Stochastic Gradient Leimkuhler–Matthews Integrator with Second Order Discretization Error

Xuda Ye[1], Yuliang Wang[2], Guang Lin[1], and Cindy Xiangrui Kong[1]

[1]Department of Mathematics, Purdue University, West Lafayette, IN 47907
[2]Department of Mathematics, Duke University, Durham, NC 27708

October 26, 2026

Abstract

Stochastic gradient Langevin dynamics (SGLD) has a discretization error of first order in the step size h, even with the full gradient. We study the stochastic gradient Leimkuhler–Matthews (SGLM) integrator, which has the same cost as SGLD, one stochastic gradient per step, and reuses each Gaussian vector in consecutive steps. With the full gradient, SGLM is the Leimkuhler–Matthews (LM) integrator, whose discretization error for the invariant distribution is of second order. For a potential strongly convex outside a ball, we prove the bound $\mathcal{O}(\frac{1}{Nh} + \Sigma^2 h^2 + h^4)$ on the mean square error (MSE) of time averages of SGLM over N steps, where Σ bounds the variance of the gradient noise. The proof uses a modified class that absorbs the non-Markovian part of SGLM. The discretization error enters this bound as h^4 and the MSE bound of SGLD as h^2, while a larger batch reduces only the term $\Sigma^2 h^2$ in both bounds. In Bayesian logistic regression on MNIST, SGLM has a smaller estimated MSE than SGLD. The coefficient of h in the numerical bias of SGLM is given explicitly by the covariance of the gradient noise and the continuous Poisson equation. The coefficient for SGLD contains in addition the Hessian of the potential, which comes from the discretization error. Through Riesz representors, the coefficients of the numerical bias determine the coefficients of the Wasserstein-1 errors of SGLM and SGLD.

Keywords: overdamped Langevin dynamics, stochastic gradient, Leimkuhler–Matthews integrator, mean square error, discrete Poisson equation, numerical bias, Riesz representor, Wasserstein-1 distance.

MSC 2020: 65C30, 60H35, 65C40.

1 Introduction

Sampling from the Boltzmann distribution $\pi(x) \propto e^{-U(x)}$ on $\mathbb{R}^d$ is a central task in statistical physics [57, 63] and in Bayesian inference [27, 33, 103]. The simplest stochastic differential equation that leaves π invariant is the overdamped Langevin dynamics

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad X_0 = x \in \mathbb{R}^d, \tag{1.1}$$

where $(B_t)_{t\geqslant 0}$ is a standard Brownian motion in $\mathbb{R}^d$ and U a confining potential; time averages along the solution $(X_t)_{t\geqslant 0}$ approximate expectations under π in the long-time limit [53, 90]. A numerical integrator of (1.1) with step size $h > 0$ produces samples $(Z_k)_{k\geqslant 0}$ that approximate the invariant distribution π; such integrators form the basis of Langevin Monte Carlo [26, 27, 33, 34, 73, 100]. Characterizing the accuracy of these samples is a basic question of numerical analysis.

The mean square error (MSE) of time averages is a natural criterion for this accuracy. For a test function f, with $\pi(f) := \int_{\mathbb{R}^d} f\,\mathrm{d}\pi$, the MSE of the time average over N steps is

$$\mathrm{MSE}(N,h) := \mathbb{E}\left[\left|\frac{1}{N}\sum_{k=0}^{N-1} f(Z_k) - \pi(f)\right|^2\right]. \tag{1.2}$$

The Wasserstein distance, the Kullback–Leibler (KL) divergence and the total variation distance compare the law of the samples with π, and are hard to evaluate in high dimension. The MSE instead concerns a single numerical trajectory, and accounts for both the fluctuation of the time average and the numerical bias, the deviation of its long-time limit from $\pi(f)$ [18, 77, 99, 102].

In Bayesian inference on large data sets, ∇U is replaced by an unbiased stochastic estimator $\nabla \hat{U}(x;\xi)$ computed from a mini-batch, with ξ a random variable independent of the Brownian motion [18, 99, 102, 103]. The standard sampler of this kind is stochastic gradient Langevin dynamics (SGLD) [103], the Euler–Maruyama (EM) integrator [50, 80] of (1.1) with the stochastic gradient,

$$Z_{k+1} = Z_k - h\,\nabla \hat{U}(Z_k;\xi_k) + \sqrt{2h}\,\chi_{k+1}, \tag{SGLD}$$

where $(\chi_k)_{k\geqslant 0}$ is a sequence of independent standard Gaussian vectors in $\mathbb{R}^d$ and ξ_k is drawn afresh at every step. The discretization error of (SGLD) is of first order in h: its numerical bias is of order h even with the full gradient (Section 2.2). The existing stochastic gradient integrators that preserve a second order discretization error are built on stochastic gradient Hamiltonian Monte Carlo (SGHMC) [22] with a symmetric splitting integrator [18] or on the underdamped Langevin dynamics [44], whose splitting integrators are of second order with one gradient evaluation per step [81]; the underdamped dynamics with stochastic gradients is studied in [58, 71, 76, 88].

We study the *stochastic gradient Leimkuhler–Matthews* (SGLM) integrator, which has the cost of (SGLD), one stochastic gradient per step, and differs from it in only one respect: the Gaussian vector χ_{k+1} is replaced by the average of χ_k and χ_{k+1}, so that each Gaussian vector is used twice, in consecutive steps,

$$Z_{k+1} = Z_k - h\,\nabla \hat{U}(Z_k;\xi_k) + \sqrt{\frac{h}{2}}\,(\chi_k + \chi_{k+1}). \tag{SGLM}$$

With the full gradient, SGLM is the Leimkuhler–Matthews (LM) integrator [56, 101],

$$Z_{k+1} = Z_k - h\,\nabla U(Z_k) + \sqrt{\frac{h}{2}}\,(\chi_k + \chi_{k+1}). \tag{LM}$$

The reuse of the Gaussian vector gives LM a second order discretization error for the invariant distribution π [56, 59, 101], although its samples $(Z_k)_{k\geqslant 0}$ do not form a Markov chain and its weak error on finite time intervals is of first order. LM evaluates ∇U once per step, while almost all other integrators of (1.1) with a second order discretization error are of predictor–corrector type, with at least two sequential evaluations of ∇U per step (Section 2.1). With a stochastic gradient, each of these evaluations carries its own noise. This raises the question whether SGLM preserves the second order discretization error of LM.

We provide a positive answer to this question. Our contributions are summarized as follows:

1. SGLM separates the stochastic gradient error from the discretization error.

Under regularity conditions on U, f and the stochastic gradient that require strong convexity of U only outside a ball, we prove in Theorem 3 the MSE bound for SGLM,

$$\text{(SGLM, variance } \Sigma) \qquad \mathrm{MSE}(N,h) \leqslant C\Big(\frac{1}{Nh} + \Sigma^2 h^2 + h^4\Big), \tag{1.3}$$

where Σ bounds the variance of the stochastic gradient, and C depends on the regularity constants of U and f, the dimension d and the initial value Z_0, but not on N, h or Σ. In (1.3), $\frac{1}{Nh}$ comes from the fluctuation of the time average, $\Sigma^2 h^2$ from the square of the stochastic gradient error, and h^4 from the square of the discretization error. The same separation holds for the UBU integrator with stochastic gradients (SG-UBU) [71], and was previously known only for integrators of the underdamped dynamics [71, 87]. The (SGLD) integrator does not have it, since its stochastic gradient error and its discretization error are both of first order in h. With the full gradient, $\Sigma = 0$ and (1.3) yields the MSE bound $\frac{1}{Nh} + h^4$ of LM. To our knowledge, this is also the first MSE bound for the LM integrator.

The separation in (1.3) is advantageous in the mini-batch regime: if each step averages P independent stochastic gradients, the variance bound Σ is replaced by Σ/P, so the contribution of the gradient noise to

the MSE becomes $\Sigma^2 h^2/P^2$, with C in (1.3) independent of P. Increasing the batch size P at a fixed step size h therefore reduces the stochastic gradient error. By contrast, the MSE of (SGLD) keeps the term h^2 from the square of its discretization error at every batch size, and only a smaller h reduces it. Since the gradients of a batch are evaluated in parallel while the steps are sequential, SGLM with a large batch size P has a smaller error bound than SGLD under the same cost.

The proof of (1.3) runs on the Markovian form of (SGLM), a Markov chain $(Y_k)_{k\geqslant 0}$ with $Z_k = Y_k + \sqrt{h/2}\,\chi_k$. It builds on the discrete Poisson equation of our earlier work [71, 106], combined with a modified class of test function, potential, generator and discrete Poisson solution that absorbs the non-Markovian part of the integrator, the Gaussian shift $\sqrt{h/2}\,\chi_k$.

2. The coefficients of the numerical bias are explicit.

The time averages of f along (SGLM) converge almost surely to $\pi_h(f)$, for a numerical invariant distribution π_h independent of the initial value, and the numerical bias $\pi_h(f)-\pi(f)$ is of order h with an explicit coefficient given in Corollary 2,

$$\text{(SGLM, numerical bias)} \qquad \lim_{h\to 0} \frac{\pi_h(f)-\pi(f)}{h} = \frac{1}{2}\,\pi\big(\nabla^2\phi_0(x) : \mathcal{C}(x)\big), \tag{1.4}$$

where $\mathcal{C}(x)$ is the covariance of the gradient noise $\nabla\hat{U}(x;\xi) - \nabla U(x)$, ϕ_0 solves the continuous Poisson equation (3.7), $-\mathcal{L}\phi_0 = f - \pi(f)$ with $\mathcal{L}$ the generator of (1.1), and : is the full contraction of two matrices. The coefficient comes entirely from the gradient noise: with the full gradient, $\mathcal{C}(x) = 0$ and the numerical bias of LM has no linear part. The gradient noise enters only through the local weak error of a single step, through its covariance $\mathcal{C}(x)$, as for SG-UBU in [71, Section 3.4].

The discrete Poisson equation also characterizes the numerical bias of (SGLD). Its time averages converge to $\bar{\pi}_h(f)$, with $\bar{\pi}_h$ the invariant distribution of SGLD, and by Corollary 3

$$\text{(SGLD, numerical bias)} \qquad \lim_{h\to 0} \frac{\bar{\pi}_h(f)-\pi(f)}{h} = \frac{1}{2}\,\pi\Big(\nabla^2\phi_0(x) : \big(\nabla^2 U(x) + \mathcal{C}(x)\big)\Big), \tag{1.5}$$

a coefficient studied formally for $d = 1$ in [102]. Besides the covariance $\mathcal{C}(x)$, the coefficient of SGLD carries the Hessian $\nabla^2 U(x)$, which comes from the discretization error of EM. With the full gradient, $\mathcal{C}(x) = 0$ and (1.5) gives the coefficient of EM; with a batch of P independent stochastic gradients, the coefficient of SGLM is divided by P, while that of SGLD tends to the coefficient of EM as P grows. In (1.4) and (1.5), the gradient noise and the discretization error both enter linearly through $\nabla^2\phi_0$, so the solution ϕ_0 of the continuous Poisson equation is an essential structure of the numerical bias of the two stochastic gradient integrators, not only a tool for upper bounds on the MSE.

3. The coefficients of the Wasserstein-1 error are explicit.

The suprema of the absolute values of the coefficients (1.4) and (1.5) over the test functions with $\|\nabla f\|_\infty \leqslant 1$ are the limits of the normalized Wasserstein-1 errors $\mathcal{W}_1(\pi_h,\pi)/h$ and $\mathcal{W}_1(\bar{\pi}_h,\pi)/h$ as $h\to 0$ (Corollary 4):

$$\text{(SGLM, } \mathcal{W}_1 \text{ coefficient)} \qquad \lim_{h\to 0} \frac{\mathcal{W}_1(\pi_h,\pi)}{h} = \frac{1}{2}\,\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}, \tag{1.6}$$

$$\text{(SGLD, } \mathcal{W}_1 \text{ coefficient)} \qquad \lim_{h\to 0} \frac{\mathcal{W}_1(\bar{\pi}_h,\pi)}{h} = \frac{1}{2}\,\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}}, \tag{1.7}$$

where $\psi_{\mathcal{C}}$ and $\psi_{\nabla^2 U}$ are the Riesz representors of $\mathcal{C}(x)$ and $\nabla^2 U(x)$ defined in (6.14), and $\|\cdot\|_{\mathrm{KR}}$ is the Kantorovich–Rubinstein norm in Section 3.1. The normalized biases $(\pi_h-\pi)/h$ and $(\bar{\pi}_h-\pi)/h$ converge to $\frac{1}{2}\psi_{\mathcal{C}}\,\pi$ and $\frac{1}{2}(\psi_{\nabla^2 U}+\psi_{\mathcal{C}})\,\pi$ in this norm at the rate $h^{\frac{1}{2}}(1+\log\frac{1}{h})$ (Theorem 7). SGLM does not always have a smaller coefficient of the Wasserstein-1 error than SGLD: for a gradient noise of large amplitude supported away from the minimizer of U, the order is reversed (Lemma 10). With a batch of P independent stochastic gradients, $\psi_{\mathcal{C}}/P$ replaces $\psi_{\mathcal{C}}$ in both limits, with $\mathcal{C}(x)$ the covariance of a single stochastic gradient. For a strongly convex potential, the coefficient of SGLM is at most that of SGLD once the batch size reaches a threshold linear in $\sup_x\|\mathcal{C}(x)\|$ (Theorem 8).

Summary of the results

Three features distinguish these results. First, SGLM separates the stochastic gradient error from the discretization error at the cost of SGLD, so that the batch size and the step size control the two errors independently; for an integrator of the overdamped dynamics, this separation is new.

Second, the results on the numerical bias are equalities, not only upper bounds. The coefficients (1.4) and (1.5) identify the numerical biases of SGLM and SGLD, and (1.6) and (1.7) their Wasserstein-1 errors, to first order in h, uniformly over the Lipschitz test functions for the Wasserstein-1 errors. The mainstream coupling methods bound these errors from above without identifying their coefficients; here the coefficients are the Kantorovich–Rubinstein norms of the signed measures given by the Riesz representors, and the shift of the invariant distribution caused by the gradient noise and by the discretization is characterized exactly.

Third, the exact coefficients make the comparison of SGLM with SGLD precise. The coefficient of SGLD keeps the Hessian of the potential at every batch size, while that of SGLM is divided by the batch size. SGLM can be less accurate than SGLD at small batch sizes (Lemma 10), and, for a strongly convex potential, Theorem 8 gives an explicit threshold of the batch size beyond which SGLM is more accurate than SGLD in the Wasserstein-1 distance, to first order in h. This threshold does not depend explicitly on the dimension d; it depends only on the eigenvalues of $\nabla^2 U(x)$ and $\mathcal{C}(x)$.

As applications of the analysis of the coefficients, Section 7 describes three approaches to bias correction. The numerical bias of SGLM is $\frac{h}{2}\,\pi(f\psi_{\mathcal{C}})$ to first order in h, so it is removed once $\pi(f\psi_{\mathcal{C}})$ is computed, with the explicit representor $\psi_{\mathcal{C}}$ for the covariances of Table 3 or by Richardson–Romberg extrapolation in the step size. The third approach, the shifting correction, replaces a sample $Z \sim \pi_h$ of SGLM by $Z + h\nabla\varphi_{\mathcal{C}}(Z)$, where $\varphi_{\mathcal{C}}$ solves (7.5), and formally removes the first order bias coefficient for all test functions at once. The function $\varphi_{\mathcal{C}}$ is approximated by minimizing the energy (7.8), which requires only the samples of SGLM at the step sizes h and $2h$, so the shifting correction is completely data-driven.

In the numerical experiments of Section 8, the evaluated coefficients of the numerical bias of SGLM and SGLD in a 4D Gaussian mixture distribution approach the exact coefficients (1.4) and (1.5) as h decreases, and the evaluated coefficients of the Wasserstein-1 errors in a radial distribution approach those of (1.6) and (1.7). In a 16D product of ring distributions, the approaches to bias correction of Section 7 reduce the numerical bias of SGLM. In Bayesian logistic regression on MNIST, SGLM has a smaller estimated MSE than SGLD with the same step size and batch size.

Section 2 reviews the related works. Section 3 states the assumptions, the Markovian forms (LM') and (SGLM'), and the estimates of the discrete Poisson equation. Section 4 constructs the modified class, Section 5 estimates the local errors of the LM and SGLM integrators against the modified dynamics, and Section 6 states the main results. Section 7 describes the approaches to bias correction, and Section 8 presents the numerical experiments. Section 9 concludes, and the proofs are in the appendices.

2 Related works

This section reviews the work related to (LM) and (SGLM). Section 2.1 reviews the integrators of (1.1) with a second order discretization error and the works on the LM integrator. Section 2.2 reviews the sampling error of (SGLD), and Section 2.3 the stochastic gradient integrators of the underdamped Langevin dynamics.

2.1 Second order integrators in the overdamped case

Among the integrators of (1.1) that evaluate ∇U alone, only a few have a discretization error of order h^2 for the invariant distribution. They include LM [56, 101] and stochastic Heun [29, 75, 80]. The large family of stochastic Runge–Kutta (SRK) integrators [29, 69, 91, 105] contains further examples, such as SRK-I/II [105, 106]. Beyond LM, almost all these integrators are of predictor–corrector type: the second evaluation of ∇U in a step is taken at a state computed from the first. With a stochastic gradient, each of these evaluations carries its own noise, while LM evaluates ∇U once per step and extends to stochastic gradients by replacing that single evaluation; this is the reason we study LM and its stochastic gradient form SGLM. A detailed comparison of these integrators is given in [106]. Integrators of high order for the invariant distribution that also use derivatives of ∇U have been constructed [1, 97].

The LM integrator was introduced in [56] as the high friction limit of the BAOAB splitting of the underdamped Langevin dynamics. Its second order discretization error for the invariant distribution was proved by Talay–Tubaro expansions [98]. These expansions were derived for the weak error, whose first order coefficient decays exponentially in time [59], and for the invariant distribution of splitting integrators on the torus [60]. The second order discretization error of LM was also proved by writing the integrator as the Markovian form (LM') followed by the Gaussian shift $\sqrt{h/2}\,\chi_k$ as a postprocessor [101]. The postprocessor argument uses the sufficient conditions for high order on the invariant distribution of [1], which rest on the Talay–Tubaro expansion [98] and on the weak backward error analysis of [30, 51]. In high dimension and under weak or sparse interactions among the variables, the bias of the unadjusted underdamped Langevin algorithm on marginals of a few coordinates depends on the dimension only through logarithmic factors [24], extending the delocalization of bias of the unadjusted overdamped Langevin algorithm [23]. This delocalization holds for all large friction parameters and therefore carries over to (LM). These works control the numerical bias of LM; none of them bounds the MSE (1.2) or treats stochastic gradients.

The modified class of this paper serves a purpose different from that of the modified equations of [1], which are used to construct integrators. Here the integrator is fixed, and the modified potential of Section 4 defines the modified dynamics (4.16), against which the Markovian form (LM') has a local weak error of order h^3. The modified class also provides the modified discrete Poisson solution, to which the MSE analysis of [71, 106] is applied. The order h^3 of this local weak error appears in the Talay–Tubaro expansions [59, 60], as the vanishing at equilibrium of its coefficient of order h, and in the postprocessor argument [101]. The expansions in [59] rely on the exponential decay in time of the derivatives of the Kolmogorov solution of Section 3.3, for which they refer to [97]. Theorem 1 proves this decay on $\mathbb{R}^d$ for a potential satisfying Assumption 1, extending [106, Theorem 1] to derivatives of order 4.

2.2 Sampling error analysis for SGLD

The (SGLD) integrator converges weakly to the solution of (1.1) as its step size tends to zero [93]. Its sampling error has been analyzed along the following lines, and Table 1 collects representative bounds; see [84] for a review.

- Recurrence relations of the integrator, combined with strong or weak error estimates, give nonasymptotic bounds in Wasserstein distance [7, 9, 17, 27, 82, 96, 107]. Weighted transportation cost inequalities give bounds in $\mathcal{W}_2$ for nonconvex potentials [89]. Under log-concavity at infinity the variance of the stochastic gradient does not change the rate of weak convergence [73], and a conductance analysis gives bounds in total variation for targets that are not log-concave [112]. For targets log-concave outside a compact set, the law of the k-th iterate of SGLD converges geometrically in $\mathcal{W}_1$ to the numerical invariant distribution $\bar{\pi}_h$ of SGLD [66].
- The continuous Poisson equation (3.7) controls the error of the time averages of numerical schemes [10, 77] and of SGLD and SGHMC, including their MSE [18, 67, 99, 102, 104]. These results require its solution ϕ_0 to have bounded or polynomially growing derivatives.
- The KL divergence between the laws of SGLD and of (1.1) is bounded uniformly in time [65]. Bounds in KL divergence also follow from convex optimization on the Wasserstein space [36] and, under a logarithmic Sobolev inequality, from the nearly Gaussian structure of the gradient noise [28].
- Variance-reduced and control-variate estimators of the gradient improve these guarantees [4, 15, 32, 49, 110]. The numerical bias of SGLD is reduced by Richardson–Romberg extrapolation [35], and its cost at a prescribed accuracy is reduced by multilevel Monte Carlo [42, 43]. The extrapolation, however, introduces negative weights and does not increase the accuracy of the trajectory itself. The cost of the gradient noise is quantified by oracle lower bounds [16], by the dependence of the MSE on the mini-batch size under a fixed budget [19], and by a comparison of SGLD with EM at a prescribed mean square accuracy [74].

The numerical bias of (SGLD) is of first order in h. The bias of the time average over N steps is $\mathcal{O}\big((Nh)^{-1} + h\big)$ [102, Theorem 9]. The uniform-in-time KL bound of [65, Corollary 3.1], of order h^2, combined with the geometric ergodicity of SGLD for a potential strongly convex outside a ball [66], yields $\mathcal{W}_1(\bar{\pi}_h, \pi) \leqslant Ch$ [65, Section 5.1]. Corollary 4 gives the exact coefficients of h in $\mathcal{W}_1(\bar{\pi}_h, \pi)$ and $\mathcal{W}_1(\pi_h, \pi)$,

reference	assumption	sampling error
Theorem 5 [27]	global	$\mathcal{W}_2(\mu_k, \pi) \lesssim e^{-\lambda k h} + h + \Sigma^{\frac{1}{2}} h^{\frac{1}{2}}$
Theorem 5, Corollary 12 [82]	global	$\mathbb{E}\lvert X_{kh} - Z_k\rvert^2 \lesssim e^{-\lambda k h} + h$
Corollary 18 [36]	global	$\mathrm{KL}(\bar{\mu}_N \Vert \pi) \lesssim (Nh)^{-1} + h$
Theorem 3.8 [96]	global	$\mathcal{W}_2(\mu_k, \pi) \lesssim e^{-\lambda k h} + h^{\frac{1}{2}}$
Proposition 1, Theorem 1 [86]	global	$\mathcal{W}_2(\mu_k, \pi) \lesssim e^{-\lambda k h} + h$
Proposition 10 [89]	non-global	$\mathcal{W}_2(\mu_k, \pi) \lesssim (\delta^{\frac{1}{4}} + h^{\frac{1}{4}})\, kh + e^{-\lambda k h}$
Theorem 4.5 [112]	non-global	$\Vert \mu_k - \pi \Vert_{\mathrm{TV}} \lesssim e^{-\lambda k h} + h^{\frac{1}{2}} + \epsilon$
Theorem 9 [102]	non-global	$\mathrm{MSE}(N, h) \lesssim (Nh)^{-1} + h^2$
Theorem 3 [18]	non-global	$\mathrm{MSE}(N, h) \lesssim (Nh)^{-1} + h^2$
Corollaries 3.1 and 5.1 [65]	log-Sobolev	$\mathrm{KL}(\mu_k \Vert \nu_k) \lesssim h^2$, $\mathcal{W}_2(\mu_k, \pi) \lesssim e^{-\lambda k h} + h$
Theorem 5 [28]	log-Sobolev	$\mathrm{KL}(\mu_k \Vert \pi) \lesssim e^{-\lambda k h} + h$
SGLM, Theorem 3	non-global	$\mathrm{MSE}(N, h) \lesssim (Nh)^{-1} + \Sigma^2 h^2 + h^4$
SGLM, Corollary 4	non-global	$\lim_{h \to 0} \frac{\mathcal{W}_1(\pi_h, \pi)}{h} = \frac{1}{2} \Vert \psi_{\mathcal{C}}\, \pi \Vert_{\mathrm{KR}}$
SGLD, Corollary 4	non-global	$\lim_{h \to 0} \frac{\mathcal{W}_1(\bar{\pi}_h, \pi)}{h} = \frac{1}{2} \Vert (\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\, \pi \Vert_{\mathrm{KR}}$

Table 1: Sampling error results for (SGLD) and for (SGLM). In the assumption column, global marks a strongly convex or convex potential, non-global a dissipative potential, one convex outside a ball, or, in [18, 102], bounds on the solution of the Poisson equation with moment conditions in place of convexity, and log-Sobolev a target satisfying a logarithmic Sobolev inequality; the references impose in addition conditions on the regularity, the initial distribution and the moments of the gradient noise. Here μ_k denotes the law of the k-th iterate of SGLD, $\bar{\pi}_h$ and π_h the numerical invariant distributions of SGLD and of SGLM, $\bar{\mu}_N$ the average of $\mu_1, \dots, \mu_N$, and ν_k the law of the solution of (1.1) at time kh. Further, X_{kh} is the solution of (1.1) coupled with the k-th iterate Z_k of SGLD, δ the relative variance of the gradient noise in [89], and $\epsilon \in (0, 1)$ the truncation parameter in [112], which fixes the radius of the truncated target on which the constants of that bound depend. Finally, Σ is the bound on the variance of the gradient noise in Assumption 3, $\psi_{\mathcal{C}}$ and $\psi_{\nabla^2 U}$ the Riesz representors (6.14) of its covariance $\mathcal{C}(x)$ in (5.8) and of $\nabla^2 U(x)$, $\Vert \cdot \Vert_{\mathrm{KR}}$ the Kantorovich–Rubinstein norm, $\mathcal{W}_p$ the Wasserstein distance of order p, and $\lambda > 0$ a rate of convergence. Other constants, depending on the dimension, the potential and the gradient noise, are suppressed.

determined through Riesz representors by the bias coefficients of Corollaries 2 and 3. For a strongly convex potential with a Lipschitz gradient and gradient noise that is Lipschitz in mean square, $\mathcal{W}_2(\bar{\pi}_h, \pi) \leqslant Ch$ when the gradient noise has a finite moment of order 4, while under a finite moment of order 2 only, the bound is of order $h^{\frac{1}{2}}$ [86, Theorem 1]. A family of spiked gradient noises shows that the order $h^{\frac{1}{2}}$ cannot be improved uniformly over gradient noises of a fixed variance [86, Lemma 1]. Each noise of this family has finite moments of all orders, while their moments of order 12 are unbounded over the family; Assumption 3, whose bound on the moment of order 12 of the stochastic gradient has a fixed constant M, excludes the family as a whole. The first order term has two sources: the gradient noise and the discretization error of EM. The discretization error of EM gives a numerical bias of first order even with the full gradient [98]; see also [102, Section 4.1].

A different line of work removes the numerical bias. The proximal sampler [25, 55], with its restricted Gaussian oracle implemented by first order rejection sampling [20], reaches a prescribed accuracy in total variation with a number of gradient queries polylogarithmic in the inverse accuracy [3, 20], and stochastic gradients with subexponential tails preserve this rate for targets satisfying a log-Sobolev or a Poincaré inequality [21]. These samplers need a warm start, an approximate proximal step and a rejection step with a random number of gradient queries in each iteration, and light-tailed gradient noise: with stochastic gradients of bounded variance only, a number of queries proportional to the inverse accuracy is necessary and sufficient [21]. SGLD and SGLM use one stochastic gradient per step and no warm start, and their numerical bias, quantified in the present paper, can be made smaller than the fluctuation of the time average

by the choice of the step size and the batch size.

2.3 Stochastic gradient integrators for underdamped Langevin dynamics

The underdamped Langevin dynamics augments the position with a velocity and is discretized by splitting integrators such as BAOAB [56, 57], OBABO [11] and UBU [85, 92]. For convex and smooth potentials, the unadjusted chain of a splitting integrator of second order with one gradient evaluation per step contracts in Wasserstein distance at a rate independent of the dimension [81].

SGHMC adds a friction to compensate for the gradient noise [22], within a general recipe for such samplers [72]. Adaptive thermostats correct for the gradient noise [31, 58, 95], with exponential convergence of the law [61] and a quantified bias from mini-batching [94]. Variance reduction accelerates SGHMC [15, 70, 108, 109, 111], and multilevel coupling gives unbiased estimators [13]. For nonconvex optimization, the complexity bounds of SGHMC can be tighter than those of (SGLD) [41]. Wasserstein bounds are proved without log-concavity [2, 14], through contraction rates of stochastic gradient integrators [62], and with dimension-free rates for log-concave targets [44].

Consider a stochastic gradient sampler such as SGHMC, with unbiased gradients and a Markov integrator whose local weak error is of order h^{K+1}. If the derivatives of the solution of the continuous Poisson equation (3.7) are bounded, the bias of its time average is of order $(Nh)^{-1} + h^K$ and its MSE of order $(Nh)^{-1} + h^{2K}$ [18, Theorems 2 and 3]. Applied to the Markovian form (LM'), which has weak order 1 [106], this framework gives a discretization term of order h^2 in the MSE, whereas the modified class of Section 4 gives h^4. The derivative bounds that this framework requires follow from the decay (3.8) of Theorem 1.

For SG-UBU and a globally convex potential, the discrete Poisson equation bounds the MSE of the time average [71]. The numerical bias is of first order in h, with a coefficient proportional to the variance of the stochastic gradient. For a finite-sum potential whose components are each strongly convex, UBU combined with a symmetric forward and backward sweep over the mini-batches reaches a bias of order h^2 with a single mini-batch per step [88]. The constant of this bound grows with the number of mini-batches, and the order rests on the symmetry of the whole sweep, which lasts a time equal to $2h$ times the number of mini-batches. For SG-UBU and strongly log-concave targets, Gaussian convolution inequalities bound the Wasserstein distance between the numerical invariant distribution and π [87]. The bound is the discretization error of the UBU integrator plus a stochastic gradient error of first order in h, which vanishes with the full gradient.

Like SG-UBU, (SGLM) evaluates one gradient per step and has a discretization error of second order, and the separation in (1.3) holds without the global convexity assumed for SG-UBU in [71]. Unlike the sweep of [88], SGLM draws ξ_k independently, so its numerical bias keeps the first order term (1.4) from the gradient noise. As for SG-UBU, this term comes from the gradient noise only and vanishes with the full gradient, whereas the coefficient (1.5) of (SGLD) also carries the Hessian $\nabla^2 U(x)$ from the discretization error of EM. UBU and LM are analyzed differently: UBU has strong error of second order [85, 92], whereas the Markovian form (LM') of LM has strong error of first order [106]. The second order discretization error of LM is recovered through the modified class of Section 4.

3 Preliminaries

3.1 Notation and assumptions

Throughout, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, $\| \cdot \|$ the operator norm of a tensor, which for a vector is the Euclidean norm $| \cdot |$, I_d is the $d \times d$ identity matrix, $A \succeq B$ means that $A - B$ is positive semidefinite, $v^{\otimes j}$ is the j-fold tensor power, $A : B$ the full contraction of two tensors of the same order, $\mathrm{Tr}[A]$ the trace of a square matrix A, and $\mathrm{Sym}\, T$ the average of a tensor T over the permutations of its indices. Derivatives of U contract with tensors as in [106]: for instance, for a matrix Q and a vector v, $\nabla^3 U(x)[Q] \in \mathbb{R}^d$ and $\nabla^3 U(x)[v] \in \mathbb{R}^{d \times d}$ have entries

$$\big(\nabla^3 U(x)[Q]\big)_i = \sum_{j,k=1}^{d} \partial_i \partial_j \partial_k U(x)\, Q_{jk}, \qquad \big(\nabla^3 U(x)[v]\big)_{ij} = \sum_{k=1}^{d} \partial_i \partial_j \partial_k U(x)\, v_k.$$

We write $\Delta U = \sum_i \partial_i^2 U$ and $\nabla \Delta U$ for its gradient, $\pi(\psi) = \int \psi \, \mathrm{d}\pi$, $\mathbb{E}^x$ for the expectation of the Langevin dynamics (1.1) started at $X_0 = x$, and $\mathbb{E}_\xi$ and $\mathbb{P}_\xi$ for the expectation and the probability over the random variable ξ of the stochastic gradient at fixed x. For a tensor-valued function g and $j \geqslant 0$, $\|\nabla^j g\|_\infty := \sup_x \|\nabla^j g(x)\|$, with $|\cdot|$ for $j = 1$; for a Lipschitz function f, $\|\nabla f\|_\infty$ is its Lipschitz constant. For a random tensor X and $p \geqslant 1$, $\|X\|_p := (\mathbb{E}\|X\|^p)^{\frac{1}{p}}$, with the Euclidean norm $|X|$ in place of $\|X\|$ when X is a vector. Most proofs do not track the dependence of the constants on the dimension d, so the norms used in the paper are equivalent, up to constants depending only on d.

For a signed measure ν with $\nu(\mathbb{R}^d) = 0$ and $\int |x| \, |\nu|(\mathrm{d}x) < \infty$, the Kantorovich–Rubinstein norm is

$$\|\nu\|_{\mathrm{KR}} := \sup_{\|\nabla f\|_\infty \leqslant 1} \int_{\mathbb{R}^d} f \, \mathrm{d}\nu,$$

with the supremum over the Lipschitz functions f with $\|\nabla f\|_\infty \leqslant 1$, so that $\mathcal{W}_1(\mu, \rho) = \|\mu - \rho\|_{\mathrm{KR}}$ for probability measures μ and ρ with finite first moments by the Kantorovich–Rubinstein duality. For a function ψ with $\pi(\psi) = 0$ and $\pi(|x| \, |\psi|) < \infty$, $\psi\pi$ is the signed measure $\psi(x)\,\pi(\mathrm{d}x)$.

Three conventions fix the notation of the paper. The solution of the discrete Poisson equation of (1.1) is written ϕ_h, the subscript h marking its dependence on the step size. Every member of the modified class of Section 4 carries a tilde: the modified test function $\tilde{f}_h$, the modified potential $\tilde{U}_h$ with its Boltzmann distribution, the modified target $\tilde{\pi}_h$, the modified generator $\tilde{\mathcal{L}}_h$, and the modified discrete Poisson solution $\tilde{\phi}_h$. So do the modified dynamics $\tilde{X}_t$, defined together with $\tilde{\mathcal{L}}_h$, and its solution $\tilde{Y}_k(t)$ started at Y_k. The hat marks the stochastic gradient: $\nabla \hat{U}(x; \xi)$ is the estimator and $U(x)$ the exact potential, and $\hat{\mathcal{F}}_k$ and $\hat{\mathbb{E}}_k$ in (3.2) are the filtration and the conditional expectation of (SGLM'). Table 2 lists the objects of (1.1) next to their modified counterparts.

	original		modified	
object	symbol	definition	symbol	definition
test function	f	Assumption 2	$\tilde{f}_h$	(4.1)
potential	U	Assumption 1	$\tilde{U}_h$	(4.11)
Langevin dynamics	X_t	(1.1)	$\tilde{X}_t$, $\tilde{Y}_k(t)$	(4.16)
generator	$\mathcal{L}$	(3.3)	$\tilde{\mathcal{L}}_h$	(4.17)
invariant distribution	π	$\pi \propto e^{-U}$	$\tilde{\pi}_h$	$\tilde{\pi}_h \propto e^{-\tilde{U}_h}$, (4.11)
discrete Poisson solution	ϕ_h	(3.5)	$\tilde{\phi}_h$	(4.18)

Table 2: Notation of the original objects and of their modified counterparts.

For the full gradient (LM) integrator, we make the following two assumptions on the potential U and the test function f. Strong convexity of U is required only outside a bounded ball.

Assumption 1 *$U \in C^7(\mathbb{R}^d)$ satisfies*

$$|\nabla U(0)| \leqslant M, \qquad \max_{2 \leqslant i \leqslant 7} \|\nabla^i U(x)\| \leqslant M, \qquad \forall x \in \mathbb{R}^d,$$

and there are constants $m > 0$ and $R \geqslant 0$ such that $\nabla^2 U(x) \succeq m I_d$ for $|x| \geqslant R$.

Assumption 2 *$f \in C^4(\mathbb{R}^d)$ satisfies*

$$\max_{1 \leqslant j \leqslant 4} \|\nabla^j f(x)\| \leqslant L, \qquad \forall x \in \mathbb{R}^d.$$

For SGLM, additional boundedness conditions are required for the stochastic gradient $\nabla \hat{U}(x; \xi)$.

Assumption 3 *The estimator is unbiased and its L^{12} norm with respect to ξ grows linearly,*

$$\mathbb{E}_\xi\big[\nabla \hat{U}(x; \xi)\big] = \nabla U(x), \qquad \Big(\mathbb{E}_\xi\big[|\nabla \hat{U}(x; \xi)|^{12}\big]\Big)^{\frac{1}{12}} \leqslant M\big(1 + |x|\big), \qquad \forall x \in \mathbb{R}^d,$$

$\hat{U}(\cdot;\xi) \in C^2(\mathbb{R}^d)$ for every ξ with $(x,\xi) \mapsto \nabla^2\hat{U}(x;\xi)$ measurable, and its Hessian has a second moment bound, uniform on unit balls,

$$\mathbb{E}_\xi\Big[\sup_{|x'-x|\leqslant 1} \|\nabla^2\hat{U}(x';\xi)\|^2\Big] \leqslant M^2, \qquad \forall x \in \mathbb{R}^d,$$

with M the constant of Assumption 1, and its variance is bounded by a constant $\Sigma \leqslant M^2$,

$$\mathbb{E}_\xi\big[|\nabla\hat{U}(x;\xi) - \nabla U(x)|^2\big] \leqslant \Sigma, \qquad \forall x \in \mathbb{R}^d.$$

The constraint $\Sigma \leqslant M^2$ in Assumption 3 is imposed for the convenience of analysis. The bound on the Hessian of $\hat{U}$ in Assumption 3 ensures the uniqueness of the invariant distribution of (SGLM'), as required by Harris's ergodic theorem [47]. A stronger condition is the almost sure uniform bound $\|\nabla^2\hat{U}(x;\xi)\| \leqslant M$ for all $x \in \mathbb{R}^d$ and almost every ξ. Although Assumption 3 requires the potential form, $\nabla\hat{U}(x;\xi)$ does not need to be a gradient: the proofs use only the boundedness of its derivative $\nabla^2\hat{U}(x;\xi)$.

From here on, the letter C denotes a positive constant depending only on m, M, R and d, but not on h, N, P, Σ or the initial state, and it changes from line to line. The letter h_0 denotes an upper bound in $(0,1]$ on the step size h, depending only on m, M, R and d. In each statement, h_0 is the smallest of the bounds its proof uses, and every statement below holds for $h \in (0, h_0]$.

Although this paper focuses on SGLM, the proofs are carried out also for the full gradient LM, as the stochastic gradient case follows directly in most places. When we speak of the LM integrators in the plural, the statement covers both the full gradient and the stochastic gradient cases.

3.2 Markovian form of the LM integrators

Although the (LM) integrator is non-Markovian, it can be written in a Markovian form for the convenience of analysis. Let $(B_t)_{t\geqslant 0}$ be a standard Brownian motion in $\mathbb{R}^d$, the one driving (1.1), and realize the Gaussian vectors as its increments over the steps,

$$\chi_k = \frac{1}{\sqrt{h}}\big(B_{(k+1)h} - B_{kh}\big), \qquad k \geqslant 0.$$

Let $\mathcal{F}_k$ be the sigma algebra generated by the Brownian motion up to the time kh, the past before the step k, and $\mathbb{E}_k$ the conditional expectation given that past,

$$\mathcal{F}_k := \sigma\big(B_s : 0 \leqslant s \leqslant kh\big), \qquad \mathbb{E}_k[\,\cdot\,] := \mathbb{E}\big[\,\cdot\,\big|\mathcal{F}_k\big], \tag{3.1}$$

so that χ_k is $\mathcal{F}_{k+1}$-measurable and independent of $\mathcal{F}_k$. We construct a Markov chain $(Y_k)_{k\geqslant 0}$ by

$$Y_{k+1} = Y_k - h\,\nabla U\Big(Y_k + \sqrt{\frac{h}{2}}\,\chi_k\Big) + \sqrt{2h}\,\chi_k. \tag{LM'}$$

Then $Z_k = Y_k + \sqrt{h/2}\,\chi_k$ satisfies (LM), so the LM integrator is equivalent to (LM'), the form in which it is analyzed as a postprocessed integrator in [101]; here Y_k is $\mathcal{F}_k$-measurable and Z_k is $\mathcal{F}_{k+1}$-measurable. Similarly, the (SGLM) integrator is equivalent to

$$Y_{k+1} = Y_k - h\,\nabla\hat{U}\Big(Y_k + \sqrt{\frac{h}{2}}\,\chi_k;\xi_k\Big) + \sqrt{2h}\,\chi_k, \tag{SGLM'}$$

where the random variables ξ_k are independent and identically distributed and independent of $(B_t)_{t\geqslant 0}$. For (SGLM') the past before the step k also contains $\xi_0,\dots,\xi_{k-1}$, and we define

$$\hat{\mathcal{F}}_k := \sigma\big(B_s,\ \xi_j : 0 \leqslant s \leqslant kh,\ 0 \leqslant j \leqslant k-1\big), \qquad \hat{\mathbb{E}}_k[\,\cdot\,] := \mathbb{E}\big[\,\cdot\,\big|\hat{\mathcal{F}}_k\big], \tag{3.2}$$

so that χ_k and ξ_k are independent of $\hat{\mathcal{F}}_k$, and Y_k is $\hat{\mathcal{F}}_k$-measurable and Z_k is $\hat{\mathcal{F}}_{k+1}$-measurable.

Since χ_k is determined by the Brownian motion on $[kh,(k+1)h]$, we call (LM') and (SGLM') the Markovian forms of LM and SGLM. The analysis through the discrete Poisson equation is carried out on the Markovian form $(Y_k)_{k\geqslant 0}$, not on the original form $(Z_k)_{k\geqslant 0}$, and their connection is made by the modified class of Section 4. The Markovian form (LM') has strong and weak order 1 [106], hence the second order discretization error of LM comes entirely from the Gaussian shift $\sqrt{h/2}\,\chi_k$ between Y_k and Z_k.

3.3 Elliptic estimates of the discrete Poisson equation

The generator of the overdamped Langevin dynamics (1.1) is

$$\mathcal{L}\psi = \Delta\psi - \nabla U \cdot \nabla\psi, \tag{3.3}$$

and $e^{t\mathcal{L}}$, $t \geqslant 0$, is its semigroup, $e^{t\mathcal{L}}f(x) = \mathbb{E}^x[f(X_t)]$, where X_t solves (1.1) with $X_0 = x$. The Dirichlet form of $\mathcal{L}$ [6] gives the integration by parts

$$\pi(g_1\,\mathcal{L}g_2) = -\pi(\nabla g_1 \cdot \nabla g_2) \tag{3.4}$$

for $g_1, g_2 \in C^2(\mathbb{R}^d)$ whose derivatives of orders 0 to 2 grow at most polynomially.

For the test function f of Assumption 2, the Kolmogorov solution $u(x,t) := e^{t\mathcal{L}}f(x) - \pi(f)$ solves the parabolic equation $\partial_t u = \mathcal{L}u$ with $u(\cdot,0) = f - \pi(f)$. By the exponential convergence of $u(x,t)$ established in Theorem 1 below, the discrete Poisson solution [71, 106]

$$\phi_h(x) := h\sum_{k=0}^{\infty} u(x,kh) = h\left(\mathcal{I} - e^{h\mathcal{L}}\right)^{-1}\big(f(x) - \pi(f)\big), \tag{3.5}$$

with $\mathcal{I}$ the identity operator, is well defined and, by telescoping the series with $e^{h\mathcal{L}}u(\cdot,kh) = u(\cdot,(k+1)h)$ and the exponential decay (3.8) of u, satisfies the discrete Poisson equation

$$\frac{\phi_h(x) - (e^{h\mathcal{L}}\phi_h)(x)}{h} = f(x) - \pi(f), \qquad x \in \mathbb{R}^d. \tag{3.6}$$

As $h \to 0$, (3.6) formally becomes the continuous Poisson equation

$$-\mathcal{L}\phi_0(x) = f(x) - \pi(f), \qquad x \in \mathbb{R}^d, \tag{3.7}$$

studied in [10, 77, 102]. By (3.8) of Theorem 1 below, $\phi_0 = \int_0^\infty u(\cdot,t)\,\mathrm{d}t$ solves (3.7), with derivatives of orders 1 to 4 bounded by CL. Among the functions with a bounded gradient, (3.7) determines ϕ_0 up to an additive constant, as the proof of Corollary 2 shows.

Theorem 1 *Let Assumptions 1 and 2 hold. There are $\lambda > 0$ and C, depending only on m, M, R and d, such that for all $x \in \mathbb{R}^d$ and $t \geqslant 0$,*

$$|u(x,t)| \leqslant CL\,(1+|x|)\,e^{-\lambda t}, \qquad \max_{1\leqslant j\leqslant 4}\|\nabla^j u(x,t)\| \leqslant CL\,e^{-\lambda t}. \tag{3.8}$$

Consequently, for every $h \in (0,\lambda^{-1}]$ the discrete Poisson solution (3.5) satisfies

$$\max_{1\leqslant j\leqslant 4}\|\nabla^j \phi_h(x)\| \leqslant CL, \qquad \forall x \in \mathbb{R}^d. \tag{3.9}$$

(Proof in Appendix A.)

Theorem 1 requires one order of derivative of u more than Theorem 1 of [106], and its proof is almost the same; the details are given in Appendix A. The proof uses the derivatives of U up to order 5 only, although Assumption 1 provides bounds up to order 7; the 2 extra orders are consumed by the modified potential of Section 4.2, whose construction loses 2 derivatives. Corollary 1 applies it to the modified potential $\tilde{U}_h$ and the modified test function $\tilde{f}_h$, whose discrete Poisson solution is written $\tilde{\phi}_h$.

4 Modified class for the LM integrator

The MSE (1.2) averages the test function along the samples $(Z_k)_{k\geqslant 0}$ of the (LM) integrator, while the analysis runs on the Markovian form (LM'). The Gaussian shift between the two is absorbed by the *modified class*: a modified test function $\tilde{f}_h$, a modified potential $\tilde{U}_h$ with its modified generator $\tilde{\mathcal{L}}_h$, and a modified discrete Poisson solution $\tilde{\phi}_h$. The modified class is local: its construction does not involve the long-time behavior. Section 4.1 defines the modified test function $\tilde{f}_h$ as the heat flow of f, which splits the MSE into the time average error of the Markovian form and a bias against a reference measure. Section 4.2 defines the modified potential $\tilde{U}_h$ by the correction that makes the bias of order h^2, cut off far from the origin so that the bounds of Assumption 1 up to order 5 hold uniformly in h, with $2M$ and $\frac{m}{2}$ in place of M and m. Section 4.3 proves that the bias against the modified target $\tilde{\pi}_h$ is of order h^2. Section 4.4 defines the modified generator $\tilde{\mathcal{L}}_h$ and the modified discrete Poisson solution $\tilde{\phi}_h$.

4.1 Modified test function $\tilde{f}_h$

The positions Z_k and Y_k in (LM) and (LM') differ by the Gaussian shift $\sqrt{h/2}\,\chi_k \sim \mathcal{N}(0, \frac{h}{2} I_d)$, which is independent of $\mathcal{F}_k$. The conditional mean of $f(Z_k)$ given $\mathcal{F}_k$ is therefore the heat semigroup applied to f at time $\frac{h}{4}$, evaluated at Y_k. This motivates the definition of the *modified test function*

$$\tilde{f}_h(y) := (e^{\frac{h}{4}\Delta} f)(y) = \mathbb{E}\Big[f\Big(y + \sqrt{\frac{h}{2}}\,\chi\Big)\Big], \quad \chi \sim \mathcal{N}(0, I_d), \tag{4.1}$$

for which $\mathbb{E}_k\big[f(Z_k)\big] = \tilde{f}_h(Y_k)$. Since $\tilde{f}_h$ is an average of translates of f, every derivative of $\tilde{f}_h$ is the same average of the corresponding derivative of f, and $\tilde{f}_h$ satisfies Assumption 2 with the same constant L for every h. Expanding f about y to the order 4 along $\sqrt{h/2}\,\chi$,

$$f\Big(y + \sqrt{\frac{h}{2}}\,\chi\Big) = f(y) + \sqrt{\frac{h}{2}}\,\nabla f(y)\cdot\chi + \frac{h}{4}\,\nabla^2 f(y)[\chi,\chi] + \frac{1}{6}\Big(\frac{h}{2}\Big)^{\frac{3}{2}}\nabla^3 f(y)\big[\chi^{\otimes 3}\big] + \varrho_4^{\chi}(y),$$

where the remainder satisfies

$$|\varrho_4^{\chi}(y)| \leqslant \frac{L}{24}\Big(\frac{h}{2}\Big)^2 |\chi|^4.$$

Under the expectation the terms of odd order in χ vanish, hence

$$\Big|\tilde{f}_h(y) - f(y) - \frac{h}{4}\,\Delta f(y)\Big| \leqslant \frac{L}{24}\Big(\frac{h}{2}\Big)^2 \mathbb{E}|\chi|^4 \leqslant CLh^2, \qquad \forall y \in \mathbb{R}^d, \tag{4.2}$$

where $\Delta f = \mathrm{Tr}[\nabla^2 f] = \sum_{i=1}^d \partial_i^2 f$. The modified test function $\tilde{f}_h$ absorbs the Gaussian shift, so that the analysis runs on the positions Y_k of the Markovian form (LM').

The modified test function $\tilde{f}_h(Y_k)$ is an unbiased estimate of $f(Z_k)$ under $\mathcal{F}_k$. Define the difference $\mathfrak{M}_k := f(Z_k) - \tilde{f}_h(Y_k)$; then $\mathbb{E}_k[\mathfrak{M}_k] = 0$ by (4.1), and since f is L-Lipschitz,

$$\mathbb{E}_k\big[\mathfrak{M}_k^2\big] \leqslant \mathbb{E}_k\Big[\big|f(Z_k) - f(Y_k)\big|^2\Big] \leqslant L^2\,\frac{h}{2}\,\mathbb{E}|\chi|^2 = \frac{dh}{2}\,L^2. \tag{4.3}$$

The $\mathfrak{M}_k$ are $\mathcal{F}_{k+1}$-measurable martingale differences, so their sum has additive variance. For every probability measure ρ on $\mathbb{R}^d$ with a finite first moment, inserting $f(Z_k) = \tilde{f}_h(Y_k) + \mathfrak{M}_k$ into the MSE (1.2), applying Cauchy–Schwarz inequality to the three resulting terms and then the additive variance with (4.3) gives

$$\begin{aligned}\mathrm{MSE}(N,h) &= \mathbb{E}\Bigg[\Bigg|\frac{1}{N}\sum_{k=0}^{N-1}\big(\tilde{f}_h(Y_k) - \rho(\tilde{f}_h)\big) + \big(\rho(\tilde{f}_h) - \pi(f)\big) + \frac{1}{N}\sum_{k=0}^{N-1}\mathfrak{M}_k\Bigg|^2\Bigg] \\ &\leqslant 3\,\mathbb{E}\Bigg[\underbrace{\Bigg|\frac{1}{N}\sum_{k=0}^{N-1}\big(\tilde{f}_h(Y_k) - \rho(\tilde{f}_h)\big)\Bigg|^2}_{\text{time average error}}\Bigg] + 3\,\underbrace{\big(\rho(\tilde{f}_h) - \pi(f)\big)^2}_{\text{modified bias}} + \frac{3dL^2}{2}\,\frac{h}{N}.\end{aligned} \tag{4.4}$$

The term $\frac{3dL^2}{2}\,\frac{h}{N}$ of (4.4) is absorbed into $\frac{1}{Nh}$, so it remains to estimate the time average error and the modified bias. The measure ρ is to be chosen so that the modified bias and the discretization part of the time average error are of order h^2.

The measure ρ cannot simply be π. With $\rho = \pi$, the expansion (4.2) gives

$$\pi(\tilde{f}_h) - \pi(f) = \frac{h}{4}\,\pi(\Delta f) + \mathcal{O}(Lh^2),$$

whose leading coefficient $\pi(\Delta f)$ is in general nonzero. The choice of ρ has to compensate for the Laplacian term, so that the modified bias $\rho(\tilde{f}_h) - \pi(f)$ is of order h^2.

4.2 Modified potential $\tilde{U}_h$

We aim to find a distribution ρ such that the modified bias $\rho(\tilde{f}_h) - \pi(f)$ is of order h^2. Since $\tilde{f}_h = f + \frac{h}{4}\Delta f + \mathcal{O}(h^2)$ by (4.2), we consider the ansatz

$$\rho \propto \exp\Big(-U - \frac{h}{4}\,V\Big), \tag{4.5}$$

where V is a function to be determined and the coefficient $\frac{h}{4}$ matches that of Δf in (4.2). We derive $V = |\nabla U|^2 - \Delta U$ by a perturbation analysis. The perturbed potential $U + \frac{h}{4}V$, once made rigorous by a cutoff far from the origin, is the *modified potential* $\tilde{U}_h$, and its Boltzmann distribution $\tilde{\pi}_h \propto \exp(-\tilde{U}_h)$ is the *modified target*. Finally, we prove the modified bias estimate $\tilde{\pi}_h(\tilde{f}_h) - \pi(f) = \mathcal{O}(h^2)$.

Derivation of V by perturbation

For a test function ψ, the ansatz (4.5) gives

$$\rho(\psi) = \frac{\pi(\psi e^{-\frac{h}{4}V})}{\pi(e^{-\frac{h}{4}V})} = \frac{\int_{\mathbb{R}^d} \psi e^{-U-\frac{h}{4}V}\,\mathrm{d}x}{\int_{\mathbb{R}^d} e^{-U-\frac{h}{4}V}\,\mathrm{d}x}.$$

Expanding $e^{-\frac{h}{4}V} = 1 - \frac{h}{4}V + \mathcal{O}(h^2)$ in the numerator and in the denominator of $\rho(\psi)$, we obtain

$$\rho(\psi) = \pi(\psi) - \frac{h}{4}\big(\pi(\psi V) - \pi(\psi)\,\pi(V)\big) + \mathcal{O}(h^2), \tag{4.6}$$

the term of order h being $-\frac{h}{4}$ times the covariance of ψ and V under π. Taking $\psi = \tilde{f}_h$ in (4.6) and inserting the expansion (4.2) of the modified test function,

$$\rho(\tilde{f}_h) - \pi(f) = \frac{h}{4}\Big(\pi(\Delta f) - \pi(fV) + \pi(f)\,\pi(V)\Big) + \mathcal{O}(h^2). \tag{4.7}$$

The modified bias is therefore of order h^2 exactly when the term of order h in (4.7) vanishes,

$$\pi(\Delta f) = \pi(fV) - \pi(f)\,\pi(V). \tag{4.8}$$

We look for a single V satisfying (4.8) for every f. The integration by parts (3.4), first with $g_1 = 1$ and $g_2 = f$ and then with $g_1 = f$ and $g_2 = U$, gives

$$\pi(\Delta f) = \pi\big(\nabla f \cdot \nabla U\big) = \pi\big(f\,|\nabla U|^2 - f\,\Delta U\big). \tag{4.9}$$

Hence we select the potential function

$$V := |\nabla U|^2 - \Delta U = -\mathcal{L}U, \tag{4.10}$$

for which $\pi(V) = -\pi(\mathcal{L}U) = 0$ by the invariance of π. This computation is formal, and the cutoff introduced below changes the measure (4.5); Lemma 2 proves the bias estimate for the modified target.

