Nonconvex Stochastic Gradient Sampling at Second Order Accuracy via Noise Correction

Xuda Ye[1] and Yuliang Wang[2]

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

November 1, 2026

Abstract

We build a class of stochastic gradient integrators with second order accuracy via noise correction. The integrators multiply the injected Gaussian noise of the stochastic gradient Leimkuhler–Matthews integrator by a noise factor, obtained by applying a scalar function, the noise corrector, to the eigenvalues of an estimate of the covariance of the batch gradient from the same batch. For nonconvex potentials that are strongly convex outside a ball, every noise corrector satisfying a polynomial bound and every batch size, we prove that the numerical bias is of second order in the step size h and that the mean square error of the time average of N samples is bounded by a constant times $(Nh)^{-1} + h^4$. Under Gaussian gradient noise and an additional polynomial bound on the square of the noise corrector, we separate the coefficient of h^2 in the numerical bias into explicit terms from discretization and noise correction.

1 Introduction

Sampling from a target distribution $\pi(x) \propto e^{-U(x)}$ on $\mathbb{R}^d$ is a basic task in Bayesian inference and statistical physics. When U is a sum over a large data set, only a stochastic gradient $g(x;\xi)$ is affordable, an unbiased estimator of $\nabla U(x)$ computed from a random sample ξ such as a mini-batch. The most popular sampler of this kind is stochastic gradient Langevin dynamics (SGLD) [28], the Euler–Maruyama discretization of the overdamped Langevin dynamics with g in place of ∇U. With a step size h, the numerical invariant distribution of SGLD has a bias of first order in h [27] with two sources [30, Corollary 3]: the stochastic gradient error, caused by the gradient noise $g(x;\xi) - \nabla U(x)$, and the discretization error of Euler–Maruyama, which is of first order even with the full gradient [26].

Several approaches reduce one of these two errors. Variance reduction [1, 2, 9] reduces the stochastic gradient error. A noise correction compensates for the gradient noise by reducing the injected Gaussian noise, with the covariance of the gradient noise in the modified SGLD [27] or with an estimate of it from an independent batch in the covariance-corrected SGLD (CC-SGLD) [8]. These variants of SGLD keep the Euler–Maruyama step and its discretization error of first order. In contrast, the stochastic gradient Leimkuhler–Matthews (SGLM) integrator of our earlier work [30] keeps the second order discretization error of the Leimkuhler–Matthews (LM) integrator [16], and its bias of first order comes from the gradient noise alone. For the underdamped Langevin dynamics, stochastic gradient Hamiltonian Monte Carlo (SGHMC) [5] corrects the injected noise in a scheme of Euler type, and the stochastic gradient UBU (SGUBU), built on the UBU integrator [24], has a bias of first order caused by the gradient noise [19, 21]. A bias of second order with stochastic gradients has been obtained only in very special settings. For finite-sum potentials under global convexity, SGUBU attains it with a forward and a backward sweep over a partition of the summands into mini-batches [22], and variance-reduced variants of SGUBU attain it with control variates built from the full gradient or from a table of all component gradients, at step sizes below a threshold inversely proportional to the number of summands [19]. Adaptive Langevin dynamics attains a bias of second order when the covariance of the gradient noise is constant, with a constant in the bias that grows with the number

of summands [25]. These constructions rely on the finite-sum structure of U or on a constant covariance, and do not apply to a general stochastic gradient $g(x;\xi)$.

Apart from these settings, all the stochastic gradient integrators studied earlier have a bias of first order, which comes from the stochastic gradient error, the discretization error, or both. Is it possible to design a stochastic gradient integrator with second order accuracy under mild conditions?

The answer is positive: we present a general noise correction approach for SGLM, which achieves a bias of second order in h. At each step, the covariance of the batch gradient is estimated from the same batch of P stochastic gradients, and the injected Gaussian noise is multiplied by a noise factor, obtained by applying a scalar function, the noise corrector, to the eigenvalues of this estimate. The resulting noise-corrected SGLM (NC-SGLM) integrators use twice as many stochastic gradients per step as SGLD with the same batch size. The error analysis requires the noise corrector to satisfy only two explicit bounds, which together form the polynomial bound, and no global bound on the covariance of the gradient noise is needed. To the best of our knowledge, NC-SGLM is the first class of stochastic gradient integrators that achieves second order accuracy for nonconvex potentials with a general stochastic gradient.

