU Markov chain and extending [140] to global strong error estimates in Appendix B.4.1 requires Wasserstein convergence. We provide this in Appendix B.3 in the full gradient setting using the methods of [105]. We provide L^4 Lyapunov drift inequalities in the full gradient setting. We can then bound the average distance to the minimizer non-asymptotically, the key result needed to get complexity bounds in the big data setting. We also provide the proof of the central limit theorem of the estimator in Appendix B.2.

In Appendix B.5 we describe the initialization and the OHO scheme for the approximate and stochastic gradient methods and some estimates of the distance between the initial measure and the target measure. We then use the techniques of [168] to provide global strong error estimates of the SVRG method. We combine and extend the techniques of [140] and [168] to prove new non-asymptotic stochastic gradient error bounds for the UBU integrator. From this we obtain in Appendix B.4 variance bounds and estimates on our estimator UBUBU with exact gradients. This is then extended to providing estimates of the variance of our multilevel estimator in the SVRG stochastic gradient setting in Appendix B.6.

We further develop bounds for our new approximate gradient UBU method in Appendix B.7 using the same techniques, in the approximate gradient setting. In general, Appendices B.6, B.7 follow similarly where one requires bounds on the variance of the quantity D_0 and $D_{l,l+1}$. However, we use an interpolation argument to improve the results in Appendix B.7 as opposed to the methods used in Appendix B.6. We also use some classical results from the theory of ODEs to establish bounds between continuous diffusions to establish the variance of $D_{0,1}$ in Appendix B.6 and B.7. Finally, we provide some auxiliary results in Appendix B.8.

B.2 Unbiased multilevel estimators

Proposition 3.3.4. *Suppose that Assumptions 3.3.1, 3.3.2 and 3.3.3 hold, and that* $2 < \phi_N < \phi_D$. *Then* S *as defined in* (3.3.16) *is an unbiased estimator of* $\mu(f)$ *that has finite variance*

$$\mathrm{Var}(S) \leq \frac{\mathrm{Var}(D_0)}{N} + \frac{V_D}{N\underline{c}_N\left(1-\frac{\phi_N}{\phi_D}\right)},$$

and finite expected computational cost.

Similarly, for any $c_R \in [0,1)$, $S(c_R)$ *as defined in* (3.3.17) *is also an unbiased estimator of* $\mu(f)$ *with finite variance*

$$\mathrm{Var}(S(c_R)) \leq \\ \frac{\mathrm{Var}(D_0)}{N} + \frac{V_D}{N\underline{c}_N\left(1-\frac{\phi_N}{\phi_D}\right)} + \frac{V_D}{N\underline{c}_N\left(1-\frac{\phi_N}{\phi_D}\right)}\frac{2}{(1-c_R)^2}\left(\frac{\phi_N}{\phi_D}\right)^{\log(2\underline{c}_N N/\phi_N)/\log(\phi_N)},$$

and finite expected computational cost.

Proof of Proposition 3.3.4. From Assumption 3.3.3, and the definition of S, it follows that the expected computational cost of S is upper bounded as follows:

$$\begin{aligned}&\mathcal{O}\left(N + \sum_{l=0}^{\infty}\mathbb{E}(N_{l,l+1})2^l(K+lB+B_0)\right)\\ &\leq \mathcal{O}\left(N\left(1+\bar{c}_N\sum_{l=0}^{\infty}\left(\frac{2}{\phi_N}\right)^l(K+lB+B_0)\right)\right) < \infty.\end{aligned}$$

From Assumption 3.3.1, and the independence of the terms, we have that

$$\begin{aligned}\mathrm{Var}(S) &\leq \frac{\mathrm{Var}(D_0)}{N} + \sum_{l=0}^{\infty}\frac{\mathbb{E}(D_{l,l+1}^2)}{\mathbb{E}N_{l,l+1}} \leq \frac{\mathrm{Var}(D_0)}{N} + \frac{V_D}{\underline{c}_N N}\sum_{l=0}^{\infty}\left(\frac{\phi_D}{\phi_N}\right)^{-l}\\ &= \frac{\mathrm{Var}(D_0)}{N} + \frac{V_D}{N\underline{c}_N\left(1-\frac{\phi_N}{\phi_D}\right)} < \infty.\end{aligned}$$

By Jensen's inequality, and Assumption 3.3.1, $\mathbb{E}(|S_0| + \sum_{l=0}^{\infty}|S_{l,l+1}|) < \infty$, hence by the dominated convergence theorem,

$$\mathbb{E}(S) = \mathbb{E}(S_0) + \sum_{l=0}^{\infty}\mathbb{E}(S_{l,l+1}) = \tilde{\mu}_{h_0}(f) + \sum_{l=0}^{\infty}\tilde{\mu}_{h_{l+1}}(f) - \tilde{\mu}_{h_l}(f) = \mu(f),$$

which concludes the proof for S.

For $S(c_R)$, the computational cost is the same as for S, so it has finite expectation. For the variance, we have

$$\mathrm{Var}(S(c_R)) \leq \frac{\mathrm{Var}(D_0)}{N} + \sum_{l=0}^{L(N)-1} \frac{\mathbb{E}(D_{l,l+1}^2)}{\mathbb{E}N_{l,l+1}} + \mathrm{Var}\left(\frac{S_{L(N),L(N)+1}}{1-c_R} + \sum_{l=L(N)+1}^{\infty} \overline{S}_{l,l+1} \right).$$

The last term can be bounded as

$$\begin{aligned}
&\mathrm{Var}\left(\frac{S_{L(N),L(N)+1}}{1-c_R} + \sum_{l=L(N)+1}^{\infty} \overline{S}_{l,l+1} \right) \\
&= \mathrm{Var}\left(S_{L(N),L(N)+1} \left(\frac{1}{1-c_R} - \sum_{l=L(N)+1}^{\infty} \frac{\mathbb{1}[N_{l,l+1}=1]}{\mathbb{E}(N_{l,l+1})} c_R^{l-L(N)} \right) + \sum_{l=L(N)+1}^{\infty} S_{l,l+1} \right) \\
&\leq 2 \cdot \mathrm{Var}\left(S_{L(N),L(N)+1} \left(\frac{1}{1-c_R} - \sum_{l=L(N)+1}^{\infty} \frac{\mathbb{1}[N_{l,l+1}=1]}{\mathbb{E}(N_{l,l+1})} c_R^{l-L(N)} \right) \right) \\
&+ 2 \cdot \mathrm{Var}\left(\sum_{l=L(N)+1}^{\infty} S_{l,l+1} \right) \leq \frac{2}{(1-c_R)^2} \mathbb{E}(S_{L(N),L(N)+1}^2) + 2 \sum_{l=L(N)+1}^{\infty} \mathbb{E}(S_{l,l+1}^2).
\end{aligned}$$

As before, we have $\mathbb{E}(S_{l,l+1}^2) \leq \frac{\mathbb{E}(D_{l,l+1}^2)}{\mathbb{E}N_{l,l+1}} \leq \frac{V_D}{\underline{c}_N N} \cdot \left(\frac{\phi_D}{\phi_N}\right)^{-l}$ for any $l \geq 0$. Using the fact that $\phi_N^{-L(N)-1} \leq \frac{1}{2\underline{c}_N N}$, we have $\phi_N^{L(N)} \geq \frac{2\underline{c}_N N}{\phi_N}$, hence $L(N) \geq \frac{\log(2\underline{c}_N N/\phi_N)}{\log(\phi_N)}$. After some rearrangement, we obtain that

$$\begin{aligned}
&\mathrm{Var}(S(c_R)) \leq \\
&\frac{\mathrm{Var}(D_0)}{N} + \frac{V_D}{N\underline{c}_N\left(1-\frac{\phi_N}{\phi_D}\right)} + \frac{V_D}{N\underline{c}_N\left(1-\frac{\phi_N}{\phi_D}\right)} \frac{2}{(1-c_R)^2} \left(\frac{\phi_N}{\phi_D}\right)^{\log(2\underline{c}_N N/\phi_N)/\log(\phi_N)},
\end{aligned}$$

where the factor $\left(\frac{\phi_N}{\phi_D}\right)^{\log(2\underline{c}_N N/\phi_N)/\log(\phi_N)}$ tends to 0 as $N \to \infty$. Finally, unbiasedness can be shown as before using the dominated convergence theorem. □

We show below that a central limit theorem holds for these estimators.

Theorem 3.3.5. *Under the assumptions of Proposition 3.3.4, we have that, as $N \to \infty$,*

$$\sqrt{N}(S - \mu(f)) \Rightarrow \mathcal{N}(0, \sigma_S^2) \quad \textit{and} \quad \sqrt{N}(S(c_R) - \mu(f)) \Rightarrow \mathcal{N}(0, \sigma_S^2),$$

where

$$\sigma_S^2 := \mathrm{Var}(D_0) + \sum_{l=0}^{\infty} \frac{\mathrm{Var}(D_{l,l+1})}{c_{l,l+1}}. \tag{3.3.18}$$

Proof of Theorem 3.3.5. First, we prove the result for $H := \sqrt{N}(S - \mu(f))$. For $l_{\max} \geq 0$, let

$$\begin{aligned} H^{l_{\max}} &:= \sqrt{N}(S_0 - \mathbb{E}(S_0)) + \sum_{l=0}^{l_{\max}} (S_{l,l+1} - \mathbb{E}(S_{l,l+1})) \\ &= \frac{1}{\sqrt{N}} \sum_{r=1}^{N} \left(D_0^{(r)} - \mathbb{E}(D_0^{(r)}) \right) + \sum_{l=0}^{l_{\max}} \frac{\sqrt{N}}{\mathbb{E}(N_{l,l+1})} \sum_{r=1}^{N_{l,l+1}} D_{l,l+1}^{(r)} - \mathbb{E}(D_{l,l+1}) \\ &:= H_0 + \sum_{l=0}^{l_{\max}} H_{l,l+1}. \end{aligned}$$

Then by using independence, and the fact that $\left(\frac{\sqrt{N}}{\mathbb{E}(N_{l,l+1})}\right) / \left(\frac{1}{\sqrt{N}}\right) \to \frac{1}{c_{l,l+1}}$, by the proof of the central limit theorem (see Sections 3.3-3.4 of [68]), for every $t \in \mathbb{R}$, H_0 and $H_{l,l+1}$ satisfies

$$\begin{aligned} &\mathbb{E}(e^{itH_0}) \to e^{-t^2 \mathcal{V}_0/2} \text{ as } N \to \infty \text{ for } \mathcal{V}_0 = \mathrm{Var}(D_0), \\ &\mathbb{E}(e^{itH_l}) \to e^{-t^2 \mathcal{V}_{l,l+1}/2} \text{ as } N \to \infty \text{ for } \mathcal{V}_{l,l+1} = \frac{\mathrm{Var}(D_{l,l+1})}{c_{l,l+1}}. \end{aligned}$$

Using independence, we can multiply these together to obtain that for any $t \in \mathbb{R}$,

$$\mathbb{E}\left(e^{itH^{l_{\max}}}\right) \to e^{-t^2(\mathcal{V}_0 + \sum_{l=0}^{l_{\max}} \mathcal{V}_{l,l+1})/2} \text{ as } N \to \infty.$$

By Lemma 3.3.19 of [68], it follows that for a random variable X with $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^2) < \infty$, we have

$$|\mathbb{E}(e^{itX}) - 1| \leq \frac{t^2 \mathbb{E}(X^2)}{2}.$$

For $X = \sqrt{N}(S - \mu(f)) - H^{l_{\max}}$, we have

$$\mathbb{E}(X^2) = \mathrm{Var}(X) \leq N \sum_{l=l_{\max}+1}^{\infty} \frac{\mathbb{E}(D_{l,l+1}^2)}{\mathbb{E}N_{l,l+1}} \leq \frac{V_D}{\underline{c}_N} \sum_{l=l_{\max}+1}^{\infty} \left(\frac{\phi_D}{\phi_N}\right)^{-l} \leq \frac{V_D}{\underline{c}_N} \frac{\left(\frac{\phi_N}{\phi_D}\right)^{l_{\max}}}{1 - \frac{\phi_N}{\phi_D}}.$$

Using independence of $H^{l_{\max}}$ and $H - H^{l_{\max}}$, $\mathbb{E}(e^{itH}) = \mathbb{E}(e^{itH^{l_{\max}}}) \cdot \mathbb{E}(e^{it(H - H^{l_{\max}})})$, so

$$\begin{aligned} &\limsup_{N \to \infty} \left| \mathbb{E}(e^{itH}) - e^{-t^2(\mathcal{V}_0 + \sum_{l=0}^{l_{\max}} \mathcal{V}_{l,l+1})/2} \right| \\ &\leq e^{-t^2(\mathcal{V}_0 + \sum_{l=0}^{l_{\max}} \mathcal{V}_{l,l+1})/2} \cdot \frac{V_D}{\underline{c}_N} \sum_{l=l_{\max}+1}^{\infty} \left(\frac{\phi_D}{\phi_N}\right)^{-l} \leq \frac{V_D}{\underline{c}_N} \frac{\left(\frac{\phi_N}{\phi_D}\right)^{l_{\max}}}{1 - \frac{\phi_N}{\phi_D}}. \end{aligned}$$

By letting $l_{\max} \to \infty$, it follows that $\limsup_{N \to \infty} \mathbb{E}(e^{itH}) = e^{-t^2 \sigma_S^2}$, hence the convergence follows by the Lévy-Cramér continuity theorem (see Theorem 3.3.17 of [68]).

The proof for $S(c_R)$ follows the same lines, except that the variances of the terms for $l \geq L(N)$ need to be controlled separately using the same bounds as in the proof of Proposition 3.3.4, we omit the details. □

B.3 Convergence results

The first set of results we prove are provided below for the convergence of the UBU scheme. Proving contraction of a coupling has been a popular method for establishing convergence rates both in the continuous time setting and for the discretization for Langevin dynamics (underdamped/kinetic) and Hamiltonian Monte Carlo (see for example [22–24, 52, 55, 57, 65, 70, 80, 115, 116, 140, 141] and many more).

Our approach to obtain convergence rates is based on proving contraction for a synchronous coupling. We need an appropriate metric to attain convergence, and contraction of the UBU scheme. We introduce the Wasserstein distance in this metric.

Definition B.3.1 (Weighted Euclidean norm)**.** *For $z = (x, v) \in \mathbb{R}^{2d}$ we introduce the weighted Euclidean norm*

$$||z||_{a,b}^2 = ||x||^2 + 2b\,\langle x, v\rangle + a\,||v||^2\,,$$

for $a, b > 0$ with $b^2 < a$.

Remark B.3.2. *Using the assumption $b^2 < a$, we can show that this is equivalent to the Euclidean norm on $\mathbb{R}^{2d}$. Under the condition $b^2 \leq a/4$, we have*

$$\frac{1}{2}\min(a,1)\|z\|^2 \leq \frac{1}{2}||z||_{a,0}^2 \leq ||z||_{a,b}^2 \leq \frac{3}{2}||z||_{a,0}^2 \leq \frac{3}{2}\max(a,1)\|z\|^2. \tag{B.1}$$

Definition B.3.3 (p-Wasserstein distance)**.** *Let us define $\mathcal{P}_p(\mathbb{R}^{2d})$ to be the set of probability measures which have p-th moment for $p \in [1, \infty)$ (i.e. $\mathbb{E}(\|Z\|^p) < \infty$). Then the p-Wasserstein distance in norm $\|\cdot\|_{a,b}$ between two measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{2d})$ is defined as*

$$\mathcal{W}_{p,a,b}(\nu,\mu) = \Big(\inf_{\xi\in\Gamma(\nu,\mu)} \int_{\mathbb{R}^{2d}} \|z_1 - z_2\|_{a,b}^p d\xi(z_1, z_2)\Big)^{1/p}, \tag{B.2}$$

where $\|\cdot\|_{a,b}$ is the norm introduced before and that $\Gamma(\nu,\mu)$ is the set of measures with respective marginals of ν and μ.

Before we proceed, we need to introduce the concept of Wasserstein convergence, which most of the results rely upon.

Lemma B.3.4 (Wasserstein convergence)**.** *Let $1 \leq p \leq \infty$, $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{2d})$, and $a, b > 0$ with $b^2 < a$. Let us assume that $(z_k)_{k\geq 0} = (x_k, v_k)_{k\geq 0}$ and $(\tilde{z}_k)_{k\geq 0} = (\tilde{x}_k, \tilde{v}_k)_{k\geq 0}$ are two Markov chains with state space Λ and kernel P_h defined on the same probability space (a coupling) such that $z_0 \sim \nu$, $\tilde{z}_0 \sim \mu$, and $\mathbb{E}(\|z_0 - \tilde{z}_0\|^p) = [\mathcal{W}_{p,a,b}(\nu,\mu)]^p$. If the following contractive property holds,*

$$\Big[\mathbb{E}(\|\tilde{z}_{k+1} - z_{k+1}\|_{a,b}^p | z_{0:k}, \tilde{z}_{0:k})\Big]^{1/p} \leq (1 - c(h))\|\tilde{z}_k - z_k\|_{a,b} \quad \text{for every } k \geq 0, \tag{B.3}$$

then we have

$$\mathcal{W}_{p,a,b}\left(\nu P_h^n, \mu P_h^n\right) \leq (1 - c(h))^n \mathcal{W}_{p,a,b}(\nu,\mu) \quad \text{for every } n \geq 0.$$

Remark B.3.5. *The existence of an optimal coupling satisfying that $\mathbb{E}(\|z_0 - \tilde{z}_0\|_{a,b}^p) = [\mathcal{W}_{p,a,b}(\nu,\mu)]^p$ follows by Theorem 4.1 of [159].*

Proof. By induction, we have $\mathbb{E}(\|\tilde{z}_n - z_n\|_{a,b}^p | z_0, \tilde{z}_0) \leq (1 - c(h))^n \|z_0 - \tilde{z}_0\|_{a,b}^p$, and the result follows by taking expectations and using Definition B.2. □

Now, we present our first proposition, a convergence result of the UBU scheme with full gradients.

Proposition B.3.6. *Suppose that U is m-strongly convex and M-∇Lipschitz. Let*

$$a = \frac{1}{M}, \quad b = \frac{1}{\gamma}, \quad c_2(h) = \frac{mh}{4\gamma}, \quad c(h) = \frac{mh}{8\gamma}. \tag{B.4}$$

Let P_h denote the transition kernel for a step of UBU with stepsize h. For all $\gamma \geq \sqrt{8M}$, $h < \frac{1}{2\gamma}$, $1 \leq p \leq \infty$, $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{2d})$, (B.2) *holds. Hence for all $n \in \mathbb{N}$,*

$$\mathcal{W}_{p,a,b}\left(\nu P_h^n, \mu P_h^n\right) \leq (1 - c_2(h))^{n/2} \mathcal{W}_{p,a,b}\left(\nu, \mu\right) \leq (1 - c(h))^n \mathcal{W}_{p,a,b}\left(\nu, \mu\right).$$

Further to this, P_h has a unique invariant measure π_h satisfying that $\pi_h \in \mathcal{P}_p(\mathbb{R}^{2d})$ for all $1 \leq p \leq \infty$.

Remark B.3.7. *We are going to use the same choices of a and b as stated in* (B.4) *everywhere in the paper.*

Corollary B.3.8. *Suppose that U is an m-strongly convex M-∇Lipschitz potential, $\gamma \geq \sqrt{8M}$, $1 \leq p \leq 2$, $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{2d})$. Suppose that $(X_0, V_0) \sim \mu$, then the solution of* (3.1.1) *exists in the strong sense for any $t \geq 0$, and the corresponding Markov kernel P_t^{cont} satisfies*

$$\mathcal{W}_{p,a,b}\left(\nu P_t^{\text{cont}}, \mu P_t^{\text{cont}}\right) \leq \exp\left(-\frac{mt}{8\gamma}\right) \mathcal{W}_{p,a,b}\left(\nu, \mu\right) \quad \text{for} \quad a = \frac{1}{M}, b = \frac{1}{\gamma}. \tag{B.5}$$

Remark B.3.9. *One can improve the restriction on γ slightly by writing the potential as a perturbation of a quadratic as in [141]. Due to the restrictions on the stepsize h and the friction parameter γ in Proposition B.3.6, $c(h) = \mathcal{O}\left(\frac{m}{M}\right)$ for all allowed parameter choices. In general, for ∇Lipschitz, strongly-convex potentials, it may be impossible to prove contraction using such a quadratic form argument and synchronous coupling for $\gamma \leq \mathcal{O}(\sqrt{M})$ as explained in [116]. In the continuous time dynamics, $\gamma = \mathcal{O}(\sqrt{m})$ seems to yield the fastest convergence rate, as explained in [38]. In Example B.3.11 in the Appendix, we show that for Gaussian targets, UBU has an accelerated convergence rate $c(h) = \mathcal{O}(\sqrt{\frac{m}{M}})$ with the choice $\gamma = \mathcal{O}(\sqrt{m})$ and $h = \mathcal{O}(1/\sqrt{M})$.*

Proof of Proposition B.3.6. We follow the approach of [115, Corollary 20]. It is sufficient to prove contraction of a synchronous coupling of Markov chains in an appropriate norm, we will use the $\|\cdot\|_{a,b}$ norm of Definition B.3.1 with $a = \frac{1}{M}$, $b = \frac{1}{\gamma}$. Based on the assumptions, we have $b^2 < a/4$. Hence, (B.1) holds.

We aim to show that contraction occurs in this norm for two Markov chains simulated by the same discretization $z_n = (x_n, v_n) \in \mathbb{R}^{2d}$ and $\tilde{z}_n = (\tilde{x}_n, \tilde{v}_n) \in \mathbb{R}^{2d}$ that are synchronously coupled (i.e. share the same Gaussian random variables $\xi^{(1)}, \ldots, \xi^{(4)}$ in (3.2.9)), that is,

$$||\tilde{z}_{k+1} - z_{k+1}||_{a,b}^2 < \left(1 - c\left(h\right)\right)^2 ||\tilde{z}_k - z_k||_{a,b}^2. \tag{B.6}$$

Let $c_2(h) = 1 - (1 - c(h))^2$, $z_j^{\triangle} = \tilde{z}_j - z_j$ for $j \in \mathbb{N}$, then (B.6) is equivalent to showing that

$$\left(z_k^{\triangle}\right)^T \left(\left(1 - c_2\left(h\right)\right) \mathcal{M} - \mathcal{P}^T \mathcal{M} \mathcal{P}\right) z_k^{\triangle} > 0, \quad \text{where} \quad \mathcal{M} = \begin{pmatrix} I_d & bI_d \\ bI_d & aI_d \end{pmatrix}, \tag{B.7}$$

and $z_{k+1}^{\triangle} = \mathcal{P} z_k^{\triangle}$ ($\mathcal{P}$ depends on z_k and $\tilde{z}_k$, but we omit this in the notation).

Proving contraction for a general scheme is equivalent to showing that the matrix $\mathcal{H} := (1 - c_2(h))\,\mathcal{M} - \mathcal{P}^T \mathcal{M} \mathcal{P} \succ 0$ is positive definite. The matrix $\mathcal{H}$ is symmetric and hence of the block form

$$\mathcal{H} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}, \tag{B.8}$$

where A, B, C are $d \times d$ matrices, then

$$\mathcal{H} \succ 0 \quad \Leftrightarrow \quad A \succ 0 \quad \text{and} \quad C - BA^{-1}B \succ 0, \tag{B.9}$$

as shown in Theorem 7.7.7 of [83]. Further it is straightforward to show that if A, B and C commute then

$$\mathcal{H} \succ 0 \quad \Leftrightarrow \quad A \succ 0 \quad \text{and} \quad AC - B^2 \succ 0. \tag{B.10}$$

Considering two synchronously coupled trajectories of the UBU scheme, such that they have common noise and consider the difference process $x^{\triangle} := (\tilde{x}_j - x_j)$, $v^{\triangle} = (\tilde{v}_j - v_j)$ and $z^{\triangle} = \left(x^{\triangle}, v^{\triangle}\right)$, where $z_j^{\triangle} = \left(x_j^{\triangle}, v_j^{\triangle}\right)$ for $j = k, k+1$ for $k \in \mathbb{N}$. Let $\eta = \exp\{-\gamma h/2\}$, and

$$Q = \int_0^1 \nabla^2 U\left(\tilde{x}_k + t(x_k - \tilde{x}_k\right) dt.$$

By convexity, we have $mI_d \preceq Q \preceq MI_d$. Using the definition of the UBU scheme in (3.2.9), we can show that $z_{k+1}^{\triangle} = \mathcal{P} z_k^{\triangle}$ and $\mathcal{H} := (1 - c_2(h))\,\mathcal{M} - \mathcal{P}^T \mathcal{M} \mathcal{P} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ has elements of the form

$$\begin{aligned}
A &= -c_2(h) I_d + Q\left(2bh\eta + \frac{2h\left(1-\eta\right)}{\gamma}\right) + Q^2\left(-ah^2\eta^2 - \frac{h^2\left(1-\eta\right)^2}{\gamma^2} - \frac{2bh^2\eta\left(1-\eta\right)}{\gamma}\right) \\
B &= \left(\left(1-\eta^2\right)\left(b - \frac{1}{\gamma}\right) - bc_2(h)\right) I_d + Q^2\left(-\frac{ah^2\eta^2\left(1-\eta\right)}{\gamma} - \frac{2bh^2\eta\left(1-\eta\right)^2}{\gamma^2} - \frac{h^2\left(1-\eta\right)^3}{\gamma^3}\right) \\
&\quad + Q\left(ah\eta^3 + \frac{h\left(\eta+1\right)\left(1-\eta\right)^2}{\gamma^2} + \frac{h\left(1-\eta\right)^2}{\gamma^2} + \frac{bh\eta^2\left(1-\eta\right)}{\gamma} + \frac{bh\eta\left(1-\eta\right)}{\gamma} + \frac{bh\eta\left(1-\eta^2\right)}{\gamma}\right) \\
C &= \left(a(1-\eta^4) - \frac{2b\eta^2\left(1-\eta^2\right)}{\gamma} - \frac{\left(1-\eta^2\right)^2}{\gamma^2} - ac_2\left(h\right)\right) I_d \\
&\quad + Q^2\left(-\frac{ah^2\eta^2\left(1-\eta\right)^2}{\gamma^2} - \frac{2bh^2\eta\left(1-\eta\right)^3}{\gamma^3} - \frac{h^2\left(1-\eta\right)^4}{\gamma^4}\right) \\
&\quad + Q\left(\frac{2ah\eta^3\left(1-\eta\right)}{\gamma} + \frac{2bh\eta^2\left(1-\eta\right)^2}{\gamma^2} + \frac{2bh\eta\left(\eta+1\right)\left(1-\eta\right)^2}{\gamma^2} + \frac{2h\left(\eta+1\right)\left(1-\eta\right)^3}{\gamma^3}\right).
\end{aligned}$$

We will now check that $\mathcal{H} \succ 0$ using (B.10). By firstly considering A we wish to show that all its eigenvalues are positive which can be precisely stated as

$$\begin{aligned} P_A(\lambda) &\geq -c_2(h) + \frac{2h\lambda}{\gamma} + \left(-\frac{1}{M} - \frac{2h}{\gamma}\right) h^2\lambda^2 \\ &\geq \frac{7h\lambda}{4\gamma} + \left(-\frac{1}{M} - \frac{1}{\gamma^2}\right) h^2\lambda^2 > 0, \end{aligned}$$

where λ is an eigenvalue of Q ($m \leq \lambda \leq M$), $P_A(\lambda)$ denotes the eigenvalue of A according to the same eigenvector (Q, A, B, C are all symmetric and have the same eigenvectors here). We used our assumptions that $\gamma^2 \geq M$, $1 - \eta \leq h\gamma/2$, and $h < \frac{1}{2\gamma}$. Hence, we have $A \succ 0$.

Now it remains to prove that $AC - B^2 \succ 0$, now we have that $AC - B^2$ is a polynomial of Q, which we denote $P_{AC-B^2}(Q)$ and hence has eigenvalues dictated by the eigenvalues λ of Q. Because the terms are more complicated than the previous discretizations, we choose a convenient way of expanding the expression, which can obtain positive definiteness. That is to expand the expression in terms of a. Therefore one can show that $P_{AC-B^2}(\lambda) = c_0 + c_1 a + c_2 a^2$, where

$$\begin{aligned} c_1 + c_2 a &= \frac{h^2 c_2(h)\lambda^2\eta^4}{\gamma^2} - \frac{2h^2 c_2(h)\lambda^2\eta^2}{\gamma^2} - \frac{h^2\lambda^2\eta^4}{\gamma^2} + \frac{2h^2\lambda^2\eta^2}{\gamma^2} + \frac{2hc_2(h)\lambda\eta^4}{\gamma} - \frac{2h\lambda\eta^4}{\gamma} \\ &+ c_2(h)\eta^4 + \frac{h^2 c_2(h)\lambda^2}{\gamma^2} - \frac{h^2\lambda^2}{\gamma^2} - \frac{2hc_2(h)\lambda}{\gamma} + \frac{2h\lambda}{\gamma} + c_2(h)^2 - c_2(h) \\ &+ a\left(-\eta^2 h^2\lambda^2 + \eta^2 h^2 c_2(h)\lambda^2\right) \\ &\geq \frac{h\lambda}{\gamma}(1 - c_2(h))\left(\frac{7}{4}(1-\eta^4) - \frac{4h\lambda}{\gamma} - h\gamma\right). \end{aligned}$$

Furthermore, we have that

$$\begin{aligned} c_0 &= \frac{h^2(1-c_2(h))\lambda^2\eta^4}{\gamma^4} - \frac{2h^2(1-c_2(h))\lambda^2\eta^2}{\gamma^4} + \frac{2h(1-c_2(h))\lambda\eta^4}{\gamma^3} + \frac{c_2(h)(1-\eta^4)}{\gamma^2} \\ &- \frac{c_2(h)^2}{\gamma^2} + \frac{h^2\lambda^2(1-c_2(h))}{\gamma^4} - \frac{2h\lambda(1-c_2(h))}{\gamma^3} \\ &> \frac{h\lambda}{\gamma^3}(1-c_2(h))\left(\frac{h\lambda}{\gamma^3}\left(1-\eta^2\right)^2 - 2\left(1-\eta^4\right)\right), \end{aligned}$$

where now we combine this with the previous estimate

$$P_{AC-B^2}(\lambda) > \frac{h(1-c_2(h))}{\gamma}\left(\frac{7}{4}(1-\eta^4) - \frac{4h\lambda}{\gamma} - h\gamma - \frac{2\lambda\left(1-\eta^4\right)}{\gamma^2}\right) > 0,$$

which is true when $\gamma \geq \sqrt{8M}$ and we have used the fact that $1 - \eta^4 \geq h\gamma$. Hence $AC - B^2 \succ 0$ and our contraction results hold. All computations can be checked using `Mathematica`. The first claim follows by Lemma B.3.4 using (B.6). The existence of a unique invariant distribution $\pi_h \in \mathcal{P}_p(\mathbb{R}^{2d})$ follows by the same argument as in [115, Corollary 20]. □

Proof of Corollary B.3.8. By the triangle inequality, we have that for $a = \frac{1}{M}$, $b = \frac{1}{\gamma}$, any $n \in \mathbb{N}$ such that $n > t \cdot 2\gamma$,

$$\begin{aligned}&\mathcal{W}_{p,a,b}\left(\nu P_t^{\mathrm{cont}}, \mu P_t^{\mathrm{cont}}\right)\\&\leq \mathcal{W}_{p,a,b}\left(\nu P_{t/n}^n, \mu P_{t/n}^n\right) + \mathcal{W}_{p,a,b}\left(\nu P_t^{\mathrm{cont}}, \nu P_{t/n}^n\right) + \mathcal{W}_{p,a,b}\left(\mu P_t^{\mathrm{cont}}, \mu P_{t/n}^n\right).\end{aligned}$$

The first term can be bounded using Proposition B.3.6, and the upper bound can be shown to converge to $\exp\left(-\frac{mt}{8\gamma}\right)$ as $n \to \infty$. The second and third terms can be shown to converge to 0 as $n \to \infty$ using the strong convergence of the UBU discretization towards the diffusion (strong order 1 under these assumptions), which was established in Section 7.7 of [140], and the claim of the corollary now follows. □

Proposition B.3.10. *Consider the UBU scheme using stochastic gradients, where the underlying potential U is m-strongly convex and M-∇Lipschitz. Assume a stochastic gradient approximation defined by $(\mathcal{G}, \rho)$ (see Definition 3.2.1) satisfying Assumption 3.2.2 with constant C_G. We use P_h^n to denote the marginal transition kernel of the numerical schemes. We have for any two synchronously coupled chains, (x_k, v_k) and $(\tilde{x}_k, \tilde{v}_k)$ under the same assumptions as imposed in Proposition B.3.6 we have for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^{2d})$, and all $n \in \mathbb{N}$,*

$$\mathcal{W}_2^2\left(\nu P_h^n, \mu P_h^n\right) \leq 3\max\left\{M, \frac{1}{M}\right\}\left(1 - \frac{mh}{4\gamma} + \frac{5h^2 C_G}{M}\right)^n \mathcal{W}_2^2\left(\nu, \mu\right).$$

Proof. Using the technique of [103]. For stochastic gradients, we synchronously couple Brownian noise as well as the stochastic gradients. We wish to instead consider expected contraction of the update rule we used to prove contraction in the full gradient setting, i.e. for synchronously coupled (in stochastic gradient and Brownian increment) iterates $(x_l, v_l), (\tilde{x}_l, \tilde{v}_l) \in \mathbb{R}^{2d}$ for $l \in \mathbb{N}$ and $(x_l^{\triangle}, v_l^{\triangle}) = (\tilde{x}_l, \tilde{v}_l) - (x_l, v_l)$ and for $k \in \mathbb{N}$,

$$\mathbb{E}\|(x_{k+1}^{\triangle}, v_{k+1}^{\triangle})\|_{a,b}^2 \leq (1 - c(h))\,\|(x_k^{\triangle}, v_k^{\triangle})\|_{a,b}^2,$$

then we have

$$\mathbb{E}\left(\left(z_k^{\triangle}\right)^T P^T M P z_k^{\triangle}\right) \leq (1 - c(h))\left(z_k^{\triangle}\right)^T M z_k^{\triangle}.$$

Now if $\tilde{Q}$ is defined through the mean value theorem of $D_x\mathcal{G}$ (the Jacobian of $\mathcal{G}$) and is a random variable in W, such that $\mathbb{E}(\tilde{Q}) = Q$. Then $P^T M P$ is of the form

$$\mathcal{P}(\tilde{Q}) = \begin{pmatrix} P_1(\tilde{Q}) & P_2(\tilde{Q}) \\ P_2(\tilde{Q}) & P_3(\tilde{Q}) \end{pmatrix},$$

where P_1, P_2 and P_3 are quadratics in $\tilde{Q}$ of the form

$$\begin{aligned}P_1(\tilde{Q}) &= a_0 + a_1\tilde{Q} + a_2\tilde{Q}^2,\\P_2(\tilde{Q}) &= b_0 + b_1\tilde{Q} + b_2\tilde{Q}^2,\\P_3(\tilde{Q}) &= c_0 + c_1\tilde{Q} + c_2\tilde{Q}^2.\end{aligned}$$

Then we have

$$\mathbb{E}(P^T M P) = \mathcal{P}(Q) + \begin{pmatrix} a_2\mathbb{E}(\tilde{Q}-Q)^2 & b_2\mathbb{E}(\tilde{Q}-Q)^2 \\ b_2\mathbb{E}(\tilde{Q}-Q)^2 & c_2\mathbb{E}(\tilde{Q}-Q)^2 \end{pmatrix},$$

in combination with the Proposition B.3.6 result we have that

$$\begin{aligned}\mathbb{E}\|(x_{k+1}^{\triangle}, v_{k+1}^{\triangle})\|_{a,b}^2 &\leq (1-c(h))\,\|(x_k^{\triangle}, v_k^{\triangle})\|_{a,b}^2 + \left(z_k^{\triangle}\right)^T \begin{pmatrix} a_2\mathbb{E}(\tilde{Q}-Q)^2 & b_2\mathbb{E}(\tilde{Q}-Q)^2 \\ b_2\mathbb{E}(\tilde{Q}-Q)^2 & c_2\mathbb{E}(\tilde{Q}-Q)^2 \end{pmatrix} z_k^{\triangle} \\ &= (1-c(h))\,\|(x_k^{\triangle}, v_k^{\triangle})\|_{a,b}^2 + \left(z_k^{\triangle}\right)^T \mathcal{R}(\tilde{Q}) z_k^{\triangle},\end{aligned}$$

where we use the notation

$$\mathcal{R}(\tilde{Q}) := \begin{pmatrix} a_2\mathbb{E}(\tilde{Q}-Q)^2 & b_2\mathbb{E}(\tilde{Q}-Q)^2 \\ b_2\mathbb{E}(\tilde{Q}-Q)^2 & c_2\mathbb{E}(\tilde{Q}-Q)^2 \end{pmatrix}.$$

Then we will bound the remainder term $z^T\mathcal{R}(\tilde{Q})z$ for the UBU scheme. We have that

$$\begin{aligned}z^T\mathcal{R}(\tilde{Q})z &= \left(ah^2\eta + \frac{h^2(1-\eta)^2}{\gamma^2} + \frac{2bh^2\eta(1-\eta)}{\gamma}\right)\left(x + \frac{1-\eta}{\gamma}v\right)^T \\ &\quad \times \mathsf{Var}(\tilde{Q})\left(x + \frac{1-\eta}{\gamma}v\right) \\ &\leq 5ah^2 C_G \|(x,v)\|_{a,b}^2,\end{aligned}$$

for $\gamma^2 \geq 8M$ and $h < \frac{1}{2\gamma}$ and where we define $\mathsf{Var}(\tilde{Q}) := \mathbb{E}(\tilde{Q}-Q)^2$. The claim follows using our choice $a = \frac{1}{M}$. □

Example B.3.11. *Considering the anisotropic Gaussian distribution on $\mathbb{R}^2$ with a m-strongly convex and M-∇ Lipschitz potential $U : \mathbb{R}^2 \mapsto \mathbb{R}$ given by*

$$U(x,y) = \frac{1}{2}mx^2 + \frac{1}{2}My^2.$$

For the BU scheme the transition matrix for the difference chain of synchronously coupled chains is given by the matrix

$$P = \begin{pmatrix} I - h\left(\frac{1-\eta^2}{\gamma}\right)Q & \frac{1-\eta^2}{\gamma}I \\ -h\eta^2 Q & \eta^2 I \end{pmatrix}, \text{ where } Q = \begin{pmatrix} m & 0 \\ 0 & M \end{pmatrix},$$

with eigenvalues

$$\frac{1+\eta^2 - h\frac{1-\eta^2}{\gamma}\lambda \pm \sqrt{-4\eta^2 + \left(1+\eta^2 - h\frac{1-\eta^2}{\gamma}\lambda\right)^2}}{2},$$

for $\lambda = m, M$. For stability and contraction, we require that

$$\lambda_{\max} := \max_{\lambda\in\{m,M\}} \left| \frac{1+\eta^2 - h\frac{1-\eta^2}{\gamma}\lambda \pm \sqrt{-4\eta^2 + \left(1+\eta^2 - h\frac{1-\eta^2}{\gamma}\lambda\right)^2}}{2} \right| < 1. \tag{B.11}$$

From this, we can compute the stepsize restrictions and the best convergence rate as, by Gelfand's formula, the asymptotic contraction rate exactly equals $1-\lambda_{\max}$. *Due to the convexity of the absolute value function it is necessary that* $\frac{1}{2}|1+\eta^2-h\frac{1-\eta^2}{\gamma}M|<1$, *therefore* $h<\sqrt{\frac{8}{M}}$, *when* $h<\frac{1}{2\gamma}$. *In the moderate to high friction regime, the contraction rate can be written as*

$$c=\frac{1-\eta^2+h\frac{1-\eta^2}{\gamma}m-\sqrt{-4h\left(\frac{1-\eta^2}{\gamma}\right)m+\left(1-\eta^2+h\frac{1-\eta^2}{\gamma}m\right)^2}}{2}$$

which can be shown to be $\mathcal{O}(mh/\gamma)$ *for* $\gamma\geq\mathcal{O}(\sqrt{M})$ *and* $h<\mathcal{O}(\frac{1}{\gamma})$ *for appropriate constants. In the low friction regime, we set* γ *such that* $-4\eta^2+\left(1+\eta^2-h\frac{1-\eta^2}{\gamma}m\right)^2=0$, *noting that the solution to this yields* γ *to be* $\mathcal{O}(\sqrt{m})$. *In this case, the eigenvalues of* P *are*

$$\eta,\qquad\frac{1}{2}\left(1+\eta^2-h\frac{1-\eta^2}{\gamma}M\pm\sqrt{-4\eta^2+\left(1+\eta^2-h\frac{1-\eta^2}{\gamma}M\right)^2}\right),$$

with modulus η *when* $\left(1+\eta^2-h\frac{1-\eta^2}{\gamma}M\right)^2<4\eta^2$. *This restriction implies that* h *is* $\mathcal{O}(1/\sqrt{M})$. *The contraction rate is therefore given by*

$$c=1-\eta\geq\frac{h\gamma}{4}=\mathcal{O}\left(\sqrt{\frac{m}{M}}\right),$$

where h *is* $\mathcal{O}(1/\sqrt{M})$. *We have the corresponding contraction rate results for UBU as well due to the fact that* $(UBU)^n=U(BU)^{n-1}U$ *and* U *is Lipschitz.*

A key ingredient to establishing some variance bounds for the inexact gradient methods is to establish non-asymptotic bounds on the fourth moment of the distance to the minimizer. To do this we use a Lyapunov function similar to the one used for kinetic Langevin dynamics in [70] and inspired by [109]. Related Lyapunov functions have also been used in [63] for discretized kinetic Langevin dynamics and [92] for optimizers based on Langevin dynamic methods. These bounds provide novel drift conditions in L^4 for UBU scheme and can be extended to the case of stochastic gradients.