Cutoff in the modified potential

The naive construction $U + \frac{h}{4}V$ of the modified potential does not satisfy Assumption 1, because the derivatives of $|\nabla U|^2$ grow with $|x|$. Since Theorem 1 must be applied to the modified potential with constants independent of h, the modification is cut off far from the origin. Fix a smooth cutoff function $\theta \in C^\infty(\mathbb{R}^d)$ with $0 \leqslant \theta \leqslant 1$, $\theta = 1$ on $\{|x| \leqslant 1\}$ and $\theta = 0$ on $\{|x| \geqslant 2\}$. With $\delta \in (0, 1]$ fixed by (B.9) in the proof of Lemma 1 below, we define the modified potential by

$$\tilde{U}_h := U + \frac{h}{4}\,V\,\theta_h, \qquad \theta_h(x) := \theta\Big(\frac{hx}{\delta}\Big), \tag{4.11}$$

so that $\tilde{U}_h = U + \frac{h}{4}V$ on the ball $\{|x| \leqslant \delta/h\}$ and $\tilde{U}_h = U$ on $\{|x| \geqslant 2\delta/h\}$. In particular, since $0 \leqslant \theta_h \leqslant 1$ and $|V(y)| \leqslant C(1+|y|)^2$ by Assumption 1,

$$|\tilde{U}_h(x) - U(x)| \leqslant \frac{h}{4} \max_{|y| \leqslant 2\delta/h} |V(y)| \leqslant \frac{C}{h}, \qquad x \in \mathbb{R}^d. \tag{4.12}$$

The modified target $\tilde{\pi}_h \propto e^{-\tilde{U}_h}$ is therefore a perturbation of π by a bounded potential, although the bound (4.12) is not uniform in h. The derivatives of $\tilde{U}_h - U$ of orders 1 to 5, however, are bounded uniformly in x and h: by (4.14) in Lemma 1 below they are at most $Ch(1+|x|)$, and they vanish for $|x| > 2\delta/h$.

Lemma 1 *Let Assumption 1 hold. There are $\delta \in (0,1]$ and $h_0 \in (0,\delta]$ such that for every $h \in (0,h_0]$, the modified potential $\tilde{U}_h \in C^5(\mathbb{R}^d)$ satisfies*

$$|\nabla \tilde{U}_h(0)| \leqslant 2M, \qquad \max_{2 \leqslant i \leqslant 5} \|\nabla^i \tilde{U}_h(x)\| \leqslant 2M, \qquad \forall x \in \mathbb{R}^d, \tag{4.13}$$

the convexity condition $\nabla^2 \tilde{U}_h(x) \succeq \frac{m}{2} I_d$ for $|x| \geqslant R$, and

$$\max_{1 \leqslant i \leqslant 5} \|\nabla^i \tilde{U}_h(x) - \nabla^i U(x)\| \leqslant Ch\big(1+|x|\big), \qquad \forall x \in \mathbb{R}^d. \tag{4.14}$$

(Proof in Appendix B.1.)

Theorem 1 therefore applies to $\tilde{U}_h$, since its proof uses the bounds of Assumption 1 up to order 5 only (Appendix A). The constants, in particular the rate λ, depend only on m, M, R and d, not on h. In the rest of the paper, the number δ is fixed by Lemma 1 and absorbed into the constant C.

4.3 Estimation of the modified bias $\tilde{\pi}_h(\tilde{f}_h) - \pi(f)$

The cancellation (4.8) between the heat flow of the test function and the modification of the potential holds for the modified target $\tilde{\pi}_h$ as it does for the ansatz (4.5), since the cutoff acts only where π has exponentially small mass. Lemma 2 carries the perturbation argument that produced V over to the modified target.

Lemma 2 *Let Assumption 1 hold. For every $h \in (0,h_0]$ and every $f \in C^1(\mathbb{R}^d)$ with $\|\nabla f\|_\infty < \infty$,*

$$\big|\tilde{\pi}_h(\tilde{f}_h) - \pi(f)\big| \leqslant Ch^2 \|\nabla f\|_\infty. \tag{4.15}$$

(Proof in Appendix B.2.)

Remark 1 *The Gaussian potential $U = \frac{1}{2}|x|^2$, taken without the cutoff, illustrates the modification. Here $V = |x|^2 - d$, so the modified potential equals $(1+\frac{h}{2})\frac{|x|^2}{2}$ up to an additive constant and $\tilde{\pi}_h$ has variance $(1+\frac{h}{2})^{-1}$ in each coordinate. The chain becomes $Y_{k+1} = (1-h)Y_k + \sqrt{2h}\,(1-\frac{h}{2})\,\chi_k$, whose stationary variance $1-\frac{h}{2}$ agrees with that of $\tilde{\pi}_h$ up to $\mathcal{O}(h^2)$. The shifted variable $Z_k = Y_k + \sqrt{h/2}\,\chi_k$ has stationary variance 1, in agreement with the exactness of (LM) for Gaussian targets established in [101, Remark 4.1].*

4.4 Modified discrete Poisson equation $\tilde{\phi}_h$

Let the *modified dynamics* be the overdamped Langevin dynamics driven by the potential $\tilde{U}_h$,

$$\mathrm{d}\tilde{X}_t = -\nabla \tilde{U}_h(\tilde{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{4.16}$$

with the *modified generator*

$$\tilde{\mathcal{L}}_h \psi = \Delta\psi - \nabla \tilde{U}_h \cdot \nabla\psi, \tag{4.17}$$

and its semigroup $e^{t\tilde{\mathcal{L}}_h}$. The invariant distribution of (4.16) is the modified target $\tilde{\pi}_h$. The time average error in (4.4) with $\rho = \tilde{\pi}_h$ is analyzed through the discrete Poisson equation of the modified generator (4.17). Let $\tilde{\phi}_h$ be the discrete Poisson solution (3.5) with $\tilde{\mathcal{L}}_h$, $\tilde{\pi}_h$ and $\tilde{f}_h$ in place of $\mathcal{L}$, π and f, so that

$$\frac{\tilde{\phi}_h(x) - (e^{h\tilde{\mathcal{L}}_h}\tilde{\phi}_h)(x)}{h} = \tilde{f}_h(x) - \tilde{\pi}_h(\tilde{f}_h), \qquad x \in \mathbb{R}^d. \tag{4.18}$$

Corollary 1 *Let Assumptions 1 and 2 hold. For every $h \in (0, h_0]$, $\tilde{\phi}_h$ is well defined and*

$$\max_{1 \leqslant j \leqslant 4} \|\nabla^j \tilde{\phi}_h(x)\| \leqslant CL, \qquad \forall x \in \mathbb{R}^d. \tag{4.19}$$

Corollary 1 is Theorem 1 applied to the modified class. By Lemma 1, $\tilde{U}_h$ satisfies the bounds that the proof of Theorem 1 uses, and $\tilde{f}_h$ satisfies Assumption 2 with the constant L. The constants in (4.19) therefore depend only on m, M, R and d, and the threshold h_0 includes λ^{-1}.

5 Local error of the LM integrators

This section compares one step of the Markovian form (LM') with one step of the modified dynamics (4.16), and derives from this comparison the local weak error used in the MSE analysis. Section 5.1 bounds the moments of the chain uniformly in time, so that the conditional one-step estimates can be averaged over the chain. It then bounds the sizes of a step and matches the moments of orders 1 to 3 of the chain and of the modified dynamics, with a mismatch of order h^3. Section 5.2 turns the matching into the local weak error of the LM integrator, and of the SGLM integrator, whose gradient noise adds an explicit term of order h^2 and a remainder of order h^3.

Taking $\rho = \tilde{\pi}_h$ in (4.4), Lemma 2 bounds the modified bias by CLh^2, and the MSE satisfies

$$\mathrm{MSE}(N, h) \leqslant 3\,\mathbb{E}\left[\underbrace{\left|\frac{1}{N}\sum_{k=0}^{N-1}\big(\tilde{f}_h(Y_k) - \tilde{\pi}_h(\tilde{f}_h)\big)\right|}_{\text{time average error}}^2\right] + CL^2\Big(h^4 + \frac{dh}{N}\Big). \tag{5.1}$$

The time average error is centered at $\tilde{\pi}_h(\tilde{f}_h)$, so it is controlled through the discrete Poisson equation of a dynamics that preserves $\tilde{\pi}_h$. This is the modified dynamics (4.16), whose discrete Poisson solution $\tilde{\phi}_h$ in (4.18) has the right side $\tilde{f}_h - \tilde{\pi}_h(\tilde{f}_h)$. The discrete Poisson solution transfers local errors into the time average error [71, 106]. The local error studied in this section is therefore the error of one step of the Markovian form (LM') against one step of the modified dynamics from the same state.

Since the chains (LM') and (SGLM') are time-homogeneous, statements about a single step are made at the index 0 and apply at any k; Lemmas 3, 6 and 7 are stated at the index k, in the form the MSE analysis uses. One step of the (LM') chain is

$$Y_1 - Y_0 = \sqrt{2h}\,\chi_0 - h\,\nabla U(Z_0), \qquad Z_0 = Y_0 + \sqrt{\frac{h}{2}}\,\chi_0, \tag{5.2}$$

where $\chi_0 \sim \mathcal{N}(0, I_d)$ is independent of $\mathcal{F}_0$, and one step of the (SGLM') chain is

$$Y_1 - Y_0 = \sqrt{2h}\,\chi_0 - h\,\nabla \hat{U}(Z_0; \xi_0), \tag{5.3}$$

where Z_0 is given by (5.2), and χ_0 and ξ_0 are independent of $\hat{\mathcal{F}}_0$. For each $k \geqslant 0$, let $(\tilde{Y}_k(t))_{t \geqslant 0}$ solve the modified dynamics (4.16) from $\tilde{Y}_k(0) = Y_k$, driven by $(B_{kh+t} - B_{kh})_{t \geqslant 0}$, whose increment over $[0, h]$ is $\sqrt{h}\,\chi_k$. Then $\mathbb{E}_k[\psi(\tilde{Y}_k(h))] = (e^{h\tilde{\mathcal{L}}_h}\psi)(Y_k)$. From the same state Y_0, the chains (LM') and (SGLM') have the same Z_0 and the same $(\tilde{Y}_0(t))_{t \geqslant 0}$, and differ only in Y_1. For both chains, the local error of one step against the modified dynamics is $Y_1 - \tilde{Y}_0(h)$.

5.1 Moment bounds and moment matching

The MSE analysis squares the polynomial weight of the local weak errors of Section 5.2, and therefore needs moments of order 12 of the chain, uniformly in time. The following lemma provides them. Its proof is standard and follows the Lyapunov function arguments in [106], with Lemma 1 for the modified dynamics.

Lemma 3 *Let Assumption 1 hold. There are $h_0 > 0$ and C such that for the chain (LM'), every $h \in (0, h_0]$ and every $k \geqslant 0$,*

$$\mathbb{E}|Y_k|^{12} \leqslant C\big(1 + |Y_0|\big)^{12}, \qquad \max\big\{\mathbb{E}|\tilde{Y}_k(h) - Y_k|^{12},\ \mathbb{E}|Y_{k+1} - Y_k|^{12}\big\} \leqslant Ch^6\big(1 + |Y_0|\big)^{12}. \tag{5.4}$$

If Assumption 3 additionally holds, the same moment bounds hold for the chain (SGLM'), with the same h_0 and C. (Proof in Appendix C.1.)

The next lemma bounds the size of one step of the chain and of the modified dynamics, and the size of the local error between them; these sizes control the remainders of all the expansions below. Lemmas 4 and 5 are stated for (LM') only; the local weak error estimate (5.9) of (SGLM') relies on them only through the full gradient step (5.2).

Lemma 4 *Let Assumption 1 hold. Then, for the chain (LM') and every $s \in [0,h]$,*

$$\|\!|Y_1 - Y_0|\!\|_4 \leqslant C\sqrt{h}\big(1+|Y_0|\big), \quad \|\!|\tilde{Y}_0(s) - Y_0|\!\|_4 \leqslant C\sqrt{h}\big(1+|Y_0|\big), \quad \|\!|Y_1 - \tilde{Y}_0(h)|\!\|_4 \leqslant Ch^{\frac{3}{2}}\big(1+|Y_0|\big), \tag{5.5}$$

where $\|\!|\cdot|\!\|_4$ is taken under $\mathbb{E}_0$. (Proof in Appendix C.2.)

The moments of one step of the modified dynamics and of the chain are expanded in Lemmas 18 and 19 of Appendix C. Compared with the exact dynamics (1.1), the chain matches the moments of orders 2 and 3 to $\mathcal{O}(h^3)$ but misses the mean by $\frac{h^2}{4}\nabla V$, with V as in (4.10); the drift correction of the modified dynamics removes exactly this mismatch. The cutoff keeps the cancellation on the ball $\{|Y_0| \leqslant \delta/h\}$, and outside the ball the remaining mismatch is again of order h^3, because $h|Y_0|/\delta \geqslant 1$ there.

Lemma 5 *Let Assumption 1 hold. Then, for the chain (LM'),*

$$\Big\|\mathbb{E}_0\big[(Y_1-Y_0)^{\otimes j}\big] - \mathbb{E}_0\big[(\tilde{Y}_0(h)-Y_0)^{\otimes j}\big]\Big\| \leqslant Ch^3\big(1+|Y_0|\big)^6, \qquad j=1,2,3. \tag{5.6}$$

(Proof in Appendix C.3.)

5.2 Local weak error expansion

The local weak error of the LM integrator on a test function G is $\mathbb{E}_0\big[G(Y_1) - G(\tilde{Y}_0(h))\big]$, and the MSE analysis takes $G = \tilde{\phi}_h$. Because the local error is of size $h^{\frac{3}{2}}$, against h^2 for the SRK integrators of [106], G is expanded along a step to the order 3: the terms of orders 1 to 3 are matched by Lemma 5, and the remainder is controlled by Lemma 4. This is why Assumption 2 asks for 4 derivatives of f.

Lemma 6 *Let Assumption 1 hold and let $G \in C^4(\mathbb{R}^d)$ have bounded derivatives of orders 1 to 4. Then, for the chain (LM') and every $k \geqslant 0$,*

$$\Big|\mathbb{E}_k\big[G(Y_{k+1}) - G(\tilde{Y}_k(h))\big]\Big| \leqslant Ch^3 \sum_{j=1}^{4} \|\nabla^j G\|_\infty \big(1+|Y_k|\big)^6. \tag{5.7}$$

(Proof in Appendix C.4.)

With the stochastic gradient, the gradient noise at x has the covariance

$$\mathcal{C}(x) := \mathbb{E}_\xi\Big[\big(\nabla\hat{U}(x;\xi) - \nabla U(x)\big) \otimes \big(\nabla\hat{U}(x;\xi) - \nabla U(x)\big)\Big], \qquad \mathrm{Tr}\big[\mathcal{C}(x)\big] \leqslant \Sigma, \tag{5.8}$$

where the bound on the trace is Assumption 3. The step (5.3) adds to the step (5.2) the gradient noise at Z_0, multiplied by $-h$. Given $\hat{\mathcal{F}}_0$ and χ_0 this noise has mean 0, so the mean of the step is unchanged, while its moment of order 2 gains $h^2\,\hat{\mathbb{E}}_0[\mathcal{C}(Z_0)]$, as for (SGLD).

Lemma 7 keeps this gain as an explicit term of order h^2, bounded through (C.27) by $\frac{\Sigma h^2}{2}\|\nabla^2 G\|_\infty$, against the order h^3 of (5.7); the rest of the gradient noise also contributes only a remainder of order h^3.

Lemma 7 *Let Assumptions 1 and 3 hold and let $G \in C^4(\mathbb{R}^d)$ have bounded derivatives of orders 1 to 4. Then, for the chain (SGLM') and every $k \geqslant 0$,*

$$\Big|\hat{\mathbb{E}}_k\big[G(Y_{k+1}) - G(\tilde{Y}_k(h))\big] - \frac{h^2}{2}\nabla^2 G(Y_k) : \hat{\mathbb{E}}_k\big[\mathcal{C}(Z_k)\big]\Big| \leqslant Ch^3 \sum_{j=1}^{4} \|\nabla^j G\|_\infty \big(1+|Y_k|\big)^6. \tag{5.9}$$

(Proof in Appendix C.5.)

6 Main results on the LM integrators

This section states the main results: the MSE bounds for the samples of the (LM) and (SGLM) integrators in Section 6.1, the numerical bias of SGLM in Section 6.2, its comparison with the numerical bias of (SGLD) in Section 6.3, and the explicit coefficients of h in $\mathcal{W}_1(\pi_h,\pi)$ and $\mathcal{W}_1(\bar\pi_h,\pi)$, given by the Riesz representors, with their comparison over the batch size, in Section 6.4. Their proofs are in Appendices D, E, F and G.

6.1 MSE analysis of SGLM

The samples of the (LM) and (SGLM) integrators start from a given $Z_0\in\mathbb{R}^d$. The analysis below runs on the Markovian form from a given Y_0, with $Z_k=Y_k+\sqrt{h/2}\,\chi_k$ for every $k\geqslant 0$; Appendix D applies it to the samples from Z_0 by starting the Markovian form at $Y_1=Z_0-h\nabla U(Z_0)+\sqrt{h/2}\,\chi_0$, with $\nabla\hat U(Z_0;\xi_0)$ in place of $\nabla U(Z_0)$ for SGLM, so that $Z_k=Y_k+\sqrt{h/2}\,\chi_k$ for every $k\geqslant 1$.

With $\rho=\tilde\pi_h$, the MSE satisfies (5.1), and it remains to bound its time average error. Telescoping $\tilde\phi_h(Y_{k+1})-\tilde\phi_h(Y_k)$ along the chain of the Markovian form, with the solution $\tilde\phi_h$ of (4.18), gives the time-averaged error formula of [71, 106],

$$\frac{1}{N}\sum_{k=0}^{N-1}\bigl(\tilde f_h(Y_k)-\tilde\pi_h(\tilde f_h)\bigr)=\frac{\tilde\phi_h(Y_0)-\tilde\phi_h(Y_N)}{Nh}+\frac{1}{N}\sum_{k=0}^{N-1}\Upsilon_k+\frac{1}{N}\sum_{k=0}^{N-1}\Psi_k, \tag{6.1}$$

with the discretization term and the martingale difference

$$\Upsilon_k:=\frac{\tilde\phi_h(Y_{k+1})-\tilde\phi_h(\tilde Y_k(h))}{h},\qquad \Psi_k:=\frac{\tilde\phi_h(\tilde Y_k(h))-\tilde\phi_h(Y_k)}{h}+\tilde f_h(Y_k)-\tilde\pi_h(\tilde f_h). \tag{6.2}$$

The conditional mean of Υ_k, under $\mathbb{E}_k$ for (LM) and $\hat{\mathbb{E}}_k$ for (SGLM), is the local weak error of Lemmas 6 and 7 with $G=\tilde\phi_h$, divided by h; the lemmas apply because (4.19) bounds the derivatives of $\tilde\phi_h$ of orders 1 to 4 by CL. The term Ψ_k has conditional mean 0 by (4.18).

Theorem 2 *Let Assumptions 1 and 2 hold, and let $(Z_k)_{k\geqslant 0}$ be the samples of (LM) from a given $Z_0\in\mathbb{R}^d$. Then, for every $h\in(0,h_0]$ and every $N\geqslant 1$,*

$$\mathrm{MSE}(N,h)\leqslant CL^2\biggl[\bigl(1+|Z_0|\bigr)^2\frac{1}{Nh}+\bigl(1+|Z_0|\bigr)^{12}h^4\biggr]. \tag{6.3}$$

(Proof in Appendix D.1.)

The term $\frac{1}{Nh}$ is the order of the variance of a time average over a trajectory of length Nh, and h^4 is the square of the bias of order h^2 of (LM), which comes from Lemma 2 and (5.7). Balancing them gives, for $N\geqslant h_0^{-5}$, the step $h=N^{-\frac15}\leqslant h_0$ and $\mathrm{MSE}(N,h)\leqslant CL^2(1+|Z_0|)^{12}N^{-\frac45}$.

For the stochastic gradient, the explicit term of Lemma 7, divided by h, adds to the conditional mean of Υ_k a term of order Σh, whose square gives the order Σ^2h^2 in (6.4).

Theorem 3 *Let Assumptions 1, 2 and 3 hold, and let $(Z_k)_{k\geqslant 0}$ be the samples of (SGLM) from a given $Z_0\in\mathbb{R}^d$. Then, for every $h\in(0,h_0]$ and every $N\geqslant 1$,*

$$\mathrm{MSE}(N,h)\leqslant CL^2\biggl[\bigl(1+|Z_0|\bigr)^2\Bigl(\frac{1}{Nh}+\Sigma^2h^2\Bigr)+\bigl(1+|Z_0|\bigr)^{12}h^4\biggr]. \tag{6.4}$$

(Proof in Appendix D.2.)

With a batch of P independent stochastic gradients, the chain (SGLM) reads

$$Z_{k+1}=Z_k-\frac{h}{P}\sum_{p=1}^{P}\nabla\hat U\bigl(Z_k;\xi_k^{(p)}\bigr)+\sqrt{\frac{h}{2}}\,\bigl(\chi_k+\chi_{k+1}\bigr),$$

with $\xi_k^{(1)},\dots,\xi_k^{(P)}$ independent and drawn afresh at every step. By Minkowski's inequality, the averaged estimator keeps the bound $M(1+|x|)$ on its L^{12} norm and the bound on its Hessian, uniformly in P, and its variance is at most Σ/P, so it satisfies Assumption 3 with Σ/P in place of Σ; the proof of Theorem 3 is unchanged, and (6.4) holds with Σ^2h^2/P^2 in place of Σ^2h^2.

6.2 Numerical bias estimate of SGLM

The MSE bounds control the time average along one trajectory. This subsection identifies the long-time limit of that average for (SGLM), and its distance to π, the numerical bias. The Markovian form (SGLM') is ergodic by Harris's ergodic theorem [47], applied with the drift condition behind Lemma 3 and a minorization that uses the Hessian bound of Assumption 3.

Lemma 8 *Let Assumptions 1 and 3 hold. For every $h \in (0, h_0]$, (SGLM') has a unique invariant distribution μ_h, with $\mu_h(|y|^{12}) \leqslant C$. Let $\pi_h := \mu_h * \mathcal{N}(0, \frac{h}{2} I_d)$. Then, for every f satisfying Assumption 2 and every given $Z_0 \in \mathbb{R}^d$, the samples of (SGLM) satisfy*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} f(Z_k) = \pi_h(f) \qquad \textit{almost surely.} \tag{6.5}$$

(Proof in Appendix E.1.)

The measure π_h is determined by the chain (SGLM') alone, through μ_h, and does not depend on the initial value Z_0. Since (SGLM) is not Markovian, it has no invariant distribution in the usual sense; by (6.5), the measure π_h can be viewed as the numerical invariant distribution of SGLM.

The next theorem expands the numerical bias $\pi_h(f) - \pi(f)$ to the order h, with a coefficient given by $\tilde{\phi}_h$ and the covariance of the gradient noise, defined in (5.8) as

$$\mathcal{C}(x) = \mathbb{E}_\xi\Big[\big(\nabla\hat{U}(x;\xi) - \nabla U(x)\big) \otimes \big(\nabla\hat{U}(x;\xi) - \nabla U(x)\big)\Big], \qquad x \in \mathbb{R}^d.$$

The bias estimate follows the same strategy as that of SG-UBU in [71, Section 3.4]. Since $\mathcal{C}(x)$ is positive semidefinite, Assumption 3 gives $\|\mathcal{C}(x)\| \leqslant \mathrm{Tr}\big[\mathcal{C}(x)\big] \leqslant \Sigma$.

Theorem 4 *Let Assumptions 1, 2 and 3 hold. For every $h \in (0, h_0]$, with $Y_0 \sim \mu_h$, $\chi_0 \sim \mathcal{N}(0, I_d)$ independent of Y_0, and $Z_0 = Y_0 + \sqrt{h/2}\,\chi_0$, so that $Z_0 \sim \pi_h$, the numerical invariant distribution of (SGLM),*

$$\Big|\pi_h(f) - \pi(f) - \frac{h}{2}\,\mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big]\Big| \leqslant CLh^2. \tag{6.6}$$

In particular, $|\pi_h(f) - \pi(f)| \leqslant CL(\Sigma h + h^2)$. (Proof in Appendix E.2.)

The bound $CL(\Sigma h + h^2)$ of Theorem 4 follows from the uniform bound (4.19) on the modified discrete Poisson solution $\tilde{\phi}_h$, together with $\|\mathcal{C}(x)\| \leqslant \Sigma$. As $h \to 0$, the coefficient $\frac{1}{2}\,\mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big]$ of h in (6.6) converges to an expression in the solution ϕ_0 of the continuous Poisson equation (3.7).

Corollary 2 *Let Assumptions 1, 2 and 3 hold. Then, for (SGLM),*

$$\lim_{h\to 0} \frac{\pi_h(f) - \pi(f)}{h} = \frac{1}{2}\,\pi\big(\nabla^2\phi_0(x) : \mathcal{C}(x)\big). \tag{6.7}$$

(Proof in Appendix E.3.)

Corollary 2 gives a rigorous characterization of the linear part of the numerical bias of (SGLM): its coefficient of order h comes entirely from the covariance $\mathcal{C}(x)$ of the stochastic gradient, and vanishes with the full gradient. This is intrinsically different from (SGLD), whose numerical bias carries a first order discretization error even with the full gradient, so that its coefficient of order h remains nonzero for some test functions as the variance of the stochastic gradient tends to zero (Corollary 4). For SG-UBU, the corresponding limit is identified only formally in [71]; for SGLM, Corollary 2 makes the coefficient explicit and proves the limit.

Linear response theory [46, 52] also explains the coefficient (6.7). To first order in h, the gradient noise of (SGLM) adds the diffusion $\frac{h}{2}\,\mathcal{C} : \nabla^2$ to the generator (3.3), and the invariant distribution responds to this perturbation by $\frac{h}{2}\,\pi(\nabla^2\phi_0 : \mathcal{C})$ for the test function f. The invariant distributions of Langevin integrators are viewed in the same way, as perturbations of π depending on the step size, in [60]. With the solution of

the continuous Poisson equation (3.7) written as $\phi_0 = \int_0^\infty e^{t\mathcal{L}}\big(f - \pi(f)\big)\mathrm{d}t$, two integrations by parts give the Green–Kubo form [45, 52] of the coefficient,

$$\frac{1}{2}\,\pi\big(\nabla^2\phi_0 : \mathcal{C}\big) = \frac{1}{2}\int_0^\infty \pi\Big(\pi^{-1}\nabla^2 : (\mathcal{C}\pi)\; e^{t\mathcal{L}}\big(f - \pi(f)\big)\Big)\mathrm{d}t, \qquad \nabla^2 : (\mathcal{C}\pi) = \sum_{i,j=1}^{d} \partial_i\partial_j\big(\mathcal{C}_{ij}\,\pi\big),$$

that is, the time integral of the correlation between $\pi^{-1}\nabla^2 : (\mathcal{C}\pi)$ at the initial value and f at the time t, along (1.1) started from π.

6.3 Numerical bias comparison with SGLD

The MSE analysis through the discrete Poisson equation applies to (SGLD) without extra effort: the samples of SGLD form a Markov chain, so no modified class is needed, and the time-averaged error formula (6.1) holds with Z_k, f, $\pi(f)$, the discrete Poisson solution ϕ_h of (3.5) and the solution X_h of (1.1) from $X_0 = Z_k$ in place of Y_k, $\tilde{f}_h$, $\tilde{\pi}_h(\tilde{f}_h)$, $\tilde{\phi}_h$ and $\tilde{Y}_k(h)$. For the samples of SGLD from a given $Z_0 \in \mathbb{R}^d$, it gives

$$\mathrm{MSE}(N,h) = \mathcal{O}\Big(\frac{1}{Nh} + \Sigma^2 h^2 + h^2\Big), \tag{6.8}$$

with the constant depending also on L and Z_0. The term $\Sigma^2 h^2$ comes from the gradient noise, as in (6.4), and the last term h^2 from the discretization error of EM, where (6.4) has h^4. Since $\Sigma \leqslant M^2$, (6.8) reduces to $\mathrm{MSE}(N,h) = \mathcal{O}\big((Nh)^{-1} + h^2\big)$, the bound already proved through the continuous Poisson equation (3.7) [102, Theorem 9], [18, Theorem 3], and we do not restate its proof.

The arguments of Lemma 8, based on Harris's ergodic theorem, show that (SGLD) has a unique invariant distribution $\bar{\pi}_h$, with $\bar{\pi}_h(|x|^{12}) \leqslant C$, and that the time averages of f along SGLD converge almost surely to $\bar{\pi}_h(f)$ (Lemma 20 in Appendix F). The invariant distribution $\bar{\pi}_h$ differs from the numerical invariant distribution π_h of (SGLM) in the coefficient of order h of the numerical bias. In the next theorem, this coefficient carries, besides the covariance $\mathcal{C}(x)$ of the gradient noise, the one-step error of EM against (1.1), through $\nabla^2 U\,\nabla U - \nabla\Delta U$ in the mean of the step and the Hessian $\nabla^2 U$ in its moment of order 2.

Theorem 5 *Let Assumptions 1, 2 and 3 hold. For every $h \in (0, h_0]$, with $Z_0 \sim \bar{\pi}_h$, the invariant distribution of (SGLD),*

$$\begin{aligned}\Big|\bar{\pi}_h(f) - \pi(f) - \frac{h}{2}\,\mathbb{E}\Big[\nabla^2\phi_h(Z_0) : \big(2\nabla^2 U(Z_0) + \mathcal{C}(Z_0)\big)\Big] \\ + \frac{h}{2}\,\mathbb{E}\Big[\nabla\phi_h(Z_0)\cdot\big(\nabla^2 U(Z_0)\,\nabla U(Z_0) - \nabla\Delta U(Z_0)\big)\Big]\Big| \leqslant CLh^2.\end{aligned} \tag{6.9}$$

In particular, $|\bar{\pi}_h(f) - \pi(f)| \leqslant CLh$. (Proof in Appendix F.2.)

As $h \to 0$, the coefficient of h in (6.9) converges to an expression in the solution ϕ_0 of (3.7), in which only the Hessian $\nabla^2\phi_0$ remains.

Corollary 3 *Let Assumptions 1, 2 and 3 hold. Then, for (SGLD),*

$$\lim_{h\to 0}\frac{\bar{\pi}_h(f) - \pi(f)}{h} = \frac{1}{2}\,\pi\Big(\nabla^2\phi_0(x) : \big(\nabla^2 U(x) + \mathcal{C}(x)\big)\Big). \tag{6.10}$$

(Proof in Appendix F.3.)

Corollary 3 is not entirely new: the coefficient of order h of the numerical bias of (SGLD) has been studied formally in [102, Section 4.1], for $d = 1$ and under the assumption that the chain is ergodic, where the gradient noise adds to the coefficient of EM a term in the variance of the stochastic gradient, and the modified SGLD of [102, Section 4.2] removes this term. Corollary 3 proves the limit in $\mathbb{R}^d$ and writes the coefficient in terms of the solution ϕ_0 of (3.7). The (SGLM) integrator treats the discretization bias in a different and clean way: it preserves the second order discretization error of the LM integrator, so its coefficient (6.7) carries the gradient noise only.

Remark 2 *Corollary 3 also requires the stochastic gradient to have a finite moment of high order, as in Assumption 3. For (SGLD) with a strongly convex potential, [86, Theorem 1] bounds the numerical bias in $\mathcal{W}_2$ by a term of order h if the moment of order 4 of the gradient noise under π is finite; if only the moment of order 2 is finite, the bound could deteriorate to the order $h^{\frac{1}{2}}$.*

The coefficients (6.7) of (SGLM) and (6.10) of (SGLD) differ by $\frac{1}{2}\,\pi\big(\nabla^2\phi_0(x) : \nabla^2 U(x)\big)$, the contribution of the discretization error of EM. This term does not depend on the gradient noise: it is the coefficient of SGLD with the full gradient, and it remains at every batch size. The coefficient of SGLM, by contrast, vanishes with $\mathcal{C}(x)$, and with a batch of P independent stochastic gradients, whose average has the covariance $\mathcal{C}(x)/P$, it is divided by P. For a given test function, the two coefficients are not ordered in absolute value, since $\nabla^2\phi_0$ is indefinite in general and the terms in $\nabla^2 U(x)$ and in $\mathcal{C}(x)$ can have opposite signs. Section 6.4 compares the suprema of their absolute values over the test functions with $\|\nabla f\|_\infty \leqslant 1$, and Lemma 10 gives a gradient noise for which even the supremum of the absolute value of the coefficient of SGLD over these test functions is smaller than that of SGLM.

The expectation in (6.9) contains the gradient $\nabla\phi_h$, through the error of EM in the mean of the step, while the limit (6.10) of Corollary 3 contains the Hessian $\nabla^2\phi_0$ only. The two forms are connected by an integration by parts that holds under π and is not available under $\bar\pi_h$. By (3.3), for each component i,

$$\mathcal{L}(\partial_i U) = \Delta\partial_i U - \nabla U\cdot\nabla\partial_i U = \big(\nabla\Delta U - \nabla^2 U\,\nabla U\big)_i,$$

so the vector multiplying $\nabla\phi_h$ in (6.9) is, componentwise, minus the generator applied to ∇U,

$$\nabla^2 U\,\nabla U - \nabla\Delta U = -\mathcal{L}\nabla U. \tag{6.11}$$

For every $\psi \in C^3(\mathbb{R}^d)$ with bounded derivatives of orders 1 to 3, (3.4) with $g_1 = \partial_i\psi$ and $g_2 = \partial_i U$, summed over i, and (6.11) give

$$\begin{aligned}\pi\Big(\nabla\psi(x)\cdot\big(\nabla^2 U(x)\,\nabla U(x) - \nabla\Delta U(x)\big)\Big) &= -\sum_{i=1}^d \pi\big(\partial_i\psi(x)\,\mathcal{L}\partial_i U(x)\big)\\ &= \sum_{i=1}^d \pi\big(\nabla\partial_i\psi(x)\cdot\nabla\partial_i U(x)\big) = \pi\big(\nabla^2\psi(x) : \nabla^2 U(x)\big).\end{aligned} \tag{6.12}$$

At a fixed step size, the expectation in (6.9) is taken under $\bar\pi_h$, while (6.12) with $\psi = \phi_h$ holds under π. Moving the terms of order h from $\bar\pi_h$ to π within the remainder CLh^2 requires Theorem 5 for a test function built from $\nabla\phi_h$ and $\nabla^2\phi_h$, whose derivatives of orders 1 to 4 involve the derivatives of ϕ_h up to order 6, beyond the order 4 of Theorem 1. The gradient $\nabla\phi_h$ therefore stays in (6.9).

As $h \to 0$, $\bar\pi_h$ converges weakly to π with moments bounded uniformly in h, and $\nabla\phi_h$ and $\nabla^2\phi_h$ converge to $\nabla\phi_0$ and $\nabla^2\phi_0$ uniformly on compact sets, so the expectation in (6.9) converges to the same expression under π. There (6.12) with $\psi = \phi_0$ applies, and the term in $\nabla\phi_0$ cancels half of the term in $2\nabla^2 U$. Theorem 5 thus holds at every step size $h \in (0, h_0]$ with a remainder of order h^2 but keeps both $\nabla\phi_h$ and $\nabla^2\phi_h$, while the Hessian form of the coefficient in Corollary 3 is reached only in the limit $h \to 0$, without a rate.

The gap between Theorem 5 and Corollary 3 also reveals a feature of the discrete Poisson equation. The solution ϕ_h of (3.6) compares one step of (SGLD) with one step of (1.1), so only the differences of their moments enter, and the derivatives of ϕ_h up to order 4 suffice for the nonasymptotic bound (6.9). The solution ϕ_0 of (3.7) compares one step of SGLD with the generator $\mathcal{L}$ instead: the coefficient of order h then involves the derivatives of ϕ_0 up to order 4, and a remainder of order h^2 requires those up to order 6, beyond Theorem 1. With ϕ_0 alone, the terms that do not cancel at a fixed step size are therefore hard to characterize under the present assumptions.

6.4 Explicit coefficient in Wasserstein-1 error

The coefficients (6.7) and (6.10) depend on the test function f through $\nabla^2\phi_0$. The suprema of their absolute values over the test functions with $\|\nabla f\|_\infty \leqslant 1$ are the limits of $\mathcal{W}_1(\pi_h,\pi)/h$ and $\mathcal{W}_1(\bar\pi_h,\pi)/h$ as $h \to 0$ (Corollary 4). To evaluate these suprema, both coefficients are written as $\pi(f\psi)$, with a function ψ that does not depend on f. The construction of ψ uses the Poincaré inequality of π.

Lemma 9 *Let Assumption 1 hold. There is a constant C_P, depending only on m, M, R and d, such that*

$$\pi\big((g-\pi(g))^2\big) \leqslant C_P\,\pi\big(|\nabla g|^2\big) \tag{6.13}$$

for every $g \in C^1(\mathbb{R}^d)$ with a bounded gradient.

(Directly derived from [5, Corollary 1.6 and Theorem 1.4]; see Appendix G.1 for further discussion.)

For a bounded measurable function $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ such that $\mathcal{A}(x)$ is a symmetric matrix for every $x \in \mathbb{R}^d$, define the Riesz representor of $\mathcal{A}$ as the function $\psi_{\mathcal{A}} \in L^2(\pi)$ with $\pi(\psi_{\mathcal{A}}) = 0$ and

$$\pi\big(\psi_{\mathcal{A}}(x)\,(-\mathcal{L}\omega)(x)\big) = \pi\big(\nabla^2\omega(x) : \mathcal{A}(x)\big) \tag{6.14}$$

for every $\omega \in C^4(\mathbb{R}^d)$ with bounded derivatives of orders 1 to 4. The next theorem shows that the Riesz representor is well defined.

Theorem 6 *Let Assumption 1 hold, and let $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be bounded and measurable, with $\mathcal{A}(x)$ a symmetric matrix for every $x \in \mathbb{R}^d$. There is a unique $\psi_{\mathcal{A}} \in L^2(\pi)$ with $\pi(\psi_{\mathcal{A}}) = 0$ that satisfies (6.14) for every $\omega \in C^4(\mathbb{R}^d)$ with bounded derivatives of orders 1 to 4, the map $\mathcal{A} \mapsto \psi_{\mathcal{A}}$ is linear, and*

$$\pi\big(\psi_{\mathcal{A}}^2\big) \leqslant d\,(1 + MC_P)\|\mathcal{A}\|_\infty^2, \tag{6.15}$$

with C_P the constant of Lemma 9. (Proof in Appendix G.1.)

For f satisfying Assumption 2, (6.14) with $\omega = \phi_0$ gives $\pi(\nabla^2\phi_0 : \mathcal{A}) = \pi(f\psi_{\mathcal{A}})$, so the coefficients (6.7) and (6.10) read

$$\lim_{h\to 0}\frac{\pi_h(f)-\pi(f)}{h} = \frac{1}{2}\,\pi(f\psi_{\mathcal{C}}), \qquad \lim_{h\to 0}\frac{\bar\pi_h(f)-\pi(f)}{h} = \frac{1}{2}\,\pi\big(f\,(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\big). \tag{6.16}$$

The representor of the Hessian $\nabla^2 U$ is explicit. The vector $\nabla^2 U\,\nabla U - \nabla\Delta U$ in (6.11) is the gradient of $\frac{1}{2}|\nabla U|^2 - \Delta U$, so (6.12) and (3.4) give

$$\pi\big(\nabla^2\omega : \nabla^2 U\big) = \pi\Big(\nabla\omega\cdot\nabla\Big(\frac{1}{2}|\nabla U|^2 - \Delta U\Big)\Big) = \pi\Big(\Big(\frac{1}{2}|\nabla U|^2 - \Delta U\Big)(-\mathcal{L}\omega)\Big).$$

Since $\pi(\mathcal{L}\omega) = 0$, the mean of $\frac{1}{2}|\nabla U|^2 - \Delta U$ may be subtracted; it equals $-\frac{1}{2}\pi(\Delta U)$, since $\pi(\mathcal{L}U) = 0$ gives $\pi(|\nabla U|^2) = \pi(\Delta U)$. Hence (6.14) holds for $\mathcal{A} = \nabla^2 U$ with

$$\psi_{\nabla^2 U}(x) = \frac{1}{2}\,|\nabla U(x)|^2 - \Delta U(x) + \frac{1}{2}\,\pi(\Delta U). \tag{6.17}$$

Through the representors, the normalized biases $(\pi_h - \pi)/h$ and $(\bar\pi_h - \pi)/h$ converge to the signed measures $\frac{1}{2}\psi_{\mathcal{C}}\,\pi$ and $\frac{1}{2}(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi$ in the Kantorovich–Rubinstein norm, at the rate $h^{\frac{1}{2}}(1 + \log\frac{1}{h})$.

Theorem 7 *Let Assumptions 1 and 3 hold. Then, for every $h \in (0, h_0]$, the numerical invariant distribution π_h of (SGLM) and the invariant distribution $\bar\pi_h$ of (SGLD) satisfy*

$$\Big\|\pi_h - \pi - \frac{h}{2}\,\psi_{\mathcal{C}}\,\pi\Big\|_{\mathrm{KR}} \leqslant Ch^{\frac{3}{2}}\Big(1 + \log\frac{1}{h}\Big), \tag{6.18}$$

$$\Big\|\bar\pi_h - \pi - \frac{h}{2}\,(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\Big\|_{\mathrm{KR}} \leqslant Ch^{\frac{3}{2}}\Big(1 + \log\frac{1}{h}\Big). \tag{6.19}$$

(Proof in Appendix G.3.)

Since $\mathcal{W}_1(\pi_h, \pi) = \|\pi_h - \pi\|_{\mathrm{KR}}$ and $\mathcal{W}_1(\bar\pi_h, \pi) = \|\bar\pi_h - \pi\|_{\mathrm{KR}}$, the triangle inequality for $\|\cdot\|_{\mathrm{KR}}$ and Theorem 7 give, for every $h \in (0, h_0]$,

$$\Big|\mathcal{W}_1(\pi_h,\pi) - \frac{h}{2}\,\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}\Big| + \Big|\mathcal{W}_1(\bar\pi_h,\pi) - \frac{h}{2}\,\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}}\Big| \leqslant Ch^{\frac{3}{2}}\Big(1 + \log\frac{1}{h}\Big). \tag{6.20}$$

The bound (6.20) determines the Wasserstein-1 errors to first order in h, but it is not guaranteed to be sharp in the discretization error: its remainder is only of order $h^{\frac{3}{2}}$, up to the factor $1 + \log\frac{1}{h}$. Enhancing this order to h^2 is left for future studies; Remark 3 discusses the globally strongly convex case.

Remark 3 *When U is globally strongly convex, that is, $R = 0$ in Assumption 1, and the covariance $\mathcal{C}(x)$ is smooth, we expect the remainder of (6.18), (6.19) and (6.20) to improve to Ch^2. The factor $h^{\frac{1}{2}}(1 + \log \frac{1}{h})$ is lost where a step of the chain acts on a Lipschitz test function and where $\psi_{\mathcal{C}}$ is only known to lie in $L^2(\pi)$. Under global strong convexity, the derivatives of the solution of (1.1) with respect to its initial point decay at the rate e^{-mt}, so $\psi_{\mathcal{C}}$, like the explicit $\psi_{\nabla^2 U}$ of (6.17), should be smooth with polynomially growing derivatives. The steps of the chain can then be applied to the smooth measures $\psi_{\mathcal{C}}\,\pi$ and $\psi_{\nabla^2 U}\,\pi$ instead of the test function, and each of these losses would become a factor h.*

Since the remainder in (6.20) is $o(h)$, dividing (6.20) by h and letting $h \to 0$ gives the explicit limits of $\mathcal{W}_1(\pi_h, \pi)/h$ and $\mathcal{W}_1(\bar{\pi}_h, \pi)/h$.

Corollary 4 *Let Assumptions 1 and 3 hold. Then the numerical invariant distribution π_h of (SGLM) and the invariant distribution $\bar{\pi}_h$ of (SGLD) satisfy*

$$\lim_{h\to 0} \frac{\mathcal{W}_1(\pi_h, \pi)}{h} = \frac{1}{2}\,\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}, \qquad \lim_{h\to 0} \frac{\mathcal{W}_1(\bar{\pi}_h, \pi)}{h} = \frac{1}{2}\,\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}}.$$

Moreover, $\|\psi_{\nabla^2 U}\,\pi\|_{\mathrm{KR}} > 0$, and $\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}} > 0$ if $\mathcal{C}(x)$ does not vanish identically. (Proof in Appendix G.4.)

Corollary 4 is nontrivial: it gives an accurate characterization of the coefficient of h in the Wasserstein-1 error, as an equality. This makes it distinct from the mainstream coupling methods of numerical analysis, which bound the Wasserstein-1 error from above without identifying its coefficient. Although the representor $\psi_{\mathcal{C}}$ is difficult to evaluate in practice, Corollary 4 and (6.20) are the decisive answer to how much SGLM and SGLD shift the invariant distribution: up to the remainder $Ch^{\frac{3}{2}}(1 + \log \frac{1}{h})$ of (6.20), the shift in the Wasserstein-1 distance is $\frac{h}{2}\,\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}$ for SGLM and $\frac{h}{2}\,\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}}$ for SGLD.

Corollary 4 thus compares SGLM with SGLD through the representors: the limit of $\mathcal{W}_1(\pi_h, \pi)/h$ is at most that of $\mathcal{W}_1(\bar{\pi}_h, \pi)/h$ exactly when

$$\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}} \geqslant \|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}. \tag{6.21}$$

This inequality fails in general: Lemma 10 gives a gradient noise of large amplitude, supported away from the minimizer of U, that violates it.

Lemma 10 *Let $d = 1$, $U(x) = \frac{1}{2}x^2$ and $\nabla\hat{U}(x;\xi) = x + 10\,\xi\,a(x)$, where ξ takes the values 1 and -1 with probability $\frac{1}{2}$ each and*

$$a(x) := \exp\Big(1 - \frac{1}{1 - 4(|x| - 2)^2}\Big) \quad \text{if } \big||x| - 2\big| < \frac{1}{2}, \qquad a(x) := 0 \quad \text{otherwise.} \tag{6.22}$$

Then Assumptions 1 and 3 hold, and

$$\frac{1}{2}\,\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}} \approx 1.230 > 1.115 \approx \frac{1}{2}\,\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}},$$

so (6.21) fails. (Proof in Appendix G.5.)

With a batch of P independent stochastic gradients, $\mathcal{C}(x)/P$ replaces the covariance $\mathcal{C}(x)$ of a single stochastic gradient. As P grows, the limit of $\mathcal{W}_1(\pi_h, \pi)/h$ tends to 0, while that of $\mathcal{W}_1(\bar{\pi}_h, \pi)/h$ tends to $\frac{1}{2}\|\psi_{\nabla^2 U}\,\pi\|_{\mathrm{KR}} > 0$; for a strongly convex potential, (6.21) holds once P reaches an explicit threshold.

Theorem 8 *Let Assumptions 1 with $R = 0$ and 3 hold, so that $\nabla^2 U(x) \succeq m I_d$ for every $x \in \mathbb{R}^d$, and let $\mathcal{C}(x)$ be the covariance of a single stochastic gradient. Then, for every batch size P satisfying*

$$P \geqslant \frac{9\,\|\nabla^2 U\|_\infty\,\|\mathcal{C}\|_\infty}{m^2} = \frac{9}{m^2}\sup_{x\in\mathbb{R}^d}\|\nabla^2 U(x)\|\sup_{x\in\mathbb{R}^d}\|\mathcal{C}(x)\|, \tag{6.23}$$

the limit of $\mathcal{W}_1(\pi_h, \pi)/h$ for SGLM with a batch of size P is at most that of $\mathcal{W}_1(\bar{\pi}_h, \pi)/h$ for SGLD with the same batch, which by Corollary 4, with $\psi_{\mathcal{C}}/P$ in place of $\psi_{\mathcal{C}}$, reads

$$\Big\|\Big(\psi_{\nabla^2 U} + \frac{1}{P}\psi_{\mathcal{C}}\Big)\pi\Big\|_{\mathrm{KR}} \geqslant \frac{1}{P}\,\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}. \tag{6.24}$$

(Proof in Appendix G.6.)

The norm $\|\cdot\|$ in (6.23) is the operator norm of Section 3.1, so the lower bound of P does not depend explicitly on the dimension d; it depends only on the eigenvalues of $\nabla^2 U(x)$ and $\mathcal{C}(x)$.

7 Bias correction via the Riesz representors

By (6.16), the numerical bias $\pi_h(f) - \pi(f)$ of SGLM equals $\frac{h}{2}\,\pi(f\psi_{\mathcal{C}})$ to first order in h, so it is removed once $\pi(f\psi_{\mathcal{C}})$ is computed. Subtracting $\frac{h}{2}\,\pi(f\psi_{\mathcal{C}})$ from $\pi_h(f)$ requires the value of $\pi(f\psi_{\mathcal{C}})$ for the given test function f, not the representor $\psi_{\mathcal{C}}$ itself. This section describes three approaches to bias correction. For the covariances of Section 7.1, $\pi(f\psi_{\mathcal{C}})$ is computed with the explicit representor $\psi_{\mathcal{C}}$. The Richardson–Romberg extrapolation of Section 7.2 requires only the time averages at the step sizes h and $2h$. The shifting correction of Section 7.3 improves the samples themselves by solving the function $\varphi_{\mathcal{C}}$ from a variational form. Section 7.4 describes the computation of the Kantorovich–Rubinstein norms that give the coefficients of the Wasserstein-1 errors. The three approaches also apply to SGLD, with $\bar{\pi}_h$ and $\psi_{\nabla^2 U} + \psi_{\mathcal{C}}$ in place of π_h and $\psi_{\mathcal{C}}$ by (6.16), where $\psi_{\nabla^2 U}$ is explicit by (6.17); we focus on SGLM for simplicity.

7.1 Covariances with an explicit representor

Under Assumptions 1 and 3, define the vector function $b_{\mathcal{C}}(x)$ by

$$b_{\mathcal{C}}(x) := \mathcal{C}(x)\,\nabla U(x) - \operatorname{div}\mathcal{C}(x), \tag{7.1}$$

with $(\operatorname{div}\mathcal{C})_j = \sum_{i=1}^{d} \partial_i \mathcal{C}_{ij}$. An integration by parts against $\pi \propto e^{-U}$ gives, for ω as in (6.14),

$$\pi\big(\nabla^2\omega : \mathcal{C}\big) = \pi\big(\nabla\omega \cdot b_{\mathcal{C}}\big). \tag{7.2}$$

When $b_{\mathcal{C}} = \nabla g$ for a function g whose derivatives grow at most polynomially, (3.4) and the uniqueness in Theorem 6 give $\psi_{\mathcal{C}} = g - \pi(g)$. Table 3 lists covariances and potentials for which $b_{\mathcal{C}}$ is a gradient.

covariance $\mathcal{C}(x)$	potential U	representor $\psi_{\mathcal{C}}(x)$, up to a constant
$a\big(U(x)\big)\, I_d$	any	$\displaystyle\int_0^{U(x)} a(s)\,\mathrm{d}s - a\big(U(x)\big)$
$\mathcal{C}(x) \in \mathbb{R}$, $d = 1$	any	$\displaystyle\int_0^{x} \mathcal{C}(y)\,U'(y)\,\mathrm{d}y - \mathcal{C}(x)$
$\operatorname{diag}\big(c_1(x_1), \dots, c_d(x_d)\big)$	$U(x) = \displaystyle\sum_{i=1}^{d} U_i(x_i)$	$\displaystyle\sum_{i=1}^{d}\Big(\int_0^{x_i} c_i(y)\,U_i'(y)\,\mathrm{d}y - c_i(x_i)\Big)$
$a_1(\lvert x\rvert)\, I_d + a_2(\lvert x\rvert)\, \frac{x\otimes x}{\lvert x\rvert^2}$	$U(x) = u(\lvert x\rvert)$	$\displaystyle\int_0^{\lvert x\rvert}\Big((a_1+a_2)(r)\,u'(r) - \frac{(d-1)\,a_2(r)}{r}\Big)\mathrm{d}r - (a_1+a_2)(\lvert x\rvert)$

Table 3: Covariances whose Riesz representor $\psi_{\mathcal{C}}$ is explicit, with $\psi_{\mathcal{C}}$ given up to the additive constant fixed by $\pi(\psi_{\mathcal{C}}) = 0$. The functions a, a_1, a_2 and $c_1, \dots, c_d$ are of class C^2, and $a_2(r) = O(r^2)$ as $r \to 0$.

With a constant $a = \sigma^2$, the row $a(U(x))\, I_d$ of Table 3 gives $\psi_{\mathcal{C}} = \sigma^2\big(U - \pi(U)\big)$. For a general covariance in dimension $d \geqslant 2$, the field $b_{\mathcal{C}}$ is not a gradient and $\psi_{\mathcal{C}}$ has no closed form; the extrapolation of Section 7.2 and the energy (7.8) of Section 7.3 do not require $\psi_{\mathcal{C}}$ directly.

7.2 Richardson–Romberg extrapolation

Without the representor, the term of order h of the numerical bias of a given test function is removed by extrapolation in the step size, as for the invariant distribution of an ergodic diffusion [64] and for SGLD [35]. For $h \in (0, h_0/2]$ and a function f with $\|\nabla f\|_\infty < \infty$, (6.18) at the step sizes h and $2h$ gives

$$\left|\frac{\pi_{2h}(f) - \pi_h(f)}{h} - \frac{1}{2}\,\pi(f\psi_{\mathcal{C}})\right| \leqslant C\|\nabla f\|_\infty\, h^{\frac{1}{2}}\Big(1 + \log\frac{1}{h}\Big), \tag{7.3}$$

so the coefficient $\frac{1}{2}\,\pi(f\psi_{\mathcal{C}})$ is approximated by $(\pi_{2h}(f) - \pi_h(f))/h$, computed from the time averages of f along runs of (SGLM) with the step sizes h and $2h$; the error is of order h when the remainder of (6.18) is

of order h^2, as expected in Remark 3. Subtracting $\pi_{2h}(f) - \pi_h(f)$, which is h times the approximation of $\frac{1}{2}\,\pi(f\psi_{\mathcal{C}})$ in (7.3), from $\pi_h(f)$ gives the approximation $2\pi_h(f) - \pi_{2h}(f)$ of $\pi(f)$, and (6.18) gives

$$\big|2\pi_h(f) - \pi_{2h}(f) - \pi(f)\big| \leqslant C\|\nabla f\|_\infty\, h^{\frac{3}{2}}\Big(1 + \log\frac{1}{h}\Big), \tag{7.4}$$

and the error is of order h^2 when the remainder of (6.18) is of order h^2.

Extrapolation corrects the expectations of given test functions but not the samples, whose laws remain π_h and π_{2h}; it does not raise the order of the Wasserstein-1 error in (6.20). Moreover, (7.3) holds for each given test function and does not mean that the empirical samples of π_h and π_{2h} approximate $\psi_{\mathcal{C}}$ itself: for the gradient noise $\sigma\xi$ with $\xi \sim \mathcal{N}(0, I_d)$, $\psi_{\mathcal{C}} = \sigma^2\big(U - \pi(U)\big)$ by Table 3, which is not read off the samples.