Our proof is built upon the modified discrete Poisson equation developed in [30]. The proof requires the potential U to be strongly convex only outside a ball and the moments of the stochastic gradient to grow at most linearly in $|x|$, together with regularity conditions. Under these conditions, for every noise corrector satisfying polynomial bounds, the bias of NC-SGLM is $\mathcal{O}(h^2)$, and the mean square error (MSE) of the time average of N samples is bounded by a constant times $(Nh)^{-1} + h^4$ at every batch size $P \geqslant 2$, with constants that do not depend on the noise corrector. An MSE of ε^2 is then reached with N of order $\varepsilon^{-5/2}$, against ε^{-3} for SGLD [27, Theorem 9], which makes NC-SGLM advantageous when a high accuracy is desired. The noise corrector of CC-SGLD does not satisfy the polynomial bound, and the same results hold for the corresponding NC-SGLM integrator when the moments of the gradient noise are bounded uniformly in x. The second order accuracy at every batch size $P \geqslant 2$ is a limit behavior as $h \to 0$; at a finite step size, the accuracy of NC-SGLM is formally related to the probability of capping.

A parallel line of work bounds the number of queries to a stochastic or noisy oracle needed to reach a prescribed accuracy [3, 4, 6, 7, 10, 14, 23]. In these works, the step size, the batch size and the number of iterations are set from the target accuracy and from the constants of the problem, such as the smoothness constant of U, its convexity or logarithmic Sobolev constant, the dimension and the variance of the gradient noise. Several guarantees also depend on the distance from the initial value to a minimizer of U, or assume a warm start. Some of these samplers are built from implicit steps: an inner sampler run to a prescribed accuracy at each step [14], or rejection sampling around the output of an approximate proximal oracle [4]. Implementing these samplers as analyzed therefore requires the constants of the problem, and in the numerical experiments of [14], the parameters are tuned by a grid search instead. In contrast, NC-SGLM is an explicit update whose definition involves no constant of U or of the gradient noise, and the user chooses only the step size, the batch size and the noise corrector.

Section 2 introduces NC-SGLM and the noise correctors, analyzes the bias formally, and describes the computation of the noise factor. Section 3 states the main theorems and compares NC-SGLM with SGLM. Section 4 presents the numerical experiments, and Section 5 concludes. Appendices A–E contain the proofs.

2 Noise correction in the SGLM integrator

This section introduces the noise correction of SGLM with a general noise corrector, together with four choices of the noise corrector. The stochastic gradient $g(x;\xi)$ satisfies the unbiasedness condition $\mathbb{E}_\xi[g(x;\xi)] = \nabla U(x)$, where ξ is the random variable that carries the randomness of the stochastic gradient, and $\mathbb{E}_\xi$ is the expectation over ξ at fixed x.

2.1 SGLD and SGLM for the overdamped Langevin dynamics

The overdamped Langevin dynamics

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \tag{2.1}$$

with $(B_t)_{t\geqslant 0}$ a standard Brownian motion in $\mathbb{R}^d$, has the invariant distribution π. For a batch size $P \geqslant 2$, we draw a batch $\boldsymbol{\xi} := \big(\xi^{(1)}, \dots, \xi^{(P)}\big)$ of P independent copies of ξ, and form the batch gradient,

$$\hat{g}_P(x; \boldsymbol{\xi}) := \frac{1}{P} \sum_{p=1}^{P} g(x; \xi^{(p)}) \in \mathbb{R}^d, \tag{2.2}$$

where the hat and the subscript P distinguish it from a single stochastic gradient $g(x;\xi)$.

Let $h > 0$ be the step size, and let $(\boldsymbol{\xi}_k)_{k\geqslant 0}$ be batches and $(\chi_k)_{k\geqslant 0}$ be vectors drawn from $\mathcal{N}(0, I_d)$ with I_d the $d \times d$ identity matrix, all mutually independent. SGLD is the Euler–Maruyama discretization of (2.1) with the batch gradient in place of ∇U,

$$Z_{k+1} = Z_k - h\, \hat{g}_P(Z_k; \boldsymbol{\xi}_k) + \sqrt{2h}\, \chi_k. \tag{SGLD}$$