The following lemma will be useful for the argument.

Lemma B.3.12 (Convexity bound)**.** *For all* $x\in\mathbb{R}^d$ *and for a* $m-$*strongly convex,* M*-*∇*Lipschitz potential* $U:\mathbb{R}^d\to\mathbb{R}$ *with minimizer* $x^*\in\mathbb{R}^d$ *such that* $\nabla U(x^*)=0$, *we have*

$$(x-x^*)\cdot\left(\nabla U(x)-\nabla U(x^*)\right)/2\geq\lambda\left(U(x)-U(x^*)+\gamma^2\|x-x^*\|^2/4\right)$$

for

$$\lambda=\min\left(\frac{1}{4},\frac{m}{\gamma^2}\right).\tag{B.12}$$

Proof. By convexity, it follows that $(x-x^*)\cdot\left(\nabla U(x)-\nabla U(x^*)\right)/4\geq\left(U(x)-U(x^*)\right)/4$, and by m-strong convexity, we have $(x-x^*)\cdot\left(\nabla U(x)-\nabla U(x^*)\right)/4\geq m\|x-x^*\|^2/4$. We obtain the result by adding up these two inequalities. □

Proposition B.3.13. *Consider the UBU scheme with the underlying potential $U : \mathbb{R}^d \to \mathbb{R}$ is M-∇Lipschitz and m-strongly convex. Denote $x^* \in \mathbb{R}^d$ to be the minimizer of U such that $\nabla U(x^*) = 0$ and $(x_k, v_k, \overline{x}_k)_{k\in\mathbb{N}}$ to be defined by* (B.41)-(B.43) *the iterates of the full gradient UBU scheme and the points of gradient evaluation within each iteration. Further assume that $h < \min\left(1, \frac{1}{2\gamma}, \frac{\lambda}{8\gamma(4+\lambda)}\right)$ and $\gamma^2 \geq M$, then we have*

$$\mathbb{E}\left[\|\overline{x}_k - x^*\|^4 \mid x_0, v_0\right] \leq \frac{4}{m^2}\Bigg[4\left(1 - \frac{c_4(h)}{2}\right)^k \left(\gamma^4\|x_0 - x^*\|^4 + \|v_0\|^4 + 122\gamma^2h^2d^2\right) + 2\frac{\frac{(6h\gamma d + 160h\gamma(1+\lambda^2))^2}{4c_4(h)} + 24h^2\gamma^2d^2}{c_4(h)}\Bigg],$$

where

$$c_4(h) := h\lambda\gamma - 8h^2\gamma^2(4+\lambda) \tag{B.13}$$

Proof. Using the fact that $(\mathcal{UBU})^n = \mathcal{U}(\mathcal{BU})^{n-1}\mathcal{BU}$ we can consider convergence of $\mathcal{BU}$, We have that the $\mathcal{BU}$ function can be written as the update rule

$$\overline{x}_{k+1} = \overline{x}_k + \frac{1-\eta^2}{\gamma}\left(\overline{v}_k - h\nabla U(\overline{x}_k)\right) + \sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(h, \xi_{k+1}^{(1)}\right) - \mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right)\right), \tag{B.14}$$

$$\overline{v}_{k+1} = \eta^2\left(\overline{v}_k - h\nabla U(\overline{x}_k)\right) + \sqrt{2\gamma}\mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right), \tag{B.15}$$

where we used the notation $(\overline{x}_k)_{k\in\mathbb{N}}$ because this is the point of the gradient evaluation at each step of UBU and is the same as the $(\overline{x}_k)_{k\in\mathbb{N}}$ in (B.42). As a reminder,

$$\mathcal{Z}^{(1)}\left(h, \xi_{k+1}^{(1)}\right) = \sqrt{h}\xi_{k+1}^{(1)}$$

$$\mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right) = \sqrt{\frac{1-\eta^4}{2\gamma}}\left(\sqrt{\frac{1-\eta^2}{1+\eta^2}\cdot\frac{2}{\gamma h}}\xi_{k+1}^{(1)} + \sqrt{1 - \frac{1-\eta^2}{1+\eta^2}\cdot\frac{2}{\gamma h}}\xi_{k+1}^{(2)}\right).$$

We choose our Lyapunov function $\mathcal{V} : \mathbb{R}^{2d} \to \mathbb{R}$, defined for $(x, v) \in \mathbb{R}^{2d}$ by

$$\mathcal{V}(x, v) := U(x) - U(x^*) + \frac{1}{4}\gamma^2\left(\|x - x^* + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x - x^*\|^2\right). \tag{B.16}$$

It is easy to check that for all $(x, v) \in \mathbb{R}^{2d}$, $\|x - x^* + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 \geq \frac{1}{2}\|x - x^*\|^2$ and hence using (B.12),

$$\mathcal{V}(x, v) \geq \left(\frac{m}{2} + \frac{1}{16}\gamma^2\right)\|x - x^*\|^2. \tag{B.17}$$

In order to have control over fourth moments $\mathbb{E}[\|\overline{x}_k - x^*\|^4]$, we start with

$$\mathbb{E}\left[\mathcal{V}(\overline{x}_{k+1}, \overline{v}_{k+1})^2 \mid \overline{x}_k, \overline{v}_k\right] = \mathbb{E}\Bigg[\Bigg(U(\overline{x}_{k+1}) - U(x^*) + \frac{1}{4}\gamma^2\Big(\|\overline{x}_{k+1} - x^* + \gamma^{-1}\overline{v}_{k+1}\|^2 + \|\gamma^{-1}\overline{v}_{k+1}\|^2 - \lambda\|\overline{x}_{k+1} - x^*\|^2\Big)\Bigg)^2 \mid \overline{x}_k, \overline{v}_k\Bigg],$$

and using [121][Lemma 1.2.3] we have

$$U(\overline{x}_{k+1}) - U(x^*) \leq U(\overline{x}_k) - U(x^*) + [\nabla U(\overline{x}_k) \cdot (\overline{x}_{k+1} - \overline{x}_k)] + \frac{M}{2} \|\overline{x}_{k+1} - \overline{x}_k\|^2$$

and

$$\mathbb{E}\left[\mathcal{V}(\overline{x}_{k+1}, \overline{v}_{k+1})^2 \mid \overline{x}_k, \overline{v}_k\right] \leq \mathbb{E}\Bigg[\Bigg(U(\overline{x}_k) - U(x^*) + [\nabla U(\overline{x}_k) \cdot (\overline{x}_{k+1} - \overline{x}_k)] + \frac{M}{2} \|\overline{x}_{k+1} - \overline{x}_k\|^2$$
$$+ \frac{1}{4}\gamma^2 \left(\|\overline{x}_{k+1} - x^* + \gamma^{-1}\overline{v}_{k+1}\|^2 + \|\gamma^{-1}\overline{v}_{k+1}\|^2 - \lambda\|\overline{x}_{k+1} - x^*\|^2\right)\Bigg)^2 \mid \overline{x}_k, \overline{v}_k\Bigg].$$

Now, we can decompose the right-hand side in the form

$$\mathbb{E}\left(\left(r(\overline{x}_k, \overline{v}_k) + \mathbf{s}(\overline{x}_k, \overline{v}_k) \cdot (\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}) + (\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})^T \mathcal{T} (\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\right)^2 \mid \overline{x}_k, \overline{v}_k\right),$$

for $r : \mathbb{R}^{2d} \to \mathbb{R}$, $\mathbf{s} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ and $\mathcal{T} \in \mathbb{R}^{2d \times 2d}$. We then have

$$\mathbb{E}\left[\mathcal{V}(\overline{x}_{k+1}, \overline{v}_{k+1})^2 \mid \overline{x}_k, \overline{v}_k\right] \leq r^2(\overline{x}_k, \overline{v}_k) + \mathbb{E}_{\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}}\Big(\left(\mathbf{s}(\overline{x}_k, \overline{v}_k) \cdot (\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\right)^2$$
$$+ \left((\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})^T \mathcal{T} (\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\right)^2 + 2r(\overline{x}_k, \overline{v}_k)(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})^T \mathcal{T} (\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\Big),$$

using the fact that $\xi_{k+1}^{(1)}$ and $\xi_{k+1}^{(2)}$ are independently distributed and have zero first and third moments. The terms $r, \mathbf{s}$ and $\mathcal{T}$ are given by

$$r(\overline{x}_k, \overline{v}_k) = \mathcal{V}(\overline{x}_k, \overline{v}_k) - \frac{h\gamma}{2}\nabla U(\overline{x}_k) \cdot (\overline{x}_k - x^* + \gamma^{-1}\overline{v}_k) + \frac{1-\eta^2}{\gamma}\overline{v}_k \cdot \nabla U(\overline{x}_k)$$
$$- \frac{\lambda(1-\eta^2)\gamma}{2}\langle \overline{x}_k - x^*, \overline{v}_k\rangle - \frac{1-\eta^4}{4}\|\overline{v}_k\|^2 - \frac{h\eta^4}{2}\overline{v}_k \cdot \nabla U(\overline{x}_k) - h\frac{1-\eta^2}{\gamma}\nabla U(\overline{x}_k) \cdot \nabla U(\overline{x}_k)$$
$$+ h^2\frac{(1+\eta^4)}{4}\|\nabla U(\overline{x}_k)\|^2 + \left(\frac{M}{2} - \frac{\gamma^2\lambda}{4}\right)\left(\frac{1-\eta^2}{\gamma}\right)^2 \|\overline{v}_k - h\nabla U(\overline{x}_k)\|^2$$
$$+ h\frac{\lambda(1-\eta^2)\gamma}{2}\langle \overline{x}_k - x^*, \nabla U(\overline{x}_k)\rangle,$$
$$\mathbf{s}(\overline{x}_k, \overline{v}_k) \cdot \left(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right)$$
$$= ((\sqrt{h} - a_1)\xi_{k+1}^{(1)} - a_2\xi_{k+1}^{(2)}) \cdot \left(\frac{M\sqrt{2\gamma}}{2\gamma}\frac{1-\eta^2}{\gamma}(\overline{v}_k - h\nabla U(\overline{x}_k)) + \sqrt{\frac{2}{\gamma}}\nabla U(\overline{x}_k)\right)$$
$$+ \frac{\sqrt{2\gamma h}}{4}\gamma\left(\overline{x}_k - x^* + \gamma^{-1}\overline{v}_k - \frac{h}{\gamma}\nabla U(\overline{x}_k)\right) \cdot \xi_{k+1}^{(1)} + \frac{\eta^2\sqrt{2\gamma}}{4}(\overline{v}_k - h\nabla U(\overline{x}_k)) \cdot \left(a_1\xi_{k+1}^{(1)} + a_2\xi_{k+1}^{(2)}\right)$$
$$- \frac{\lambda\gamma\sqrt{2\gamma}}{4}\left(\overline{x}_k - x^* + \frac{1-\eta^2}{\gamma}(\overline{v}_k - h\nabla U(\overline{x}_k)) \cdot \left((\sqrt{h} - a_1)\xi_{k+1}^{(1)} - a_2\xi_{k+1}^{(2)}\right)\right),$$
$$(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})^T \mathcal{T} (\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}) = \left(\frac{M}{\gamma} - \frac{\lambda\gamma}{2}\right)\left\|\left(\sqrt{h} - a_1\right)\xi_{k+1}^{(1)} - a_2\xi_{k+1}^{(2)}\right\|^2$$
$$+ \frac{h\gamma}{2}\left\|\xi_{k+1}^{(1)}\right\|^2 + \frac{\gamma}{2}\left\|a_1\xi_{k+1}^{(1)} + a_2\xi_{k+1}^{(2)}\right\|^2,$$

where we have defined $\mathcal{Z}^{(2)}(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}) := a_1\xi_{k+1}^{(1)} + a_2\xi_{k+1}^{(2)}$ and $\mathcal{Z}^{(1)}(h, \xi_{k+1}^{(1)}) := \sqrt{h}\xi_{k+1}^{(1)}$ and $\mathcal{Z}^{(1)}(h, \xi_{k+1}^{(1)}) - \mathcal{Z}^{(2)}(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}) = (\sqrt{h} - a_1)\xi_{k+1}^{(1)} - a_2\xi_{k+1}^{(2)}$ with $|\sqrt{h} - a_1| \leq 2\sqrt{h}$, $|a_2| \leq \sqrt{h}$ and $|a_1| \leq \sqrt{h}$.

We start by bounding the deterministic component r:

$$\begin{aligned} r(\overline{x}_k, \overline{v}_k) = \mathcal{V}(\overline{x}_k, \overline{v}_k) - \frac{h\gamma}{2}\nabla U(\overline{x}_k) \cdot (\overline{x}_k - x^* + \gamma^{-1}\overline{v}_k) + \frac{1-\eta^2}{\gamma}\overline{v}_k \cdot \nabla U(\overline{x}_k) \\ - \frac{\lambda(1-\eta^2)\gamma}{2}\langle \overline{x}_k - x^*, \overline{v}_k\rangle - \frac{1-\eta^4}{4}\|\overline{v}_k\|^2 - \frac{h\eta^4}{2}\overline{v}_k \cdot \nabla U(\overline{x}_k) + \mathcal{O}(h^2) \end{aligned}$$

where the higher-order terms are given by

$$\begin{aligned} &- h\frac{1-\eta^2}{\gamma}\nabla U(\overline{x}_k) \cdot \nabla U(\overline{x}_k) + \left(\frac{M}{2} - \frac{\gamma^2\lambda}{4}\right)\left(\frac{1-\eta^2}{\gamma}\right)^2 \|\overline{v}_k - h\nabla U(\overline{x}_k)\|^2 \\ &+ h\frac{\lambda(1-\eta^2)\gamma}{2}\langle \overline{x}_k - x^*, \nabla U(\overline{x}_k)\rangle + h^2\frac{(1+\eta^4)}{4}\|\nabla U(\overline{x}_k)\|^2. \end{aligned}$$

Using Lemma B.3.12 we have

$$\begin{aligned} r(\overline{x}_k, \overline{v}_k) &\leq \mathcal{V}(\overline{x}_k, \overline{v}_k) \\ &\quad - h\gamma\lambda\left(U(\overline{x}_k) - U(x^*) + \frac{\gamma^2}{4}\|\overline{x}_k - x^*\|^2 + \frac{1-\eta^2}{2h}\langle \overline{x}_k - x^*, \overline{v}_k\rangle + \frac{1-\eta^4}{4h\gamma\lambda}\|\overline{v}_k\|^2\right) + \mathcal{O}(h^2) \\ &\quad \leq (1 - h\lambda\gamma)\,\mathcal{V}(\overline{x}_k, \overline{v}_k) + h\gamma\lambda\left(\frac{1-\eta^4}{4h\gamma\lambda} - \frac{1}{2\lambda}\right)\|\overline{v}_k\|^2 \\ &\quad + h\gamma\lambda\left(\frac{1-\eta^2}{2h} - \frac{\gamma}{2}\right)\langle \overline{x}_k - x^*, \overline{v}_k\rangle + \mathcal{O}(h^2) \\ &\quad \leq (1 - h\lambda\gamma)\,\mathcal{V}(\overline{x}_k, \overline{v}_k) + h\gamma\lambda\left(\frac{1-\eta^2}{2h} - \frac{\gamma}{2}\right)\langle \overline{x}_k - x^*, \overline{v}_k\rangle + \mathcal{O}(h^2), \end{aligned}$$

where we have used

$$\left(-\frac{h}{2} + \frac{1-\eta^2}{\gamma} - \frac{h\eta^4}{2}\right)\overline{v}_k \cdot \nabla U(\overline{x}_k) \leq \frac{h^2\gamma}{2}\left|\overline{v}_k \cdot \nabla U(\overline{x}_k)\right|,$$

due to the fact that for all $0 < x < 1$, $0 \leq -x + 2(1 - e^{-x}) - xe^{-2x} \leq x^2$ and $0 < h\gamma < 1$. We group this term into higher-order terms and use the fact that $1 - \eta^2 \geq h\gamma - \frac{(h\gamma)^2}{2}$ to arrive at

$$h\gamma\lambda\left(\frac{1-\eta^2}{2h} - \frac{\gamma}{2}\right)\langle \overline{x}_k - x^*, \overline{v}_k\rangle \leq h\gamma\lambda\left(\frac{\gamma}{2} - \frac{1-\eta^2}{2h}\right)|\langle \overline{x}_k - x^*, \overline{v}_k\rangle| \leq \lambda\frac{h^2\gamma^3}{4}\left|\langle \overline{x}_k - x^*, \overline{v}_k\rangle\right|.$$

We again group this into the higher-order terms. Assuming $h < 1$, we find that the second-order terms are bounded by

$$\begin{aligned} &Mh^2\left(\|\overline{v}_k\|^2 + h^2M^2\|\overline{x}_k - x^*\|^2\right) + h^2\frac{\gamma^2\lambda}{2}M\|\overline{x}_k - x^*\|^2 \\ &+ \frac{h^2\gamma}{2}\left(\sqrt{M}\|\overline{v}_k\|^2 + M^{3/2}\|\overline{x}_k - x^*\|^2\right) + \lambda\frac{h^2\gamma^3}{4}\left(\gamma\|\overline{x}_k - x^*\|^2 + \frac{1}{\gamma}\|\overline{v}_k\|^2\right) + \frac{h^2M^2}{2}\|\overline{x}_k - x^*\|^2. \end{aligned}$$

Assuming that $\lambda \leq \frac{1}{4}$ we have, for all $x, v \in \mathbb{R}^d$,

$$8\mathcal{V}(x,v) \geq \|v\|^2 \qquad 16\mathcal{V}(x,v) \geq \gamma^2\|x-x^*\|^2$$

and using $h < \frac{1}{2\sqrt{M}}$, the $\mathcal{O}(h^2)$ terms are bounded by

$$\begin{aligned}&h^2\left(\gamma^2 + \frac{\gamma^2\lambda}{4} + \frac{\gamma^2}{2}\right)\|\overline{v}_k\|^2 + \gamma^2 h^2\left(\frac{M}{4} + \frac{M\lambda}{2} + \frac{M}{2} + \frac{\gamma^2\lambda}{4} + \frac{M^2}{2\gamma^2}\right)\|\overline{x}_k - x^*\|^2 \\ &\leq 8h^2\gamma^2\left(4+\lambda\right)\mathcal{V}(\overline{x}_k, \overline{v}_k).\end{aligned}$$

Therefore

$$r(\overline{x}_k, \overline{v}_k) \leq \left(1 - h\lambda\gamma + 8h^2\gamma^2\left(4+\lambda\right)\right)\mathcal{V}(\overline{x}_k, \overline{v}_k).$$

Now let us define $c_4(h) := h\lambda\gamma - 8h^2\gamma^2\left(4+\lambda\right)$, then we have that

$$r^2(\overline{x}_k, \overline{v}_k) \leq (1 - c_4(h))^2\mathcal{V}^2(\overline{x}_k, \overline{v}_k)$$

and

$$\begin{aligned}&2r(\overline{x}_k, \overline{v}_k)\mathbb{E}_{\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}}\left[(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})^T\mathcal{T}(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\right] \leq \\ &2(1 - c_4(h))\mathcal{V}(\overline{x}_k, \overline{v}_k)\mathbb{E}_{\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}}\left[(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})^T\mathcal{T}(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\right].\end{aligned}$$

From the fact that $\lambda\gamma/2 \leq M/\gamma$ (due to Lemma B.3.12) and $\gamma^2 \geq 8M$ we have the estimates

$$\begin{aligned}&\mathbb{E}_{\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}}\left[(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})^T\mathcal{T}(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\right] = \mathbb{E}_{\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}}\Bigg[\left(\frac{M}{\gamma} - \frac{\lambda\gamma}{2}\right)\left\|\left(\sqrt{h} - a_1\right)\xi_{k+1}^{(1)} - a_2\xi_{k+1}^{(2)}\right\|^2 \\ &+ \frac{h\gamma}{2}\left\|\xi_{k+1}^{(1)}\right\|^2 + \frac{\gamma}{2}\left\|a_1\xi_{k+1}^{(1)} + a_2\xi_{k+1}^{(2)}\right\|^2\Bigg] \leq 3h\gamma d \\ &\mathbb{E}_{\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}}\left[\left((\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})^T\mathcal{T}(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\right)^2\right] \\ &\leq \mathbb{E}_{\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}}\left[\left(2h\gamma\|\xi_{k+1}^{(1)}\|^2 + 2h\gamma\|\xi_{k+1}^{(2)}\|^2\right)^2\right] \leq 24h^2\gamma^2d^2.\end{aligned}$$

Therefore the remaining term we need to bound is $\mathbb{E}_{\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}}\left(\mathbf{s}(\overline{x}_k, \overline{v}_k)\cdot(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)})\right)^2 = \|\mathbf{s}(\overline{x}_k, \overline{v}_k)\|^2$, where

$$\begin{aligned}&\mathbf{s}(\overline{x}_k, \overline{v}_k)\cdot\left(\xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right) \\ &= ((\sqrt{h} - a_1)\xi_{k+1}^{(1)} - a_2\xi_{k+1}^{(2)})\cdot\left(\frac{M\sqrt{2\gamma}}{2\gamma}\frac{1-\eta^2}{\gamma}(\overline{v}_k - h\nabla U(\overline{x}_k)) + \sqrt{\frac{2}{\gamma}}\nabla U(\overline{x}_k)\right) \\ &+ \frac{\sqrt{2\gamma h}}{4}\gamma\left(\overline{x}_k - x^* + \gamma^{-1}\overline{v}_k - \frac{h}{\gamma}\nabla U(\overline{x}_k)\right)\cdot\xi_{k+1}^{(1)} + \frac{\eta^2\sqrt{2\gamma}}{4}(\overline{v}_k - h\nabla U(\overline{x}_k))\cdot\left(a_1\xi_{k+1}^{(1)} + a_2\xi_{k+1}^{(2)}\right) \\ &- \frac{\lambda\gamma\sqrt{2\gamma}}{4}\left(\overline{x}_k - x^* + \frac{1-\eta^2}{\gamma}(\overline{v}_k - h\nabla U(\overline{x}_k))\cdot\left((\sqrt{h} - a_1)\xi_{k+1}^{(1)} - a_2\xi_{k+1}^{(2)}\right)\right),\end{aligned}$$

using that $\gamma^2\|x-x^*\|^2 \leq 16\mathcal{V}(x,v)$ and $\|v\|^2 \leq 8\mathcal{V}(x,v)$ for all $x,v \in \mathbb{R}^d$ we have

$$\|\mathbf{s}(\overline{x}_k,\overline{v}_k)\|^2 = \|s_1(\overline{x}_k,\overline{v}_k)\|^2 + \|s_2(\overline{x}_k,\overline{v}_k)\|^2,$$

where

$$\begin{aligned}
\|s_1(\overline{x}_k,\overline{v}_k)\|^2 &\leq h\Bigg(\left(2\frac{M(1-\eta^2)}{\gamma\sqrt{2\gamma}} + \frac{\eta^2\sqrt{2\gamma}}{4} + 2\frac{\lambda\sqrt{2\gamma}(1-\eta^2)}{4}\right)\|\overline{v}_k\| \\
&+ \left(2\frac{hM^2(1-\eta^2)}{\gamma\sqrt{2\gamma}} + 2\sqrt{\frac{2}{\gamma}}M + \frac{\sqrt{2\gamma}hM}{4} + \frac{\eta^2\sqrt{2\gamma}hM}{4} + 2\frac{\lambda\sqrt{2\gamma}(1-\eta^2)hM}{4}\right)\|\overline{x}_k - x^*\| \\
&+ \frac{\sqrt{2}\gamma^{3/2}}{4}\|\overline{x}_k - x^* + \gamma^{-1}\overline{v}_k\|\Bigg)^2 \\
&\leq h\left((2\sqrt{\gamma}+\lambda\sqrt{\gamma})\sqrt{\mathcal{V}(\overline{x}_k,\overline{v}_k)} + \left(\frac{2\gamma^{3/2}}{\sqrt{8}} + \frac{\lambda\gamma^{3/2}}{32}\right)\frac{4}{\gamma}\sqrt{\mathcal{V}(\overline{x}_k,\overline{v}_k)} + \frac{5}{2}\sqrt{\gamma}\sqrt{\mathcal{V}(\overline{x}_k,\overline{v}_k)}\right)^2 \\
&\leq 110h\gamma\left(1+\lambda^2\right)\mathcal{V}(\overline{x}_k,\overline{v}_k)
\end{aligned}$$

for $\gamma^2 \geq \sqrt{8M}$ and $h < \frac{1}{2\gamma}$ and

$$\begin{aligned}
\|s_2(\overline{x}_k,\overline{v}_k)\|^2 &\leq h\Bigg(\left(2\frac{M(1-\eta^2)}{\gamma\sqrt{2\gamma}} + \frac{\eta^2\sqrt{2\gamma}}{4} + 2\frac{\lambda\sqrt{2\gamma}(1-\eta^2)}{4}\right)\|\overline{v}_k\| \\
&+ \left(2\frac{hM^2(1-\eta^2)}{\gamma\sqrt{2\gamma}} + 2\sqrt{\frac{2}{\gamma}}M + \frac{\eta^2\sqrt{2\gamma}hM}{4} + 2\frac{\lambda\sqrt{2\gamma}(1-\eta^2)hM}{4}\right)\|\overline{x}_k - x^*\|\Bigg)^2 \\
&\leq h\left((2\sqrt{\gamma}+\lambda\sqrt{\gamma})\sqrt{\mathcal{V}(\overline{x}_k,\overline{v}_k)} + \left(\frac{2\gamma^{3/2}}{\sqrt{8}} + \frac{\lambda\gamma^{3/2}}{32}\right)\frac{4}{\gamma}\sqrt{\mathcal{V}(\overline{x}_k,\overline{v}_k)}\right)^2 \\
&\leq 50h\gamma\left(1+\lambda^2\right)\mathcal{V}(\overline{x}_k,\overline{v}_k).
\end{aligned}$$

Therefore $\mathbb{E}_{\xi_{k+1}^{(1)},\xi_{k+1}^{(2)}}\left[\left(\mathbf{s}(\overline{x}_k,\overline{v}_k)\cdot(\xi_{k+1}^{(1)},\xi_{k+1}^{(2)})\right)^2\right] \leq 160h\gamma(1+\lambda^2)\mathcal{V}(\overline{x}_k,\overline{v}_k)$. Combining estimates, we have the drift inequality

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{V}(\overline{x}_{k+1},\overline{v}_{k+1})^2 \mid \overline{x}_k,\overline{v}_k\right] \leq (1-c_4(h))^2\mathcal{V}^2(\overline{x}_k,\overline{v}_k) + 6h\gamma d\left(1-c_4(h)\right)\mathcal{V}(\overline{x}_k,\overline{v}_k) \\
&+ 160h\gamma(1+\lambda^2)\mathcal{V}(\overline{x}_k,\overline{v}_k) + 24h^2\gamma^2d^2.
\end{aligned}$$

We will now use the quadratic property that states, for $b_1, b_2 > 0$,

$$b_2x^2 + \frac{b_1^2}{4b_2} \geq b_1x,$$

for all $x \in \mathbb{R}$ and therefore

$$c_4(h)\mathcal{V}^2(\overline{x}_k,\overline{v}_k) + \frac{(6h\gamma d + 160h\gamma(1+\lambda^2))^2}{4c_4(h)} \geq 6h\gamma d\left(1-c_4(h)\right)\mathcal{V}(\overline{x}_k,\overline{v}_k) + 160h\gamma(1+\lambda^2)\mathcal{V}(\overline{x}_k,\overline{v}_k)$$

and therefore for $c_4(h) < \frac{1}{2}$ (which is satisfied when $h < \frac{1}{2\gamma}$ and $\lambda < 1$, which is satisfied as $\lambda \leq M/2\gamma^2 \leq 1$ for $\gamma^2 \geq \frac{M}{2}$) we have

$$\mathbb{E}\left[\mathcal{V}(\overline{x}_{k+1}, \overline{v}_{k+1})^2 \mid \overline{x}_k, \overline{v}_k\right] \leq \left(1 - \frac{c_4(h)}{2}\right) \mathcal{V}^2(\overline{x}_k, \overline{v}_k) + \frac{(6h\gamma d + 160h\gamma(1+\lambda^2))^2}{4c_4(h)} + 24h^2\gamma^2 d^2, \tag{B.18}$$

then globally, we have

$$\begin{aligned}\frac{m^2}{4}\mathbb{E}\left[\|\overline{x}_k - x^*\|^4 \mid y_0, v_0\right] &\leq \mathbb{E}\left[\mathcal{V}(\overline{x}_k, v_k)^2 \mid y_0, v_0\right] \\ &\leq \left(1 - \frac{c_4(h)}{2}\right)^k \mathcal{V}^2(\overline{x}_0, \overline{v}_0) + 2\frac{\frac{(6h\gamma d+160h\gamma(1+\lambda^2))^2}{4c_4(h)} + 24h^2\gamma^2 d^2}{c_4(h)}.\end{aligned}$$

Now, we have proved this for the iterates of $\mathcal{BU}$, where we wish to use the relation $(\mathcal{UBU})^k = \mathcal{U}(\mathcal{BU})^{k-1}\mathcal{BU}$. In this case, we have that $\overline{x}_k$, the $(k+1)$-th point of approximate gradient/full gradient evaluation, is precisely the position after $\mathcal{U}(\mathcal{BU})^k$. It follows that

$$\frac{m^2}{4}\mathbb{E}\left[\|\overline{x}_k - x^*\|^4 \mid \overline{x}_0, \overline{v}_0\right] \leq \left(1 - \frac{c_4(h)}{2}\right)^k \mathcal{V}^2(\overline{x}_0, \overline{v}_0) + 2\frac{\frac{(6h\gamma d+160h\gamma(1+\lambda^2))^2}{4c_4(h)} + 24h^2\gamma^2 d^2}{c_4(h)},$$

where $(\overline{x}_0, \overline{v}_0) = \mathcal{U}(x_0, v_0, h/2, \xi_0^{(1)}, \xi_0^{(2)})$. It is easy to show that $\mathcal{V}(x, v) \leq \gamma^2\|x - x^*\|^2 + \|v\|^2$ for all $(x, v) \in \mathbb{R}^{2d}$ using [121][Lemma 1.2.3] and that $\gamma^2 \geq M$. Therefore

$$\begin{aligned}&\mathbb{E}\left[\mathcal{V}^2(\mathcal{U}(x_0, v_0, h/2, \xi_0^{(1)}, \xi_0^{(2)})) \mid x_0, v_0\right] \leq \mathbb{E}\left[2\gamma^4\|\overline{x}_0 - x^*\|^4 + 2\|\overline{v}_0\|^4 \mid x_0, v_0\right] \\ &\leq 2\gamma^4\mathbb{E}\left\|x_0 - x^* + \frac{1-\eta^2}{\gamma}v_0 + \sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}(h/2, \xi_0^{(1)}) - \mathcal{Z}^{(2)}(h/2, \xi_0^{(1)}, \xi_0^{(2)})\right)\right\|^4 \\ &+ 2\mathbb{E}\left\|\eta v_0 + \sqrt{2\gamma}\mathcal{Z}^{(2)}(h/2, \xi_0^{(1)}, \xi_0^{(2)})\right\|^4 \\ &\leq 4\gamma^4\|x_0 - x^*\|^4 + 4\|v_0\|^4 + \mathbb{E}\left[32\gamma^2\|\mathcal{Z}^{(1)}(h/2, \xi_0^{(1)})\|^4 + 40\gamma^2\|\mathcal{Z}^{(2)}(h/2, \xi_0^{(1)}, \xi_0^{(2)})\|^4\right] \\ &\leq 4\gamma^4\|x_0 - x^*\|^4 + 4\|v_0\|^4 + 8\gamma^2h^2d^2 + 480\gamma^2h^2d^2,\end{aligned}$$

where we have used that $U(\overline{x}_0) - U(x^*) \leq \frac{M}{2}\|\overline{x}_0 - x^*\|^2$ in the first inequality and naive bounds on the fourth moments of the Gaussian increments. Hence, we arrive at the estimate

$$\begin{aligned}\mathbb{E}\left[\|\overline{x}_k - x^*\|^4 \mid x_0, v_0\right] \leq \frac{4}{m^2}\Bigg[&4\left(1 - \frac{c_4(h)}{2}\right)^k \left(\gamma^4\|x_0 - x^*\|^4 + \|v_0\|^4 + 122\gamma^2h^2d^2\right) \\ &+ 2\frac{\frac{(6h\gamma d+160h\gamma(1+\lambda^2))^2}{4c_4(h)} + 24h^2\gamma^2d^2}{c_4(h)}\Bigg],\end{aligned}$$

for the UBU scheme with full gradients.