7.3 Postprocessed shifting correction

The extrapolation of Section 7.2 improves the expectations $\pi_h(f)$, but not the samples of SGLM themselves. We instead define a vector field that shifts the samples of SGLM to more accurate ones. The vector field is $h\,\nabla\varphi_{\mathcal{C}}$, with $\varphi_{\mathcal{C}}$ the solution of

$$\mathcal{L}\varphi_{\mathcal{C}} = \frac{1}{2}\,\psi_{\mathcal{C}}, \qquad \pi(\varphi_{\mathcal{C}}) = 0. \tag{7.5}$$

By (3.4), $\varphi_{\mathcal{C}}$ is the minimizer of the functional

$$\mathcal{E}(\varphi) := \frac{1}{2}\,\pi\big(|\nabla\varphi|^2\big) + \frac{1}{2}\,\pi\big(\psi_{\mathcal{C}}\,\varphi\big), \qquad \pi(\varphi) = 0, \tag{7.6}$$

and since $\psi_{\mathcal{C}} \in L^2(\pi)$ and $\pi(\psi_{\mathcal{C}}) = 0$, the minimizer over the functions φ with $\varphi, |\nabla\varphi| \in L^2(\pi)$ exists and is unique by the Poincaré inequality (6.13) and the Riesz representation theorem. The shifting correction replaces a sample $Z \sim \pi_h$ of SGLM by the postprocessed sample

$$Z + h\,\nabla\varphi_{\mathcal{C}}(Z), \tag{7.7}$$

where the map $x \mapsto x + h\,\nabla\varphi_{\mathcal{C}}(x)$ is a postprocessor in the sense of [101], that is, a change of variables applied to the samples to raise the order for the invariant distribution. Formally, let f be a smooth test function, and assume that the remainder of (6.18) is of order h^2, as expected in Remark 3. A Taylor expansion of f at x gives

$$f\big(x + h\nabla\varphi_{\mathcal{C}}(x)\big) = f(x) + h\,\nabla f(x)\cdot\nabla\varphi_{\mathcal{C}}(x) + \mathcal{O}(h^2),$$

and the expectation under π_h gives

$$\pi_h\big(f(x + h\nabla\varphi_{\mathcal{C}}(x))\big) = \pi_h(f) + h\,\pi_h(\nabla f\cdot\nabla\varphi_{\mathcal{C}}) + \mathcal{O}(h^2).$$

By (6.18) applied to f and to $\nabla f\cdot\nabla\varphi_{\mathcal{C}}$,

$$\pi_h(f) = \pi(f) + \frac{h}{2}\,\pi(f\psi_{\mathcal{C}}) + \mathcal{O}(h^2), \qquad \pi_h(\nabla f\cdot\nabla\varphi_{\mathcal{C}}) = \pi(\nabla f\cdot\nabla\varphi_{\mathcal{C}}) + \mathcal{O}(h),$$

so that

$$\pi_h\big(f(x + h\nabla\varphi_{\mathcal{C}}(x))\big) = \pi(f) + \frac{h}{2}\,\pi(f\psi_{\mathcal{C}}) + h\,\pi(\nabla f\cdot\nabla\varphi_{\mathcal{C}}) + \mathcal{O}(h^2).$$

By (3.4) with $g_1 = f$ and $g_2 = \varphi_{\mathcal{C}}$, and by (7.5),

$$\pi(\nabla f\cdot\nabla\varphi_{\mathcal{C}}) = -\pi(f\,\mathcal{L}\varphi_{\mathcal{C}}) = -\frac{1}{2}\,\pi(f\psi_{\mathcal{C}}),$$

so the terms $\frac{h}{2}\,\pi(f\psi_{\mathcal{C}})$ and $h\,\pi(\nabla f\cdot\nabla\varphi_{\mathcal{C}})$ cancel, and formally

$$\pi_h\big(f(x + h\nabla\varphi_{\mathcal{C}}(x))\big) = \pi(f) + \mathcal{O}(h^2).$$

The error of the postprocessed samples is thus of order h^2 for every test function at once. The shifting correction formally raises the order of the Wasserstein-1 error in (6.20) to h^2. The rigorous study of the shifting correction is left for future work.

When $\psi_{\mathcal{C}}$ is explicit, as in Section 7.1, $\varphi_{\mathcal{C}}$ is approximated by minimizing (7.6) over a class of functions, with π replaced by the average over the empirical samples of π_h. In general, the energy (7.6) uses $\psi_{\mathcal{C}}$ only through $\frac{1}{2}\,\pi(\psi_{\mathcal{C}}\,\varphi)$, the coefficient (6.16) of the test function φ, which (7.3) approximates. Replacing $\frac{1}{2}\,\pi(\psi_{\mathcal{C}}\,\varphi)$ by $(\pi_{2h}(\varphi) - \pi_h(\varphi))/h$, and π by π_h in $\frac{1}{2}\,\pi(|\nabla\varphi|^2)$, gives

$$\mathcal{E}_h(\varphi) := \frac{1}{2}\,\pi_h\big(|\nabla\varphi|^2\big) + \frac{\pi_{2h}(\varphi) - \pi_h(\varphi)}{h}. \tag{7.8}$$

The energy $\mathcal{E}_h$ requires only the empirical samples of π_h and π_{2h}, from runs of (SGLM) with the step sizes h and $2h$, so it is completely data-driven: it involves neither U, ∇U nor $\mathcal{C}(x)$, and one minimization gives the shift. It does not change when a constant is added to φ, so its minimization needs no constraint on φ. A minimizer φ of $\mathcal{E}_h$ satisfies $h\,\pi_h(\nabla f \cdot \nabla\varphi) = \pi_h(f) - \pi_{2h}(f)$ for every f, so the shift $x \mapsto x + h\nabla\varphi(x)$ formally carries $\pi_h(f)$ to the approximation $2\pi_h(f) - \pi_{2h}(f)$ of (7.4), up to terms of order h^2, for every f at once. The step size h cannot be too small for the minimization to be stable: the fluctuation of the empirical averages that replace $\pi_{2h}(\varphi)$ and $\pi_h(\varphi)$ is divided by h in (7.8), and the minimizer of $\mathcal{E}_h$ with these empirical averages fits this fluctuation.

Remark 4 *The loss function $\mathcal{E}_h$ of (7.8) is closely related to the loss in action matching [83], which learns, from samples of a family of distributions $(q_t)_{t\in[0,1]}$, a vector field ∇s_t that transports them, by minimizing*

$$q_0(s_0) - q_1(s_1) + \int_0^1 q_t\Big(\frac{1}{2}\,|\nabla s_t|^2 + \partial_t s_t\Big)\,\mathrm{d}t$$

over functions $s_t(x)$. With $q_0 = \pi_{2h}$, $q_1 = \pi_h$, the function $s_t = h\varphi$ independent of t, and q_t replaced by π_h in the integral, this loss equals $h^2\,\mathcal{E}_h(\varphi)$. A loss of the same form, whose 2 terms are expectations estimated by sampling, is also written in [12]. In these works, the vector field transports the samples between the given distributions, while the shifting correction applies it beyond them.

Since $\pi_{2h} - \pi_h$ is of order h, a minimizer φ of (7.8) formally gives $\pi_{2h}\big(f(x + h\nabla\varphi(x))\big) = \pi_h(f) + \mathcal{O}(h^2)$, so the map $x \mapsto x + h\nabla\varphi(x)$ carries π_{2h} to π_h, and the shifting correction applies the same map once more, from π_h toward π. Formally, up to terms of order h^2, the shifting correction is thus the extrapolation at the time 2 of the Wasserstein geodesic from π_{2h} at the time 0 to π_h at the time 1, the operation used in [39, 40] for time discretizations of Wasserstein gradient flows.

7.4 Calculation of the Kantorovich–Rubinstein norm

The coefficients of the Wasserstein-1 errors in Corollary 4 are $\frac{1}{2}\|\psi\,\pi\|_{\mathrm{KR}}$, with $\psi = \psi_{\mathcal{C}}$ for SGLM and $\psi = \psi_{\nabla^2 U} + \psi_{\mathcal{C}}$ for SGLD, so they require the norm itself and not only the representor. On a convex compact domain, with no flux through its boundary, the supremum that defines the norm in Section 3.1 equals the least cost of a flux F that carries the negative part of $\psi\pi$ onto its positive part [68, (2)]:

$$\|\psi\pi\|_{\mathrm{KR}} = \min\Big\{\int_{\mathbb{R}^d} |F(x)|\,\mathrm{d}x : \operatorname{div} F = -\psi\,\pi\Big\}. \tag{7.9}$$

For the computation, $\psi\,\pi$ is restricted to a box centered at the origin that carries all but a negligible part of π, with n cells per side: the function f of the supremum is placed at the centers of the cells, with forward differences of Euclidean norm at most 1; the flux F is placed on the faces; and $\psi\,\pi$ is replaced by its cell averages. The primal–dual iteration of Li et al. [68] solves the two discrete problems together, and refining the grid approximates (7.9). The grid has n^d cells, so the cost grows exponentially with the dimension: in $d = 4$, $n = 100$ already gives 10^8 cells. Computing Kantorovich–Rubinstein norms in high dimensions is intrinsically difficult, and it is possibly unnecessary in the current setting, because the Wasserstein-1 error is not a *quantity of interest* in real computations: samples are used to estimate the expectations $\pi(f)$ of given test functions, and the bias of each of them is given to first order by the coefficient (6.16), which (7.3) approximates without the norm or the representor.

8 Numerical experiments

The numerical experiments study the convergence of SGLM and SGLD and compare the two integrators in various tasks. Sections 8.1 and 8.2 verify the leading terms of order h of the numerical bias and of the Wasserstein-1 error in Section 6, Section 8.3 applies the bias corrections of Section 7 to SGLM, and Section 8.4 compares the MSE of SGLM and SGLD in a Bayesian logistic regression on MNIST.

8.1 Evaluating the coefficients in a 4D Gaussian mixture distribution

The target π is a mixture of 5 Gaussian distributions on $\mathbb{R}^4$ with unequal weights, means $\vartheta_1, \dots, \vartheta_5$ near the vertices of a regular simplex of edge 4.5, and covariance matrices of random orientation, so that $U = -\log \pi$ is not convex. The stochastic gradient $\nabla \hat{U}(x;\xi)$ adds to $\nabla U(x)$ either a Gaussian noise $\xi \sim \mathcal{N}(0, I_4)$, with $\mathcal{C}(x) = I_4$, or a non-Gaussian noise, a vector of independent centered exponential random variables of unit variance, rotated and scaled depending on x, with $\mathcal{C}(x) = \big(2 - \pi(x)/\pi(\vartheta_1)\big)\, I_4$. Both covariances belong to the first row of Table 3, so the exact coefficients of the numerical bias follow from (6.16) by quadrature. The test functions $f_1, \dots, f_4$, given in the panel titles of Figure 1 with vectors a_1, a_2 of lengths 1, 0.6 and the phase φ_2, have explicit expectations under π; they are taken from a pool of 12, among the functions whose exact coefficients exceed 0.024 in absolute value for both schemes and both noises.

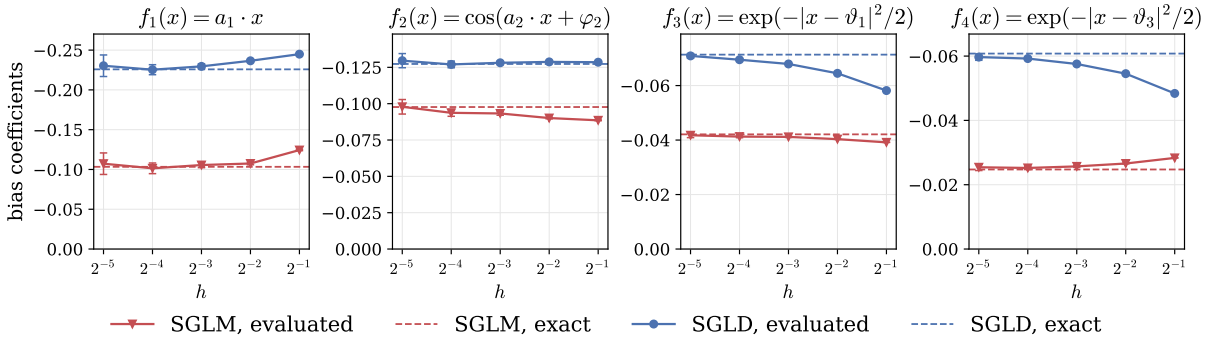


Figure 1: Gaussian noise: the evaluated coefficients $(\pi_h(f) - \pi(f))/h$ of SGLM and $(\bar{\pi}_h(f) - \pi(f))/h$ of SGLD, with error bars of 2 standard errors, and the exact coefficients (6.16), dashed. The vertical axis starts at 0 and is reversed for negative exact coefficients.

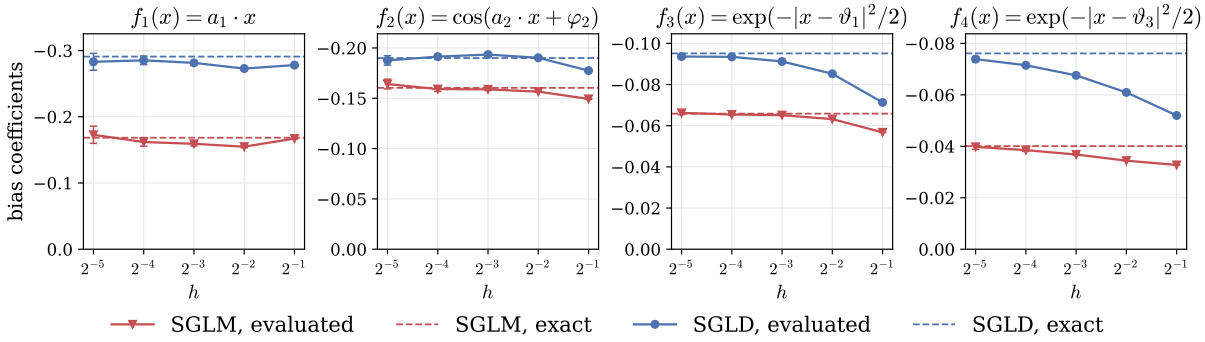


Figure 2: Non-Gaussian noise: the evaluated coefficients $(\pi_h(f) - \pi(f))/h$ of SGLM and $(\bar{\pi}_h(f) - \pi(f))/h$ of SGLD, with error bars of 2 standard errors, and the exact coefficients (6.16), dashed. The vertical axis starts at 0 and is reversed for negative exact coefficients.

The evaluated coefficients $(\pi_h(f) - \pi(f))/h$ of SGLM and $(\bar{\pi}_h(f) - \pi(f))/h$ of SGLD are obtained from the time averages of 16 chains, each over the time 5×10^7, at $h = 2^{-1}, \dots, 2^{-5}$. In Figures 1 and 2, they approach the exact coefficients as h decreases, for every test function, both schemes and both noises; at

$h = 2^{-5}$, the relative difference is at most 3.8%. The agreement under the non-Gaussian noise, whose third moments do not vanish, is in line with (6.7) and (6.10), in which the gradient noise enters only through its covariance $\mathcal{C}(x)$. For each of these test functions and both noises, the exact coefficient of SGLM is smaller in absolute value than that of SGLD, between 41% and 84% of it.

noise	integrator	$\frac{1}{2}\|\psi\,\pi\|_{\mathrm{KR}}$	f_1	f_2	f_3	f_4
Gaussian	SGLM	0.5643	18.3%	28.9%	12.3%	7.2%
	SGLD	0.9824	23.0%	21.6%	12.0%	10.2%
non-Gaussian	SGLM	0.9701	17.4%	27.5%	11.2%	6.8%
	SGLD	1.3636	21.3%	23.2%	11.5%	9.2%

Table 4: Exact coefficients $\frac{1}{2}\|\psi\,\pi\|_{\mathrm{KR}}$ of the Wasserstein-1 errors in Corollary 4, with $\psi = \psi_{\mathcal{C}}$ for SGLM and $\psi = \psi_{\nabla^2 U} + \psi_{\mathcal{C}}$ for SGLD. For each test function f of Figure 1, the percentage in the table shows $|\pi(f\psi)|/\big(\|\nabla f\|_\infty \|\psi\,\pi\|_{\mathrm{KR}}\big) \leqslant 1$, which is the ratio of its exact coefficient of the numerical bias, normalized by $\|\nabla f\|_\infty$, to the exact coefficient of the Wasserstein-1 error.

Table 4 lists the exact coefficients of the Wasserstein-1 errors in Corollary 4, calculated as in Section 7.4 and extrapolated from grids with $n = 25, 50, 100$. The exact coefficient of SGLM is 57% of that of SGLD under the Gaussian noise and 71% under the non-Gaussian noise, so (6.21) holds for both noises. The ratios in Table 4 are at most 0.29, so the Wasserstein-1 error, a supremum over all test functions f with $\|\nabla f\|_\infty \leqslant 1$, exceeds the numerical bias of each of these test functions, normalized by $\|\nabla f\|_\infty$, by a factor larger than 3.

8.2 Convergence of Wasserstein-1 error in a radial distribution

For a radial potential $U(x) = u(|x|)$ and the gradient noise $\sigma\xi$ with $\xi \sim \mathcal{N}(0, I_d)$, the covariance is $\mathcal{C}(x) = \sigma^2 I_d$, so both integrators commute with rotations. Hence the distributions π_h, $\bar{\pi}_h$ and π are rotation invariant, and the Wasserstein-1 distance between two of them is determined by their radial densities. The radial parts of (SGLD) and of the Markovian form (SGLM') are Markov chains on $[0, \infty)$, whose invariant distributions are computed by discretizing their stationarity equations, and the representors $\psi_{\mathcal{C}} = \sigma^2\big(U - \pi(U)\big)$ and $\psi_{\nabla^2 U}$ of (6.17) are radial, with $\|\psi\,\pi\|_{\mathrm{KR}} = \int_0^\infty \big|\int_{|x|\leqslant r} \psi\,\mathrm{d}\pi\big|\,\mathrm{d}r$. The evaluated coefficients $\mathcal{W}_1(\pi_h, \pi)/h$ and $\mathcal{W}_1(\bar{\pi}_h, \pi)/h$ at a given h, and the exact coefficients $\frac{1}{2}\|\psi\,\pi\|_{\mathrm{KR}}$ of Corollary 4, are thus computed without sampling in every dimension. The potential is

$$u(r) = (r^2 - r)\,\frac{r^2}{r^2 + 1/4}, \tag{8.1}$$

which grows quadratically and is not globally convex, with a ring of minima at $|x| \approx 0.64$. Since $u(r) = -4r^3 + \mathcal{O}(r^4)$ near $r = 0$, U is not three times differentiable at the origin and does not satisfy Assumption 1; the example mainly serves to observe the convergence of the Wasserstein-1 error numerically.

In Figure 3, left, the evaluated coefficients converge to the exact coefficients for $\sigma = 1.0, 2.0$, and their distance to the exact coefficients halves with h; at $h = 2^{-8}$ it is between 0.09% and 0.32% of the exact coefficients. In this example the remainder of (6.20) is therefore of order h^2, the order expected in Remark 3 under global strong convexity, although U is not convex.

In Figure 3, right, the exact coefficients grow like $\sqrt{d}$ for $d = 2, 4, \ldots, 256$: from $d = 128$ to $d = 256$ with $\sigma = 1.0$, they are multiplied by 1.416 for SGLM and by 1.434 for SGLD, close to $\sqrt{2}$. This is the growth in the Gaussian case $u(r) = r^2$, where both exact coefficients are proportional to $\mathbb{E}|X| \sim \sqrt{d}$ with $X \sim \mathcal{N}(0, I_d)$.

The exact coefficient of SGLM is below that of SGLD for every d and both σ, so (6.21) holds throughout, and the advantage of SGLM is larger when the variance of the gradient noise is smaller: the ratio of the exact coefficient of SGLM to that of SGLD is 0.46 at $d = 2$ and 0.34 at $d = 256$ for $\sigma = 1.0$, against 0.81 and 0.68 for $\sigma = 2.0$. The exact coefficient of SGLM is proportional to σ^2, while that of SGLD contains the representor $\psi_{\nabla^2 U}$ of the discretization error of EM, which does not depend on σ.

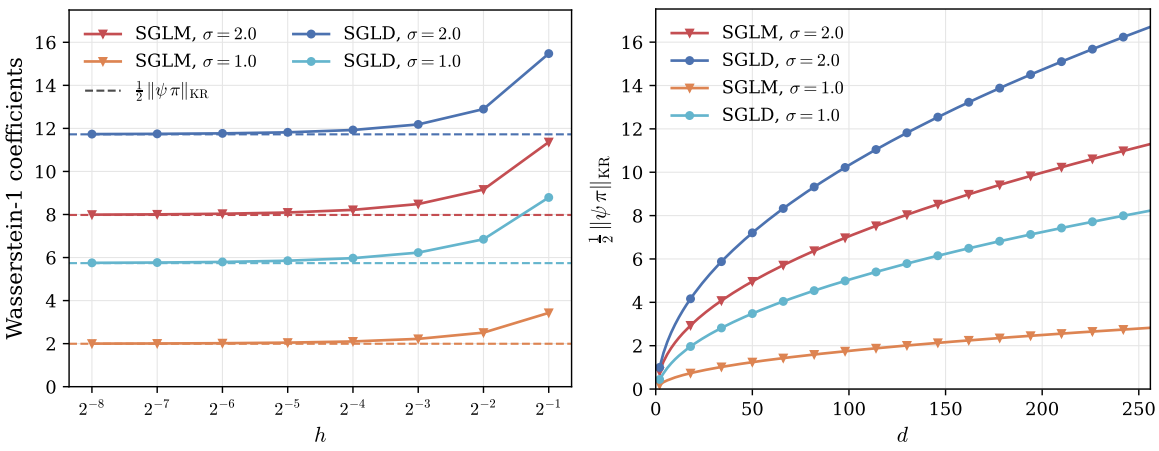


Figure 3: Radial potential (8.1) with the gradient noise $\sigma\xi$, $\xi \sim \mathcal{N}(0, I_d)$. Left: the evaluated coefficients $\mathcal{W}_1(\pi_h, \pi)/h$ of SGLM and $\mathcal{W}_1(\bar{\pi}_h, \pi)/h$ of SGLD in dimension $d = 128$, with the exact coefficients $\frac{1}{2}\|\psi\,\pi\|_{\mathrm{KR}}$ of Corollary 4, dashed. Right: the exact coefficients against $d = 2, 4, \ldots, 256$.

8.3 Bias correction of SGLM in a 16D product of ring distributions

The target π on $\mathbb{R}^{16}$ is a product of 8 ring distributions in rotated coordinates,

$$U(x) = \sum_{i=1}^{8} u_i\Big(\sqrt{y_{2i-1}^2 + y_{2i}^2}\Big), \qquad y = Qx, \qquad u_i(r) = \frac{r^2}{2} + \beta_i\, e^{-r^2/2}, \tag{8.2}$$

with a random orthogonal matrix Q and $\beta_1, \ldots, \beta_8$ drawn uniformly from $[2, 4]$. Each u_i has a ring of minima at $r = \sqrt{2\log\beta_i}$, so U is not convex. The gradient noise added to $\nabla U(x)$ is either a Gaussian one $\xi \sim \mathcal{N}(0, I_{16})$, or a non-Gaussian one: on each pair (x_{2i-1}, x_{2i}), a vector of 2 independent centered exponential random variables of unit variance, rotated by the angle $x_{2i-1} + x_{2i}$. Both noises have the covariance $\mathcal{C}(x) = I_{16}$, so $\psi_{\mathcal{C}} = U - \pi(U)$ by the first row of Table 3. The third moments of the non-Gaussian noise do not vanish and depend on x.

SGLM is run at the step sizes $h = 2^{-4}$ and $2h = 2^{-3}$, with 32 chains, each over the time 2×10^7, at each step size. The expectations $\pi_h(f)$ and $\pi_{2h}(f)$ are replaced by the time averages of f over the chains. The means $\pi(f)$ and the exact coefficients $\frac{1}{2}\,\pi(f\psi_{\mathcal{C}}) = \frac{1}{2}\big(\pi(fU) - \pi(f)\,\pi(U)\big)$ are each computed with 1.024×10^{10} exact samples of π. The test functions for evaluating the numerical bias are chosen as

$$\begin{aligned}
f_1(x) &= x_1, & f_2(x) &= \frac{1}{4}\sum_{j=1}^{16} x_j,\\
f_3(x) &= \cos x_1, & f_4(x) &= \sin(x_1 + x_2),\\
f_5(x) &= \sqrt{1 + x_1^2}, & f_6(x) &= e^{-(x_1^2 + x_2^2)/2},\\
f_7(x) &= \log\big(1 + e^{x_1 - x_2}\big), & f_8(x) &= \sqrt{1 + |x|^2},\\
f_9(x) &= \cos(x_1 + x_2), & f_{10}(x) &= \sqrt{1 + x_1^2 + x_2^2},\\
f_{11}(x) &= \frac{x_1 x_2}{\sqrt{1 + x_1^2 + x_2^2}}, & f_{12}(x) &= \log\big(1 + x_1^2 + x_2^2 + x_3^2 + x_4^2\big).
\end{aligned}$$

For each test function, Tables 6 and 7 list the approximations of $\pi(f)$ in Table 5 at $h = 2^{-4}$, each with $\pi(f)$ subtracted, where $\varphi_{\mathcal{C}}$ of (7.5) is approximated by minimizing (7.8).

The entry of SGLM is the numerical bias $\pi_h(f) - \pi(f)$, and the entries of explicit, extrapolation and shift are the errors of the approaches of Section 7. The column explicit uses the exact coefficient, as in Section 7.1, and the column extrapolation is the error bounded in (7.4).

method	approximation of $\pi(f)$
SGLM	$\pi_h(f)$
explicit coefficient	$\pi_h(f) - \frac{h}{2}\,\pi(f\psi_{\mathcal{C}})$
extrapolation	$2\pi_h(f) - \pi_{2h}(f)$
shifting correction	$\pi_h\big(f(x + h\nabla\varphi_{\mathcal{C}}(x))\big)$

Table 5: The approximations of $\pi(f)$ by SGLM and by the approaches of Section 7, listed in the columns SGLM, explicit, extrapolation and shift of Tables 6 and 7.

The column shift uses the postprocessed samples (7.7), with $\varphi_{\mathcal{C}}$ approximated by a neural network with 3 hidden layers of width 64, trained on 1.68×10^7 states of SGLM at each of the step sizes h and $2h$, stored once per unit of time along 32 chains and thus covering the total time 1.68×10^7. The change $\pi_h\big(f(x+h\nabla\varphi_{\mathcal{C}}(x))\big) - \pi_h(f)$ is computed on states of runs at h independent of the training states and added to the time average of f at h.

	SGLM	explicit	extrapolation	shift
f_1	-8.64×10^{-5}	$\mathbf{-8.57 \times 10^{-5}}$	-3.00×10^{-4}	-5.36×10^{-4}
f_2	$\mathbf{-1.83 \times 10^{-4}}$	-1.85×10^{-4}	-5.42×10^{-4}	-8.86×10^{-4}
f_3	-4.86×10^{-3}	$\mathbf{-6.73 \times 10^{-5}}$	1.49×10^{-4}	1.86×10^{-4}
f_4	$\mathbf{-2.63 \times 10^{-5}}$	-2.68×10^{-5}	-9.72×10^{-5}	-2.61×10^{-4}
f_5	5.70×10^{-3}	$\mathbf{1.08 \times 10^{-4}}$	-2.63×10^{-4}	-2.29×10^{-4}
f_6	-2.99×10^{-3}	$\mathbf{-1.51 \times 10^{-5}}$	4.67×10^{-5}	1.50×10^{-4}
f_7	3.86×10^{-3}	6.38×10^{-5}	-2.60×10^{-4}	$\mathbf{-1.71 \times 10^{-6}}$
f_8	3.61×10^{-2}	$\mathbf{5.07 \times 10^{-4}}$	-1.48×10^{-3}	-1.99×10^{-3}
f_9	-3.97×10^{-3}	$\mathbf{-2.48 \times 10^{-7}}$	2.29×10^{-5}	1.11×10^{-4}
f_{10}	1.00×10^{-2}	$\mathbf{1.59 \times 10^{-4}}$	-3.98×10^{-4}	-5.42×10^{-4}
f_{11}	9.84×10^{-5}	$\mathbf{-3.18 \times 10^{-5}}$	-1.08×10^{-4}	1.35×10^{-4}
f_{12}	1.11×10^{-2}	$\mathbf{1.19 \times 10^{-4}}$	-3.47×10^{-4}	-3.33×10^{-4}

Table 6: Gaussian noise: the numerical bias $\pi_h(f) - \pi(f)$ of SGLM at $h = 2^{-4}$ and the errors of the explicit coefficient, the extrapolation and the shifting correction of Section 7. The entry of smallest absolute value in each row is in bold.

In general, the explicit coefficient has the smallest errors: it removes the term of order h of the bias with the exact coefficient, and what remains is the remainder of (6.18) and the fluctuation of the time averages. The extrapolation and the shifting correction have larger errors for some test functions, mainly from the additional error of the run at $2h = 2^{-3}$. Formally, the term of order h^2 of $\pi_{2h}(f) - \pi(f)$ is 4 times that of $\pi_h(f) - \pi(f)$, so $2\pi_h(f) - \pi_{2h}(f) - \pi(f)$ carries the term of order h^2 of $\pi_h(f) - \pi(f)$ with the factor -2, against 1 for the explicit coefficient; for f_4 under the non-Gaussian noise, Table 7 gives -9.99×10^{-4} for the extrapolation against 4.94×10^{-4} for the explicit coefficient. The shifting correction inherits the additional error of the run at $2h$, since the term of order h of $\pi_h\big(f(x + h\nabla\varphi_{\mathcal{C}}(x))\big) - \pi_h(f)$ is formally $h\,\pi_h(\nabla f \cdot \nabla\varphi_{\mathcal{C}}) = \pi_h(f) - \pi_{2h}(f)$ by the minimization of (7.8); the entries of shift for f_8, f_{10} and f_{12} have, in both tables, the sign and size of those of extrapolation. The largest error of the shifting correction, 2.44×10^{-3} for f_1 under the non-Gaussian noise, comes from the fluctuation of the stored states on which $\varphi_{\mathcal{C}}$ is trained, which cover the total time 1.68×10^7, against 6.4×10^8 for the time averages: on the stored states, $(\pi_{2h}(f_1) - \pi_h(f_1))/h$ is -3.70×10^{-2}, with the standard error 1.69×10^{-2}, against -7.08×10^{-3} from the time averages, and the change caused by the shift, 2.24×10^{-3}, is close to $-h$ times -3.70×10^{-2}. Nevertheless, the largest absolute error over $f_1, \dots, f_{12}$ is 3.61×10^{-2} for SGLM under both noises, against at most 5.43×10^{-4}, 1.48×10^{-3} and 2.44×10^{-3} for the explicit coefficient, the extrapolation and the shifting correction; the corrections reduce the largest absolute error of SGLM by factors between 14.8 and 71.1, so the bias correction works in this example.

	SGLM	explicit	extrapolation	shift
f_1	$\mathbf{1.99 \times 10^{-4}}$	2.00×10^{-4}	6.42×10^{-4}	2.44×10^{-3}
f_2	5.56×10^{-5}	5.38×10^{-5}	$\mathbf{-3.11 \times 10^{-5}}$	1.32×10^{-3}
f_3	-4.76×10^{-3}	3.22×10^{-5}	$\mathbf{4.79 \times 10^{-6}}$	3.09×10^{-4}
f_4	$\mathbf{4.94 \times 10^{-4}}$	$\mathbf{4.94 \times 10^{-4}}$	-9.99×10^{-4}	5.61×10^{-4}
f_5	5.63×10^{-3}	$\mathbf{3.87 \times 10^{-5}}$	-1.91×10^{-4}	-4.42×10^{-4}
f_6	-2.98×10^{-3}	$\mathbf{-5.15 \times 10^{-6}}$	5.24×10^{-5}	7.70×10^{-5}
f_7	4.02×10^{-3}	$\mathbf{2.27 \times 10^{-4}}$	2.61×10^{-4}	1.13×10^{-3}
f_8	3.61×10^{-2}	$\mathbf{5.43 \times 10^{-4}}$	-1.44×10^{-3}	-1.76×10^{-3}
f_9	-4.00×10^{-3}	$\mathbf{-3.50 \times 10^{-5}}$	-8.72×10^{-5}	-1.78×10^{-4}
f_{10}	1.00×10^{-2}	$\mathbf{1.47 \times 10^{-4}}$	-4.08×10^{-4}	-3.72×10^{-4}
f_{11}	2.07×10^{-4}	$\mathbf{7.66 \times 10^{-5}}$	1.37×10^{-4}	3.47×10^{-4}
f_{12}	1.11×10^{-2}	$\mathbf{7.75 \times 10^{-5}}$	-4.24×10^{-4}	-6.31×10^{-4}

Table 7: Non-Gaussian noise: the same quantities as in Table 6, with the entry of smallest absolute value in each row in bold.

8.4 Bayesian logistic regression on MNIST

Bayesian logistic regression is applied to the handwritten digits 4 and 9 of the MNIST data set [54], shown in Figure 4, with the training set restricted to these digits. Each of the $n = 10{,}791$ training images, 5,342 of the digit 4 and 5,449 of the digit 9, is flattened, its pixel intensities are divided by 255, and a constant 1 is appended for the intercept, giving the vector $a_i \in \mathbb{R}^{785}$, with the label $y_i = 0$ for the digit 4 and $y_i = 1$ for the digit 9. With the prior $\mathcal{N}(0, \sigma_0^2 I_{785})$, $\sigma_0^2 = 10^{-3}$, the target is $\pi \propto e^{-U}$ on $\mathbb{R}^{785}$ with

$$U(x) = \frac{|x|^2}{2\sigma_0^2} + \sum_{i=1}^{n} \Big(\log\big(1 + e^{a_i \cdot x}\big) - y_i \, a_i \cdot x \Big). \tag{8.3}$$

The stochastic gradient averages P indices $\xi = (\xi_1, \dots, \xi_P)$, drawn independently and uniformly from $\{1, \dots, n\}$ at every step, independently of the Gaussian variables of the integrator:

$$\nabla \hat{U}(x; \xi) = \frac{x}{\sigma_0^2} + \frac{n}{P} \sum_{p=1}^{P} \Big(\frac{1}{1 + e^{-a_{\xi_p} \cdot x}} - y_{\xi_p} \Big) \, a_{\xi_p}. \tag{8.4}$$

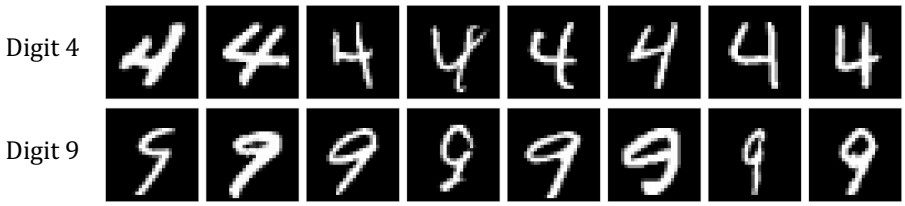


Figure 4: Examples of the digits 4 and 9 of the MNIST data set.

The test functions measure the posterior in directions of different curvature. Let x_* be the minimizer of U, computed by L-BFGS with the full gradient, and let $\lambda_1 \geqslant \dots \geqslant \lambda_8$ be the 8 largest eigenvalues of $\nabla^2 U(x_*)$, with orthonormal eigenvectors $v_1, \dots, v_8$. In the standardized coordinates

$$z_j(x) = \sqrt{\lambda_j} \, v_j \cdot (x - x_*), \qquad j = 1, \dots, 8, \tag{8.5}$$

the 16 bounded test functions are

$$f_{C_j}(x) = \cos z_j(x), \qquad j = 1, \dots, 8, \tag{8.6}$$

$$f_{G_j}(x) = e^{-z_j(x)^2/2}, \qquad j = 1, \dots, 4, \tag{8.7}$$

$$f_{G_{j,k}}(x) = e^{-(z_j(x)^2 + z_k(x)^2)/2}, \qquad (j,k) \in \{(1,2), (3,4), (5,6), (7,8)\}. \tag{8.8}$$

The cosines f_{C_j} probe single directions, and the Gaussians f_{G_j} and $f_{G_{j,k}}$ the concentration near x_* in 1 or 2 coordinates.

The step size follows Brosse et al. [9, Section 4] for logistic regression with the Gaussian prior, $h = (\sigma_0^{-2} + \delta/4)^{-1} \approx 8.5099 \times 10^{-6}$, where δ is the largest eigenvalue of $\sum_{i=1}^n a_i \otimes a_i$. For each batch size $P = 1024, 2048, 4096$, SGLD and SGLM are run in 20 pairs of trajectories, the 2 trajectories of a pair using the same indices ξ and the same Gaussian variables. Every trajectory starts at x_*, discards its first 2^{14} steps and retains the next 2^{17} states. The reference values $\widehat{\pi}_{\mathrm{ref}}(f)$ are the pooled averages of 4 independent chains of NUTS [48] with the full gradient, which leaves π invariant, each with 2^{19} retained states after 2^{14} adaptation steps. The chains use the implementation of NUTS in Pyro [8], in the coordinates whitened by the Cholesky factor of $\nabla^2 U(x_*)$, with the target acceptance rate 0.9, the maximal tree depth 10, and the step size and a diagonal mass matrix adapted during the adaptation steps. The 4 chains agree closely for all 16 test functions, their Monte Carlo errors are small compared with the errors below, and these Monte Carlo errors are included in the resampling intervals below.

For each integrator, batch size and test function f, the estimates use the first N retained states, for every multiple N of 8192 up to 2^{17}. With $Z_{r,k}$ the kth retained state of the repetition $r = 1, \dots, 20$, let

$$\widehat{\mu}_r = \frac{1}{N} \sum_{k=0}^{N-1} f(Z_{r,k}), \qquad \overline{\mu} = \frac{1}{20} \sum_{r=1}^{20} \widehat{\mu}_r$$

be the time average and its mean over the repetitions. The estimated finite-time bias and the variance of the time averages are

$$\widehat{b} = \overline{\mu} - \widehat{\pi}_{\mathrm{ref}}(f), \qquad \widehat{V} = \frac{1}{20} \sum_{r=1}^{20} (\widehat{\mu}_r - \overline{\mu})^2.$$

The reported estimate of the MSE (1.2) is

$$\widehat{\mathrm{MSE}}_f(N, h) = \frac{1}{20} \sum_{r=1}^{20} \big(\widehat{\mu}_r - \widehat{\pi}_{\mathrm{ref}}(f)\big)^2 = \widehat{V} + \widehat{b}^2. \tag{8.9}$$

The variance is that of the time averages over the repetitions, not of single states, and the bias is that of time averages of finite length, not the numerical bias of the invariant distribution.

Figure 5 shows 2 test functions of each form, $f_{C_1}, f_{C_4}, f_{G_2}, f_{G_3}, f_{G_{5,6}}$ and $f_{G_{7,8}}$, which together involve all 8 coordinates (8.5). For f_{C_1} the estimated MSE is nearly constant in N, so longer averages bring little improvement in the tested range. For f_{C_4} and $f_{G_{7,8}}$, several curves decrease markedly at small N and then level off. For f_{G_2}, f_{G_3} and $f_{G_{5,6}}$, longer averages are particularly beneficial for SGLM at $P = 4096$, whose errors fall below those of SGLD. The leveling off of several curves indicates a part of the error from the numerical bias, which longer averages at fixed h and P cannot remove.

At the largest retained length $N = 2^{17}$, SGLM has the smaller estimated MSE for all 16 test functions at each of the 3 batch sizes, that is, in all 48 comparisons, and in each of them the 95% interval of $\widehat{\mathrm{MSE}}_{\mathrm{SGLD}} - \widehat{\mathrm{MSE}}_{\mathrm{SGLM}}$ obtained by resampling the 20 pairs with replacement lies above 0. For each test function and batch size, the comparison at $N = 2^{17}$ uses the quantities

$$\text{MSE Reduction} = 100\left(1 - \frac{\widehat{\mathrm{MSE}}_{\mathrm{SGLM}}}{\widehat{\mathrm{MSE}}_{\mathrm{SGLD}}}\right)\%,$$

$$\text{Bias Ratio} = \frac{|\widehat{b}_{\mathrm{SGLM}}|}{|\widehat{b}_{\mathrm{SGLD}}|}, \qquad \text{Variance Ratio} = \frac{\widehat{V}_{\mathrm{SGLM}}}{\widehat{V}_{\mathrm{SGLD}}}.$$

A positive MSE reduction and ratios below 1 favor SGLM. Table 8 gives the median of each quantity over the 16 test functions.

The median MSE reduction grows with the batch size, from 34.6% to 63.3%, while the median variance ratio stays close to 1 and the median bias ratio decreases from 0.784 to 0.516: the gain of SGLM comes from smaller finite-time biases, not from smaller fluctuations. This trend in P is consistent with the separation of the errors in Theorem 3: a batch of P independent stochastic gradients reduces the contribution of the

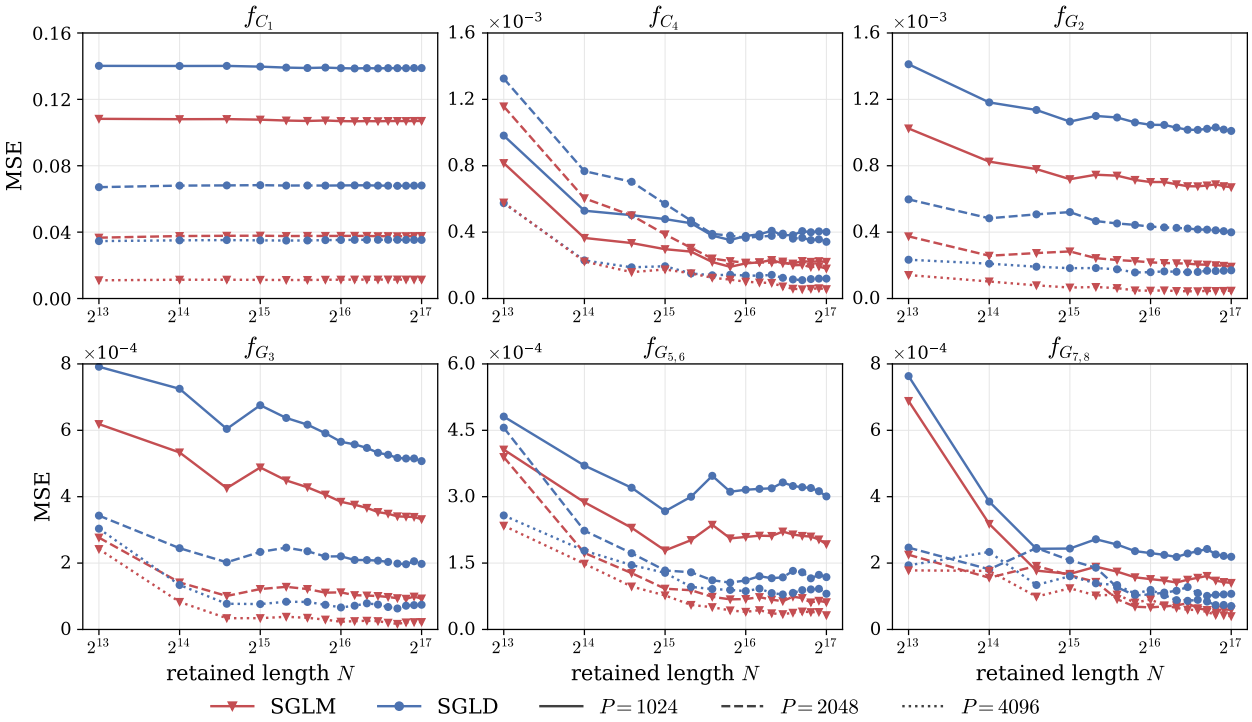


Figure 5: Estimated MSE (8.9) of the time averages of SGLD and SGLM for $f_{C_1}, f_{C_4}, f_{G_2}, f_{G_3}, f_{G_{5,6}}$ and $f_{G_{7,8}}$ against the retained length N, with 20 pairs of trajectories for each batch size P.

P	MSE Reduction	Bias Ratio	Variance Ratio
1024	34.6%	0.784	1.020
2048	46.5%	0.677	1.018
4096	63.3%	0.516	1.032

Table 8: Comparison of SGLM with SGLD at $N = 2^{17}$: the median over the 16 test functions of each quantity, computed for each test function.

gradient noise to $\Sigma^2 h^2/P^2$, while the discretization error contributes h^4 for SGLM and h^2 for SGLD in (6.8). In the bias, the coefficient (6.7) of SGLM is divided by P, while the coefficient (6.10) of SGLD keeps the contribution of the discretization error of EM at every batch size.

9 Conclusion

For a potential strongly convex outside a ball, this paper proves MSE bounds for the SGLM integrator and its full gradient form LM, and gives the explicit coefficients of h in the numerical bias and in the Wasserstein-1 error of SGLM and SGLD. The samples of (SGLM) and (LM) do not form a Markov chain; the modified class of Section 4 moves the analysis to the Markovian forms (SGLM') and (LM'), and the discrete Poisson equation of [71, 106] converts their local weak errors against the modified dynamics into the bound $\frac{1}{Nh} + \Sigma^2 h^2 + h^4$ of Theorem 3, with the bound $\frac{1}{Nh} + h^4$ of Theorem 2 for the full gradient. The stochastic gradient error and the discretization error thus separate for an integrator of the overdamped dynamics, as for SG-UBU [71] and unlike (SGLD).

The numerical bias of SGLM is of first order in h, with the coefficient (6.7) given by the covariance $\mathcal{C}(x)$ of the gradient noise. For (SGLD), the same discrete Poisson equation gives the coefficient (6.10), which carries the Hessian $\nabla^2 U(x)$ besides the covariance $\mathcal{C}(x)$. This contribution of the discretization error of EM remains at every batch size, while the coefficient of SGLM is divided by the batch size. For a given test function, the

two coefficients are not ordered in absolute value, since $\nabla^2\phi_0$ is indefinite in general. In both coefficients, the gradient noise and the discretization error enter linearly through $\nabla^2\phi_0$: the Poisson equation describes the structure of the numerical bias of both stochastic gradient integrators, beyond its role in the upper bounds on the MSE. The gap between Theorem 5, which holds at every step size with ϕ_h, and Corollary 3, which reaches the Hessian form with ϕ_0 only in the limit $h \to 0$, also shows that the discrete Poisson equation requires fewer derivatives than the continuous one for nonasymptotic bounds on the numerical bias.

Through the Riesz representors of $\mathcal{C}(x)$ and $\nabla^2 U(x)$, the coefficients of h in the Wasserstein-1 errors are explicit: $\mathcal{W}_1(\pi_h, \pi)/h$ and $\mathcal{W}_1(\bar{\pi}_h, \pi)/h$ converge to $\frac{1}{2}\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}$ and $\frac{1}{2}\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}}$ (Corollary 4), as a consequence of the convergence of the normalized biases $(\pi_h - \pi)/h$ and $(\bar{\pi}_h - \pi)/h$ in the Kantorovich–Rubinstein norm (Theorem 7). For a strongly convex potential, the coefficient of SGLM is at most that of SGLD once the batch size reaches a threshold linear in $\sup_x \|\mathcal{C}(x)\|$, with no explicit dependence on the dimension (Theorem 8), and Lemma 10 shows that the order can be reversed at small batch sizes.

The coefficients also lead to approaches to bias correction in Section 7: the explicit coefficient, Richardson–Romberg extrapolation and the shifting correction. The shifting correction replaces a sample $Z \sim \pi_h$ of SGLM by $Z + h\nabla\varphi_{\mathcal{C}}(Z)$ and formally removes the first order bias for all test functions at once. It is completely data-driven, since $\varphi_{\mathcal{C}}$ is approximated by minimizing the energy (7.8) with the samples of SGLM at the step sizes h and $2h$.

The numerical experiments of Section 8 illustrate the coefficients, the bias correction and the comparison with SGLD. In a 4D Gaussian mixture, the evaluated coefficients of the numerical bias approach the exact coefficients (6.7) and (6.10), within 3.8% at $h = 2^{-5}$. For a radial potential, the evaluated coefficients of the Wasserstein-1 errors converge to those of Corollary 4 with a remainder of order h^2, and the coefficient of SGLM is below that of SGLD for $d = 2, 4, \ldots, 256$. In a 16D product of ring distributions, the bias corrections reduce the largest absolute numerical bias of SGLM over the test functions by factors between 14.8 and 71.1. In Bayesian logistic regression on MNIST, SGLM has a smaller estimated MSE than SGLD for all test functions and batch sizes, and the median reduction grows with the batch size, from 34.6% at $P = 1024$ to 63.3% at $P = 4096$, mainly through a smaller bias, in line with the coefficient of SGLM being divided by P.

Four extensions are possible. The first is variance reduction: with variance-reduced gradients, the first order term of the bias of SG-UBU becomes of second order once h is below a threshold [71], and the same construction applied to (SGLM) would give an integrator of the overdamped dynamics with a bias of second order. The second is the remainder $Ch^{\frac{3}{2}}(1 + \log\frac{1}{h})$ of (6.20), which is not guaranteed to be sharp in the discretization error, and the error order could be enhanced to h^2. The third is the underdamped Langevin dynamics, where a modified class for its stochastic gradient integrators would extend the analysis to that setting. The fourth is a rigorous study of the shifting correction, whose order h^2 is formal, together with the stability of the minimization of (7.8), in which the fluctuation of the empirical averages is divided by h.

Funding statement

Lin gratefully acknowledges support from the National Science Foundation (DMS-2533878, DMS-2053746, DMS-2134209, ECCS-2328241, CBET-2347401 and OAC-2311848), and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program under the "Uncertainty Quantification for Multifidelity Operator Learning (MOLUcQ)" project (Project No. 81739), DE-SC0023161, the SciDAC LEADS Institute, and DOE–Fusion Energy Science, under grant number: DE-SC0024583.

Acknowledgments

The authors thank Dr. He Zhang for valuable suggestions on this paper.

Data availability

The code of the numerical experiments is available at https://github.com/xuda-ye-math/Poisson-SGLM, with the commands that regenerate the numerical results.

Declaration of AI use

The authors used Claude Opus 5.5 and GPT-6 Astra in deriving part of the results, in checking the correctness of the proofs, and in writing the code of the numerical experiments. Claude Opus 5.5 provided the counterexample in Lemma 10. The authors hold full responsibility for the content of this paper.

References

[1] Assyr Abdulle, Gilles Vilmart, and Konstantinos C. Zygalakis. High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM Journal on Numerical Analysis*, 52(4):1600–1622, 2014. doi: 10.1137/130935616.

[2] Ömer Deniz Akyildiz and Sotirios Sabanis. Nonasymptotic analysis of stochastic gradient Hamiltonian Monte Carlo under local conditions for nonconvex optimization. *Journal of Machine Learning Research*, 25(113):1–34, 2024.

[3] Jason M. Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 71(3):1–55, 2024. doi: 10.1145/3653446.

[4] Jack Baker, Paul Fearnhead, Emily B. Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29(3):599–615, 2019. doi: 10.1007/s11222-018-9826-2.

[5] Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electronic Communications in Probability*, 13:60–66, 2008. doi: 10.1214/ECP.v13-1352.

[6] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2014. doi: 10.1007/978-3-319-00227-9.

[7] Mathias Barkhagen, Ngoc Huy Chau, Éric Moulines, Miklós Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021. doi: 10.3150/19-BEJ1187.

[8] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019.

[9] Nicolas Brosse, Alain Durmus, and Éric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[10] Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019. doi: 10.1016/j.spa.2018.10.002.

[11] Giovanni Bussi and Michele Parrinello. Accurate sampling using Langevin dynamics. *Physical Review E*, 75(5):056707, 2007. doi: 10.1103/PhysRevE.75.056707.

[12] Yu Cao and Eric Vanden-Eijnden. Learning optimal flows for non-equilibrium importance sampling. In *Advances in Neural Information Processing Systems*, volume 35, pages 12352–12364, 2022.

[13] Neil K. Chada, Benedict Leimkuhler, Daniel Paulin, and Peter A. Whalley. Unbiased kinetic Langevin Monte Carlo with inexact gradients. *The Annals of Statistics*, 54(4):1895–1922, 2026. doi: 10.1214/26-AOS2630.

[14] Martin Chak and Pierre Monmarché. Reflection coupling for unadjusted generalized Hamiltonian Monte Carlo in the nonconvex stochastic gradient case. *IMA Journal of Numerical Analysis*, 46(4):2216–2266, 2026. doi: 10.1093/imanum/draf045.

[15] Niladri S. Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L. Bartlett, and Michael I. Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 764–773, 2018.

[16] Niladri S. Chatterji, Peter L. Bartlett, and Philip M. Long. Oracle lower bounds for stochastic gradient sampling algorithms. *Bernoulli*, 28(2):1074–1092, 2022. doi: 10.3150/21-BEJ1377.

[17] Ngoc Huy Chau, Éric Moulines, Miklós Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021. doi: 10.1137/20M1355392.

[18] Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[19] Changyou Chen, Wenlin Wang, Yizhe Zhang, Qinliang Su, and Lawrence Carin. A convergence analysis for a class of practical variance-reduction stochastic gradient MCMC. *Science China Information Sciences*, 62(1):12101, 2019. doi: 10.1007/s11432-018-9656-y.

[20] Fan Chen, Sinho Chewi, Constantinos Daskalakis, and Alexander Rakhlin. High-accuracy sampling for diffusion models and log-concave distributions. In *International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 13835–13869, 2026.

[21] Fan Chen, Sinho Chewi, Constantinos Daskalakis, and Alexander Rakhlin. High-accuracy log-concave sampling with stochastic queries. In *Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 1341–1372, 2026.

[22] Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1683–1691, 2014.

[23] Yifan Chen, Xiaoou Cheng, Jonathan Niles-Weed, and Jonathan Weare. Convergence of unadjusted Langevin in high dimensions: delocalization of bias. *Communications on Pure and Applied Mathematics*, 79(6):1467–1491, 2026. doi: 10.1002/cpa.70032.

[24] Yifan Chen, Xiaoou Cheng, Jonathan Niles-Weed, and Jonathan Weare. Delocalization of bias in unadjusted Hamiltonian Monte Carlo and underdamped Langevin. arXiv:2607.15208, 2026.

[25] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2984–3014, 2022.

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

[27] Arnak S. Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019. doi: 10.1016/j.spa.2019.02.016.

[28] Aniket Das, Dheeraj M. Nagaraj, and Anant Raj. Utilising the CLT structure in stochastic gradient based sampling: improved analysis and faster algorithms. In *Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4072–4129, 2023.

[29] Kristian Debrabant. Runge–Kutta methods for third order weak approximation of SDEs with multidimensional additive noise. *BIT Numerical Mathematics*, 50(3):541–558, 2010. doi: 10.1007/s10543-010-0276-2.

[30] Arnaud Debussche and Erwan Faou. Weak backward error analysis for SDEs. *SIAM Journal on Numerical Analysis*, 50(3):1735–1752, 2012. doi: 10.1137/110831544.

[31] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[32] Kumar Avinava Dubey, Sashank J. Reddi, Sinead A. Williamson, Barnabás Póczos, Alexander J. Smola, and Eric P. Xing. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[33] Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017. doi: 10.1214/16-AAP1238.