SGLM is the stochastic gradient version of the LM integrator [16], and is studied as a variant of SGLD in [30]. SGLM is the Markov chain $(Y_k)_{k\geqslant 0}$ in $\mathbb{R}^d$ with the update

$$Z_k = Y_k + \sqrt{\frac{h}{2}}\, \chi_k, \quad Y_{k+1} = Y_k - h\, \hat{g}_P(Z_k; \boldsymbol{\xi}_k) + \sqrt{2h}\, \chi_k. \tag{SGLM}$$

Although $(Y_k)_{k\geqslant 0}$ is the Markov chain, the samples that approximate the target distribution are $(Z_k)_{k\geqslant 0}$. With the full gradient ∇U in place of $\hat{g}_P$, (SGLM) is the Markovian form of LM, and the samples have a discretization error of order h^2.

2.2 A general noise correction approach

With $v^{\otimes 2} := vv^\top$ for $v \in \mathbb{R}^d$, the covariance matrix of $g(x;\xi)$ is

$$\mathcal{C}(x) := \mathbb{E}_\xi\Big[\big(g(x;\xi) - \nabla U(x)\big)^{\otimes 2}\Big] \in \mathbb{R}^{d\times d}. \tag{2.3}$$

With $\mathbb{E}_{\boldsymbol{\xi}}$ the expectation over the batch at fixed x, the covariance of $\hat{g}_P(x;\boldsymbol{\xi})$ in (2.2) is

$$\mathbb{E}_{\boldsymbol{\xi}}\Big[\big(\hat{g}_P(x;\boldsymbol{\xi}) - \nabla U(x)\big)^{\otimes 2}\Big] = \frac{1}{P^2} \sum_{p=1}^{P} \mathbb{E}_{\boldsymbol{\xi}}\Big[\big(g(x;\xi^{(p)}) - \nabla U(x)\big)^{\otimes 2}\Big] = \frac{1}{P}\, \mathcal{C}(x), \tag{2.4}$$

where the cross terms vanish since $\xi^{(1)}, \dots, \xi^{(P)}$ are independent and each $g(x;\xi^{(p)}) - \nabla U(x)$ has mean 0 by the unbiasedness condition. The batch approximation of $\mathcal{C}(x)/P$ is

$$\hat{\mathcal{C}}_P(x;\boldsymbol{\xi}) := \frac{1}{P(P-1)} \sum_{p=1}^{P} \big(g(x;\xi^{(p)}) - \hat{g}_P(x;\boldsymbol{\xi})\big)^{\otimes 2} \in \mathbb{R}^{d\times d}, \qquad \mathbb{E}_{\boldsymbol{\xi}}\big[\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big] = \frac{1}{P}\, \mathcal{C}(x), \tag{2.5}$$

where the identity for the mean follows again from the independence of $\xi^{(1)}, \dots, \xi^{(P)}$.

Noise correction multiplies the Gaussian noise injected in each step by a matrix formed from $\hat{\mathcal{C}}_P$ through a *noise corrector*, a measurable function $\mathcal{S}_h : [0,\infty) \to \mathbb{R}$. The error analysis requires the noise corrector to satisfy the following *polynomial bound*.

For every $h > 0$ and $\lambda \geqslant 0$, the noise corrector $\mathcal{S}_h$ satisfies

$$\big|\mathcal{S}_h(\lambda) - 1\big| \leqslant \min\Big\{1, \frac{h\lambda}{2}\Big\}, \qquad \Big|\mathcal{S}_h(\lambda) - 1 + \frac{h\lambda}{4}\Big| \leqslant \frac{(h\lambda)^2}{8}. \tag{2.6}$$

The noise corrector $\mathcal{S}_h$ acts on the symmetric positive semidefinite matrix $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$ through its eigenvalues: the noise factor $\mathcal{S}_h\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) \in \mathbb{R}^{d\times d}$ has the eigenvectors of $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$, and each eigenvalue λ of $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$ is replaced by $\mathcal{S}_h(\lambda)$. The NC-SGLM integrator multiplies the injected noise $\sqrt{2h}\,\chi_k$ of (SGLM) by the noise factor at Y_k, and is the Markov chain $(Y_k)_{k\geqslant 0}$ in $\mathbb{R}^d$ with the update

$$Z_k = Y_k + \sqrt{\frac{h}{2}}\, \chi_k, \quad Y_{k+1} = Y_k - h\, \hat{g}_P(Z_k; \boldsymbol{\xi}_k) + \sqrt{2h}\, \mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_k; \boldsymbol{\xi}_k)\big)\chi_k, \tag{NC-SGLM}$$