□

B.4 Variance bounds for UBUBU estimator with exact gradients

B.4.1 Variance bound of $D_{l,l+1}$

To bound the variance of $D_{l,l+1}$ we use strong error estimates for the UBU integrator using the results of [140].

In this analysis we define for random vectors $z_1, z_2 \in \mathbb{R}^{2d}$ the L^2 norm $\|z_1\|_{L^2,a,b} = \mathbb{E}\left(\|z_1\|_{a,b}^2\right)^{1/2}$ and respective inner product $\langle z_1, z_2\rangle_{L^2,a,b} = \mathbb{E}\left(z_1^T \mathcal{M} z_2\right)$, where

$$\mathcal{M} = \begin{pmatrix} I_d & bI_d \\ bI_d & aI_d \end{pmatrix}.$$

Assumption B.4.1 (Local Strong Error [140]). *Let $\phi\left(z, t, (W_s)_{s=0}^t\right)$ be the solution of the continuous dynamics* (3.1.1) *with initial condition $z \in \mathbb{R}^{2d}$ up to time t, with Brownian motion $(W_s)_{s=0}^t$. Let $\psi_h\left(z, t, (W_s)_{s=0}^t\right)$ be the solution of a numerical discretization with initial condition $z \in \mathbb{R}^{2d}$ up to time t, with Brownian motion $(W_s)_{s=0}^t$ and stepsize h. Let $z' \sim \pi$, then we assume that*

$$\psi_h\left(z', h, (W_s)_{s=0}^h\right) - \phi\left(z', h, (W_s)_{s=0}^h\right) = \alpha_h\left(z', (W_s)_{s=0}^h\right) + \beta_h\left(z', (W_s)_{s=0}^h\right),$$

where

$$\left\|\alpha_h\left(z', (W_s)_{s=0}^h\right)\right\|_{L^2,a,b} \leq C_1 h^{q+1/2},$$
$$\left\|\beta_h\left(z', (W_s)_{s=0}^h\right)\right\|_{L^2,a,b} \leq C_2 h^{q+1},$$

and

$$\left|\left\langle \psi_h\left(z', h, (W_s)_{s=0}^h\right) - \psi_h\left(z, h, (W_s)_{s=0}^h\right), \alpha_h\left(z', (W_s)_{s=0}^h\right)\right\rangle_{L^2,a,b}\right|$$
$$\leq C_0 h \left\|z' - z\right\|_{L^2,a,b} \left\|\alpha_h\left(z', (W_s)_{s=0}^h\right)\right\|_{L^2,a,b}.$$

for some $C_0, C_1, C_2 > 0$.

We restate Assumptions 3.3.6-3.3.13 here for easier readability.

Assumption 3.3.6 (M-∇ Lipschitz). *$U : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable and there exists $M > 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\|\nabla U(x) - \nabla U(y)\| \leq M\|x - y\|.$$

Assumption 3.3.7 (m-strong convexity). *$U : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and there exists $m > 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\langle \nabla U(x) - \nabla U(y), x - y\rangle \geq m|x - y|^2.$$

Assumption 3.3.10 (M_1^s-strongly Hessian Lipschitz). *$U : \mathbb{R}^d \to \mathbb{R}$ is three times continuously differentiable and M_1^s-strongly Hessian Lipschitz if there exists $M_1^s > 0$ such that*

$$\|\nabla^3 U(x)\|_{\{1,2\}\{3\}} \leq M_1^s$$

for all $x \in \mathbb{R}^d$.

Assumption 3.3.12 (1-Lipschitzness of f)**.** *f is a 1-Lipschitz function with respect to the Euclidean distance on $\mathbb{R}^{2d}$, that only depends on x, not v (i.e. $f(x,v) = f(x,v')$ for any $x, v, v' \in \mathbb{R}^d$).*

Assumption 3.3.13 (Distance of initial distribution from target)**.** *The initial distribution on $\Lambda = \mathbb{R}^{2d}$ satisfy that $\mathcal{W}_2(\pi, \mu_0) \leq c_{\mu_0}\sqrt{\frac{d}{m}}$, for some $c_{\mu_0} > 0$.*

We make use of the following proposition, essentially due to [140].

Proposition B.4.2. *Suppose a numerical scheme approximating* (3.1.1) *satisfies Assumption B.4.1, with a potential which satisfies Assumptions 3.3.6-3.3.10, and $\psi_h(z, h, (W_s)_0^h) \sim P_h(z, \cdot)$ satisfies the Wasserstein contractivity condition* (B.3) *for $p = 2$, and some $a, b > 0$, $b^2 < a$.*

Let $\phi\left(z, t, (W_s)_{s=0}^t\right)$ be the solution of the continuous dynamics (3.1.1) *with initial condition $z \in \mathbb{R}^{2d}$ up to time t, with Brownian motion $(W_s)_{s=0}^t$. Let $\psi_h\left(z, t, (W_s)_{s=0}^t\right)$ be the solution of a numerical discretization with initial condition $z \in \mathbb{R}^{2d}$ up to time t, with Brownian motion $(W_s)_{s=0}^t$ and stepsize $h > 0$ satisfying that*

$$(1 - c(h))^2 + C_0^2 h^2 < 1. \tag{B.19}$$

Then for any $k \geq 0$, any z_0 such that $\|z_0\|_{L^2,a,b} < \infty$, and $Z^0 \sim \pi$, we have

$$\begin{aligned}&\left\|\psi_h\left(z_0, kh, (W_s)_{s=0}^{kh}\right) - \phi\left(Z^0, kh, (W_s)_{s=0}^{kh}\right)\right\|_{L^2,a,b} \\ &\leq (1 - R(h))^k \|z_0 - Z^0\|_{L^2,a,b} + \sqrt{2}C_1 \frac{h^{q+1/2}}{\sqrt{R(h)}} + \frac{2C_2 h^{q+1}}{R(h)},\end{aligned}$$

where $R(h) = 1 - \sqrt{(1 - c(h))^2 + C_0^2 h^2}$.

In particular, the discretization scheme admits a stationary distribution π_h, and its bias can be bounded as

$$\mathcal{W}_{2,a,b}(\pi_h, \pi) \leq \sqrt{2}C_1 \frac{h^{q+1/2}}{\sqrt{R(h)}} + \frac{2C_2 h^{q+1}}{R(h)}. \tag{B.20}$$

Proof. Introduce the notation

$$Z^n := \phi\left(Z^0, nh, (W_s)_{s=0}^{nh}\right), \quad z_n := \psi_h\left(z_0, nh, (W_s)_{s=0}^{nh}\right)$$

for all $n \in \mathbb{N}$. Using the assumption $Z^0 \sim \pi$, we also have $Z^n \sim \pi$, since the kinetic Langevin dynamics keeps π invariant. By Assumption B.4.1, we then have

$$
\begin{aligned}
&\left\|z_k - Z^k\right\|_{L^2,a,b} = \left\|\psi_h\left(z_{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \phi\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right)\right\|_{L^2,a,b} \\
&= \Bigg\|\psi_h\left(z_{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \psi_h\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) \\
&+ \psi_h\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \phi\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right)\Bigg\|_{L^2,a,b} \\
&= \Bigg\|\psi_h\left(z_{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \psi_h\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) \\
&+ \alpha_h\left(Z^{k-1}, (W_s)_{(k-1)h}^{kh}\right) + \beta_h\left(Z^{k-1}, (W_s)_{(k-1)h}^{kh}\right)\Bigg\|_{L^2,a,b} \\
&\leq \left\|\beta^{k-1}\right\|_{L^2,a,b} + \left\|\psi_h\left(z_{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \psi_h\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) + \alpha^{k-1}\right\|_{L^2,a,b},
\end{aligned}
\tag{B.21}
$$

where α^{k-1} and β^{k-1} are defined as

$$
\begin{aligned}
&\psi_h\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \phi\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) \\
&= \alpha_h\left(Z^{k-1}, (W_s)_{(k-1)h}^{kh}\right) + \beta_h\left(Z^{k-1}, (W_s)_{(k-1)h}^{kh}\right) \\
&:= \alpha^{k-1} + \beta^{k-1}.
\end{aligned}
$$

Assumption B.4.1, and the Wasserstein contractivity condition (B.3) then together imply

$$
\begin{aligned}
&\left\|\psi_h\left(z_{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \psi_h\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) + \alpha^{k-1}\right\|_{L^2,a,b} \\
&= \Bigg(\left\|\alpha^{k-1}\right\|_{L^2,a,b}^2 + \left\|\psi_h\left(z_{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \psi_h\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right)\right\|_{L^2,a,b}^2 \\
&+ 2\left\langle\alpha^{k-1}, \psi_h\left(z_{k-1}, h, (W_s)_{(k-1)h}^{kh}\right) - \psi_h\left(Z^{k-1}, h, (W_s)_{(k-1)h}^{kh}\right)\right\rangle_{L^2,a,b}\Bigg)^{1/2} \\
&\leq \Bigg(\left\|\alpha^{k-1}\right\|_{L^2,a,b}^2 + (1-c(h))^2\left\|z_{k-1} - Z^{k-1}\right\|_{L^2,a,b}^2 \\
&+ 2C_0 h\left\|\alpha^{k-1}\right\|_{L^2,a,b}\left\|z_{k-1} - Z^{k-1}\right\|_{L^2,a,b}\Bigg)^{1/2} \\
&\leq \left(2\left\|\alpha^{k-1}\right\|_{L^2,a,b}^2 + ((1-c(h))^2 + C_0^2 h^2)\left\|z_{k-1} - Z^{k-1}\right\|_{L^2,a,b}^2\right)^{1/2} \\
&\leq \left(2C_1^2 h^{2q+1} + ((1-c(h))^2 + C_0^2 h^2)\left\|z_{k-1} - Z^{k-1}\right\|_{L^2,a,b}^2\right)^{1/2}.
\end{aligned}
$$

Lemma 28 of [140] states that if a sequence of nonnegative real numbers $(a_n)_{n\geq 0}$ satisfies that $a_{n+1} \leq \sqrt{(1-A)^2 a_n^2 + B} + C$ with $A \in (0,1)$, $B \geq 0$, $C \geq 0$, then for every $n \geq 0$,

$$
a_n \leq (1-A)^n a_0 + \sqrt{\frac{B}{A}} + \frac{C}{A}.
$$

Using this for $a_n = \|z_n - Z^n\|_{L^2,a,b}$, we have that

$$\left\|z_k - Z^k\right\|_{L^2,a,b} \leq (1 - R(h))^k \|z_0 - Z^0\|_{L^2,a,b} + \sqrt{2}\frac{C_1 h^{q+1/2}}{\sqrt{R(h)}} + \frac{2C_2 h^{q+1}}{R(h)},$$

where $R(h) = 1 - \sqrt{(1 - c(h))^2 + C_0^2 h^2}$, which is our first claim.

The existence of a stationary distribution π_h follows from Lemma B.3.4. The bound on the bias follows by letting $k \to \infty$. □

We now are in a position to present our first result related to the variance of our unbiased scheme, which is a bound on the variance related to the global strong error or convergence. This is given below.

Proposition B.4.3. *Suppose a numerical scheme approximating* (3.1.1) *satisfies the same assumptions as in Proposition B.4.2, and f satisfies Assumption 3.3.12. If we have two chains at coarser and finer discretization levels l and $l+1$ using stepsizes h_l and $h_{l+1} = \frac{h_l}{2}$ satisfying* (B.19) *with synchronously coupled Brownian motions $(z_k)_{k\in\mathbb{N}}$ and $(z'_k)_{k\in\mathbb{N}}$, such that $z_0 \sim \pi_0$ and $z'_0 \sim \pi'_0$, then we have*

$$\begin{aligned}
&\mathrm{Var}\left(f(z'_k) - f(z_k)\right) \leq \mathbb{E}\left[\left(f(z'_k) - f(z_k)\right)^2\right] \leq \mathbb{E}\|z'_k - z_k\|^2_{a,b} \\
&\leq \left(\exp\left(-\frac{mkh_l}{8\gamma}\right)\left(\|z'_0 - z_0\|_{L^2,a,b} + \mathcal{W}_{2,a,b}(\pi_0,\pi) + \mathcal{W}_{2,a,b}(\pi'_0,\pi)\right)\right. \\
&+ (1 - R(h_l))^k \mathcal{W}_{2,a,b}(\pi_0,\pi) + (1 - R(h_{l+1}))^{2k}\mathcal{W}_{2,a,b}(\pi'_0,\pi) \\
&\left.+ \sqrt{2}C_1\left(\frac{h_{l+1}^{q+1/2}}{\sqrt{R(h_{l+1})}} + \frac{h_l^{q+1/2}}{\sqrt{R(h_l)}}\right) + 2C_2\left(\frac{h_{l+1}^{q+1}}{R(h_{l+1})} + \frac{h_l^{q+1}}{R(h_l)}\right)\right)^2,
\end{aligned}$$

where $R(h_i) = 1 - \sqrt{(1 - c(h_i))^2 + C_0^2 h_i^2}$ for $i = l, l+1$.

Proof of Proposition B.4.3. Consider the following variance bound:

$$\mathrm{Var}\left(f(z'_k) - f(z_k)\right) \leq \mathbb{E}\left[\left(f(z'_k) - f(z_k)\right)^2\right] \leq \mathbb{E}\|z'_k - z_k\|^2_{a,b}.$$

Let $\tilde{Z}_0 \sim \pi$ be such that $\|\tilde{Z}_0 - z_0\|_{L^2,a,b} = \mathcal{W}_{2,a,b}(\pi_0,\pi)$, and $\tilde{Z}'_0 \sim \pi$ be such that $\|\tilde{Z}'_0 - z'_0\|_{L^2,a,b} = \mathcal{W}_{2,a,b}(\pi'_0,\pi)$ (the existence of optimal couplings was shown in Theorem 4.1 of [159]). We use the estimate

$$\begin{aligned}
&\sqrt{\mathbb{E}\|z'_k - z_k\|^2_{a,b}} = \|z'_k - z_k\|_{L^2,a,b} \\
&\leq \left\|z_k - \phi\left(\tilde{Z}_0, kh_l, (W_s)_{s=0}^{kh_l}\right)\right\|_{L^2,a,b} \\
&+ \left\|\phi\left(\tilde{Z}_0, kh_l, (W_s)_{s=0}^{kh_l}\right) - \phi\left(\tilde{Z}'_0, kh_l, (W_s)_{s=0}^{kh_l}\right)\right\|_{L^2,a,b} \\
&+ \left\|z'_k - \phi\left(\tilde{Z}'_0, kh_l, (W_s)_{s=0}^{kh_l}\right)\right\|_{L^2,a,b} \\
&=: (\mathsf{I}) + (\mathsf{II}) + (\mathsf{III}).
\end{aligned}$$

We split this into two global error terms (I) and (III) and a contraction term (II). We estimate the second term by Corollary B.3.8 as

$$\begin{aligned}(\text{II}) &\leq \exp\left(-\frac{mkh_l}{8\gamma}\right)\|\tilde{Z}_0' - \tilde{Z}_0\|_{L^2,a,b} \\ &\leq \exp\left(-\frac{mkh_l}{8\gamma}\right)\left(\|z_0' - z_0\|_{L^2,a,b} + \mathcal{W}_{2,a,b}(\pi_0,\pi) + \mathcal{W}_{2,a,b}(\pi_0',\pi)\right).\end{aligned}$$

By Proposition B.4.2, we have

$$(\text{I}) \leq (1-R(h_l))^k \mathcal{W}_{2,a,b}(\pi_0,\pi) + \sqrt{2}C_1\frac{2h_l^{q+1/2}}{\sqrt{R(h_l)}} + \frac{2C_2h_l^{q+1}}{R(h_l)}.$$

The same argument can be applied to (III) to obtain

$$(\text{III}) \leq (1-R(h_{l+1}))^{2k} \mathcal{W}_{2,a,b}(\pi_0',\pi) + \sqrt{2}C_1\frac{2h_{l+1}^{q+1/2}}{\sqrt{R(h_{l+1})}} + \frac{2C_2h_{l+1}^{q+1}}{R(h_{l+1})}.$$

Combining these we get the required result. □

Below are a number of useful remarks to highlight from the above theorem.

Remark B.4.4. *The local error, which arises from [140] is demonstrated through the bound on $\alpha_h+\beta_h$ from Assumption B.4.1. This indicates there is an order of local strong order $q+1/2$. However, when we go to the global strong order, the order is only reduced by $1/2$ as it is order q. As stated in [140], this is similar to the Euler–Maruyama scheme with local strong order $3/2$, but global strong order 1 [114, Theorem 1.1].*

Remark B.4.5. *Proposition B.4.3 holds for $q = 2$ for the UBU scheme; [140] showed that the assumptions are true. For the UBU scheme we have for $\gamma^2 \geq M$ and $h < \frac{1}{2\gamma}$ that $C_2 \leq \sqrt{d}\left(\frac{7}{10}\gamma^2 + \frac{M_1^s}{10\sqrt{M}}\right)$, $C_1 = \frac{\sqrt{6dM\gamma}}{24}$ and $C_0 \leq 4\sqrt{2M}$. These constants can be computed by following [140, Section 7.6] where all computations are done with arbitrary γ, the constant c we consider to be set to 1 in their estimates. Constants C_1 and C_2 are estimated in the second and third step, whilst C_0 is estimated in the fourth step and fifth step. We remark that there is a missing term and a stronger assumption is needed in [140, Section 7.6, fifth step] which has been corrected in [123]. The additional term can be treated by the same argument as in the fourth step to arrive at the C_0 bound.*

Corollary B.4.6. *Suppose that Assumptions 3.3.6, 3.3.7, 3.3.10, and 3.3.13 hold, $\gamma \geq \sqrt{8M}$ and*

$$h_0 \leq \frac{1}{\gamma}\cdot\frac{m}{264M}. \tag{B.22}$$

Assume that the burn-in periods $B \geq \frac{16\log(4)\gamma}{mh_0}$, $B_0 \geq \frac{16\gamma}{mh_0}\log\left(\frac{c_{\mu_0}+1}{\sqrt{M}\gamma h_0^2}\right)$. *Then for every* $l \geq 0$, $1 \leq k \leq K$, *the UBUBU samples satisfy*

$$\mathrm{Var}\left(f(z_k'^{(l,l+1)}) - f(z_k^{(l,l+1)})\right) \leq \mathbb{E}\left[\left(f(z_k'^{(l,l+1)}) - f(z_k^{(l,l+1)})\right)^2\right] \leq \mathbb{E}\|z_k'^{(l,l+1)} - z_k^{(l,l+1)}\|_{a,b}^2$$
$$\leq Cd\left(\left(\gamma^4 + \frac{(M_1^s)^2}{M}\right)\left(\frac{\gamma}{m}\right)^2 + \frac{M\gamma^2}{m}\right)h_l^4.$$

Proof of Corollary B.4.6. We have $(B_0 + Bl)2^l$ burn-in steps at level l, and $(B_0 + B(l+1))2^{l+1}$ burn-in steps at level $l+1$. Let $\delta_* = \delta_{x^*} \times \delta_{0_d}$ be a distribution on Λ that fixes $x = x^*$ and $v = 0_d$. Using the assumptions, we have

$$R(h_i) = 1 - \sqrt{(1-c(h_i))^2 + C_0^2h_i^2} = 1 - \sqrt{\left(1 - \frac{mh_i}{8\gamma}\right)^2 + C_0^2h_i^2}$$
$$= 1 - \sqrt{1 - \frac{mh_i}{4\gamma} + \left(\left(\frac{m}{8\gamma}\right)^2 + C_0^2\right)h_i^2} \geq 1 - \sqrt{1 - \frac{mh_i}{8\gamma}} \geq \frac{mh_i}{16\gamma},$$
$$\mathcal{W}_{2,a,b}(\pi_0, \pi) = \mathcal{W}_{2,a,b}(\mu_0, \pi) \leq c_{\mu_0}\sqrt{\frac{d}{m}},$$
$$\mathcal{W}_{2,a,b}(\pi_0', \pi) \leq \mathcal{W}_{2,a,b}\left(\mu_0 R_{l+1}^B, \pi_{h_{l+1}}\right) + \mathcal{W}_{2,a,b}\left(\pi_{h_{l+1}}, \pi\right),$$
$$\leq \mathcal{W}_{2,a,b}(\mu_0, \pi) + 2\mathcal{W}_{2,a,b}\left(\pi_{h_{l+1}}, \pi\right) \leq c_{\mu_0}\sqrt{\frac{d}{m}} + 2\sqrt{2}C_1\frac{h_{l+1}^{q+1/2}}{\sqrt{R(h_{l+1})}} + \frac{4C_2h_{l+1}^{q+1}}{R(h_{l+1})},$$

and

$$\|z_0' - z_0\|_{L^2,a,b} \leq \mathcal{W}_{2,a,b}(\pi_0', \delta_*) + \mathcal{W}_{2,a,b}(\pi_0, \delta_*)$$
$$\leq \mathcal{W}_{2,a,b}(\pi_0', \pi) + \mathcal{W}_{2,a,b}(\mu_0, \pi) + 2\mathcal{W}_{2,a,b}(\pi, \delta_*)$$
$$\leq (3c_{\mu_0} + 3)\sqrt{\frac{d}{m}} + 2\sqrt{2}C_1\frac{h_{l+1}^{q+1/2}}{\sqrt{R(h_{l+1})}} + \frac{4C_2h_{l+1}^{q+1}}{R(h_{l+1})}.$$

It is easy to check that (B.22) together with $C_0 \leq 4\sqrt{2M}$ implies that the condition (B.19) of Proposition B.4.3 is satisfied, and we have

$$\mathrm{Var}\left(f(z_k'^{(l,l+1)}) - f(z_k^{(l,l+1)})\right) \leq \mathbb{E}\left[\left(f(z_k'^{(l,l+1)}) - f(z_k^{(l,l+1)})\right)^2\right]$$
$$\leq \left(\exp\left(-\frac{m(B_0+lB)h_0}{8\gamma}\right)\left(\|z_0' - z_0\|_{L^2,a,b} + \mathcal{W}_{2,a,b}(\pi_0, \pi) + \mathcal{W}_{2,a,b}(\pi_0', \pi)\right)\right.$$
$$+ \exp\left(-\frac{m(B_0+lB)h_0}{16\gamma}\right)\left(\mathcal{W}_{2,a,b}(\pi_0, \pi) + \mathcal{W}_{2,a,b}(\pi_0', \pi)\right)$$
$$\left. + \sqrt{2}C_1\left(\frac{h_{l+1}^{q+1/2}}{\sqrt{R(h_{l+1})}} + \frac{h_l^{q+1/2}}{\sqrt{R(h_l)}}\right) + 2C_2\left(\frac{h_{l+1}^{q+1}}{R(h_{l+1})} + \frac{h_l^{q+1}}{R(h_l)}\right)\right)^2$$
$$\leq \left(\exp\left(-\frac{m(B_0+lB)h_0}{16\gamma}\right)(7c_{\mu_0}+3)\sqrt{\frac{d}{m}} + 10\sqrt{2}C_1\left(\frac{h_l^{5/2}}{\sqrt{\frac{mh_l}{16\gamma}}}\right) + 20C_2\left(\frac{h_l^3}{\frac{mh_l}{16\gamma}}\right)\right)^2$$

using the assumptions on B_0 and B

$$\leq C\left(C_1^2\frac{\gamma}{m}+C_2^2\left(\frac{\gamma}{m}\right)^2\right)h_l^4 \leq Cd\left(\left(\gamma^4+\frac{(M_1^s)^2}{M}\right)\left(\frac{\gamma}{m}\right)^2+\frac{M\gamma^2}{m}\right)h_l^4.$$

□

Proposition B.4.7. *Suppose that the assumptions of Proposition B.3.6 hold for $h = h_l$. Let $R_{l,l+1} = (P_{h_l,h_{l+1}})^{2^l}$ be the Markov kernel defined in Section 3.3.1 for two synchronously coupled UBU chains at discretization levels $l, l+1$. This chain is moving on state space Λ^2. Let $\overline{z}_1, \ldots, \overline{z}_K$ be a Markov chain with kernel $R_{l,l+1}$. Let $F : \Lambda^2 \to \mathbb{R}$ be 1-Lipschitz in norm $\|\|_{a,b}$ on Λ^2, defined as $\|z_1, z_2\|_{a,b}^2 = \|z_1\|_{a,b}^2 + \|z_2\|_{a,b}^2$. Then we have*

$$\mathrm{Var}\left(\frac{\sum_{i=1}^K F(\overline{z}_i)}{K}\right) \leq \frac{2}{K^2}\sum_{i=1}^K\sum_{k=0}^{K-i}\min\left(\frac{\mathrm{Var}(F(\overline{z}_i))+\mathrm{Var}(F(\overline{z}_{i+k}))}{2},\right.$$
$$\left.\sqrt{\mathrm{Var}(F(\overline{z}_i))\mathbb{E}\left[\|\overline{z}_i-\mathbb{E}\overline{z}_i\|_{a,b}^2\right]}\cdot\exp\left(-\frac{mh_0}{8\gamma}\cdot k\right)\right),$$

Proof. We need to bound

$$\mathrm{Cov}(F(\overline{z}_i), F(\overline{z}_{i+k})) = \mathbb{E}[(F(\overline{z}_i)-\mathbb{E}(F(\overline{z}_i)))(F(\overline{z}_{i+k})-\mathbb{E}(F(\overline{z}_{i+k})))].$$

Let $\tilde{z}_i$ be an independent identically distributed copy of $\overline{z}_i$. For $0 \leq l \leq K-i-1$, and assume that conditioned on $\tilde{z}_{i:i+l}$ and $\overline{z}_{1:i+l}$, $\tilde{z}_{i+l+1} \sim P(\tilde{z}_{i+l}, \cdot)$, and $(\overline{z}_{i+l+1}, \tilde{z}_{i+l+1})$ are synchronously coupled, i.e. $\tilde{z}_{i+l+1}$ is defined based on (3.3.25) using the same Gaussian variables that were used to move from $\overline{z}_{i+l}$ to $\overline{z}_{i+l+1}$. Since we have also used synchronous couplings in the proof of Proposition B.3.6, it follows from Proposition B.3.6 that

$$\mathbb{E}\left(\|\tilde{z}_{i+l+1}-\overline{z}_{i+l+1}\|_{a,b}^2\middle|\,\overline{z}_{i:i+l},\tilde{z}_{i:i+l}\right)$$
$$\leq \max\left(\left(1-\frac{mh_l}{8\gamma}\right)^{2\cdot 2^l},\left(1-\frac{mh_{l+1}}{8\gamma}\right)^{2\cdot 2^{l+1}}\right)\|\tilde{z}_{i+l}-\overline{z}_{i+l}\|_{a,b}^2$$

using that $1-x \leq \exp(-x)$ for $x \geq 0$,

$$\leq \exp\left(-\frac{mh_0}{4\gamma}\right)\|\tilde{z}_{i+l}-\overline{z}_{i+l}\|_{a,b}^2.$$

By using this bound recursively, we have

$$\mathbb{E}\left(\|\tilde{z}_{i+k}-\overline{z}_{i+k}\|_{a,b}^2\middle|\,\overline{z}_i,\tilde{z}_i\right) \leq \exp\left(-\frac{mh_0}{4\gamma}\cdot k\right)\|\overline{z}_i-\tilde{z}_i\|_{a,b}^2.$$

Since $\tilde{z}_i$ is independent of $\overline{z}_i$, and $\tilde{z}_i + 1, \ldots, \tilde{z}_{i+k}$ was constructed using $\tilde{z}_i$ and Gaussians that are independent of $\overline{z}_i$ (synchronous coupling with $\overline{z}_{i+1}, \ldots, \overline{z}_{i+k}$), it follows that $\tilde{z}_{i+k}$ is still independent of $\overline{z}_i$. Using this and the 1-Lipschitz property of F, we have

$$\begin{aligned}
&\mathrm{Cov}(F(\overline{z}_i), F(\overline{z}_{i+k})) = \mathbb{E}[(F(\overline{z}_i) - \mathbb{E}(F(\overline{z}_i)))(F(\overline{z}_{i+k}) - \mathbb{E}(F(\overline{z}_{i+k})))\\
&= \mathbb{E}[(F(\overline{z}_i) - \mathbb{E}(F(\overline{z}_i)))(F(\overline{z}_{i+k}) - F(\tilde{z}_{i+k}))]\\
&= \mathbb{E}[(F(\overline{z}_i) - \mathbb{E}(F(\overline{z}_i)))\mathbb{E}(F(\overline{z}_{i+k}) - F(\tilde{z}_{i+k})|\overline{z}_i, \tilde{z}_i)]\\
&\leq \sqrt{\mathrm{Var}(F(\overline{z}_i))\mathbb{E}\left[\|\overline{z}_{i+k} - \tilde{z}_{i+k}\|_{a,b}^2\right]}\\
&\leq \exp\left(-\frac{mh_0}{8\gamma} \cdot k\right)\sqrt{\mathrm{Var}(F(\overline{z}_i)) \cdot \mathbb{E}\left[\|\overline{z}_i - \tilde{z}_i\|_{a,b}^2\right]}\\
&= \exp\left(-\frac{mh_0}{8\gamma} \cdot k\right)\sqrt{2\mathrm{Var}(F(\overline{z}_i)) \cdot \mathbb{E}\left[\|\overline{z}_i - \mathbb{E}\overline{z}_i\|_{a,b}^2\right]},
\end{aligned}$$

and the claim follows by summation. □

Proposition B.4.8. *Under the same assumptions as in Corollary B.4.6, the UBUBU samples satisfy that*

$$\mathrm{Var}(D_{l,l+1}) \leq \frac{1}{K}C(\gamma, m, M, M_1^s)dh_l^4\left(C(\gamma, m, M, M_1^s) - 2\log(h_0) + \log(4)l + \frac{4\gamma}{mh_0}\right).$$

Proof. Note that the function $F(z_1, z_1) = f(z_1) - f(z_2)$ is 1-Lipschitz with respect to $\|(z_1, z_2)\|_{a,b} = \|z_1\|_{a,b} + \|z_2\|_{a,b}$. Let $\overline{z}_i = (z_i^{(l,l+1)}, z_i'^{(l,l+1)})$, then by Proposition B.4.7, we have that

$$\begin{aligned}
\mathrm{Var}(D_{l,l+1}) \leq \frac{1}{K^2}\sum_{i=1}^{K}\sum_{k=0}^{K-i}\min\Bigg(&\frac{\mathrm{Var}(F(\overline{z}_i)) + \mathrm{Var}(F(\overline{z}_{i+k}))}{2},\\
&\sqrt{\mathrm{Var}(F(\overline{z}_i))\mathbb{E}\left[\|\overline{z}_i - \mathbb{E}\overline{z}_i\|_{a,b}^2\right]} \cdot \exp\left(-\frac{mh_0}{8\gamma} \cdot k\right)\Bigg).
\end{aligned}$$

By a similar argument as in the proof of Corollary B.4.6, using our assumptions on B and B_0, we can show that

$$\left(\mathbb{E}\left[\|\overline{z}_i - \mathbb{E}\overline{z}_i\|_{a,b}^2\right]\right)^{1/2} \leq \left(\mathbb{E}\left[\|\overline{z}_i - (x^*, 0_d, x^*, 0_d)\|_{a,b}^2\right]\right)^{1/2} \leq C\sqrt{\frac{d}{m}},$$

and by Proposition B.4.3, we have

$$\mathrm{Var}(F(\overline{z}_i)) \leq C(\gamma, m, M, M_1^s)dh_l^4.$$

Let

$$k^*(l) := \max\left(\log\left(C\sqrt{\frac{1}{m}}\right) - \frac{1}{2}\log\left(C(\gamma, m, M, M_1)h_l^4\right), 0\right),$$

then for $k \geq \lceil k^*(l) \rceil$, we have

$$\sqrt{\mathrm{Var}(F(\overline{z}_i))\mathbb{E}\left[\|\overline{z}_i - \mathbb{E}\overline{z}_i\|_{a,b}^2\right]} \cdot \exp\left(-\frac{mh_0}{8\gamma} \cdot k\right) \\ \leq C(\gamma, m, M, M_1^s) d h_l^4 \exp\left(-\frac{mh_0}{8\gamma} \cdot (k - \lceil k^*(l) \rceil)\right).$$

It is clear that $\lceil k^*(l) \rceil \leq C(\gamma, m, M, M_1^s) - 2\log(h_0) + \log(4)l$, and after some rearrangement, we have

$$\mathrm{Var}(D_{l,l+1}) \leq \frac{1}{K} C(\gamma, m, M, M_1^s) d h_l^4 \left(C(\gamma, m, M, M_1^s) - 2\log(h_0) + \log(4)l + \frac{4\gamma}{mh_0}\right).$$

□

B.4.2 Variance bound of D_0

Proposition B.4.9. *Consider an m-strongly convex M-∇Lipschitz potential U and let P_h be the transition kernel of UBU with stepsize h. Suppose that $f : \Omega \to \mathbb{R}$ only depends on x and is a 1-Lipschitz function. Suppose that $\gamma \geq \sqrt{8M}$, and $h < \frac{1}{2\gamma}$. Let μ_0 be a distribution on Λ, and the Markov chain $z_{-B_0}^{(0)} \sim \mu_0$, $z_{-B_0+1}^{(0)} \sim P_h(z_{-B_0}^{(0)}, \cdot), \ldots, z_K \sim P_h(z_{K-1}^{(0)}, \cdot)$. Then D_0 as defined in* (3.3.21) *satisfies that*

$$\mathrm{Var}(D_0) \leq \frac{C}{c(h)K}\left(1 + \frac{1}{c(h)K}\right)\left(\frac{1}{\gamma} + \frac{\gamma}{M}\right)\frac{h}{c(h)} + \frac{(1-c(h))^{2(B_0+1)}}{2c(h)^2K^2}\sigma_{\mu_0}^2,$$

where

$$c(h) = \frac{mh}{8\gamma}, \quad \sigma_{\mu_0}^2 = \int\int \|w - \tilde{w}\|_{a,b}^2 d\mu_0(w) d\mu_0(\tilde{w}),$$

for some absolute constant C.