[34] Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019. doi: 10.3150/18-BEJ1073.

[35] Alain Durmus, Umut Simsekli, Éric Moulines, Roland Badeau, and Gaël Richard. Stochastic gradient Richardson–Romberg Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[36] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

[37] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3–4):851–886, 2016. doi: 10.1007/s00440-015-0673-1.

[38] K. David Elworthy and Xue-Mei Li. Formulae for the derivatives of heat semigroups. *Journal of Functional Analysis*, 125(1):252–286, 1994. doi: 10.1006/jfan.1994.1124.

[39] Thomas O. Gallouët, Andrea Natale, and Gabriele Todeschi. From geodesic extrapolation to a variational BDF2 scheme for Wasserstein gradient flows. *Mathematics of Computation*, 93(350):2769–2810, 2024. doi: 10.1090/mcom/3951.

[40] Thomas O. Gallouët, Andrea Natale, and Gabriele Todeschi. Metric extrapolation in the Wasserstein space. *Calculus of Variations and Partial Differential Equations*, 64(5):147, 2025. doi: 10.1007/s00526-025-02996-6.

[41] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient Hamiltonian Monte Carlo for nonconvex stochastic optimization: nonasymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5):2931–2947, 2022. doi: 10.1287/opre.2021.2162.

[42] Michael B. Giles, Tigran Nagapetyan, Łukasz Szpruch, Sebastian J. Vollmer, and Konstantinos C. Zygalakis. Multilevel Monte Carlo for scalable Bayesian computations. arXiv:1609.06144, 2016.

[43] Michael B. Giles, Mateusz B. Majka, Łukasz Szpruch, Sebastian J. Vollmer, and Konstantinos C. Zygalakis. Multi-level Monte Carlo methods for the approximation of invariant measures of stochastic differential equations. *Statistics and Computing*, 30(3):507–524, 2020. doi: 10.1007/s11222-019-09890-0.

[44] Nicolaï Gouraud, Pierre Le Bris, Adrien Majka, and Pierre Monmarché. HMC and underdamped Langevin united in the unadjusted convex smooth case. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):278–303, 2025. doi: 10.1137/23M1608963.

[45] Melville S. Green. Markoff random processes and the statistical mechanics of time-dependent phenomena. II. Irreversible processes in fluids. *The Journal of Chemical Physics*, 22(3):398–413, 1954. doi: 10.1063/1.1740082.

[46] Martin Hairer and Andrew J. Majda. A simple framework to justify linear response theory. *Nonlinearity*, 23(4):909–922, 2010. doi: 10.1088/0951-7715/23/4/008.

[47] Martin Hairer and Jonathan C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In Robert C. Dalang, Marco Dozzi, and Francesco Russo, editors, *Seminar on Stochastic Analysis, Random Fields and Applications VI*, volume 63 of *Progress in Probability*, pages 109–117. Birkhäuser/Springer Basel, 2011. doi: 10.1007/978-3-0348-0021-1_7.

[48] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

[49] Yuri Kinoshita and Taiji Suzuki. Improved convergence rate of stochastic gradient Langevin dynamics with variance reduction and its application to optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[50] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23 of *Applications of Mathematics (New York)*. Springer, 1992. doi: 10.1007/978-3-662-12616-5.

[51] Marie Kopec. Weak backward error analysis for overdamped Langevin processes. *IMA Journal of Numerical Analysis*, 35(2):583–614, 2015. doi: 10.1093/imanum/dru016.

[52] Ryogo Kubo. Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, 12(6): 570–586, 1957. doi: 10.1143/JPSJ.12.570.

[53] Damien Lamberton and Gilles Pagès. Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3):367–405, 2002.

[54] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[55] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2993–3050, 2021.

[56] Benedict Leimkuhler and Charles Matthews. Rational construction of stochastic numerical methods for molecular sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56, 2013. doi: 10.1093/amrx/abs010.

[57] Benedict Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, volume 39 of *Interdisciplinary Applied Mathematics*. Springer, 2015. doi: 10.1007/978-3-319-16375-8.

[58] Benedict Leimkuhler and Xiaocheng Shang. Adaptive thermostats for noisy gradient systems. *SIAM Journal on Scientific Computing*, 38(2):A712–A736, 2016. doi: 10.1137/15M102318X.

[59] Benedict Leimkuhler, Charles Matthews, and Michael V. Tretyakov. On the long-time integration of stochastic gradient systems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 470(2170):20140120, 2014. doi: 10.1098/rspa.2014.0120.

[60] Benedict Leimkuhler, Charles Matthews, and Gabriel Stoltz. The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA Journal of Numerical Analysis*, 36(1): 13–79, 2016. doi: 10.1093/imanum/dru056.

[61] Benedict Leimkuhler, Matthias Sachs, and Gabriel Stoltz. Hypocoercivity properties of adaptive Langevin dynamics. *SIAM Journal on Applied Mathematics*, 80(3):1197–1222, 2020. doi: 10.1137/19M1291649.

[62] Benedict Leimkuhler, Daniel Paulin, and Peter A. Whalley. Contraction rate estimates of stochastic gradient kinetic Langevin integrators. *ESAIM: Mathematical Modelling and Numerical Analysis*, 58 (6):2255–2286, 2024. doi: 10.1051/m2an/2024038.

[63] Tony Lelièvre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016. doi: 10.1017/S0962492916000039.

[64] Vincent Lemaire, Gilles Pagès, and Fabien Panloup. Invariant measure of duplicated diffusions and application to Richardson–Romberg extrapolation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 51(4):1562–1596, 2015. doi: 10.1214/13-AIHP591.

[65] Lei Li and Yuliang Wang. A sharp uniform-in-time error estimate for stochastic gradient Langevin dynamics. *CSIAM Transactions on Applied Mathematics*, 6(4):711–759, 2025. doi: 10.4208/csiam-am.SO-2024-0039.

[66] Lei Li, Jian-Guo Liu, and Yuliang Wang. Geometric ergodicity of SGLD via reflection coupling. *Stochastics and Dynamics*, 24(5):2450035, 2024. doi: 10.1142/S0219493724500357.

[67] Ruilin Li, Xin Wang, Hongyuan Zha, and Molei Tao. Improving sampling accuracy of stochastic gradient MCMC methods via non-uniform subsampling of gradients. *Discrete and Continuous Dynamical Systems – Series S*, 16(2):329–360, 2023. doi: 10.3934/dcdss.2021157.

[68] Wuchen Li, Ernest K. Ryu, Stanley Osher, Wotao Yin, and Wilfrid Gangbo. A parallel method for Earth Mover's distance. *Journal of Scientific Computing*, 75(1):182–197, 2018. doi: 10.1007/s10915-017-0529-1.

[69] Xuechen Li, Yi Wu, Lester Mackey, and Murat A. Erdogdu. Stochastic Runge–Kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[70] Zhize Li, Tianyi Zhang, Shuyu Cheng, Jun Zhu, and Jian Li. Stochastic gradient Hamiltonian Monte Carlo with variance reduction for Bayesian inference. *Machine Learning*, 108(8–9):1701–1727, 2019. doi: 10.1007/s10994-019-05825-y.

[71] Jianfeng Lu, Xuda Ye, and Zhennan Zhou. Mean square error analysis of stochastic gradient and variance-reduced sampling algorithms. arXiv:2511.04413, 2025.

[72] Yi-An Ma, Tianqi Chen, and Emily B. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[73] Mateusz B. Majka, Aleksandar Mijatović, and Łukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, 30(4):1534–1581, 2020. doi: 10.1214/19-AAP1535.

[74] Mateusz B. Majka, Tigran Nagapetyan, Łukasz Szpruch, Yue Wu, and Danqi Zhuang. On the computational cost of stochastic gradient Langevin dynamics. arXiv:2609.17750, 2026.

[75] Riccardo Mannella. Numerical integration of stochastic differential equations: the Heun algorithm revisited and the Itô–Stratonovich calculus. *Entropy*, 27(9):910, 2025. doi: 10.3390/e27090910.

[76] Charles Matthews and Jonathan Weare. Langevin Markov chain Monte Carlo with stochastic gradients. arXiv:1805.08863, 2018.

[77] Jonathan C. Mattingly, Andrew M. Stuart, and Michael V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM Journal on Numerical Analysis*, 48 (2):552–577, 2010. doi: 10.1137/090770527.

[78] Sean P. Meyn and Richard L. Tweedie. Stability of Markovian processes I: Criteria for discrete-time chains. *Advances in Applied Probability*, 24(3):542–574, 1992. doi: 10.2307/1427479.

[79] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993. doi: 10.1007/978-1-4471-3267-7.

[80] Grigori N. Milstein and Michael V. Tretyakov. *Stochastic Numerics for Mathematical Physics*. Scientific Computation. Springer, 2004. doi: 10.1007/978-3-662-10063-9.

[81] Pierre Monmarché. High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electronic Journal of Statistics*, 15(2):4117–4166, 2021. doi: 10.1214/21-EJS1888.

[82] Tigran Nagapetyan, Andrew B. Duncan, Leonard Hasenclever, Sebastian J. Vollmer, Łukasz Szpruch, and Konstantinos C. Zygalakis. The true cost of stochastic gradient Langevin dynamics. arXiv:1706.02692, 2017.

[83] Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25858–25889, 2023.

[84] Christopher Nemeth and Paul Fearnhead. Stochastic gradient Markov chain Monte Carlo. *Journal of the American Statistical Association*, 116(533):433–450, 2021. doi: 10.1080/01621459.2020.1847120.

[85] Daniel Paulin and Peter A. Whalley. Correction to "Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations". *Journal of Machine Learning Research*, 25(376):1–9, 2024.

[86] Daniel Paulin and Peter A. Whalley. Theoretical guarantees for stochastic gradient Langevin dynamics. arXiv:2610.01651, 2026.

[87] Daniel Paulin and Peter A. Whalley. Theoretical guarantees for stochastic gradient sampling methods via Gaussian convolution inequalities. arXiv:2604.24632, 2026.

[88] Daniel Paulin, Peter A. Whalley, Neil K. Chada, and Benedict Leimkuhler. Sampling from Bayesian neural network posteriors with symmetric minibatch splitting Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 5014–5022, 2025.

[89] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, 2017.

[90] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. doi: 10.2307/3318418.

[91] Andreas Rößler. Runge–Kutta methods for the strong approximation of solutions of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 48(3):922–952, 2010. doi: 10.1137/09076636X.

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

[93] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient Langevin dynamics by using Fokker–Planck equation and Ito process. In *International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 982–990, 2014.

[94] Inass Sekkat and Gabriel Stoltz. Mini-batching error and adaptive Langevin dynamics. *Journal of Machine Learning Research*, 24(329):1–58, 2023.

[95] Xiaocheng Shang, Zhanxing Zhu, Benedict Leimkuhler, and Amos J. Storkey. Covariance-controlled adaptive Langevin thermostat for large-scale Bayesian sampling. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[96] Luke Shaw and Peter A. Whalley. Random reshuffling for stochastic gradient Langevin dynamics. arXiv:2501.16055, 2025.

[97] Denis Talay. Second-order discretization schemes of stochastic differential systems for the computation of the invariant law. *Stochastics and Stochastic Reports*, 29(1):13–36, 1990. doi: 10.1080/17442509008833606.

[98] Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990. doi: 10.1080/07362999008809220.

[99] Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33, 2016.

[100] Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[101] Gilles Vilmart. Postprocessed integrators for the high order integration of ergodic SDEs. *SIAM Journal on Scientific Computing*, 37(1):A201–A220, 2015. doi: 10.1137/140974328.

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

[103] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

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

[105] Bin Yang and Xiaojie Wang. Accelerating Langevin Monte Carlo via efficient stochastic Runge–Kutta methods beyond log-concavity. In *International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 146170–146197, 2026.

[106] Xuda Ye. Mean square error analysis of stochastic Runge–Kutta integrators. arXiv:2609.11528, 2026.

[107] Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics and Optimization*, 87(2):25, 2023. doi: 10.1007/s00245-022-09932-6.

[108] Difan Zou and Quanquan Gu. On the convergence of Hamiltonian Monte Carlo with stochastic gradients. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 13012–13022, 2021.

[109] Difan Zou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced Hamilton Monte Carlo methods. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 6028–6037, 2018.

[110] Difan Zou, Pan Xu, and Quanquan Gu. Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2936–2945, 2019.

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

[112] Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient Langevin dynamics for non-log-concave sampling. In *Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1152–1162, 2021.

A Proof of Theorem 1

This appendix proves Theorem 1: the exponential decay of the Kolmogorov solution u and of its derivatives of orders up to 4. The decay of u and of its derivatives of orders up to 3 is proved in [106, Theorem 1], so essentially only the derivative of order 4 needs to be estimated.[1] The bound (3.9) on ϕ_h then follows by summation over the time steps. Appendix A.1 writes out the variation processes and their pathwise bounds. Appendix A.2 states the estimates on the semigroup $e^{t\mathcal{L}}$ that the proof uses and proves the short-time bound (A.12). For a function g with bounded derivatives of orders 1 to 4, Appendix A.3 bounds $\nabla^4(e^{t\mathcal{L}}g)$ by differentiating the derivative formula (A.10) 3 times. Appendix A.4 combines this bound with the decay (A.8) and the short-time bound (A.12).

The proof is written for the potential U and the test function f of Assumptions 1 and 2, and it uses the derivatives of U up to order 5. Throughout this appendix, the constant C depends only on m, M, R and d, unless otherwise specified.

A.1 Variation processes

Let X_t be the solution of (1.1) started at $X_0 = x$, and write ∂_w for the derivative with respect to the initial state x along a vector w. For a function g on $\mathbb{R}^d$, the derivatives of $e^{t\mathcal{L}}g$ in the initial state are expressed through the variation processes. The variation processes of orders 1 to 4 are

$$J_t := \partial_x X_t, \qquad K_t[w] := \partial_w J_t, \qquad \Theta_t[w,w'] := \partial_{w'} K_t[w], \qquad \Xi_t[w,w',w''] := \partial_{w''}\Theta_t[w,w'], \tag{A.1}$$

where the variations of orders 2 to 4 are taken along vectors $w, w', w'' \in \mathbb{R}^d$, so that, like J_t, they take values in $\mathbb{R}^{d\times d}$. From the definition (A.1),

$$J_t w = \partial_w X_t, \qquad K_t[w]w' = \partial_w\partial_{w'} X_t,$$
$$\Theta_t[w,w']w'' = \partial_w\partial_{w'}\partial_{w''} X_t, \qquad \Xi_t[w,w',w'']w''' = \partial_w\partial_{w'}\partial_{w''}\partial_{w'''} X_t,$$

hence $K_t[w]w'$, $\Theta_t[w,w']w''$ and $\Xi_t[w,w',w'']w'''$ are symmetric in all their directions. Since the noise of (1.1) is additive, differentiating the dynamics in its initial state gives ordinary differential equations along the path, with the initial values $J_0 = I_d$ and $K_0 = \Theta_0 = \Xi_0 = 0$:

$$\frac{\mathrm{d}J_t}{\mathrm{d}t} = -\nabla^2 U(X_t)\, J_t, \tag{A.2}$$

$$\frac{\mathrm{d}K_t[w]}{\mathrm{d}t} = -\nabla^2 U(X_t)\, K_t[w] - \nabla^3 U(X_t)[J_t w]\, J_t, \tag{A.3}$$

$$\begin{aligned}\frac{\mathrm{d}\Theta_t[w,w']}{\mathrm{d}t} = {}& -\nabla^2 U(X_t)\,\Theta_t[w,w'] - \nabla^3 U(X_t)[J_t w]\, K_t[w'] - \nabla^3 U(X_t)[J_t w']\, K_t[w] \\ & - \nabla^3 U(X_t)\big[K_t[w]w'\big]\, J_t - \nabla^4 U(X_t)[J_t w, J_t w']\, J_t.\end{aligned} \tag{A.4}$$

These equations are derived in [106, Equations (A.1), (A.8) and (A.9)]. Differentiating (A.4) once more, along w'', gives the equation of the variation Ξ_t of order 4,

$$\begin{aligned}\frac{\mathrm{d}\Xi_t[w,w',w'']}{\mathrm{d}t} = {}& -\nabla^2 U(X_t)\,\Xi_t[w,w',w''] - \sum_{(2,1)} \nabla^3 U(X_t)[J_t w'']\,\Theta_t[w,w'] - \sum_{(2,1)} \nabla^3 U(X_t)\big[K_t[w]w'\big]\, K_t[w''] \\ & - \nabla^3 U(X_t)\big[\Theta_t[w,w']w''\big]\, J_t - \sum_{(2,1)} \nabla^4 U(X_t)[J_t w, J_t w']\, K_t[w''] \\ & - \sum_{(2,1)} \nabla^4 U(X_t)\big[K_t[w]w', J_t w''\big]\, J_t - \nabla^5 U(X_t)[J_t w, J_t w', J_t w'']\, J_t,\end{aligned} \tag{A.5}$$

where, for an expression F of the 3 directions in which the underlined pair forms a block, the sum runs over the partitions of $\{w, w', w''\}$ into a pair and a singleton,

$$\sum_{(2,1)} F(\underline{w, w'}, w'') := F(\underline{w, w'}, w'') + F(\underline{w, w''}, w') + F(\underline{w', w''}, w). \tag{A.6}$$

[1] The paper [106] is an earlier work of the first author, Xuda Ye, and is under review at the IMA Journal of Numerical Analysis. Beyond it, the proofs of this paper use no self-cited results that are not yet peer-reviewed.

In each sum $\sum_{(2,1)}$ of (A.5), the summand written out has the pair $\{w, w'\}$ and the singleton w''.

The bounds of the next lemma are Grönwall estimates, which grow exponentially in t. The factor t in the bounds on K_t, Θ_t and Ξ_t is what the proof of Lemma 16 uses at short times.

Lemma 11 *Let Assumption 1 hold. Then, pathwise, for all $t \geqslant 0$ and unit vectors w, w', w'',*

$$\begin{gathered}\|J_t\| \leqslant e^{Mt}, \qquad \|K_t[w]\| \leqslant Mt\, e^{2Mt}, \qquad \|\Theta_t[w,w']\| \leqslant M(1+3Mt)\, t\, e^{3Mt}, \\ \|\Xi_t[w,w',w'']\| \leqslant C(1+t)^2\, t\, e^{4Mt},\end{gathered} \tag{A.7}$$

with C depending only on M.

Proof. The bounds on J_t, K_t and Θ_t are [106, Lemma 6]. For $\Xi_t[w,w',w'']$, the matrix $J_t J_s^{-1}$ solves (A.2) on $[s,t]$ from I_d, so $\|J_t J_s^{-1}\| \leqslant e^{M(t-s)}$ by Grönwall's inequality, and Duhamel's principle with $\Xi_0 = 0$ gives

$$\Xi_t[w,w',w''] = \int_0^t J_t J_s^{-1} \left(\frac{\mathrm{d}\Xi_s[w,w',w'']}{\mathrm{d}s} + \nabla^2 U(X_s)\, \Xi_s[w,w',w''] \right) \mathrm{d}s.$$

The factor $\frac{\mathrm{d}\Xi_s[w,w',w'']}{\mathrm{d}s} + \nabla^2 U(X_s)\, \Xi_s[w,w',w'']$ of the integrand is the sum of the 14 terms of (A.5) other than $-\nabla^2 U(X_s)\, \Xi_s[w,w',w'']$. The bounds (A.7) on J_s, K_s and Θ_s, with $\|\nabla^i U\| \leqslant M$ for $i = 3,4,5$, bound these terms group by group: the 5 summands of the next display bound, in this order, the sum carrying $\nabla^3 U$ and Θ_s, the sum carrying $\nabla^3 U$ and 2 factors K_s, the term carrying $\nabla^3 U$ and $\Theta_s[w,w']w''$, the 2 sums carrying $\nabla^4 U$ together, and the term carrying $\nabla^5 U$:

$$\begin{aligned}&\left\| \frac{\mathrm{d}\Xi_s[w,w',w'']}{\mathrm{d}s} + \nabla^2 U(X_s)\, \Xi_s[w,w',w''] \right\| \\ &\leqslant 3M^2(1+3Ms)\, s\, e^{4Ms} + 3M^3 s^2 e^{4Ms} + M^2(1+3Ms)\, s\, e^{4Ms} + 6M^2 s\, e^{4Ms} + M e^{4Ms} \\ &= M\big(1 + 10Ms + 15M^2 s^2\big) e^{4Ms} \leqslant 15M(1+Ms)^2 e^{4Ms}.\end{aligned}$$

Hence

$$\|\Xi_t[w,w',w'']\| \leqslant \int_0^t e^{M(t-s)}\, 15M(1+Ms)^2 e^{4Ms}\, \mathrm{d}s \leqslant 15M(1+Mt)^2\, t\, e^{4Mt} \leqslant C(1+t)^2\, t\, e^{4Mt}.$$

□

A.2 Semigroup estimates

The estimates of this subsection hold for every $g \in C^4(\mathbb{R}^d)$ with bounded derivatives of orders 1 to 4; the rate $\lambda > 0$ depends only on m, M, R and d. Lemmas 12, 13 and 14 are proved in [106, Appendix A] under the bounds of Assumption 1 of orders up to 4, and are restated without proof, since the proof of Lemma 16 uses them explicitly. Lemma 15 extends the short-time bounds of [106, Equations (A.11) and (A.12)] to the derivative of order 4 and is proved here.

Lemma 12 *Let Assumption 1 hold. Then, for all $t \geqslant 0$,*

$$\big\|\nabla(e^{t\mathcal{L}} g)\big\|_\infty \leqslant C e^{-\lambda t} \|\nabla g\|_\infty; \tag{A.8}$$

and for $0 < t \leqslant 1$,

$$\big\|\nabla^2(e^{t\mathcal{L}} g)\big\|_\infty \leqslant C\, t^{-\frac{1}{2}}\, \|\nabla g\|_\infty, \qquad \big\|\nabla^3(e^{t\mathcal{L}} g)\big\|_\infty \leqslant C\, t^{-1}\, \|\nabla g\|_\infty. \tag{A.9}$$

The decay (A.8) is [106, Equation (A.7)], obtained from the reflection coupling of the overdamped Langevin dynamics [37]. The smoothing bounds (A.9) are [106, Lemmas 8 and 9], obtained by differentiating the derivative formula (A.10) below up to 2 times in the initial state.

The derivative formula [38], in the form of [106, Lemma 7], expresses the derivative of $e^{2\tau\mathcal{L}} g$ along a vector v through a stochastic integral of the variation process J_s of order 1.

Lemma 13 *Let Assumption 1 hold. Then, for $\tau > 0$, $\psi := e^{\tau\mathcal{L}} g$ and $v \in \mathbb{R}^d$,*

$$v \cdot \nabla(e^{2\tau\mathcal{L}} g)(x) = \frac{1}{\sqrt{2}\,\tau}\, \mathbb{E}^x\left[\psi(X_\tau) \int_0^\tau (J_s v) \cdot \mathrm{d}B_s\right], \qquad \forall x \in \mathbb{R}^d. \tag{A.10}$$

The next lemma, [106, Equation (A.17)] in the proof of [106, Lemma 8], bounds the variance of $\psi(X_\tau)$.

Lemma 14 *Let Assumption 1 hold. Then, for $\tau > 0$, $\psi := e^{\tau\mathcal{L}} g$ and $X_0 = x \in \mathbb{R}^d$,*

$$\mathrm{Var}\big(\psi(X_\tau)\big) \leqslant 2d\,\tau\, e^{4M\tau}\, \|\nabla g\|_\infty^2. \tag{A.11}$$

The randomness of $\psi(X_\tau)$ comes from the Brownian motion $(B_t)_{t\geqslant 0}$ on $[0,\tau]$, which drives X_t from $X_0 = x$.

The last estimate of this subsection bounds the derivatives of $e^{t\mathcal{L}} g$ over a finite time; it covers the times $t \leqslant 1$ in the proof of Theorem 1 and supplies the derivative bounds needed in the proof of Lemma 16 and in the case $t \geqslant 1$ of that theorem.

Lemma 15 *Let Assumption 1 hold. There is C, depending only on M, such that for every $g \in C^4(\mathbb{R}^d)$ with bounded derivatives of orders 1 to 4 and all $t \geqslant 0$,*

$$\big\|\nabla^j (e^{t\mathcal{L}} g)\big\|_\infty \leqslant C(1+t)^3 e^{4Mt} \sum_{i=1}^{j} \|\nabla^i g\|_\infty, \qquad j = 1,2,3,4. \tag{A.12}$$

Proof. For $j \leqslant 3$, (A.12) follows from the bounds [106, Equations (A.11) and (A.12)],

$$\begin{aligned}
&\big\|\nabla(e^{t\mathcal{L}} g)\big\|_\infty \leqslant e^{Mt}\|\nabla g\|_\infty, \qquad \big\|\nabla^2(e^{t\mathcal{L}} g)\big\|_\infty \leqslant e^{2Mt}\|\nabla^2 g\|_\infty + Mt\, e^{2Mt}\|\nabla g\|_\infty,\\
&\big\|\nabla^3(e^{t\mathcal{L}} g)\big\|_\infty \leqslant e^{3Mt}\|\nabla^3 g\|_\infty + 3Mt\, e^{3Mt}\|\nabla^2 g\|_\infty + M(1+3Mt)\, t\, e^{3Mt}\|\nabla g\|_\infty,
\end{aligned}$$

whose coefficients are each at most $C(1+t)^3 e^{4Mt}$.

For $j = 4$, fix $x \in \mathbb{R}^d$ and unit vectors w, w', w'', w'''. By (A.1), the derivatives of X_t in its initial state along these vectors are $J_t w$, $K_t[w]w'$, $\Theta_t[w,w']w''$ and $\Xi_t[w,w',w'']w'''$, together with the same expressions along the other subsets of $\{w, w', w'', w'''\}$. The chain rule for higher derivatives applied to $g(X_t)$, and the exchange of the derivatives with the expectation, justified by the pathwise bounds (A.7) and the bounded derivatives of g, give

$$\begin{aligned}
\nabla^4(e^{t\mathcal{L}} g)(x)[w,w',w'',w'''] = \mathbb{E}^x\Big[&\nabla^4 g(X_t)[J_t w, J_t w', J_t w'', J_t w'''] + \sum_{(2,1,1)} \nabla^3 g(X_t)\big[K_t[w]w', J_t w'', J_t w'''\big]\\
&+ \sum_{(3,1)} \nabla^2 g(X_t)\big[\Theta_t[w,w']w'', J_t w'''\big] + \sum_{(2,2)} \nabla^2 g(X_t)\big[K_t[w]w', K_t[w'']w'''\big]\\
&+ \nabla g(X_t)\cdot \Xi_t[w,w',w'']w'''\Big],
\end{aligned} \tag{A.13}$$

where, for an expression F of the 4 directions in which each underlined group of directions forms a block, the sums run over the partitions of $\{w, w', w'', w'''\}$ into blocks of the sizes in the subscript:

$$\begin{aligned}
\sum_{(2,1,1)} F(\underline{w,w'},w'',w''') &:= F(\underline{w,w'},w'',w''') + F(\underline{w,w''},w',w''') + F(\underline{w,w'''},w',w'')\\
&\quad + F(\underline{w',w''},w,w''') + F(\underline{w',w'''},w,w'') + F(\underline{w'',w'''},w,w'),\\
\sum_{(3,1)} F(\underline{w,w',w''},w''') &:= F(\underline{w,w',w''},w''') + F(\underline{w,w',w'''},w'') + F(\underline{w,w'',w'''},w') + F(\underline{w',w'',w'''},w),\\
\sum_{(2,2)} F(\underline{w,w'},\underline{w'',w'''}) &:= F(\underline{w,w'},\underline{w'',w'''}) + F(\underline{w,w''},\underline{w',w'''}) + F(\underline{w,w'''},\underline{w',w''}).
\end{aligned}$$

By (A.7), the terms of the 5 groups in (A.13) are bounded by

$$e^{4Mt}\|\nabla^4 g\|_\infty, \qquad Mt\,e^{4Mt}\|\nabla^3 g\|_\infty, \qquad M(1+3Mt)\,t\,e^{4Mt}\|\nabla^2 g\|_\infty,$$
$$M^2t^2e^{4Mt}\|\nabla^2 g\|_\infty, \qquad C(1+t)^2\,t\,e^{4Mt}\|\nabla g\|_\infty,$$

each at most $C(1+t)^3e^{4Mt}$ times the norm of the derivative of g it carries. Summing the 15 terms and taking the supremum over x and over the unit vectors gives (A.12) for $j=4$. □

A.3 The derivative of order 4

This subsection extends the smoothing bounds (A.9) to the derivative of order 4; the proof uses Lemmas 11, 12, 13, 14 and 15.

Lemma 16 *Let Assumption 1 hold. For every $g \in C^4(\mathbb{R}^d)$ with bounded derivatives of orders 1 to 4 and every $0 < t \leqslant 1$,*

$$\big\|\nabla^4(e^{t\mathcal{L}}g)(x)\big\| \leqslant C\,t^{-\frac{3}{2}}\,\|\nabla g\|_\infty, \qquad \forall x \in \mathbb{R}^d. \tag{A.14}$$

Proof. Set $\tau = t/2 \leqslant \frac{1}{2}$ and $\psi = e^{\tau\mathcal{L}}g$, so that $e^{2\tau\mathcal{L}}g = e^{t\mathcal{L}}g$, and fix $x \in \mathbb{R}^d$ and unit vectors v, w, w', w''. By (A.12), ψ has bounded derivatives of orders 1 to 4. Differentiate (A.10) in Lemma 13 along w, w' and w''. By (A.1), a derivative that falls on the stochastic integral raises the order of the variation process in its integrand, from J_sv to $K_s[w]v$, to $\Theta_s[w,w']v$ and to $\Xi_s[w,w',w'']v$. Distributing the 3 derivatives between $\psi(X_\tau)$ and the stochastic integral gives

$$\begin{aligned}
\sqrt{2}\,\tau\,\nabla^4(e^{t\mathcal{L}}g)(x)[v,w,w',w''] &= \mathbb{E}^x\Big[\partial_w\partial_{w'}\partial_{w''}\psi(X_\tau)\int_0^\tau (J_sv)\cdot \mathrm{d}B_s\Big] \\
&+ \sum_{(2,1)} \mathbb{E}^x\Big[\partial_w\partial_{w'}\psi(X_\tau)\int_0^\tau \big(K_s[w'']v\big)\cdot \mathrm{d}B_s\Big] + \sum_{(2,1)} \mathbb{E}^x\Big[\partial_{w''}\psi(X_\tau)\int_0^\tau \big(\Theta_s[w,w']v\big)\cdot \mathrm{d}B_s\Big] \\
&+ \mathbb{E}^x\Big[\big(\psi(X_\tau) - \mathbb{E}^x\big[\psi(X_\tau)\big]\big)\int_0^\tau \big(\Xi_s[w,w',w'']v\big)\cdot \mathrm{d}B_s\Big],
\end{aligned} \tag{A.15}$$

where $\sum_{(2,1)}$ is the sum over the partitions of $\{w,w',w''\}$ defined in (A.6).

In the term of (A.15) carrying Ξ_s, $\psi(X_\tau)$ is centered at its mean $\mathbb{E}^x\big[\psi(X_\tau)\big] = e^{t\mathcal{L}}g(x)$. The centering does not change the value of this term, since the stochastic integral has mean 0, and it makes (A.11) applicable. The exchanges of the derivatives with the expectation and with the stochastic integral are justified as in the proof of [106, Lemma 8]. The variation processes are differentiable in the initial state along each path and bounded by (A.7) uniformly in x, so the difference quotients of the integrands converge in L^2 over $[0,\tau]$ and the paths, and by the Itô isometry those of the stochastic integrals converge in L^2. Since ψ has bounded derivatives of orders 1 to 4, the difference quotients of the factors in front of the stochastic integrals also converge in L^2, and the Cauchy–Schwarz inequality passes the limits through the expectation.

The derivatives of $\psi(X_\tau)$ in (A.15) follow from the chain rule and (A.1):

$$\partial_{w''}\psi(X_\tau) = \nabla\psi(X_\tau)\cdot J_\tau w'', \qquad \partial_w\partial_{w'}\psi(X_\tau) = \nabla^2\psi(X_\tau)[J_\tau w, J_\tau w'] + \nabla\psi(X_\tau)\cdot K_\tau[w]w',$$
$$\partial_w\partial_{w'}\partial_{w''}\psi(X_\tau) = \nabla^3\psi(X_\tau)[J_\tau w, J_\tau w', J_\tau w''] + \sum_{(2,1)} \nabla^2\psi(X_\tau)\big[K_\tau[w]w', J_\tau w''\big] + \nabla\psi(X_\tau)\cdot\Theta_\tau[w,w']w''.$$

Since $\tau \leqslant \frac{1}{2}$, (A.12) and (A.9) at time τ give

$$\|\nabla\psi\|_\infty \leqslant C\|\nabla g\|_\infty, \qquad \|\nabla^2\psi\|_\infty \leqslant C\tau^{-\frac{1}{2}}\|\nabla g\|_\infty, \qquad \|\nabla^3\psi\|_\infty \leqslant C\tau^{-1}\|\nabla g\|_\infty,$$

and (A.7) gives $\|J_s\| \leqslant C$ and $\|K_s[w]\|, \|\Theta_s[w,w']\|, \|\Xi_s[w,w',w'']\| \leqslant Cs$ for $s \leqslant \tau$. The bounds on $\nabla^j\psi$ and on the variation processes then give

$$\big|\partial_{w''}\psi(X_\tau)\big| \leqslant C\|\nabla g\|_\infty, \quad \big|\partial_w\partial_{w'}\psi(X_\tau)\big| \leqslant C\tau^{-\frac{1}{2}}\|\nabla g\|_\infty, \quad \big|\partial_w\partial_{w'}\partial_{w''}\psi(X_\tau)\big| \leqslant C\tau^{-1}\|\nabla g\|_\infty, \tag{A.16}$$

and (A.11) bounds the centered factor of the term carrying Ξ_s:

$$\operatorname{Var}\big(\psi(X_\tau)\big)^{\frac12} = \Big(\mathbb{E}^x\Big|\psi(X_\tau) - \mathbb{E}^x\big[\psi(X_\tau)\big]\Big|^2\Big)^{\frac12} \leqslant C\tau^{\frac12}\|\nabla g\|_\infty. \tag{A.17}$$

By the Itô isometry and (A.7),

$$\begin{aligned} \mathbb{E}^x\Big|\int_0^\tau (J_s v)\cdot \mathrm{d}B_s\Big|^2 &\leqslant C\tau, \qquad \mathbb{E}^x\Big|\int_0^\tau \big(K_s[w'']v\big)\cdot \mathrm{d}B_s\Big|^2 \leqslant C\tau^3, \\ \mathbb{E}^x\Big|\int_0^\tau \big(\Theta_s[w,w']v\big)\cdot \mathrm{d}B_s\Big|^2 &\leqslant C\tau^3, \qquad \mathbb{E}^x\Big|\int_0^\tau \big(\Xi_s[w,w',w'']v\big)\cdot \mathrm{d}B_s\Big|^2 \leqslant C\tau^3, \end{aligned} \tag{A.18}$$

and these bounds hold for every term of the sums $\sum_{(2,1)}$ in (A.15), since w, w' and w'' are arbitrary unit vectors. The Cauchy–Schwarz inequality, with (A.16), (A.18) and (A.17), bounds the terms of the 4 groups in (A.15) by

$$C\tau^{-1}\cdot\tau^{\frac12}, \qquad C\tau^{-\frac12}\cdot\tau^{\frac32}, \qquad C\cdot\tau^{\frac32}, \qquad C\tau^{\frac12}\cdot\tau^{\frac32}$$

times $\|\nabla g\|_\infty$. For $\tau \leqslant \frac12$, each is at most $C\tau^{-\frac12}\|\nabla g\|_\infty$, so dividing (A.15) by $\sqrt{2}\,\tau$ and taking the supremum over the unit vectors gives

$$\big\|\nabla^4(e^{t\mathcal{L}}g)(x)\big\| \leqslant C\tau^{-\frac32}\|\nabla g\|_\infty,$$

which is (A.14) with $\tau = t/2$. □

A.4 Conclusion of the proof

Proof of Theorem 1. The bound on $|u(x,t)|$ in (3.8) and the bounds on $\nabla^j u(x,t)$ for $1 \leqslant j \leqslant 3$ are [106, Theorem 1], whose assumptions on U and f are the bounds of Assumptions 1 and 2 up to the orders 5 and 3, respectively; decreasing λ if necessary, these bounds and (A.8) hold with the same rate λ. For the order 4, since $\pi(f)$ is a constant, $\nabla^4 u(x,t) = \nabla^4(e^{t\mathcal{L}}f)(x)$, and we split at $t = 1$. For $t \leqslant 1$, the short-time bound (A.12) with $g = f$ and Assumption 2 give

$$\|\nabla^4 u(x,t)\| \leqslant C\sum_{i=1}^4 \|\nabla^i f\|_\infty \leqslant CL \leqslant CL\,e^{\lambda}\,e^{-\lambda t}.$$

For $t \geqslant 1$, the semigroup property gives $e^{t\mathcal{L}}f = e^{\mathcal{L}}\big(e^{(t-1)\mathcal{L}}f\big)$, and $e^{(t-1)\mathcal{L}}f$ has bounded derivatives of orders 1 to 4, since (A.12) at time $t-1$ bounds them. Hence (A.14) at time 1, with $e^{(t-1)\mathcal{L}}f$ in place of g, bounds $\|\nabla^4 u(x,t)\|$ by $C\|\nabla(e^{(t-1)\mathcal{L}}f)\|_\infty$, and (A.8) at time $t-1$ bounds this gradient, so that

$$\|\nabla^4 u(x,t)\| \leqslant C\,\big\|\nabla(e^{(t-1)\mathcal{L}}f)\big\|_\infty \leqslant C\,e^{-\lambda(t-1)}\,\|\nabla f\|_\infty \leqslant CL\,e^{\lambda}\,e^{-\lambda t}.$$

Together, the cases $t \leqslant 1$ and $t \geqslant 1$ give the bound on $\nabla^4 u(x,t)$ in (3.8).

For the discrete Poisson solution (3.5), the geometric bounds (3.8) on the summands justify differentiating the series term by term. Summing the derivative bound in (3.8) over $t = kh$ gives, for $\lambda h \leqslant 1$ and $1 \leqslant j \leqslant 4$,

$$\|\nabla^j \phi_h(x)\| \leqslant h\sum_{k=0}^\infty \|\nabla^j u(x,kh)\| \leqslant CL\,h\sum_{k=0}^\infty e^{-\lambda kh} = \frac{CL\,h}{1-e^{-\lambda h}} \leqslant \frac{2CL}{\lambda},$$

which is (3.9) after $2/\lambda$ is absorbed into C. □

B Proofs for Section 4

Throughout this appendix, the constant C depends only on m, M, R and d. The next lemma records the dissipativity of U used in the moment estimates, and the Gaussian tail of π that follows from it.

Lemma 17 *Let Assumption 1 hold. For all $x \in \mathbb{R}^d$,*

$$\langle x, \nabla U(x)\rangle \geqslant \frac{m}{4}|x|^2 - C, \qquad U(x) - U(0) \geqslant \frac{m}{8}|x|^2 - C, \tag{B.1}$$

and consequently $\pi(x) \leqslant Ce^{-\frac{m}{8}|x|^2}$ and $\int_{\mathbb{R}^d} |x|^q \, \pi(\mathrm{d}x) \leqslant C_q$ for every $q \geqslant 1$.

Proof. By [106, Lemma 4], with $R_1 := 4R(1 + M/m)$,

$$\langle \nabla U(x) - \nabla U(y), x - y\rangle \geqslant -M|x-y|^2, \quad \langle \nabla U(x) - \nabla U(y), x - y\rangle \geqslant \frac{m}{2}|x-y|^2 \ \text{ for } |x-y| \geqslant R_1. \tag{B.2}$$

Applying (B.2) to the pair $(x, 0)$ and using $|\nabla U(0)| \leqslant M$,

$$\langle x, \nabla U(x)\rangle = \langle x, \nabla U(x) - \nabla U(0)\rangle + \langle x, \nabla U(0)\rangle \geqslant \begin{cases} \frac{m}{2}|x|^2 - M|x|, & |x| \geqslant R_1, \\ -M|x|^2 - M|x|, & |x| < R_1, \end{cases}$$

and both cases are at least $\frac{m}{4}|x|^2 - C$, which is the drift bound in (B.1). Integrating along the segment from 0 to x, with (B.2) at the pair $(rx, 0)$ and the integral split at $r|x| = R_1$ when $|x| \geqslant R_1$,

$$U(x) - U(0) = \int_0^1 \langle \nabla U(rx), x\rangle \, \mathrm{d}r \geqslant \frac{m}{4}\big(|x|^2 - R_1^2\big) - \frac{M}{2}R_1^2 - M|x|,$$

which gives the lower bound on $U(x) - U(0)$ in (B.1). The upper bound $U(y) - U(0) \leqslant M|y| + \frac{M}{2}|y|^2$ and the lower bound on $U(x) - U(0)$ in (B.1) give

$$\pi(x) = \frac{e^{-(U(x)-U(0))}}{\int_{\mathbb{R}^d} e^{-(U(y)-U(0))}\mathrm{d}y} \leqslant \frac{e^{C - \frac{m}{8}|x|^2}}{\int_{\mathbb{R}^d} e^{-M|y| - \frac{M}{2}|y|^2}\mathrm{d}y} \leqslant Ce^{-\frac{m}{8}|x|^2},$$

and hence $\int_{\mathbb{R}^d} |x|^q \, \pi(\mathrm{d}x) \leqslant C_q$ for every $q \geqslant 1$. □

B.1 Proof of Lemma 1: modified potential estimates

Proof of Lemma 1. The proof bounds the derivatives of V and of $V\theta_h$, which gives (4.14), and then chooses δ so that $\tilde{U}_h$ satisfies (4.13) and the convexity condition. Under Assumption 1, $|\nabla U(x)| \leqslant M(1 + |x|)$ and $\|\nabla^j U(x)\| \leqslant M$ for $2 \leqslant j \leqslant 7$. The derivatives of $|\nabla U|^2$ follow from the Leibniz rule,

$$\nabla^i |\nabla U|^2 = \sum_{l=0}^{i} \binom{i}{l} \operatorname{Sym} \big\langle \nabla^{l+1} U, \, \nabla^{i-l+1} U \big\rangle, \qquad 0 \leqslant i \leqslant 5, \tag{B.3}$$

where $\langle \cdot, \cdot \rangle$ contracts one index of each factor, so that both sides are tensors of order i. The symmetrization is needed: $\langle \nabla^{l+1} U, \nabla^{i-l+1} U\rangle$ is symmetric within its first l indices and within its last $i - l$ indices, but in general not across these groups.

For $i \geqslant 1$, each term of (B.3) contains the factor ∇U at most once and is therefore bounded by $C(1+|x|)$; the derivatives $\nabla^i \Delta U$, contractions of $\nabla^{i+2} U$, are bounded by dM for $0 \leqslant i \leqslant 5$. Hence $V = |\nabla U|^2 - \Delta U$ of (4.10) satisfies

$$|V(x)| \leqslant C\big(1 + |x|\big)^2, \qquad \|\nabla^i V(x)\| \leqslant C\big(1 + |x|\big), \quad 1 \leqslant i \leqslant 5, \qquad \forall x \in \mathbb{R}^d, \tag{B.4}$$

where C depends only on M and d.

Let $\delta \in (0,1]$, fixed by (B.9) at the end of the proof, and let $h \leqslant \delta$ throughout. By (4.11), $\tilde{U}_h - U = \frac{h}{4} V\theta_h$ with $\theta_h(x) = \theta(hx/\delta)$. The function θ_h is supported in $\{|x| \leqslant 2\delta/h\}$, where

$$\frac{h}{\delta}\big(1+|x|\big) \leqslant \frac{h}{\delta} + 2 \leqslant 3, \tag{B.5}$$

and $\nabla^l \theta_h(x) = (h/\delta)^l \, (\nabla^l \theta)(hx/\delta)$ is bounded by $C(h/\delta)^l$ for $0 \leqslant l \leqslant 5$, with C depending on the fixed function θ only. By the Leibniz rule again,

$$\nabla^i (V\theta_h) = \sum_{l=0}^{i} \binom{i}{l} \operatorname{Sym}\big(\nabla^{i-l} V \otimes \nabla^l \theta_h\big), \qquad 1 \leqslant i \leqslant 5. \tag{B.6}$$

On the support of θ_h, (B.4), $h \leqslant \delta$ and (B.5) give

$$\|\nabla^{i-l} V(x) \otimes \nabla^l \theta_h(x)\| \leqslant \begin{cases} C\big(1+|x|\big), & 0 \leqslant l \leqslant i-1, \\ C(h/\delta)^i \big(1+|x|\big)^2 \leqslant 3C\big(1+|x|\big), & l = i. \end{cases}$$

Hence, by (B.6), the derivatives of $V\theta_h$ of orders 1 to 5 grow at most linearly on the support of θ_h,

$$\big\|\nabla^i (V\theta_h)(x)\big\| \leqslant C\big(1+|x|\big), \qquad 1 \leqslant i \leqslant 5, \qquad \forall x \in \mathbb{R}^d : |x| \leqslant \frac{2\delta}{h}. \tag{B.7}$$

Since $\nabla^i (V\theta_h)$ vanishes for $|x| > 2\delta/h$, (B.7) gives (4.14), and, by (B.5),

$$\|\nabla^i \tilde{U}_h(x) - \nabla^i U(x)\| = \frac{h}{4} \big\|\nabla^i (V\theta_h)(x)\big\| \leqslant C\delta, \qquad 1 \leqslant i \leqslant 5, \qquad \forall x \in \mathbb{R}^d, \tag{B.8}$$

with C independent of δ and h. Choose $\delta \in (0,1]$ so small that this constant satisfies

$$C\delta \leqslant \min\Big\{\frac{m}{2}, M\Big\}, \tag{B.9}$$

and let $h_0 \leqslant \delta$. Then (B.8) gives

$$\nabla^2 \tilde{U}_h(x) \succeq \nabla^2 U(x) - C\delta \, I_d, \qquad \|\nabla^i \tilde{U}_h(x)\| \leqslant \|\nabla^i U(x)\| + C\delta, \quad 1 \leqslant i \leqslant 5, \qquad \forall x \in \mathbb{R}^d, \tag{B.10}$$

which, with Assumption 1 and (B.9), gives (4.13) and the convexity condition of Lemma 1. Finally, $\tilde{U}_h \in C^5(\mathbb{R}^d)$, since $\theta_h \in C^\infty(\mathbb{R}^d)$ and $U \in C^7(\mathbb{R}^d)$ gives $V \in C^5(\mathbb{R}^d)$. □

B.2 Proof of Lemma 2: modified bias estimate

Proof of Lemma 2. The proof bounds the distance between $\tilde{\pi}_h * \mathcal{N}(0, \frac{h}{2} I_d)$ and π, weighted by $1+|x|$, by Ch^2. It writes $\tilde{\pi}_h$ as $\pi \, e^{-(\tilde{U}_h - U)}$ divided by its normalizing constant, expands this constant to the order h, where, since $\pi(V) = 0$, the term of order h comes only from the cutoff and is of order h^2, and expands $\tilde{\pi}_h * \mathcal{N}(0, \frac{h}{2} I_d)$ to the order 4 along the Gaussian shift. Since $\tilde{\pi}_h \propto e^{-\tilde{U}_h} = e^{-U} e^{-(\tilde{U}_h - U)}$, where $\tilde{U}_h - U = \frac{h}{4} V\theta_h$ by (4.11), we have

$$\tilde{\pi}_h(x) = \frac{\pi(x) \, e^{-(\tilde{U}_h - U)(x)}}{\pi\big(e^{-(\tilde{U}_h - U)}\big)}. \tag{B.11}$$

By (B.4), $|\tilde{U}_h - U| \leqslant Ch(1+|x|^2)$, while $\pi(x) \leqslant C e^{-\frac{m}{8}|x|^2}$ by Lemma 17. Shrinking h_0 so that $Ch_0 \leqslant \frac{m}{16}$, with C the constant of the bound on $|\tilde{U}_h - U|$, we have for $h \leqslant h_0$,

$$e^{|\tilde{U}_h - U|} \leqslant C e^{\frac{m}{16}|x|^2}, \qquad \pi\big((1+|x|)^5 \, e^{|\tilde{U}_h - U|}\big) \leqslant C. \tag{B.12}$$

Taylor's formula for the exponential gives

$$e^{-(\tilde{U}_h - U)} = 1 - \big(\tilde{U}_h - U\big) + \varrho_2(x), \quad |\varrho_2(x)| \leqslant \frac{1}{2}\big(\tilde{U}_h - U\big)^2 e^{|\tilde{U}_h - U|} \leqslant Ch^2 \big(1+|x|\big)^4 e^{|\tilde{U}_h - U|}, \tag{B.13}$$

hence the remainder $\varrho_2(x)$ satisfies

$$\pi\big((1+|x|)\,|\varrho_2(x)|\big) \leqslant Ch^2\,\pi\big((1+|x|)^5 e^{|\tilde U_h - U|}\big) \leqslant Ch^2. \tag{B.14}$$

Next, $\pi(V) = -\pi(\mathcal{L}U) = 0$ by (4.9) with $f = 1$ and (4.10). Replacing V by $V\theta_h$ changes $\pi(V)$ by an exponentially small amount: since $1-\theta_h$ vanishes on $\{|x| \leqslant \delta/h\}$,

$$\pi\big((1+|x|)\,|V|\,(1-\theta_h)\big) \leqslant C\,\pi\big((1+|x|)^3\,\mathbf{1}_{\{|x|\geqslant\delta/h\}}\big) \leqslant Ce^{-\frac{m\delta^2}{16h^2}} \leqslant Ch^2, \tag{B.15}$$

by (B.4) and $\pi(x) \leqslant Ce^{-\frac{m}{8}|x|^2} \leqslant Ce^{-\frac{m\delta^2}{16h^2}}\,e^{-\frac{m}{16}|x|^2}$ on $\{|x| \geqslant \delta/h\}$. Inserting (B.13) into the denominator of (B.11), and using $\pi(V) = 0$, (B.14) and (B.15), we obtain

$$\pi\big(e^{-(\tilde U_h - U)}\big) = 1 - \frac{h}{4}\,\pi(V\theta_h) + \pi(\varrho_2) = 1 + \frac{h}{4}\,\pi\big(V(1-\theta_h)\big) + \pi(\varrho_2) = 1 + \mathcal{O}(h^2), \tag{B.16}$$

where $\mathcal{O}(h^2)$ is bounded by Ch^2. After shrinking h_0, $\pi\big(e^{-(\tilde U_h - U)}\big) \geqslant 1 - Ch^2 \geqslant \frac12$, hence

$$\left|\frac{1}{\pi\big(e^{-(\tilde U_h - U)}\big)} - 1\right| \leqslant Ch^2. \tag{B.17}$$

Let $\chi \sim \mathcal{N}(0, I_d)$, so that

$$\Big(\tilde\pi_h * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big)\Big)(x) = \mathbb{E}\Big[\tilde\pi_h\Big(x - \sqrt{\frac{h}{2}}\,\chi\Big)\Big].$$

Since $\tilde U_h \in C^5$ by Lemma 1, $\tilde\pi_h \in C^5$. Expanding $\tilde\pi_h$ about x to the order 4 along $-\sqrt{h/2}\,\chi$,

$$\tilde\pi_h\Big(x - \sqrt{\frac{h}{2}}\,\chi\Big) = \tilde\pi_h(x) - \sqrt{\frac{h}{2}}\,\nabla\tilde\pi_h(x)\cdot\chi + \frac{h}{4}\,\nabla^2\tilde\pi_h(x)[\chi,\chi] - \frac16\Big(\frac{h}{2}\Big)^{\frac32}\nabla^3\tilde\pi_h(x)\big[\chi^{\otimes 3}\big] + \varrho_4^\chi(x),$$

with

$$|\varrho_4^\chi(x)| \leqslant \frac{h^2}{24}\,|\chi|^4 \int_0^1 (1-s)^3 \left\|\nabla^4\tilde\pi_h\Big(x - s\sqrt{\frac{h}{2}}\,\chi\Big)\right\| \mathrm{d}s. \tag{B.18}$$

Under the expectation the terms of odd order in χ vanish and the term of order 2 has mean $\frac{h}{4}\,\Delta\tilde\pi_h(x)$, hence

$$\Big(\tilde\pi_h * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big)\Big)(x) = \tilde\pi_h(x) + \frac{h}{4}\,\Delta\tilde\pi_h(x) + \mathbb{E}\big[\varrho_4^\chi(x)\big]. \tag{B.19}$$

Differentiating $\tilde\pi_h \propto e^{-\tilde U_h}$ to the order 2,

$$\Delta\tilde\pi_h = \tilde\pi_h\,\big(|\nabla\tilde U_h|^2 - \Delta\tilde U_h\big). \tag{B.20}$$

By (B.11), (B.19) and (B.20),

$$\Big(\tilde\pi_h * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big)\Big)(x) = \frac{\pi(x)\,e^{-(\tilde U_h - U)(x)}}{\pi\big(e^{-(\tilde U_h - U)}\big)}\Big(1 + \frac{h}{4}\big(|\nabla\tilde U_h(x)|^2 - \Delta\tilde U_h(x)\big)\Big) + \mathbb{E}\big[\varrho_4^\chi(x)\big]. \tag{B.21}$$

By (4.13), (4.14) and (4.10), $\big||\nabla\tilde U_h|^2 - \Delta\tilde U_h\big| \leqslant C(1+|x|)^2$ and

$$\big||\nabla\tilde U_h|^2 - \Delta\tilde U_h - V\big| \leqslant |\nabla\tilde U_h - \nabla U|\,|\nabla\tilde U_h + \nabla U| + |\Delta\tilde U_h - \Delta U| \leqslant Ch\big(1+|x|\big)^2. \tag{B.22}$$

Taylor's formula for the exponential, with $\tilde U_h - U = \frac{h}{4}\,V\theta_h$, the bound $|V| \leqslant C(1+|x|)^2$ of (B.4), the bound on $|\nabla\tilde U_h|^2 - \Delta\tilde U_h$ and the bound on $e^{|\tilde U_h - U|}$ in (B.12) give

$$\left|e^{-(\tilde U_h - U)}\Big(1 + \frac{h}{4}\big(|\nabla\tilde U_h|^2 - \Delta\tilde U_h\big)\Big) - 1 - \frac{h}{4}\big(|\nabla\tilde U_h|^2 - \Delta\tilde U_h - V\theta_h\big)\right| \leqslant Ch^2\big(1+|x|\big)^6 e^{\frac{m}{16}|x|^2}.$$

Since $V\theta_h = V - V(1-\theta_h)$, (B.22) gives

$$\left| e^{-(\tilde U_h - U)} \Big(1 + \frac{h}{4}\big(|\nabla \tilde U_h|^2 - \Delta \tilde U_h\big)\Big) - 1 \right| \leqslant Ch^2 \big(1+|x|\big)^6 e^{\frac{m}{16}|x|^2} + \frac{h}{4}\,|V|\big(1-\theta_h\big). \tag{B.23}$$

Multiplying (B.23) by $(1+|x|)\pi(x)$ and integrating over $\mathbb{R}^d$, the term $Ch^2(1+|x|)^6 e^{\frac{m}{16}|x|^2}$ contributes at most $Ch^2 \int_{\mathbb{R}^d} (1+|x|)^7 e^{-\frac{m}{16}|x|^2}\,\mathrm{d}x \leqslant Ch^2$ by the density bound $\pi(x) \leqslant Ce^{-\frac{m}{8}|x|^2}$ of Lemma 17, and the term $\frac{h}{4}\,|V|(1-\theta_h)$ contributes at most Ch^3 by (B.15), so

$$\int_{\mathbb{R}^d} \big(1+|x|\big) \left| e^{-(\tilde U_h - U)} \Big(1 + \frac{h}{4}\big(|\nabla \tilde U_h|^2 - \Delta \tilde U_h\big)\Big) - 1 \right| \pi(x)\,\mathrm{d}x \leqslant Ch^2. \tag{B.24}$$

Differentiating $\tilde\pi_h \propto e^{-\tilde U_h}$ to the order 4 and using (4.13) with $|\nabla \tilde U_h(y)| \leqslant 2M(1+|y|)$,

$$\|\nabla^4 \tilde\pi_h(y)\| \leqslant C\big(1+|y|\big)^4 \tilde\pi_h(y), \qquad y \in \mathbb{R}^d.$$

By (B.11), (B.12), the lower bound $\pi(e^{-(\tilde U_h - U)}) \geqslant \frac12$ and the density bound of Lemma 17, $\tilde\pi_h\big((1+|x|)^q\big) \leqslant 2C\,\pi\big((1+|x|)^q e^{\frac{m}{16}|x|^2}\big) \leqslant C_q$ for $q \geqslant 1$. For each fixed $s \in [0,1]$, the substitution $y = x - s\sqrt{h/2}\,\chi$ and the bound on $\tilde\pi_h\big((1+|x|)^q\big)$ with $q=5$ give

$$\int_{\mathbb{R}^d} \big(1+|x|\big) \left(1 + \Big|x - s\sqrt{\frac{h}{2}}\,\chi\Big|\right)^4 \tilde\pi_h\Big(x - s\sqrt{\frac{h}{2}}\,\chi\Big)\,\mathrm{d}x \leqslant \big(1+|\chi|\big) \int_{\mathbb{R}^d} \big(1+|y|\big)^5 \tilde\pi_h(y)\,\mathrm{d}y \leqslant C\big(1+|\chi|\big).$$

Hence, by (B.18), Tonelli's theorem and $\mathbb{E}\big[|\chi|^4(1+|\chi|)\big] \leqslant C$,

$$\begin{aligned} &\int_{\mathbb{R}^d} \big(1+|x|\big)\mathbb{E}|\varrho_4^\chi(x)|\,\mathrm{d}x \\ &\leqslant Ch^2\,\mathbb{E}\left[|\chi|^4 \int_0^1 (1-s)^3 \int_{\mathbb{R}^d} \big(1+|x|\big) \left(1 + \Big|x - s\sqrt{\frac{h}{2}}\,\chi\Big|\right)^4 \tilde\pi_h\Big(x - s\sqrt{\frac{h}{2}}\,\chi\Big)\,\mathrm{d}x\,\mathrm{d}s\right] \leqslant Ch^2. \end{aligned} \tag{B.25}$$

By (B.21), the difference between $\tilde\pi_h * \mathcal{N}(0, \frac{h}{2} I_d)$ and π splits into the error $\mathcal{T}_1$ of the normalizing constant, the error $\mathcal{T}_2$ of the factor and the remainder $\mathcal{T}_3$ of the expansion along the Gaussian shift,

$$\begin{aligned} &\Big(\tilde\pi_h * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big)\Big)(x) - \pi(x) \\ &= \underbrace{\left(\frac{1}{\pi\big(e^{-(\tilde U_h - U)}\big)} - 1\right)\pi(x)}_{=:\mathcal{T}_1} + \underbrace{\frac{\pi(x)}{\pi\big(e^{-(\tilde U_h - U)}\big)} \left(e^{-(\tilde U_h - U)(x)} \Big(1 + \frac{h}{4}\big(|\nabla \tilde U_h(x)|^2 - \Delta \tilde U_h(x)\big)\Big) - 1\right)}_{=:\mathcal{T}_2} + \underbrace{\mathbb{E}\big[\varrho_4^\chi(x)\big]}_{=:\mathcal{T}_3}. \end{aligned} \tag{B.26}$$

Integrated against $1+|x|$, each term of (B.26) is at most Ch^2: $\mathcal{T}_1$ by (B.17) and $\pi(1+|x|) \leqslant C$, $\mathcal{T}_2$ by the lower bound $\pi\big(e^{-(\tilde U_h - U)}\big) \geqslant \frac12$ and (B.24), and $\mathcal{T}_3$ by (B.25), which gives

$$\int_{\mathbb{R}^d} \big(1+|x|\big) \left| \Big(\tilde\pi_h * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big)\Big)(x) - \pi(x) \right| \mathrm{d}x \leqslant Ch^2. \tag{B.27}$$

Finally, the symmetry of $\mathcal{N}(0, \frac{h}{2} I_d)$ gives

$$\tilde\pi_h(\tilde f_h) = \int_{\mathbb{R}^d} f(x) \Big(\tilde\pi_h * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big)\Big)(x)\,\mathrm{d}x,$$

and since both $\tilde\pi_h * \mathcal{N}(0, \frac{h}{2} I_d)$ and π are probability densities, f may be replaced by $f - f(0)$, with $|f(x) - f(0)| \leqslant \|\nabla f\|_\infty |x|$; (B.27) then gives (4.15). □

C Proofs for Section 5

Statements about a single step are made at the index 0, with the step $D_0 = Y_1 - Y_0$ of (5.2) and $\chi_0 \sim \mathcal{N}(0, I_d)$ independent of $\mathcal{F}_0$, so that under $\mathbb{E}_0$ the state Y_0 is fixed and

$$\mathbb{E}_0[\chi_0] = 0, \qquad \mathbb{E}_0\big[\chi_0^{\otimes 2}\big] = I_d, \qquad \mathbb{E}_0\big[\chi_0^{\otimes 3}\big] = 0, \qquad \mathbb{E}_0|\chi_0|^p \leqslant C, \quad p \geqslant 1.$$

The displacement of the modified dynamics started at Y_0 is written $\tilde{W}_s := \tilde{Y}_0(s) - Y_0$, $0 \leqslant s \leqslant h$. Throughout this appendix, the constant C depends only on m, M, R and d, and also on p where a moment order p appears. The largest moment order used is $p = 12$, for the bounds (5.4) in the proof of Lemma 3; all other proofs in this appendix use $p \leqslant 6$.