whose samples are $(Z_k)_{k\geqslant 0}$. NC-SGLM is thus a class of integrators, one for each noise corrector $\mathcal{S}_h$. For a test function f and $\pi(f) := \int_{\mathbb{R}^d} f \,\mathrm{d}\pi$, the MSE of the time average of N samples is

$$\mathrm{MSE}(N,h) := \mathbb{E}\left[\left|\frac{1}{N}\sum_{k=0}^{N-1} f(Z_k) - \pi(f)\right|^2\right], \tag{2.7}$$

where the expectation is over the batches and the Gaussian vectors, from a given initial value $Y_0 \in \mathbb{R}^d$.

The covariance correction (CC) of CC-SGLD [8] is the noise corrector

$$\mathcal{S}_h^{\mathrm{CC}}(\lambda) = 1 - \frac{h\lambda}{4}, \tag{2.8}$$

with the linear noise factor $I_d - \frac{h}{4}\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$. The corresponding integrator (NC-SGLM) is called CC-SGLM. CC does not satisfy the polynomial bound (2.6) for $h\lambda > 4$ and amplifies the injected noise for $h\lambda > 8$. With unbounded variance of the gradient noise, the eigenvalues of $\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)$ can exceed $4/h$ for large $|Y_k|$, which could lead to the instability of CC. CC-SGLD sets the covariance estimate to 0 at large $|x|$ as a safeguard. We do not employ this safeguard because it introduces a discontinuous correction. CC without safeguard also enables a clearer comparison with other noise correctors.

We further consider three noise correctors, the root correction (RC), the stable correction (SC) and the inverse correction (IC),

$$\mathcal{S}_h^{\mathrm{RC}}(\lambda) = \left(1 - \frac{h\lambda}{2}\right)_+^{\frac{1}{2}}, \qquad \mathcal{S}_h^{\mathrm{SC}}(\lambda) = \left(1 - \frac{h\lambda}{4}\right)_+, \qquad \mathcal{S}_h^{\mathrm{IC}}(\lambda) = \left(1 + \frac{h\lambda}{4}\right)^{-1}, \tag{2.9}$$

where $t_+ := \max\{t, 0\}$ for $t \in \mathbb{R}$. Each of the three noise correctors satisfies the polynomial bound (2.6) and takes values in $[0,1]$, so that its noise factor never amplifies the injected noise. The corresponding integrators (NC-SGLM) are called RC-SGLM, SC-SGLM and IC-SGLM, respectively.

RC is the default choice: when $h\lambda \leqslant 2$ for every eigenvalue λ of $\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)$, the identity

$$2h\,\mathcal{S}_h^{\mathrm{RC}}\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)^2 = 2hI_d - h^2\,\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k) \tag{2.10}$$

holds exactly. For an eigenvalue $\lambda > 2/h$, the positive part sets the corresponding eigenvalue of $\mathcal{S}_h^{\mathrm{RC}}\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)$ to 0, and we say that $\mathcal{S}_h^{\mathrm{RC}}\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)$ is capped. SC is CC with the positive part, $\mathcal{S}_h^{\mathrm{SC}}(\lambda) = \mathcal{S}_h^{\mathrm{CC}}(\lambda)_+$, so that $\mathcal{S}_h^{\mathrm{SC}}\big(\hat{\mathcal{C}}_P\big) = \mathcal{S}_h^{\mathrm{CC}}\big(\hat{\mathcal{C}}_P\big)$ when $h\lambda \leqslant 4$ for every eigenvalue λ of $\hat{\mathcal{C}}_P$. IC needs no positive part, and its noise factor is $\mathcal{S}_h^{\mathrm{IC}}\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) = \big(I_d + \frac{h}{4}\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big)^{-1}$.