Proof. The bound is based on Theorem 2 of [89]. We need to control the following quantities for every $z \in \Lambda$:

$$\sigma(z)^2 := \frac{1}{2}\int\int \|w - \tilde{w}\|_{a,b}^2 P_h(z, dw) P_h(z, d\tilde{w}), \tag{B.23}$$

$$n(z) := \inf_{g:\Lambda\to\mathbb{R}, \|g\|_{a,b,Lip} \leq 1} \frac{\int\int \|w - \tilde{w}\|_{a,b}^2 P_h(z, dw) P_h(z, d\tilde{w})}{\int\int (g(w) - g(\tilde{w}))^2 P_h(z, dw) P_h(z, d\tilde{w})}. \tag{B.24}$$

Here we choose $a = \frac{1}{M}$, and $b = \frac{1}{\gamma}$ as in Proposition B.3.6. To control $\sigma^2(z)$, let us define two independent identically distributed random variables $w(z) \sim P_h(z,\cdot)$ and $\tilde{w}(z) \sim P_h(z,\cdot)$. Using the definition of UBU in (3.2.9), we have

$$
\begin{aligned}
\sigma(z)^2 &= \frac{1}{2}\mathbb{E}(\|w(z) - \tilde{w}(z)\|_{a,b}^2) \\
&= \frac{1}{2}\mathbb{E}\left(\left\| \mathcal{UBU}\left(z, h, \xi^{(1)}, \xi^{(2)}, \xi^{(3)}, \xi^{(4)}\right) - \mathcal{UBU}\left(z, h, \tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}, \tilde{\xi}^{(3)}, \tilde{\xi}^{(4)}\right) \right\|_{a,b}^2 \right) \\
&\leq \mathbb{E}\Bigg(\Bigg\| \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z, h/2, \xi^{(1)}, \xi^{(2)}\right), h\right), h/2, \xi^{(3)}, \xi^{(4)}\right) \\
&\quad - \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z, h/2, \xi^{(1)}, \xi^{(2)}\right), h\right), h/2, \tilde{\xi}^{(3)}, \tilde{\xi}^{(4)}\right) \Bigg\|_{a,b}^2 \Bigg) \\
&\quad + \mathbb{E}\Bigg(\Bigg\| \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z, h/2, \xi^{(1)}, \xi^{(2)}\right), h\right), h/2, \tilde{\xi}^{(3)}, \tilde{\xi}^{(4)}\right) \\
&\quad - \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z, h/2, \tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}\right), h\right), h/2, \tilde{\xi}^{(3)}, \tilde{\xi}^{(4)}\right) \Bigg\|_{a,b}^2 \Bigg).
\end{aligned}
$$

Recalling the definitions of $\mathcal{U}$ and $\mathcal{B}$ from equations (3.2.6-3.2.7), we have

$$
\begin{aligned}
\mathcal{B}(x, v, h) &= (x, v - h\nabla U(x)), \\
\mathcal{U}(x, v, h/2, \xi^{(1)}, \xi^{(2)}) &= \Big(x + \frac{1-\eta}{\gamma}v + \sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(h/2, \xi^{(1)}\right) - \mathcal{Z}^{(2)}\left(h/2, \xi^{(1)}, \xi^{(2)}\right)\right), \\
&\qquad \eta v + \sqrt{2\gamma}\mathcal{Z}^{(2)}\left(h/2, \xi^{(1)}, \xi^{(2)}\right)\Big), \\
\mathcal{Z}^{(1)}\left(h/2, \xi^{(1)}\right) &= \sqrt{\frac{h}{2}}\xi^{(1)}, \\
\mathcal{Z}^{(2)}\left(h/2, \xi^{(1)}, \xi^{(2)}\right) &= \sqrt{\frac{1-\eta^2}{2\gamma}}\left(\sqrt{\frac{1-\eta}{1+\eta}\cdot\frac{4}{\gamma h}}\xi^{(1)} + \sqrt{1 - \frac{1-\eta}{1+\eta}\cdot\frac{4}{\gamma h}}\xi^{(2)}\right).
\end{aligned}
$$

First,

$$
\begin{aligned}
&\mathbb{E}\Bigg(\Bigg\| \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z, h/2, \xi^{(1)}, \xi^{(2)}\right), h\right), h/2, \xi^{(3)}, \xi^{(4)}\right) \\
&- \mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z, h/2, \xi^{(1)}, \xi^{(2)}\right), h\right), h/2, \tilde{\xi}^{(3)}, \tilde{\xi}^{(4)}\right) \Bigg\|_{a,b}^2 \Bigg) \\
&= \mathbb{E}\Bigg(\Bigg\| \left(\sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(\frac{h}{2}, \xi^{(1)}\right) - \mathcal{Z}^{(2)}\left(\frac{h}{2}, \xi^{(1)}, \xi^{(2)}\right)\right) - \sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(\tilde{\xi}^{(1)}\right) - \mathcal{Z}^{(2)}\left(\frac{h}{2}, \xi^{(1)}, \tilde{\xi}^{(2)}\right)\right), \\
&\sqrt{2\gamma}\mathcal{Z}^{(2)}\left(\frac{h}{2}, \xi^{(1)}, \xi^{(2)}\right) - \sqrt{2\gamma}\mathcal{Z}^{(2)}\left(\frac{h}{2}, \tilde{\xi}^{(1)}, \tilde{\xi}^{(2)}\right)\Bigg) \Bigg\|_{a,b}^2 \Bigg)
\end{aligned}
$$

using (B.1), and the fact that $a = \frac{1}{M}$

$$\leq \frac{6}{\gamma}\mathbb{E}\left(\left\|\mathcal{Z}^{(1)}\left(\frac{h}{2},\xi^{(1)}\right)-\mathcal{Z}^{(2)}\left(\frac{h}{2},\xi^{(1)},\tilde{\xi}^{(2)}\right)\right\|^2\right)+\frac{6\gamma}{M}\mathbb{E}\left(\left\|\mathcal{Z}^{(2)}\left(\frac{h}{2},\xi^{(1)},\tilde{\xi}^{(2)}\right)\right\|^2\right)$$
$$\leq \left(\frac{3}{\gamma}+\frac{3\gamma}{M}\right)dh.$$

Second, using the assumptions $\gamma \geq \sqrt{8M}$ and $h \leq \frac{1}{\sqrt{M}}$, for any x, v, x', v',

$$\begin{aligned}
&\left\|\mathcal{U}\left(x,v,h/2,\tilde{\xi}^{(3)},\tilde{\xi}^{(4)}\right)-\mathcal{U}\left(x',v',h/2,\tilde{\xi}^{(3)},\tilde{\xi}^{(4)}\right)\right\|_{a,b}^2\\
&\leq \frac{3}{2}\left\|\mathcal{U}\left(x,v,h/2,\tilde{\xi}^{(3)},\tilde{\xi}^{(4)}\right)-\mathcal{U}\left(x',v',h/2,\tilde{\xi}^{(3)},\tilde{\xi}^{(4)}\right)\right\|_{a,0}^2\\
&\leq \frac{3}{2}\left(\frac{1}{M}\exp(-\gamma h)\|v-v'\|^2+2\|x-x'\|^2+\frac{2(1-\exp(-\gamma h/2))^2}{\gamma^2}\|v-v'\|^2\right)\\
&\leq 3\|x-x'\|^2+\frac{3}{2}\frac{1}{M}\|v-v'\|^2 \leq 6\|(x-x',v-v')\|_{a,b}^2, \qquad \text{(B.25)}\\
&\|\mathcal{B}(x,v,h)-\mathcal{B}(x',v',h)\|_{a,b}^2 \leq \frac{3}{2}\|(x-x',v-v'+h\nabla U(x')-h\nabla U(x))\|_{0,a}^2\\
&\leq \frac{3}{2}\|x-x'\|^2+\frac{3}{M}\|v-v'\|^2+\frac{3h^2}{M}\|\nabla U(x')-\nabla U(x)\|^2\\
&\leq \left(\frac{3}{2}+3h^2M\right)\|x-x'\|^2+\frac{3}{M}\|v-v'\|^2 \leq 6\|(x-x',v-v')\|_{a,b}^2 \qquad \text{(B.26)}
\end{aligned}$$

hence

$$\begin{aligned}
&\mathbb{E}\Bigg(\Bigg\|\mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z,h/2,\xi^{(1)},\xi^{(2)}\right),h\right),h/2,\tilde{\xi}^{(3)},\tilde{\xi}^{(4)}\right)\\
&-\mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z,h/2,\tilde{\xi}^{(1)},\tilde{\xi}^{(2)}\right),h\right),h/2,\tilde{\xi}^{(3)},\tilde{\xi}^{(4)}\right)\Bigg\|_{a,b}^2\Bigg)\\
&\leq 36\mathbb{E}\left(\left\|\mathcal{U}\left(z,h/2,\xi^{(1)},\xi^{(2)}\right)-\mathcal{U}\left(z,h/2,\tilde{\xi}^{(1)},\tilde{\xi}^{(2)}\right)\right\|_{a,b}^2\right)\\
&\leq 36\mathbb{E}\Bigg(\Bigg\|\left(\sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(\frac{h}{2},\xi^{(1)}\right)-\mathcal{Z}^{(2)}\left(\frac{h}{2},\xi^{(1)},\xi^{(2)}\right)\right)\right.\\
&-\sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(\tilde{\xi}^{(1)}\right)-\mathcal{Z}^{(2)}\left(\frac{h}{2},\xi^{(1)},\tilde{\xi}^{(2)}\right)\right),\\
&\left.\sqrt{2\gamma}\mathcal{Z}^{(2)}\left(\frac{h}{2},\xi^{(1)},\xi^{(2)}\right)-\sqrt{2\gamma}\mathcal{Z}^{(2)}\left(\frac{h}{2},\tilde{\xi}^{(1)},\tilde{\xi}^{(2)}\right)\right)\Bigg\|_{a,b}^2\Bigg) \leq 36\left(\frac{3}{\gamma}+\frac{3\gamma}{M}\right)dh,
\end{aligned}$$

using the same argument as for the previous term. Hence by summing up the above bounds, we have

$$\sigma(z)^2 \leq 37\left(\frac{3}{\gamma}+\frac{3\gamma}{M}\right)dh. \qquad \text{(B.27)}$$

Now, we will lower bound $n(z)$ as defined in (B.24). By (B.1), we have

$$\begin{aligned}
&\mathbb{E}(\|w(z)-\tilde{w}(z)\|_{a,b}^2) \geq \frac{1}{2}\mathbb{E}(\|w(z)-\tilde{w}(z)\|_{a,0}^2) \\
&= \mathbb{E}\Bigg(\Bigg\|\mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z,h/2,\xi^{(1)},\xi^{(2)}\right),h\right),h/2,\xi^{(3)},\xi^{(4)}\right) \qquad \text{(B.28)} \\
&- \mathbb{E}\mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z,h/2,\xi^{(1)},\xi^{(2)}\right),h\right),h/2,\xi^{(3)},\xi^{(4)}\right)\Bigg\|_{a,0}^2\Bigg) \\
&\geq \mathbb{E}\left(\left\|\left(\sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(\frac{h}{2},\xi^{(3)}\right)-\mathcal{Z}^{(2)}\left(\frac{h}{2},\xi^{(3)},\xi^{(4)}\right)\right),\sqrt{2\gamma}\mathcal{Z}^{(2)}\left(\frac{h}{2},\xi^{(3)},\xi^{(4)}\right)\right)\right\|_{a,0}^2\right) \\
&\geq \frac{\gamma}{M}dh. \qquad \text{(B.29)}
\end{aligned}$$

For the denominator, we have

$$\begin{aligned}
&\int\int (g(w)-g(\tilde{w}))^2 P_h(z,dw)P_h(z,d\tilde{w}) = 2\cdot \mathrm{Var}_{w\sim P_h(z,\cdot)}(g(w)) \\
&= 2\cdot \mathrm{Var}\left(g\left(\mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z,h/2,\xi^{(1)},\xi^{(2)}\right),h\right),h/2,\xi^{(3)},\xi^{(4)}\right)\right)\right)
\end{aligned}$$

by the Efron-Stein inequality [31, 150]

$$\begin{aligned}
&\leq 2\mathbb{E}(\mathrm{Var}_{\xi^{(1)},\xi^{(2)}}\left(g\left(\mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z,h/2,\xi^{(1)},\xi^{(2)}\right),h\right),h/2,\xi^{(3)},\xi^{(4)}\right)\right)\right) \\
&+ 2\mathbb{E}(\mathrm{Var}_{\xi^{(3)},\xi^{(4)}}\left(g\left(\mathcal{U}\left(\mathcal{B}\left(\mathcal{U}\left(z,h/2,\xi^{(1)},\xi^{(2)}\right),h\right),h/2,\xi^{(3)},\xi^{(4)}\right)\right)\right),
\end{aligned}$$

where $\mathrm{Var}_{\xi^{(1)},\xi^{(2)}}(\cdot)$ means that we compute the conditional variance with respect to $\xi^{(3)},\xi^{(4)}$ (so the $\xi^{(3)},\xi^{(4)}$ are kept constant, and only the variance with respect to $\xi^{(1)},\xi^{(2)}$ is considered). Let

$$\begin{aligned}
J_{\mathcal{U}}(h) &:= \frac{\partial \mathcal{U}(x,v,h,\xi^{(1)},\xi^{(2)})}{\partial(\xi^{(1)},\xi^{(2)})} \\
&= \begin{pmatrix} \left(\sqrt{\frac{2h}{\gamma}}-\frac{\sqrt{2}(1-e^{-\gamma h})}{\gamma^{3/2}\sqrt{h}}\right)I_d, & -\frac{(1-e^{-\gamma h})\sqrt{2}}{\sqrt{\gamma h}}I_d \\ -\frac{1}{\gamma}\sqrt{1-e^{-2\gamma h}-\frac{2(1-e^{-\gamma h})^2}{\gamma h}}I_d, & \sqrt{1-e^{-2\gamma h}-\frac{2(1-e^{-\gamma h})^2}{\gamma h}}I_d \end{pmatrix}, \\
\tilde{g}_h(z) &:= g\left(\mathcal{U}\left(\mathcal{B}\left(z,h\right),h/2,\xi^{(3)},\xi^{(4)}\right)\right).
\end{aligned}$$

Using the assumption that g in 1-Lipschitz in (B.24), and the bounds (B.25-B.26), it follows that $\tilde{g}_h$ is a 6-Lipschitz function in $\|\cdot\|_{a,b}$, and (B.1) implies that it is a 12-Lipschitz function in $\|\cdot\|_{a,0}$. Since the continuously differentiable Lipschitz functions are dense amongst Lipschitz functions (see [7]), we can assume without loss of generality that g and thus $\tilde{g}_h$ are continuously differentiable. Note that

$$\begin{aligned}
&\tilde{g}_h(z)-\tilde{g}_h(z') = \left\langle \nabla\tilde{g}_h(z), z-z'\right\rangle + o(\|z-z'\|_{a,0}) \\
&= \left\langle \begin{pmatrix} I_d & 0 \\ 0 & a^{-1/2}I_d \end{pmatrix}\nabla\tilde{g}_h(z), \begin{pmatrix} I_d & 0 \\ 0 & a^{1/2}I_d \end{pmatrix}(z-z')\right\rangle + o(\|z-z'\|_{a,0}).
\end{aligned}$$

Using this, it is easy to show that the 12-Lipschitz property of $\tilde{g}_h$ in $\|\cdot\|_{a,0}$ implies that $\|\nabla \tilde{g}_h(z)\|_{1/a,0} \leq 12$ for every $z \in \Lambda$. Hence, we obtain

$$\begin{aligned}
&\left\| \frac{\partial}{\partial(\xi^{(1)}, \xi^{(2)})} \left(g \left(\mathcal{U} \left(\mathcal{B} \left(\mathcal{U} \left(z, h/2, \xi^{(1)}, \xi^{(2)} \right), h \right), h/2, \xi^{(3)}, \xi^{(4)} \right) \right) \right) \right\| \\
&= \left\| \frac{\partial}{\partial(\xi^{(1)}, \xi^{(2)})} \tilde{g}_h \left(\mathcal{U} \left(z, h/2, \xi^{(1)}, \xi^{(2)} \right) \right) \right\| = \left\| J_U(h/2) \nabla \tilde{g}_h \left(\mathcal{U} \left(z, h/2, \xi^{(1)}, \xi^{(2)} \right) \right) \right\| \\
&\leq 12 \sup_{w \in \Lambda : \|w\|_{1/a,0} \leq 1} \|J_U(h/2) w\| = 12 \sup_{w \in \Lambda : \|w\| \leq 1} \left\| J_U(h/2) \begin{pmatrix} I_d & 0_d \\ 0_d & \frac{1}{\sqrt{M}} I_d \end{pmatrix} w \right\| \\
&= 12 \left\| J_U(h/2) \begin{pmatrix} I_d & 0_d \\ 0_d & \frac{1}{\sqrt{M}} I_d \end{pmatrix} \right\| \\
&= 12 \left\| \begin{pmatrix} \left(\sqrt{\frac{h}{\gamma}} - \frac{2(1-e^{-\gamma h/2})}{\gamma^{3/2}\sqrt{h}} \right), & -\frac{2(1-e^{-\gamma h/2})}{\sqrt{M\gamma h}} \\ -\frac{1}{\gamma} \sqrt{1 - e^{-\gamma h} - \frac{4(1-e^{-\gamma h/2})^2}{\gamma h}}, & \frac{1}{\sqrt{M}} \sqrt{1 - e^{-\gamma h} - \frac{4(1-e^{-\gamma h/2})^2}{\gamma h}} \end{pmatrix} \right\|,
\end{aligned}$$

using the fact that $-(1-e^{-x})^2 \leq -x^2 + x^3$ for $x \geq 0$, and that $\gamma h \leq 1$

$$\leq 12 \left(\sqrt{\frac{h}{\sqrt{M}}} + \frac{2\gamma h}{\sqrt{M}} \right).$$

From the Gaussian Poincaré inequality (see e.g. Theorem 3.20 of [31]), and the fact that $\xi^{(1)}, \xi^{(2)}$ are standard normal, it follows that

$$\begin{aligned}
&2\mathbb{E}(\mathrm{Var}_{\xi^{(1)}, \xi^{(2)}} \left(g \left(\mathcal{U} \left(\mathcal{B} \left(\mathcal{U} \left(z, h/2, \xi^{(1)}, \xi^{(2)} \right), h \right), h/2, \xi^{(3)}, \xi^{(4)} \right) \right) \right) \\
&\leq 2 \cdot 12^2 \left(\sqrt{\frac{h}{\sqrt{M}}} + \frac{2\gamma h}{\sqrt{M}} \right)^2 \leq 576 \left(\frac{h}{\sqrt{M}} + 4\frac{\gamma^2 h^2}{M} \right).
\end{aligned}$$

We can bound the second term similarly, since

$$\begin{aligned}
&\left\| \frac{\partial}{\partial(\xi^{(3)}, \xi^{(4)})} \left(g \left(\mathcal{U} \left(\mathcal{B} \left(\mathcal{U} \left(z, h/2, \xi^{(1)}, \xi^{(2)} \right), h \right), h/2, \xi^{(3)}, \xi^{(4)} \right) \right) \right) \right\| \\
&= \left\| J_U(h/2) \nabla g \left(\mathcal{U} \left(\mathcal{B} \left(\mathcal{U} \left(z, h/2, \xi^{(1)}, \xi^{(2)} \right), h \right), h/2, \xi^{(3)}, \xi^{(4)} \right) \right) \right\|
\end{aligned}$$

using the fact that g is 2-Lipschitz with respect to $\|\cdot\|_{a,0}$,

$$\leq 2 \left\| J_U(h/2) \begin{pmatrix} I_d & 0_d \\ 0_d & \frac{1}{\sqrt{M}} I_d \end{pmatrix} \right\| \leq 2\sqrt{\frac{h}{\sqrt{M}}} + \frac{2\gamma h}{\sqrt{M}},$$

and thus by the Gaussian Poincaré inequality,

$$\begin{aligned}
&2\mathbb{E}(\mathrm{Var}_{\xi^{(3)}, \xi^{(4)}} \left(g \left(\mathcal{U} \left(\mathcal{B} \left(\mathcal{U} \left(z, h/2, \xi^{(1)}, \xi^{(2)} \right), h \right), h/2, \xi^{(3)}, \xi^{(4)} \right) \right) \right) \\
&\leq 8 \left(\sqrt{\frac{h}{\sqrt{M}}} + \frac{2\gamma h}{\sqrt{M}} \right)^2 \leq 16 \left(\frac{h}{\sqrt{M}} + 4\frac{\gamma^2 h^2}{M} \right).
\end{aligned}$$

By adding these up, we obtain

$$\int\int (g(w)-g(\tilde{w}))^2 P_h(z,dw)P_h(z,d\tilde{w}) \leq 592\left(\frac{h}{\sqrt{M}}+4\frac{\gamma^2h^2}{M}\right),$$

and hence by (B.24) and (B.29), we have

$$n(z) \geq \frac{\frac{\gamma}{M}dh}{592\left(\frac{h}{\sqrt{M}}+4\frac{\gamma^2h^2}{M}\right)} \geq \frac{\frac{\gamma}{M}dh}{592\cdot 5\left(\frac{\gamma h}{M}\right)} \geq \frac{d}{3000}. \tag{B.30}$$

Combining this with (B.27), we have that

$$\sup_{z\in\Lambda}\frac{\sigma(z)^2}{n(z)} \leq \left(37\left(\frac{3}{\gamma}+\frac{3\gamma}{M}\right)dh\right)\cdot\frac{3000}{d} \leq 333000\left(\frac{1}{\gamma}+\frac{\gamma}{M}\right)h,$$

and the claim now follows by Theorem 2 of [89] and the bound on $\mathrm{Var}[\mathbb{E}(\hat{\pi}(f))|X_0)]$ on page 2427 of [89], using the fact that $\kappa \geq 1-\sqrt{1-\frac{mh}{4\gamma}} \geq \frac{mh}{8\gamma}$ by Proposition B.3.6. □

B.4.3 Variance of $S(c_R)$

Theorem 3.3.15. *Suppose that Assumptions 3.3.6, 3.3.7, 3.3.10, 3.3.12, 3.3.13 hold, and in addition,*

$$\gamma \geq \sqrt{8M}, \quad h_0 \leq \frac{1}{\gamma}\cdot\frac{m}{264M}, \quad B \geq \frac{16\log(4)\gamma}{mh_0}, \quad B_0 \geq \frac{16\gamma}{mh_0}\log\left(\frac{c_{\mu_0}+1}{\sqrt{M}\gamma h_0^2}\right).$$

Suppose that $c_R \in [0,1)$, and $2<\phi_N<16$. Then for any $N\geq 1$, the UBUBU estimator $S(c_R)$ has finite expected computational cost, $\mathbb{E}S(c_R)=\pi(f)$, and it has finite variance. Moreover, it satisfies a CLT as $N\to\infty$, and the asymptotic variance σ_S^2 defined in (3.3.18) *can be bounded as*

$$\sigma_S^2 \leq \frac{C(m,M,M_1^s,\gamma,c_N,\phi_N)}{Kh_0}\left(1+\frac{1}{h_0K}+dh_0^4\right).$$

Proof of Theorem 3.3.15. By Corollary B.4.6, and the fact that

$$\mathbb{E}(D_{l,l+1}^2) \leq \max_{1\leq k\leq K}\mathbb{E}\left[\left(f(z_k'^{(l,l+1)})-f(z_k^{(l,l+1)})\right)^2\right],$$

it follows that under the assumptions of Corollary B.4.6, we have

$$\mathbb{E}(D_{l,l+1}^2) \leq Cd\left(\left(\gamma^4+\frac{(M_1^s)^2}{M}\right)\left(\frac{\gamma}{m}\right)^2+\frac{M\gamma^2}{m}\right)h_l^4 \leq V_D\phi_D^{-l},$$

for $V_D = Ch_0^4d\left(\left(\gamma^4+\frac{(M_1^s)^2}{M}\right)\left(\frac{\gamma}{m}\right)^2+\frac{M\gamma^2}{m}\right)$ and $\phi_D=16$. From Proposition B.4.9, and using the fact that $c(h_0)=\frac{h_0m}{8\gamma}$, and our assumptions on B_0, we have

$$\begin{aligned}\mathrm{Var}(D_0) &\leq \frac{C}{c(h_0)K}\left(1+\frac{1}{c(h_0)K}\right)\left(\frac{1}{\gamma}+\frac{\gamma}{M}\right)\frac{h_0}{c(h_0)}+\frac{(1-c(h_0))^{2(B_0+1)}}{2c(h_0)^2K^2}\sigma_{\mu_0}^2 \\ &\leq \frac{C}{K}\cdot\frac{1}{h_0}\left(\frac{8\gamma}{m}\right)^2\left(\frac{1}{\gamma}+\frac{\gamma}{M}\right)\left(1+\frac{8\gamma}{h_0m}\cdot\frac{1}{K}\right).\end{aligned} \tag{B.31}$$

The computational cost at each level satisfies the assumptions of Proposition 3.3.4, so if we fix $2 < \phi_N < 16$, all assumptions of this proposition are satisfied. Hence $S(c_R)$ is an unbiased estimator with finite variance and computational cost.

The claim about the asymptotic variance follows by using the bounds in (B.31) and in Proposition B.4.8, and adding up all terms according to (3.3.18). □

Proposition B.4.10. *Suppose that the assumptions of Proposition B.3.6 hold for $h = h_l$. Let $R_{l,l+1} = (P_{h_l,h_{l+1}})^{2^l}$ be the Markov kernel defined in Section 3.3.1 for two synchronously coupled UBU chains at discretization levels $l, l+1$. This chain is moving on state space Λ^2. Let $\overline{z}_1, \dots, \overline{z}_K$ be a Markov chain with kernel $R_{l,l+1}$. Let $F : \Lambda^2 \to \mathbb{R}$ be of the form $F(z, z') = f(z) - f(z')$, where f is of the form* (3.3.28)*. Suppose that the target π is a product distribution, satisfying the same conditions as in Proposition 3.3.18. Then we have*

$$\mathrm{Var}\left(\frac{\sum_{i=1}^K F(\overline{z}_i)}{K}\right) \leq \frac{2}{K^2}\sum_{i=1}^{K}\sum_{k=0}^{K-i}\min\left(\frac{\mathrm{Var}(F(\overline{z}_i)) + \mathrm{Var}(F(\overline{z}_{i+k}))}{2},\right.$$
$$\left.\sqrt{4r\left(\sum_{s=1}^{r}\|w^{(s)}\|^2\right)}\sqrt{\mathrm{Var}(F(\overline{z}_i))\max_{1\leq j\leq d}\mathbb{E}\left[\|\overline{z}_{i,j} - \mathbb{E}\overline{z}_{i,j}\|_{a,b}^2\right]}\cdot\exp\left(-\frac{mh_0}{8\gamma}\cdot k\right)\right).$$

Proof. We proceed similarly to the proof of Proposition B.4.7.

$$\mathrm{Cov}(F(\overline{z}_i), F(\overline{z}_{i+k})) = \mathbb{E}[(F(\overline{z}_i) - \mathbb{E}(F(\overline{z}_i)))(F(\overline{z}_{i+k}) - \mathbb{E}(F(\overline{z}_{i+k})))].$$

Let $\tilde{z}_i$ be an independent identically distributed copy of $\overline{z}_i$, and define $(\overline{z}_{i:i+k}, \tilde{z}_{i:i+k})$ as synchronously coupled, in the same way as in the proof of Proposition B.4.7. It follows from applying Proposition B.3.6 on each coordinate, and using independence that for every coordinate $1 \leq j \leq d$,

$$\mathbb{E}\left(\|\tilde{z}_{i+k,j} - \overline{z}_{i+k,j}\|_{a,b}^2 \middle| \overline{z}_{i,j}, \tilde{z}_{i,j}\right) \leq \exp\left(-\frac{mh_0}{4\gamma}\cdot k\right)\|\overline{z}_{i,j} - \tilde{z}_{i,j}\|_{a,b}^2.$$

With a slight abuse of notation, index j here refers to both position and velocity components, hence $\tilde{z}_{i,j} = (\tilde{x}_{i,j}, \tilde{v}_{i,j}, \tilde{x}'_{i,j}, \tilde{v}'_{i,j}) \in \mathbb{R}^4$. As previously, $\overline{z}_i$ and $\tilde{z}_{i+k}$ are independent, and

$$\mathrm{Cov}(F(\overline{z}_i), F(\overline{z}_{i+k})) = \mathbb{E}[(F(\overline{z}_i) - \mathbb{E}(F(\overline{z}_i)))\mathbb{E}(F(\overline{z}_{i+k}) - F(\tilde{z}_{i+k})|\overline{z}_i, \tilde{z}_i)]$$
$$\leq \sqrt{\mathrm{Var}\left(F(\overline{z}_i)\right)\mathrm{Var}\left(F(\overline{z}_{i+k}) - F(\tilde{z}_{i+k})\right)}$$

By the Efron-Stein inequality [31, 150], and some rearrangement, we have

$$\mathrm{Var}\left(F(\overline{z}_{i+k}) - F(\tilde{z}_{i+k})\right) \leq 2\sum_{j=1}^{d}\left(\sum_{s=1}^{r}|w_j^{(s)}|\right)^2\mathbb{E}\left(\|\tilde{z}_{i+k,j} - \overline{z}_{i+k,j}\|_{a,b}^2\right)$$
$$\leq 2r\left(\sum_{s=1}^{r}\|w^{(s)}\|^2\right)\exp\left(-\frac{mh_0}{4\gamma}\cdot k\right)\max_{1\leq j\leq d}\mathbb{E}\left[\|\overline{z}_{i,j} - \tilde{z}_{i,j}\|_{a,b}^2\right]$$
$$= 4r\left(\sum_{s=1}^{r}\|w^{(s)}\|^2\right)\exp\left(-\frac{mh_0}{4\gamma}\cdot k\right)\max_{1\leq j\leq d}\mathbb{E}\left[\|\overline{z}_{i,j} - \mathbb{E}\overline{z}_{i,j}\|_{a,b}^2\right],$$

and the claim follows by rearrangement and summation. □

Proposition 3.3.18. *Suppose that Assumption 3.3.17 holds, and denote the potential U as $U(x) = \sum_{i=1}^{d} U_i(x_i)$. Suppose that Assumptions 3.3.6, 3.3.7, and 3.3.10 hold for each component $(U_i)_{1\le i\le d}$, and that*

$$\gamma \ge \sqrt{8M}, \quad h_0 \le \frac{1}{\gamma}\cdot\frac{m}{264M}, \quad B \ge \frac{16\log(4)\gamma}{mh_0}, \quad B_0 \ge \frac{16\gamma}{mh_0}\log\left(\frac{c_{\mu_0}+1}{\sqrt{M}\gamma h_0^2}\right).$$

Suppose that f is of the form

$$f(x,v) = g(\langle w^{(1)}, x\rangle, \dots \langle w^{(r)}, x\rangle), \tag{3.3.28}$$

where $g : \mathbb{R}^r \to \mathbb{R}$ is 1-Lipschitz, and $w^{(1)}, \dots, w^{(r)} \in \mathbb{R}^d$. Suppose that $c_R \in [0,1)$ and $2 < \phi_N < 16$. Then for any $N \ge 1$, the UBUBU estimator $S(c_R)$ has finite expected computational cost, $\mathbb{E}S(c_R) = \pi(f)$, and it has finite variance. Moreover, it satisfies a CLT as $N \to \infty$, and the asymptotic variance can be bounded as

$$\sigma_S^2 \le \frac{C(m, M, M_1, \gamma, r, c_N, \phi_N)}{Kh_0}\sum_{1\le i\le r}\|w^{(i)}\|^2.$$

Proof of Proposition 3.3.18. Unbiasedness, finite variance, and finite computational cost follow from Theorem 3.3.15. By (3.3.18), the asymptotic variance can be expressed as

$$\sigma_S^2 := \mathrm{Var}(D_0) + \sum_{l=0}^{\infty}\mathrm{Var}(D_{l,l+1})\cdot\frac{\phi_N^l}{c}.$$

It is easy to show that f is $\sum_{s=1}^{r}\|w^{(s)}\|$-Lipschitz, so the variance term $\mathrm{Var}(D_0)$ can be bounded using Proposition B.4.9, relying on the burn-in assumptions.

To control $\mathrm{Var}(D_{l,l+1})$, we first need to bound terms of the form $\mathrm{Var}(f(z_k') - f(z_k))$. Let $z_{k,j} = (x_{k,j}, v_{k,j}) \in \mathbb{R}^2$ denote components j in both x and v. Using the Efron-Stein inequality [31, 150], and independence of the components, we have

$$\begin{aligned}\mathrm{Var}(f(z_k^{\prime(l,l+1)}) - f(z_k^{(l,l+1)})) &\le 2\sum_{j=1}^{d}\left(\sum_{s=1}^{r}|w_j^{(s)}|\right)^2\mathbb{E}\left(\left\|z_{k,j}^{\prime(l,l+1)} - z_{k,j}^{(l,l+1)}\right\|_{a,b}^2\right)\\ &\le 2r\left(\sum_{s=1}^{r}\|w^{(s)}\|^2\right)\max_{1\le j\le d}\mathbb{E}\left(\left\|z_{k,j}^{\prime(l,l+1)} - z_{k,j}^{(l,l+1)}\right\|_{a,b}^2\right).\end{aligned}$$

By applying Corollary B.4.6 component-wise, it follows that under our assumptions,

$$\max_{1\le j\le d}\mathbb{E}\left(\left\|z_{k,j}^{\prime(l,l+1)} - z_{k,j}\right\|_{a,b}^2\right) \le C(m, M, \gamma, M_1)h_l^4,$$

hence

$$\mathrm{Var}(f(z_k^{\prime(l,l+1)}) - f(z_k^{(l,l+1)})) \le C(m, M, M_1, \gamma, r)\sum_{1\le i\le r}\|w^{(i)}\|^2 h_l^4.$$

Using this, and Proposition B.4.10, by a similar argument as in the proof of Theorem 3.3.15, we can show that

$$\mathrm{Var}(D_{l,l+1}) \leq \frac{C(m, M, M_1, \gamma, r)}{K} \left(\sum_{1 \leq i \leq r} \|w^{(i)}\|^2 \right) h_l^4 \left(1 + \frac{4\gamma}{m h_0} + \log(4) l \right),$$

and the claim follows by summation and rearrangement. □

B.5 Initialization and Gaussian approximation

We will use the following assumptions in this section and sections B.6 and B.7, which we restate here for easier readability. We will consider potentials of the form

$$U(x) = U_0(x) + \sum_{i=1}^{N_D} U_i(x), \tag{B.32}$$

where we aim to understand the scaling of the computational complexity when inexact gradients are used within the UBUBU framework in the large N_D case. We assume that the potential has the form (B.32) in this section and sections B.6 and B.7, and we impose the following assumptions on the potential.

Assumption 3.3.21 (∇Lipschitz property)**.** *For every $1 \leq i \leq N_D$, $U_i : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable and there exists a $\tilde{M} > 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla U_i(x) - \nabla U_i(y)\| \leq \tilde{M} \|x - y\|,$$

for every $1 \leq i \leq N_D$ and moreover,

$$\|\nabla U(x) - \nabla U(y)\| \leq M \|x - y\| \quad \textit{for} \quad M = N_D \tilde{M}.$$

Assumption 3.3.22 ($N_D \tilde{m}$-strong convexity)**.** *There exists a $\tilde{m} > 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m \|x - y\|^2 \quad \textit{for} \quad m = N_D \tilde{m}.$$

Assumption 3.3.23 (strongly Hessian Lipschitz property)**.** *$U : \mathbb{R}^d \to \mathbb{R}$ is three times continuously differentiable and M_1^s-strongly Hessian Lipschitz if there exists $M_1^s > 0$ such that*

$$\|\nabla^3 U(x)\|_{\{1,2\}\{3\}} \leq M_1^s \quad \textit{for} \quad M_1^s = N_D \tilde{M}_1^s,$$

for all $x \in \mathbb{R}^d$.

For a better understanding of the scaling in terms of N_D, we also introduce

$$\tilde{\gamma} = \frac{\gamma}{\sqrt{N_D}}, \tag{B.33}$$

so that $\gamma = \sqrt{N_D} \tilde{\gamma}$.

B.5.1 OHO scheme

In this section, we detail some results for the OHO scheme we use for initialization in (3.3.34)-(3.3.36). We state results for a potential that satisfies Assumptions 3.3.6 and 3.3.7 that can be applied in the case of Gaussian approximation. In particular, we show strong error results using similar techniques to [105] and [142].

We define the solution map $\mathcal{H}$ to have update rule

$$\mathcal{H} : (x, v) \to \phi_h(x, v), \tag{B.34}$$

where $\phi_h(x, v)$ is the solution to the ODE

$$dX_t = V_t dt, \qquad dV_t = -\nabla U(X_t)dt,$$

initialized at $(X_0, V_0) := (x, v) \in \mathbb{R}^{2d}$ at time $h > 0$. We then define the OHO scheme with stepsize $h > 0$ as a half step of $\mathcal{O}$ with stepsize $h/2$ (defined in (3.3.35)), followed by a full step of $\mathcal{H}$ with stepsize h and a half step of $\mathcal{O}$ with stepsize $h/2$, which exactly preserves the invariant measure.