C.1 Proof of Lemma 3

The proof uses the bounds (C.1) on one step of the modified dynamics. Lemma 18 states them first, together with the moments of that step, which are used in Appendix C.3.

Lemma 18 *Let Assumption 1 hold and let $\tilde{U}_h$ be the modified potential (4.11). Then, for $0 \leqslant s \leqslant h$ and $p \geqslant 1$, with C depending only on M, d and p and $|||\cdot|||_p$ taken under $\mathbb{E}_0$,*

$$|||\tilde{Y}_0(s) - Y_0|||_p \leqslant C\sqrt{h}\big(1 + |Y_0|\big), \qquad |||\tilde{Y}_0(h) - Y_0 - \sqrt{2h}\,\chi_0|||_p \leqslant Ch\big(1 + |Y_0|\big), \tag{C.1}$$

and, with C depending only on M and d, all derivatives evaluated at Y_0 and $1 \leqslant i, j, k \leqslant d$,

$$\Big|\mathbb{E}_0\big[\tilde{Y}_0(h) - Y_0\big] + h\,\nabla\tilde{U}_h - \frac{h^2}{2}\big(\nabla^2\tilde{U}_h\,\nabla\tilde{U}_h - \nabla\Delta\tilde{U}_h\big)\Big| \leqslant Ch^3\big(1 + |Y_0|\big)^3, \tag{C.2}$$

$$\Big\|\mathbb{E}_0\big[(\tilde{Y}_0(h) - Y_0)^{\otimes 2}\big] - 2h\,I_d - h^2\big(\nabla\tilde{U}_h^{\otimes 2} - 2\,\nabla^2\tilde{U}_h\big)\Big\| \leqslant Ch^3\big(1 + |Y_0|\big)^4, \tag{C.3}$$

$$\Big|\big(\mathbb{E}_0\big[(\tilde{Y}_0(h) - Y_0)^{\otimes 3}\big]\big)_{ijk} + 2h^2\big(\delta_{ij}\,\partial_k\tilde{U}_h + \delta_{ik}\,\partial_j\tilde{U}_h + \delta_{jk}\,\partial_i\tilde{U}_h\big)\Big| \leqslant Ch^3\big(1 + |Y_0|\big)^6. \tag{C.4}$$

The same statements hold with U in place of $\tilde{U}_h$ and X_s in place of $\tilde{Y}_0(s)$, for the exact dynamics (1.1) started at $X_0 = Y_0$, with the same constants.

Proof. By Lemma 1, $|\nabla\tilde{U}_h(0)| \leqslant 2M$ and $\|\nabla^2\tilde{U}_h\| \leqslant 2M$, so that $|\nabla\tilde{U}_h(x)| \leqslant 2M(1 + |x|)$ for every x. The displacement of the modified dynamics (4.16) started at Y_0 is

$$\tilde{W}_s = \tilde{Y}_0(s) - Y_0 = -\int_0^s \nabla\tilde{U}_h\big(Y_0 + \tilde{W}_r\big)\mathrm{d}r + \sqrt{2}\,B_s, \qquad 0 \leqslant s \leqslant h,$$

where $(B_t)_{t\geqslant 0}$ is the Brownian motion of the step, so that $\sqrt{2}\,B_h = \sqrt{2h}\,\chi_0$. Hence

$$|\tilde{W}_s| \leqslant 2M\int_0^s \big(1 + |Y_0| + |\tilde{W}_r|\big)\mathrm{d}r + \sqrt{2}\sup_{r\leqslant h}|B_r|,$$

and Grönwall's inequality gives

$$\sup_{s\leqslant h}|\tilde{W}_s| \leqslant e^{2Mh}\Big(2Mh\big(1 + |Y_0|\big) + \sqrt{2}\sup_{r\leqslant h}|B_r|\Big).$$

Since $\mathbb{E}\big[\sup_{r\leqslant h}|B_r|^p\big] \leqslant Ch^{\frac{p}{2}}$ by the Burkholder–Davis–Gundy inequality and $h \leqslant 1$, the bound on $\tilde{Y}_0(s) - Y_0$ in (C.1) follows. For the bound on $\tilde{Y}_0(h) - Y_0 - \sqrt{2h}\,\chi_0$,

$$\big|\tilde{W}_h - \sqrt{2h}\,\chi_0\big| = \Big|\int_0^h \nabla\tilde{U}_h\big(Y_0 + \tilde{W}_r\big)\mathrm{d}r\Big| \leqslant 2Mh\Big(1 + |Y_0| + \sup_{r\leqslant h}|\tilde{W}_r|\Big),$$

and the Grönwall bound on $\sup_{s\leqslant h}|\tilde{W}_s|$ above applies. This proves the bound on $\tilde{Y}_0(h) - Y_0 - \sqrt{2h}\,\chi_0$ in (C.1). In particular, $\mathbb{E}_0|\tilde{Y}_0(s)|^q \leqslant C(1 + |Y_0|)^q$ for $s \leqslant h$ and every $q \geqslant 1$.

Next we consider the moments of $\tilde W_h = \tilde Y_0(h) - Y_0$ shown in (C.2)–(C.4). Since $\tilde Y_0(h)$ is driven by the modified dynamics (4.16) with the generator $\tilde{\mathcal{L}}_h\psi = \Delta\psi - \nabla\tilde U_h\cdot\nabla\psi$ of (4.17), the estimate of these moments relies on a Taylor expansion with respect to $\tilde{\mathcal{L}}_h$.

Let $\ell(z) := z - Y_0$, a linear function of $z\in\mathbb{R}^d$ with components ℓ_i, so that $\tilde W_h = \ell(\tilde Y_0(h))$ and the entries of $\mathbb{E}_0\big[\tilde W_h^{\otimes n}\big]$, $n=1,2,3$, are $\mathbb{E}_0\big[\psi(\tilde Y_0(h))\big]$ for the polynomials $\psi = \ell_i$, $\ell_i\ell_j$, $\ell_i\ell_j\ell_k$. Here and in the rest of the proof, z denotes the variable of ℓ and of these polynomials.

For each such ψ, since $\tilde U_h\in C^5(\mathbb{R}^d)$ by Lemma 1, the functions $\tilde{\mathcal{L}}_h\psi$ and $\tilde{\mathcal{L}}_h^2\psi$ are of class C^2 and $\tilde{\mathcal{L}}_h^3\psi$ is continuous. By (4.13), $\tilde{\mathcal{L}}_h\psi$, $\tilde{\mathcal{L}}_h^2\psi$, $\tilde{\mathcal{L}}_h^3\psi$ and the gradients of ψ, $\tilde{\mathcal{L}}_h\psi$, $\tilde{\mathcal{L}}_h^2\psi$ have polynomial growth, so that, by $\mathbb{E}_0|\tilde Y_0(s)|^q \leqslant C(1+|Y_0|)^q$, the stochastic integral in Itô's formula is a martingale. Dynkin's formula, applied to ψ, then to $\tilde{\mathcal{L}}_h\psi$ and $\tilde{\mathcal{L}}_h^2\psi$ inside the integrals, gives with Fubini's theorem the Taylor expansion to the order 2 with integral remainder, where $\psi(Y_0)=0$ is omitted since $\ell(Y_0)=0$,

$$\mathbb{E}_0\big[\psi(\tilde Y_0(h))\big] = h\,(\tilde{\mathcal{L}}_h\psi)(Y_0) + \frac{h^2}{2}\,(\tilde{\mathcal{L}}_h^2\psi)(Y_0) + \frac12\int_0^h (h-s)^2\,\mathbb{E}_0\big[(\tilde{\mathcal{L}}_h^3\psi)(\tilde Y_0(s))\big]\mathrm{d}s. \tag{C.5}$$

Here $\tilde{\mathcal{L}}_h\psi(Y_0)$ and $\tilde{\mathcal{L}}_h^2\psi(Y_0)$ are computed with the product rule

$$\tilde{\mathcal{L}}_h(g_1g_2) = g_1\tilde{\mathcal{L}}_hg_2 + g_2\tilde{\mathcal{L}}_hg_1 + 2\,\nabla g_1\cdot\nabla g_2, \qquad g_1,g_2\in C^2(\mathbb{R}^d). \tag{C.6}$$

Next we evaluate the explicit $\tilde{\mathcal{L}}_h$ and $\tilde{\mathcal{L}}_h^2$ operations on $\psi=\ell_i$, $\ell_i\ell_j$, $\ell_i\ell_j\ell_k$, to give the explicit h and h^2 coefficients in the moment estimates (C.2)–(C.4).

1. For $\psi=\ell_i$, the linearity of ℓ_i gives $\tilde{\mathcal{L}}_h\ell_i = -\partial_i\tilde U_h$, and

$$\tilde{\mathcal{L}}_h^2\ell_i = -\tilde{\mathcal{L}}_h\partial_i\tilde U_h = -\partial_i\Delta\tilde U_h + \sum_j \partial_j\tilde U_h\,\partial_i\partial_j\tilde U_h = \big(\nabla^2\tilde U_h\,\nabla\tilde U_h - \nabla\Delta\tilde U_h\big)_i.$$

   At $z=Y_0$, $h\,\tilde{\mathcal{L}}_h\ell_i + \frac{h^2}{2}\,\tilde{\mathcal{L}}_h^2\ell_i$ is the component i of $-h\,\nabla\tilde U_h + \frac{h^2}{2}\big(\nabla^2\tilde U_h\,\nabla\tilde U_h - \nabla\Delta\tilde U_h\big)$, the approximation of $\mathbb{E}_0[\tilde W_h]$ in (C.2).

2. For $\psi=\ell_i\ell_j$, (4.17) gives

$$\tilde{\mathcal{L}}_h(\ell_i\ell_j) = 2\delta_{ij} - \partial_i\tilde U_h\,\ell_j - \partial_j\tilde U_h\,\ell_i,$$

   which equals $2\delta_{ij}$ at $z=Y_0$. Since $\tilde{\mathcal{L}}_h$ annihilates the constant $2\delta_{ij}$, $\tilde{\mathcal{L}}_h^2(\ell_i\ell_j) = -\tilde{\mathcal{L}}_h(\partial_i\tilde U_h\,\ell_j) - \tilde{\mathcal{L}}_h(\partial_j\tilde U_h\,\ell_i)$, and (C.6) with $g_1=\partial_i\tilde U_h$ and $g_2=\ell_j$ gives, at $z=Y_0$, where $\ell_j=0$,

$$\tilde{\mathcal{L}}_h\big(\partial_i\tilde U_h\,\ell_j\big)(Y_0) = \partial_i\tilde U_h\,\tilde{\mathcal{L}}_h\ell_j + 2\,\nabla\partial_i\tilde U_h\cdot\nabla\ell_j = -\partial_i\tilde U_h\,\partial_j\tilde U_h + 2\,\partial_i\partial_j\tilde U_h.$$

   This identity and its copy with i and j exchanged give

$$\tilde{\mathcal{L}}_h^2(\ell_i\ell_j)(Y_0) = 2\,\partial_i\tilde U_h\,\partial_j\tilde U_h - 4\,\partial_i\partial_j\tilde U_h,$$

   so that, at $z=Y_0$, $h\,\tilde{\mathcal{L}}_h(\ell_i\ell_j) + \frac{h^2}{2}\,\tilde{\mathcal{L}}_h^2(\ell_i\ell_j)$ is the entry (i,j) of $2h\,I_d + h^2\big(\nabla\tilde U_h^{\otimes2} - 2\,\nabla^2\tilde U_h\big)$, the approximation of $\mathbb{E}_0[\tilde W_h^{\otimes2}]$ in (C.3).

3. For $\psi=\ell_i\ell_j\ell_k$,

$$\tilde{\mathcal{L}}_h\big(\ell_i\ell_j\ell_k\big) = 2\big(\delta_{ij}\,\ell_k + \delta_{ik}\,\ell_j + \delta_{jk}\,\ell_i\big) - \big(\partial_i\tilde U_h\,\ell_j\ell_k + \partial_j\tilde U_h\,\ell_i\ell_k + \partial_k\tilde U_h\,\ell_i\ell_j\big),$$

   which vanishes at $z=Y_0$. In $\tilde{\mathcal{L}}_h^2(\ell_i\ell_j\ell_k)(Y_0)$, the terms $2\delta_{ij}\,\ell_k$, $2\delta_{ik}\,\ell_j$, $2\delta_{jk}\,\ell_i$ contribute $-2\delta_{ij}\,\partial_k\tilde U_h$, $-2\delta_{ik}\,\partial_j\tilde U_h$, $-2\delta_{jk}\,\partial_i\tilde U_h$, since $\tilde{\mathcal{L}}_h\ell_k = -\partial_k\tilde U_h$. By (C.6) with $g_1=\partial_i\tilde U_h$ and $g_2=\ell_j\ell_k$, whose value and gradient vanish at Y_0, the term $-\partial_i\tilde U_h\,\ell_j\ell_k$ contributes $-\partial_i\tilde U_h\,\tilde{\mathcal{L}}_h(\ell_j\ell_k)(Y_0) = -2\delta_{jk}\,\partial_i\tilde U_h$, and in the same way the terms $-\partial_j\tilde U_h\,\ell_i\ell_k$ and $-\partial_k\tilde U_h\,\ell_i\ell_j$ contribute $-2\delta_{ik}\,\partial_j\tilde U_h$ and $-2\delta_{ij}\,\partial_k\tilde U_h$. Hence

$$\tilde{\mathcal{L}}_h^2\big(\ell_i\ell_j\ell_k\big)(Y_0) = -4\big(\delta_{ij}\,\partial_k\tilde U_h + \delta_{ik}\,\partial_j\tilde U_h + \delta_{jk}\,\partial_i\tilde U_h\big),$$

   so that $\frac{h^2}{2}\,\tilde{\mathcal{L}}_h^2(\ell_i\ell_j\ell_k)(Y_0)$ is the entry (i,j,k) of the approximation of $\mathbb{E}_0[\tilde W_h^{\otimes3}]$ in (C.4).

It remains to bound the integral term of (C.5). Each of ψ, $\tilde{\mathcal{L}}_h\psi$, $\tilde{\mathcal{L}}_h^2\psi$ and $\tilde{\mathcal{L}}_h^3\psi$ is a linear combination of products of components of ℓ, components of $\nabla\tilde{U}_h$ and derivatives of $\tilde{U}_h$ of orders 2 to 5; call the degree of such a product the number of its factors that are components of ℓ or of $\nabla\tilde{U}_h$. By the linear growth of $\nabla\tilde{U}_h$ and (4.13), a product of degree q is bounded by $C(1+|z|+|Y_0|)^q$. In $\tilde{\mathcal{L}}_h$, the Laplacian does not raise the degree, and $-\nabla\tilde{U}_h\cdot\nabla$ raises it by at most 1, and only where it differentiates a derivative of $\tilde{U}_h$ of order at least 2. No such derivative occurs in ψ or $\tilde{\mathcal{L}}_h\psi$, so for ψ a product of n components of ℓ the products in $\tilde{\mathcal{L}}_h\psi$ and $\tilde{\mathcal{L}}_h^2\psi$ have degree at most n, those in $\tilde{\mathcal{L}}_h^3\psi$ have degree at most $n+1$, and

$$\big|(\tilde{\mathcal{L}}_h^3\psi)(z)\big| \leqslant C\big(1+|z|+|Y_0|\big)^{n+1}.$$

With $\mathbb{E}_0(1+|\tilde{Y}_0(s)|+|Y_0|)^{n+1} \leqslant C(1+|Y_0|)^{n+1}$ for $s\leqslant h$, the integral term is bounded by $Ch^3(1+|Y_0|)^{n+1}$, which is at most the remainders stated in (C.2)–(C.4).

For the exact Langevin dynamics (1.1), the same argument applies with U and X_s in place of $\tilde{U}_h$ and $\tilde{Y}_0(s)$, since by Assumption 1 the derivatives of U satisfy (4.13) with M in place of $2M$. □

Proof of Lemma 3. We prove the bounds with the exponents $2q$ and q in place of 12 and 6 for an integer $q\geqslant 1$, the case $q=6$ being (5.4); in this proof C may depend on q as well. The proof is a Lyapunov argument: a bound on the displacement over one step and a drift condition give a contraction of $\mathbb{E}_k|Y_{k+1}|^{2q}$ against $|Y_k|^{2q}$, which is iterated over k. By (LM'), the step of the chain at the index k is

$$D_k := Y_{k+1}-Y_k = \sqrt{2h}\,\chi_k - h\,\nabla U(Z_k), \qquad Z_k = Y_k + \sqrt{\frac{h}{2}}\,\chi_k,$$

and Assumption 1 gives $|\nabla U(Z_k)|\leqslant M(1+|Y_k|+\sqrt{h/2}\,|\chi_k|)$.

We first bound the displacement over one step. Since $\mathbb{E}_k|\sqrt{2h}\,\chi_k|^p \leqslant Ch^{\frac{p}{2}}$, the bound on $|\nabla U(Z_k)|$ gives, for every $p\geqslant 1$,

$$\mathbb{E}_k|D_k|^p \leqslant C\big(\mathbb{E}_k|\sqrt{2h}\,\chi_k|^p + h^p\,\mathbb{E}_k|\nabla U(Z_k)|^p\big) \leqslant Ch^{\frac{p}{2}} + Ch^p\big(1+|Y_k|\big)^p, \tag{C.7}$$

and, since (C.1) at the index k gives $\big\|\tilde{Y}_k(h)-Y_k-\sqrt{2h}\,\chi_k\big\|_p \leqslant Ch(1+|Y_k|)$ under $\mathbb{E}_k$,

$$\mathbb{E}_k|\tilde{Y}_k(h)-Y_k|^p \leqslant Ch^{\frac{p}{2}} + Ch^p\big(1+|Y_k|\big)^p. \tag{C.8}$$

Next we derive the drift condition. Splitting $\langle Y_k,\nabla U(Z_k)\rangle = \langle Z_k,\nabla U(Z_k)\rangle - \sqrt{h/2}\,\langle\chi_k,\nabla U(Z_k)\rangle$ and applying the drift bound in (B.1) at Z_k and the bound on $|\nabla U(Z_k)|$,

$$\langle Y_k,\nabla U(Z_k)\rangle \geqslant \frac{m}{4}|Z_k|^2 - C - M\sqrt{\frac{h}{2}}\,|\chi_k|\Big(1+|Y_k|+\sqrt{\frac{h}{2}}\,|\chi_k|\Big).$$

Taking $\mathbb{E}_k$, with $\mathbb{E}_k|Z_k|^2 = |Y_k|^2+\frac{dh}{2}$ and $\mathbb{E}_k|\chi_k|^2=d$,

$$\langle Y_k,\mathbb{E}_k[\nabla U(Z_k)]\rangle \geqslant \frac{m}{4}|Y_k|^2 - C - C\sqrt{h}\big(1+|Y_k|\big) \geqslant \frac{m}{8}|Y_k|^2 - C, \tag{C.9}$$

since $C\sqrt{h}\,|Y_k| \leqslant \frac{m}{8}|Y_k|^2 + C$ by Young's inequality and $h\leqslant 1$.

Next we combine (C.7) and (C.9) into the Lyapunov estimate. Let $S_k := |Y_{k+1}|^2-|Y_k|^2 = 2\langle Y_k,D_k\rangle + |D_k|^2$. Since χ_k is centered and independent of $\mathcal{F}_k$, (C.9) and (C.7) with $p=2$ give

$$\mathbb{E}_kS_k = -2h\langle Y_k,\mathbb{E}_k[\nabla U(Z_k)]\rangle + \mathbb{E}_k|D_k|^2 \leqslant -\frac{m}{4}h|Y_k|^2 + Ch + Ch^2\big(1+|Y_k|\big)^2 \leqslant -\frac{m}{8}h|Y_k|^2 + Ch$$

for $h\leqslant h_0$ with $Ch_0\leqslant\frac{m}{8}$, while $|S_k|\leqslant 2|Y_k||D_k|+|D_k|^2$ and (C.7) give, for $1\leqslant j\leqslant q$,

$$\mathbb{E}_k|S_k|^j \leqslant C\big(|Y_k|^j\,\mathbb{E}_k|D_k|^j + \mathbb{E}_k|D_k|^{2j}\big) \leqslant Ch^{\frac{j}{2}}|Y_k|^j + Ch^j\big(1+|Y_k|\big)^{2j}.$$

Expanding

$$|Y_{k+1}|^{2q} = \big(|Y_k|^2+S_k\big)^q = \sum_{j=0}^{q}\binom{q}{j}|Y_k|^{2(q-j)}S_k^j$$

and inserting the bounds on $\mathbb{E}_k S_k$ and $\mathbb{E}_k|S_k|^j$, with $h \leqslant 1$,

$$q\,|Y_k|^{2q-2}\,\mathbb{E}_k S_k \leqslant -\frac{qm}{8}h|Y_k|^{2q} + Ch|Y_k|^{2q-2}, \qquad |Y_k|^{2(q-j)}\,\mathbb{E}_k|S_k|^j \leqslant Ch|Y_k|^{2q-j} + Ch^2(1+|Y_k|)^{2q}, \quad j \geqslant 2.$$

By Young's inequality, the terms of degree below $2q$ in $|Y_k|$ satisfy

$$Ch|Y_k|^{2q-2} + Ch\sum_{j=2}^{q}|Y_k|^{2q-j} \leqslant \frac{qm}{32}\,h|Y_k|^{2q} + Ch,$$

and, for $h \leqslant h_0$ with Ch_0 small enough, the terms of order h^2 satisfy

$$Ch^2\big(1+|Y_k|\big)^{2q} \leqslant \frac{qm}{32}\,h|Y_k|^{2q} + Ch.$$

With these bounds on the terms of degree below $2q$ and of order h^2, the expansion of $\mathbb{E}_k|Y_{k+1}|^{2q}$ gives

$$\mathbb{E}_k|Y_{k+1}|^{2q} \leqslant |Y_k|^{2q} - \frac{qm}{8}\,h|Y_k|^{2q} + 2\cdot\frac{qm}{32}\,h|Y_k|^{2q} + Ch,$$

that is,

$$\mathbb{E}_k|Y_{k+1}|^{2q} \leqslant \Big(1 - \frac{qm}{16}h\Big)|Y_k|^{2q} + Ch, \tag{C.10}$$

with $h_0 \leqslant 16/(qm)$ making the factor lie in $[0,1)$. Taking expectations in (C.10) and iterating over k gives the bound on $\mathbb{E}|Y_k|^{2q}$ in (5.4), and averaging (C.7) and (C.8) with $p = 2q$ over the law of Y_k gives the displacement bounds in (5.4):

$$\mathbb{E}|Y_k|^{2q} \leqslant |Y_0|^{2q} + \frac{16C}{qm}, \qquad \max\big\{\mathbb{E}|D_k|^{2q},\ \mathbb{E}|\tilde{Y}_k(h) - Y_k|^{2q}\big\} \leqslant Ch^q + Ch^{2q}\,\mathbb{E}\big(1+|Y_k|\big)^{2q} \leqslant Ch^q\big(1+|Y_0|\big)^{2q}.$$

The argument carries over to the stochastic gradient chain (SGLM'), with $\hat{\mathbb{E}}_k$ of (3.2) in place of $\mathbb{E}_k$, once the displacement bounds and the drift condition are verified for it. The step (5.3) of this chain, at the index k, is

$$D_k = Y_{k+1} - Y_k = \sqrt{2h}\,\chi_k - h\,\nabla\hat{U}(Z_k;\xi_k).$$

By the moment bound of Assumption 3 and Jensen's inequality, for $1 \leqslant p \leqslant 12$,

$$\hat{\mathbb{E}}_k|\nabla\hat{U}(Z_k;\xi_k)|^p \leqslant M^p\,\hat{\mathbb{E}}_k\big(1+|Z_k|\big)^p \leqslant C\big(1+|Y_k|\big)^p,$$

which is the bound used for $\nabla U(Z_k)$ in (C.7), with the same constant, so that

$$\hat{\mathbb{E}}_k|D_k|^p \leqslant Ch^{\frac{p}{2}} + Ch^p\big(1+|Y_k|\big)^p, \qquad 1 \leqslant p \leqslant 12.$$

The bound (C.8) holds with $\hat{\mathbb{E}}_k$ in place of $\mathbb{E}_k$, since $\tilde{Y}_k(h) - Y_k$ is driven by $(B_{kh+t} - B_{kh})_{0\leqslant t\leqslant h}$, which is independent of $\hat{\mathcal{F}}_k$. Since the estimator is unbiased and ξ_k is independent of $\hat{\mathcal{F}}_k$ and χ_k, the derivation of (C.9) with $\hat{\mathbb{E}}_k$ in place of $\mathbb{E}_k$ gives

$$\big\langle Y_k, \hat{\mathbb{E}}_k[\nabla\hat{U}(Z_k;\xi_k)]\big\rangle = \big\langle Y_k, \hat{\mathbb{E}}_k[\nabla U(Z_k)]\big\rangle \geqslant \frac{m}{8}|Y_k|^2 - C.$$

The Lyapunov estimate uses only $\hat{\mathbb{E}}_k[\chi_k] = 0$, the drift condition and the bound on $\hat{\mathbb{E}}_k|D_k|^p$ with $p \leqslant 2q$, so for $q \leqslant 6$ it gives (C.10) with $\hat{\mathbb{E}}_k$ in place of $\mathbb{E}_k$ and the same constants. Averaging the bound on $\hat{\mathbb{E}}_k|D_k|^p$ and (C.8) under $\hat{\mathbb{E}}_k$, with $p = 2q \leqslant 12$, over the law of Y_k then gives (5.4) for (SGLM'). □

C.2 Proof of Lemma 4

Proof of Lemma 4. We prove the bounds in $\|\!|\cdot|\!\|_p$ for every $p \geqslant 1$, with C depending also on p; the case $p = 4$ is (5.5). The bounds on $Y_1 - Y_0$ and on $\tilde{Y}_0(s) - Y_0$ in (5.5) are direct. For the step $D_0 = Y_1 - Y_0 = \sqrt{2h}\,\chi_0 - h\,\nabla U(Z_0)$ of (5.2), the moments of χ_0 and Assumption 1 give

$$\|\!|\sqrt{2h}\,\chi_0|\!\|_p \leqslant C\sqrt{h}, \qquad \|\!|h\,\nabla U(Z_0)|\!\|_p \leqslant Mh\Big(1 + |Y_0| + \sqrt{\frac{h}{2}}\,\|\!|\chi_0|\!\|_p\Big) \leqslant Ch(1 + |Y_0|),$$

and the bound on $\tilde{Y}_0(s) - Y_0$ is that in (C.1).

Next we bound the local error $Y_1 - \tilde{Y}_0(h) = D_0 - \tilde{W}_h$. Subtracting the displacement $\tilde{W}_h = \sqrt{2h}\,\chi_0 - \int_0^h \nabla\tilde{U}_h(Y_0 + \tilde{W}_r)\,\mathrm{d}r$ of (4.16) from D_0 gives

$$Y_1 - \tilde{Y}_0(h) = \underbrace{\int_0^h \Big(\nabla U\big(Y_0 + \tilde{W}_r\big) - \nabla U(Z_0)\Big)\mathrm{d}r}_{=:\mathcal{T}_1} + \underbrace{\int_0^h \Big(\nabla\tilde{U}_h\big(Y_0 + \tilde{W}_r\big) - \nabla U\big(Y_0 + \tilde{W}_r\big)\Big)\mathrm{d}r}_{=:\mathcal{T}_2}. \tag{C.11}$$

The integrands of $\mathcal{T}_1$ and $\mathcal{T}_2$ satisfy

$$\big|\nabla U(Y_0 + \tilde{W}_r) - \nabla U(Z_0)\big| \leqslant M\Big|\tilde{W}_r - \sqrt{\frac{h}{2}}\,\chi_0\Big|, \qquad \big|(\nabla\tilde{U}_h - \nabla U)(Y_0 + \tilde{W}_r)\big| \leqslant Ch\big(1 + |Y_0| + |\tilde{W}_r|\big),$$

by $\|\nabla^2 U\| \leqslant M$ of Assumption 1 with $Z_0 = Y_0 + \sqrt{h/2}\,\chi_0$, and by (4.14), respectively. Minkowski's inequality for the integrals over $r \in [0, h]$ and the bound $\|\!|\tilde{W}_r|\!\|_p \leqslant C\sqrt{h}\,(1 + |Y_0|)$ of (C.1) then give

$$\begin{aligned}
\|\!|\mathcal{T}_1|\!\|_p &\leqslant M\int_0^h \Big(\|\!|\tilde{W}_r|\!\|_p + \sqrt{\frac{h}{2}}\,\|\!|\chi_0|\!\|_p\Big)\mathrm{d}r \leqslant Ch^{\frac{3}{2}}(1 + |Y_0|),\\
\|\!|\mathcal{T}_2|\!\|_p &\leqslant Ch\int_0^h \big(1 + |Y_0| + \|\!|\tilde{W}_r|\!\|_p\big)\mathrm{d}r \leqslant Ch^2(1 + |Y_0|).
\end{aligned}$$

Adding the bounds on $\|\!|\mathcal{T}_1|\!\|_p$ and $\|\!|\mathcal{T}_2|\!\|_p$ gives $\|\!|Y_1 - \tilde{Y}_0(h)|\!\|_p \leqslant Ch^{\frac{3}{2}}(1 + |Y_0|)$, as stated in (5.5) for $p = 4$. □

C.3 Proof of Lemma 5

The proof compares the moments of $\tilde{W}_h$ in Lemma 18 with the following expansion of the moments of the step $Y_1 - Y_0$ of the chain.

Lemma 19 *Let Assumption 1 hold and let $h \in (0, 1]$. Then, with all derivatives of U evaluated at Y_0 and C depending only on M and d, the moments of the step (5.2) of the chain satisfy, for $1 \leqslant i, j, k \leqslant d$,*

$$\Big|\mathbb{E}_0[Y_1 - Y_0] + h\,\nabla U + \frac{h^2}{4}\,\nabla\Delta U\Big| \leqslant Ch^3, \tag{C.12}$$

$$\Big\|\mathbb{E}_0\big[(Y_1 - Y_0)^{\otimes 2}\big] - 2h\,I_d - h^2\big(\nabla U^{\otimes 2} - 2\,\nabla^2 U\big)\Big\| \leqslant Ch^3\big(1 + |Y_0|\big), \tag{C.13}$$

$$\Big|\big(\mathbb{E}_0\big[(Y_1 - Y_0)^{\otimes 3}\big]\big)_{ijk} + 2h^2\big(\delta_{ij}\,\partial_k U + \delta_{ik}\,\partial_j U + \delta_{jk}\,\partial_i U\big)\Big| \leqslant Ch^3\big(1 + |Y_0|\big)^3. \tag{C.14}$$

Proof. The step (5.2) reads $Y_1 - Y_0 = \sqrt{2h}\,\chi_0 - h\,\nabla U(Z_0)$ with $Z_0 - Y_0 = \sqrt{h/2}\,\chi_0$, so its moments follow from expansions of $\nabla U(Z_0)$ about Y_0 in powers of χ_0. In this proof, $\mathcal{O}(h^q)$ denotes a tensor of norm at most Ch^q. Taylor's formula to the order 3, with the remainder controlled by $\|\nabla^5 U\| \leqslant M$, gives

$$\nabla U(Z_0) = \nabla U + \sqrt{\frac{h}{2}}\,\nabla^2 U\,\chi_0 + \frac{h}{4}\,\nabla^3 U\big[\chi_0^{\otimes 2}\big] + \frac{1}{6}\Big(\frac{h}{2}\Big)^{\frac{3}{2}}\nabla^4 U\big[\chi_0^{\otimes 3}\big] + \varrho_4, \qquad |\varrho_4| \leqslant \frac{M}{24}\Big(\frac{h}{2}\Big)^2|\chi_0|^4, \tag{C.15}$$

and Taylor's formula to the order 1, with the remainder controlled by $\|\nabla^3 U\| \leqslant M$, gives

$$\nabla U(Z_0) = \nabla U + \sqrt{\frac{h}{2}}\,\nabla^2 U\,\chi_0 + \varrho_2, \qquad |\varrho_2| \leqslant \frac{M}{2}\,\frac{h}{2}\,|\chi_0|^2. \tag{C.16}$$

The expansion (C.15) is used where $\nabla U(Z_0)$ carries a factor h or $h^{\frac{3}{2}}$, namely in the mean and in the cross term of the moment of order 2, and (C.16) where it carries a factor h^2 or $h^{\frac{5}{2}}$. In addition, $|\nabla U(x)| \leqslant M(1+|x|)$ gives $|\nabla U(Z_0)| \leqslant M(1+|Y_0| + \sqrt{h/2}\,|\chi_0|)$, and under $\mathbb{E}_0$ every term of odd degree in χ_0 in (C.15) and (C.16), and in their products with powers of χ_0, has mean zero.

Next we compute the moments of orders 1, 2, 3 of $Y_1 - Y_0$ from (C.15) and (C.16), to give the explicit h and h^2 coefficients in (C.12)–(C.14).

1. For the mean, the terms of degrees 1 and 3 in χ_0 of (C.15) have mean zero, and $\nabla^3 U[\mathbb{E}_0 \chi_0^{\otimes 2}] = \nabla^3 U[I_d] = \nabla \Delta U$ by the contraction convention of Section 3.1, so that

$$\mathbb{E}_0\big[\nabla U(Z_0)\big] = \nabla U + \frac{h}{4}\,\nabla \Delta U + \mathbb{E}_0[\varrho_4], \qquad |\mathbb{E}_0[\varrho_4]| \leqslant C h^2.$$

   Multiplied by $-h$, this is (C.12), since $\mathbb{E}_0[Y_1 - Y_0] = -h\,\mathbb{E}_0[\nabla U(Z_0)]$.

2. For the moment of order 2,

$$(Y_1 - Y_0)^{\otimes 2} = 2h\,\chi_0^{\otimes 2} - 2h\sqrt{2h}\,\mathrm{Sym}\big(\chi_0 \otimes \nabla U(Z_0)\big) + h^2\,\nabla U(Z_0)^{\otimes 2}.$$

   The term $2h\,\chi_0^{\otimes 2}$ has conditional mean $2hI_d$. For the cross term $-2h\sqrt{2h}\,\mathrm{Sym}(\chi_0 \otimes \nabla U(Z_0))$, insert (C.15) into $\chi_0 \otimes \nabla U(Z_0)$. The terms $\chi_0 \otimes \nabla U$ and $\chi_0 \otimes \nabla^3 U[\chi_0^{\otimes 2}]$ have odd degree and mean zero, $\mathbb{E}_0[\chi_0 \otimes \nabla^2 U\,\chi_0] = \nabla^2 U$ since $\mathbb{E}_0[\chi_0^{\otimes 2}] = I_d$ and $\nabla^2 U$ is symmetric, and the terms $\frac{1}{6}(\frac{h}{2})^{\frac{3}{2}}\,\chi_0 \otimes \nabla^4 U[\chi_0^{\otimes 3}]$ and $\chi_0 \otimes \varrho_4$ have conditional means of norm at most $Ch^{\frac{3}{2}}$. Hence

$$\mathbb{E}_0\big[\chi_0 \otimes \nabla U(Z_0)\big] = \sqrt{\frac{h}{2}}\,\nabla^2 U + \mathcal{O}\big(h^{\frac{3}{2}}\big), \qquad -2h\sqrt{2h}\,\mathrm{Sym}\,\mathbb{E}_0\big[\chi_0 \otimes \nabla U(Z_0)\big] = -2h^2\,\nabla^2 U + \mathcal{O}(h^3).$$

   For the term $h^2\,\nabla U(Z_0)^{\otimes 2}$, (C.16) gives

$$\begin{aligned}\nabla U(Z_0)^{\otimes 2} = \nabla U^{\otimes 2} &+ 2\,\mathrm{Sym}\Big(\nabla U \otimes \sqrt{\frac{h}{2}}\,\nabla^2 U\,\chi_0\Big) + 2\,\mathrm{Sym}\big(\nabla U \otimes \varrho_2\big) \\ &+ \frac{h}{2}\big(\nabla^2 U\,\chi_0\big)^{\otimes 2} + 2\,\mathrm{Sym}\Big(\sqrt{\frac{h}{2}}\,\nabla^2 U\,\chi_0 \otimes \varrho_2\Big) + \varrho_2^{\otimes 2}.\end{aligned} \tag{C.17}$$

   The term $2\,\mathrm{Sym}(\nabla U \otimes \sqrt{h/2}\,\nabla^2 U\,\chi_0)$ has mean zero. By $|\nabla U| \leqslant M(1+|Y_0|)$ and the bound on ϱ_2 in (C.16), the conditional means of the remaining 4 terms of (C.17) satisfy

$$\begin{aligned}&\Big\|\mathbb{E}_0\big[2\,\mathrm{Sym}(\nabla U \otimes \varrho_2)\big]\Big\| \leqslant Ch(1+|Y_0|), \qquad \Big\|\mathbb{E}_0\Big[\frac{h}{2}(\nabla^2 U\,\chi_0)^{\otimes 2}\Big]\Big\| \leqslant Ch, \\ &\Big\|\mathbb{E}_0\big[2\,\mathrm{Sym}\big(\sqrt{h/2}\,\nabla^2 U\,\chi_0 \otimes \varrho_2\big)\big]\Big\| \leqslant Ch^{\frac{3}{2}}, \qquad \Big\|\mathbb{E}_0\big[\varrho_2^{\otimes 2}\big]\Big\| \leqslant Ch^2,\end{aligned}$$

   so that

$$\Big\|h^2\,\mathbb{E}_0\big[\nabla U(Z_0)^{\otimes 2}\big] - h^2\,\nabla U^{\otimes 2}\Big\| \leqslant Ch^3(1+|Y_0|).$$

   Adding the conditional means of the 3 terms of $(Y_1 - Y_0)^{\otimes 2}$, namely $2hI_d$, $-2h^2\,\nabla^2 U + \mathcal{O}(h^3)$ and $h^2\,\nabla U^{\otimes 2}$ up to $Ch^3(1+|Y_0|)$, gives (C.13).

3. For the moment of order 3,

$$(Y_1 - Y_0)^{\otimes 3} = (2h)^{\frac{3}{2}}\,\chi_0^{\otimes 3} - 6h^2\,\mathrm{Sym}\big(\chi_0^{\otimes 2} \otimes \nabla U(Z_0)\big) + 3h^2\sqrt{2h}\,\mathrm{Sym}\big(\chi_0 \otimes \nabla U(Z_0)^{\otimes 2}\big) - h^3\,\nabla U(Z_0)^{\otimes 3}.$$

The term $(2h)^{\frac{3}{2}}\,\chi_0^{\otimes 3}$ has mean zero. For the term $-6h^2\,\mathrm{Sym}(\chi_0^{\otimes 2}\otimes\nabla U(Z_0))$, insert (C.16) into the entry $\mathbb{E}_0[\chi_{0,i}\,\chi_{0,j}\,\nabla U(Z_0)_k]$: the term ∇U gives $\delta_{ij}\,\partial_k U$, the term linear in χ_0 has mean zero, and the remainder ϱ_2 contributes at most Ch, so that

$$\mathbb{E}_0\big[\chi_{0,i}\,\chi_{0,j}\,\nabla U(Z_0)_k\big] = \delta_{ij}\,\partial_k U + \mathcal{O}(h),$$

and, since $\delta_{ij}\,\partial_k U$ is symmetric in i and j, its symmetrization is $\frac{1}{3}(\delta_{ij}\,\partial_k U+\delta_{ik}\,\partial_j U+\delta_{jk}\,\partial_i U)$, so that

$$-6h^2\,\mathrm{Sym}\,\mathbb{E}_0\big[\chi_0^{\otimes 2}\otimes\nabla U(Z_0)\big]_{ijk} = -2h^2\big(\delta_{ij}\,\partial_k U+\delta_{ik}\,\partial_j U+\delta_{jk}\,\partial_i U\big)+\mathcal{O}(h^3).$$

For the term $3h^2\sqrt{2h}\,\mathrm{Sym}(\chi_0\otimes\nabla U(Z_0)^{\otimes 2})$, insert (C.17) into $\chi_0\otimes\nabla U(Z_0)^{\otimes 2}$: the term $\chi_0\otimes\nabla U^{\otimes 2}$ has mean zero, and every other term carries a factor $\sqrt{h/2}\,\nabla^2 U\,\chi_0$ or ϱ_2, so its conditional mean has norm at most $C\sqrt{h}(1+|Y_0|)$. For the term $h^3\,\nabla U(Z_0)^{\otimes 3}$, the bound on $|\nabla U(Z_0)|$ suffices. Hence

$$\begin{aligned}
&\Big\|3h^2\sqrt{2h}\,\mathrm{Sym}\,\mathbb{E}_0\big[\chi_0\otimes\nabla U(Z_0)^{\otimes 2}\big]\Big\| \leqslant Ch^3(1+|Y_0|),\\
&\Big\|h^3\,\mathbb{E}_0\big[\nabla U(Z_0)^{\otimes 3}\big]\Big\| \leqslant h^3\,\mathbb{E}_0|\nabla U(Z_0)|^3 \leqslant Ch^3(1+|Y_0|)^3.
\end{aligned}$$

Adding the conditional means of the 4 terms of $(Y_1-Y_0)^{\otimes 3}$ gives (C.14).

□

Proof of Lemma 5. The moments of $\tilde W_h=\tilde Y_0(h)-Y_0$ are expanded in (C.2)–(C.4) with coefficients in $\tilde U_h$, and those of $D_0=Y_1-Y_0$ in (C.12)–(C.14) with coefficients in U; Lemma 19 applies since $h\leqslant h_0\leqslant 1$, and all derivatives are evaluated at Y_0. Next we compare the two expansions. Their coefficients differ through $\tilde U_h-U$, and by (4.14) and the linear growth of ∇U, and of $\nabla\tilde U_h$ from Lemma 1,

$$|\nabla\tilde U_h-\nabla U|\leqslant Ch\big(1+|Y_0|\big),\quad \|\nabla\tilde U_h^{\otimes 2}-\nabla U^{\otimes 2}\|\leqslant Ch\big(1+|Y_0|\big)^2,\quad \max_{2\leqslant i\leqslant 3}\|\nabla^i\tilde U_h-\nabla^i U\|\leqslant Ch\big(1+|Y_0|\big). \tag{C.18}$$

Next we prove (5.6) for $j=1,2,3$ with (C.18), first for $j=2,3$ and then for $j=1$.

1. For $j=2$ and $j=3$, replacing $\tilde U_h$ by U in the terms of order h^2 of (C.3) and (C.4) changes them in norm by at most $Ch^3(1+|Y_0|)^2$, by (C.18). After this replacement, the approximations of $\mathbb{E}_0[\tilde W_h^{\otimes j}]$ in (C.3) and (C.4) are those of $\mathbb{E}_0[D_0^{\otimes j}]$ in (C.13) and (C.14), which proves (5.6) for $j=2$ and $j=3$.

2. For $j=1$, the term $-h\,\nabla\tilde U_h$ of the approximation of $\mathbb{E}_0[\tilde W_h]$ in (C.2) is kept, since (C.18) bounds $h(\nabla\tilde U_h-\nabla U)$ only by $Ch^2(1+|Y_0|)$. In the term of order h^2, the replacement of $\tilde U_h$ by U introduces an additional error of at most $Ch^3(1+|Y_0|)^2$ in norm, by (C.18), the linear growth of $\nabla\tilde U_h$ and $\|\nabla^2 U\|\leqslant M$, the bound on $\nabla^3\tilde U_h-\nabla^3 U$ in (C.18) controlling $\nabla\Delta\tilde U_h-\nabla\Delta U$. Subtracting (C.12) then gives

$$\Big|\mathbb{E}_0[\tilde W_h]-\mathbb{E}_0[D_0]+h\big(\nabla\tilde U_h-\nabla U\big)-\frac{h^2}{2}\big(\nabla^2 U\,\nabla U-\nabla\Delta U\big)-\frac{h^2}{4}\,\nabla\Delta U\Big|\leqslant Ch^3\big(1+|Y_0|\big)^3. \tag{C.19}$$

It remains to bound the terms of (C.19) other than $\mathbb{E}_0[\tilde W_h]-\mathbb{E}_0[D_0]$ by $Ch^3(1+|Y_0|)^2$, and here the cutoff θ_h of (4.11) enters. On the ball $\{|Y_0|\leqslant\delta/h\}$, where $\theta_h=1$ and $\nabla\theta_h=0$, $\nabla\tilde U_h-\nabla U=\frac{h}{4}\nabla V=\frac{h}{4}(2\,\nabla^2 U\,\nabla U-\nabla\Delta U)$ by (4.10), so that

$$\begin{aligned}
&h\big(\nabla\tilde U_h-\nabla U\big)-\frac{h^2}{2}\big(\nabla^2 U\,\nabla U-\nabla\Delta U\big)-\frac{h^2}{4}\,\nabla\Delta U\\
&=\frac{h^2}{4}\big(2\,\nabla^2 U\,\nabla U-\nabla\Delta U\big)-\frac{h^2}{2}\,\nabla^2 U\,\nabla U+\frac{h^2}{4}\,\nabla\Delta U=0.
\end{aligned} \tag{C.20}$$

Outside that ball, (C.18) and Assumption 1 give

$$h\,\big|\nabla\tilde U_h(Y_0)-\nabla U(Y_0)\big|\leqslant Ch^2(1+|Y_0|),\qquad \Big|\frac{h^2}{2}\big(\nabla^2 U\,\nabla U-\nabla\Delta U\big)+\frac{h^2}{4}\,\nabla\Delta U\Big|\leqslant Ch^2(1+|Y_0|),$$

and $1 \leqslant h|Y_0|/\delta$ there, so the left side of (C.20) is at most

$$Ch^2(1+|Y_0|)\cdot\frac{h|Y_0|}{\delta} \leqslant Ch^3(1+|Y_0|)^2$$

in norm. On the ball and outside it, (C.19) therefore gives

$$\big|\mathbb{E}_0[\tilde{W}_h] - \mathbb{E}_0[D_0]\big| \leqslant Ch^3(1+|Y_0|)^3 \leqslant Ch^3(1+|Y_0|)^6,$$

which proves (5.6) for $j=1$.

□

C.4 Proof of Lemma 6

Proof of Lemma 6. The proof expands $G(Y_1)-G(\tilde{Y}_0(h))$ about Y_0 to the order 3, so that the terms of orders 1 to 3 are differences of the moments of D_0 and $\tilde{W}_h$, bounded by Lemma 5, and the remainder is bounded by Lemma 4.