2.3 Formal analysis of the second order bias

The numerical biases of SGLD and SGLM are of first order in h, and their coefficients of h separate the discretization error from the gradient noise. Let $\mathcal{L} := \Delta - \nabla U \cdot \nabla$ be the generator of (2.1), and let ϕ_0 solve the Poisson equation

$$-\mathcal{L}\phi_0(x) = f(x) - \pi(f), \qquad x \in \mathbb{R}^d. \tag{2.11}$$

The Poisson equation is widely used in the analysis of integrators for stochastic differential equations [20, 29]. Our earlier work [30] proves the existence of the numerical invariant distributions π_h^{SGLD} and π_h^{SGLM} of SGLD and SGLM, and derives the explicit coefficients of h in their numerical biases. With the batch gradient $\hat{g}_P$, whose covariance is $\mathcal{C}/P$ by (2.4), the coefficients are

$$\lim_{h\to 0} \frac{\pi_h^{\mathrm{SGLD}}(f) - \pi(f)}{h} = \frac{1}{2}\,\pi\big(\nabla^2\phi_0 : \nabla^2 U\big) + \frac{1}{2P}\,\pi\big(\nabla^2\phi_0 : \mathcal{C}\big), \tag{2.12}$$

$$\lim_{h\to 0} \frac{\pi_h^{\mathrm{SGLM}}(f) - \pi(f)}{h} = \frac{1}{2P}\,\pi\big(\nabla^2\phi_0 : \mathcal{C}\big), \tag{2.13}$$

where : denotes the full contraction of matrices. In (2.12), the term with $\nabla^2 U$ comes from the error of the Euler–Maruyama discretization and remains with the full gradient, and the term with $\mathcal{C}$ comes from the

gradient noise. SGLM removes the term with $\nabla^2 U$, since LM has a discretization error of second order, and the coefficient (2.13) of SGLM comes from the gradient noise alone.

The term with $\mathcal{C}$ in (2.13) reflects a broken fluctuation–dissipation relation [15]: the overdamped Langevin dynamics keeps π invariant by the balance between the drift $-\nabla U \cdot \nabla$ and the diffusion Δ in $\mathcal{L}$. In a step of SGLM from Y_k, the batch noise $\hat{g}_P(Z_k;\boldsymbol{\xi}_k) - \nabla U(Z_k)$ has mean 0 given Y_k and χ_k, and the term $-h\,(\hat{g}_P(Z_k;\boldsymbol{\xi}_k) - \nabla U(Z_k))$ that it contributes to the step $Y_{k+1} - Y_k$ has, by (2.4) at $x = Z_k$, the covariance

$$h^2\,\mathbb{E}_{\boldsymbol{\xi}_k}\Big[\big(\hat{g}_P(Z_k;\boldsymbol{\xi}_k) - \nabla U(Z_k)\big)^{\otimes 2}\Big] = \frac{h^2}{P}\,\mathcal{C}(Z_k), \tag{2.14}$$

which is $h^2\,\mathcal{C}(Y_k)/P$ up to higher order terms. This covariance adds the diffusion $\frac{h}{2P}\,\mathcal{C} : \nabla^2$ to $\mathcal{L}$ without a matching drift, and the linear response to this extra diffusion gives the coefficient (2.13).

Noise correction restores the fluctuation–dissipation relation by reducing the injected noise. Given Y_k and $\boldsymbol{\xi}_k$, the injected noise $\sqrt{2h}\,\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)\chi_k$ of NC-SGLM has the covariance $2h\,\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)^2$. By the polynomial bound (2.6) and the mean in (2.5),

$$\mathbb{E}_{\boldsymbol{\xi}_k}\Big[2h\,\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)^2\Big] = 2hI_d - \frac{h^2}{P}\,\mathcal{C}(Y_k) + \mathcal{O}(h^3). \tag{2.15}$$

Up to higher order terms, the reduction $h^2\,\mathcal{C}(Y_k)/P$ in (2.15) cancels the covariance (2.14). The other terms of the step leave this cancellation intact: the noise correction $\sqrt{2h}\,\big(\mathcal{S}_h(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)) - I_d\big)\chi_k$ is of order h by the bound on $|\mathcal{S}_h(\lambda) - 1|$ in the polynomial bound (2.6), with a coefficient independent of χ_k given Y_k, and the batch noise has mean 0 given Y_k and χ_k, so that their products with the other terms of the step have expectations of higher order. The extra diffusion of SGLM is thus removed, the coefficient (2.13) vanishes for NC-SGLM, and the bias of NC-SGLM is formally of second order, the order of the discretization error of LM. RC exactly restores the fluctuation–dissipation relation: when $\mathcal{S}_h^{\mathrm{RC}}\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)$, with $\mathcal{S}_h^{\mathrm{RC}}$ in (2.9), is not capped for almost every batch $\boldsymbol{\xi}_k$, (2.10) gives (2.15) without the term $\mathcal{O}(h^3)$.