Remark B.5.1. *We remark that the OHO scheme is a special case of the scheme studied in [117] using a hypocoercivity approach. It has also been considered as an exact splitting for discretization analysis in [25, 28, 80, 115]. In practice this scheme is only applicable when the Hamiltonian dynamics can be solved exactly, for example for a Gaussian target.*

Proposition B.5.2. *Let $h < 1/2\gamma$, $\gamma^2 \geq 4M$, $k \in \mathbb{N}$ and $(X_t, V_t)_{t\geq 0} := (Z_t)_{t\geq 0}$ be the solution of* (3.1.1) *and $(x_t, v_t)_{t\geq 0} := (z_t)_{t\geq 0}$ be the solution to the $\mathcal{OHO}$ scheme with stepsize $h > 0$, with synchronously coupled Brownian motion and where both are initialized at $z_0 = Z_0 \sim \pi$. We have that*

$$\|Z_{kh} - z_{kh}\|_{L^2,a,b} \leq \sqrt{\frac{3}{2}} e^{\frac{3}{2}hk\sqrt{M}} \left(\frac{3h\gamma\sqrt{k\gamma hd} + 5k(h\gamma)^2\sqrt{d}}{\sqrt{M}} \right).$$

Proof. Considering the OHO scheme given by

$$(x_h, v_h) := \Bigg(x + h\left(\eta v + \sqrt{1-\eta^2}\xi_1\right) - \int_0^h \nabla U(x(t))(h-t)dt,$$
$$\eta\left(\eta v + \sqrt{1-\eta^2}\xi_1 - h\nabla U(x) - \int_0^h \nabla^2 U(x(t))\overline{v}(t)(h-t)dt\right) + \sqrt{1-\eta^2}\xi_2 \Bigg),$$

where the Hamiltonian dynamics $(x(t), v(t))_{t=0}^h$ is initialized at $(x, \eta v + \sqrt{1-\eta^2}\xi_1)$. The kinetic Langevin dynamics for one step can be written as

$$V_h = \mathcal{E}(h)V_0 - \int_0^h \mathcal{E}(h-s)\nabla U(X_s)ds + \sqrt{2\gamma}\int_0^h \mathcal{E}(h-s)dW_s, \tag{B.35}$$

$$X_h = X_0 + \mathcal{F}(h)V_0 - \int_0^h \mathcal{F}(h-s)\nabla U(X_s)ds + \sqrt{2\gamma}\int_0^h \mathcal{F}(h-s)dW_s, \tag{B.36}$$

where $\mathcal{E}(h) = e^{-\gamma h}, \mathcal{F}(h) = \frac{1-e^{-\gamma h}}{\gamma}$, and we couple the noises such that $\sqrt{1-\eta^2}\xi_1 = \sqrt{2\gamma}\int_0^{h/2}\mathcal{E}(h/2-s)dW_s$ and $\sqrt{1-\eta^2}\xi_2 = \sqrt{2\gamma}\int_{h/2}^{h}\mathcal{E}(h-s)dW_s$. Considering the velocity component we have

$$
\begin{aligned}
\|V_h - v_h\|_{L^2} &\leq \left\| \eta^2 (V_0 - v_0) - \int_0^h \mathcal{E}(h-s)\left(\nabla U(X_s) - \nabla U(x)\right) ds + \right. \\
&+ \left. \sqrt{2\gamma}(1-\eta)\int_0^{h/2} \mathcal{E}(h/2-s) dW_s \right\|_{L^2} + \sqrt{Md}\left|\frac{1-\eta^2-h\gamma\eta}{\gamma}\right| + \frac{h^2 M\sqrt{d}}{2} \\
&\leq \left\| \eta^2 (V_0 - v_0) - \int_0^h \mathcal{E}(h-s)\left(\nabla U(X_0) - \nabla U(x)\right) ds + \sqrt{2\gamma}(1-\eta)\int_0^{h/2} \mathcal{E}(h/2-s) dW_s \right\|_{L^2} \\
&+ \left\| \int_0^h \mathcal{E}(h-s)\left(\nabla U(X_s) - \nabla U(X_0)\right) ds \right\|_{L^2} + \left(\gamma\sqrt{M} + \frac{M}{2}\right) h^2\sqrt{d} \\
&\leq \left\| \eta^2 (V_0 - v_0) - \int_0^h \mathcal{E}(h-s)\left(\nabla U(X_0) - \nabla U(x)\right) ds + \sqrt{2\gamma}(1-\eta)\int_0^{h/2} \mathcal{E}(h/2-s) dW_s \right\|_{L^2} \\
&+ \left(\gamma\sqrt{M} + 3M\right) h^2\sqrt{d},
\end{aligned}
$$

where the final estimate is a rough estimate due to (B.35), $(X_s, V_s) \sim \pi$ for all $s \in [0,h]$, the fact that U is M-∇Lipschitz, $h < \frac{1}{2\gamma}$ and $\gamma^2 \geq 4M$. Now considering time $kh \geq 0$ for $k \in \mathbb{N}$ and iteratively applying the argument whilst keeping Brownian components in the same L^2 norm, we have

$$
\begin{aligned}
\|V_{kh} - v_{kh}\|_{L^2} &\leq \sum_{i=1}^{k} hMa_i + \sqrt{2\gamma}\left\| \sum_{i=1}^{k} \eta^{2(k-i)}(1-\eta)\int_{(i-1)h}^{(i-1/2)h} \mathcal{E}((i-1/2)h - s) dW_s \right\|_{L^2} \\
&+ 3k\gamma\sqrt{M}h^2\sqrt{d} \\
&\leq \sum_{i=1}^{k} hMa_i + \frac{h\gamma\sqrt{2\gamma}}{2}\left\| \sum_{i=1}^{k} \int_{(i-1)h}^{(i-1/2)h} \mathcal{E}((i-1/2)h - s) dW_s \right\|_{L^2} \\
&+ k\left(h\gamma\right)^2\sqrt{d} + 3k\gamma\sqrt{M}h^2\sqrt{d} \\
&\leq \sum_{i=1}^{k} hMa_i + h\gamma\sqrt{k\gamma h d} + 3k\left(h\gamma\right)^2\sqrt{d},
\end{aligned}
$$

where $a_i := \|X_{ih} - x_{ih}\|_{L^2}$ for $i \in \mathbb{N}$ and $b_i := \|V_{ih} - v_{ih}\|_{L^2}$ for $i \in \mathbb{N}$. We have also used the independence of the Brownian motion over independent time intervals.

Now considering the position components we have

$$\begin{aligned}\|X_h - x_h\|_{L^2} &\leq \Bigg\| X_0 - x_0 + \mathcal{F}(h)\,(V_0 - v_0) - \int_0^h \mathcal{F}(h-s)\,(\nabla U(X_s) - \nabla U(x_s))\,ds \\ &\quad + \int_0^h \sqrt{2\gamma}\mathcal{F}(h-s)dW_s - h\sqrt{2\gamma}\int_0^{h/2}\mathcal{E}(h/2-s)dW_s \Bigg\|_{L^2} + 2\gamma h^2\sqrt{d} \\ &\leq \Bigg\| X_0 - x_0 + \mathcal{F}(h)(V_0 - v_0) - \int_0^h \mathcal{F}(h-s)\,(\nabla U(X_0) - \nabla U(x_0))\,ds \\ &\quad + \int_0^h \sqrt{2\gamma}\mathcal{F}(h-s)dW_s - h\sqrt{2\gamma}\int_0^{h/2}\mathcal{E}(h/2-s)dW_s \Bigg\|_{L^2} + 3\gamma h^2\sqrt{d}.\end{aligned}$$

Then as before we consider time $kh \geq 0$ for $k \in \mathbb{N}$ and we have

$$\begin{aligned}&\|X_{kh} - x_{kh}\|_{L^2} \leq \sum_{i=1}^k \left(hb_i + h^2 M a_i\right) + 3k\gamma h^2\sqrt{d} \\ &+ \sqrt{2\gamma}\left\|\sum_{i=1}^k \int_{(i-1)h}^{ih} \mathcal{F}(ih-s)dW_s - h\int_{(i-1)h}^{(i-1/2)h}\mathcal{E}((i-1/2)h-s)dW_s\right\|_{L^2} \\ &\leq \sum_{i=1}^k \left(hb_i + h^2 M a_i\right) + 3k\gamma h^2\sqrt{d} + 2\sqrt{2\gamma}h\sqrt{hkd}.\end{aligned}$$

In $\|\cdot\|_{L^2,a,b}$ using the preceding estimates we have

$$\begin{aligned}\|Z_{kh} - z_{kh}\|_{L^2,a,0} &\leq \frac{3}{2}\sum_{i=1}^k h\sqrt{M}\|Z_{ih} - z_{ih}\|_{L^2,a,0} + \frac{3h\gamma\sqrt{k\gamma hd} + 5k(h\gamma)^2\sqrt{d}}{\sqrt{M}} \\ &\leq e^{\frac{3}{2}hk\sqrt{M}}\left(\frac{3h\gamma\sqrt{k\gamma hd} + 5k(h\gamma)^2\sqrt{d}}{\sqrt{M}}\right).\end{aligned}$$

□

Theorem B.5.3. *Let $h < 1/2\gamma$, $\gamma^2 \geq 4M$, $l \in \mathbb{N}$ and $(X_t, V_t)_{t\geq 0} := (Z_t)_{t\geq 0}$ be the solution of kinetic Langevin dynamics and $(x_l, v_l)_{l\in\mathbb{N}} := (z_l)_{l\in\mathbb{N}}$ be the iterates of the solution to the $\mathcal{OHO}$ scheme with stepsize $h > 0$, where both are initialized at the same point according to the invariant measure. We have that*

$$\|Z_{lh} - z_l\|_{L^2,a,b} \leq 104h\gamma^2\sqrt{d}\left(\frac{3\sqrt{2\gamma/\sqrt{M}} + 10\gamma/\sqrt{M}}{m}\right).$$

Proof. We use an approach used in [105] to remove the exponential constant in Proposition B.5.2. We define a sequence of interpolating variants $Z_l^{(k)}$ for every $k = 0, ..., l$ as follows. We first define $Z_0^{(k)} = Z_0$, and then $(Z_i^{(k)})_{i=1}^k$ are defined by $\mathcal{OHO}$ steps followed by $(Z_i^{(k)})_{i=k+1}^l$ steps of kinetic Langevin dynamics with stepsize $h > 0$. We break the l steps into blocks of size $\tilde{l} = \lceil \frac{2}{3h\sqrt{M}} \rceil$, then

we have

$$\begin{aligned}\|Z_{lh} - z_l\|_{L^2,a,b} &= \|Z_l^{(0)} - Z_l^{(l)}\|_{L^2,a,b} \\ &\leq \sum_{j=0}^{\lfloor l/\tilde{l}\rfloor - 1} \left\|\left(Z_l^{(j\tilde{l})} - Z_l^{((j+1)\tilde{l})}\right)\right\|_{L^2,a,b} + \left\|\left(Z^{(\lfloor l/\tilde{l}\rfloor l)} - Z^{(l)}\right)\right\|_{L^2,a,b},\end{aligned}$$

where we bound the terms using the fact that the first $j\tilde{l}$ steps according to OHO keep the stationary distribution invariant and they only deviate in the following $\tilde{l}$ steps, where we will use Proposition B.5.2 with l chosen as $\tilde{l}$. We finally use contraction of the continuous dynamics (Corollary B.3.8) in the remaining steps and we have

$$\left\|\left(Z_l^{(j\tilde{l})} - Z_l^{((j+1)\tilde{l})}\right)\right\|_{L^2,a,b} \leq 4e^{-(l-1-(j+1)\tilde{l})c(h)} \left(\frac{3h\gamma\sqrt{\tilde{l}\gamma hd} + 5\tilde{l}(h\gamma)^2\sqrt{d}}{\sqrt{M}}\right).$$

Then summing up the terms we have that

$$\begin{aligned}\|Z_{lh} - z_l\|_{L^2,a,b} &\leq 4\left(\frac{3h\gamma\sqrt{\tilde{l}\gamma hd} + 5\tilde{l}(h\gamma)^2\sqrt{d}}{\sqrt{M}}\right)\left(1 + \frac{1}{1 - e^{-c(h)\tilde{l}}}\right) \\ &\leq 8\left(\frac{3h\gamma\sqrt{\tilde{l}\gamma hd} + 5\tilde{l}(h\gamma)^2\sqrt{d}}{\sqrt{M}}\right)\left(1 + \frac{12\gamma\sqrt{M}}{m}\right) \\ &\leq 104h\gamma^2\sqrt{d}\left(\frac{3\sqrt{2\gamma/\sqrt{M}} + 10\gamma/\sqrt{M}}{m}\right),\end{aligned}$$

as required. □

B.5.2 Initialization and bounds

For convex potentials, we can approximate the gradient with the Hessian at the minimizer by

$$\mathcal{Q}(x \mid x^*) = \nabla U(x^*) + H^*(x - x^*) = H^*(x - x^*), \tag{B.37}$$

where $x^* \in \mathbb{R}^d$ is the minimizer of U and $H^* = \nabla^2 U(x^*)$.

Lemma B.5.4. *Considering the gradient approximation $\mathcal{Q}$ given by* (B.37)*, where the potential U satisfies Assumption 3.3.23, and has a minimizer $x^* \in \mathbb{R}^d$ we then have the property*

$$\mathbb{E}\,\|\nabla U(x) - \mathcal{Q}(x \mid x^*)\|^p \leq (\tilde{M}_1^s)^p N_D^p \|x - x^*\|_{L^{2p}}^{2p},$$

for any $x \in \mathbb{R}^d$.

Proof. Follows from Taylor expansion. □

We then define the measure $\mu_G = \mathcal{N}(x^*, (H^*)^{-1}) \times \mathcal{N}(0_d, I_d)$ to be the Gaussian approximation of the target as in (3.3.33), which is the invariant measure of the OHO scheme and continuous kinetic Langevin dynamics with the use of the gradient approximation (B.37).

Proposition B.5.5. *Let $p = 2$ or 4, then we have the following Wasserstein bound between a potential U which satisfies Assumptions 3.3.22,3.3.23 and 3.3.28*

$$\mathcal{W}_{p,a,b}(\pi, \mu_G) \leq \sqrt{\frac{3}{2}} \left(\frac{(2p)!}{2^p p!} \right)^{1/p} \frac{\tilde{M}_1 d}{\tilde{m}^2 N_D}.$$

Proof. If we let π_x denote the marginal in the position of π and $(\mu_G)_x$ denote the marginal in position of μ_G. We have from the equivalence of norms that for $p = 2, 4$

$$\begin{aligned}
\mathcal{W}_{p,a,b}(\pi, \mu_G) &\leq \sqrt{\frac{3}{2}} \mathcal{W}_{p,a,0}(\pi, \mu_G) \leq \sqrt{\frac{3}{2}} \mathcal{W}_p(\pi_x, (\mu_G)_x) \\
&\leq \sqrt{\frac{3}{2}} \frac{\|\nabla U - \mathcal{Q}(\cdot \mid x^*)\|_{L^p}}{m} \leq \sqrt{\frac{3}{2}} \frac{M_1^s \|x - x^*\|_{L^{2p}}^2}{m} \\
&\leq \sqrt{\frac{3}{2}} \left(\frac{(2p)!}{2^p p!} \right)^{1/p} \frac{M_1^s d}{m^2}
\end{aligned}$$

where the third inequality is due to Proposition 22 of [161], the fourth due to Lemma B.5.4 and the final inequality by Lemma B.8.2. □

B.6 Variance bounds for UBU with SVRG

For this section, we make use of the technique of the recent work of Hu et al [168], related to using stochastic variance reduced gradient (SVRG).

A stochastic gradient version of the UBU scheme is simply constructed by replacing the $\mathcal{B}$ operator with

$$\mathcal{B}_{\mathcal{G}}(x, v, h, \omega | \hat{x}) = (x, v - h\mathcal{G}(x, \omega | \hat{x})),$$

where $\mathcal{G}$ is a stochastic gradient approximation of the potential U as defined in the approximation given by (3.3.30).

We start with an alternative formula for the kinetic Langevin dynamics introduced in (3.1.1). This is used for the analysis of the UBU scheme in the full gradient setting in [140] and alternative schemes with the SVRG approximation in (3.3.30). The convenient way of expressing kinetic Langevin dynamics is to use Itô's formula on the product $e^{\gamma t} V_t$. This results in the following set of equations for (3.1.1) with initial condition $(X_0, V_0) \in \mathbb{R}^{2d}$:

$$V_t = \mathcal{E}(t) V_0 - \int_0^t \mathcal{E}(t-s) \nabla U(X_s) ds + \sqrt{2\gamma} \int_0^t \mathcal{E}(t-s) dW_s, \tag{B.38}$$

$$X_t = X_0 + \mathcal{F}(t) V_0 - \int_0^t \mathcal{F}(t-s) \nabla U(X_s) ds + \sqrt{2\gamma} \int_0^t \mathcal{F}(t-s) dW_s, \tag{B.39}$$

where

$$\mathcal{E}(t) = e^{-\gamma t} \qquad \mathcal{F}(t) = \frac{1 - e^{-\gamma t}}{\gamma}. \tag{B.40}$$

Then the UBU scheme (as in [140]) can be expressed as

$$v_{k+1} = \mathcal{E}(h)v_k - h\mathcal{E}(h/2)\nabla U(\overline{x}_k) + \sqrt{2\gamma}\int_{kh}^{(k+1)h} \mathcal{E}((k+1)h - s)dW_s, \tag{B.41}$$

$$\overline{x}_k = x_k + \mathcal{F}(h/2)v_k + \sqrt{2\gamma}\int_{kh}^{(k+1/2)h} \mathcal{F}((k+1/2)h - s)dW_s, \tag{B.42}$$

$$x_{k+1} = x_k + \mathcal{F}(h)v_k - h\mathcal{F}(h/2)\nabla U(\overline{x}_k) + \sqrt{2\gamma}\int_{kh}^{(k+1)h} \mathcal{F}((k+1)h - s)dW_s, \tag{B.43}$$

which can be more easily compared to the true dynamics via (B.38) and (B.39). We will refer to $(\overline{x}_k)_{k\in\mathbb{N}}$ as the gradient evalution points of the scheme. Similarly, stochastic gradient UBU can be expressed as (B.41)-(B.43) by replacing the gradients with stochastic gradient approximations,

$$v_{k+1} = \mathcal{E}(h)v_k - h\mathcal{E}(h/2)\mathcal{G}(\overline{x}_k, \omega_{k+1}|\hat{x}_k) + \sqrt{2\gamma}\int_{kh}^{(k+1)h} \mathcal{E}((k+1)h - s)dW_s, \tag{B.44}$$

$$\overline{x}_k = x_k + \mathcal{F}(h/2)v_k + \sqrt{2\gamma}\int_{kh}^{(k+1/2)h} \mathcal{F}((k+1/2)h - s)dW_s, \tag{B.45}$$

$$\hat{x}_k = \overline{x}_{L(k)} \quad \text{for} \quad L(k) = \tau\lfloor k/\tau \rfloor, \tag{B.46}$$

$$x_{k+1} = x_k + \mathcal{F}(h)v_k - h\mathcal{F}(h/2)\mathcal{G}(\overline{x}_k, \omega_{k+1}|\hat{x}_k) + \sqrt{2\gamma}\int_{kh}^{(k+1)h} \mathcal{F}((k+1)h - s)dW_s, \tag{B.47}$$

If we are using a stochastic gradient approximation of the UBU dynamics, additional bias is introduce by the use of gradient approximations. We wish to measure the local error caused by the stochastic gradient approximation.

B.6.1 Variance bound of $D_{l,l+1}$

Suppose now we have two UBU schemes, a UBU scheme $(z_k)_{k\in\mathbb{N}} = (x_k, v_k)_{k\in\mathbb{N}}$ which uses a stochastic gradient approximation as defined in Definition 3.2.1 with $(\omega_k)_{k\in\mathbb{N}}$ such that $\omega_k \sim \mathcal{SWR}(N_D, N_b)$ for each $k \in \mathbb{N}$. Further at iteration k define $z_k^h := (x_k^h, v_k^h) := \psi_h\left(z_k, h, (W_{t'})_{t'=kh}^{(k+1)h}\right)$ to be a step of the full gradient UBU scheme at iteration z_k, with synchronously coupled Brownian motion. Then the local error after one step is

$$\mathbb{E}\|(x_{k+1} - x_k^h, v_{k+1} - v_k^h)\|^2 = h^2\left(\mathcal{E}(h/2) + \mathcal{F}(h/2)\right)^2 \mathbb{E}\|\nabla U(\overline{x}_k) - \mathcal{G}(\overline{x}_k, \omega_{k+1})\|^2,$$

and

$$\mathbb{E}\|x_{k+1} - x_k^h\|^2 \leq \frac{h^4}{4}\mathbb{E}\|\nabla U(\overline{x}_k) - \mathcal{G}(\overline{x}_k, \omega_{k+1})\|^2,$$

where expectations are taken over stochastic gradient approximation and Brownian increments. The sequence $(\overline{x}_k)_{k\in\mathbb{N}}$ are the points where the stochastic gradient approximations are evaluated defined by (B.45). We now wish to bound the term $\mathbb{E}\|\nabla U(\overline{x}_k) - \mathcal{G}(\overline{x}_k, \omega_{k+1})\|^2$, uniformly in k to control the error due to the stochastic gradient. For this, we state Lemma 1 of [168] with our notations, together with its proof.

Lemma B.6.1. *Considering iterates $(x_k, v_k, \overline{x}_k)_{k\in\mathbb{N}}$ of stochastic gradient UBU with the SVRG $(\mathcal{G}, \mathcal{SWR}(N_D, N_b))$ for a potential U which has the form* (3.3.29)*, with data size N_D and batch size N_b, epoch length $\tau = \lceil N_D/N_b \rceil$, and initial condition $(x_0, v_0) \in \mathbb{R}^{2d}$, then we have the property*

$$\begin{aligned}&\mathbb{E}\left(\left\|\mathcal{G}(\overline{x}_k, \omega_{k+1} \mid \overline{x}_{L(k)}) - \nabla U(\overline{x}_k)\right\|^2\right)\\ &\leq \frac{N_D(N_D - N_b)(\tau-1)^2}{N_b(N_D-1)}.\max_{j<k}\sum_{i=1}^{N_D}\mathbb{E}\left(\|\nabla U_i(\overline{x}_{j+1}) - \nabla U_i(\overline{x}_j)\|^2\right).\end{aligned}$$

Corollary B.6.2. *Suppose that Assumption 3.3.21 holds. For UBU with SVRG updates as defined by* (B.44)-(B.47)*, we have*

$$\mathbb{E}\left(\left\|\mathcal{G}(\overline{x}_k, \omega_{k+1} \mid \overline{x}_{L(k)}) - \nabla U(\overline{x}_k)\right\|^2\right) \leq \Theta \max_{j<k}\mathbb{E}\|\overline{x}_{j+1} - \overline{x}_j\|^2, \tag{B.48}$$

$$\Theta = \frac{\tilde{M}^2 N_D^2 (N_D - N_b)(\tau-1)^2}{N_b(N_D-1)}. \tag{B.49}$$

Proof of Lemma B.6.1. For the potential of the form $U(x) = U_0(x) + \sum_{i=1}^{N_D} U_i(x)$ and for $k \geq 1$ we define $\overline{X}_i = \nabla U_i(\overline{x}_k) - \nabla U_i(\overline{x}_{L(k)})$ and we define $Y_i = N_D\overline{X}_i - \sum_{j=1}^{N_D}\overline{X}_j$ for $i = 1, ..., N_D$. Then we have that $\sum_{i=1}^{N_D} Y_i = 0$ and that

$$\mathcal{G}(\overline{x}_k, \omega_{k+1} \mid \overline{x}_{L(k)}) - \nabla U(\overline{x}_k) = \frac{1}{N_b}\sum_{i\in\omega_{k+1}} Y_i.$$

Therefore our aim is to establish a bound on $\frac{1}{N_b}\sum_{i\in\omega_{k+1}} Y_i$. We have that

$$\begin{aligned}\mathbb{E}_{\omega_{k+1}}\left\|\frac{1}{N_b}\sum_{i\in\omega_{k+1}} Y_i\right\|^2 &= \frac{1}{N_b^2}\mathbb{E}_{\omega_{k+1}}\left(\sum_{i\in\omega_{k+1}}\|Y_i\|^2 + \sum_{i\neq j\in\omega_{k+1}}\langle Y_i, Y_j\rangle\right)\\ &= \frac{1}{N_b N_D}\sum_{i=1}^{N_D}\|Y_i\|^2 + \frac{b-1}{N_b N_D(N_D-1)}\sum_{i\neq j}\langle Y_i, Y_j\rangle\\ &= \frac{N_D - N_b}{N_D - 1}\frac{1}{N_b N_D}\sum_{i=1}^{N_D}\|Y_i\|^2,\end{aligned}$$

where the last line is due to the fact that $\sum_{i=1}^{N_D} Y_i = 0$. Then using the fact that $\sum_{i=1}^{N_D} \|Y_i\|^2 \leq N_D^2 \sum_{i=1}^{N_D} \|\overline{X}_i\|^2$ and the last full gradient evaluation is at $k - \tau + 1 \leq L(k) \leq k$ we have that

$$\begin{aligned}
&\mathbb{E}\left(\|\nabla U(\overline{x}_k) - \mathcal{G}(\overline{x}_k, \omega_{k+1} \mid \overline{x}_{L(k)})\|^2\right) = \frac{N_D - N_b}{N_b N_D (N_D - 1)} \sum_{i=1}^{N_D} \mathbb{E}\|Y_i\|^2 \\
&\leq \frac{N_D(N_D - N_b)}{N_b(N_D - 1)} \sum_{i=1}^{N_D} \mathbb{E}\|\overline{X}_i\|^2 \\
&\leq \frac{N_D(N_D - N_b)}{N_b(N_D - 1)} \sum_{i=1}^{N_D} \mathbb{E}\|\nabla U_i(\overline{x}_k) - \nabla U_i(\overline{x}_{L(k)})\|^2 \\
&\leq \frac{N_D(N_D - N_b)(k - L(k))}{N_b(N_D - 1)} \sum_{j=L(k)}^{k-1} \sum_{i=1}^{N_D} \mathbb{E}\|\nabla U_i(\overline{x}_{j+1}) - \nabla U_i(\overline{x}_j)\|^2 \\
&\leq \frac{N_D(N_D - N_b)(\tau - 1)^2}{N_b(N_D - 1)} \max_{j<k} \sum_{i=1}^{N_D} \mathbb{E}\|\nabla U_i(\overline{x}_{j+1}) - \nabla U_i(\overline{x}_j)\|^2,
\end{aligned}$$

which concludes the proof. □

Hence it is sufficient to bound $\mathbb{E}\|\overline{x}_{k+1} - \overline{x}_k\|^2$ uniformly in $k \in \mathbb{N}$, which will be done in the following lemma.

Lemma B.6.3 (Displacement Lemma)**.** *Let a stochastic gradient UBU integrator defined by* (B.44)-(B.45) *with stochastic gradient* $(\mathcal{G}, \rho)$ *satisfy*

$$\mathbb{E}\left(\|\mathcal{G}(\overline{x}_k, \omega_{k+1} | \hat{x}_k) - \nabla U(\overline{x}_k)\|^2\right) \leq \Theta \max_{j<k} \mathbb{E}\|\overline{x}_{j+1} - \overline{x}_j\|^2,$$

for some $\Theta > 0$*. If* U *satisfies Assumptions 3.3.21, 3.3.22,* $h < 1/2\gamma$ *and* $\gamma^2 \geq 8M$*, then*

$$\|\overline{x}_{k+1} - \overline{x}_k\|_{L^2} \leq h^2\sqrt{\Theta} \max_{0 \leq i < k} \|\overline{x}_{i+1} - \overline{x}_i\|_{L^2} + 7h\sqrt{M}\|z_{k-1} - Z^{k-1}\|_{L^2,a,b} + 6h\sqrt{d},$$

where $Z^k := Z_{kh} = \phi(Z_0, kh, (W_{t'})_{t'=0}^{kh}) \in \mathbb{R}^{2d}$ *is the solution to* (3.1.1) *initialized at the invariant measure* $Z_0 \sim \pi$ *at time* kh *for* $k \in \mathbb{N}$*.*

Proof. Then we use the following estimate

$$\begin{aligned}
\|\overline{x}_{k+1} - \overline{x}_k\|_{L^2} &= \|\overline{x}_{k+1} - x_{k+1} + x_{k+1} - x_k + x_k - \overline{x}_k\|_{L^2} \\
&\leq \|\mathcal{F}(h/2)(v_k - v_{k-1})\|_{L^2} + \|x_{k+1} - x_k\|_{L^2} \\
&+ \sqrt{2\gamma} \left\| \int_{(k+1)h}^{(k+3/2)h} \mathcal{F}((k+3/2)h - s) dW_s \right\|_{L^2} \\
&+ \sqrt{2\gamma} \left\| \int_{kh}^{(k+1/2)h} \mathcal{F}((k+1/2)h - s) dW_s \right\|_{L^2} \\
&=: \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)},
\end{aligned}$$

and we bound (I), (II), (III) and (IV) separately. (III) and (IV) can be bounded above by $\sqrt{\gamma h^3 d}$. Firstly, we will bound (II), but first we denote

$$A_j = \|\overline{x}_{j+1} - \overline{x}_j\|_{L^2}^2,$$

for $j \in \mathbb{N}$, and $z_k^h = (x_k^h, v_k^h) := \psi_h(z_k, h, (W_{t'})_{t'=kh}^{(k+1)h})$ is an iterate with stepsize h and initial point (x_k, v_k) of the full gradient UBU scheme and synchronously coupled Brownian motion to the stochastic gradient scheme. We then estimate

$$\begin{aligned}\|x_{k+1} - x_k\|_{L^2} &\leq \|x_{k+1} - x_k^h\|_{L^2} + \|x_k^h - x_k\|_{L^2} \\ &\leq \frac{h^2}{2}\sqrt{\Theta}\max_{j<k}\sqrt{A_j} + \|x_k^h - x_k\|_{L^2},\end{aligned}$$

then if we define the notation $Z_k^t = (X_k^t, V_k^t) := \phi(z_k, t, (W_{t'})_{t'=kh}^{kh+t}) \in \mathbb{R}^{2d}$ for $k \in \mathbb{N}$ and $t \geq 0$ to be the continuous dynamics solution with initial condition (x_k, v_k) at time t defined by (B.38) and (B.39). Then we can estimate the second term by splitting it up into discretization error and one-step displacement and bounding each of these terms separately as

$$\|x_k^h - x_k\|_{L^2} \leq \|x_k^h - X_k^h\|_{L^2} + \|X_k^h - x_k\|_{L^2}.$$

Then using [140, Section 7.6] we have that

$$\begin{aligned}&\|x_k^h - X_k^h\|_{L^2} \leq \\ &\left\|\int_0^h \mathcal{F}(h/2)\left(\nabla U(X_k^s) - \nabla U(\overline{x}_k)\right)ds + \int_0^h \left(\mathcal{F}(h-s) - \mathcal{F}(h/2)\right)\nabla U(X_k^s)ds\right\|_{L^2} \\ &\leq \frac{h}{2}\int_0^h \|\nabla U(X_k^s) - \nabla U(\overline{x}_k)\|_{L^2}\,ds + h^2 \max_{0\leq s\leq h}\|\nabla U(X_k^s)\|_{L^2} \\ &\leq \frac{hM}{2}\int_0^h \left\|X_k^s - \left(x_k + \mathcal{F}\left(h/2\right)v_k + \sqrt{2\gamma}\int_{kh}^{(k+1/2)h}\mathcal{F}\left((k+1/2)h - s\right)dW_s\right)\right\|_{L^2} ds \\ &+ h^2 \max_{0\leq s\leq h}\|\nabla U(X_k^s)\|_{L^2} \\ &\leq h^2 \max_{0\leq s\leq h}\|\nabla U(X_k^s)\|_{L^2} + \frac{hM}{2}\int_0^h \|X_k^s - x_k\|_{L^2}ds + \frac{h^3 M}{4}\max_{0\leq s\leq h}\|V_k^s\|_{L^2} + \frac{h^{7/2}M\sqrt{\gamma d}}{4}.\end{aligned}$$

Now, we bound

$$\int_0^h \|X_k^s - x_k\|_{L^2}ds \leq h^2\|v_k\|_{L^2} + h^3 \max_{0\leq s\leq h}\|\nabla U(X_k^s)\|_{L^2} + h^{5/2}\sqrt{2\gamma d},$$

and using the fact that $h < \min\{\frac{1}{5\sqrt{M}}, \frac{1}{2\gamma}\}$ we have

$$\|x_k^h - X_k^h\|_{L^2} \leq \frac{3h^2}{2}\max_{0\leq s\leq h}\|\nabla U(X_k^s)\|_{L^2} + h\max_{0\leq s\leq h}\|V_k^s\|_{L^2} + h\sqrt{d},$$

and using (3.1.1) we have that

$$\begin{aligned}\|X_k^h - x_k\|_{L^2} &= \left\|\mathcal{F}(h)v_k - \int_0^h \mathcal{F}(h)\nabla U(X_k^s)ds + \sqrt{2\gamma}\int_0^h \mathcal{F}(h-s)dW_s\right\|_{L^2} \\ &\leq h\|v_k\|_{L^2} + h^2 \max_{0\leq s\leq h}\|\nabla U(X_k^s)\|_{L^2} + \sqrt{2\gamma h^3 d}.\end{aligned}$$

To bound the maximum terms we introduce $(Z_t)_{t\geq 0} = (X_t, V_t)_{t\geq 0}$ to be the solution to (3.1.1) initialized at the invariant measure with synchronously coupled Brownian motion. We also define $Z^k := Z_{kh}$ for $k \in \mathbb{N}$. Then we have, in expectation, for any $0 \leq s \leq h$,

$$\begin{aligned}\|V_k^s\|_{L^2} &\leq \|V_k^s - V_{kh+s}\|_{L^2} + \|V_{kh+s}\|_{L^2} \\ &\leq \sqrt{2M}\|z_k - Z^k\|_{L^2,a,b} + \sqrt{d},\end{aligned}$$

and for any $0 \leq s \leq h$ we have

$$\begin{aligned}\|\nabla U(X_k^s)\|_{L^2} &\leq \|\nabla U(X_k^s) - \nabla U(X_{kh+s})\|_{L^2} + \|\nabla U(X_{kh+s})\|_{L^2} \\ &\leq M\|X_k^s - X_{kh+s}\|_{L^2} + \sqrt{Md} \\ &\leq \sqrt{2}M\|z_k - Z^k\|_{L^2,a,b} + \sqrt{Md},\end{aligned}$$

where we have used contraction of the continuous dynamics under synchronous coupling provided in Corollary B.3.8 and [51, Lemma 2] to bound $\|\nabla U(X_{kh+s})\|_{L^2}$. Therefore we have the following bound on (II)

$$\text{(II)} \leq \frac{h^2}{2}\sqrt{\Theta}\max_{0\leq i<k}\sqrt{A_i} + 4h\sqrt{M}\|z_k - Z^k\|_{L^2,a,b} + \frac{9h\sqrt{d}}{2},$$

where $h < \frac{1}{5\sqrt{M}}$ due to the fact that $\gamma^2 \geq 8M$ and $h < \frac{1}{2\gamma}$.

Next, we consider (I) and we can estimate

$$\begin{aligned}\text{(I)} &\leq \frac{h}{2}\|v_k - v_{k-1}\|_{L^2} \leq \frac{h}{2}\|v_k - v_{k-1}^h\|_{L^2} + \frac{h}{2}\|v_{k-1}^h - v_{k-1}\|_{L^2} \\ &\leq \frac{h^2}{2}\sqrt{\Theta}\max_{0\leq i<k}\sqrt{A_i} + \frac{h}{2}\|v_{k-1}^h - v_{k-1}\|_{L^2},\end{aligned}$$

where

$$\|v_{k-1}^h - v_{k-1}\|_{L^2} \leq \|v_{k-1}^h - V_{k-1}^h\|_{L^2} + \|V_{k-1}^h - v_{k-1}\|_{L^2}.$$

Then we can bound

$$\begin{aligned}
&\|v_{k-1}^h - V_{k-1}^h\|_{L^2} \leq \\
&\left\| \int_0^h \mathcal{E}(h/2) \left(\nabla U(X_{k-1}^s) ds - \nabla U(\overline{x}_{k-1}) \right) + \left(\mathcal{E}(h-s) - \mathcal{E}(h/2) \right) \nabla U(X_{k-1}^s) ds \right\|_{L^2} \\
&\leq M \int_0^h \left\| X_{k-1}^s - \overline{x}_{k-1} \right\|_{L^2} ds + h \max_{0 \leq s \leq h} \|\nabla U(X_{k-1}^s)\|_{L^2} \\
&\leq \frac{3}{50} \max_{0 \leq s \leq h} \|V_{k-1}^s\|_{L^2} + 2\frac{51}{50} h \max_{0 \leq s \leq h} \|\nabla U(X_{k-1}^s)\|_{L^2} + \frac{2}{25}\sqrt{d},
\end{aligned}$$

where we have used the estimate of $\int_0^h \|X_{k-1}^s - \overline{x}_{k-1}\|_{L^2}$ from the $\|x_{k-1}^h - X_{k-1}^h\|_{L^2}$ bound and the fact that $h < 1/5\sqrt{M}$. Using (B.38) we have

$$\begin{aligned}
\|V_{k-1}^h - v_{k-1}\|_{L^2} &\leq \left\| (\mathcal{E}(h) - 1) v_{k-1} - \int_0^h \mathcal{E}(h-s) \nabla U(X_{k-1}^s) ds + \sqrt{2\gamma} \int_0^h \mathcal{E}(h-s) dW_s \right\|_{L^2} \\
&\leq h\gamma \|v_{k-1}\|_{L^2} + h \max_{0 \leq s \leq h} \|\nabla U(X_{k-1}^s)\|_{L^2} + \sqrt{2\gamma h d} \\
&\leq 2h\sqrt{M} \left(\gamma + \sqrt{M} \right) \|z_{k-1} - Z^{k-1}\|_{L^2,a,b} + h\sqrt{d} \left(\gamma + \sqrt{M} \right) + \sqrt{2\gamma h d},
\end{aligned}$$

and we can combine terms to get the following bound on (I)

$$\text{(I)} \leq \frac{h^2}{2} \max_{0 \leq i < k} \sqrt{\Theta A_i} + 3h\sqrt{M} \|z_{k-1} - Z^{k-1}\|_{L^2,a,b} + \frac{5}{4} h\sqrt{d},$$

and summing all terms we have that

$$\|\overline{x}_{k+1} - \overline{x}_k\|_{L^2} \leq h^2 \max_{0 \leq i < k} \sqrt{\Theta A_i} + 7h\sqrt{M} \|z_{k-1} - Z^{k-1}\|_{L^2,a,b} + 6h\sqrt{d}.$$

□

Proposition B.6.4. *For a stochastic gradient UBU integrator with iterates $(z_k)_{k \in \mathbb{N}}$, gradient evaluation points $(\overline{x}_k)_{k \in \mathbb{N}}$, transition kernel P_h and potential U satisfying Assumptions 3.3.21-3.3.22, where we approximate the gradient using a unbiased stochastic gradient $(\mathcal{G}, \rho)$ satisfying*

$$\mathbb{E}\left(\|\mathcal{G}(\overline{x}_k, \omega_{k+1} | \hat{x}_k) - \nabla U(\overline{x}_k)\|^2 \right) \leq \Theta \max_{j<k} \mathbb{E} \|\overline{x}_{j+1} - \overline{x}_j\|^2.$$

Consider the continuous solution to (3.1.1) *initialized at the invariant measure, for $k \in \mathbb{N}$ define $Z^k := Z_{kh} = \phi(Z_0, kh, (W_{t'})_{t'=0}^{kh}) \in \mathbb{R}^{2d}$ with synchronously coupled Brownian motion to $(z_k)_{k \in \mathbb{N}}$, then for all*

$$h < \min \left\{ \frac{1}{2\tilde{\gamma} N_D^{1/2}}, 1/2, \frac{\tilde{m}^{1/3} N_D^{1/6}}{24(\Theta\tilde{\gamma})^{1/3}}, \frac{1}{4\Theta^{1/4}}, \frac{\tilde{m}}{256 \tilde{M} \tilde{\gamma} N_D^{1/2}} \right\},$$

we have

$$\begin{aligned}
&\|z_k - Z^k\|_{L^2,a,b} \leq 4(1 - R_2(h)/2)^k \left(\|z_0 - Z^0\|_{L^2,a,b} + \sqrt{2} C_1 h^{5/2} \right) \\
&+ C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s) h^{3/2} d^{1/2} N_D^{-3/4} \Theta^{1/2},
\end{aligned}$$

where $R_2(h) = 1 - c_2(h) + C_0^2 h^2$.