Work at the index 0, with D_0 and $\tilde{W}_h$ as in the opening of Appendix C. Write $Q_\varsigma := \varsigma Y_1 + (1-\varsigma)\tilde{Y}_0(h)$ for $\varsigma \in [0,1]$, the interpolated position of [106], so that $\frac{\mathrm{d}}{\mathrm{d}\varsigma}Q_\varsigma = Y_1 - \tilde{Y}_0(h)$, the local error, and

$$G(Y_1) - G(\tilde{Y}_0(h)) = \int_0^1 \big\langle \nabla G(Q_\varsigma),\, Y_1 - \tilde{Y}_0(h)\big\rangle \,\mathrm{d}\varsigma. \tag{C.21}$$

Expand ∇G about Y_0 to the order 2 with the integral remainder,

$$\nabla G(Y_0+v) = \nabla G(Y_0) + \nabla^2 G(Y_0)\,v + \frac{1}{2}\nabla^3 G(Y_0)\big[v^{\otimes 2}\big] + \varrho_3(v), \qquad |\varrho_3(v)| \leqslant \frac{1}{6}\,\|\nabla^4 G\|_\infty\,|v|^3,$$

and insert it into (C.21) with $v = Q_\varsigma - Y_0$. Since $\nabla^2 G(Y_0)$ and $\nabla^3 G(Y_0)$ are symmetric, $\nabla^j G(Y_0) : T = \nabla^j G(Y_0) : \operatorname{Sym} T$ for every tensor T of order $j = 2, 3$. Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}\varsigma}\big(Q_\varsigma - Y_0\big)^{\otimes j} = j\,\operatorname{Sym}\big((Q_\varsigma - Y_0)^{\otimes(j-1)} \otimes (Y_1 - \tilde{Y}_0(h))\big), \qquad j = 2,3,$$

and integrating over $\varsigma \in [0,1]$, with $Q_0 - Y_0 = \tilde{W}_h$ and $Q_1 - Y_0 = D_0$,

$$\operatorname{Sym}\bigg(\int_0^1 (Q_\varsigma - Y_0)\,\mathrm{d}\varsigma \otimes \big(Y_1 - \tilde{Y}_0(h)\big)\bigg) = \frac{1}{2}\Big(D_0^{\otimes 2} - \tilde{W}_h^{\otimes 2}\Big),$$
$$\operatorname{Sym}\bigg(\int_0^1 (Q_\varsigma - Y_0)^{\otimes 2}\,\mathrm{d}\varsigma \otimes \big(Y_1 - \tilde{Y}_0(h)\big)\bigg) = \frac{1}{3}\Big(D_0^{\otimes 3} - \tilde{W}_h^{\otimes 3}\Big).$$

These identities turn (C.21) into

$$G(Y_1) - G(\tilde{Y}_0(h)) = \sum_{j=1}^{3} \frac{1}{j!}\,\nabla^j G(Y_0) : \Big(D_0^{\otimes j} - \tilde{W}_h^{\otimes j}\Big) + \varrho_4, \tag{C.22}$$

where the remainder $\varrho_4 := \int_0^1 \big\langle \varrho_3(Q_\varsigma - Y_0),\, Y_1 - \tilde{Y}_0(h)\big\rangle\,\mathrm{d}\varsigma$ satisfies

$$|\varrho_4| \leqslant \frac{1}{6}\,\|\nabla^4 G\|_\infty\,\big|Y_1 - \tilde{Y}_0(h)\big|\,\max\big\{|D_0|,\,|\tilde{W}_h|\big\}^3.$$

In (C.22), $G(Y_1)$ and $G(\tilde{Y}_0(h))$ are both expanded about the initial value Y_0, which is why Lemma 5 estimates the moments of D_0 and $\tilde{W}_h$ up to the order 3. Taking $\mathbb{E}_0$, (5.6) bounds the terms of orders 1 to 3 in (C.22):

$$\bigg|\sum_{j=1}^{3} \frac{1}{j!}\,\nabla^j G(Y_0) : \mathbb{E}_0\big[D_0^{\otimes j} - \tilde{W}_h^{\otimes j}\big]\bigg| \leqslant Ch^3 \sum_{j=1}^{3} \|\nabla^j G\|_\infty \big(1+|Y_0|\big)^6.$$

For the remainder ϱ_4, (5.5) gives

$$\|\!|Y_1 - \tilde{Y}_0(h)|\!\|_4 \leqslant Ch^{\frac{3}{2}}(1+|Y_0|), \qquad \|\!|D_0|\!\|_4 + \|\!|\tilde{W}_h|\!\|_4 \leqslant C\sqrt{h}(1+|Y_0|),$$

and Hölder's inequality gives

$$\mathbb{E}_0|\varrho_4| \leqslant \frac{1}{6}\|\nabla^4 G\|_\infty \, \|\!|Y_1 - \tilde{Y}_0(h)|\!\|_4 \big(\|\!|D_0|\!\|_4 + \|\!|\tilde{W}_h|\!\|_4\big)^3 \leqslant C\|\nabla^4 G\|_\infty \, h^3\big(1+|Y_0|\big)^4.$$

Adding the bounds on the terms of orders 1 to 3 and on ϱ_4 proves (5.7) at the index 0, and by time homogeneity at every k. □

C.5 Proof of Lemma 7

The proof uses two bounds on the covariance of the gradient noise, which also serve Appendices E to G. We write the gradient noise and its covariance (5.8) as

$$\eta(x;\xi) := \nabla\hat{U}(x;\xi) - \nabla U(x), \qquad \mathcal{C}(x) = \mathbb{E}_\xi\big[\eta(x;\xi)\otimes\eta(x;\xi)\big].$$

Let

$$S(x;\xi) := \sup_{|x'-x|\leqslant 1} \|\nabla^2\hat{U}(x';\xi)\|, \qquad \mathbb{E}_\xi\big[S(x;\xi)^2\big] \leqslant M^2, \tag{C.23}$$

the bound being the Hessian condition of Assumption 3. Let $|x-x'| \leqslant 1$. Then

$$\eta(x;\xi) - \eta(x';\xi) = \int_0^1 \big(\nabla^2\hat{U}(x_s;\xi) - \nabla^2 U(x_s)\big)(x-x')\mathrm{d}s, \qquad x_s := x' + s(x-x'),$$

so the definition of S and $\|\nabla^2 U\| \leqslant M$ give

$$|\eta(x;\xi) - \eta(x';\xi)| \leqslant \big(S(x;\xi) + M\big)|x-x'|. \tag{C.24}$$

The identity

$$\eta(x;\xi)^{\otimes 2} - \eta(x';\xi)^{\otimes 2} = \eta(x;\xi)\otimes\big(\eta(x;\xi)-\eta(x';\xi)\big) + \big(\eta(x;\xi)-\eta(x';\xi)\big)\otimes\eta(x';\xi)$$

gives

$$\big\|\eta(x;\xi)^{\otimes 2} - \eta(x';\xi)^{\otimes 2}\big\| \leqslant \big(|\eta(x;\xi)| + |\eta(x';\xi)|\big)|\eta(x;\xi)-\eta(x';\xi)|.$$

Minkowski's inequality, with Assumption 3 for $\eta(x;\xi)$ and $\eta(x';\xi)$ and with (C.23) for S, gives

$$\mathbb{E}_\xi\Big[\big(|\eta(x;\xi)| + |\eta(x';\xi)|\big)^2\Big] \leqslant 4\Sigma, \qquad \mathbb{E}_\xi\big[(S(x;\xi)+M)^2\big] \leqslant 4M^2, \tag{C.25}$$

so (5.8), the bound on $\eta(x;\xi)^{\otimes 2} - \eta(x';\xi)^{\otimes 2}$, (C.24) and the Cauchy–Schwarz inequality with (C.25) give

$$\begin{aligned}\|\mathcal{C}(x) - \mathcal{C}(x')\| &= \Big\|\mathbb{E}_\xi\big[\eta(x;\xi)^{\otimes 2} - \eta(x';\xi)^{\otimes 2}\big]\Big\| \leqslant \mathbb{E}_\xi\Big[\big(|\eta(x;\xi)| + |\eta(x';\xi)|\big)|\eta(x;\xi)-\eta(x';\xi)|\Big] \\ &\leqslant \mathbb{E}_\xi\Big[\big(|\eta(x;\xi)| + |\eta(x';\xi)|\big)\big(S(x;\xi)+M\big)\Big]\,|x-x'| \leqslant 2\Sigma^{\frac{1}{2}}\cdot 2M\,|x-x'|.\end{aligned}$$

For $|x-x'| > 1$, since $\mathcal{C}(x)$ and $\mathcal{C}(x')$ are positive semidefinite, (5.8) gives

$$\|\mathcal{C}(x) - \mathcal{C}(x')\| \leqslant \mathrm{Tr}\big[\mathcal{C}(x)\big] + \mathrm{Tr}\big[\mathcal{C}(x')\big] \leqslant 2\Sigma \leqslant 2\Sigma\,|x-x'|.$$

Since $\Sigma \leqslant M^2$, both cases give

$$\|\mathcal{C}(x) - \mathcal{C}(x')\| \leqslant 4M^2|x-x'|, \qquad x, x' \in \mathbb{R}^d. \tag{C.26}$$

Writing $\mathcal{C}(x) = \sum_{i=1}^{d} \lambda_i\, u_i \otimes u_i$ with $\lambda_i \geqslant 0$ and (u_i) orthonormal, (5.8) gives

$$|A : \mathcal{C}(x)| = \Big|\sum_{i=1}^{d} \lambda_i \langle u_i, A u_i\rangle\Big| \leqslant \|A\| \operatorname{Tr}\big[\mathcal{C}(x)\big] \leqslant \Sigma \|A\|, \qquad A \in \mathbb{R}^{d\times d},\ x \in \mathbb{R}^d. \tag{C.27}$$

Proof of Lemma 7. Work at the index 0, where $\hat{\mathbb{E}}_0$ and $\mathbb{E}_0$ agree on functions of Y_0 and $(B_s)_{s\leqslant h}$, both fixing Y_0 and averaging over $(B_s)_{s\leqslant h}$. With $D_0 = \sqrt{2h}\,\chi_0 - h\,\nabla U(Z_0)$, the right side of (5.2), the step (5.3) adds the gradient noise η_0 at Z_0, multiplied by $-h$,

$$Y_1 = Y_0 + D_0 - h\eta_0, \qquad \eta_0 := \eta(Z_0;\xi_0) = \nabla\hat{U}(Z_0;\xi_0) - \nabla U(Z_0).$$

The proof expands $G(Y_1)$ about $Y_0 + D_0$ to the order 2 in $h\eta_0$, compares $G(Y_0 + D_0)$ with $G(\tilde{Y}_0(h))$ by Lemma 6, and finally replaces $\nabla^2 G(Y_0 + D_0)$ by $\nabla^2 G(Y_0)$ in the term of order h^2.

Let $\hat{\mathcal{H}}_0 := \sigma\big(\hat{\mathcal{F}}_0, (B_s)_{s\leqslant h}\big)$. Then Y_0, D_0, Z_0 and $\tilde{Y}_0(h)$ are $\hat{\mathcal{H}}_0$-measurable, and since ξ_0 is independent of $\hat{\mathcal{H}}_0$, Assumption 3 and (5.8) give

$$\mathbb{E}\big[\eta_0 \,\big|\, \hat{\mathcal{H}}_0\big] = 0, \qquad \mathbb{E}\big[\eta_0 \otimes \eta_0 \,\big|\, \hat{\mathcal{H}}_0\big] = \mathcal{C}(Z_0).$$

Taylor's formula for G at $Y_0 + D_0$, to the order 2 in $h\eta_0$, gives

$$G(Y_1) = G(Y_0 + D_0) - h\,\nabla G(Y_0 + D_0)\cdot\eta_0 + \frac{h^2}{2}\,\nabla^2 G(Y_0 + D_0)[\eta_0,\eta_0] + \varrho_3^{\eta_0}, \qquad |\varrho_3^{\eta_0}| \leqslant \frac{h^3}{6}\,\|\nabla^3 G\|_\infty\,|\eta_0|^3.$$

Under $\mathbb{E}[\,\cdot\,|\,\hat{\mathcal{H}}_0]$, the term linear in η_0 vanishes and the term quadratic in η_0 becomes $\frac{h^2}{2}\,\nabla^2 G(Y_0+D_0) : \mathcal{C}(Z_0)$. For the remainder, $|\eta_0| \leqslant |\nabla\hat{U}(Z_0;\xi_0)| + |\nabla U(Z_0)|$, and the moment bound of Assumption 3, applied given $\hat{\mathcal{H}}_0$, with $|\nabla U(x)| \leqslant M(1+|x|)$ and $h \leqslant 1$, gives

$$\hat{\mathbb{E}}_0|\eta_0|^3 \leqslant C\,\hat{\mathbb{E}}_0\big(1+|Z_0|\big)^3 \leqslant C\big(1+|Y_0|\big)^3.$$

Hence

$$\Big|\hat{\mathbb{E}}_0\big[G(Y_1)\big] - \mathbb{E}_0\big[G(Y_0+D_0)\big] - \frac{h^2}{2}\,\hat{\mathbb{E}}_0\big[\nabla^2 G(Y_0+D_0) : \mathcal{C}(Z_0)\big]\Big| \leqslant Ch^3\|\nabla^3 G\|_\infty\big(1+|Y_0|\big)^3. \tag{C.28}$$

Next, $Y_0 + D_0$ is the state after one step of (LM') from Y_0, so $\mathbb{E}_0[G(Y_0 + D_0)] - \mathbb{E}_0[G(\tilde{Y}_0(h))]$ is the local weak error of the chain with the full gradient, and (5.7) at $k = 0$ gives

$$\big|\mathbb{E}_0\big[G(Y_0+D_0) - G(\tilde{Y}_0(h))\big]\big| \leqslant Ch^3 \sum_{j=1}^{4} \|\nabla^j G\|_\infty \big(1+|Y_0|\big)^6. \tag{C.29}$$

It remains to replace $\nabla^2 G(Y_0+D_0)$ by $\nabla^2 G(Y_0)$ in the term carrying $\mathcal{C}(Z_0)$, at the cost Ch^3. By Taylor's formula,

$$\nabla^2 G(Y_0+D_0) = \nabla^2 G(Y_0) + \nabla^3 G(Y_0)[D_0] + \varrho_2^{D_0}, \qquad \|\varrho_2^{D_0}\| \leqslant \frac{1}{2}\|\nabla^4 G\|_\infty |D_0|^2,$$

and, with $D_0 = \sqrt{2h}\,\chi_0 - h\,\nabla U(Z_0)$,

$$\begin{aligned} &\big(\nabla^2 G(Y_0+D_0) - \nabla^2 G(Y_0)\big) : \mathcal{C}(Z_0) \\ &\quad = \varrho_2^{D_0} : \mathcal{C}(Z_0) - h\,\nabla^3 G(Y_0)[\nabla U(Z_0)] : \mathcal{C}(Z_0) + \sqrt{2h}\,\nabla^3 G(Y_0)[\chi_0] : \mathcal{C}(Z_0). \end{aligned} \tag{C.30}$$

Assumption 1 gives

$$|\nabla U(Z_0)| \leqslant M\big(1+|Z_0|\big) \leqslant M\Big(1+|Y_0| + \sqrt{\frac{h}{2}}\,|\chi_0|\Big).$$

By (C.27), $\Sigma \leqslant M^2$, the bound on $|\nabla U(Z_0)|$ and (5.5) for the step D_0 of (5.2), the terms of (C.30) carrying $\varrho_2^{D_0}$ and $\nabla U(Z_0)$ contribute

$$\begin{aligned}
&\big|\hat{\mathbb{E}}_0[\varrho_2^{D_0} : \mathcal{C}(Z_0)]\big| \leqslant \frac{1}{2}\|\nabla^4 G\|_\infty \, \Sigma \, \hat{\mathbb{E}}_0|D_0|^2 \leqslant Ch\|\nabla^4 G\|_\infty (1+|Y_0|)^2, \\
&h\,\big|\hat{\mathbb{E}}_0\big[\nabla^3 G(Y_0)[\nabla U(Z_0)] : \mathcal{C}(Z_0)\big]\big| \leqslant \Sigma h\|\nabla^3 G\|_\infty \, \hat{\mathbb{E}}_0|\nabla U(Z_0)| \leqslant Ch\|\nabla^3 G\|_\infty (1+|Y_0|).
\end{aligned} \tag{C.31}$$

For the term of (C.30) carrying χ_0, split

$$\nabla^3 G(Y_0)[\chi_0] : \mathcal{C}(Z_0) = \nabla^3 G(Y_0)[\chi_0] : \mathcal{C}(Y_0) + \nabla^3 G(Y_0)[\chi_0] : \big(\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\big). \tag{C.32}$$

Since χ_0 has mean 0 and is independent of $\hat{\mathcal{F}}_0$, while $\nabla^3 G(Y_0)$ and $\mathcal{C}(Y_0)$ are $\hat{\mathcal{F}}_0$-measurable,

$$\hat{\mathbb{E}}_0\big[\nabla^3 G(Y_0)[\chi_0] : \mathcal{C}(Y_0)\big] = 0, \tag{C.33}$$

and, since $Z_0 - Y_0 = \sqrt{h/2}\,\chi_0$, (C.26) gives

$$\|\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\| \leqslant 4M^2\sqrt{\frac{h}{2}}\,|\chi_0|,$$

and the facts

$$|A : B| \leqslant d\|A\|\|B\|, \qquad \|\nabla^3 G(Y_0)[\chi_0]\| \leqslant \|\nabla^3 G\|_\infty |\chi_0|, \qquad \hat{\mathbb{E}}_0|\chi_0|^2 = d$$

give

$$\begin{aligned}
&\sqrt{2h}\,\Big|\hat{\mathbb{E}}_0\big[\nabla^3 G(Y_0)[\chi_0] : \big(\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\big)\big]\Big| \leqslant \sqrt{2h}\,d\,\|\nabla^3 G\|_\infty \,\hat{\mathbb{E}}_0\big[|\chi_0|\,\|\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\|\big] \\
&\quad\leqslant \sqrt{2h}\,d\,\|\nabla^3 G\|_\infty \cdot 4M^2\sqrt{\frac{h}{2}}\,\hat{\mathbb{E}}_0|\chi_0|^2 = 4d^2M^2h\|\nabla^3 G\|_\infty.
\end{aligned} \tag{C.34}$$

Inserting (C.31), (C.33) and (C.34) into (C.30) and (C.32), and multiplying by $\frac{h^2}{2}$, gives

$$\frac{h^2}{2}\Big|\hat{\mathbb{E}}_0\big[\big(\nabla^2 G(Y_0+D_0) - \nabla^2 G(Y_0)\big) : \mathcal{C}(Z_0)\big]\Big| \leqslant Ch^3\big(\|\nabla^3 G\|_\infty + \|\nabla^4 G\|_\infty\big)(1+|Y_0|)^2. \tag{C.35}$$

Finally, the left side of (5.9) at the index 0 splits as

$$\begin{aligned}
&\hat{\mathbb{E}}_0\big[G(Y_1) - G(\tilde{Y}_0(h))\big] - \frac{h^2}{2}\,\nabla^2 G(Y_0) : \hat{\mathbb{E}}_0\big[\mathcal{C}(Z_0)\big] \\
&= \Big(\hat{\mathbb{E}}_0\big[G(Y_1)\big] - \mathbb{E}_0\big[G(Y_0+D_0)\big] - \frac{h^2}{2}\,\hat{\mathbb{E}}_0\big[\nabla^2 G(Y_0+D_0) : \mathcal{C}(Z_0)\big]\Big) + \mathbb{E}_0\big[G(Y_0+D_0) - G(\tilde{Y}_0(h))\big] \\
&\quad + \frac{h^2}{2}\,\hat{\mathbb{E}}_0\big[\big(\nabla^2 G(Y_0+D_0) - \nabla^2 G(Y_0)\big) : \mathcal{C}(Z_0)\big].
\end{aligned} \tag{C.36}$$

On the right side of (C.36), (C.28) bounds the first bracket, (C.29) the second term, and (C.35) the third term. Adding these bounds gives

$$\begin{aligned}
&\Big|\hat{\mathbb{E}}_0\big[G(Y_1) - G(\tilde{Y}_0(h))\big] - \frac{h^2}{2}\,\nabla^2 G(Y_0) : \hat{\mathbb{E}}_0\big[\mathcal{C}(Z_0)\big]\Big| \\
&\leqslant Ch^3\|\nabla^3 G\|_\infty (1+|Y_0|)^3 + Ch^3\sum_{j=1}^{4}\|\nabla^j G\|_\infty (1+|Y_0|)^6 + Ch^3\big(\|\nabla^3 G\|_\infty + \|\nabla^4 G\|_\infty\big)(1+|Y_0|)^2 \\
&\leqslant Ch^3\sum_{j=1}^{4}\|\nabla^j G\|_\infty (1+|Y_0|)^6,
\end{aligned}$$

which is (5.9) at the index 0; by time homogeneity, (5.9) holds at every k. □

D Proofs of the MSE analysis

Throughout this appendix, C depends only on m, M, R and d, and $h \leqslant h_0$ with $h_0 \leqslant \min\{\lambda^{-1}, \frac{1}{2}\}$, λ the rate of Theorem 1 for $\tilde{U}_h$, so that Corollary 1 applies. By (4.19), $\tilde{\phi}_h$ is CL-Lipschitz.

D.1 Proof of Theorem 2: MSE of LM

Proof of Theorem 2. Let first $Nh \geqslant 1$, so that $N \geqslant 2$ since $h \leqslant \frac{1}{2}$. The first step of (LM) gives

$$Z_k = Y_k + \sqrt{\frac{h}{2}}\,\chi_k, \quad k \geqslant 1, \qquad Y_1 := Z_0 - h\nabla U(Z_0) + \sqrt{\frac{h}{2}}\,\chi_0,$$

with $(Y_k)_{k\geqslant 1}$ the chain (LM') started from Y_1. Here Y_1 is independent of $(B_t - B_h)_{t\geqslant h}$, and $\mathbb{E}|Y_1|^{12} \leqslant C(1+|Z_0|)^{12}$. Splitting off the term $k=0$ in (1.2),

$$\mathrm{MSE}(N,h) \leqslant \frac{2}{N^2}\big|f(Z_0) - \pi(f)\big|^2 + 2\Big(\frac{N-1}{N}\Big)^2 \mathbb{E}\left[\left|\frac{1}{N-1}\sum_{k=1}^{N-1} f(Z_k) - \pi(f)\right|^2\right],$$

where

$$\frac{2}{N^2}\big|f(Z_0) - \pi(f)\big|^2 \leqslant \frac{CL^2}{N^2}\big(1+|Z_0|\big)^2 \leqslant \frac{CL^2}{Nh}\big(1+|Z_0|\big)^2,$$

and the expectation of the time average over $Z_1, \dots, Z_{N-1}$ is the MSE of $N-1$ samples of a chain of the same form started from Y_1, with $\frac{1}{(N-1)h} \leqslant \frac{2}{Nh}$ for $N \geqslant 2$. Conditioning on Y_1, it therefore suffices to prove (6.3) for the samples $Z_k = Y_k + \sqrt{h/2}\,\chi_k$, $k \geqslant 0$, of the chain (LM') started from a given Y_0, with Y_0 in place of Z_0, and then to apply it with Y_1 in place of Y_0 and take the expectation, with $\mathbb{E}(1+|Y_1|)^2 \leqslant C(1+|Z_0|)^2$ and $\mathbb{E}(1+|Y_1|)^{12} \leqslant C(1+|Z_0|)^{12}$. By (5.1) and $\frac{h}{N} \leqslant \frac{1}{Nh}$ for $h \leqslant 1$, it suffices to prove

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{k=0}^{N-1}\big(\tilde{f}_h(Y_k) - \tilde{\pi}_h(\tilde{f}_h)\big)\right|^2\right] \leqslant CL^2\Big(\big(1+|Y_0|\big)^2\frac{1}{Nh} + \big(1+|Y_0|\big)^{12}h^4\Big). \tag{D.1}$$

By (6.1) and Cauchy's inequality, the left side of (D.1) is at most

$$\frac{3}{N^2h^2}\,\mathbb{E}\big|\tilde{\phi}_h(Y_0) - \tilde{\phi}_h(Y_N)\big|^2 + \frac{3}{N^2}\,\mathbb{E}\left|\sum_{k=0}^{N-1}\Upsilon_k\right|^2 + \frac{3}{N^2}\,\mathbb{E}\left|\sum_{k=0}^{N-1}\Psi_k\right|^2. \tag{D.2}$$

The terms Υ_k and Ψ_k are $\mathcal{F}_{k+1}$-measurable, since $\tilde{Y}_k(h)$ is driven by the Brownian motion on $[kh,(k+1)h]$, and $\mathbb{E}_k\Psi_k = 0$ by (4.18). Hence the Ψ_k, and the centered terms $\Upsilon_k - \mathbb{E}_k\Upsilon_k$, are martingale differences with respect to $(\mathcal{F}_k)_{k\geqslant 0}$, and their sums have additive second moments. Lemmas 3 and 4 also give

$$\mathbb{E}|Y_k|^2 \leqslant C\big(1+|Y_0|\big)^2, \qquad \mathbb{E}|Y_{k+1} - Y_k|^2 + \mathbb{E}|\tilde{Y}_k(h) - Y_k|^2 \leqslant Ch\big(1+|Y_0|\big)^2. \tag{D.3}$$

For the boundary term of (D.2), the Lipschitz bound on $\tilde{\phi}_h$ and (D.3) give

$$\frac{3}{N^2h^2}\,\mathbb{E}\big|\tilde{\phi}_h(Y_0) - \tilde{\phi}_h(Y_N)\big|^2 \leqslant \frac{CL^2}{N^2h^2}\,\mathbb{E}|Y_N - Y_0|^2 \leqslant \frac{CL^2}{N^2h^2}\big(1+|Y_0|\big)^2 \leqslant \frac{CL^2}{Nh}\big(1+|Y_0|\big)^2, \tag{D.4}$$

using $Nh \geqslant \frac{1}{2}$. This holds for the $N-1$ samples after the first step, since $(N-1)h \geqslant 1-h \geqslant \frac{1}{2}$.

For the discretization term of (D.2), carrying Υ_k with $h\Upsilon_k = \tilde{\phi}_h(Y_{k+1}) - \tilde{\phi}_h(\tilde{Y}_k(h))$ by (6.2), Lemma 6 with $G = \tilde{\phi}_h$ and (4.19) give

$$|\mathbb{E}_k\Upsilon_k| \leqslant CLh^2\big(1+|Y_k|\big)^6,$$

so by Cauchy's inequality and Lemma 3,

$$\mathbb{E}\left|\sum_{k=0}^{N-1}\mathbb{E}_k\Upsilon_k\right|^2 \leqslant N\sum_{k=0}^{N-1}\mathbb{E}\big|\mathbb{E}_k\Upsilon_k\big|^2 \leqslant CL^2N^2h^4\big(1+|Y_0|\big)^{12}.$$

For its fluctuation, the Lipschitz bound on $\tilde{\phi}_h$ and Lemma 4 give

$$h|\Upsilon_k| \leqslant CL\,\big|Y_{k+1} - \tilde{Y}_k(h)\big|, \qquad \mathbb{E}_k|\Upsilon_k|^2 \leqslant CL^2 h\big(1+|Y_k|\big)^2,$$

so by the orthogonality of martingale differences,

$$\mathbb{E}\Big|\sum_{k=0}^{N-1}\big(\Upsilon_k - \mathbb{E}_k\Upsilon_k\big)\Big|^2 = \sum_{k=0}^{N-1}\mathbb{E}\big|\Upsilon_k - \mathbb{E}_k\Upsilon_k\big|^2 \leqslant CL^2Nh\big(1+|Y_0|\big)^2.$$

The discretization term of (D.2) is therefore at most

$$\frac{3}{N^2}\,\mathbb{E}\Big|\sum_{k=0}^{N-1}\Upsilon_k\Big|^2 \leqslant CL^2\Big(\big(1+|Y_0|\big)^{12}h^4 + \big(1+|Y_0|\big)^2\frac{h}{N}\Big). \tag{D.5}$$

For the martingale term of (D.2), carrying Ψ_k with $h\Psi_k = \tilde{\phi}_h(\tilde{Y}_k(h)) - \mathbb{E}_k[\tilde{\phi}_h(\tilde{Y}_k(h))]$ by (6.2) and (4.18), the Lipschitz bound and (D.3) give

$$h^2\,\mathbb{E}|\Psi_k|^2 \leqslant CL^2\,\mathbb{E}\big|\tilde{Y}_k(h) - Y_k\big|^2 \leqslant CL^2h\big(1+|Y_0|\big)^2,$$

and by orthogonality,

$$\frac{3}{N^2}\,\mathbb{E}\Big|\sum_{k=0}^{N-1}\Psi_k\Big|^2 = \frac{3}{N^2}\sum_{k=0}^{N-1}\mathbb{E}|\Psi_k|^2 \leqslant \frac{CL^2}{Nh}\big(1+|Y_0|\big)^2. \tag{D.6}$$

Adding (D.4), (D.5) and (D.6), and using $\frac{h}{N} \leqslant \frac{1}{Nh}$ for $h \leqslant 1$, proves (D.1).

Let finally $Nh < 1$. By Assumption 2 and the moments of π in Lemma 17, $|f(x) - \pi(f)| \leqslant CL(1+|x|)$, and $\mathbb{E}|Z_k|^2 \leqslant C(1+|Z_0|)^2$ for every $k \geqslant 0$, by the moment bound on Y_1 and Lemma 3. Hence (1.2) and Cauchy's inequality give

$$\mathrm{MSE}(N,h) \leqslant \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\big|f(Z_k) - \pi(f)\big|^2 \leqslant CL^2\big(1+|Z_0|\big)^2 \leqslant \frac{CL^2}{Nh}\big(1+|Z_0|\big)^2,$$

which is at most the right side of (6.3). □

D.2 Proof of Theorem 3: MSE of SGLM

Proof of Theorem 3. As in the proof of Theorem 2, with $Y_1 := Z_0 - h\nabla\hat{U}(Z_0;\xi_0) + \sqrt{h/2}\,\chi_0$, it suffices to bound the MSE of the samples $Z_k = Y_k + \sqrt{h/2}\,\chi_k$, $k \geqslant 0$, of the chain (SGLM') started from a given Y_0 by the right side of (6.4) with Y_0 in place of Z_0, and then to apply it with Y_1 in place of Y_0 and take the expectation, with $\mathbb{E}(1+|Y_1|)^2 \leqslant C(1+|Z_0|)^2$ and $\mathbb{E}(1+|Y_1|)^{12} \leqslant C(1+|Z_0|)^{12}$ by Assumption 3. The bound (5.1) and the formula (6.1) involve the modified test function, the modified dynamics and $\tilde{\phi}_h$ only, and hold for this chain with $\hat{\mathcal{F}}_k$ and $\hat{\mathbb{E}}_k$ in place of $\mathcal{F}_k$ and $\mathbb{E}_k$. The proof of Theorem 2 is repeated with two changes.

First, the moment bounds of Lemma 3 hold for (SGLM') with the same constants, so the boundary term and the martingale term are bounded by $CL^2(1+|Y_0|)^2/(Nh)$ as before.

Second, in the discretization term, Lemma 7 with $G = \tilde{\phi}_h$, (4.19) and (C.27) give

$$\big|\hat{\mathbb{E}}_k\Upsilon_k\big| \leqslant CL\Big(\Sigma h + h^2\big(1+|Y_k|\big)^6\Big),$$

since $\frac{h^2}{2}|\nabla^2\tilde{\phi}_h(Y_k) : \hat{\mathbb{E}}_k[\mathcal{C}(Z_k)]| \leqslant CL\Sigma h^2$. Cauchy's inequality and Lemma 3 for the term $h^2(1+|Y_k|)^6$ give

$$\mathbb{E}\Big|\sum_{k=0}^{N-1}\hat{\mathbb{E}}_k\Upsilon_k\Big|^2 \leqslant CL^2N^2\Big(\Sigma^2h^2 + h^4\big(1+|Y_0|\big)^{12}\Big).$$

For its fluctuation, the step of (SGLM') differs from that of (LM') by $-h\big(\nabla\hat{U}(Z_k;\xi_k)-\nabla U(Z_k)\big)$, and

$$\hat{\mathbb{E}}_k\big|h\big(\nabla\hat{U}(Z_k;\xi_k)-\nabla U(Z_k)\big)\big|^2 \leqslant \Sigma h^2 \leqslant M^2h^2.$$

With the Lipschitz bound on $\tilde{\phi}_h$ and Lemma 4, this gives

$$\hat{\mathbb{E}}_k|\Upsilon_k|^2 \leqslant CL^2\Big(h\big(1+|Y_k|\big)^2+\Sigma\Big) \leqslant CL^2\big(1+|Y_k|\big)^2,$$

and, by orthogonality and $h\leqslant 1$,

$$\frac{3}{N^2}\,\mathbb{E}\Big|\sum_{k=0}^{N-1}\big(\Upsilon_k-\hat{\mathbb{E}}_k\Upsilon_k\big)\Big|^2 \leqslant \frac{CL^2}{N}\big(1+|Y_0|\big)^2 \leqslant \frac{CL^2}{Nh}\big(1+|Y_0|\big)^2.$$

Adding the bounds proves (6.4) for $Nh\geqslant 1$. For $Nh<1$, it follows as at the end of the proof of Theorem 2, since the moment bounds of Lemma 3 hold for (SGLM') and $\mathbb{E}(1+|Y_1|)^2\leqslant C(1+|Z_0|)^2$.

For the average of P independent estimators, the average is unbiased, its moment of order 12 satisfies the bound of Assumption 3 with the same M by Minkowski's inequality, and its variance is at most $\Sigma/P\leqslant M^2$. Minkowski's inequality in L^2 also gives the Hessian bound of Assumption 3 with the same M. Assumption 3 therefore holds with Σ/P in place of Σ, and (6.4) gives the term Σ^2h^2/P^2. □

E Proofs of the numerical bias estimate

Throughout this appendix, $(Y_k)_{k\geqslant 0}$ is the chain (SGLM'), $\mathcal{P}_h$ its transition kernel, and $h\leqslant h_0$ with $Mh_0\leqslant 1$. The local Hessian bound (C.23) and the bounds (C.26) and (C.27) on the covariance $\mathcal{C}(x)$ of the gradient noise are those of Appendix C.5.

E.1 Proof of Lemma 8: invariant distribution

Proof of Lemma 8. The proof obtains the existence and uniqueness of μ_h from Harris's ergodic theorem, deduces the moment bound of μ_h from Lemma 3, and proves the limit of time averages first for the chain (SGLM') and then, in (6.5), for the samples of (SGLM). Harris's ergodic theorem in the form of [47, Theorem 1.2] assumes that a measurable $W\colon\mathbb{R}^d\to[0,\infty)$, constants $\gamma,\alpha_h\in(0,1)$, $K\geqslant 0$, a level $R_W>0$ and a probability measure ν satisfy the drift condition, the minorization on the level set $\{W\leqslant R_W\}$, and the condition on the level,

$$(\mathcal{P}_hW)(y)\leqslant\gamma W(y)+K,\quad y\in\mathbb{R}^d,\qquad \mathcal{P}_h(y,\cdot)\geqslant\alpha_h\,\nu,\quad W(y)\leqslant R_W,\qquad R_W>\frac{2K}{1-\gamma}. \tag{E.1}$$

That theorem then gives a unique invariant distribution μ_h and constants $C_h>0$ and $\bar{\gamma}\in(0,1)$, which may depend on h, with

$$\sup_{y\in\mathbb{R}^d}\frac{\big|(\mathcal{P}_h^n\varphi)(y)-\mu_h(\varphi)\big|}{1+W(y)}\leqslant C_h\,\bar{\gamma}^{\,n}\sup_{y\in\mathbb{R}^d}\frac{|\varphi(y)-\mu_h(\varphi)|}{1+W(y)},\qquad n\geqslant 0, \tag{E.2}$$

for every measurable φ with $\sup_{y\in\mathbb{R}^d}|\varphi(y)|/(1+W(y))<\infty$.

For the drift condition in (E.1), take $W(y):=|y|^2$. The Lyapunov estimate (C.10) with $q=1$, which holds for (SGLM') as shown in the proof of Lemma 3, gives

$$(\mathcal{P}_hW)(y)\leqslant\Big(1-\frac{m}{16}h\Big)W(y)+Ch,\qquad y\in\mathbb{R}^d.$$

This is the drift condition in (E.1) with

$$\gamma:=1-\frac{m}{16}h,\qquad K:=Ch,$$

where $\gamma \in (0,1)$ since $m \leqslant M$ by Assumption 1 and $Mh \leqslant 1$. With the constant C in $K = Ch$, set

$$R_W := \frac{64C}{m} > \frac{32C}{m} = \frac{2K}{1-\gamma},$$

which depends only on m, M, R and d.

Next we prove the minorization on the level set $\{W \leqslant R_W\}$, with a constant α_h that may depend on h. Let h_0 also satisfy $4Mh_0(2 + R_W^{\frac{1}{2}}) \leqslant 1$, let $\mathbb{B}$ be the closed unit ball and ν the uniform distribution on $\mathbb{B}$, and fix y with $|y|^2 \leqslant R_W$. From $Y_0 = y$, the step (SGLM') reads

$$Y_1 = T_{\xi_0}\Big(\frac{y}{2} + \sqrt{\frac{h}{2}}\,\chi_0\Big), \qquad T_\xi(v) := 2v - h\nabla\hat{U}\Big(\frac{y}{2} + v; \xi\Big), \qquad v \in \mathbb{R}^d. \tag{E.3}$$

By Markov's inequality, (C.23) and the moment bound of Assumption 3, the event $E_0 := \{\xi : S(\frac{y}{2}; \xi) \leqslant 4M,\ |\nabla\hat{U}(\frac{y}{2}; \xi)| \leqslant 4M(1 + \frac{|y|}{2})\}$ satisfies $\mathbb{P}_\xi(E_0^c) \leqslant \frac{1}{16} + 4^{-12} \leqslant \frac{1}{2}$. For $\xi \in E_0$ and $v, v' \in \mathbb{B}$,

$$h\Big|\nabla\hat{U}\Big(\frac{y}{2} + v; \xi\Big) - \nabla\hat{U}\Big(\frac{y}{2} + v'; \xi\Big)\Big| \leqslant 4Mh\,|v - v'|, \qquad h\Big|\nabla\hat{U}\Big(\frac{y}{2} + v; \xi\Big)\Big| \leqslant 4Mh\Big(2 + \frac{|y|}{2}\Big) \leqslant 1.$$

Hence T_ξ is Lipschitz with constant 3 on $\mathbb{B}$, and for every $w \in \mathbb{B}$ the map

$$v \mapsto \frac{1}{2}\Big(w + h\nabla\hat{U}\Big(\frac{y}{2} + v; \xi\Big)\Big)$$

maps $\mathbb{B}$ into itself with Lipschitz constant $\frac{1}{2}$, so, by the contraction mapping theorem, it has a fixed point, which solves $T_\xi(v) = w$; hence $\mathbb{B} \subset T_\xi(\mathbb{B})$. For a Borel set A, the set $A \cap \mathbb{B}$ is therefore contained in $T_\xi(T_\xi^{-1}(A) \cap \mathbb{B})$, whose outer Lebesgue measure is at most $3^d \operatorname{Leb}(T_\xi^{-1}(A) \cap \mathbb{B})$ by the Lipschitz bound for the d-dimensional Hausdorff measure, which equals the Lebesgue measure Leb on $\mathbb{R}^d$; on $\mathbb{B}$, whose points lie within $1 + \frac{1}{2}R_W^{\frac{1}{2}}$ of $\frac{y}{2}$, the density of $\frac{y}{2} + \sqrt{h/2}\,\chi_0 \sim \mathcal{N}(\frac{y}{2}, \frac{h}{2}I_d)$ is at least $\beta_h := (\pi h)^{-\frac{d}{2}} \exp\big(-(1 + \frac{1}{2}R_W^{\frac{1}{2}})^2/h\big)$. Conditioning (E.3) on ξ_0, which is independent of χ_0, gives

$$\mathcal{P}_h(y, A) \geqslant \mathbb{E}_\xi\Big[\mathbf{1}_{\{\xi \in E_0\}}\,\mathbb{P}\Big(\frac{y}{2} + \sqrt{\frac{h}{2}}\,\chi_0 \in T_\xi^{-1}(A) \cap \mathbb{B}\Big)\Big] \geqslant \frac{\beta_h}{2 \cdot 3^d}\operatorname{Leb}(A \cap \mathbb{B}) = \frac{\beta_h \operatorname{Leb}(\mathbb{B})}{2 \cdot 3^d}\,\nu(A),$$

which is the minorization with $\alpha_h := \min\{\frac{1}{2}, \beta_h \operatorname{Leb}(\mathbb{B})/(2 \cdot 3^d)\}$, and (E.2) holds.

For the moment bound, let $\varphi_\ell(y) := \min\{|y|^{12}, \ell\}$, $\ell \geqslant 1$. Since $0 \leqslant \varphi_\ell \leqslant \ell$, (E.2) with $\varphi = \varphi_\ell$ and $n = k$ gives, for the chain from $Y_0 = 0$,

$$\big|\mathbb{E}\big[\varphi_\ell(Y_k)\big] - \mu_h(\varphi_\ell)\big| = \big|(\mathcal{P}_h^k\varphi_\ell)(0) - \mu_h(\varphi_\ell)\big| \leqslant \ell\, C_h\, \bar{\gamma}^{\,k} \to 0, \qquad k \to \infty.$$

With Lemma 3 at $Y_0 = 0$, this gives

$$\mu_h(\varphi_\ell) = \lim_{k\to\infty} \mathbb{E}\big[\varphi_\ell(Y_k)\big] \leqslant \sup_{k \geqslant 0} \mathbb{E}|Y_k|^{12} \leqslant C.$$

Monotone convergence as $\ell \to \infty$ gives $\mu_h(|y|^{12}) \leqslant C$.

Next we prove the limit of time averages for the chain (SGLM'). By the minorization, the level set $\{W \leqslant R_W\}$ is petite [78, Section 3.1]. Outside this set, $W > R_W > \frac{2K}{1-\gamma}$, so the drift condition in (E.1) gives $\mathcal{P}_h W \leqslant \frac{1+\gamma}{2}\,W$. For $Y_0 = y$ with $W(y) > R_W$ and the entrance time τ of the chain into $\{W \leqslant R_W\}$, the process $(\frac{1+\gamma}{2})^{-(k\wedge\tau)}\,W(Y_{k\wedge\tau})$ is therefore a nonnegative supermartingale; it exceeds $(\frac{1+\gamma}{2})^{-k}R_W$ on $\{\tau > k\}$ and has mean at most $W(y)$, so

$$\mathbb{P}(\tau > k) \leqslant \Big(\frac{1+\gamma}{2}\Big)^k \frac{W(y)}{R_W} \to 0, \qquad k \to \infty.$$

By the Markov property at the time 1, the chain started from any initial value returns to $\{W \leqslant R_W\}$ almost surely, so it is Harris recurrent by [78, Theorem 4.1(i)], and, since it has the invariant probability μ_h, positive

Harris recurrent [78, Section 1.3]. Let g be measurable and of linear growth, so that $\mu_h(|g|) < \infty$ by the moment bound. The strong law of large numbers for positive Harris recurrent chains [79, Theorem 17.0.1(i)] gives, for every initial value Y_0,

$$\frac{1}{N}\sum_{k=0}^{N-1} g(Y_k) \to \mu_h(g) \qquad \text{almost surely.} \tag{E.4}$$

For the samples of (SGLM) from the given Z_0 of (6.5), start (SGLM'), as in Section 6.1, at

$$Y_1 = Z_0 - h\nabla\hat{U}(Z_0;\xi_0) + \sqrt{\frac{h}{2}}\,\chi_0, \qquad \text{so that} \qquad Z_k = Y_k + \sqrt{\frac{h}{2}}\,\chi_k, \quad k \geqslant 1.$$

Since Y_1 is independent of $(\chi_k, \xi_k)_{k\geqslant 1}$, the sequence $(Y_{j+1})_{j\geqslant 0}$ is the chain (SGLM') started at Y_1, and (E.4), which holds for every initial value, gives by Fubini's theorem, almost surely,

$$\frac{1}{N}\sum_{j=0}^{N-1} g(Y_{j+1}) \to \mu_h(g), \qquad N \to \infty. \tag{E.5}$$

For $k \geqslant 1$, as in Section 4.1, $\mathfrak{M}_k = f(Z_k) - \tilde{f}_h(Y_k)$, where $\hat{\mathbb{E}}_k[\mathfrak{M}_k] = 0$ by (4.1) and $\mathbb{E}[\mathfrak{M}_k^2] \leqslant \frac{dh}{2}L^2$ by (4.3), both with $\hat{\mathbb{E}}_k$ of (3.2) in place of $\mathbb{E}_k$, since Y_k is $\hat{\mathcal{F}}_k$-measurable and χ_k is independent of $\hat{\mathcal{F}}_k$; the $\mathfrak{M}_k$ are therefore $\hat{\mathcal{F}}_{k+1}$-measurable martingale differences. By (E.5) with $g = \tilde{f}_h$, which is Lipschitz with constant L by Section 4.1, hence of linear growth,

$$\frac{1}{N}\sum_{k=1}^{N-1} \tilde{f}_h(Y_k) \to \mu_h(\tilde{f}_h) \qquad \text{almost surely.}$$

By the bound on $\mathbb{E}[\mathfrak{M}_k^2]$, the martingale $\sum_{k=1}^{N}\mathfrak{M}_k/k$ is bounded in L^2 and converges almost surely, and Kronecker's lemma gives

$$\frac{1}{N}\sum_{k=1}^{N-1}\mathfrak{M}_k \to 0 \qquad \text{almost surely.}$$

Let $Y \sim \mu_h$ and $\chi \sim \mathcal{N}(0, I_d)$ be independent. Fubini's theorem, (4.1) and the definition of π_h in Lemma 8 give

$$\mu_h(\tilde{f}_h) = \mathbb{E}\left[f\left(Y + \sqrt{\frac{h}{2}}\,\chi\right)\right] = \pi_h(f). \tag{E.6}$$

Combining the limits of the averages of $\tilde{f}_h(Y_k)$ and of $\mathfrak{M}_k$ with (E.6) and with $f(Z_0)/N \to 0$, since Z_0 is given,

$$\frac{1}{N}\sum_{k=0}^{N-1} f(Z_k) = \frac{f(Z_0)}{N} + \frac{1}{N}\sum_{k=1}^{N-1}\tilde{f}_h(Y_k) + \frac{1}{N}\sum_{k=1}^{N-1}\mathfrak{M}_k \to \pi_h(f) \qquad \text{almost surely,}$$

which is (6.5). □

E.2 Proof of Theorem 4: numerical bias

Proof of Theorem 4. The proof writes $\mu_h(\tilde{f}_h) - \tilde{\pi}_h(\tilde{f}_h)$ through the modified discrete Poisson equation (4.18) as the local weak error of Lemma 7 averaged over μ_h, and then adds the bound of Lemma 2 on $\tilde{\pi}_h(\tilde{f}_h) - \pi(f)$.

Let $Y_0 \sim \mu_h$, the invariant distribution of (SGLM') given by Lemma 8, independent of $(B_t)_{t\geqslant 0}$ and $(\xi_k)_{k\geqslant 0}$. Then $Y_1 \sim \mu_h$. Add Y_0 to the generators of $\hat{\mathcal{F}}_k$ in (3.2), so that $\hat{\mathcal{F}}_0 = \sigma(Y_0)$, while χ_k and ξ_k remain independent of $\hat{\mathcal{F}}_k$. The process $\tilde{Y}_0(s)$ evolves by the modified dynamics (4.16), with the generator $\tilde{\mathcal{L}}_h$ of (4.17), from $\tilde{Y}_0(0) = Y_0$, so that

$$(e^{h\tilde{\mathcal{L}}_h}\tilde{\phi}_h)(Y_0) = \hat{\mathbb{E}}_0\big[\tilde{\phi}_h(\tilde{Y}_0(h))\big].$$

The modified discrete Poisson equation (4.18) at Y_0 therefore reads

$$\tilde{f}_h(Y_0) - \tilde{\pi}_h(\tilde{f}_h) = \frac{1}{h}\,\hat{\mathbb{E}}_0\big[\tilde{\phi}_h(Y_0) - \tilde{\phi}_h(\tilde{Y}_0(h))\big].$$

Taking expectations, with $\mathbb{E}\,\tilde{f}_h(Y_0) = \mu_h(\tilde{f}_h)$, gives

$$\mu_h(\tilde{f}_h) - \tilde{\pi}_h(\tilde{f}_h) = \frac{1}{h}\,\mathbb{E}\big[\tilde{\phi}_h(Y_0) - \tilde{\phi}_h(\tilde{Y}_0(h))\big] = \frac{1}{h}\,\mathbb{E}\big[\tilde{\phi}_h(Y_1) - \tilde{\phi}_h(\tilde{Y}_0(h))\big], \tag{E.7}$$

where $\mathbb{E}\,\tilde{\phi}_h(Y_0) = \mathbb{E}\,\tilde{\phi}_h(Y_1)$ since $Y_1 \sim \mu_h$. All expectations in (E.7) are finite, since $\tilde{\phi}_h$ is Lipschitz with constant CL by (4.19), $\mu_h(|y|^{12}) \leqslant C$ by Lemma 8, and $\mathbb{E}|\tilde{Y}_0(h) - Y_0| \leqslant C\sqrt{h}\,(1 + \mathbb{E}|Y_0|)$ by the bound on $\tilde{Y}_0(s) - Y_0$ in (C.1).

Next we expand $\mathbb{E}\big[\tilde{\phi}_h(Y_1) - \tilde{\phi}_h(\tilde{Y}_0(h))\big]$ in (E.7) by Lemma 7. By (4.19), the derivatives of $\tilde{\phi}_h$ of orders 1 to 4 are bounded by CL, so (5.9) with $G = \tilde{\phi}_h$ at $k = 0$ gives

$$\Big|\hat{\mathbb{E}}_0\big[\tilde{\phi}_h(Y_1) - \tilde{\phi}_h(\tilde{Y}_0(h))\big] - \frac{h^2}{2}\,\nabla^2\tilde{\phi}_h(Y_0) : \hat{\mathbb{E}}_0\big[\mathcal{C}(Z_0)\big]\Big| \leqslant CLh^3\big(1 + |Y_0|\big)^6.$$

Take expectations in the bound on $\hat{\mathbb{E}}_0\big[\tilde{\phi}_h(Y_1) - \tilde{\phi}_h(\tilde{Y}_0(h))\big]$. Its term of order h^2 carries the $\hat{\mathcal{F}}_0$-measurable factor $\nabla^2\tilde{\phi}_h(Y_0)$, so the tower property gives

$$\mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \hat{\mathbb{E}}_0[\mathcal{C}(Z_0)]\big] = \mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big],$$

and hence

$$\Big|\mathbb{E}\big[\tilde{\phi}_h(Y_1) - \tilde{\phi}_h(\tilde{Y}_0(h))\big] - \frac{h^2}{2}\,\mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big]\Big| \leqslant CLh^3\,\mathbb{E}\big(1 + |Y_0|\big)^6.$$

Dividing by h and inserting (E.7),

$$\Big|\mu_h(\tilde{f}_h) - \tilde{\pi}_h(\tilde{f}_h) - \frac{h}{2}\,\mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big]\Big| \leqslant CLh^2\,\mathbb{E}\big(1 + |Y_0|\big)^6 \leqslant CLh^2, \tag{E.8}$$

by the moment bound $\mu_h(|y|^{12}) \leqslant C$ of Lemma 8. By (E.6), $\pi_h(f) = \mu_h(\tilde{f}_h)$, so

$$\begin{aligned}
&\pi_h(f) - \pi(f) - \frac{h}{2}\,\mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big] \\
&= \Big(\mu_h(\tilde{f}_h) - \tilde{\pi}_h(\tilde{f}_h) - \frac{h}{2}\,\mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big]\Big) + \big(\tilde{\pi}_h(\tilde{f}_h) - \pi(f)\big).
\end{aligned} \tag{E.9}$$

The first bracket of (E.9) is bounded by (E.8), and the second by (4.15) of Lemma 2, which proves (6.6). Finally, by (C.27) and (4.19),

$$\frac{h}{2}\Big|\mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big]\Big| \leqslant \frac{\Sigma h}{2}\,\|\nabla^2\tilde{\phi}_h\|_\infty \leqslant CL\Sigma h,$$

which with (6.6) gives

$$|\pi_h(f) - \pi(f)| \leqslant CL(\Sigma h + h^2).$$

□

E.3 Proof of Corollary 2: coefficient of the numerical bias

Proof of Corollary 2. With $Y_0 \sim \mu_h$ and Z_0 as in Theorem 4, let

$$a_h := \mathbb{E}\big[\nabla^2\tilde{\phi}_h(Y_0) : \mathcal{C}(Z_0)\big]. \tag{E.10}$$

Dividing (6.6) by h gives

$$\Big|\frac{\pi_h(f) - \pi(f)}{h} - \frac{1}{2}\,a_h\Big| \leqslant CLh,$$

so it suffices to prove

$$a_h \to \pi\big(\nabla^2\phi_0(x) : \mathcal{C}(x)\big), \qquad h \to 0.$$

Since $Z_0 = Y_0 + \sqrt{h/2}\,\chi_0$ with $\chi_0 \sim \mathcal{N}(0, I_d)$ independent of $Y_0 \sim \mu_h$, conditioning on Y_0 in (E.10) gives

$$a_h = \mathbb{E}\Big[\nabla^2\tilde{\phi}_h(Y_0) : \mathbb{E}\big[\mathcal{C}(Z_0) \,\big|\, Y_0\big]\Big] = \mu_h(A_h), \tag{E.11}$$

with

$$A_h(y) := \nabla^2\tilde{\phi}_h(y) : \tilde{\mathcal{C}}_h(y), \qquad \tilde{\mathcal{C}}_h(y) := \mathbb{E}\Big[\mathcal{C}\Big(y + \sqrt{\frac{h}{2}}\,\chi\Big)\Big], \quad \chi \sim \mathcal{N}(0, I_d). \tag{E.12}$$

First, $\mu_h \to \pi$ weakly as $h \to 0$. The family (μ_h) is tight: since $\mu_h(|y|^{12}) \leqslant C$ uniformly in h by Lemma 8, Markov's inequality gives

$$\mu_h\big(|y| > \rho\big) \leqslant \rho^{-12}\,\mu_h\big(|y|^{12}\big) \leqslant C\rho^{-12}, \qquad \forall \rho > 0. \tag{E.13}$$

Let $\varphi \in C_c^\infty(\mathbb{R}^d)$; it satisfies Assumption 2 with some constant L_φ. By (E.6) and (4.2),

$$\pi_h(\varphi) = \mu_h\big(e^{\frac{h}{4}\Delta}\varphi\big), \qquad \big|(e^{\frac{h}{4}\Delta}\varphi)(y) - \varphi(y)\big| \leqslant CL_\varphi h, \quad y \in \mathbb{R}^d,$$

and with Theorem 4 applied to φ these give

$$|\mu_h(\varphi) - \pi(\varphi)| \leqslant \Big|\mu_h\big(\varphi - e^{\frac{h}{4}\Delta}\varphi\big)\Big| + |\pi_h(\varphi) - \pi(\varphi)| \leqslant CL_\varphi h + CL_\varphi(\Sigma h + h^2).$$

Hence $\mu_h(\varphi) \to \pi(\varphi)$ for every $\varphi \in C_c^\infty(\mathbb{R}^d)$. Let b be a bounded continuous function on $\mathbb{R}^d$, and let $\rho > 0$ and $\varepsilon > 0$. With the cutoff θ of (4.11), which satisfies $\mathbf{1}_{\{|y| \leqslant \rho\}} \leqslant \theta(y/\rho) \leqslant 1$, choose a mollification $\varphi \in C_c^\infty(\mathbb{R}^d)$ of $b\,\theta(\cdot/\rho)$ with $\|b\,\theta(\cdot/\rho) - \varphi\|_\infty \leqslant \varepsilon$. The split

$$b = b\,\big(1 - \theta(\cdot/\rho)\big) + \big(b\,\theta(\cdot/\rho) - \varphi\big) + \varphi, \qquad \Big|b\,\big(1 - \theta(\cdot/\rho)\big)\Big| \leqslant \|b\|_\infty\,\mathbf{1}_{\{|y| > \rho\}},$$

taken under μ_h and under π, gives

$$|\mu_h(b) - \pi(b)| \leqslant 2\varepsilon + |\mu_h(\varphi) - \pi(\varphi)| + \|b\|_\infty\big(\mu_h(|y| > \rho) + \pi(|y| > \rho)\big),$$

and (E.13), with $\mu_h(\varphi) \to \pi(\varphi)$, gives

$$\limsup_{h \to 0} |\mu_h(b) - \pi(b)| \leqslant 2\varepsilon + \|b\|_\infty\big(C\rho^{-12} + \pi(|y| > \rho)\big),$$

which tends to 0 as $\varepsilon \to 0$ and $\rho \to \infty$. So $\mu_h \to \pi$ weakly.

Second, $\nabla^2\tilde{\phi}_h \to \nabla^2\phi_0$ uniformly on compact sets as $h \to 0$. Let

$$\psi_h(x) := \tilde{\phi}_h(x) - \tilde{\phi}_h(0), \qquad x \in \mathbb{R}^d, \tag{E.14}$$

so that $\psi_h(0) = 0$. By (4.19), the derivatives of ψ_h of orders 1 to 4 are bounded by CL uniformly in h; hence, for $x, x' \in \mathbb{R}^d$ and $j = 0, 1, 2, 3$,

$$\|\nabla^j\psi_h(x) - \nabla^j\psi_h(x')\| \leqslant CL|x - x'|, \qquad |\psi_h(x)| \leqslant CL|x|,$$

so, on every ball $\{|x| \leqslant \rho\}$ and uniformly in h,

$$\|\nabla^j\psi_h(x)\| \leqslant CL(1 + \rho), \qquad j = 0, 1, 2, 3.$$

The Arzelà–Ascoli theorem gives, for every sequence $h_n \to 0$, a subsequence, still written h_n, along which each $\nabla^j\psi_{h_n}$ converges uniformly on compact sets. Since the uniform limit of derivatives is the derivative of the limit, the limits of $\nabla^j\psi_{h_n}$ are $\nabla^j\psi$ for a function $\psi \in C^3(\mathbb{R}^d)$ with $\nabla\psi$, $\nabla^2\psi$, $\nabla^3\psi$ bounded by CL, and the uniform convergence on compact sets reads

$$\lim_{n \to \infty} \sup_{|x| \leqslant \rho} \Big\{ \max_{0 \leqslant j \leqslant 3} \|\nabla^j\psi_{h_n}(x) - \nabla^j\psi(x)\| \Big\} = 0, \qquad \forall \rho > 0. \tag{E.15}$$

To identify the limit ψ, fix a deterministic initial point $Y_0 \in \mathbb{R}^d$. By the modified discrete Poisson equation (4.18) at Y_0, with $(e^{h\tilde{\mathcal{L}}_h}\tilde{\phi}_h)(Y_0) = \mathbb{E}[\tilde{\phi}_h(\tilde{Y}_0(h))]$,

$$\mathbb{E}\big[\tilde{\phi}_h(\tilde{Y}_0(h))\big] - \tilde{\phi}_h(Y_0) = -h\big(\tilde{f}_h(Y_0) - \tilde{\pi}_h(\tilde{f}_h)\big). \tag{E.16}$$

Since $\tilde{Y}_0(h)$ is driven by the modified dynamics (4.16) from Y_0, with the generator $\tilde{\mathcal{L}}_h$ of (4.17), Itô's formula for $\tilde{\phi}_h$, whose stochastic integral is a martingale since $\nabla\tilde{\phi}_h$ is bounded, expands the difference on the left side of (E.16) as

$$\mathbb{E}\big[\tilde{\phi}_h(\tilde{Y}_0(h))\big] - \tilde{\phi}_h(Y_0) = \int_0^h \mathbb{E}\big[\tilde{g}_h(\tilde{Y}_0(s))\big]\mathrm{d}s, \tag{E.17}$$

where, with ψ_h of (E.14),

$$\tilde{g}_h := \tilde{\mathcal{L}}_h\tilde{\phi}_h = \Delta\psi_h - \nabla\tilde{U}_h \cdot \nabla\psi_h. \tag{E.18}$$

Compared with the modified discrete Poisson equation (4.18), $-\tilde{g}_h$ is the counterpart of $\tilde{f}_h - \tilde{\pi}_h(\tilde{f}_h)$, with the difference quotient $(\tilde{\phi}_h - e^{h\tilde{\mathcal{L}}_h}\tilde{\phi}_h)/h$ replaced by $-\tilde{\mathcal{L}}_h\tilde{\phi}_h$. Next we bound $\tilde{g}_h(Y_0) + \tilde{f}_h(Y_0) - \tilde{\pi}_h(\tilde{f}_h)$. Differentiating $\tilde{g}_h$ in (E.18),

$$\nabla\tilde{g}_h = \nabla\Delta\psi_h - \nabla^2\tilde{U}_h\,\nabla\psi_h - \nabla^2\psi_h\,\nabla\tilde{U}_h,$$

and (4.13) gives

$$\|\nabla^2\tilde{U}_h(z)\| \leqslant C, \qquad |\nabla\tilde{U}_h(z)| \leqslant C(1+|z|), \qquad z \in \mathbb{R}^d.$$

Hence the bounds of (4.19) on the derivatives of ψ_h of orders 1 to 3 give, for $z \in \mathbb{R}^d$ and $0 \leqslant s \leqslant h$,

$$|\nabla\tilde{g}_h(z)| \leqslant CL(1+|z|), \qquad |\tilde{g}_h(\tilde{Y}_0(s)) - \tilde{g}_h(Y_0)| \leqslant CL\big(1+|Y_0|+|\tilde{Y}_0(s)-Y_0|\big)|\tilde{Y}_0(s)-Y_0|.$$

The bound on $\tilde{Y}_0(s) - Y_0$ in (C.1) gives, for $0 \leqslant s \leqslant h \leqslant 1$,

$$\mathbb{E}|\tilde{Y}_0(s)-Y_0| \leqslant C\sqrt{h}(1+|Y_0|), \qquad \mathbb{E}|\tilde{Y}_0(s)-Y_0|^2 \leqslant Ch(1+|Y_0|)^2.$$

Taking the expectation of the bound on $|\tilde{g}_h(\tilde{Y}_0(s)) - \tilde{g}_h(Y_0)|$ and inserting the bounds on $\mathbb{E}|\tilde{Y}_0(s)-Y_0|$ and $\mathbb{E}|\tilde{Y}_0(s)-Y_0|^2$ gives

$$\Big|\mathbb{E}\big[\tilde{g}_h(\tilde{Y}_0(s))\big] - \tilde{g}_h(Y_0)\Big| \leqslant CL(1+|Y_0|)\,\mathbb{E}|\tilde{Y}_0(s)-Y_0| + CL\,\mathbb{E}|\tilde{Y}_0(s)-Y_0|^2 \leqslant CL\sqrt{h}\,(1+|Y_0|)^2.$$

By (E.16) and (E.17), averaging the bound on $\mathbb{E}\big[\tilde{g}_h(\tilde{Y}_0(s))\big] - \tilde{g}_h(Y_0)$ over $s \in [0,h]$ gives

$$\big|\tilde{g}_h(Y_0) + \tilde{f}_h(Y_0) - \tilde{\pi}_h(\tilde{f}_h)\big| = \left|\frac{1}{h}\int_0^h \Big(\mathbb{E}\big[\tilde{g}_h(\tilde{Y}_0(s))\big] - \tilde{g}_h(Y_0)\Big)\mathrm{d}s\right| \leqslant CL\sqrt{h}\,(1+|Y_0|)^2, \quad Y_0 \in \mathbb{R}^d. \tag{E.19}$$

To pass to the limit in (E.19) along h_n, we write, with $\mathcal{L}$ as in (3.3) and $\tilde{g}_h$ as in (E.18),

$$\tilde{g}_{h_n} - \mathcal{L}\psi = \big(\Delta\psi_{h_n} - \Delta\psi\big) - \big(\nabla\tilde{U}_{h_n} - \nabla U\big)\cdot\nabla\psi_{h_n} - \nabla U\cdot\big(\nabla\psi_{h_n} - \nabla\psi\big), \tag{E.20}$$

and, for every $x \in \mathbb{R}^d$, with $|\nabla\psi_{h_n}| \leqslant CL$ by (4.19), (E.20) gives

$$\big|\tilde{g}_{h_n}(x) - \mathcal{L}\psi(x)\big| \leqslant \big|\Delta\psi_{h_n}(x) - \Delta\psi(x)\big| + CL\,\big|\nabla\tilde{U}_{h_n}(x) - \nabla U(x)\big| + |\nabla U(x)|\,\big|\nabla\psi_{h_n}(x) - \nabla\psi(x)\big|. \tag{E.21}$$

On the right side of (E.21), the terms in $\psi_{h_n} - \psi$ vanish by (E.15) at $j=1$ and $j=2$, and the term in $\nabla\tilde{U}_{h_n} - \nabla U$ vanishes by (4.14):

$$\lim_{n\to\infty}\big|\Delta\psi_{h_n}(x) - \Delta\psi(x)\big| \leqslant d\lim_{n\to\infty}\|\nabla^2\psi_{h_n}(x) - \nabla^2\psi(x)\| = 0,$$
$$\lim_{n\to\infty}\big|\nabla\psi_{h_n}(x) - \nabla\psi(x)\big| = 0, \qquad \lim_{n\to\infty}\big|\nabla\tilde{U}_{h_n}(x) - \nabla U(x)\big| \leqslant \lim_{n\to\infty} Ch_n(1+|x|) = 0.$$

Hence $\lim_{n\to\infty}\tilde{g}_{h_n}(x) = \mathcal{L}\psi(x)$, and (4.2) and (4.15) give

$$\lim_{n\to\infty}\big(\tilde{f}_{h_n}(x) - \tilde{\pi}_{h_n}(\tilde{f}_{h_n})\big) = f(x) - \pi(f).$$

Letting $n \to \infty$ in (E.19) at $h = h_n$ shows that ψ solves the continuous Poisson equation (3.7).

It remains to show that $\nabla^2\psi = \nabla^2\phi_0$. The continuous Poisson equation (3.7) has the solution

$$\phi_0 = \int_0^\infty u(\cdot, t)\,\mathrm{d}t.$$

The integral is well defined: by (3.8), $u(\cdot,t)$ decays like $e^{-\lambda t}$ uniformly on compact sets, and its derivatives of orders 1 to 4 decay like $e^{-\lambda t}$ uniformly on $\mathbb{R}^d$. Hence

$$\|\nabla^j \phi_0\|_\infty \leqslant CL, \qquad j = 1, \dots, 4, \tag{E.22}$$

and, since $\partial_t u = \mathcal{L}u$ and $u(\cdot,0) = f - \pi(f)$, the integral solves (3.7). The difference

$$w := \psi - \phi_0 \in C^3(\mathbb{R}^d) \qquad \text{satisfies} \qquad \mathcal{L}w = 0, \qquad \|\nabla w\|_\infty < \infty.$$

Since $\mathcal{L}w = 0$ and ∇w is bounded, Itô's formula gives $(e^{t\mathcal{L}}w)(x) = \mathbb{E}^x[w(X_t)] = w(x)$ for $t \geqslant 0$. The contraction of [106, Lemma 5], which holds under Assumption 1 and requires of its initial function only a bounded gradient, gives

$$\|\nabla w\|_\infty = \|\nabla(e^{t\mathcal{L}}w)\|_\infty \leqslant Ce^{-\lambda t}\|\nabla w\|_\infty \to 0, \qquad t \to \infty.$$

So w is constant and $\nabla^2\psi = \nabla^2\phi_0$. The argument for w uses only $\mathcal{L}w = 0$ and the bounded gradient of w, so (3.7) determines ϕ_0 up to an additive constant among the functions with a bounded gradient, as stated in Section 3.3. Since $\nabla^2\psi_h = \nabla^2\tilde{\phi}_h$ and $\nabla^2\psi = \nabla^2\phi_0$, (E.15) at $j = 2$ shows that every sequence $h_n \to 0$ has a subsequence along which

$$\lim_{n\to\infty} \sup_{|x|\leqslant\rho} \|\nabla^2\tilde{\phi}_{h_n}(x) - \nabla^2\phi_0(x)\| = 0, \qquad \forall \rho > 0,$$

and hence $\nabla^2\tilde{\phi}_h \to \nabla^2\phi_0$ uniformly on compact sets as $h \to 0$.

Third, $\mu_h(A_h) \to \pi(A_0)$, with A_h of (E.12) and

$$A_0(y) := \nabla^2\phi_0(y) : \mathcal{C}(y), \qquad y \in \mathbb{R}^d. \tag{E.23}$$

By (C.26) and the definition of $\tilde{\mathcal{C}}_h$ in (E.12), for every $y \in \mathbb{R}^d$,

$$\|\tilde{\mathcal{C}}_h(y) - \mathcal{C}(y)\| \leqslant \mathbb{E}\left\|\mathcal{C}\Big(y + \sqrt{\frac{h}{2}}\,\chi\Big) - \mathcal{C}(y)\right\| \leqslant 4M^2\sqrt{\frac{h}{2}}\,\mathbb{E}|\chi|. \tag{E.24}$$

By (E.12) and (E.23),

$$A_h(y) - A_0(y) = \nabla^2\tilde{\phi}_h(y) : \big(\tilde{\mathcal{C}}_h(y) - \mathcal{C}(y)\big) + \big(\nabla^2\tilde{\phi}_h(y) - \nabla^2\phi_0(y)\big) : \mathcal{C}(y). \tag{E.25}$$

For the term carrying $\tilde{\mathcal{C}}_h(y) - \mathcal{C}(y)$ in (E.25), $|A : B| \leqslant d\|A\|\|B\|$, (4.19), (E.24) and $\mathbb{E}|\chi| \leqslant \sqrt{d}$ give

$$\Big|\nabla^2\tilde{\phi}_h(y) : \big(\tilde{\mathcal{C}}_h(y) - \mathcal{C}(y)\big)\Big| \leqslant d\,\|\nabla^2\tilde{\phi}_h\|_\infty \|\tilde{\mathcal{C}}_h(y) - \mathcal{C}(y)\| \leqslant d \cdot CL \cdot 4M^2\sqrt{\frac{h}{2}}\,\sqrt{d} \leqslant CL\sqrt{h},$$

and for the term carrying $\nabla^2\tilde{\phi}_h(y) - \nabla^2\phi_0(y)$, (C.27) gives

$$\Big|\big(\nabla^2\tilde{\phi}_h(y) - \nabla^2\phi_0(y)\big) : \mathcal{C}(y)\Big| \leqslant \Sigma\,\|\nabla^2\tilde{\phi}_h(y) - \nabla^2\phi_0(y)\|.$$

Hence

$$|A_h(y) - A_0(y)| \leqslant CL\sqrt{h} + \Sigma\,\|\nabla^2\tilde{\phi}_h(y) - \nabla^2\phi_0(y)\|, \qquad y \in \mathbb{R}^d. \tag{E.26}$$

With the uniform convergence of $\nabla^2\tilde{\phi}_h$ to $\nabla^2\phi_0$ on compact sets, (E.26) gives $A_h \to A_0$ uniformly on compact sets. Next we bound A_h and A_0 uniformly in y and h. Since $A : \tilde{\mathcal{C}}_h(y) = \mathbb{E}\big[A : \mathcal{C}(y + \sqrt{h/2}\,\chi)\big]$ by (E.12), (C.27) gives

$$|A : \tilde{\mathcal{C}}_h(y)| \leqslant \Sigma\|A\|, \qquad A \in \mathbb{R}^{d\times d},\ y \in \mathbb{R}^d.$$

With (4.19) and (E.22), the bound on $A : \tilde{\mathcal{C}}_h(y)$ and (C.27) give

$$|A_h(y)| + |A_0(y)| \leqslant \Sigma\big(\|\nabla^2 \tilde{\phi}_h\|_\infty + \|\nabla^2 \phi_0\|_\infty\big) \leqslant CL\Sigma, \qquad y \in \mathbb{R}^d,$$

and A_0 is continuous, since $\nabla^2\phi_0$ is continuous and $x \mapsto \mathcal{C}(x)$ is Lipschitz by (C.26). For every $\rho > 0$,

$$|\mu_h(A_h) - \pi(A_0)| \leqslant \Big|\mu_h\big((A_h - A_0)\,\mathbf{1}_{\{|y|\leqslant\rho\}}\big)\Big| + \Big|\mu_h\big((A_h - A_0)\,\mathbf{1}_{\{|y|>\rho\}}\big)\Big| + |\mu_h(A_0) - \pi(A_0)|,$$

and the bound $|A_h| + |A_0| \leqslant CL\Sigma$ on the term carrying $\mathbf{1}_{\{|y|>\rho\}}$, followed by (E.13), gives

$$\begin{aligned} |\mu_h(A_h) - \pi(A_0)| &\leqslant \sup_{|y|\leqslant\rho} |A_h(y) - A_0(y)| + CL\Sigma\,\mu_h\big(|y| > \rho\big) + |\mu_h(A_0) - \pi(A_0)| \\ &\leqslant \sup_{|y|\leqslant\rho} |A_h(y) - A_0(y)| + CL\Sigma\,\rho^{-12} + |\mu_h(A_0) - \pi(A_0)|. \end{aligned}$$

As $h \to 0$, the uniform convergence of A_h to A_0 on compact sets and the weak convergence $\mu_h \to \pi$, applied to the bounded continuous A_0, give

$$\sup_{|y|\leqslant\rho} |A_h(y) - A_0(y)| \to 0, \qquad |\mu_h(A_0) - \pi(A_0)| \to 0,$$

so, with $a_h = \mu_h(A_h)$ by (E.11) and $\pi(A_0) = \pi\big(\nabla^2\phi_0(x) : \mathcal{C}(x)\big)$ by (E.23),

$$\limsup_{h\to 0} \Big|a_h - \pi\big(\nabla^2\phi_0(x) : \mathcal{C}(x)\big)\Big| \leqslant CL\Sigma\,\rho^{-12}, \qquad \rho > 0.$$

Letting $\rho \to \infty$ gives $a_h \to \pi\big(\nabla^2\phi_0(x) : \mathcal{C}(x)\big)$, which proves (6.7). □

F Proofs for the numerical bias comparison with SGLD

Throughout this appendix, $(Z_k)_{k\geqslant 0}$ is the chain (SGLD) and $\mathcal{P}_h$ its transition kernel. Since $(\chi_k)_{k\geqslant 1}$ is independent of $(\xi_k)_{k\geqslant 0}$, the law of the chain is unchanged if its Gaussian increments are taken as $\sqrt{2h}\,\chi_{k+1} = \sqrt{2}\,(B_{(k+1)h} - B_{kh})$, with $(B_t)_{t\geqslant 0}$ a Brownian motion independent of $(\xi_k)_{k\geqslant 0}$; then, for a given Z_0, Z_k is $\hat{\mathcal{F}}_k$-measurable, with $\hat{\mathcal{F}}_k$ and $\hat{\mathbb{E}}_k$ of (3.2). Statements about a single step are made at the index 0, with the solution X_s of (1.1) from $X_0 = Z_0$, driven by the same Brownian motion. Then

$$Z_1 = Z_0 + D_0 - h\,\eta_0, \qquad D_0 := \sqrt{2}\,B_h - h\,\nabla U(Z_0), \qquad \eta_0 := \eta(Z_0;\xi_0) = \nabla\hat{U}(Z_0;\xi_0) - \nabla U(Z_0), \quad \text{(F.1)}$$

where $Z_0 + D_0$ is the state after one step of EM from Z_0, and η is the gradient noise of Appendix C.5.

F.1 Invariant distribution of SGLD

Lemma 20 *Let Assumptions 1 and 3 hold. For every $h \in (0, h_0]$, (SGLD) has a unique invariant distribution $\bar{\pi}_h$, with $\bar{\pi}_h(|x|^{12}) \leqslant C$. For every f satisfying Assumption 2 and every given $Z_0 \in \mathbb{R}^d$,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} f(Z_k) = \bar{\pi}_h(f) \qquad \textit{almost surely.}$$

Proof. The proof follows that of Lemma 8 in Appendix E.1, and we indicate the changes only. The step of (SGLD) from Z_k is $\sqrt{2h}\,\chi_{k+1} - h\,\nabla\hat{U}(Z_k;\xi_k)$, with the stochastic gradient evaluated at Z_k itself. The proof of Lemma 3 for (SGLM') therefore applies with Z_k and χ_{k+1} in place of Y_k and χ_k, and with the Gaussian shift $\sqrt{h/2}\,\chi_k$ replaced by 0; in (C.9) the drift bound in (B.1) applies at Z_k directly. For $q \leqslant 6$ that proof gives (C.10) and the bound on $\mathbb{E}|Y_k|^{12}$ in (5.4) for Z_k,

$$\hat{\mathbb{E}}_k|Z_{k+1}|^{2q} \leqslant \Big(1 - \frac{qm}{16}h\Big)|Z_k|^{2q} + Ch, \qquad \mathbb{E}|Z_k|^{12} \leqslant C\big(1 + |Z_0|\big)^{12},$$

and, as in Appendix E.1, the drift condition of Harris's ergodic theorem with $W(x) = |x|^2$ and a level R_W depending only on m, M, R and d.

The minorization is simpler than in Appendix E.1 and does not use the Hessian bound of Assumption 3, since the step from Z_0 is Gaussian given ξ_0,

$$Z_1 \sim \mathcal{N}\Big(Z_0 - h\,\nabla\hat{U}(Z_0;\xi_0),\, 2hI_d\Big).$$

Let h_0 also satisfy $4Mh_0 \leqslant 1$, let $|Z_0|^2 \leqslant R_W$, and let

$$E_0 := \Big\{\xi : |\nabla\hat{U}(Z_0;\xi)| \leqslant 4M\big(1+|Z_0|\big)\Big\},$$

so that Markov's inequality and the moment bound of Assumption 3 give $\mathbb{P}_\xi(E_0^c) \leqslant 4^{-12} \leqslant \frac{1}{2}$. For $\xi_0 \in E_0$ and x in the closed unit ball $\mathbb{B}$, since $4Mh \leqslant 1$,

$$\big|x - Z_0 + h\,\nabla\hat{U}(Z_0;\xi_0)\big| \leqslant 2\big(1+R_W^{\frac{1}{2}}\big),$$

and the Gaussian density of Z_1 given ξ_0, with ν the uniform distribution on $\mathbb{B}$ and $\mathbb{P}_\xi(E_0) \geqslant \frac{1}{2}$, gives for every Borel set A

$$\mathcal{P}_h(Z_0,A) \geqslant \frac{1}{2}\,(4\pi h)^{-\frac{d}{2}}\exp\Big(-\frac{(1+R_W^{\frac{1}{2}})^2}{h}\Big)\mathrm{Leb}(\mathbb{B})\,\nu(A), \qquad |Z_0|^2 \leqslant R_W,$$

which is the minorization, with α_h the minimum of $\frac{1}{2}$ and the factor of $\nu(A)$.

Harris's ergodic theorem then gives the unique invariant distribution $\bar{\pi}_h$ and (E.2) with $\bar{\pi}_h$ in place of μ_h. The moment bound follows from the truncations $\min\big\{|x|^{12},\ell\big\}$, and the almost sure limit from the Harris recurrence and the strong law of large numbers [79, Theorem 17.0.1(i)] applied to $g = f$, which is Lipschitz by Assumption 2. Both arguments are those of Appendix E.1, with the bound on $\mathbb{E}|Z_k|^{12}$ above, at $Z_0 = 0$, in place of Lemma 3. The passage from (SGLM') to the samples of (SGLM) is not needed, since $(Z_k)_{k\geqslant 0}$ is itself the chain. □

F.2 Proof of Theorem 5: numerical bias of SGLD

The proof uses the local weak error of (SGLD) against (1.1), expanded to the order h^2 with explicit coefficients. The following lemma is the counterpart of Lemma 7 for (SGLM'), with (1.1) in place of the modified dynamics (4.16), and its proof follows the same steps.

Lemma 21 *Let Assumptions 1 and 3 hold and let $G \in C^4(\mathbb{R}^d)$ have bounded derivatives of orders 1 to 4. Then, for every $h \in (0,1]$, the chain (SGLD) satisfies, for every $k \geqslant 0$,*

$$\begin{aligned}&\Big|\hat{\mathbb{E}}_k\big[G(Z_{k+1})\big] - (e^{h\mathcal{L}}G)(Z_k) - \frac{h^2}{2}\,\nabla^2 G(Z_k) : \big(2\nabla^2 U(Z_k) + \mathcal{C}(Z_k)\big)\\ &+ \frac{h^2}{2}\,\nabla G(Z_k)\cdot\big(\nabla^2 U(Z_k)\,\nabla U(Z_k) - \nabla\Delta U(Z_k)\big)\Big| \leqslant Ch^3\sum_{j=1}^{4}\|\nabla^j G\|_\infty\big(1+|Z_k|\big)^6.\end{aligned} \tag{F.2}$$

Proof. By time homogeneity, it suffices to take $k = 0$. The proof expands $G(Z_1)$ about $Z_0 + D_0$ to the order 2 in $h\eta_0$, replaces $\nabla^2 G(Z_0 + D_0)$ by $\nabla^2 G(Z_0)$ in the term of order h^2, and compares $G(Z_0 + D_0)$ with $G(X_h)$ through the moments of D_0 and $X_h - Z_0$. With $\hat{\mathcal{H}}_0 = \sigma\big(\hat{\mathcal{F}}_0, (B_s)_{s\leqslant h}\big)$ as in the proof of Lemma 7, Z_0, the increment D_0 of (F.1) and the solution X_h are $\hat{\mathcal{H}}_0$-measurable, and ξ_0 is independent of $\hat{\mathcal{H}}_0$, so Assumption 3 and (5.8) give

$$\mathbb{E}\big[\eta_0 \,\big|\, \hat{\mathcal{H}}_0\big] = 0, \qquad \mathbb{E}\big[\eta_0\otimes\eta_0 \,\big|\, \hat{\mathcal{H}}_0\big] = \mathcal{C}(Z_0).$$

Taylor's formula for G at $Z_0 + D_0$, to the order 2 in $h\eta_0$, as in the proof of Lemma 7, and the bound $\hat{\mathbb{E}}_0|\eta_0|^3 \leqslant C\big(1+|Z_0|\big)^3$ from Assumption 3 give

$$\Big|\hat{\mathbb{E}}_0\big[G(Z_1) - G(Z_0+D_0)\big] - \frac{h^2}{2}\,\hat{\mathbb{E}}_0\big[\nabla^2 G(Z_0+D_0)\big] : \mathcal{C}(Z_0)\Big| \leqslant Ch^3\|\nabla^3 G\|_\infty\big(1+|Z_0|\big)^3. \tag{F.3}$$

Next, $\nabla^2 G(Z_0+D_0)$ is replaced by $\nabla^2 G(Z_0)$ in (F.3) at the cost Ch^3. By Taylor's formula,

$$\big\|\nabla^2 G(Z_0+D_0)-\nabla^2 G(Z_0)-\nabla^3 G(Z_0)[D_0]\big\| \leqslant \frac{1}{2}\|\nabla^4 G\|_\infty |D_0|^2,$$

and, since $\nabla^3 G(Z_0)$ and $\mathcal{C}(Z_0)$ are $\hat{\mathcal{F}}_0$-measurable while B_h has mean 0 and is independent of $\hat{\mathcal{F}}_0$,

$$\hat{\mathbb{E}}_0\big[\nabla^3 G(Z_0)[D_0] : \mathcal{C}(Z_0)\big] = -h\,\nabla^3 G(Z_0)[\nabla U(Z_0)] : \mathcal{C}(Z_0).$$

With $|\nabla U(Z_0)| \leqslant M\big(1+|Z_0|\big)$ and $\hat{\mathbb{E}}_0|D_0|^2 \leqslant Ch\big(1+|Z_0|\big)^2$, (C.27) and $\Sigma \leqslant M^2$ give

$$\begin{aligned}\frac{h^2}{2}\Big|\hat{\mathbb{E}}_0\big[\nabla^2 G(Z_0+D_0)-\nabla^2 G(Z_0)\big] : \mathcal{C}(Z_0)\Big| &\leqslant \frac{\Sigma h^2}{2}\Big(h\|\nabla^3 G\|_\infty |\nabla U(Z_0)| + \frac{1}{2}\|\nabla^4 G\|_\infty\,\hat{\mathbb{E}}_0|D_0|^2\Big)\\ &\leqslant Ch^3\big(\|\nabla^3 G\|_\infty + \|\nabla^4 G\|_\infty\big)\big(1+|Z_0|\big)^2. \end{aligned} \tag{F.4}$$

The state Z_0+D_0 is compared with X_h as the state Y_1 after one step of (LM') is compared with $\tilde{Y}_0(h)$ in the proof of Lemma 6. Since $X_h = Z_0 + \sqrt{2}\,B_h - \int_0^h \nabla U(X_s)\,\mathrm{d}s$, the local error is

$$Z_0+D_0-X_h = \int_0^h \big(\nabla U(X_s)-\nabla U(Z_0)\big)\mathrm{d}s,$$

and Assumption 1 with the bound on $X_s - Z_0$ in (C.1) for (1.1) give, with $|||\cdot|||_4$ taken under $\hat{\mathbb{E}}_0$,

$$|||Z_0+D_0-X_h|||_4 \leqslant Ch^{\frac{3}{2}}\big(1+|Z_0|\big), \qquad |||D_0|||_4 + |||X_h-Z_0|||_4 \leqslant C\sqrt{h}\big(1+|Z_0|\big). \tag{F.5}$$

With Z_0, Z_0+D_0, X_h and X_h-Z_0 in place of Y_0, Y_1, $\tilde{Y}_0(h)$ and $\tilde{W}_h$, the expansion (C.22) reads

$$G(Z_0+D_0)-G(X_h) = \sum_{j=1}^{3}\frac{1}{j!}\,\nabla^j G(Z_0) : \Big(D_0^{\otimes j} - (X_h-Z_0)^{\otimes j}\Big) + \varrho_4^{D_0}, \tag{F.6}$$

with a remainder $\varrho_4^{D_0}$ that satisfies

$$|\varrho_4^{D_0}| \leqslant \frac{1}{6}\,\|\nabla^4 G\|_\infty\,\big|Z_0+D_0-X_h\big|\,\max\big\{|D_0|,\,|X_h-Z_0|\big\}^3. \tag{F.7}$$

Taking $\hat{\mathbb{E}}_0$ in (F.6), Hölder's inequality with (F.7) and (F.5) gives $\hat{\mathbb{E}}_0|\varrho_4^{D_0}| \leqslant Ch^3\|\nabla^4 G\|_\infty\big(1+|Z_0|\big)^4$, so

$$\Big|\hat{\mathbb{E}}_0\big[G(Z_0+D_0)-G(X_h)\big] - \sum_{j=1}^{3}\frac{1}{j!}\,\nabla^j G(Z_0) : \hat{\mathbb{E}}_0\Big[D_0^{\otimes j}-(X_h-Z_0)^{\otimes j}\Big]\Big| \leqslant Ch^3\|\nabla^4 G\|_\infty\big(1+|Z_0|\big)^4. \tag{F.8}$$

The moments of D_0 in (F.8) are explicit. With all derivatives of U evaluated at Z_0 and $1 \leqslant i,j,l \leqslant d$, since $\sqrt{2}\,B_h \sim \mathcal{N}(0, 2hI_d)$ and $\nabla U(Z_0)$ is fixed,

$$\begin{aligned}\hat{\mathbb{E}}_0[D_0] = -h\,\nabla U, \qquad \hat{\mathbb{E}}_0\big[D_0^{\otimes 2}\big] &= 2h\,I_d + h^2\,\nabla U^{\otimes 2},\\ \Big(\hat{\mathbb{E}}_0\big[D_0^{\otimes 3}\big]\Big)_{ijl} &= -2h^2\big(\delta_{ij}\,\partial_l U + \delta_{il}\,\partial_j U + \delta_{jl}\,\partial_i U\big) - h^3\,\partial_i U\,\partial_j U\,\partial_l U.\end{aligned}$$

The moments of $X_h - Z_0$ are expanded in (C.2)–(C.4) of Lemma 18 for (1.1), with U, Z_0, X_h and $\hat{\mathbb{E}}_0$ in place of $\tilde{U}_h$, Y_0, $\tilde{Y}_0(h)$ and $\mathbb{E}_0$. Next we subtract them from the moments of D_0, for $j=1,2,3$.

1. For $j=1$, the terms $-h\,\nabla U$ cancel, and the term of order h^2 in (C.2) remains with the opposite sign:

$$\Big|\hat{\mathbb{E}}_0\big[D_0-(X_h-Z_0)\big] + \frac{h^2}{2}\big(\nabla^2 U\,\nabla U - \nabla\Delta U\big)\Big| \leqslant Ch^3\big(1+|Z_0|\big)^3. \tag{F.9}$$

2. For $j=2$, the terms $2h\,I_d$ and $h^2\,\nabla U^{\otimes 2}$ cancel, and the term $-2h^2\,\nabla^2 U$ in (C.3) remains with the opposite sign:
$$\Big\|\hat{\mathbb{E}}_0\big[D_0^{\otimes 2}-(X_h-Z_0)^{\otimes 2}\big]-2h^2\,\nabla^2 U\Big\| \leqslant Ch^3\big(1+|Z_0|\big)^4. \tag{F.10}$$

3. For $j=3$, the terms of order h^2 cancel, and the term $-h^3\,\nabla U^{\otimes 3}$ of $\hat{\mathbb{E}}_0\big[D_0^{\otimes 3}\big]$ has norm $h^3|\nabla U(Z_0)|^3 \leqslant Ch^3\big(1+|Z_0|\big)^3$:
$$\Big\|\hat{\mathbb{E}}_0\big[D_0^{\otimes 3}-(X_h-Z_0)^{\otimes 3}\big]\Big\| \leqslant Ch^3\big(1+|Z_0|\big)^6. \tag{F.11}$$

Inserting (F.9)–(F.11) in (F.8) gives
$$\begin{aligned}&\Big|\hat{\mathbb{E}}_0\big[G(Z_0+D_0)-G(X_h)\big]-h^2\,\nabla^2 G(Z_0):\nabla^2 U(Z_0)+\frac{h^2}{2}\,\nabla G(Z_0)\cdot\big(\nabla^2 U(Z_0)\,\nabla U(Z_0)-\nabla\Delta U(Z_0)\big)\Big|\\ &\quad\leqslant Ch^3\sum_{j=1}^{4}\|\nabla^j G\|_\infty\big(1+|Z_0|\big)^6.\end{aligned} \tag{F.12}$$

Since $(e^{h\mathcal{L}}G)(Z_0)=\hat{\mathbb{E}}_0[G(X_h)]$, the sum of (F.3), (F.4) and (F.12) proves (F.2). □

Proof of Theorem 5. Here ϕ_h is the solution (3.5) of (3.6). Let $h_0 \leqslant \lambda^{-1}$, with λ the rate of Theorem 1, so that (3.9) holds. Let $Z_0 \sim \bar\pi_h$, independent of $(B_t)_{t\geqslant 0}$ and of $(\xi_k)_{k\geqslant 0}$, so that $Z_1 \sim \bar\pi_h$, and add Z_0 to the generators of $\hat{\mathcal{F}}_k$ in (3.2), as in the proof of Theorem 4, so that $\hat{\mathcal{F}}_0=\sigma(Z_0)$. The discrete Poisson equation (3.6) at Z_0, with $(e^{h\mathcal{L}}\phi_h)(Z_0)=\hat{\mathbb{E}}_0[\phi_h(X_h)]$, gives after taking expectations
$$\bar\pi_h(f)-\pi(f)=\frac{1}{h}\,\mathbb{E}\big[\phi_h(Z_0)-\phi_h(X_h)\big]=\frac{1}{h}\,\mathbb{E}\big[\phi_h(Z_1)-\phi_h(X_h)\big], \tag{F.13}$$
where $\mathbb{E}\,\phi_h(Z_0)=\mathbb{E}\,\phi_h(Z_1)$ since $Z_1\sim\bar\pi_h$. The expectations are finite, as in the proof of Theorem 4, since ϕ_h is Lipschitz with constant CL by (3.9), $\bar\pi_h(|x|^{12})\leqslant C$ by Lemma 20, and $\mathbb{E}|X_h-Z_0|\leqslant C$ by the bound on X_s-Z_0 in (C.1) for (1.1), averaged over $Z_0\sim\bar\pi_h$. By (3.9), $G=\phi_h$ has derivatives of orders 1 to 4 bounded by CL, so (F.2) of Lemma 21 at $k=0$, averaged over $Z_0\sim\bar\pi_h$, gives
$$\begin{aligned}&\Big|\mathbb{E}\big[\phi_h(Z_1)-\phi_h(X_h)\big]-\frac{h^2}{2}\,\mathbb{E}\Big[\nabla^2\phi_h(Z_0):\big(2\nabla^2 U(Z_0)+\mathcal{C}(Z_0)\big)\Big]\\ &+\frac{h^2}{2}\,\mathbb{E}\Big[\nabla\phi_h(Z_0)\cdot\big(\nabla^2 U(Z_0)\,\nabla U(Z_0)-\nabla\Delta U(Z_0)\big)\Big]\Big|\leqslant CLh^3\,\mathbb{E}\big(1+|Z_0|\big)^6\leqslant CLh^3.\end{aligned}$$
Dividing by h and inserting (F.13) gives (6.9). Finally, (C.27), $\Sigma\leqslant M^2$, (3.9) and Assumption 1 give
$$\begin{aligned}&\Big|\mathbb{E}\Big[\nabla^2\phi_h(Z_0):\big(2\nabla^2 U(Z_0)+\mathcal{C}(Z_0)\big)\Big]\Big|\\ &+\Big|\mathbb{E}\Big[\nabla\phi_h(Z_0)\cdot\big(\nabla^2 U(Z_0)\,\nabla U(Z_0)-\nabla\Delta U(Z_0)\big)\Big]\Big|\leqslant CL\,\mathbb{E}\big(1+|Z_0|\big)\leqslant CL,\end{aligned}$$
which with (6.9) gives $|\bar\pi_h(f)-\pi(f)|\leqslant CLh$. □

F.3 Proof of Corollary 3: coefficient of the numerical bias of SGLD

Proof of Corollary 3. Let
$$\begin{aligned}\bar A_h(x)&:=\nabla^2\phi_h(x):\big(2\nabla^2 U(x)+\mathcal{C}(x)\big)-\nabla\phi_h(x)\cdot\big(\nabla^2 U(x)\,\nabla U(x)-\nabla\Delta U(x)\big),\\ \bar A_0(x)&:=\nabla^2\phi_0(x):\big(2\nabla^2 U(x)+\mathcal{C}(x)\big)-\nabla\phi_0(x)\cdot\big(\nabla^2 U(x)\,\nabla U(x)-\nabla\Delta U(x)\big),\end{aligned} \tag{F.14}$$
so that dividing (6.9) by h gives
$$\Big|\frac{\bar\pi_h(f)-\pi(f)}{h}-\frac{1}{2}\,\bar\pi_h(\bar A_h)\Big|\leqslant CLh.$$

It suffices to prove $\bar\pi_h(\bar A_h) \to \pi(\bar A_0)$ and to compute $\pi(\bar A_0)$.

The measures $\bar\pi_h$ converge weakly to π as $h \to 0$. For $\varphi \in C_c^\infty(\mathbb{R}^d)$, which satisfies Assumption 2 with some constant L_φ, Theorem 5 gives $|\bar\pi_h(\varphi) - \pi(\varphi)| \leqslant C L_\varphi h$, and Lemma 20 gives, uniformly in h,

$$\bar\pi_h\Big(\big(1+|x|\big)\,\mathbf{1}_{\{|x|>\rho\}}\Big) \leqslant \rho^{-11}\,\bar\pi_h\Big(\big(1+|x|\big)^{12}\Big) \leqslant C\rho^{-11}, \qquad \rho > 0. \tag{F.15}$$

The approximation of a bounded continuous b by $\varphi \in C_c^\infty(\mathbb{R}^d)$ in Appendix E.3, with $\bar\pi_h$ and (F.15) in place of μ_h and (E.13), then gives $\bar\pi_h(b) \to \pi(b)$ for every bounded continuous b.

The gradient and the Hessian of ϕ_h converge to those of ϕ_0 uniformly on compact sets. The argument for the convergence of $\nabla^2\tilde\phi_h$ in Appendix E.3, through (E.17) and (E.19), applies with U, $\mathcal{L}$, f, $\pi(f)$, ϕ_h, the solution X_s of (1.1) and its initial value X_0 in place of $\tilde U_h$, $\tilde{\mathcal{L}}_h$, $\tilde f_h$, $\tilde\pi_h(\tilde f_h)$, $\tilde\phi_h$, $\tilde Y_0(s)$ and Y_0. It then uses (3.6), (3.9) and Assumption 1 in place of (4.18), (4.19) and Lemma 1, and the differences $\nabla\tilde U_h - \nabla U$, $\tilde f_h - f$ and $\tilde\pi_h(\tilde f_h) - \pi(f)$ estimated there are 0. It gives, for every sequence $h_n \to 0$, a subsequence along which the derivatives of ϕ_{h_n} of orders 1 to 3 converge uniformly on compact sets to those of a function that differs from ϕ_0 by a constant, hence

$$\sup_{|x|\leqslant\rho}\Big(|\nabla\phi_h(x) - \nabla\phi_0(x)| + \|\nabla^2\phi_h(x) - \nabla^2\phi_0(x)\|\Big) \to 0, \qquad h \to 0, \quad \rho > 0.$$

By (F.14), (C.27), $\Sigma \leqslant M^2$ and Assumption 1,

$$|\bar A_h(x) - \bar A_0(x)| \leqslant C\|\nabla^2\phi_h(x) - \nabla^2\phi_0(x)\| + C\big(1+|x|\big)|\nabla\phi_h(x) - \nabla\phi_0(x)|,$$

so $\bar A_h \to \bar A_0$ uniformly on compact sets, and (3.9) and (E.22) give

$$|\bar A_h(x)| + |\bar A_0(x)| \leqslant CL\big(1+|x|\big), \qquad x \in \mathbb{R}^d.$$

The function $\bar A_0$ is continuous, since $\nabla\phi_0$, $\nabla^2\phi_0$ and the derivatives of U are continuous and $\mathcal{C}(x)$ is Lipschitz by (C.26). For every $\rho > 0$,

$$\big|\bar\pi_h(\bar A_h) - \pi(\bar A_0)\big| \leqslant \sup_{|x|\leqslant\rho}|\bar A_h(x) - \bar A_0(x)| + CL\,\bar\pi_h\Big(\big(1+|x|\big)\,\mathbf{1}_{\{|x|>\rho\}}\Big) + \big|\bar\pi_h(\bar A_0) - \pi(\bar A_0)\big|.$$

The bounded continuous function $\max\big\{-CL(1+\rho), \min\{\bar A_0, CL(1+\rho)\}\big\}$ equals $\bar A_0$ on $\{|x| \leqslant \rho\}$ and differs from $\bar A_0$ by at most $CL(1+|x|)\,\mathbf{1}_{\{|x|>\rho\}}$, so the weak convergence of $\bar\pi_h$ to π, (F.15) and the moments of π in Lemma 17 give $\limsup_{h\to0}|\bar\pi_h(\bar A_0) - \pi(\bar A_0)| \leqslant CL\rho^{-11}$. With (F.15) for $CL\,\bar\pi_h\big((1+|x|)\,\mathbf{1}_{\{|x|>\rho\}}\big)$,

$$\limsup_{h\to0}\big|\bar\pi_h(\bar A_h) - \pi(\bar A_0)\big| \leqslant CL\rho^{-11}, \qquad \rho > 0,$$

and letting $\rho \to \infty$ gives $\bar\pi_h(\bar A_h) \to \pi(\bar A_0)$.

It remains to compute $\pi(\bar A_0)$. Since ϕ_0 has bounded derivatives of orders 1 to 3 by (E.22), (6.12) with $\psi = \phi_0$ gives

$$\pi(\bar A_0) = \pi\Big(\nabla^2\phi_0(x) : \big(2\nabla^2 U(x) + \mathcal{C}(x)\big)\Big) - \pi\big(\nabla^2\phi_0(x) : \nabla^2 U(x)\big) = \pi\Big(\nabla^2\phi_0(x) : \big(\nabla^2 U(x) + \mathcal{C}(x)\big)\Big),$$

which with $\bar\pi_h(\bar A_h) \to \pi(\bar A_0)$ proves (6.10). □

G Proofs for the explicit coefficient in Wasserstein-1 error

Throughout this appendix, the constant C depends only on m, M, R and d. Let

$$\mathcal{D} := \big\{\omega \in C^4(\mathbb{R}^d) : \|\nabla^j\omega\|_\infty < \infty,\ j = 1,2,3,4\big\}, \qquad L_0^2(\pi) := \big\{g \in L^2(\pi) : \pi(g) = 0\big\}, \tag{G.1}$$

so $\mathcal{D}$ is the set of the test functions ω of (6.14), and $L_0^2(\pi)$ is a Hilbert space with the inner product of $L^2(\pi)$.

G.1 Proof of Theorem 6: well-posedness of the Riesz representor

Lemma 9 follows from [5]. By (B.1), U is bounded from below, and $\langle x, \nabla U(x)\rangle \geqslant \frac{m}{4}|x|^2 - C \geqslant |x|$ outside a ball whose radius depends only on m, M, R and d. Hence U satisfies the drift condition of [5, Corollary 1.6], and [5, Theorem 1.4] gives (6.13) for smooth g. In the proof of [5, Corollary 1.6], the Lyapunov function is an exponential of $|x|$ outside a larger ball, whose radius depends only on the previous radius and on d, and the constants of the inequality [5, (1.3)] that it satisfies depend on U only through ∇U on that ball. By [5, Theorem 1.4], C_P depends only on these constants and on the Poincaré constant of π restricted to the same ball, which [5, Section 1] bounds through the oscillation of U on the ball. Since $|\nabla U(x)| \leqslant M(1+|x|)$, C_P depends only on m, M, R and d. Mollification and dominated convergence, with the moments of Lemma 17, extend (6.13) to every $g \in C^1(\mathbb{R}^d)$ with a bounded gradient. The next lemma is the identity of Bochner under π, used in the proofs of Theorems 6 and 8.

Lemma 22 *Let Assumption 1 hold. For every $\omega \in \mathcal{D}$,*

$$\pi\big((\mathcal{L}\omega)^2\big) = \pi\big(\nabla^2\omega : \nabla^2\omega\big) + \pi\big(\nabla\omega \cdot \nabla^2 U\, \nabla\omega\big). \tag{G.2}$$

Proof. Differentiating (3.3) gives $\partial_i \mathcal{L}\omega = \mathcal{L}\partial_i\omega - \sum_{j=1}^d \partial_i\partial_j U\, \partial_j\omega$ for each component i. The functions ω, $\mathcal{L}\omega$ and $\partial_i\omega$ and their derivatives of orders 1 and 2 grow at most linearly, so (3.4) with $g_1 = \mathcal{L}\omega$, $g_2 = \omega$, and then with $g_1 = g_2 = \partial_i\omega$, gives

$$\begin{aligned}\pi\big((\mathcal{L}\omega)^2\big) &= -\sum_{i=1}^d \pi\big(\partial_i\omega\, \partial_i\mathcal{L}\omega\big) = -\sum_{i=1}^d \pi\big(\partial_i\omega\, \mathcal{L}\partial_i\omega\big) + \pi\big(\nabla\omega \cdot \nabla^2 U\, \nabla\omega\big)\\ &= \sum_{i=1}^d \pi\big(|\nabla\partial_i\omega|^2\big) + \pi\big(\nabla\omega \cdot \nabla^2 U\, \nabla\omega\big),\end{aligned}$$

which is (G.2). □

Proof of Theorem 6. The proof bounds $|\pi(\nabla^2\omega : \mathcal{A})|$ by the norm of $\mathcal{L}\omega$ in $L^2(\pi)$, through (6.13) and (G.2), and applies the Riesz representation theorem on the dense set of the functions $-\mathcal{L}\omega$, $\omega \in \mathcal{D}$, in $L^2_0(\pi)$. Let $\omega \in \mathcal{D}$. Since ω grows at most linearly, (3.4) and (6.13) give

$$\pi\big(|\nabla\omega|^2\big) = \pi\big((\omega - \pi(\omega))(-\mathcal{L}\omega)\big) \leqslant C_P^{\frac12}\, \pi\big(|\nabla\omega|^2\big)^{\frac12}\, \pi\big((\mathcal{L}\omega)^2\big)^{\frac12},$$

so $\pi(|\nabla\omega|^2) \leqslant C_P\, \pi((\mathcal{L}\omega)^2)$, and (G.2) with $\nabla^2 U \succeq -MI_d$ gives

$$\pi\big(\nabla^2\omega : \nabla^2\omega\big) \leqslant \pi\big((\mathcal{L}\omega)^2\big) + M\, \pi\big(|\nabla\omega|^2\big) \leqslant (1 + MC_P)\, \pi\big((\mathcal{L}\omega)^2\big). \tag{G.3}$$

Since $\mathcal{A}(x) : \mathcal{A}(x) \leqslant d\, \|\mathcal{A}(x)\|^2$, the Cauchy–Schwarz inequality under π gives

$$\big|\pi\big(\nabla^2\omega : \mathcal{A}\big)\big| \leqslant \sqrt{d}\, \|\mathcal{A}\|_\infty\, \pi\Big(\big(\nabla^2\omega : \nabla^2\omega\big)^{\frac12}\Big) \leqslant \sqrt{d}\, \|\mathcal{A}\|_\infty\, \pi\big(\nabla^2\omega : \nabla^2\omega\big)^{\frac12},$$

and inserting the bound on $\pi(\nabla^2\omega : \nabla^2\omega)$ in (G.3) gives

$$\big|\pi\big(\nabla^2\omega : \mathcal{A}\big)\big| \leqslant \sqrt{d\,(1 + MC_P)}\, \|\mathcal{A}\|_\infty\, \pi\big((\mathcal{L}\omega)^2\big)^{\frac12}, \qquad \omega \in \mathcal{D}. \tag{G.4}$$

The functions $-\mathcal{L}\omega$, $\omega \in \mathcal{D}$, lie in $L^2_0(\pi)$ by their linear growth and (3.4) with $g_1 = 1$. Every $g \in C^\infty_c(\mathbb{R}^d)$ satisfies Assumption 2, so the solution ϕ_0 of the continuous Poisson equation (3.7) with g in place of f belongs to $\mathcal{D}$ by (E.22), and $g - \pi(g) = -\mathcal{L}\phi_0$. Since the functions $g - \pi(g)$, $g \in C^\infty_c(\mathbb{R}^d)$, are dense in $L^2_0(\pi)$, so are the functions $-\mathcal{L}\omega$, $\omega \in \mathcal{D}$. By (G.4), $-\mathcal{L}\omega \mapsto \pi(\nabla^2\omega : \mathcal{A})$ is a well-defined linear functional on the functions $-\mathcal{L}\omega$, $\omega \in \mathcal{D}$, bounded in the norm of $L^2_0(\pi)$, and linear in $\mathcal{A}$. It extends to $L^2_0(\pi)$, and the Riesz representation theorem gives the unique $\psi_\mathcal{A} \in L^2_0(\pi)$ that satisfies (6.14) for every $\omega \in \mathcal{D}$. By this uniqueness and the linearity of the functional in $\mathcal{A}$, the map $\mathcal{A} \mapsto \psi_\mathcal{A}$ is linear, and (G.4) gives (6.15). □

G.2 Auxiliary estimates for Theorem 7

Throughout this subsection and the next, $h \leqslant h_0$ with $h_0 \leqslant \lambda^{-1}$, λ the rate of Lemma 23. For a signed measure ν on $\mathbb{R}^d$ with $\nu(\mathbb{R}^d) = 0$ and $\int (1+|x|)\,|\nu|(\mathrm{d}x) < \infty$, such as $\psi\pi$ with $\psi \in L^2_0(\pi)$, the Kantorovich–Rubinstein norm of Section 3.1 satisfies

$$\|\nu\|_{\mathrm{KR}} = \sup\left\{ \int_{\mathbb{R}^d} f \,\mathrm{d}\nu : f \text{ satisfies Assumption 2},\ \|\nabla f\|_\infty \leqslant 1 \right\}. \tag{G.5}$$

The supremum on the right side of (G.5) is at most $\|\nu\|_{\mathrm{KR}}$, since its functions f are Lipschitz with $\|\nabla f\|_\infty \leqslant 1$, and it is at least $\|\nu\|_{\mathrm{KR}}$, since the mollifications of a Lipschitz function g with $\|\nabla g\|_\infty \leqslant 1$ satisfy Assumption 2 with the same bound on the gradient and converge to g uniformly.

In the proofs of Theorems 7 and 8, the test function f satisfies Assumption 2 and $\|\nabla f\|_\infty \leqslant 1$, and the estimates do not depend on the constant L of Assumption 2. The modified test function (4.1) is $\tilde f_h = f * \mathcal{N}(0, \frac{h}{2} I_d)$, where $\mathcal{N}(0, \frac{h}{2} I_d)$ also denotes the density of this Gaussian distribution, so

$$\nabla^j \tilde f_h = (\nabla f) * \nabla^{j-1} \mathcal{N}\Big(0, \frac{h}{2} I_d\Big), \qquad \int_{\mathbb{R}^d} \Big\| \nabla^{j-1} \mathcal{N}\Big(0, \frac{h}{2} I_d\Big)(y) \Big\| \,\mathrm{d}y \leqslant C h^{-\frac{j-1}{2}}, \qquad j = 1,2,3,4,$$

and

$$\|\nabla \tilde f_h\|_\infty \leqslant 1, \qquad \|\nabla^j \tilde f_h\|_\infty \leqslant C h^{-\frac{j-1}{2}}, \qquad j = 2,3,4. \tag{G.6}$$

The semigroup estimates of Appendix A.2 and Lemma 16 combine into one bound for the exact and the modified dynamics. The proof technique is essentially that of [106, Lemma 10], which splits the times into $0 < t \leqslant 1$ and $t > 1$.

Lemma 23 *Let Assumption 1 hold. There is $\lambda > 0$, depending only on m, M, R and d, such that, for every $h \in (0, h_0]$, every $g \in \mathcal{D}$ of (G.1) and every $t > 0$,*

$$\|\nabla^j (e^{t\mathcal{L}} g)\|_\infty + \|\nabla^j (e^{t\tilde{\mathcal{L}}_h} g)\|_\infty \leqslant C \min(1,t)^{-\frac{j-1}{2}}\, e^{-\lambda t}\, \|\nabla g\|_\infty, \qquad j = 1,2,3,4. \tag{G.7}$$

Proof. For $j = 1$, the bound on $e^{t\mathcal{L}} g$ is (A.8). For $j = 2,3,4$ and $0 < t \leqslant 1$, (A.9) and (A.14) give

$$\|\nabla^j (e^{t\mathcal{L}} g)\|_\infty \leqslant C\, t^{-\frac{j-1}{2}}\, \|\nabla g\|_\infty \leqslant C e^{\lambda} \min(1,t)^{-\frac{j-1}{2}}\, e^{-\lambda t}\, \|\nabla g\|_\infty,$$

since $\min(1,t) = t$ and $e^{-\lambda t} \geqslant e^{-\lambda}$, which is (G.7) for $e^{t\mathcal{L}}$ with Ce^{λ} in place of C. For $j = 2,3,4$ and $t > 1$, the semigroup property gives $e^{t\mathcal{L}} g = e^{\mathcal{L}} \big(e^{(t-1)\mathcal{L}} g\big)$, and $e^{(t-1)\mathcal{L}} g \in \mathcal{D}$, since (A.12) at time $t-1$ bounds its derivatives of orders 1 to 4. Hence (A.9) and (A.14) at time 1, with $e^{(t-1)\mathcal{L}} g$ in place of g, bound $\|\nabla^j (e^{t\mathcal{L}} g)\|_\infty$ by $C \|\nabla (e^{(t-1)\mathcal{L}} g)\|_\infty$, and (A.8) at time $t-1$ bounds this gradient, so that

$$\|\nabla^j (e^{t\mathcal{L}} g)\|_\infty \leqslant C \big\|\nabla (e^{(t-1)\mathcal{L}} g)\big\|_\infty \leqslant C e^{-\lambda(t-1)} \|\nabla g\|_\infty = C e^{\lambda} e^{-\lambda t} \|\nabla g\|_\infty.$$

Since $\min(1,t) = 1$ for $t > 1$, this is (G.7) for $e^{t\mathcal{L}}$ with Ce^{λ} in place of C.

The same argument applies to $e^{t\tilde{\mathcal{L}}_h}$. Lemmas 12, 15 and 16 use only the bounds of Assumption 1 up to the order 5 and the convexity outside the ball of radius R, and by Lemma 1, for every $h \in (0, h_0]$, $\tilde U_h$ satisfies them with $2M$ and $\frac{m}{2}$ in place of M and m. These lemmas therefore hold for $e^{t\tilde{\mathcal{L}}_h}$ with a rate and a constant that depend only on m, M, R and d, as in Corollary 1. Since the left side of (G.7) is the sum of the norms for $e^{t\mathcal{L}}$ and for $e^{t\tilde{\mathcal{L}}_h}$, (G.7) holds with λ the smaller of the rates for U and for $\tilde U_h$, and C twice the larger of the constants. □

In the proofs of Theorem 7, the bounds (G.7) at the times $t = ih$ are summed over $i \geqslant 0$ through the following elementary bounds.