Further for all $\mu \in \mathcal{P}_2(\mathbb{R}^{2d})$, *and all* $k \in \mathbb{N}$,

$$\mathcal{W}_{2,a,b}(\mu P_h^k, \pi) \leq 4(1 - R_2(h)/2)^k \left(\mathcal{W}_{2,a,b}(\mu, \pi) + \sqrt{2} C_1 h^{5/2}\right) + C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s) h^{3/2} d^{1/2} N_D^{-3/4} \Theta^{1/2}.$$

Proof. Let us define the notation $Z_k^t = (X_k^t, V_k^t) := \phi(z_k, t, (W_{t'})_{t'=kh}^{kh+t}) \in \mathbb{R}^{2d}$ for $k \in \mathbb{N}$ and $t \geq 0$ to be the continuous dynamics solution with initial condition (x_k, v_k) at time t defined by (B.38) and (B.39). Further define $z_k^h = (x_k^h, v_k^h) := \psi_h(z_k, h, (W_{t'})_{t'=kh}^{(k+1)h})$ is an iterate with stepsize h and initial point (x_k, v_k) of the full gradient UBU scheme and synchronously coupled Brownian motion to the stochastic gradient scheme.

Firstly, we split up the difference in the following way

$$\begin{aligned}\|z_k - Z^k\|_{L^2,a,b}^2 &= \left\| \left(z_k - z_{k-1}^h\right) + \left(z_{k-1}^h - Z^k\right) \right\|_{L^2,a,b}^2 \\ &= \left\| z_k - z_{k-1}^h \right\|_{L^2,a,b}^2 + 2 \left\langle z_k - z_{k-1}^h, z_{k-1}^h - Z^k \right\rangle_{L^2,a,b} + \|z_{k-1}^h - Z^k\|_{L^2,a,b}^2.\end{aligned}$$

Considering the inner product we have the expectation conditional on z_{k-1} and $(W_{t'})_{t'=(k-1)h}^{kh}$ is zero as it is independent of the Brownian motion (due to synchronous coupling) and the stochastic gradient estimator is unbiased. Therefore

$$\begin{aligned}\|z_k - Z^k\|_{L^2,a,b}^2 &\leq \|z_k - z_{k-1}^h\|_{L^2,a,b}^2 + (\|\beta_{k-1}\|_{L^2,a,b} \\ &\quad + \|z_{k-1}^h - \psi(Z^{k-1}, h, (W_{t'})_{t'=(k-1)h}^{kh}) + \alpha_{k-1}\|_{L^2,a,b})^2 \\ &= (\mathrm{I})' + (\mathrm{II})'.\end{aligned}$$

We have that

$$(\mathrm{I})' \leq \frac{2h^2\Theta}{M} \max_{j<k} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^2}^2,$$

and

$$(\mathrm{II})' \leq \left(\sqrt{(1 - c_2(h) + C_0^2 h^2)\|z_{k-1} - Z^{k-1}\|_{L^2,a,b}^2 + 2C_1^2 h^5} + C_2 h^3\right)^2.$$

Let $R_2(h) := c_2(h) - C_0^2 h^2$, then assuming that $mh/8\gamma < R_2(h) < 1/2$ (which holds for $h < \frac{1}{2\gamma}$ and $h < \frac{m}{256\gamma M}$), using Lemma B.8.1,

$$\begin{aligned}\|z_k - Z^k\|_{L^2,a,b} \leq \sqrt{2}(1 - R_2(h)/2)^k \left(\|z_0 - Z^0\|_{L^2,a,b} + \sqrt{2C_1^2 h^5}\right) + \frac{2\sqrt{2} C_2 h^3}{R_2(h)} \\ + 2\sqrt{\frac{\frac{2h^2\Theta}{M} \max_{j<k} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^2}^2 + 2C_1^2 h^5}{R_2(h)}},\end{aligned}$$

and now we wish to bound $\max_{j<k} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^2}^2$. Considering Lemma B.6.3 we have that

$$\begin{aligned}\|\overline{x}_{k+1} - \overline{x}_k\|_{L^2} &\leq h^2\sqrt{\Theta \max_{0\leq j\leq k} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^2}^2} + 6h\sqrt{d} \\ &+ 7h\sqrt{M}\left(\sqrt{2}(1 - R_2(h)/2)^k\left(\|z_0 - Z^0\|_{L^2} + \sqrt{2C_1^2h^5}\right) + \frac{16\sqrt{2}\gamma C_2h^2}{m}\right) \\ &+ 56h\sqrt{\frac{h\gamma\Theta \max_{j<k} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^2}^2 + \gamma C_1^2h^4}{m}}.\end{aligned}$$

If we assume that

$$h < \min\left\{\frac{m^{1/3}}{24(\Theta\gamma)^{1/3}}, \frac{1}{4\Theta^{1/4}}, 1\right\},$$

then

$$\begin{aligned}\max_{j<k} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^2} \leq 21h\sqrt{M}&\left(\sqrt{2}(1 - R_2(h)/2)^k\left(\|z_0 - Z^0\|_{L^2} + \sqrt{2C_1^2h^5}\right) + \frac{16\sqrt{2}\gamma C_2h^2}{m}\right) \\ &+ 18h\sqrt{d} + 168h^3\sqrt{M}C_1\sqrt{\frac{\gamma}{m}},\end{aligned}$$

and

$$\begin{aligned}\|z_k - Z^k\|_{L^2,a,b} &\leq \sqrt{2}(1 - R_2(h)/2)^k\left(\|z_0 - Z^0\|_{L^2,a,b} + \sqrt{2}C_1h^{5/2}\right) \\ &+ \frac{16\sqrt{2\gamma}h^2\left(C_2\sqrt{\gamma} + C_1\sqrt{m}\right)}{m} + 2\sqrt{\frac{16h\gamma\Theta \max_{j<k} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^2}^2}{mM}} \\ &\leq 4(1 - R_2(h)/2)^k\left(\|z_0 - Z^0\|_{L^2,a,b} + \sqrt{2}C_1h^{5/2}\right) + \frac{24\sqrt{2\gamma}h^2\left(C_2\sqrt{\gamma} + C_1\sqrt{m}\right)}{m} \\ &+ 144h^{3/2}\left(\frac{d\gamma\Theta}{mM}\right)^{1/2} + 1344h^{7/2}C_1\frac{\gamma\sqrt{\Theta}}{m},\end{aligned}$$

and the first claim follow by rewriting this bound in terms of $\tilde{m}$, $\tilde{M}$ and $\tilde{\gamma}$. For non-asymptotic Wasserstein results, we simply replace Z^{k-1} with the continuous dynamics initialized at $\tilde{Z}_{k-1} \sim \pi$ be such that $\|\tilde{Z}_{k-1} - z_{k-1}\|_{L^2,a,b} = \mathcal{W}_{2,a,b}(\mu P_h^{k-1}, \pi)$ as in [140, Theorem 23]. We can then apply Lemma B.8.1 to get the required result. □

Remark B.6.5. *To get the non-asymptotic result to have discretization error which is of order $\mathcal{O}(h^{3/2})$, the gradient approximation needs to be an unbiased estimator of the gradient, without this property the discretization error reduces to order $\mathcal{O}(h)$.*

Lemma B.6.6. *Suppose we have two Langevin diffusions governed by* (3.1.1) *with synchronously coupled Brownian motion, $(Z_t)_{t\geq 0}$ with potential U satisfying Assumptions 3.3.28-3.3.23 and $(\tilde{Z}_t)_{t\geq 0}$ with potential defined by* (B.37)*, a Gaussian approximation of U. We further assume that the diffusions are initialized at their invariant measures and $\gamma \geq \sqrt{8M}$. We then have that*

$$\|Z_t - \tilde{Z}_t\|_{L^2,a,b} \leq e^{-\frac{mt}{16\gamma}}\|Z_0 - \tilde{Z}_0\|_{L^2,a,b} + \frac{dC(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s)}{N_D}.$$

Proof. We define $(X_t, V_t)_{t\geq 0} := (Z_t)_{t\geq 0}$ to be the diffusion according to (B.38)-(B.39) with potential U and define $(\tilde{X}_t, \tilde{V}_t)_{t\geq 0} := (\tilde{Z}_t)_{t\geq 0}$ to be the diffusion according to (B.38)-(B.39) with potential $\tilde{U}$, defined by (B.37), and synchronously coupled Brownian motion. By the same argument as Corollary B.3.8 with the expectations rather than Wasserstein distance we have that for $h > 0$,

$$\begin{aligned}
&\|Z_{(k+1)h} - \tilde{Z}_{(k+1)h}\|_{L^2,a,b} \leq \left(1 - \frac{mh}{16\gamma}\right) \|Z_{kh} - \tilde{Z}_{kh}\|_{L^2,a,b} \\
&+ h \left\| \left(0, \nabla\tilde{U}(\tilde{X}_{kh}) - \nabla U(\tilde{X}_{kh})\right) \right\|_{L^2,a,b} + C(\gamma, M, d)h^2 \\
&\leq \left(1 - \frac{mh}{16\gamma}\right) \|Z_{kh} - \tilde{Z}_{kh}\|_{L^2,a,b} + \frac{hM_1^s}{\sqrt{M}} \left\|\tilde{X}_{kh} - x^*\right\|_{L^4}^2 + C(\gamma, M, d)h^2,
\end{aligned}$$

where the last inequality is due to Lemma B.5.4 and x^* is the minimizer of U and $\tilde{U}$. Then due to Proposition B.3.13 taking the limit as $h \to 0$,

$$\|\tilde{X}_{kh} - x^*\|_{L^4}^2 \leq \frac{2}{m} \left[\sqrt{4\left(1 - \frac{h\lambda\gamma}{2}\right)^k (\gamma^4 \|\tilde{X}_0 - x^*\|_{L^4}^4 + \|\tilde{V}_0\|_{L^4}^4) + \frac{(6d + 160(1+\lambda^2))^2}{2\lambda^2}} \right]$$

using Lemma B.8.2,

$$\leq \frac{2}{m} \left[\sqrt{\frac{12\gamma^4 d^2}{m^2} + 12d^2 + \frac{(6d + 160(1+\lambda^2))^2}{2\lambda^2}} \right],$$

where $\lambda = \min\left(\frac{1}{4}, \frac{m}{\gamma^2}\right)$ is defined as in (B.12). We choose $k = t/h$ and define

$$c_u := \frac{2M_1^s}{m\sqrt{M}} \left[\sqrt{\frac{12\gamma^4 d^2}{m} + 12d^2 + \frac{(6d + 160(1+\lambda^2))^2}{2\lambda^2}} \right],$$

then we have

$$\limsup_{h\to 0} \frac{\|Z_{t+h} - \tilde{Z}_{t+h}\|_{L^2,a,b} - \|Z_t - \tilde{Z}_t\|_{L^2,a,b}}{h} \leq -\frac{m}{16\gamma} \|Z_t - \tilde{Z}_t\|_{L^2,a,b} + c_u.$$

All terms are bounded on the right-hand side due to the assumptions on the initial condition, therefore due to the Denjoy–Young–Saks theorem we have the upper Dini derivative (upper right-hand derivative) is finite. Hence considering $u : \mathbb{R} \to \mathbb{R}$ to be solution to the ODE

$$\frac{d}{dt}u(t) = -\frac{m}{16\gamma}u(t) + c_u,$$

with initial condition $u(0) = \|Z_0 - \tilde{Z}_0\|_{L^2,a,b}$ which we can solve exactly. Therefore by the comparison principle for ODEs and Dini derivatives [93][Lemma 3.4] we have

$$\begin{aligned}
\|Z_t - \tilde{Z}_t\|_{L^2,a,b} \leq u(t) &\leq e^{-\frac{mt}{16\gamma}} \|Z_0 - \tilde{Z}_0\|_{L^2,a,b} + \frac{16\gamma}{m} c_u \\
&\leq e^{-\frac{mt}{16\gamma}} \|Z_0 - \tilde{Z}_0\|_{L^2,a,b} + \frac{dC(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s)}{N_D},
\end{aligned}$$

as required. □

Proposition B.6.7. *Suppose two stochastic gradient UBU chains at coarser and finer discretization levels l and $l+1$, with synchronously coupled Brownian motions $(z_k)_{k\in\mathbb{N}}$ and $(z'_k)_{k\in\mathbb{N}}$ and stepsizes h_l and $h_{l+1} = h_l/2$, satisfying the conditions of Proposition B.6.4, be such that $z_0 \sim \pi_0$ and $z'_0 \sim \pi'_0$. Then for f satisfying Assumption 3.3.12 we have the following variance bound*

$$\begin{aligned}
&\mathrm{Var}\left(f(z'_k) - f(z_k)\right) \leq \mathbb{E}\left(f(z'_k) - f(z_k)\right)^2 \leq \\
&\left(\exp\left(-\frac{mkh_l}{8\gamma}\right)\left(\|z'_0 - z_0\|_{L^2,a,b} + \mathcal{W}_{2,a,b}(\pi_0,\pi) + \mathcal{W}_{2,a,b}(\pi'_0,\pi)\right)\right. \\
&+ 4(1 - R(h_l)/2)^k\left(\mathcal{W}_{2,a,b}(\pi_0,\pi) + \sqrt{2}C_1 h_l^{5/2}\right) \\
&+ 4(1 - R(h_{l+1})/2)^{2k}\left(\mathcal{W}_{2,a,b}(\pi'_0,\pi) + \sqrt{2}C_1 h_{l+1}^{5/2}\right) \\
&\left.+ C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)d^{1/2}N_D^{-3/4}\Theta^{1/2}h_l^{3/2}\right)^2.
\end{aligned}$$

Proof. By following the same argument as Proposition B.4.3 using Proposition B.6.4 we have the desired result. □

Proposition B.6.8. *Suppose that the assumptions of Proposition B.6.4 hold for the potential U, $h_0 > 0$ and $\gamma > 0$ and the SVRG stochastic gradient approximation. Assume that*

$$h_0 \leq \frac{C(\tilde{\gamma},\tilde{m},\tilde{M},\tau/N_D,N_b)}{N_D^{3/2}}, \quad B \geq \frac{16\log(2^{3/2})\tilde{\gamma}}{\tilde{m}h_0N_D^{1/2}}, \quad B_0 \geq \frac{16\tilde{\gamma}}{\tilde{m}h_0N_D^{1/2}}\log\left(\frac{1}{N_D^{9/4}h_0^{3/2}}\right),$$

and the levels are initialized as described in Section B.5. Then for every $l \geq 1$, $1 \leq k \leq K$, for a test function f which satisfies Assumption 3.3.12 the UBUBU samples satisfy

$$\begin{aligned}
\mathrm{Var}\left(f(z_k'^{(l,l+1)}) - f(z_k^{(l,l+1)})\right) &\leq \mathbb{E}\left[\left(f(z_k'^{(l,l+1)}) - f(z_k^{(l,l+1)})\right)^2\right] \\
&\leq \mathbb{E}\|z_k'^{(l,l+1)} - z_k^{(l,l+1)}\|_{a,b}^2 \\
&\leq C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)N_D^{5/2}h_l^3 d,
\end{aligned}$$

and further

$$\mathrm{Var}(D_{l,l+1}) \leq C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)N_D^{5/2}h_l^3 d. \tag{B.50}$$

Proof of Proposition B.6.8. Following a similar proof as Corollary B.4.6, we need to be careful with bounding the distance between z_0 and z'_0. This is the reason for our construction of the initial conditions in Section B.5 using the OHO scheme. In particular, we wish to have at most $\mathcal{O}(1/N_D)$ distance in initialization. We define $(Z_t)_{t\geq 0}$ to be defined by (3.1.1) with the potential U and $Z_0 \sim \pi$ such that $\|Z_0 - z'_0\|_{L^2,a,b} = \mathcal{W}_{2,a,b}(\pi,\mu_G)$ are optimally coupled. We define $(Z_t^{\mathcal{G}})_{t\geq 0}$ to be defined by (3.1.1) with the potential being a Gaussian approximation of the potential such that $Z_0^{\mathcal{G}} = z'_0 \sim \mu_G$, $(z_t^{\mathcal{G}})_{t\geq 0}$ to be the OHO scheme with the potential being a Gaussian approximation of the potential

such that $z_0^{\mathcal{G}} = z_0' \sim \mu_G$. We therefore have $z_B^{\mathcal{G}} = z_0$ and

$$\begin{aligned}\|z_0 - z'_{B/h_{l+1}}\|_{L^2,a,b} &\leq \|z_B^{\mathcal{G}} - Z_B^{\mathcal{G}}\|_{L^2,a,b} + \|Z_B^{\mathcal{G}} - Z_B\|_{L^2,a,b} + \|Z_B - z'_{B/h_{l+1}}\|_{L^2,a,b} \\ &\leq h\sqrt{d}C(\tilde{\gamma},\tilde{m},\tilde{M}) + \frac{dC(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)}{N_D} \\ &+ 5\|Z_0 - z_0'\|_{L^2,a,b} + C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)N_D^{5/2}h_l^3 d \\ &\leq \frac{dC(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)}{N_D},\end{aligned}$$

where we have used Theorem B.5.3 for the first term, Lemma B.6.6 for the second and Proposition B.6.4 for the third.

We also have the following rough estimates for the Wasserstein distances

$$\mathcal{W}_{2,a,b}(\mu_0^{(l+1)},\pi) = \mathcal{W}_{2,a,b}(\pi_0,\pi) = \mathcal{W}_{2,a,b}(\mu_G,\pi) \leq \frac{dC(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)}{N_D},$$

which follow from Proposition B.5.5, where the estimate of $\mathcal{W}_{2,a,b}(\mu_0^{(l+1)},\pi)$ along with Proposition B.6.4 implies $\mathcal{W}_{2,a,b}(\pi_0',\pi) \leq \frac{dC(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)}{N_D}$.

We have $(B_0 + Bl)2^l$ burn-in steps at level l, and $(B_0 + B(l+1))2^{l+1}$ burn-in steps at level $l+1$. Using the assumption that $h_0 < \frac{\tilde{m}}{256\tilde{M}\tilde{\gamma}N_D^{1/2}}$, we have for all $i \in \mathbb{N}$ $R(h_i) \geq \frac{mh_i}{8\gamma}$, and using Proposition B.6.4 we have

$$\begin{aligned}&\mathrm{Var}\left(f(z_k^{\prime(l,l+1)}) - f(z_k^{(l,l+1)})\right) \leq \mathbb{E}\left[\left(f(z_k^{\prime(l,l+1)}) - f(z_k^{(l,l+1)})\right)^2\right] \\ &\leq \left(\exp\left(-\frac{\tilde{m}\sqrt{N_D}(B_0+lB)h_0}{8\tilde{\gamma}}\right)\left(\|z'_{B/h_{l+1}} - z_0\|_{L^2,a,b} + \mathcal{W}_{2,a,b}(\pi_0,\pi) + \mathcal{W}_{2,a,b}(\pi_0',\pi)\right)\right. \\ &+ 4\exp\left(-\frac{\tilde{m}\sqrt{N_D}(B_0+lB)h_0}{16\tilde{\gamma}}\right)(\mathcal{W}_{2,a,b}(\pi_0,\pi) + \mathcal{W}_{2,a,b}(\pi_0',\pi)) \\ &\left.+ C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)h_l^{3/2}d^{1/2}N_D^{5/4}\right)^2 \\ &\leq \left(\exp\left(-\frac{m(B_0+lB)h_0}{16\gamma}\right)\frac{dC(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D)}{N_D} + C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)h_l^{3/2}d^{1/2}N_D^{5/4}\right)^2\end{aligned}$$

using the assumptions on B_0 and B

$$\leq C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)N_D^{5/2}h_l^3 d.$$

We now use the simple bound

$$\begin{aligned}\mathrm{Var}(D_{l,l+1}) \leq \mathbb{E}(D_{l,l+1}^2) &\leq \max_{1\leq k\leq K}\mathbb{E}\left[\left(f(z_k^{\prime l,l+1}) - f(z_k^{(l,l+1)})\right)^2\right] \\ &\leq C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)N_D^{5/2}h_l^3 d\end{aligned}$$

as required. □

Remark B.6.9. *As an alternative, one can consider a coupling with a randomized midpoint scheme, which was utilized in the work of [168] and [26] in the context of kinetic Langevin dynamics and Hamiltonian Monte Carlo. This is beyond the scope of the work, and thus we leave this as a direction to consider for future work.*

Proposition B.6.10. *Suppose a full gradient Gaussian approximation OHO chain $(z_k)_{k\in\mathbb{N}}$ at level 0 and a stochastic gradient UBU chain $(z'_k)_{k\in\mathbb{N}}$ at level 1 using the SVRG unbiased estimator, with stepsizes h_0 and $h_1 = \frac{h_0}{2}$, respectively. Further, we assume that they have synchronously coupled Brownian motions and $z_0 \sim \pi_0 = \mu_G$ and $z'_0 \sim \pi'_0$. Assuming the same assumptions as Proposition B.4.3 for $(z_k)_{k\in\mathbb{N}}$ and Proposition B.6.4 for $(z'_k)_{k\in\mathbb{N}}$. Then for f satisfying Assumption 3.3.12 we have the following variance bound*

$$\begin{aligned}
&\mathrm{Var}\left(f(z'_k) - f(z_k)\right) \leq \mathbb{E}\left[\left(f(z'_k) - f(z_k)\right)^2\right] \\
&\leq \left(\exp\left(-\frac{\tilde{m}\sqrt{N_D}kh_0}{8\tilde{\gamma}}\right)\left(\|z'_0 - z_0\|_{L^2,a,b} + \mathcal{W}_{2,a,b}(\pi'_0,\pi)\right)\right. \\
&+ 4\left(1 - R(h_1)/2\right)^{2k}\mathcal{W}_{2,a,b}(\pi'_0,\pi) + \frac{dC(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)}{N_D} + h_0\sqrt{d}C(\tilde{\gamma},\tilde{m},\tilde{M}) \\
&\left.+ C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)d^{1/2}N_D^{-3/4}\Theta^{1/2}h_1^{3/2}\right)^2,
\end{aligned}$$

where $R(h_1) = 1 - \sqrt{(1-c(h_1))^2 + C_0^2h_0^2}$.

Proof. By following the same argument as Proposition B.4.3, but by using Proposition B.6.4 and Theorem B.5.3 we have the desired result. However, because level zero and level one are approximating different diffusions, we can't use the contraction results for the continuous dynamics to bound $(\mathbf{II})$, so we consider an alternative argument. For this component, we use Lemma B.6.6. To bound $(\mathbf{I})$ we use Theorem B.5.3 and to bound $(\mathbf{III})$ we use Proposition B.6.4 and we have the required result. □

Proposition B.6.11. *Suppose that the assumptions of Proposition B.6.10 hold for the potential U and $h_0 > 0$. Assume that*

$$h_0 \leq \frac{C(\tilde{\gamma},\tilde{m},\tilde{M},\tau/N_D,N_b)}{N_D^{3/2}}, \quad B \geq \frac{16\log(2^{3/2})\tilde{\gamma}}{\tilde{m}h_0N_D^{1/2}}, \quad B_0 \geq \frac{16\tilde{\gamma}}{\tilde{m}N_D^{1/2}h_0}\log\left(\frac{1}{N_D^{9/4}h_0^{3/2}}\right),$$

$1 < k \leq K$, the levels are initialized as described in Section B.5 and for a function f which satisfies Assumption 3.3.12 for stochastic gradient UBUBU we have

$$\mathrm{Var}(D_{0,1}) \leq \frac{C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)d^2}{N_D^2} + C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D,N_b)N_D^{5/2}h_0^3d. \tag{B.51}$$

Proof. Following the same proof as Proposition B.6.8 using the results of Proposition B.6.10. □

B.6.2 Variance of $S(c_R)$

Theorem 3.3.24. *Considering UBUBU with stochastic gradients, suppose that Assumptions 3.3.12, 3.3.21, 3.3.22, 3.3.23 hold, and in addition* $\gamma \geq \sqrt{8M}$,

$$h_0 \leq \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tau/N_D, N_b)}{N_D^{3/2}}, \quad B \geq \frac{16\log(2^{3/2})\tilde{\gamma}}{\tilde{m}h_0 N_D^{1/2}}, \quad B_0 \geq \frac{16\tilde{\gamma}}{\tilde{m}h_0 N_D^{1/2}} \log\left(\frac{1}{N_D^{9/4}h_0^{3/2}}\right).$$

Suppose that $c_R \in [0,1)$ *and* $2 < \phi_N < 8$. *Then for any* $N \geq 1$, *the UBUBU estimator* $S(c_R)$ *has finite expected computational cost,* $\mathbb{E}S(c_R) = \pi(f)$, *and it has finite variance. Moreover, it satisfies a CLT as* $N \to \infty$, *and the asymptotic variance* σ_S^2 *defined in* (3.3.18) *can be bounded as*

$$\sigma_S^2 \leq \frac{1}{\tilde{m}N_D K} + C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, N_b, \phi_N)\frac{d^2}{c_N N_D^2}.$$

Proof. By Propositions B.6.8 and B.6.11, we have for $l \geq 1$ that

$$\mathbb{E}(D_{l,l+1}^2) \leq C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, N_b)h_l^3 dN_D^{5/2} \leq V_{D_1}\phi_{D_1}^{-l},$$

for $V_{D_1} = C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s \tau/N_D, b)h_0^3 dN_D^{5/2}$ and $\phi_{D_1} = 8$. For $l = 0$ we have

$$\mathbb{E}(D_{l,l+1}^2) \leq C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, N_b)\frac{d^2}{N_D^2} \leq V_{D_2}\phi_{D_2}^{-l},$$

for $V_{D_2} = C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, N_b)\frac{d^2}{N_D^2}$ and $\phi_{D_2} = 2$.

Due to the fact that for D_0 we take K i.i.d. Gaussian samples, it is easy to show using the Gaussian Poincaré inequality that

$$\mathrm{Var}(D_0) \leq \frac{1}{\tilde{m}N_D K}.$$

The computational cost at levels $l \geq 1$ satisfies the assumptions of Proposition 3.3.4, so if we fix $2 < \phi_N < 8$, all assumptions of this proposition are satisfied. Hence $S(c_R)$ is an unbiased estimator with finite variance and computational cost.

For the asymptotic variance using (3.3.18), and the above estimates we have

$$\begin{aligned}\sigma_S^2 &\leq \frac{1}{\tilde{m}N_D} + C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, N_b)\frac{d^2}{c_N N_D^2} + \sum_{l=1}^{\infty}\frac{V_{D_1}\phi_{D_1}^{-l}}{c_N \phi_N^{-l}} \\ &\leq \frac{1}{\tilde{m}N_D} + C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, N_b, \phi_N)\left(\frac{d^2}{c_N N_D^2} + \frac{h_0^3 dN_D^{5/2}}{c_N}\right),\end{aligned}$$

if we choose h_0 to be of the order $\mathcal{O}(1/N_D^{3/2})$ then we have

$$\leq \frac{1}{\tilde{m}N_D} + C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, N_b, \phi_N)\frac{d^2}{c_N N_D^2},$$

as required.

□

B.7 Variance bounds for UBUBU estimator with approximate gradients

One can also approximate the gradient in a cheap way, which has bias, but such that the bias tends to zero with the stepsize. The multilevel estimator will still be an unbiased estimator from the target measure.

For convex potentials, we can approximate the gradient with the Hessian at the minimizer by

$$\mathcal{Q}(x \mid \hat{x}) = \nabla U(\hat{x}) + \nabla^2 U(x^*)(x - \hat{x}). \tag{B.52}$$

Despite the fact that this estimator is biased, in our multilevel approach, the overall estimator will still be unbiased.

As before, the updates in $(\overline{x}_k, \overline{v}_k)_{k\geq 0}$ form a $\mathcal{BU}$ step, so they can be expressed as

$$\overline{x}_{k+1} = \overline{x}_k + \frac{1-\eta^2}{\gamma}\left(\overline{v}_k - h\mathcal{Q}(\overline{x}_k|\overline{x}_{L(k)})\right) + \sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(h, \xi_{k+1}^{(1)}\right) - \mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right)\right), \tag{B.53}$$

$$\overline{v}_{k+1} = \eta^2\left(\overline{v}_k - h\mathcal{Q}(\overline{x}_k|\overline{x}_{L(k)})\right) + \sqrt{2\gamma}\mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right), \tag{B.54}$$

where $\hat{x}_k = \overline{x}_{L(k)}$ and $L(k) = \tau\lfloor k/\tau \rfloor$.

It turns out that at level 0, it can be advantageous to simply use the gradients of the Gaussian approximation, and never compute gradients of U. This corresponds to gradient approximation of the form

$$\mathcal{Q}^*(x) = \mathcal{Q}(x|x^*) = \nabla^2 U(x^*)(x - x^*), \tag{B.55}$$

and so (B.53)-(B.54) holds with $\hat{x}_k = x^*$ for every $k \geq 0$ in this case.

B.7.1 Non-asymptotic guarantees

Lemma B.7.1. *Considering iterates $(x_k, v_k, \overline{x}_k)_{k\in\mathbb{N}}$ of approximate gradient UBU, with epoch length τ and gradient approximation $\mathcal{Q}$ given by* (B.52)*, and initial condition $(x_0, v_0) \in \mathbb{R}^{2d}$, we have the property*

$$\mathbb{E}\left\|\nabla U(\overline{x}_k) - \mathcal{Q}(\overline{x}_k \mid \overline{x}_{L(k)})\right\|^2 \leq \tilde{M}_1^2 N_D^2 (\tau - 1)^2 \max_{j\leq k} \|\overline{x}_j - x^*\|_{L^4}^2 \cdot \max_{j<k} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^4}^2,$$

and we also have

$$\mathbb{E}\left\|\nabla U(\overline{x}_k) - \mathcal{Q}(\overline{x}_k \mid x^*)\right\|^2 \leq \tilde{M}_1^2 N_D^2 \|\overline{x}_k - x^*\|_{L^4}^4.$$

Proof. Let the last full gradient evaluation be at iteration $L(k)$, then

$$\begin{aligned}
&\mathbb{E}\|\nabla U(\overline{x}_k) - \mathcal{Q}(\overline{x}_k \mid \overline{x}_{L(k)})\|^2 = \mathbb{E}\|\nabla U(\overline{x}_k) - \nabla U(\overline{x}_{L(k)}) - \nabla^2 U(x^*)(\overline{x}_k - \overline{x}_{L(k)})\|^2 \\
&= \mathbb{E}\left\|\left(\int_{t=0}^1 \nabla^2 U(\overline{x}_k + t(\overline{x}_{L(k)} - \overline{x}_k))\mathrm{d}t - \nabla^2 U(x^*)\right)(\overline{x}_k - \overline{x}_{L(k)})\right\|^2 \\
&\leq \tilde{M}_1^2 N_D^2 \mathbb{E}\left(\left(\int_{t=0}^1 \|\overline{x}_k + t(\overline{x}_{L(k)} - \overline{x}_k) - x^*\|\right)^2 \left\|\overline{x}_k - \overline{x}_{L(k)}\right\|^2\right) \\
&\leq \frac{\tilde{M}_1^2 N_D^2}{2}\left(\|\overline{x}_k - x^*\|_{L^4}^2 + \|\overline{x}_{L(k)} - x^*\|_{L^4}^2\right)\left\|\overline{x}_k - \overline{x}_{L(k)}\right\|_{L^4}^2 \\
&\leq \tilde{M}_1^2 N_D^2 (\tau-1)^2 \max_{j\leq k}\|\overline{x}_j - x^*\|_{L^4}^2 \cdot \max_{j<k}\|\overline{x}_{j+1} - \overline{x}_j\|_{L^4}^2.
\end{aligned}$$

The second claim follows by Taylor expansion. □

Now, we are going to bound the terms $\|\overline{x}_j - x^*\|_{L^4}$ and $\|\overline{x}_{j+1} - \overline{x}_j\|_{L^4}$.

Lemma B.7.2. *When using exact gradients, we have*

$$\|\overline{x}_{k+1} - \overline{x}_k\|_{L^4} \leq 2h\sqrt{M}\|z_k - z^*\|_{L^4,a,b} + 2h\sqrt{d}.$$

With approximate gradients, we have

$$\|\overline{x}_{k+1} - \overline{x}_k\|_{L^4} \leq 2h\sqrt{M}\left(\|\overline{z}_k - z^*\|_{L^4,a,b} + \sqrt{2}(1 + M/m)\|\overline{z}_{L(k)} - z^*\|_{L^4,a,b}\right) + 2h\sqrt{d}.$$

Proof. In the case of exact gradients, we have

$$\begin{aligned}
\overline{x}_{k+1} &= \overline{x}_k + \frac{1-\eta^2}{\gamma}(\overline{v}_k - h\nabla U(\overline{x}_k)) + \sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(h, \xi_{k+1}^{(1)}\right) - \mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right)\right), \\
\overline{v}_{k+1} &= \eta^2(\overline{v}_k - h\nabla U(\overline{x}_k)) + \sqrt{2\gamma}\mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right),
\end{aligned}$$

so using $\|\nabla U(x_k)\| \leq M\|x_k - x^*\|$, Lemma B.8.4, and the fact that $\xi \sim N(0, I_d)$ satisfies that $\mathbb{E}(\|\xi\|^4) \leq 3d^2$, we have that for $h \leq 1/\sqrt{M}$, $\gamma \geq \sqrt{8M}$,

$$\|\overline{x}_{k+1} - \overline{x}_k\|_{L^4} \leq h\|\overline{v}_k\|_{L^4} + h^2 M\|x_k - x^*\|_{L^4} + 3^{1/4}2^{1/2}h\sqrt{d} \leq 2h\sqrt{M}\|z_k - z^*\|_{L^4,a,b} + 2h\sqrt{d}.$$

For approximate gradients, we have

$$\overline{x}_{k+1} = \overline{x}_k + \frac{1-\eta^2}{\gamma}\left(\overline{v}_k - h\mathcal{Q}(\overline{x}_k|\overline{x}_{L(k)})\right) + \sqrt{\frac{2}{\gamma}}\left(\mathcal{Z}^{(1)}\left(h, \xi_{k+1}^{(1)}\right) - \mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right)\right), \quad \text{(B.56)}$$

$$\overline{v}_{k+1} = \eta^2\left(\overline{v}_k - h\mathcal{Q}(\overline{x}_k|\overline{x}_{L(k)})\right) + \sqrt{2\gamma}\mathcal{Z}^{(2)}\left(h, \xi_{k+1}^{(1)}, \xi_{k+1}^{(2)}\right). \quad \text{(B.57)}$$

Let $\tilde{x}_k = x_{L(k)} - (\nabla^2 U(x^*))^{-1}\nabla U(\overline{x}_{L(k)})$, and $\tilde{U}_k(x) = \frac{1}{2}(x-\tilde{x}_k)^T\nabla^2 U(x^*)(x-\tilde{x}_k)$. Then the approximate gradient step is the same as an exact gradient step with respect to the potential $\tilde{U}_k$. So we have by the result for exact gradients that for approximate gradients,

$$\begin{aligned}\|\overline{x}_{k+1} - \overline{x}_k\|_{L^4} &\leq h\|\overline{v}_k\|_{L^4} + h^2 M\|x_k - \tilde{x}_k\|_{L^4} + 3^{1/4}2^{1/2}hd \\ &\leq 2h\sqrt{M}\left\|\overline{z}_k - \begin{pmatrix}\tilde{x}_k \\ 0_d\end{pmatrix}\right\|_{L^4,a,b} + 2h\sqrt{d}.\end{aligned}$$

Here using the triangle inequality, we have

$$\begin{aligned}\left\|z_k - \begin{pmatrix}\tilde{x}_k \\ 0_d\end{pmatrix}\right\|_{L^4,a,b} &\leq \|\overline{z}_k - z^*\|_{L^4,a,b} + \left\|\begin{pmatrix}\tilde{x}_k \\ 0_d\end{pmatrix} - z^*\right\|_{L^4,a,b} \\ &\leq \|\overline{z}_k - z^*\|_{L^4,a,b} + \sqrt{2}(1+M/m)\|\overline{z}_{L(k)} - z^*\|_{L^4,a,b},\end{aligned}$$

hence

$$\|\overline{x}_{k+1} - \overline{x}_k\|_{L^4} \leq 2h\sqrt{M}\left(\|\overline{z}_k - z^*\|_{L^4,a,b} + \sqrt{2}(1+M/m)\|\overline{z}_{L(k)} - z^*\|_{L^4,a,b}\right) + 2h\sqrt{d}.$$

□

We still need to control the evolution of $\|\overline{z}_k - z^*\|_{L^4,a,b}$. As before in (B.16), we define the Lyapunov function $\mathcal{V}$ as

$$\mathcal{V}(x,v) = U(x) - U(x^*) + \frac{1}{4}\gamma^2\left(\|x - x^* + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x-x^*\|^2\right).$$

The following lemma establishes some useful properties about this.