Lemma 24 *For every $h \in (0, h_0]$,*

$$h \sum_{i=0}^{\infty} \min\big(1, (i+1)h\big)^{-\frac{j-1}{2}} e^{-\lambda i h} \leqslant \begin{cases} C, & j = 1,2, \\ C\big(1 + \log \frac{1}{h}\big), & j = 3, \\ C h^{-\frac{1}{2}}, & j = 4. \end{cases} \tag{G.8}$$

Proof. Split the sum at $(i+1)h = 1$. The terms with $(i+1)h > 1$, that is, $i \geqslant \lfloor 1/h \rfloor$, have the minimum 1, and since $\lambda h \leqslant 1$,

$$h \sum_{i=\lfloor 1/h \rfloor}^{\infty} e^{-\lambda i h} \leqslant \frac{h}{1 - e^{-\lambda h}} \leqslant \frac{2}{\lambda}.$$

For the terms with $(i+1)h \leqslant 1$, put $n = i + 1$ and bound $e^{-\lambda i h}$ by 1:

$$h \sum_{n=1}^{\lfloor 1/h \rfloor} (nh)^{-\frac{j-1}{2}} = h^{\frac{3-j}{2}} \sum_{n=1}^{\lfloor 1/h \rfloor} n^{-\frac{j-1}{2}} \leqslant \begin{cases} 1, & j = 1, \\ 2, & j = 2, \\ 1 + \log \frac{1}{h}, & j = 3, \\ 3h^{-\frac{1}{2}}, & j = 4. \end{cases}$$

Adding the bound $2/\lambda$ gives (G.8). □

The gradient noise enters a step of (SGLD) with the factor h, but in the next lemma it contributes only at the order $h^{\frac{3}{2}}$: the Gaussian increment of the step smooths the test function, and the noise has mean 0.

Lemma 25 *Let Assumptions 1 and 3 hold. Then, for every Lipschitz function f with $\|\nabla f\|_\infty \leqslant 1$ and every $h \in (0, 1]$, the chain (SGLD) satisfies*

$$\left|\hat{\mathbb{E}}_0\big[f(Z_1)\big] - (e^{h\mathcal{L}} f)(Z_0)\right| \leqslant C h^{\frac{3}{2}} \big(1 + |Z_0|\big). \tag{G.9}$$

Proof. Write the step from Z_0 as in (F.1), with D_0, η_0 and the solution X_s of (1.1) from $X_0 = Z_0$ as in Appendix F. Then

$$\hat{\mathbb{E}}_0\big[f(Z_1)\big] - (e^{h\mathcal{L}} f)(Z_0) = \hat{\mathbb{E}}_0\big[f(Z_0 + D_0 - h\eta_0) - f(Z_0 + D_0)\big] + \hat{\mathbb{E}}_0\big[f(Z_0 + D_0) - f(X_h)\big]. \tag{G.10}$$

Since $\|\nabla f\|_\infty \leqslant 1$, the bound on $Z_0 + D_0 - X_h$ in (F.5) gives

$$\Big|\hat{\mathbb{E}}_0\big[f(Z_0 + D_0) - f(X_h)\big]\Big| \leqslant \hat{\mathbb{E}}_0 |Z_0 + D_0 - X_h| \leqslant C h^{\frac{3}{2}} \big(1 + |Z_0|\big).$$

For the expectation of $f(Z_0 + D_0 - h\eta_0) - f(Z_0 + D_0)$ in (G.10), condition on $\hat{\mathcal{F}}_0$ and η_0. Since $\sqrt{2}\, B_h \sim \mathcal{N}(0, 2hI_d)$ is independent of $\sigma(\hat{\mathcal{F}}_0, \eta_0)$,

$$\mathbb{E}\big[f(Z_0 + D_0 - h\eta_0) - f(Z_0 + D_0) \,\big|\, \hat{\mathcal{F}}_0, \eta_0\big] = (e^{h\Delta} f)\big(Z_0 - h\,\nabla U(Z_0) - h\eta_0\big) - (e^{h\Delta} f)\big(Z_0 - h\,\nabla U(Z_0)\big),$$

with $e^{h\Delta} f = f * \mathcal{N}(0, 2hI_d)$, for which $\|\nabla^2 e^{h\Delta} f\|_\infty \leqslant C h^{-\frac{1}{2}}$ as in (G.6). Taylor's formula in $h\eta_0$ gives

$$\Big|\mathbb{E}\big[f(Z_0 + D_0 - h\eta_0) - f(Z_0 + D_0) \,\big|\, \hat{\mathcal{F}}_0, \eta_0\big] + h\,\nabla(e^{h\Delta} f)\big(Z_0 - h\,\nabla U(Z_0)\big) \cdot \eta_0\Big| \leqslant C h^{\frac{3}{2}} |\eta_0|^2.$$

Since $\nabla(e^{h\Delta} f)\big(Z_0 - h\,\nabla U(Z_0)\big)$ is $\hat{\mathcal{F}}_0$-measurable and $\hat{\mathbb{E}}_0[\eta_0] = 0$, taking $\hat{\mathbb{E}}_0$ removes the term linear in η_0, and $\hat{\mathbb{E}}_0|\eta_0|^2 \leqslant \Sigma \leqslant M^2$ by Assumption 3 gives

$$\Big|\hat{\mathbb{E}}_0\big[f(Z_0 + D_0 - h\eta_0) - f(Z_0 + D_0)\big]\Big| \leqslant C h^{\frac{3}{2}}\, \hat{\mathbb{E}}_0 |\eta_0|^2 \leqslant C h^{\frac{3}{2}}.$$

With the bound $C h^{\frac{3}{2}}(1 + |Z_0|)$ on the expectation of $f(Z_0 + D_0) - f(X_h)$, (G.10) then gives (G.9). □

In the proof of (6.19), the integrand of the sum in (G.32) carries $\nabla^2 U\, \nabla U - \nabla \Delta U$, whose gradient grows linearly, so the integrand satisfies only the bound (G.11) below, and $\bar{\pi}_h$ is replaced by π in the coefficient of $\frac{h^2}{2}$ in (G.32) through the following truncation.

Lemma 26 *Let μ and ν be probability measures on $\mathbb{R}^d$ with $\mu(|x|^{12}) + \nu(|x|^{12}) < \infty$ and $\mathcal{W}_1(\mu, \nu) \leqslant 1$, and let $F : \mathbb{R}^d \to \mathbb{R}$ satisfy, for some $K \geqslant 0$,*

$$|F(x) - F(x')| \leqslant K\big(1 + |x| + |x'|\big)|x - x'|, \qquad x, x' \in \mathbb{R}^d. \tag{G.11}$$

Then $|\mu(F) - \nu(F)| \leqslant 3K\big(1 + \mu(|x|^{12}) + \nu(|x|^{12})\big)\, \mathcal{W}_1(\mu, \nu)^{\frac{10}{11}}$.

Proof. If $\mathcal{W}_1(\mu,\nu)=0$, then $\mu=\nu$. Otherwise let $\rho := \mathcal{W}_1(\mu,\nu)^{-\frac{1}{11}} \geqslant 1$, and let $T_\rho(x) := x$ for $|x| \leqslant \rho$ and $T_\rho(x) := \rho x/|x|$ for $|x| > \rho$, the retraction onto the closed ball of radius ρ, which is 1-Lipschitz. By (G.11), $F \circ T_\rho$ is Lipschitz with constant $K(1+2\rho) \leqslant 3K\rho$, so the Kantorovich–Rubinstein duality gives

$$\big|\mu(F\circ T_\rho) - \nu(F\circ T_\rho)\big| \leqslant 3K\rho\,\mathcal{W}_1(\mu,\nu).$$

For $|x|>\rho$, $|T_\rho(x)| = \rho < |x|$ and $|x - T_\rho(x)| \leqslant |x|$, so (G.11) and $|x|>\rho\geqslant 1$ give

$$|F(x) - F(T_\rho(x))| \leqslant K\big(1+2|x|\big)|x| \leqslant 3K|x|^2 \leqslant 3K\rho^{-10}|x|^{12},$$

while $F = F\circ T_\rho$ on $\{|x|\leqslant\rho\}$. Hence

$$|\mu(F)-\nu(F)| \leqslant 3K\rho\,\mathcal{W}_1(\mu,\nu) + 3K\rho^{-10}\big(\mu(|x|^{12})+\nu(|x|^{12})\big) = 3K\big(1+\mu(|x|^{12})+\nu(|x|^{12})\big)\,\mathcal{W}_1(\mu,\nu)^{\frac{10}{11}},$$

with F integrable under μ and ν since $|F(x)| \leqslant |F(0)| + K(1+|x|)|x|$. □

G.3 Proof of Theorem 7: expansions in the Kantorovich–Rubinstein norm

Since $\mathcal{C}(x)$ is positive semidefinite, $\|\mathcal{C}\|_\infty \leqslant \Sigma \leqslant M^2$ by (5.8), and $\|\nabla^2 U\|_\infty \leqslant M$. The bound (6.15) of Theorem 6, the Cauchy–Schwarz inequality and the moments of Lemma 17 then give

$$\pi\big(|\psi_{\mathcal{C}}|\,(1+|x|)^2\big) + \pi\big(|\psi_{\nabla^2 U}+\psi_{\mathcal{C}}|\,(1+|x|)^2\big) \leqslant C. \tag{G.12}$$

Proof of (6.18) in Theorem 7. Recall that μ_h is the invariant distribution of (SGLM') and $\tilde\pi_h \propto e^{-\tilde U_h}$ is the modified target, the invariant distribution of (4.16). The proof writes $\mu_h(\tilde f_h) - \tilde\pi_h(\tilde f_h)$, by the telescoping formula (G.17), as a series of local errors of Lemma 7 on the functions

$$\tilde{\mathcal{V}}_i := e^{ih\tilde{\mathcal{L}}_h}\tilde f_h, \qquad i=0,1,2,\dots, \tag{G.13}$$

moves the leading term of each to an integral against π, and sums the series through (6.14) into $\frac{h}{2}\,\pi(f\psi_{\mathcal{C}})$.

Let f satisfy Assumption 2 and $\|\nabla f\|_\infty \leqslant 1$. Let $Y_0 \sim \mu_h$, independent of $(B_t)_{t\geqslant 0}$ and $(\xi_k)_{k\geqslant 0}$, so that $Y_1 \sim \mu_h$. As in the proof of Theorem 4, add Y_0 to the generators of $\hat{\mathcal{F}}_k$ in (3.2), so that $\hat{\mathcal{F}}_0 = \sigma(Y_0)$, while χ_0 and ξ_0 remain independent of $\hat{\mathcal{F}}_0$. Then $Z_0 = Y_0 + \sqrt{h/2}\,\chi_0 \sim \pi_h$. By (E.6), $\pi_h(f) = \mu_h(\tilde f_h)$, so

$$\pi_h(f) - \pi(f) = \big(\mu_h(\tilde f_h) - \tilde\pi_h(\tilde f_h)\big) + \big(\tilde\pi_h(\tilde f_h) - \pi(f)\big), \tag{G.14}$$

and (4.15) bounds the bracket $\tilde\pi_h(\tilde f_h) - \pi(f)$ of (G.14) by Ch^2.

For $i=0$, $\tilde{\mathcal{V}}_0 = \tilde f_h$ and (G.6) bounds its derivatives, and for $i\geqslant 1$, (G.7) at $t=ih$ with $\min(1,ih) \geqslant \frac12\min(1,(i+1)h)$ bounds those of $\tilde{\mathcal{V}}_i$, so that, for $i\geqslant 0$ and $j=1,2,3,4$,

$$\|\nabla^j \tilde{\mathcal{V}}_i\|_\infty \leqslant C\min\big(1,(i+1)h\big)^{-\frac{j-1}{2}} e^{-\lambda ih}, \tag{G.15}$$

and $\tilde{\mathcal{V}}_i \in \mathcal{D}$. Since $\tilde Y_0(h)$ solves the modified dynamics (4.16) from Y_0, driven by Brownian increments independent of $\hat{\mathcal{F}}_0$, $\hat{\mathbb{E}}_0[\tilde{\mathcal{V}}_i(\tilde Y_0(h))] = \tilde{\mathcal{V}}_{i+1}(Y_0)$, and since Y_0 and Y_1 have the law μ_h,

$$\mathbb{E}\big[\tilde{\mathcal{V}}_i(Y_1) - \tilde{\mathcal{V}}_i(\tilde Y_0(h))\big] = \mu_h(\tilde{\mathcal{V}}_i) - \mu_h(\tilde{\mathcal{V}}_{i+1}).$$

The modified target $\tilde\pi_h$ is invariant under $e^{t\tilde{\mathcal{L}}_h}$ and has a moment of order 1 bounded uniformly in h, as in the proof of Lemma 2, so (G.15) at $j=1$ gives

$$\big|\tilde{\mathcal{V}}_N(y) - \tilde\pi_h(\tilde f_h)\big| \leqslant \|\nabla\tilde{\mathcal{V}}_N\|_\infty \int_{\mathbb{R}^d} |y-y'|\,\tilde\pi_h(\mathrm{d}y') \leqslant Ce^{-\lambda Nh}\big(1+|y|\big), \tag{G.16}$$

and $\mu_h(\tilde{\mathcal{V}}_N) \to \tilde\pi_h(\tilde f_h)$ as $N\to\infty$. Summing over i,

$$\mu_h(\tilde f_h) - \tilde\pi_h(\tilde f_h) = \sum_{i=0}^{\infty} \mathbb{E}\big[\tilde{\mathcal{V}}_i(Y_1) - \tilde{\mathcal{V}}_i(\tilde Y_0(h))\big]. \tag{G.17}$$

By (5.9) at $k=0$ with $G=\tilde{\mathcal{V}}_i$, the tower property for the $\hat{\mathcal{F}}_0$-measurable factor $\nabla^2\tilde{\mathcal{V}}_i(Y_0)$, and the moment bound $\mu_h(|y|^{12})\leqslant C$ of Lemma 8,

$$\Big|\mathbb{E}\big[\tilde{\mathcal{V}}_i(Y_1)-\tilde{\mathcal{V}}_i(\tilde{Y}_0(h))\big]-\frac{h^2}{2}\,\mathbb{E}\big[\nabla^2\tilde{\mathcal{V}}_i(Y_0):\mathcal{C}(Z_0)\big]\Big|\leqslant Ch^3\sum_{j=1}^{4}\|\nabla^j\tilde{\mathcal{V}}_i\|_\infty. \tag{G.18}$$

By (G.15) and (G.8), the remainders of (G.18) satisfy

$$Ch^3\sum_{i=0}^{\infty}\sum_{j=1}^{4}\|\nabla^j\tilde{\mathcal{V}}_i\|_\infty\leqslant Ch^2\Big(1+\log\frac{1}{h}+h^{-\frac12}\Big)\leqslant Ch^{\frac32}.$$

With (G.14), (4.15) and (G.17),

$$\Big|\pi_h(f)-\pi(f)-\frac{h^2}{2}\sum_{i=0}^{\infty}\mathbb{E}\big[\nabla^2\tilde{\mathcal{V}}_i(Y_0):\mathcal{C}(Z_0)\big]\Big|\leqslant Ch^{\frac32}. \tag{G.19}$$

By (C.27), (G.15) and (G.8) at $j=2$, the sum in (G.19) satisfies

$$\frac{h^2}{2}\sum_{i=0}^{\infty}\Big|\mathbb{E}\big[\nabla^2\tilde{\mathcal{V}}_i(Y_0):\mathcal{C}(Z_0)\big]\Big|\leqslant\frac{\Sigma h^2}{2}\sum_{i=0}^{\infty}\|\nabla^2\tilde{\mathcal{V}}_i\|_\infty\leqslant C\Sigma h.$$

Hence $|\pi_h(f)-\pi(f)|\leqslant C(\Sigma h+h^{\frac32})$ for every f satisfying Assumption 2 and $\|\nabla f\|_\infty\leqslant 1$, and (G.5) gives

$$\mathcal{W}_1(\pi_h,\pi)\leqslant C\big(\Sigma h+h^{\frac32}\big)\leqslant Ch. \tag{G.20}$$

In the coefficient of $\frac{h^2}{2}$ in (G.19), the expectation can be replaced by an integral against π. Since $Z_0-Y_0=\sqrt{h/2}\,\chi_0$, (C.27) gives

$$\Big|\mathbb{E}\big[\nabla^2\tilde{\mathcal{V}}_i(Y_0):\mathcal{C}(Z_0)\big]-\pi_h\big(\nabla^2\tilde{\mathcal{V}}_i:\mathcal{C}\big)\Big|=\Big|\mathbb{E}\big[\big(\nabla^2\tilde{\mathcal{V}}_i(Y_0)-\nabla^2\tilde{\mathcal{V}}_i(Z_0)\big):\mathcal{C}(Z_0)\big]\Big|\leqslant C\sqrt{h}\,\|\nabla^3\tilde{\mathcal{V}}_i\|_\infty. \tag{G.21}$$

By (C.27), (C.26) and $|A:B|\leqslant d\|A\|\|B\|$ for $A,B\in\mathbb{R}^{d\times d}$, the function $\nabla^2\tilde{\mathcal{V}}_i:\mathcal{C}$ is Lipschitz,

$$\big|\nabla^2\tilde{\mathcal{V}}_i(x):\mathcal{C}(x)-\nabla^2\tilde{\mathcal{V}}_i(x'):\mathcal{C}(x')\big|\leqslant\big(\Sigma\|\nabla^3\tilde{\mathcal{V}}_i\|_\infty+4dM^2\|\nabla^2\tilde{\mathcal{V}}_i\|_\infty\big)|x-x'|,$$

so the Kantorovich–Rubinstein duality and (G.20) give

$$\big|\pi_h(\nabla^2\tilde{\mathcal{V}}_i:\mathcal{C})-\pi(\nabla^2\tilde{\mathcal{V}}_i:\mathcal{C})\big|\leqslant Ch\big(\|\nabla^2\tilde{\mathcal{V}}_i\|_\infty+\|\nabla^3\tilde{\mathcal{V}}_i\|_\infty\big). \tag{G.22}$$

By (G.21), (G.22), (G.15) and (G.8) at $j=2,3$, the replacements of Y_0 by Z_0 and of π_h by π in the coefficient of $\frac{h^2}{2}$ in (G.19) introduce an additional error of at most

$$\begin{aligned}&\frac{h^2}{2}\sum_{i=0}^{\infty}\Big(C\sqrt{h}\,\|\nabla^3\tilde{\mathcal{V}}_i\|_\infty+Ch\big(\|\nabla^2\tilde{\mathcal{V}}_i\|_\infty+\|\nabla^3\tilde{\mathcal{V}}_i\|_\infty\big)\Big)\\&\quad\leqslant Ch^{\frac32}\Big(1+\log\frac{1}{h}\Big)+Ch^2\Big(1+\log\frac{1}{h}\Big)\leqslant Ch^{\frac32}\Big(1+\log\frac{1}{h}\Big),\end{aligned}$$

and (G.19) gives

$$\Big|\pi_h(f)-\pi(f)-\frac{h^2}{2}\sum_{i=0}^{\infty}\pi\big(\nabla^2\tilde{\mathcal{V}}_i:\mathcal{C}\big)\Big|\leqslant Ch^{\frac32}\Big(1+\log\frac{1}{h}\Big). \tag{G.23}$$

Since $\tilde{\mathcal{V}}_i\in\mathcal{D}$, (6.14) with $\mathcal{A}=\mathcal{C}$ and $\omega=\tilde{\mathcal{V}}_i$ gives

$$\pi\big(\nabla^2\tilde{\mathcal{V}}_i:\mathcal{C}\big)=\pi\big(\psi_{\mathcal{C}}\,(-\mathcal{L}\tilde{\mathcal{V}}_i)\big). \tag{G.24}$$

The generators (3.3) and (4.17) differ only in the drift, so

$$-\mathcal{L}\tilde{\mathcal{V}}_i = -\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i - \big(\nabla\tilde{U}_h - \nabla U\big)\cdot\nabla\tilde{\mathcal{V}}_i, \tag{G.25}$$

and (4.14) bounds the drift term of (G.25):

$$\big|\big(\nabla\tilde{U}_h(x) - \nabla U(x)\big)\cdot\nabla\tilde{\mathcal{V}}_i(x)\big| \leqslant Ch\big(1+|x|\big)\|\nabla\tilde{\mathcal{V}}_i\|_\infty, \qquad x\in\mathbb{R}^d.$$

Inserting (G.25) into (G.24) and using the bound on $\pi\big(|\psi_{\mathcal{C}}|\,(1+|x|)^2\big)$ in (G.12) gives

$$\Big|\pi\big(\nabla^2\tilde{\mathcal{V}}_i : \mathcal{C}\big) - \pi\big(\psi_{\mathcal{C}}\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)\big)\Big| \leqslant Ch\,\pi\big(|\psi_{\mathcal{C}}|\,(1+|x|)\big)\|\nabla\tilde{\mathcal{V}}_i\|_\infty \leqslant Ch\|\nabla\tilde{\mathcal{V}}_i\|_\infty.$$

By (G.15) and (G.8) at $j=1$, $h\sum_{i=0}^{\infty}\|\nabla\tilde{\mathcal{V}}_i\|_\infty \leqslant C$, so

$$\frac{h^2}{2}\sum_{i=0}^{\infty}\Big|\pi\big(\nabla^2\tilde{\mathcal{V}}_i : \mathcal{C}\big) - \pi\big(\psi_{\mathcal{C}}\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)\big)\Big| \leqslant Ch^3\sum_{i=0}^{\infty}\|\nabla\tilde{\mathcal{V}}_i\|_\infty \leqslant Ch^2,$$

and (G.23) gives

$$\Big|\pi_h(f) - \pi(f) - \frac{h^2}{2}\sum_{i=0}^{\infty}\pi\big(\psi_{\mathcal{C}}\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)\big)\Big| \leqslant Ch^{\frac{3}{2}}\Big(1+\log\frac{1}{h}\Big). \tag{G.26}$$

It remains to sum $h(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)$ over i in (G.26). For this sum, $Y_0\in\mathbb{R}^d$ is a deterministic initial point of $\tilde{Y}_0(s)$, in place of $Y_0\sim\mu_h$. By (E.17), with $\tilde{\mathcal{V}}_i$ and $\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i$ in place of $\tilde{\phi}_h$ and $\tilde{g}_h$, which applies since $\nabla\tilde{\mathcal{V}}_i$ is bounded, and $e^{h\tilde{\mathcal{L}}_h}\tilde{\mathcal{V}}_i = \tilde{\mathcal{V}}_{i+1}$,

$$\tilde{\mathcal{V}}_i(Y_0) - \tilde{\mathcal{V}}_{i+1}(Y_0) = \int_0^h \mathbb{E}\big[(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)(\tilde{Y}_0(s))\big]\mathrm{d}s,$$

hence

$$h\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)(Y_0) - \big(\tilde{\mathcal{V}}_i(Y_0) - \tilde{\mathcal{V}}_{i+1}(Y_0)\big) = \int_0^h \mathbb{E}\big[(\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)(\tilde{Y}_0(s)) - (\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)(Y_0)\big]\mathrm{d}s. \tag{G.27}$$

Differentiating (4.17) gives

$$\nabla(\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i) = \nabla\Delta\tilde{\mathcal{V}}_i - \nabla^2\tilde{U}_h\,\nabla\tilde{\mathcal{V}}_i - \nabla^2\tilde{\mathcal{V}}_i\,\nabla\tilde{U}_h,$$

and (4.13) gives

$$\|\nabla^2\tilde{U}_h(z)\| \leqslant 2M, \qquad |\nabla\tilde{U}_h(z)| \leqslant 2M(1+|z|), \qquad z\in\mathbb{R}^d.$$

Hence, for $z\in\mathbb{R}^d$ and $0\leqslant s\leqslant h$,

$$\big|\nabla(\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)(z)\big| \leqslant C(1+|z|)\sum_{j=1}^{3}\|\nabla^j\tilde{\mathcal{V}}_i\|_\infty,$$

$$\big|(\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)(\tilde{Y}_0(s)) - (\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)(Y_0)\big| \leqslant C\big(1+|Y_0|+|\tilde{Y}_0(s)-Y_0|\big)|\tilde{Y}_0(s)-Y_0|\sum_{j=1}^{3}\|\nabla^j\tilde{\mathcal{V}}_i\|_\infty.$$

The bound on $\tilde{Y}_0(s)-Y_0$ in (C.1) with $p=2$ gives

$$\mathbb{E}\big[(1+|Y_0|+|\tilde{Y}_0(s)-Y_0|)\,|\tilde{Y}_0(s)-Y_0|\big] \leqslant C\sqrt{h}\,(1+|Y_0|)^2, \qquad 0\leqslant s\leqslant h,$$

so integrating over $s\in[0,h]$ in (G.27) gives, for every $Y_0\in\mathbb{R}^d$,

$$\Big|h\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)(Y_0) - \big(\tilde{\mathcal{V}}_i(Y_0) - \tilde{\mathcal{V}}_{i+1}(Y_0)\big)\Big| \leqslant Ch^{\frac{3}{2}}\big(1+|Y_0|\big)^2\sum_{j=1}^{3}\|\nabla^j\tilde{\mathcal{V}}_i\|_\infty.$$

Integrating this bound in Y_0 against $|\psi_{\mathcal{C}}|\,\pi$, with (G.12), (G.15) and (G.8) at $j=1,2,3$,

$$\frac{h}{2}\sum_{i=0}^{\infty}\Big|\pi\Big(\psi_{\mathcal{C}}\,\big(h\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i) - \tilde{\mathcal{V}}_i + \tilde{\mathcal{V}}_{i+1}\big)\Big)\Big| \leqslant Ch^{\frac{3}{2}}\Big(1+\log\frac{1}{h}\Big).$$

The partial sums of the telescoping series are

$$\sum_{i=0}^{N-1} \pi\big(\psi_{\mathcal{C}}\,(\tilde{\mathcal{V}}_i - \tilde{\mathcal{V}}_{i+1})\big) = \pi\big(\psi_{\mathcal{C}}\tilde{f}_h\big) - \pi\big(\psi_{\mathcal{C}}\tilde{\mathcal{V}}_N\big),$$

and $\pi(\psi_{\mathcal{C}}) = 0$, (G.16) and (G.12) give $\pi(\psi_{\mathcal{C}}\tilde{\mathcal{V}}_N) \to 0$ as $N \to \infty$. The series of $\pi(\psi_{\mathcal{C}}\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i))$ converges absolutely: by $|\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i(x)| \leqslant C(1+|x|)(\|\nabla\tilde{\mathcal{V}}_i\|_\infty + \|\nabla^2\tilde{\mathcal{V}}_i\|_\infty)$, (G.12) and (G.15),

$$\sum_{i=0}^{\infty}\Big|\pi\big(\psi_{\mathcal{C}}\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)\big)\Big| \leqslant C\sum_{i=0}^{\infty}\big(\|\nabla\tilde{\mathcal{V}}_i\|_\infty + \|\nabla^2\tilde{\mathcal{V}}_i\|_\infty\big) \leqslant Ch^{-\frac{1}{2}}\sum_{i=0}^{\infty} e^{-\lambda ih} < \infty.$$

Therefore

$$\Big|\frac{h^2}{2}\sum_{i=0}^{\infty}\pi\big(\psi_{\mathcal{C}}\,(-\tilde{\mathcal{L}}_h\tilde{\mathcal{V}}_i)\big) - \frac{h}{2}\,\pi\big(\psi_{\mathcal{C}}\tilde{f}_h\big)\Big| \leqslant Ch^{\frac{3}{2}}\Big(1+\log\frac{1}{h}\Big). \tag{G.28}$$

Finally, (4.1) and $\|\nabla f\|_\infty \leqslant 1$ give $|\tilde{f}_h - f| \leqslant \sqrt{h/2}\,\mathbb{E}|\chi| \leqslant \sqrt{dh/2}$, so $\frac{h}{2}\,|\pi(\psi_{\mathcal{C}}\,(\tilde{f}_h - f))| \leqslant Ch^{\frac{3}{2}}$ by (G.12). With (G.26) and (G.28),

$$\Big|\pi_h(f) - \pi(f) - \frac{h}{2}\,\pi(f\psi_{\mathcal{C}})\Big| \leqslant Ch^{\frac{3}{2}}\Big(1+\log\frac{1}{h}\Big).$$

Since $\pi_h(f) - \pi(f) - \frac{h}{2}\,\pi(f\psi_{\mathcal{C}}) = \int f\,\mathrm{d}(\pi_h - \pi - \frac{h}{2}\,\psi_{\mathcal{C}}\,\pi)$, the supremum over the functions f satisfying Assumption 2 and $\|\nabla f\|_\infty \leqslant 1$, with (G.5), proves (6.18). □

Proof of (6.19) *in Theorem 7.* The proof follows that of (6.18), with the local error (F.2) in place of (5.9), Lemma 25 for the term $i = 0$, and Lemma 26 for the move from $\bar{\pi}_h$ to π. Let f satisfy Assumption 2 and $\|\nabla f\|_\infty \leqslant 1$, and let

$$\mathcal{V}_i := e^{ih\mathcal{L}}f, \qquad i = 0, 1, 2, \dots, \tag{G.29}$$

so that $\mathcal{V}_0 = f$ and $\mathcal{V}_i \in \mathcal{D}$ by (A.12). By (G.7) at $t = ih$, as for (G.15), for $i \geqslant 1$ and $j = 1, 2, 3, 4$,

$$\|\nabla^j\mathcal{V}_i\|_\infty \leqslant C\min\big(1, (i+1)h\big)^{-\frac{j-1}{2}}e^{-\lambda ih}. \tag{G.30}$$

With the transition kernel $\mathcal{P}_h$ of (SGLD), as in Appendix F, $(\mathcal{P}_h g)(Z_0) = \hat{\mathbb{E}}_0\big[g(Z_1)\big]$ for a given Z_0. Since $\bar{\pi}_h$ is invariant under $\mathcal{P}_h$ and $e^{h\mathcal{L}}\mathcal{V}_i = \mathcal{V}_{i+1}$,

$$\bar{\pi}_h(\mathcal{V}_i) - \bar{\pi}_h(\mathcal{V}_{i+1}) = \bar{\pi}_h\big(\mathcal{P}_h\mathcal{V}_i - e^{h\mathcal{L}}\mathcal{V}_i\big),$$

and $|\mathcal{V}_N(x) - \pi(f)| \leqslant Ce^{-\lambda Nh}(1+|x|)$ by (G.7) at $j = 1$ and the invariance of π under $e^{t\mathcal{L}}$, as in (G.16). Summing over i,

$$\bar{\pi}_h(f) - \pi(f) = \sum_{i=0}^{\infty}\bar{\pi}_h\big(\mathcal{P}_h\mathcal{V}_i - e^{h\mathcal{L}}\mathcal{V}_i\big). \tag{G.31}$$

For the term $i = 0$ of (G.31), averaging (G.9) over $Z_0 \sim \bar{\pi}_h$, with $\bar{\pi}_h(|x|) \leqslant C$ from Lemma 20, gives

$$\big|\bar{\pi}_h\big(\mathcal{P}_h f - e^{h\mathcal{L}}f\big)\big| \leqslant Ch^{\frac{3}{2}}\big(1 + \bar{\pi}_h(|x|)\big) \leqslant Ch^{\frac{3}{2}}.$$

For $i \geqslant 1$, averaging (F.2) at $k = 0$, with $G = \mathcal{V}_i$, over $Z_0 \sim \bar{\pi}_h$, with $\bar{\pi}_h(|x|^{12}) \leqslant C$, gives

$$\Big|\bar{\pi}_h\big(\mathcal{P}_h\mathcal{V}_i - e^{h\mathcal{L}}\mathcal{V}_i\big) - \frac{h^2}{2}\,\bar{\pi}_h\Big(\nabla^2\mathcal{V}_i : \big(2\nabla^2 U + \mathcal{C}\big) - \nabla\mathcal{V}_i\cdot\big(\nabla^2 U\,\nabla U - \nabla\Delta U\big)\Big)\Big| \leqslant Ch^3\sum_{j=1}^{4}\|\nabla^j\mathcal{V}_i\|_\infty,$$

and, by (G.30) and (G.8),

$$Ch^3\sum_{i=1}^{\infty}\sum_{j=1}^{4}\|\nabla^j\mathcal{V}_i\|_\infty \leqslant Ch^2\Big(1+\log\frac{1}{h} + h^{-\frac{1}{2}}\Big) \leqslant Ch^{\frac{3}{2}}.$$

Hence

$$\left|\bar\pi_h(f)-\pi(f)-\frac{h^2}{2}\sum_{i=1}^{\infty}\bar\pi_h\Big(\nabla^2\mathcal{V}_i:\big(2\nabla^2U+\mathcal{C}\big)-\nabla\mathcal{V}_i\cdot\big(\nabla^2U\,\nabla U-\nabla\Delta U\big)\Big)\right|\leqslant Ch^{\frac32}. \tag{G.32}$$

By Assumption 1 and (C.27), for every $x\in\mathbb{R}^d$ and every symmetric matrix A,

$$\begin{aligned}\big|\nabla^2U(x)\,\nabla U(x)-\nabla\Delta U(x)\big|&\leqslant M^2(1+|x|)+dM\leqslant C(1+|x|),\\ \big|A:\big(2\nabla^2U(x)+\mathcal{C}(x)\big)\big|&\leqslant(2dM+\Sigma)\|A\|.\end{aligned}$$

Hence the integrand in (G.32) satisfies

$$\left|\nabla^2\mathcal{V}_i:\big(2\nabla^2U+\mathcal{C}\big)-\nabla\mathcal{V}_i\cdot\big(\nabla^2U\,\nabla U-\nabla\Delta U\big)\right|\leqslant C\big(1+|x|\big)\big(\|\nabla\mathcal{V}_i\|_\infty+\|\nabla^2\mathcal{V}_i\|_\infty\big),$$

and the moment bound $\bar\pi_h(|x|)\leqslant C$ of Lemma 20, (G.30) and (G.8) at $j=1,2$ give

$$\frac{h^2}{2}\sum_{i=1}^{\infty}\left|\bar\pi_h\Big(\nabla^2\mathcal{V}_i:\big(2\nabla^2U+\mathcal{C}\big)-\nabla\mathcal{V}_i\cdot\big(\nabla^2U\,\nabla U-\nabla\Delta U\big)\Big)\right|\leqslant Ch^2\sum_{i=1}^{\infty}\big(\|\nabla\mathcal{V}_i\|_\infty+\|\nabla^2\mathcal{V}_i\|_\infty\big)\leqslant Ch.$$

With (G.32), $|\bar\pi_h(f)-\pi(f)|\leqslant C(h+h^{\frac32})\leqslant Ch$ for every f satisfying Assumption 2 and $\|\nabla f\|_\infty\leqslant1$, and (G.5) gives

$$\mathcal{W}_1(\bar\pi_h,\pi)\leqslant Ch. \tag{G.33}$$

In the coefficient of $\frac{h^2}{2}$ in (G.32), the measure $\bar\pi_h$ can be replaced by π. By Assumption 1, the derivatives of U of orders 2 to 4 are bounded, so ∇^2U and $\nabla\Delta U$ are bounded and Lipschitz, and

$$\big|\big(\nabla^2U\,\nabla U-\nabla\Delta U\big)(x)-\big(\nabla^2U\,\nabla U-\nabla\Delta U\big)(x')\big|\leqslant C\big(1+|x|+|x'|\big)|x-x'|,\qquad x,x'\in\mathbb{R}^d. \tag{G.34}$$

With (G.34), (C.27), (C.26) and $|A:B|\leqslant d\|A\|\|B\|$, for each i, the integrand in the $\bar\pi_h$ bracket of (G.32) satisfies (G.11) with $K=C\sum_{j=1}^3\|\nabla^j\mathcal{V}_i\|_\infty$. Lemma 26, with $\mu=\bar\pi_h$, $\nu=\pi$, the moments of Lemmas 20 and 17, and $\mathcal{W}_1(\bar\pi_h,\pi)\leqslant Ch\leqslant1$ by (G.33) with $Ch_0\leqslant1$, gives, for $i\geqslant1$,

$$\begin{aligned}&\left|(\bar\pi_h-\pi)\Big(\nabla^2\mathcal{V}_i:\big(2\nabla^2U+\mathcal{C}\big)-\nabla\mathcal{V}_i\cdot\big(\nabla^2U\,\nabla U-\nabla\Delta U\big)\Big)\right|\\ &\qquad\leqslant C\,\mathcal{W}_1(\bar\pi_h,\pi)^{\frac{10}{11}}\sum_{j=1}^{3}\|\nabla^j\mathcal{V}_i\|_\infty\leqslant Ch^{\frac{10}{11}}\sum_{j=1}^{3}\|\nabla^j\mathcal{V}_i\|_\infty.\end{aligned} \tag{G.35}$$

By (G.30) and (G.8) at $j=1,2,3$, $h\sum_{i=1}^\infty\sum_{j=1}^3\|\nabla^j\mathcal{V}_i\|_\infty\leqslant C(1+\log\frac1h)$, so, by (G.35), replacing $\bar\pi_h$ by π in the coefficient of $\frac{h^2}{2}$ in (G.32) introduces an additional error of at most

$$\frac{h^2}{2}\sum_{i=1}^{\infty}Ch^{\frac{10}{11}}\sum_{j=1}^{3}\|\nabla^j\mathcal{V}_i\|_\infty\leqslant Ch^{1+\frac{10}{11}}\Big(1+\log\frac1h\Big)\leqslant Ch^{\frac32}.$$

Under π, (6.12) with $\mathcal{V}_i$ in place of ψ and (6.14) give

$$\pi\Big(\nabla^2\mathcal{V}_i:\big(2\nabla^2U+\mathcal{C}\big)-\nabla\mathcal{V}_i\cdot\big(\nabla^2U\,\nabla U-\nabla\Delta U\big)\Big)=\pi\big(\nabla^2\mathcal{V}_i:(\nabla^2U+\mathcal{C})\big)=\pi\big((\psi_{\nabla^2U}+\psi_{\mathcal{C}})\,(-\mathcal{L}\mathcal{V}_i)\big),$$

so we obtain

$$\left|\bar\pi_h(f)-\pi(f)-\frac{h^2}{2}\sum_{i=1}^{\infty}\pi\big((\psi_{\nabla^2U}+\psi_{\mathcal{C}})\,(-\mathcal{L}\mathcal{V}_i)\big)\right|\leqslant Ch^{\frac32}. \tag{G.36}$$

The sum in (G.36) telescopes as in the proof of (6.18), with $\mathcal{L}$, $\mathcal{V}_i$ of (G.29), the solution X_s of (1.1) and its initial value X_0 in place of $\tilde{\mathcal{L}}_h$, $\tilde{\mathcal{V}}_i$ of (G.13), $\tilde Y_0(s)$ and Y_0, with $\psi_{\nabla^2U}+\psi_{\mathcal{C}}$ in place of $\psi_{\mathcal{C}}$, and with (G.30) and Assumption 1 in place of (G.15) and (4.13). The identity (G.27) holds in this form, and so does

the bound on its left side by $Ch^{\frac{3}{2}}(1+|X_0|)^2\sum_{j=1}^{3}\|\nabla^j\mathcal{V}_i\|_\infty$. Since the sum starts at $i=1$, the partial sums of the telescoping series are $\pi((\psi_{\nabla^2 U}+\psi_{\mathcal{C}})\,\mathcal{V}_1)-\pi((\psi_{\nabla^2 U}+\psi_{\mathcal{C}})\,\mathcal{V}_N)$. By $\pi(\psi_{\nabla^2 U}+\psi_{\mathcal{C}})=0$, the bound $|\mathcal{V}_N(x)-\pi(f)|\leqslant Ce^{-\lambda Nh}(1+|x|)$ and (G.12), $\pi((\psi_{\nabla^2 U}+\psi_{\mathcal{C}})\,\mathcal{V}_N)\to 0$ as $N\to\infty$. Hence

$$\left|\frac{h^2}{2}\sum_{i=1}^{\infty}\pi\big((\psi_{\nabla^2 U}+\psi_{\mathcal{C}})\,(-\mathcal{L}\mathcal{V}_i)\big)-\frac{h}{2}\,\pi\big((\psi_{\nabla^2 U}+\psi_{\mathcal{C}})\,\mathcal{V}_1\big)\right|\leqslant Ch^{\frac{3}{2}}\Big(1+\log\frac{1}{h}\Big).$$

Finally, the bound on X_s-X_0 in (C.1) for (1.1) gives $|\mathcal{V}_1(x)-f(x)|=|\mathbb{E}^x[f(X_h)-f(X_0)]|\leqslant\mathbb{E}^x|X_h-X_0|\leqslant C\sqrt{h}\,(1+|x|)$, so $\frac{h}{2}\,|\pi((\psi_{\nabla^2 U}+\psi_{\mathcal{C}})\,(\mathcal{V}_1-f))|\leqslant Ch^{\frac{3}{2}}$ by (G.12). With (G.36),

$$\left|\bar\pi_h(f)-\pi(f)-\frac{h}{2}\,\pi\big(f\,(\psi_{\nabla^2 U}+\psi_{\mathcal{C}})\big)\right|\leqslant Ch^{\frac{3}{2}}\Big(1+\log\frac{1}{h}\Big),$$

and the supremum over the functions f satisfying Assumption 2 and $\|\nabla f\|_\infty\leqslant 1$, with (G.5), proves (6.19). □

G.4 Proof of Corollary 4: limits of the Wasserstein-1 errors

Proof of Corollary 4. Both limits of Corollary 4 follow from (6.20) as in Section 6.4, since π_h, $\bar\pi_h$ and π have finite first moments by Lemmas 8, 20 and 17.

For the positivity, a function $\psi\in L^2_0(\pi)$ with $\|\psi\,\pi\|_{\mathrm{KR}}=0$ has $\pi(g\psi)=0$ for every $g\in C_c^\infty(\mathbb{R}^d)$, hence $\psi=0$ π-almost everywhere. By (6.17), $\psi_{\nabla^2 U}$ is continuous and unbounded, since $|\Delta U|\leqslant dM$ while $|\nabla U(x)|\geqslant\langle x,\nabla U(x)\rangle/|x|\to\infty$ as $|x|\to\infty$ by (B.1). Hence $\psi_{\nabla^2 U}$ is nonzero on a nonempty open set, which has positive π-measure since $\pi(x)>0$ for every $x\in\mathbb{R}^d$, and $\|\psi_{\nabla^2 U}\,\pi\|_{\mathrm{KR}}>0$. For $\psi_{\mathcal{C}}$, let $r>0$ and

$$\varphi_r(x):=\big(r^2+|x|^2\big)^{\frac{1}{2}},\qquad x\in\mathbb{R}^d,\tag{G.37}$$

so that $\varphi_r\in\mathcal{D}$ and

$$\begin{aligned}&|\nabla\varphi_r|\leqslant 1,\qquad \nabla^2\varphi_r=\frac{\varphi_r^2 I_d-x\otimes x}{\varphi_r^3}\succeq\frac{r^2 I_d}{\varphi_r^3},\qquad \|\nabla^2\varphi_r\|\leqslant\frac{1}{\varphi_r},\\ &\Delta\varphi_r=\frac{d-1}{\varphi_r}+\frac{r^2}{\varphi_r^3},\qquad |\nabla\Delta\varphi_r|=\frac{(d-1)\,|x|}{\varphi_r^3}+\frac{3r^2|x|}{\varphi_r^5}\leqslant\frac{d+1}{2r^2}.\end{aligned}\tag{G.38}$$

Since $|\nabla U(x)|\leqslant M(1+|x|)\leqslant\sqrt{2}\,M\varphi_1(x)$, the gradient of $\mathcal{L}\varphi_1=\Delta\varphi_1-\nabla U\cdot\nabla\varphi_1$ satisfies, by (G.38), $|\nabla(\mathcal{L}\varphi_1)|\leqslant\frac{d+1}{2}+M+\sqrt{2}\,M\leqslant d+3M$. With (6.14) for $\omega=\varphi_1$ and $\mathcal{C}(x)$ positive semidefinite,

$$\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}\geqslant\frac{\pi\big(\psi_{\mathcal{C}}\,(-\mathcal{L}\varphi_1)\big)}{d+3M}=\frac{\pi\big(\nabla^2\varphi_1:\mathcal{C}\big)}{d+3M}\geqslant\frac{1}{d+3M}\,\pi\bigg(\frac{\mathrm{Tr}\big[\mathcal{C}(x)\big]}{(1+|x|^2)^{\frac{3}{2}}}\bigg).$$

The trace $\mathrm{Tr}\big[\mathcal{C}(x)\big]$ is nonnegative and continuous by (C.26); if $\mathcal{C}(x)$ does not vanish identically, the trace is positive on a nonempty open set, and $\pi\Big(\mathrm{Tr}\big[\mathcal{C}(x)\big]\,(1+|x|^2)^{-\frac{3}{2}}\Big)>0$, so $\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}>0$. □

G.5 Proof of Lemma 10: failure of the comparison

Proof of Lemma 10. The proof checks the assumptions, writes the Kantorovich–Rubinstein norm in dimension 1 as the integral (G.39), computes $\psi_{\mathcal{C}}$ explicitly, and evaluates both norms of Lemma 10 by quadrature. Assumption 1 holds with $m=1$, $R=0$ and every $M\geqslant 1$. The estimator is unbiased, $|\nabla\hat U(x;\xi)|\leqslant|x|+10$, its Hessian $1+10\,\xi\,a'(x)$ is bounded by $1+10\|a'\|_\infty$, and its variance is $100\,a(x)^2\leqslant 100$, so Assumption 3 holds for M large enough.

Here π is the standard Gaussian distribution, with $\pi'(x)=-x\,\pi(x)$. For $\psi\in L^2_0(\pi)$ and $s\in\mathbb{R}$, the Cauchy–Schwarz inequality on $(-\infty,s]$ for $s\leqslant 0$, and on $[s,\infty)$ for $s>0$ since $\pi(\psi)=0$, gives

$$\left|\int_{-\infty}^{s}\psi(y)\,\pi(y)\,\mathrm{d}y\right|\leqslant\pi(\psi^2)^{\frac{1}{2}}e^{-\frac{1}{4}s^2},$$

so for f Lipschitz with $|f'| \leqslant 1$ an integration by parts gives

$$\pi(f\psi) = -\int_{\mathbb{R}} f'(s) \int_{-\infty}^{s} \psi(y)\,\pi(y)\,\mathrm{d}y\,\mathrm{d}s \leqslant \int_{\mathbb{R}} \left| \int_{-\infty}^{s} \psi(y)\,\pi(y)\,\mathrm{d}y \right| \mathrm{d}s,$$

with equality for $f(s) = -\int_0^s \operatorname{sign}\bigl(\int_{-\infty}^{s'} \psi(y)\,\pi(y)\,\mathrm{d}y\bigr)\mathrm{d}s'$. Hence

$$\|\psi\,\pi\|_{\mathrm{KR}} = \int_{\mathbb{R}} \left| \int_{-\infty}^{s} \psi(y)\,\pi(y)\,\mathrm{d}y \right| \mathrm{d}s. \tag{G.39}$$

By (6.17), $\psi_{\nabla^2 U}(x) = \frac{1}{2}(x^2 - 1)$, and

$$\int_{-\infty}^{s} \psi_{\nabla^2 U}(y)\,\pi(y)\,\mathrm{d}y = -\frac{1}{2}\,s\,\pi(s), \qquad s \in \mathbb{R}. \tag{G.40}$$

For $\mathcal{A} = a^2$, let $J(x) := \int_{-\infty}^{x} y\,a(y)^2\,\mathrm{d}y$ and $\zeta := \pi(a^2) - \pi(J)$. The function $J - a^2 + \zeta$ is smooth and bounded, satisfies $\pi(J - a^2 + \zeta) = 0$, and has the derivative $x\,a^2 - (a^2)'$, so for $\omega \in \mathcal{D}$, (3.4) gives

$$\pi\bigl((J - a^2 + \zeta)(-\mathcal{L}\omega)\bigr) = \pi\bigl((x\,a^2 - (a^2)')\,\omega'\bigr) = -\int_{\mathbb{R}} \omega'(x)\,\bigl(a(x)^2\pi(x)\bigr)'\,\mathrm{d}x = \pi\bigl(\omega''\,a^2\bigr),$$

which is (6.14) for $\mathcal{A} = a^2$. By the uniqueness in Theorem 6,

$$\psi_{a^2} = J - a^2 + \zeta. \tag{G.41}$$

Since $\mathcal{C} = 100\,a^2$, the linearity in Theorem 6 gives $\psi_{\mathcal{C}} = 100\,\psi_{a^2}$. Quadrature in (G.39), with (G.40) and (G.41), then gives the values in Lemma 10, so (6.21) fails. □

G.6 Proof of Theorem 8: threshold on the batch size

Proof of Theorem 8. The proof bounds $\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}}$ from below by testing (6.14) with a function of the distance to the minimizer of U, bounds $\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}$ from above through (G.2), and compares the 2 bounds. Let x_* be the minimizer of U, and let φ_r be the function of (G.37) with $r^2 = d/m$ and $x - x_*$ in place of x. The gradient of $\mathcal{L}\varphi_r = \Delta\varphi_r - \nabla U \cdot \nabla\varphi_r$ satisfies

$$|\nabla(\mathcal{L}\varphi_r)| \leqslant 2\,\|\nabla^2 U\|_\infty + \frac{d+1}{2r^2} \leqslant 3\,\|\nabla^2 U\|_\infty, \tag{G.42}$$

since (G.38) bounds $|\nabla\Delta\varphi_r|$ by $\frac{d+1}{2r^2} = \frac{(d+1)\,m}{2d} \leqslant m \leqslant \|\nabla^2 U\|_\infty$, and, by $|\nabla\varphi_r| \leqslant 1$, $\|\nabla^2\varphi_r\| \leqslant \varphi_r^{-1}$ and $|\nabla U(x)| = |\nabla U(x) - \nabla U(x_*)| \leqslant \|\nabla^2 U\|_\infty\,\varphi_r(x)$,

$$|\nabla(\nabla U \cdot \nabla\varphi_r)| \leqslant \|\nabla^2 U\|\,|\nabla\varphi_r| + \|\nabla^2\varphi_r\|\,|\nabla U| \leqslant 2\|\nabla^2 U\|_\infty.$$

By (G.42), the function $-\mathcal{L}\varphi_r/(3\,\|\nabla^2 U\|_\infty)$ is 1-Lipschitz, and (6.14) with $\omega = \varphi_r$ gives $\pi((-\mathcal{L}\varphi_r)\,\psi_{\mathcal{A}}) = \pi(\nabla^2\varphi_r : \mathcal{A})$ for $\mathcal{A} = \nabla^2 U$ and $\mathcal{A} = \mathcal{C}$. As $\nabla^2\varphi_r$ and $\mathcal{C}(x)$ are positive semidefinite, $\pi(\nabla^2\varphi_r : \mathcal{C}) \geqslant 0$, and as $\nabla^2 U \succeq m I_d$, $\nabla^2\varphi_r : \nabla^2 U \geqslant m\,\Delta\varphi_r$. Hence

$$\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}} \geqslant \frac{\pi\bigl((-\mathcal{L}\varphi_r)\,(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\bigr)}{3\,\|\nabla^2 U\|_\infty} \geqslant \frac{m\,\pi(\Delta\varphi_r)}{3\,\|\nabla^2 U\|_\infty}. \tag{G.43}$$

The identity (3.4) with $g_1 = 1$ and $g_2 = \frac{1}{2}|x - x_*|^2$ gives $\pi(\nabla U \cdot (x - x_*)) = d$, and $\nabla U(x) \cdot (x - x_*) \geqslant m|x - x_*|^2$, so $\pi(|x - x_*|^2) \leqslant d/m$. As functions of $|x - x_*|^2$, $(r^2 + |x - x_*|^2)^{-\frac{1}{2}}$ and $(r^2 + |x - x_*|^2)^{-\frac{3}{2}}$ are convex and decreasing, so Jensen's inequality and $r^2 = d/m$ give

$$\pi(\Delta\varphi_r) \geqslant \frac{d-1}{(2d/m)^{\frac{1}{2}}} + \frac{d/m}{(2d/m)^{\frac{3}{2}}} = \left(\frac{m}{2d}\right)^{\frac{1}{2}} \left(d - \frac{1}{2}\right) \geqslant \left(\frac{md}{8}\right)^{\frac{1}{2}},$$

and (G.43) gives

$$\|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}} \geqslant \frac{m^{\frac{3}{2}}\sqrt{d}}{6\sqrt{2}\,\|\nabla^2 U\|_\infty} \geqslant \frac{m^{\frac{3}{2}}\sqrt{d}}{9\,\|\nabla^2 U\|_\infty}. \tag{G.44}$$

For $\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}}$ in (6.21), let f satisfy Assumption 2 and $\|\nabla f\|_\infty \leqslant 1$, and let ϕ_0 solve (3.7). By (G.2) with $\omega = \phi_0 \in \mathcal{D}$ and $\nabla^2 U \succeq m I_d$, and (3.4),

$$\pi\big(\nabla^2\phi_0 : \nabla^2\phi_0\big) + m\,\pi\big(|\nabla\phi_0|^2\big) \leqslant \pi\big((f - \pi(f))^2\big) = \pi\big(\nabla f \cdot \nabla\phi_0\big) \leqslant \pi\big(|\nabla\phi_0|^2\big)^{\frac{1}{2}}. \tag{G.45}$$

By (G.45), $\pi(|\nabla\phi_0|^2)^{\frac{1}{2}} \leqslant m^{-1}$ and then $\pi(\nabla^2\phi_0 : \nabla^2\phi_0) \leqslant m^{-1}$. As for (G.4), (6.14) with $\omega = \phi_0$ and the Cauchy–Schwarz inequality give $|\pi(f\psi_{\mathcal{C}})| = |\pi(\nabla^2\phi_0 : \mathcal{C})| \leqslant \sqrt{d}\,\|\mathcal{C}\|_\infty\, m^{-\frac{1}{2}}$, and the supremum over f with (G.5) gives

$$\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}} \leqslant \frac{\sqrt{d}\,\|\mathcal{C}\|_\infty}{m^{\frac{1}{2}}}. \tag{G.46}$$

Hence, whenever $\|\mathcal{C}\|_\infty \leqslant m^2/(9\,\|\nabla^2 U\|_\infty)$, (G.46) and (G.44) give

$$\|\psi_{\mathcal{C}}\,\pi\|_{\mathrm{KR}} \leqslant \frac{\sqrt{d}\,\|\mathcal{C}\|_\infty}{m^{\frac{1}{2}}} \leqslant \frac{m^{\frac{3}{2}}\sqrt{d}}{9\,\|\nabla^2 U\|_\infty} \leqslant \|(\psi_{\nabla^2 U} + \psi_{\mathcal{C}})\,\pi\|_{\mathrm{KR}},$$

which is (6.21). For SGLM and SGLD with a batch of P independent stochastic gradients, the averaged estimator satisfies Assumption 3 with Σ/P in place of Σ (Section 6.1), so Corollary 4 applies with its covariance $\mathcal{C}(x)/P$, and $\|\mathcal{C}\|_\infty/P \leqslant m^2/(9\,\|\nabla^2 U\|_\infty)$ exactly when P satisfies (6.23). The representor of $\mathcal{C}(x)/P$ is $\psi_{\mathcal{C}}/P$ by the linearity in Theorem 6, and (6.21) with $\psi_{\mathcal{C}}/P$ in place of $\psi_{\mathcal{C}}$ is (6.24). □