Lemma B.7.3. *Suppose that $\gamma \geq \sqrt{8M}$, and that Assumptions 3.3.6 and 3.3.7 hold for U. Then for any $z = (x,v) \in \Lambda$, $\mathcal{V}(x,v) \geq 0$, and*

$$\mathcal{V}^{1/2}(x,v) \geq \frac{1}{8}(\gamma\|x-x^*\| + \|v\|) \geq \frac{\sqrt{M}}{8}\|z - z^*\|_{a,b}. \tag{B.58}$$

Moreover, $\mathcal{V}^{1/2}$ is 8γ-Lipschitz with respect to the $\|\cdot\|_{a,b}$ norm.

Proof. Using the strong convexity of U,

$$\begin{aligned}\mathcal{V}(x,v) &= U(x) - U(x^*) + \frac{1}{4}\gamma^2\left(\|x - x^* + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x-x^*\|^2\right) \\ &\geq \frac{m}{2}\|x-x^*\|^2 + \frac{1}{4}\gamma^2\left((1-\lambda)\|x-x^*\|^2 + 2\gamma^{-2}\|v\|^2 + 2\left\langle x-x^*, \gamma^{-1}v\right\rangle\right)\end{aligned}$$

using that $\left|2\left\langle x-x^*, \gamma^{-1}v\right\rangle\right| \leq \frac{\|x-x^*\|^2}{c} + c\|\gamma^{-1}v\|^2$ with $c = 8/5$, and that $0 < \lambda \leq \frac{1}{4}$,

$$\geq \frac{1}{4}\gamma^2\left(\frac{1}{8}\|x-x^*\|^2 + \frac{2}{5}\|\gamma^{-1}v\|^2\right) \geq \frac{1}{64}\left(\gamma\|x-x^*\| + \|v\|\right)^2 \geq \frac{M}{64}\|z-z^*\|_{a,b}^2,$$

and our first claim follows by taking square-root.

For the second claim, note that $\nabla\mathcal{V}^{1/2}(x,v) = \frac{1}{2}\frac{\nabla\mathcal{V}}{\mathcal{V}^{1/2}(x,v)}$. Here

$$\begin{aligned}
\nabla_x\mathcal{V}(x,v) &= \nabla U(x) + \frac{1}{2}\gamma^2((1-\lambda)(x-x^*) + \gamma^{-1}v),\\
\|\nabla_x\mathcal{V}(x,v)\| &\leq \left(M + \frac{\gamma^2(1-\lambda)}{2}\right)\|x-x^*\| + \frac{\gamma}{2}\|v\| \leq \gamma^2\|x-x^*\| + \frac{\gamma}{2}\|v\|,\\
\nabla_v\mathcal{V}(x,v) &= \frac{1}{2}\gamma^2(\gamma^{-1}((x-x^*) + \gamma^{-1}v) + \gamma^{-2}v)\\
\|\nabla_v\mathcal{V}(x,v)\| &\leq \frac{\gamma}{2}\|x-x^*\| + \|v\|,
\end{aligned}$$

so we have

$$\begin{aligned}
\|\nabla_x\mathcal{V}^{1/2}(x,v)\| &= \frac{\|\nabla_x\mathcal{V}(x,v)\|}{2\mathcal{V}^{1/2}(x,v)} \leq 4\gamma,\\
\|\nabla_v\mathcal{V}^{1/2}(x,v)\| &= \frac{\|\nabla_v\mathcal{V}(x,v)\|}{2\mathcal{V}^{1/2}(x,v)} \leq 4,
\end{aligned}$$

and since $\gamma \geq \sqrt{8M}$, for any $(x,v),(x',v') \in \Lambda$, we have

$$\begin{aligned}
&|\mathcal{V}^{1/2}(x,v) - \mathcal{V}^{1/2}(x',v')| = \left\langle \int_{t=0}^{1} \nabla\mathcal{V}^{1/2}(x+t(x'-x), v+t(x'-v))dt, (x'-x, v'-v)\right\rangle\\
&\leq 4\sqrt{2}\gamma\|(x,v)-(x',v')\|_{a,0} \leq 8\gamma\|(x,v)-(x',v')\|_{a,b}.
\end{aligned}$$

□

As previously, let $\lambda = \min\left(\frac{1}{4}, \frac{m}{\gamma^2}\right)$, and $c_4(h) = h\lambda\gamma - 8h^2\gamma^2(4+\lambda)$. By (B.18), for the exact gradient scheme, if $c_4(h) < \frac{1}{2}$, we have

$$\mathbb{E}\left[\mathcal{V}(\overline{x}_{k+1},\overline{v}_{k+1})^2 \mid \overline{x}_k,\overline{v}_k\right] \leq \left(1-\frac{c_4(h)}{2}\right)\mathcal{V}^2(\overline{x}_k,\overline{v}_k) + \frac{(6h\gamma d + 160h\gamma(1+\lambda^2))^2}{4c_4(h)} + 24h^2\gamma^2d^2.$$

Let $C_{\mathcal{V}}(h) := \frac{(6h\gamma d+160h\gamma(1+\lambda^2))^2}{4c_4(h)} + 24h^2\gamma^2d^2$, then by applying this j times, we have

$$\mathbb{E}\left[\mathcal{V}(\overline{x}_{k+j},\overline{v}_{k+j})^2 \mid \overline{x}_k,\overline{v}_k\right] \leq \left(1-\frac{c_4(h)}{2}\right)^j\mathcal{V}^2(\overline{x}_k,\overline{v}_k) + C_{\mathcal{V}}(h)\frac{1-(c_4(h)/2)^j}{1-c_4(h)/2}. \tag{B.59}$$

Now we are going to generalise this result to the approximate gradient scheme.

Lemma B.7.4. *Consider iterates $(x_k,v_k,\overline{x}_k)_{k\in\mathbb{N}}$ of approximate gradient UBU, with epoch length τ and gradient approximation $\mathcal{Q}$ given by* (B.52)*, and initial condition $(x_0,v_0) \in \mathbb{R}^{2d}$. Suppose that $L(k) = k$ (i.e. k is divisible by τ), and $c_4(h) > 0$, then for any $1 \leq j \leq \tau$, we have*

$$\begin{aligned}
\|\mathcal{V}^{1/2}(\overline{z}_{k+j})\|_{L^4} &\leq \left[\left(1-\frac{c_4(h)}{2}\right)^j\|\mathcal{V}^{1/2}(\overline{z}_k)\|_{L^4}^4 + C_{\mathcal{V}}(h)j\right]^{1/4}\\
&+ 8\gamma\left(48h^2\sqrt{M}\cdot j^2(\|\mathcal{V}^{1/2}(\overline{z}_k)\|_{L^4}^4 + C_{\mathcal{V}}(h)j)^{1/4} + 6h^2j^2\sqrt{dM}\right).
\end{aligned}$$

Proof. We use an interpolation argument, inspired by the interpolation to independence coupling in [44]. For $0 \leq i \leq j$, let $\overline{z}_{k+j}^{(i)} = \left(\overline{x}_{k+j}^{(i)}, \overline{v}_{k+j}^{(i)}\right)$ be defined by performing $j - i$ $\mathcal{BU}$ steps with exact gradients starting from $(\overline{x}_k, \overline{v}_k)$ according to (B.14)-(B.15), followed by i steps with approximate gradients according to (B.53)-(B.54). Then we have $\overline{z}_{k+j} = \overline{z}_{k+j}^{(j)}$, and $\overline{z}_{k+j}^{(0)}$ corresponds to taking j steps with exact gradients. By the triangle inequality, we have

$$\|\overline{z}_{k+j} - \overline{z}_{k+j}^{(0)}\|_{a,b} \leq \sum_{i=0}^{j-1} \|\overline{z}_{k+j}^{(i+1)} - \overline{z}_{k+j}^{(i)}\|_{a,b}.$$

Using Proposition B.3.6, we have a contraction according to $\|\cdot\|_{a,b}$ with synchronous coupling when using the approximate gradients (because these are exact gradients with respect to a Gaussian), so we have

$$\|\overline{z}_{k+j}^{(i+1)} - \overline{z}_{k+j}^{(i)}\|_{L^4,a,b} \leq \|\overline{z}_{k+i+1}^{(1)} - \overline{z}_{k+i+1}^{(0)}\|_{L^4,a,b},$$

which is the one-step error of the approximate gradient scheme versus the exact gradient scheme.

$$\begin{aligned}
&\|\overline{z}_{k+i+1}^{(1)} - \overline{z}_{k+i+1}^{(0)}\|_{L^4,a,b} \\
&= \left\| \left(\frac{(1-\eta^2)h}{\gamma} (\mathcal{Q}(\overline{x}_{k+i}^{(0)}|\overline{x}_k) - \nabla U(\overline{x}_{k+i}^{(0)})), \eta^2 h (\mathcal{Q}(\overline{x}_{k+i}^{(0)}|\overline{x}_k) - \nabla U(\overline{x}_{k+i}^{(0)})) \right) \right\|_{L^4,a,b} \\
&\leq \sqrt{2} \|\overline{x}_{k+i}^{(0)} - \overline{x}_k\|_{L^4} \cdot M \left(h^2 + \frac{h}{\sqrt{M}} \right).
\end{aligned}$$

So, for $h < \frac{1}{\sqrt{M}}$, we have

$$\begin{aligned}
\|\overline{z}_{k+j} - \overline{z}_{k+j}^{(0)}\|_{L^4,a,b} &\leq 3h\sqrt{M} \sum_{i=0}^{j-1} \|\overline{x}_{k+i}^{(0)} - \overline{x}_k\|_{L^4} \\
&\leq 3h\sqrt{M} \cdot j \cdot \sum_{0 \leq i \leq j-1} \|\overline{x}_{k+i+1}^{(0)} - \overline{x}_{k+i}^{(0)}\|_{L^4}
\end{aligned}$$

using Lemma B.7.2,

$$\leq 6h^2 M \cdot j \cdot \sum_{0 \leq i \leq j-1} \|\overline{z}_{k+i}^{(0)} - z^*\|_{L^4,a,b} + 6h^2 j^2 \sqrt{dM}$$

using Lemma B.7.3,

$$\leq 48h^2\sqrt{M} \cdot j \cdot \sum_{0 \leq i \leq j-1} \|\mathcal{V}^{1/2}(\overline{z}_{k+i}^{(0)})\|_{L^4} + 6h^2 j^2 \sqrt{dM}$$

using (B.59)

$$\leq 48h^2\sqrt{M} \cdot j^2 (\|\mathcal{V}^{1/2}(\overline{z}_k)\|_{L^4}^4 + C_{\mathcal{V}}(h)j)^{1/4} + 6h^2 j^2 \sqrt{dM}.$$

We do know that

$$\|\mathcal{V}^{1/2}(\overline{z}_{k+j}^{(0)})\|_{L^4}^4 \leq \left(1 - \frac{c_4(h)}{2}\right)^j \|\mathcal{V}^{1/2}(\overline{z}_k)\|_{L^4}^4 + C_{\mathcal{V}}(h),$$

so by the 8γ-Lipschitz property of $\mathcal{V}^{1/2}$ in $\|\cdot\|_{a,b}$ by Lemma B.7.3, we have

$$\begin{aligned}\|\mathcal{V}^{1/2}(\overline{z}_{k+j})\|_{L^4} \leq &\left[\left(1 - \frac{c_4(h)}{2}\right)^j \|\mathcal{V}^{1/2}(\overline{z}_k)\|_{L^4}^4 + C_{\mathcal{V}}(h)j\right]^{1/4} \\ &+ 8\gamma\left(48h^2\sqrt{M}\cdot j^2(\|\mathcal{V}^{1/2}(\overline{z}_k)\|_{L^4}^4 + C_{\mathcal{V}}(h)j)^{1/4} + 6h^2j^2\sqrt{dM}\right).\end{aligned}$$

□

Corollary B.7.5. *Consider iterates $(x_k, v_k, \overline{x}_k)_{k\in\mathbb{N}}$ of approximate gradient UBU, with epoch length τ and gradient approximation $\mathcal{Q}$ given by* (B.52) *approximating a potential U which satisfies Assumptions 3.3.28 and 3.3.22 with $z_0 \sim \mu_G$. Assume that*

$$h < \min\left\{2/\tau\gamma, 1, 1/2\gamma, \frac{\lambda\tau}{64(432\sqrt{M}\tau^2 + \gamma(1+\lambda)\tau)}\right\}, \quad \gamma \geq \sqrt{M},$$

then

$$\|\overline{z}_k - z^*\|_{L^4,a,b} \leq \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M})\sqrt{d}}{\sqrt{N_D}}.$$

Proof. If we define $b_k := \|\mathcal{V}^{1/2}(\overline{z}_{\tau k})\|_{L^4}$, then for $\gamma \geq \sqrt{8M}$ and $h < \frac{8}{\tau\gamma}$, we have $c_4(h) \leq 2/\tau$ (here $c_4(h)$ and λ are defined as in (B.13) and (B.12)), and so

$$\begin{aligned}b_{k+1} &\leq \left[\left(1 - \frac{c_4(h)}{2}\right)^\tau b_k^4 + C_{\mathcal{V}}(h)\tau\right]^{1/4} + 384h^2\gamma\sqrt{M}\tau^2\left(b_k^4 + C_{\mathcal{V}}(h)\tau\right)^{1/4} + 48\gamma\sqrt{M}h^2\tau^2\sqrt{d} \\ &\leq \left[\left(1 - \frac{c_4(h)\tau}{4}\right) b_k^4 + C_{\mathcal{V}}(h)\tau\right]^{1/4} + 384h^2\gamma\sqrt{M}\tau^2\left(b_k^4 + C_{\mathcal{V}}(h)\tau\right)^{1/4} + 48\gamma\sqrt{M}h^2\tau^2\sqrt{d}.\end{aligned} \tag{B.60}$$

Using this, for $b_k < \max\left\{\left(\frac{8C_{\mathcal{V}}(h)}{c_4(h)}\right)^{1/4}, \sqrt{d}\right\}$ we have that

$$b_{k+1} \leq (1 + 384h^2\gamma\sqrt{M}\tau^2)\left[\frac{8C_{\mathcal{V}}(h)}{c_4(h)} + d^2 + C_{\mathcal{V}}(h)\tau\right]^{1/4} + 48\gamma\sqrt{M}h^2\tau^2\sqrt{d}. \tag{B.61}$$

For $b_k \geq \max\left\{\left(\frac{8C_{\mathcal{V}}(h)}{c_4(h)}\right)^{1/4}, \sqrt{d}\right\}$, using that $(1+x)^{1/4} \leq 1 + \frac{x}{4}$ for $x \in [-1, \infty)$, we have

$$\begin{aligned}b_{k+1} &\leq b_k\left[\left[\left(1 - \frac{c_4(h)\tau}{4}\right) + \frac{\mathcal{C}_{\mathcal{V}}(h)\tau}{b_k^4}\right]^{1/4} + 384h^2\gamma\sqrt{M}\tau^2\left(1 + \frac{C_{\mathcal{V}}(h)\tau}{b_k^4}\right)^{1/4} + \frac{48\gamma\sqrt{M}h^2\tau^2\sqrt{d}}{b_k}\right] \\ &\leq \left[1 - \frac{c_4(h)\tau}{32} + 432h^2\gamma\sqrt{M}\tau^2\right] b_k\end{aligned}$$

using the definition $c_4(h) = h\lambda\gamma - 8h^2\gamma^2(4+\lambda)$

$$\leq \left[1 - h\frac{\lambda\gamma\tau}{32} + h^2(432\gamma\sqrt{M}\tau^2 + \gamma^2(1+\lambda)\tau)\right] b_k$$

using the assumption $h \leq \frac{\lambda\gamma\tau}{64(432\gamma\sqrt{M}\tau^2+\gamma^2(1+\lambda)\tau)}$

$$\leq \left[1 - h\frac{\lambda\gamma\tau}{64}\right] b_k.$$

Therefore we have that for all $k \in \mathbb{N}$

$$b_k \leq \left[1 - h\frac{\lambda\gamma\tau}{64}\right]^k b_0 + (1 + 384h^2\gamma\sqrt{M}\tau^2)\left[\frac{8C_{\mathcal{V}}(h)}{c_4(h)} + d^2 + C_{\mathcal{V}}(h)\tau\right]^{1/4} + 48\gamma\sqrt{M}h^2\tau^2\sqrt{d}.$$

Now considering $b_{k,j} := \|\mathcal{V}^{1/2}(\overline{z}_{\tau k+j})\|_{L^4}$ we have that by the same argument replacing τ by j that

$$b_{k,j} \leq \left[1 - h\frac{\lambda\gamma j}{64}\right]^j b_k + (1 + 384h^2\gamma\sqrt{M}j^2)\left[\frac{8C_{\mathcal{V}}(h)}{c_4(h)} + d^2 + C_{\mathcal{V}}(h)j\right]^{1/4} + 48\gamma\sqrt{M}h^2j^2\sqrt{d}.$$

Therefore considering the iterates of the approximate gradient UBU scheme we have

$$\frac{\sqrt{\tilde{M}N_D}}{8}\|\overline{z}_k - z^*\|_{L^4,a,b} \leq C(\tilde{\gamma},\tilde{m},\tilde{M})\left(\|\mathcal{V}^{1/2}(\overline{z}_0)\|_{L^4} + h\tau\sqrt{N_D} + \sqrt{d}\right)$$

and therefore using Lemma B.8.2 for the initial distribution we have

$$\|\overline{z}_k - z^*\|_{L^4,a,b} \leq \frac{C(\tilde{\gamma},\tilde{m},\tilde{M})\sqrt{d}}{\sqrt{N_D}}.$$

□

Proposition B.7.6. *For an approximate gradient UBU integrator with iterates $(z_k)_{k\in\mathbb{N}}$, transition kernel P_h and a potential U satisfying Assumptions 3.3.6-3.3.10 and $z_0 \sim \mu_G$, where we approximate the gradient using the gradient approximation $\mathcal{Q}$ given in* (B.52)*. Consider the continuous solution to* (3.1.1) *$(Z_t)_{t\geq 0}$, and define $Z^k := Z_{kh}$ for $k \in \mathbb{N}$, where $Z^0 \sim \pi$ is initialized at the invariant measure with synchronously coupled Brownian motion to $(z_k)_{k\in\mathbb{N}}$, then for all*

$$h < \min\left\{2/\tau\gamma, 1, 1/2\gamma, \frac{\lambda\tau}{64(432\sqrt{M}\tau^2 + \gamma(1+\lambda)\tau)}\right\}, \quad \gamma \geq \sqrt{8M},$$

$k, l \in \mathbb{N}$ *such that* $k > l$

$$\|z_k - Z^k\|_{L^2,a,b} \leq (1 - c(h))^{k-l}\|z_l - Z^l\|_{L^2,a,b} + \frac{h\left((\tau-1)\sqrt{d} + \sqrt{N_D}\right)C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)\sqrt{d}}{\sqrt{N_D}},$$

and further

$$\mathcal{W}_{2,a,b}(\mu_G P_h^k, \pi) \leq (1 - c(h))^{k-l}\mathcal{W}_{2,a,b}(\mu_G P_h^l, \pi) + \frac{h\left((\tau-1)\sqrt{d} + \sqrt{N_D}\right)C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)\sqrt{d}}{\sqrt{N_D}}.$$

Proof. Firstly, we introduce the notation $z_k^h := (x_k^h, v_k^h) := \psi_h\left(z_k, h, (W_{t'})_{t'=kh}^{(k+1)h}\right)$ for all $k \in \mathbb{N}$, an iteration of the full gradient scheme with stepsize $h > 0$ and initial point z_k with synchronously coupled Brownian motion to the approximate gradient scheme. We split up the difference in the following way

$$\begin{aligned}\|z_k - Z^k\|_{L^2,a,b} &\leq \left\|z_k - z_{k-1}^h\right\|_{L^2,a,b} + \|z_{k-1}^h - Z^k\|_{L^2,a,b} \\ &\leq \|z_k - z_{k-1}^h\|_{L^2,a,b} + \|\psi(Z^{k-1}, h, (W_{t'})_{t'=(k-1)h}^{kh}) - Z^k\|_{L^2,a,b} \\ &+ \|z_{k-1}^h - \psi(Z^{k-1}, h, (W_{t'})_{t'=(k-1)h}^{kh})\|_{L^2,a,b} \\ &= (\mathsf{I})' + (\mathsf{II})' + (\mathsf{III})'.\end{aligned}$$

We have by Corollary B.7.5, Lemma B.7.2 and Lemma B.7.1 that

$$\begin{aligned}(\mathsf{I})' &\leq \frac{\sqrt{2}h}{\sqrt{M}} M_1^s(\tau-1) \max_{j\leq k-1} \|\overline{x}_j - x^*\|_{L^4} \cdot \max_{j<k-1} \|\overline{x}_{j+1} - \overline{x}_j\|_{L^4} \\ &\leq (\tau-1)h^2 C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s)d,\end{aligned}$$

by the discretization results in Appendix B.8

$$(\mathsf{II})' \leq \tilde{C}h^2 \leq \frac{3}{7}\sqrt{d}\left(\sqrt{M} + \gamma\right)h^2$$

and

$$(\mathsf{III})' \leq (1 - c(h))\|z_{k-1} - Z^{k-1}\|_{L^2,a,b},$$

where the inequality for $(\mathsf{II})'$ is shown in Appendix B.8. Therefore going from local to global we have that

$$\begin{aligned}\|z_k - Z^k\|_{L^2,a,b} &\leq (1-c(h))^{k-l}\|z_l - Z^l\|_{L^2,a,b} + \frac{h^2\left((\tau-1)d + \sqrt{N_D d}\right)C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s)}{c(h)} \\ &= (1-c(h))^{k-l}\|z_l - Z^l\|_{L^2,a,b} + \frac{h\left((\tau-1)d + \sqrt{N_D d}\right)C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s)}{\sqrt{N_D}}.\end{aligned}$$

For non-asymptotic Wasserstein results, we simply replace Z^{k-1} with the continuous dynamics initialized at $\tilde{Z}_{k-1} \sim \pi$ be such that $\|\tilde{Z}_{k-1} - z_{k-1}\|_{L^2,a,b} = \mathcal{W}_{2,a,b}(\mu P_h^{k-1}, \pi)$ as in [140, Theorem 23]. We can then apply Lemma B.8.1 to get the required result. □

B.7.2 Variance bound of $D_{l,l+1}$

Proposition B.7.7. *Suppose two approximate gradient UBU chains at coarser and finer discretization levels l and $l+1$, with synchronously coupled Brownian motions $(z_k)_{k\in\mathbb{N}}$ and $(z'_k)_{k\in\mathbb{N}}$ and stepsizes h_l and $h_{l+1} = h_l/2$, satisfying the conditions of Proposition B.7.6, be such that $z_0 \sim \pi_0 = \mu_G$ and $z'_0 \sim \pi'_0 = \mu_G(P_{h_{l+1}}^A)^{B/h_{l+1}}$. Then for f satisfying Assumption 3.3.12 we have the following variance*

bound

$$\begin{aligned}
&\mathrm{Var}\left(f(z_k') - f(z_k)\right) \leq \mathbb{E}\left(f(z_k') - f(z_k)\right)^2 \leq \\
&\left(\exp\left(-\frac{\tilde{m}\sqrt{N_D}kh_l}{8\tilde{\gamma}}\right)\left(\|z_0' - z_0\|_{L^2,a,b} + \mathcal{W}_{2,a,b}(\pi_0,\pi) + \mathcal{W}_{2,a,b}(\pi_0',\pi)\right)\right. \\
&+ (1-c(h_l))^k \mathcal{W}_{2,a,b}(\pi_0,\pi) + (1-c(h_{l+1}))^{2k}\mathcal{W}_{2,a,b}(\pi_0',\pi) \\
&\left.+ \frac{h_l\left((\tau-1)d + \sqrt{N_D d}\right)C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)}{\sqrt{N_D}}\right)^2.
\end{aligned}$$

Proof. By following the same argument as Proposition B.4.3 using Proposition B.7.6 we have the desired result. □

Corollary B.7.8. *Suppose that the assumptions of Proposition B.7.7 hold for the potential U and $h_0 > 0$. Assume that*

$$h_0 \leq \frac{C(\tilde{\gamma},\tilde{m},\tilde{M},\tau/N_D)}{N_D^{3/2}}, \quad B \geq \frac{16\log(2)\tilde{\gamma}}{\tilde{m}h_0 N_D^{1/2}}, \quad B_0 \geq \frac{16\tilde{\gamma}}{\tilde{m}N_D^{1/2}h_0}\log\left(\frac{1}{N_D^3 h_0^2}\right),$$

and the levels are initialized as described in Section B.5, Let $l \geq 1$, $1 \leq k \leq K$, and a test function f satisfy Assumption 3.3.12 then for approximate gradient UBUBU with $\tau = N_D$ we have

$$\mathrm{Var}(D_{l,l+1}) \leq C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s,\tau/N_D)d^2h_l^2N_D. \tag{B.62}$$

Proof. Following the proof of Proposition B.6.8, but using the results of Proposition B.7.6 you get the required result. □

Proposition B.7.9. *Suppose a OHO chain at level 0 using a full gradient Gaussian approximation and a approximate gradient UBU chain at level 1, with synchronously coupled Brownian motions $(z_k)_{k\in\mathbb{N}}$ and $(z_k')_{k\in\mathbb{N}}$ and stepsizes h_0 and $h_1 = h_0/2$, satisfying the conditions of Proposition B.7.6, be such that $z_0 \sim \pi_0 = \mu_G$ and $z_0' \sim \pi_0' = \mu_G(P_{h_1}^A)^{B/h_1}$. Then for f satisfying Assumption 3.3.12 we have the following variance bound*

$$\begin{aligned}
&\mathrm{Var}\left(f(z_k') - f(z_k)\right) \leq \mathbb{E}\left[\left(f(z_k') - f(z_k)\right)^2\right] \leq \mathbb{E}\|z_k' - z_k\|_{a,b}^2 \\
&\leq \left(\exp\left(-\frac{mkh_0}{16\gamma}\right)\left(\|z_0' - z_0\|_{L^2,a,b} + \mathcal{W}_{2,a,b}(\pi_0',\pi)\right)\right. \\
&+ (1-c(h_1))^k\mathcal{W}_{2,a,b}(\pi_0',\pi) + \frac{dC(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s)}{N_D} + C(\tilde{\gamma},\tilde{m},\tilde{M})h_0\sqrt{d} \\
&\left.+ \frac{h_1((\tau-1)d + \sqrt{N_D d})C(\tilde{\gamma},\tilde{m},\tilde{M},\tilde{M}_1^s))}{\sqrt{N_D}}\right)^2.
\end{aligned}$$

Proof. We aim to consider the same argument as Proposition B.6.10 using the results from Proposition B.7.6, Lemma B.6.6 and Proposition B.5.3. □

Corollary B.7.10. *Suppose that the assumptions of Proposition B.7.9 hold for the potential U and $h_0 > 0$. Assume that*

$$h_0 \leq \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tau/N_D)}{N_D^{3/2}}, \qquad B \geq \frac{16\log(2)\tilde{\gamma}}{\tilde{m}h_0 N_D^{1/2}}, \quad B_0 \geq \frac{16\tilde{\gamma}}{\tilde{m}N_D^{1/2}h_0}\log\left(\frac{1}{N_D^3 h_0^2}\right),$$

and the levels are initialized as described in Section B.5. Let $1 \leq k \leq K$, and a test function f satisfy Assumption 3.3.12 then for approximate gradient UBUBU with $\tau = N_D$ we have

$$\mathsf{Var}(D_{0,1}) \leq \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D)d^2}{N_D^2}. \tag{B.63}$$

Proof. Following the same argument as Corollary B.7.8, but using Proposition B.7.9. □

B.7.3 Variance bound of $S(c_R)$

Theorem 3.3.29. *Considering UBUBU-Approx method, suppose that Assumptions 3.3.12, 3.3.22, 3.3.23, 3.3.28 hold, and in addition $\gamma \geq \sqrt{8M}$,*

$$h_0 \leq \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tau/N_D)}{N_D^{3/2}}, \quad B \geq \frac{16\log(2)\tilde{\gamma}}{\tilde{m}h_0 N_D^{1/2}}, \quad B_0 \geq \frac{16\tilde{\gamma}}{\tilde{m}N_D^{1/2}h_0}\log\left(\frac{1}{N_D^3 h_0^2}\right).$$

Suppose that $c_R \in [0,1)$ and $2 < \phi_N < 4$. Then for any $N \geq 1$, $S(c_R)$ has finite expected computational cost, $\mathbb{E}S(c_R) = \pi(f)$, and it has finite variance. Moreover, it satisfies a CLT as $N \to \infty$, and the asymptotic variance σ_S^2 defined in (3.3.18) *can be bounded as*

$$\sigma_S^2 \leq \frac{1}{\tilde{m}N_D K} + \frac{C(\tilde{\gamma}, \tilde{m}, \tilde{M}, \tilde{M}_1^s, \tau/N_D, \phi_N)d^2}{c_N N_D^2}.$$

Proof. Following the same argument as Theorem 3.3.24 using Corollaries B.7.8 and B.7.10.

□

B.8 Auxiliary results

Lemma B.8.1. *If we have a sequence of non-negative numbers $(r_k)_{k\in\mathbb{N}}$ such that for constants $A \in (0, 1/2)$, $B, C, D \in \mathbb{R}_{\geq 0}$ such that*

$$r_{k+1}^2 \leq \left(\left((1-A)\, r_k^2 + B\right)^{1/2} + C\right)^2 + D$$

then

$$r_k \leq \sqrt{2}\left(1 - \frac{A}{2}\right)^k (r_0 + \sqrt{B}) + \frac{2\sqrt{2}C}{A} + 2\sqrt{\frac{D+B}{A}}.$$

Proof. If we define $\tilde{r}_k := \sqrt{(1-A)r_k^2 + B}$, then we have that

$$\begin{aligned}\tilde{r}_{k+1}^2 &\leq (1-A)\left(\tilde{r}_k + C\right)^2 + (1-A)D + B \\ &\leq ((1-A/2)\tilde{r}_k + C)^2 + D + B.\end{aligned}$$

Then using [51, Lemma 7] we have that

$$\tilde{r}_k \leq (1-A/2)^k \tilde{r}_0 + \frac{2C}{A} + \sqrt{\frac{2(D+B)}{A}},$$

then

$$r_k\sqrt{1-A} \leq \tilde{r}_k \leq (1-A/2)^k\left(r_0 + \sqrt{B}\right) + \frac{2C}{A} + \sqrt{\frac{2(D+B)}{A}},$$

and, using the fact that $A \leq 1/2$, we obtain the required result. □

Lemma B.8.2. *If a potential $U : \mathbb{R}^d \to \mathbb{R}$ is such that $\nabla^2 U \succ mI$, and $\nabla U(x^*) = 0$ then for $x \sim \pi \propto e^{-U(x)}$ we have*

$$\mathbb{E}\left[\|x - x^*\|^4\right]^{1/4} \leq 3^{1/4}\sqrt{\frac{d}{m}} \quad \text{and} \quad \mathbb{E}\left[\|x - x^*\|^8\right]^{1/8} \leq 105^{1/8}\sqrt{\frac{d}{m}}.$$

Proof. By using integration by parts and the convexity of U we have that

$$\begin{aligned}&\int_{x\in\mathbb{R}^d} \|x - x^*\|^4 e^{-U(x)}dx \leq \int_{x\in\mathbb{R}^d} \sum_{i=1}^d \sum_{j=1}^d \left(x_i - x_i^*\right)^2 \left(x_j - x_j^*\right)^2 e^{-U(x)}dx \\ &\leq d\sum_{i=1}^d \int_{x\in\mathbb{R}^d} \left(x_i - x_i^*\right)^4 e^{-U(x)}dx \\ &\leq \frac{d}{m}\sum_{i=1}^d \int_{x_{-i}\in\mathbb{R}^{d-1}} \int_{x_i\in\mathbb{R}} \left(x_i - x_i^*\right)^3 \partial_i U(x) e^{-U(x)}dx_i dx_{-i} \\ &= \frac{3d}{m}\sum_{i=1}^d \int_{x_{-i}\in\mathbb{R}^{d-1}} \int_{x_i\in\mathbb{R}} \left(x_i - x_i^*\right)^2 e^{-U(x)}dx_i dx_{-i} \\ &\leq \frac{3d}{m^2}\sum_{i=1}^d \int_{x_{-i}\in\mathbb{R}^{d-1}} \int_{x_i\in\mathbb{R}} \left(x_i - x_i^*\right) \partial_i U(x) e^{-U(x)}dx_i dx_{-i} \\ &= \frac{3d}{m^2}\sum_{i=1}^d \int_{x\in\mathbb{R}^d} e^{-U(x)}dx,\end{aligned}$$

and similarly, we have

$$\begin{aligned}
&\int_{x\in\mathbb{R}^d} \|x-x^*\|^8 e^{-U(x)}dx \\
&\leq \int_{x\in\mathbb{R}^d} \sum_{i,j,k,l=1}^{d} (x_i-x_i^*)^2\left(x_j-x_j^*\right)^2(x_k-x_k^*)^2(x_l-x_l^*)^2\, e^{-U(x)}dx \\
&\leq \int_{x\in\mathbb{R}^d} d^3\sum_{i=1}^{d}(x_i-x_i^*)^8\, e^{-U(x)}dx \\
&\leq \frac{105d^3}{m^4}\sum_{i=1}^{d}\int_{x\in\mathbb{R}^d} e^{-U(x)}dx,
\end{aligned}$$

as required.

□

Proposition B.8.3 (Local error bounds for UBU)**.** *Suppose we have a potential U which satisfies Assumptions 3.3.6 and 3.3.7. Let $\phi\left(\xi,h,(W_{t'})_{t'=0}^{h}\right)$ be the solution to* (3.1.1) *at time h with initial condition $\xi\sim\pi$, using Brownian motion $(W_{t'})_{t'=0}^{t}$. Let $\psi_h\left(\xi,h,(W_{t'})_{t'=0}^{t}\right)$ to be the solution of the numerical discretization UBU step, defined in Sec. 3.2.1, with stepsize h and the same initial condition and Brownian motion. Then we have the following local error bound*

$$\|\phi(\xi,h,(W_{t'})_{t'=0}^{h})-\psi_h(\xi,h,(W_{t'})_{t'=0}^{h})\|_{L^2,a,b}\leq\frac{3}{7}\sqrt{d}\left(\sqrt{M}+\gamma\right)h^2,$$

for $h<\min\left\{\frac{1}{5\sqrt{M}},\frac{1}{2\gamma}\right\}$.

Proof. Using the method of [140] we wish to bound the local error of the UBU scheme, when initialized at the target measure of the continuous dynamics. When considering (B.38) and (B.41) we have that for $\xi\sim\pi$

$$\phi(\xi,h,(W_{t'})_{t'=0}^{h})-\psi_h(\xi,h,(W_{t'})_{t'=0}^{h})=(\triangle_x,\triangle_v),$$

$$\triangle_x=-\int_0^h\mathcal{F}(h-s)\nabla U(x(s))ds+h\mathcal{F}(h/2)\nabla U(y).$$

and

$$\triangle_v=-\int_0^h\mathcal{E}(h-s)\nabla U(x(s))ds+h\mathcal{E}(h/2)\nabla U(y).$$

Next, we use the fundamental theorem of calculus

$$\begin{aligned}
\mathcal{E}(h-s)\nabla U(x(s))=\mathcal{E}(h/2)\nabla U(x(h/2))&\\
+\int_{h/2}^{s}\left(\mathcal{E}(h-s')\nabla^2U(x(s'))v(s')+\gamma\mathcal{E}(h-s')\nabla U(x(s'))\right)ds'.&
\end{aligned}$$

Then

$$\triangle_v=-h\mathcal{E}(h/2)\left(\nabla U(x(h/2))-\nabla U(y)\right)+\tilde{I}_1+\tilde{I}_2,$$

where

$$\tilde{I}_1=-\int_0^h\int_{h/2}^{s}\mathcal{E}(h-s')\nabla^2U(x(s'))v(s')ds'ds,$$

and

$$\tilde{I}_2 = -\int_0^h \int_{h/2}^s \gamma \mathcal{E}(h-s')\nabla U(x(s'))ds'ds.$$

Hence

$$\|h\mathcal{E}(h/2)\left(\nabla U(x(h/2)) - \nabla U(y)\right)\|_{L^2} \leq \frac{h^3 M^{3/2}\sqrt{d}}{\sqrt{48}}$$

from [140, Eq. 36]. Now, we estimate $\tilde{I}_1$ as

$$\begin{aligned}\mathbb{E}\left(\|\tilde{I}_1\|^2\right) &\leq \mathbb{E}\left[\left(\int_0^h \left|\int_{h/2}^s \mathcal{E}(h-s')^2 ds'\right| ds\right) \times \left(\int_0^h \left|\int_{h/2}^s \|\nabla^2 U(x(s'))v(s')\|^2 ds'\right| ds\right)\right] \\ &\leq \frac{\mathcal{F}(h)^2}{4} \times \frac{h^2 M^2 d}{4} \leq \frac{h^4 M^2 d}{16},\end{aligned}$$

and we estimate $\tilde{I}_2$ as

$$\begin{aligned}\mathbb{E}\left(\|\tilde{I}_2\|^2\right) &\leq \gamma^2 \mathbb{E}\left[\left(\int_0^h \left|\int_{h/2}^s \mathcal{E}(h-s')^2 ds'\right| ds\right) \times \left(\int_0^h \left|\int_{h/2}^s \left\|\nabla U(x(s'))\right\|^2 ds'\right| ds\right)\right] \\ &\leq \gamma^2 \frac{\mathcal{F}(h)^2}{4} \times \frac{h^2 M d}{4} \leq \frac{h^4 M \gamma^2 d}{16},\end{aligned}$$

then

$$\|\triangle_v\|_{L^2} \leq \frac{h^3 M^{3/2}\sqrt{d}}{\sqrt{48}} + \frac{h^2 M\sqrt{d}}{4} + \frac{h^2\gamma\sqrt{Md}}{4}.$$

Using [140, Eq 42 Estimate] we get the bound

$$\|\triangle_x\|_{L^2} \leq \frac{h^3}{24}\left(\sqrt{3}hM^{3/2} + \left(\frac{\sqrt{42}}{2} + 1\right)M + \gamma M^{1/2}\right)\sqrt{d}.$$

In the modified Euclidean norm we have

$$\begin{aligned}\|(\triangle_x, \triangle_v)\|_{L^2,a,b} &\leq \sqrt{\frac{3}{2}}\left(\|\triangle_x\|_{L^2} + \frac{1}{\sqrt{M}}\|\triangle_v\|_{L^2}\right) \\ &\leq \sqrt{\frac{3d}{2}}h^2\left(\frac{h}{24}\left(\sqrt{3}hM^{3/2} + \frac{9}{2}M + \gamma M^{1/2}\right) + \frac{\sqrt{M}}{4} + \frac{\gamma}{4}\right),\end{aligned}$$

and under the assumption that $h < \min\{\frac{1}{5\sqrt{M}}, \frac{1}{2\gamma}\}$ we see that

$$\|(\triangle_x, \triangle_v)\|_{L^2,a,b} \leq \frac{3}{7}\sqrt{d}\left(\sqrt{M} + \gamma\right)h^2.$$

□

The following lemma will be bound the variances of $\mathcal{Z}^{(1)}$, $\mathcal{Z}^{(2)}$ and $\mathcal{Z}^{(1)} - \mathcal{Z}^{(2)}$.

Lemma B.8.4. *For $\mathcal{Z}^{(1)}$ and $\mathcal{Z}^{(2)}$ as defined in* (3.2.8)*, we have*

$$\begin{aligned}\mathrm{Cov}\left(\mathcal{Z}^{(1)}\left(h,\xi^{(1)}\right)\right) &= hI_d,\\ \mathrm{Cov}\left(\mathcal{Z}^{(2)}\left(h,\xi^{(1)},\xi^{(2)}\right)\right) &\preceq hI_d,\\ \mathrm{Cov}\left(\mathcal{Z}^{(1)}\left(h,\xi^{(1)}\right)-\mathcal{Z}^{(2)}\left(h,\xi^{(1)},\xi^{(2)}\right)\right) &\preceq \frac{\gamma h^2}{4}I_d.\end{aligned}$$

Proof. From the definitions

$$\begin{aligned}\mathcal{Z}^{(1)}\left(h,\xi^{(1)}\right) &= \sqrt{h}\xi^{(1)},\\ \mathcal{Z}^{(2)}\left(h,\xi^{(1)},\xi^{(2)}\right) &= \sqrt{\frac{1-\eta^4}{2\gamma}}\left(\sqrt{\frac{1-\eta^2}{1+\eta^2}\cdot\frac{2}{\gamma h}}\xi^{(1)}+\sqrt{1-\frac{1-\eta^2}{1+\eta^2}\cdot\frac{2}{\gamma h}}\xi^{(2)}\right),\end{aligned}$$

it is clear that $\mathrm{Cov}\left(\mathcal{Z}^{(1)}\left(h,\xi^{(1)}\right)\right)=hI_d$ and $\mathrm{Cov}\left(\mathcal{Z}^{(2)}\left(h,\xi^{(1)},\xi^{(2)}\right)\right)=\frac{1-\eta^4}{2\gamma}I_d\preceq hI_d$. For the last claim, we have

$$\begin{aligned}&\mathrm{Cov}\left(\mathcal{Z}^{(2)}\left(h,\xi^{(1)},\xi^{(2)}\right)-\mathcal{Z}^{(1)}\left(h,\xi^{(1)}\right)\right)\\ &=\left(\sqrt{\frac{1-\eta^4}{2\gamma}}\sqrt{\frac{1-\eta^2}{1+\eta^2}\cdot\frac{2}{\gamma h}}-\sqrt{h}\right)^2+\frac{1-\eta^4}{2\gamma}\left(1-\frac{1-\eta^2}{1+\eta^2}\cdot\frac{2}{\gamma h}\right)\\ &=\frac{1-\eta^4}{2\gamma}+h-2\frac{(1-\eta^2)}{\gamma}=\frac{1-e^{-2\gamma h}-4(1-e^{-\gamma h})+2\gamma h}{2\gamma}\leq\frac{(\gamma h)^2}{2}\cdot\frac{1}{2\gamma}\leq\frac{\gamma h^2}{4}.\end{aligned}$$

□

Lemma B.8.5. *Let $A=\sum_{l=1}^n A^{(l)}$, with $A^{(l)}\in\mathbb{R}^{d\times d}$ for every $1\leq l\leq n$. Then we have*

$$\|A\|_{\{12\}\{3\}}=\left\|\sum_{i_1,l,m}(A^{(l)}_{i_1,\cdot,\cdot})^T\cdot A^{(m)}_{i_1,\cdot,\cdot}\right\|^{1/2}.$$

Proof. This follows by expanding the formula $\|A\|_{\{12\}\{3\}}=\left\|\sum_{i_1}A^T_{i_1,\cdot,\cdot}\cdot A_{i_1,\cdot,\cdot}\right\|^{1/2}$ shown in Lemma 7 of [123]. □

The following lemma shows some bounds for the gradient-Lipschitz constant M and strongly Hessian Lipschitz constant M_1^s for the Bayesian multinomial regression example.

Lemma B.8.6. *Consider the Bayesian multinomial regression likelihood of the form,*

$$p(y^j|q)=\frac{\exp(\langle x^j,q^{y^j}\rangle)}{\sum_{1\leq k\leq m}\exp(\langle x^j,q^k\rangle)},\tag{B.64}$$

where the posterior potential is given as

$$U(q)=-\log(p_0(q))-\sum_{k=1}^{N_D}\log\left(p(y^j|q)\right),\tag{B.65}$$

with $p_0(q) = \frac{\exp(-\|q\|^2/(2\sigma_0^2))}{(\pi\sigma_0^2)^{d/2}}$. *This satisfies the following bounds,*

$$\sup_{q\in\mathbb{R}^d} \|\nabla^2 U(q)\| \leq \sigma_0^{-2} + \left\|\sum_{l=1}^{N_D}(x^l)(x^l)^T\right\|,$$
$$\sup_{q\in\mathbb{R}^d} \|\nabla^3 U(q)\|_{\{12\}\{3\}} \leq 6\left\|\sum_{l=1}^{N_D}\left[(x^l)(x^l)^T\left(\sum_{m=1}^{N_D}\langle x^l, x^m\rangle^2\right)\right]\right\|^{1/2}.$$

Remark B.8.7. *If* $N_D \to \infty$, *and* $(x^l)_{1\leq l\leq N_D}$ *are i.i.d. samples from a continuous* d*-dimensional distribution that is non-degenerate with* $\mathbb{E}(\|x^l\|^6) = \mathcal{O}(1)$, *then we would expect* $\|\nabla^2 U(q)\| \propto \frac{N_D}{d}$, *and* $\|\nabla^3 U(q)\|_{\{12\}\{3\}} \propto \frac{N_D}{d}$.

Proof. For $1 \leq i \leq m$, let $E^i = \begin{pmatrix} 0_{d_o} \\ \vdots \\ I_{d_o} \\ \vdots \\ 0_{d_o} \end{pmatrix}$ be an $d \times d_o$ block matrix with an identity matrix at block i. Let

$$S(x,q) = \sum_{1\leq l\leq m} \exp(\langle x, q^l\rangle).$$

Let $x \otimes y \otimes z \in \mathbb{R}^{d\times d\times d}$ denote the tensor product of 3 vectors, i.e. $(x\otimes y\otimes z)_{ijk} = x_i y_j z_k$. Then we can express the likelihood term $\log(p(y|q))$ and its derivatives as follows.

$$\begin{aligned}
&\log(p(y|q)) = \langle x, q^y\rangle - \log(S(x,q)) \\
&\nabla_q \log(p(y|q)) = \sum_{i=1}^m (E^i x)\left(\mathbb{I}[y=i] - \exp(\langle x, q^i\rangle)(S(x,q))^{-1}\right) \\
&\nabla_q^2 \log(p(y|q)) = \sum_{i,j=1}^m (E^i x)(E^j x)^T \exp(\langle x, q^i\rangle + \langle x, q^j\rangle)(S(x,q))^{-2} \\
&- \sum_{i=1}^m (E^i x)(E^i x)^T \exp(\langle x, q^i\rangle)(S(x,q))^{-1} \\
&\nabla_q^3 \log(p(y|q)) = -\sum_{i=1}^m (S(x,q))^{-1}\exp(\langle x, q^i\rangle)(E^i x)\otimes(E^i x)\otimes(E^i x) \\
&+ \sum_{i,j=1}^m (S(x,q))^{-2}\exp(\langle x, q^i\rangle + \langle x, q^j\rangle)\cdot \\
&\cdot\left((E^i x)\otimes(E^i x)\otimes(E^j x) + (E^i x)\otimes(E^j x)\otimes(E^i x) + (E^i x)\otimes(E^j x)\otimes(E^j x)\right) \\
&- 2(S(x,q))^{-3}\sum_{i,j,k=1}^m (E^i x)\otimes(E^j x)\otimes(E^k x)\exp(\langle x, q^i\rangle + \langle x, q^j\rangle + \langle x, q^k\rangle).
\end{aligned}$$

The first claim of the lemma bounding $\|\nabla^2 U(q)\|$ follows from the fact that

$$0_d \preceq -\nabla_q^2 \log(p(y|q)) \preceq \sum_i^m (E^i x)(E^i x)^T \exp(\langle x, q^i\rangle)(S(x,q))^{-1} \preceq \sum_i^m (E^i x)(E^i x)^T,$$

here $\preceq$ denotes the semidefinite order.

For the second claim, note that

$$\left\| -\sum_{l=1}^{N_D}\sum_{i=1}^{m}(S(x^l,q))^{-1}\exp(\langle x^l,q^i\rangle)(E^ix^l)\otimes(E^ix^l)\otimes(E^ix^l)\right\|_{\{12\}\{3\}}$$

using Lemma B.8.5

$$\begin{aligned}
&\leq \left\|\sum_{l,m=1}^{N_D}\langle x^l,x^m\rangle\left((x^l)(x^m)^T\right)^2\right\|^{1/2} = \left\|\sum_{l,m=1}^{N_D}\langle x^l,x^m\rangle^2(x^l)(x^m)^T\right\|^{1/2}\\
&\leq \left\|\frac{1}{2}\sum_{l,m=1}^{N_D}\langle x^l,x^m\rangle^2[(x^l)(x^l)^T+(x^m)(x^m)^T]\right\|^{1/2}\\
&= \left\|\sum_{l=1}^{N_D}\left[(x^l)(x^l)^T\left(\sum_{m=1}^{N_D}\langle x^l,x^m\rangle^2\right)\right]\right\|^{1/2}.
\end{aligned}$$

The other terms in the sum can be bounded similarly as

$$\begin{aligned}
&\left\|\sum_{i,j=1}^{m}(S(x,q))^{-2}\exp(\langle x,q^i\rangle+\langle x,q^j\rangle)\cdot\right.\\
&\left.\cdot\left((E^ix)\otimes(E^ix)\otimes(E^jx)+(E^ix)\otimes(E^jx)\otimes(E^ix)+(E^ix)\otimes(E^jx)\otimes(E^jx)\right)\right\|_{\{12\}\{3\}}\\
&\leq 3\left\|\sum_{l=1}^{N_D}\left[(x^l)(x^l)^T\left(\sum_{m=1}^{N_D}\langle x^l,x^m\rangle^2\right)\right]\right\|^{1/2},\\
&\left\|-2(S(x,q))^{-3}\sum_{i,j,k=1}^{m}(E^ix)\otimes(E^jx)\otimes(E^kx)\exp(\langle x,q^i\rangle+\langle x,q^j\rangle+\langle x,q^k\rangle)\right\|_{\{12\}\{3\}}\\
&\leq 2\left\|\sum_{l=1}^{N_D}\left[(x^l)(x^l)^T\left(\sum_{m=1}^{N_D}\langle x^l,x^m\rangle^2\right)\right]\right\|^{1/2},
\end{aligned}$$

and the claim follows by the triangle inequality. □

Bibliography

[1] Assyr Abdulle, Gilles Vilmart, and Konstantinos C. Zygalakis. Long time accuracy of Lie–Trotter splitting methods for Langevin dynamics. *SIAM Journal on Numerical Analysis*, 53(1):1–16, 2015.

[2] Steven A. Adelman and Jimmie D. Doll. Generalized Langevin equation approach for atom/solid-surface scattering: General formulation for classical scattering off harmonic solids. , 64(6):2375–2388, March 1976.

[3] Alfonso Alamo and Jesús María Sanz-Serna. A technique for studying strong and weak local errors of splitting stochastic integrators. *SIAM Journal on Numerical Analysis*, 54(6):3239–3257, 2016.

[4] Jason M. Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2169–2176, 2023.

[5] Simon Apers, Sander Gribling, and Dániel Szilágyi. Hamiltonian Monte Carlo for efficient Gaussian sampling: long and random steps. *arXiv preprint arXiv:2209.12771*, 2022.

[6] Elsiddig Awadelkarim, Ajay Jasra, and Hamza Ruzayqat. Unbiased parameter estimation for partially observed diffusions. *arXiv preprint arXiv:2309.10589*, 2023.

[7] Daniel Azagra, Juan Ferrera, Fernando López-Mesas, and Yenny Rangel. Smooth approximation of Lipschitz functions on Riemannian manifolds. *Journal of Mathematical Analysis and Applications*, 326(2):1370–1378, 2007.

[8] Jack Baker, Paul Fearnhead, Emily B. Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Stat. Comput.*, 29(3):599–615, 2019.

[9] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84*, volume 1123 of *Lecture Notes in Math.*, pages 177–206. Springer, Berlin, 1985.

[10] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, Cham, 2014.

[11] Fabrice Baudoin. Wasserstein contraction properties for hypoelliptic diffusions. *arXiv preprint arXiv:1602.04177*, 2016.

[12] Fabrice Baudoin. Bakry-Émery meet Villani. *J. Funct. Anal.*, 273(7):2275–2291, 2017.

[13] Andrea Bertazzi, Joris Bierkens, and Paul Dobson. Approximations of piecewise deterministic Markov processes and their convergence properties. *Stochastic Process. Appl.*, 154:91–153, 2022.

[14] Andrea Bertazzi, Paul Dobson, and Pierre Monmarché. Splitting schemes for second order approximations of piecewise-deterministic Markov processes. *arXiv preprint arXiv:2301.02537*, 2023.

[15] Julian Besag. Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1734–1741, 1994.

[16] Julian E. Besag. Comments on 'Representations of knowledge in complex systems' by U. Grenander and M.I. Miller. *J. Roy. Statist. Soc. Ser. B*, (56):591–592, 1994.

[17] Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

[18] Joris Bierkens, Paul Fearnhead, and Gareth Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Statist.*, 47(3):1288–1320, 2019.

[19] Joris Bierkens, Sebastiano Grazzi, Kengo Kamatani, and Gareth Roberts. The boomerang sampler. In *International conference on machine learning*, pages 908–918. PMLR, 2020.

[20] François Bolley, Arnaud Guillin, and Florent Malrieu. Trend to equilibrium and particle approximation for a weakly selfconsistent Vlasov-Fokker-Planck equation. *M2AN Math. Model. Numer. Anal.*, 44(5):867–884, 2010.

[21] Stephen D. Bond and Benedict J. Leimkuhler. Molecular dynamics and the accuracy of numerically computed averages. *Acta Numer.*, 16:1–65, 2007.

[22] Nawaf Bou-Rabee and Andreas Eberle. Couplings for Andersen dynamics. *Ann. Inst. Henri Poincaré Probab. Stat.*, 58(2):916–944, 2022.

[23] Nawaf Bou-Rabee and Andreas Eberle. Mixing time guarantees for unadjusted Hamiltonian Monte Carlo. *Bernoulli*, 29(1):75–104, 2023.

[24] Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *Ann. Appl. Probab.*, 30(3):1209–1250, 2020.

[25] Nawaf Bou-Rabee and Tore Selland Kleppe. Randomized Runge-Kutta-Nyström. *arXiv preprint arXiv:2310.07399*, 2023.

[26] Nawaf Bou-Rabee and Milo Marsden. Unadjusted Hamiltonian MCMC with stratified Monte Carlo time integration. *arXiv preprint arXiv:2211.11003*, 2022.

[27] Nawaf Bou-Rabee and Stefan Oberdörster. Mixing of Metropolis-Adjusted Markov Chains via Couplings: The High Acceptance Regime. *arXiv preprint arXiv:2308.04634*, 2023.

[28] Nawaf Bou-Rabee and Houman Owhadi. Long-run accuracy of variational integrators in the stochastic context. *SIAM J. Numer. Anal.*, 48(1):278–297, 2010.

[29] Nawaf Bou-Rabee and Jesús María Sanz-Serna. Randomized Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 27(4):2159 – 2194, 2017.

[30] Alexandre Bouchard-Côté, Sebastian J. Vollmer, and Arnaud Doucet. The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018.

[31] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[32] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[33] Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278, 2018.

[34] Axel Brünger, Charles L. Brooks, and Martin Karplus. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chemical Physics Letters*, 105(5):495–500, 1984.

[35] Giovanni Bussi and Michele Parrinello. Accurate sampling using Langevin dynamics. *Phys. Rev. E*, 75:056707, May 2007.

[36] Evan Camrud, Alain Oliviero Durmus, Pierre Monmarché, and Gabriel Stoltz. Second order quantitative bounds for unadjusted generalized Hamiltonian Monte Carlo. *arXiv preprint arXiv:2306.09513*, 2023.

[37] Yu Cao, Jianfeng Lu, and Lihan Wang. Complexity of randomized algorithms for underdamped Langevin dynamics. *Communications in Mathematical Sciences*, 19:1827–1853, 2021.

[38] 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, Aug 2023.

[39] Neil K. Chada, Jordan Franks, Ajay Jasra, Kody J. Law, and Matti Vihola. Unbiased inference for discretely observed hidden Markov model diffusions. *SIAM/ASA J. Uncertain. Quantif.*, 9(2):763–787, 2021.

[40] 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.

[41] Martin Chak and Pierre Monmarché. Reflection coupling for unadjusted generalized Hamiltonian Monte Carlo in the nonconvex stochastic gradient case. *arXiv preprint arXiv:2310.18774*, 2023.

[42] Subrahmanyan Chandrasekhar. Stochastic problems in Physics and Astronomy. *Rev. Mod. Phys.*, 15:1–89, Jan 1943.

[43] 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.

[44] Louis H. Y. Chen and Adrian Röllin. Stein couplings for normal approximation. *arXiv preprint arXiv:1003.6039*, 2010.

[45] Yuansi Chen and Khashayar Gatmiry. When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm? *arXiv preprint arXiv:2304.04724*, 2023.

[46] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR, 2018.

[47] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

[48] Singo Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-Adjusted Langevin Algorithm. *Proceedings of Machine Learning Research*, 134:1–41, 2021.

[49] Adrien Corenflos, Matthew Sutton, and Nicolas Chopin. Debiasing piecewise deterministic Markov process samplers using couplings. *arXiv preprint arXiv:2306.15422*, 2023.

[50] Rob Cornish, Paul Vanetti, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Scalable Metropolis-Hastings for exact Bayesian inference with large datasets. In *International Conference on Machine Learning*, pages 1351–1360. PMLR, 2019.

[51] Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017.

[52] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):651–676, 2017.

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

[54] Arnak S. Dalalyan, Avetik Karagulyan, and Lionel Riou-Durand. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *J. Mach. Learn. Res.*, 23:Paper No. [235], 38, 2022.

[55] Arnak S. Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

[56] Pierre Del Moral, Shulan Hu, Ajay Jasra, Hamza Ruzayqat, and Xinyu Wang. Bayesian Parameter Inference for Partially Observed Diffusions using Multilevel Stochastic Runge-Kutta Methods. *arXiv preprint arXiv:2309.13557*, 2023.

[57] George Deligiannidis, Daniel Paulin, Alexandre Bouchard-Côté, and Arnaud Doucet. Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates. *Ann. Appl. Probab.*, 31(6):2612–2662, 2021.

[58] Roland L. Dobrushin. Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486, 1970.

[59] Jean Dolbeault, Clément Mouhot, and Christian Schmeiser. Hypocoercivity for kinetic equations with linear relaxation terms. *C. R. Math. Acad. Sci. Paris*, 347(9-10):511–516, 2009.

[60] Jean Dolbeault, Clément Mouhot, and Christian Schmeiser. Hypocoercivity for linear kinetic equations conserving mass. *Trans. Amer. Math. Soc.*, 367(6):3807–3828, 2015.

[61] Randall Douc, Pierre E. Jacob, Anthony Lee, and Dootika Vats. Solving the Poisson equation using coupled Markov chains. *arXiv preprint arXiv:2206.05691*, 2022.

[62] Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

[63] Alain Durmus, Aurélien Enfroy, Éric Moulines, and Gabriel Stoltz. Uniform minorization condition and convergence bounds for discretizations of kinetic Langevin dynamics. *arXiv preprint arXiv:2107.14542*, 2021.

[64] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *J. Mach. Learn. Res.*, 20:Paper No. 73, 46, 2019.

[65] Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

[66] Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[67] Alain Oliviero Durmus and Andreas Eberle. Asymptotic bias of inexact Markov Chain Monte Carlo methods in high dimension. *Ann. Appl. Probab.*, 2024.

[68] Rick Durrett. *Probability: theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fifth edition, 2019.

[69] Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference On Learning Theory*, pages 793–797. PMLR, 2018.

[70] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Ann. Probab.*, 47(4):1982–2010, 2019.

[71] Donald L. Ermak and Helen Buckholz. Numerical integration of the Langevin equation: Monte Carlo simulation. *Journal of Computational Physics*, 35(2):169–182, 1980.

[72] Joshua Finkelstein, Giacomo Fiorin, and Benjamin Seibold. Comparison of modern Langevin integrators for simulations of coarse-grained polymer melts. *Molecular Physics*, 118(6):e1649493, 2020.

[73] James Foster, Terry Lyons, and Harald Oberhauser. The shifted ODE method for underdamped Langevin MCMC. *arXiv preprint arXiv:2101.03446*, 2021.

[74] James M. Foster, Gonçalo dos Reis, and Calum Strange. High order splitting methods for SDEs satisfying a commutativity condition. *SIAM J. Numer. Anal.*, 62(1):500–532, 2024.

[75] Harry Furstenberg and Harry Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, 1960.

[76] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[77] Michael B. Giles. Multilevel Monte Carlo methods. *Acta Numer.*, 24:259–328, 2015.

[78] Michael B. Giles, Mateusz B. Majka, Lukasz Szpruch, Sebastian J. Vollmer, and Konstantinos C. Zygalakis. Multi-level Monte Carlo methods for the approximation of invariant measures of stochastic differential equations. *Stat. Comput.*, 30(3):507–524, 2020.

[79] Peter W. Glynn and Chang-Han Rhee. Exact estimation for markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

[80] Nicolaï Gouraud, Pierre Le Bris, Adrien Majka, and Pierre Monmarché. HMC and underdamped Langevin united in the unadjusted convex smooth case. *arXiv preprint arXiv:2202.00977*, 2023.

[81] David Griffeath. A maximal coupling for Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31(2):95–106, 1975.

[82] Jeremy Heng and Pierre E Jacob. Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2), 2019.

[83] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[84] Roger A. Horn and Fuzhen Zhang. *Basic Properties of the Schur Complement*, pages 17–46. Springer US, Boston, MA, 2005.

[85] Jesús A. Izaguirre, Daniel P. Catarello, Justin Mx. Wozniak, and Robert D. Skeel. Langevin stabilization of molecular dynamics. *The Journal of Chemical Physics*, 114(5):2090–2098, 2001.

[86] Pierre E. Jacob, John O'Leary, and Yves F. Atchadé. Unbiased Markov chain Monte Carlo methods with couplings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(3):543–600, 2020.

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

[88] Andrew Jones and Benedict Leimkuhler. Adaptive stochastic methods for sampling driven molecular systems. *The Journal of chemical physics*, 135(8):084125, 2011.

[89] Aldéric Joulin and Yann Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.*, 38(6):2418–2442, 2010.

[90] Nabil Kahalé. Unbiased time-average estimators for Markov chains. *Mathematics of Operations Research*, 2023.

[91] Vladislav Kargin. Products of random matrices: Dimension and growth in norm. *The Annals of Applied Probability*, pages 890–906, 2010.

[92] Aikaterini Karoni, Benedict Leimkuhler, and Gabriel Stoltz. Friction-adaptive descent: a family of dynamics-based optimization methods. *J. Comput. Dyn.*, 10(4):450–484, 2023.

[93] Hassan K. Khalil. *Nonlinear systems*. Prentice-Hall: Upper Saddle River, New Jersey, 2002.

[94] Siem J. Koopman and Rutger Lit. A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society. Series A*, 178(1):167–186, 2015.

[95] Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. An introduction with applications.

[96] Yann LeCun, Corinna Cortes, Chris Burges, et al. MNIST handwritten digit database, 2010.

[97] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Lower bounds on Metropolized sampling methods for well-conditioned distributions. *Advances in Neural Information Processing Systems*, 34:18812–18824, 2021.

[98] Benedict Leimkuhler and Charles Matthews. Rational construction of stochastic numerical methods for molecular sampling. *Appl. Math. Res. Express. AMRX*, (1):34–56, 2013.

[99] Benedict Leimkuhler and Charles Matthews. Robust and efficient configurational molecular sampling via Langevin dynamics. *Journal of Chemical Physics*, 138:174102, 2013.

[100] Benedict Leimkuhler and Charles Matthews. Molecular dynamics. *Interdisciplinary applied mathematics*, 39:443, 2015.

[101] Benedict Leimkuhler, Charles Matthews, and Gabriel Stoltz. The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA J. Numer. Anal.*, 36(1):13–79, 2016.

[102] Benedict Leimkuhler, Charles Matthews, and Michael V. Tretyakov. On the long-time integration of stochastic gradient systems. *Proc. R. Soc. A.*, 470(2170):20140120, 16, 2014.

[103] Benedict Leimkuhler, Daniel Paulin, and Peter A. Whalley. Contraction Rate Estimates of Stochastic Gradient Kinetic Langevin Integrators. *ESIAM: Mathematical Modelling and Numerical Analysis*, (Accepted), 2024.

[104] Benedict Leimkuhler and Matthias Sachs. Efficient numerical algorithms for the generalized Langevin equation. *SIAM J. Sci. Comput.*, 44(1):A364–A388, 2022.

[105] Benedict J. Leimkuhler, Daniel Paulin, and Peter A. Whalley. Contraction and Convergence Rates for Discretized Kinetic Langevin Dynamics. *SIAM Journal on Numerical Analysis*, 62(3):1226–1258, 2024.

[106] Jianfeng Lu and Lihan Wang. On explicit L^2-convergence rate estimate for piecewise deterministic Markov processes in MCMC algorithms. *The Annals of Applied Probability*, 32(2):1333 – 1361, 2022.

[107] Michael J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[108] Mateusz B. Majka, Aleksandar Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. *Annals of Applied Probability*, 30(4):1534–1581, 2020.

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

[110] Robert I. McLachlan and G. Reinout W. Quispel. Splitting methods. *Acta Numerica*, 11:341–434, 2002.

[111] Simone Melchionna. Design of quasisymplectic propagators for Langevin dynamics. *The Journal of chemical physics*, 127(4):044108, 2007.

[112] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[113] Lawrence Middleton, George Deligiannidis, Arnaud Doucet, and Pierre E. Jacob. Unbiased Markov chain Monte Carlo for intractable target distributions. *Electron. J. Stat.*, 14(2):2842–2891, 2020.

[114] Grigori N. Milstein and Michael V. Tretyakov. *Stochastic numerics for mathematical physics*. Scientific Computation. Springer-Verlag, Berlin, 2004.

[115] Pierre Monmarché. High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electron. J. Stat.*, 15(2):4117–4166, 2021.

[116] Pierre Monmarché. Almost sure contraction for diffusions on $\mathbb{R}^d$. Application to generalized Langevin diffusions. *Stochastic Process. Appl.*, 161:316–349, 2023.

[117] Pierre Monmarché. An entropic approach for Hamiltonian Monte Carlo: the idealized case. *Ann. Appl. Probab.*, 34(2):2243–2293, 2024.

[118] Eike H. Müller, Rob Scheichl, and Tony Shardlow. Improving multilevel Monte Carlo for stochastic differential equations with application to the Langevin equation. *Proc. A.*, 471(2176):20140679, 20, 2015.

[119] Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 113–162. CRC Press, Boca Raton, FL, 2011.

[120] Christopher Nemeth and Paul Fearnhead. Stochastic gradient Markov chain Monte Carlo. *Journal of the American Statistical Association*, 116(533):433–450, 2021.

[121] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[122] Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[123] Daniel Paulin and Peter A. Whalley. Correction to "Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations". *arXiv preprint arXiv:2402.08711*, 2024.

[124] Grigorios A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

[125] Elias A. J. F. Peters and Gijsbertus de With. Rejection-free Monte Carlo sampling for general potentials. *Physical Review E*, 85(2):026703, 2012.

[126] Qian Qin and James P. Hobert. Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions. volume 58, pages 872–889, 2022.

[127] Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding up MCMC by efficient data subsampling. *J. Amer. Statist. Assoc.*, 114(526):831–843, 2019.

[128] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[129] Chang-Han Rhee and Peter W. Glynn. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015.

[130] Lewis F. Richardson. The approximate arithmetical solution by finite differences with an application to stresses in masonry dams. *Philosophical Transactions of the Royal Society of America*, 210:307–357, 1911.

[131] Lionel Riou-Durand and Jure Vogrinc. Metropolis Adjusted Langevin Trajectories: a robust alternative to Hamiltonian Monte Carlo. *arXiv preprint arXiv:2202.13230*, 2022.

[132] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[133] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

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

[135] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

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

[137] Peter J. Rossky, Jimmie D. Doll, and Harold L. Friedman. Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

[138] Hamza Ruzayqat, Neil K. Chada, and Ajay Jasra. Unbiased estimation using underdamped Langevin dynamics. *SIAM J. Sci. Comput.*, 45(6):A3047–A3070, 2023.

[139] Jesús María Sanz Serna and Konstantinos C. Zygalakis. Contractivity of Runge–Kutta methods for convex gradient systems. *SIAM Journal on Numerical Analysis*, 58(4):2079–2092, 2020.

[140] Jesus María Sanz-Serna and Konstantinos C. Zygalakis. Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *J. Mach. Learn. Res.*, 22:Paper No. 242, 37, 2021.

[141] Katharina Schuh. Global contractivity for Langevin dynamics with distribution-dependent forces and uniform in time propagation of chaos. *Ann. Inst. Henri Poincaré Probab. Stat.*, 2024.

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

[143] Inass Sekkat and Gabriel Stoltz. Quantifying the mini-batching error in Bayesian inference for adaptive Langevin dynamics. *J. Mach. Learn. Res.*, 24:Paper No. [329], 58, 2023.

[144] Mar Serrano, Gianni De Fabritiis, Pep Español, and Peter V. Coveney. A stochastic Trotter integration scheme for dissipative particle dynamics. *Math. Comput. Simulation*, 72(2-6):190–194, 2006.

[145] Tony Shardlow. Splitting for dissipative particle dynamics. *SIAM J. Sci. Comput.*, 24(4):1267–1282, 2003.

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

[147] Chunmei Shi, Yu Xiao, and Chiping Zhang. The convergence and MS stability of exponential Euler method for semilinear stochastic differential equations. *Abstr. Appl. Anal.*, pages Art. ID 350407, 19, 2012.

[148] Robert D. Skeel. Integration schemes for molecular dynamics and related applications. *The Graduate Student's Guide to Numerical Analysis' 98: Lecture Notes from the VIII EPSRC Summer School in Numerical Analysis*, pages 119–176, 1999.

[149] Robert D. Skeel and Jesüs A. Izaguirre. An impulse integrator for Langevin dynamics. *Molecular Physics*, 100(24):3885–3891, 2002.

[150] J. Michael Steele. An Efron-Stein inequality for nonsymmetric statistics. *Ann. Statist.*, 14(2):753–758, 1986.

[151] 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.

[152] Adam Telatovich and Xiantao Li. The strong convergence of operator-splitting methods for the Langevin dynamics model. *arXiv preprint arXiv:1706.04237*, 2017.

[153] Fabrice Thalmann and Jean Farago. Trotter derivation of algorithms for Brownian and dissipative particle dynamics. *The Journal of Chemical Physics*, 127(12):124109, 09 2007.

[154] Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G. Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, and Christopher Yau. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1, 2021.

[155] Leonid Vaserstein. Markovian processes on countable space product describing large systems of automata. *Probl. Peredachi Inf*, 5(3):64–72, 1969.

[156] Dootika Vats, James M. Flegal, and Galin L. Jones. Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337, 2019.

[157] Matti Vihola. Unbiased estimators and multilevel Monte Carlo. *Operations Research*, 66(2):448–462, 2018.

[158] Cédric Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950):iv+141, 2009.

[159] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[160] Sebastian J. Vollmer, Konstantinos C. Zygalakis, and Yee Whye Teh. Exploration of the (non-)asymptotic bias and variance of stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.*, 17:Paper No. 159, 45, 2016.

[161] Maxime Vono, Daniel Paulin, and Arnaud Doucet. Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting. *J. Mach. Learn. Res.*, 23:Paper No. [25], 69, 2022.

[162] Tianze Wang and Guanyang Wang. Unbiased multilevel Monte Carlo methods for intractable distributions: MLMC meets MCMC. *J. Mach. Learn. Res.*, 24:Paper No. [249], 40, 2023.

[163] Christian H. Weiß, Fukang Zhu, and Aisouda Hoshiyar. Softplus ingarch models. *Statistica Sinica*, 32(2):1099–1120, 2022.

[164] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

[165] Paul F. V. Wiemann, Thomas Kneib, and Julien Hambuckers. Using the softplus function to construct alternative link functions in generalized linear models and beyond. *Statistical Papers*, pages 1–26, 2023.

[166] Tim Zajic. Non-asymptotic error bounds for scaled underdamped Langevin MCMC. *arXiv preprint arXiv:1912.03154*, 2019.

[167] Shunshi Zhang, Sinho Chewi, Mufan Li, Krishna Balasubramanian, and Murat A. Erdogdu. Improved discretization analysis for underdamped Langevin Monte Carlo. In *Conference On Learning Theory*, pages 36–71. PMLR, 2023.

[168] Hu Zhengmian, Feihu Huang, and Heng Huang. Optimal Underdamped Langevin MCMC Method. *Advances in Neural Information Processing Systems*, 34:19363–19374, 2021.

[169] Alfonso Álamo Zapatero. *Word Series for the Numerical Integration of Stochastic Differential Equations*. PhD thesis, Universidad de Valladolid, 2021.