In (NC-SGLM), the batch gradient is evaluated at Z_k and the noise factor at Y_k, so the stochastic gradients of each batch $\boldsymbol{\xi}_k$ are evaluated at both points, and NC-SGLM uses twice as many stochastic gradients as (SGLM) with the same batch size. Evaluating the noise factor at Y_k keeps the mean of the step: since Y_k and $\boldsymbol{\xi}_k$ are independent of χ_k, the noise correction has mean 0 given Y_k, and the step of NC-SGLM has the mean of the step of (SGLM). Evaluated at Z_k, the noise factor would depend on χ_k, and the noise correction would in general shift the mean of the step and leave a bias of first order.

2.4 Computation of the noise corrector

Each step of (NC-SGLM) uses the noise factor only through the vector $\mathcal{S}_h\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big)\chi \in \mathbb{R}^d$. Since $\mathcal{S}_h$ acts on the eigenvalues of $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$, a matrix of rank at most $\min\{d, P-1\}$, this product requires in general the eigendecomposition of $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$, which is expensive when $\min\{d, P-1\}$ is large. The costs below count arithmetic operations and exclude the $2P$ evaluations of the stochastic gradient in each step.

The exact computation forms the $d \times d$ matrix $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$ at a cost of order d^2P, diagonalizes it at a cost of order d^3, and applies $\mathcal{S}_h$ to its eigenvalues, at a total cost of order $d^2P + d^3$. When $P < d$, the nonzero eigenvalues of $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$ and their eigenvectors follow from the eigendecomposition of the $P \times P$ matrix with the entries $\big(g(x;\xi^{(p)}) - \hat{g}_P(x;\boldsymbol{\xi})\big)\cdot\big(g(x;\xi^{(q)}) - \hat{g}_P(x;\boldsymbol{\xi})\big)$, $p, q = 1, \dots, P$, at a total cost of order $dP^2 + P^3$.

The grouping trick reduces the cost for a large batch. Let $P_0 \geqslant 2$ divide P, and split the batch into P_0 groups of P/P_0 elements,

$$\boldsymbol{\xi} = \big(\boldsymbol{\xi}^{[1]}, \dots, \boldsymbol{\xi}^{[P_0]}\big), \qquad \boldsymbol{\xi}^{[j]} := \big(\xi^{((j-1)P/P_0+1)}, \dots, \xi^{(jP/P_0)}\big).$$

The group means $\hat{g}_{P/P_0}(x;\boldsymbol{\xi}^{[j]})$, $j = 1, \dots, P_0$, given by (2.2) with $\boldsymbol{\xi}^{[j]}$ in place of $\boldsymbol{\xi}$ and P/P_0 in place of P, are independent, each with the mean $\nabla U(x)$ and, by (2.4) with P/P_0 in place of P, the covariance $\frac{P_0}{P}\,\mathcal{C}(x)$, and their average is $\hat{g}_P(x;\boldsymbol{\xi})$. The grouped approximation of $\mathcal{C}(x)/P$ is

$$\hat{\mathcal{C}}_{P,P_0}(x;\boldsymbol{\xi}) := \frac{1}{P_0(P_0-1)}\sum_{j=1}^{P_0}\big(\hat{g}_{P/P_0}(x;\boldsymbol{\xi}^{[j]}) - \hat{g}_P(x;\boldsymbol{\xi})\big)^{\otimes 2}, \qquad \mathbb{E}_{\boldsymbol{\xi}}\big[\hat{\mathcal{C}}_{P,P_0}(x;\boldsymbol{\xi})\big] = \frac{1}{P}\,\mathcal{C}(x), \tag{2.16}$$

of rank at most $P_0 - 1$. With $\hat{\mathcal{C}}_{P,P_0}$ in place of $\hat{\mathcal{C}}_P$ in (NC-SGLM), forming the group means costs order dP, and the eigendecomposition of the $P_0 \times P_0$ matrix formed in the same way from the P_0 group means costs order $dP_0^2 + P_0^3$, so that the total cost is of order $dP + dP_0^2 + P_0^3$, linear in d and in P for a fixed P_0, such as $P_0 = 16$. Since (2.2) and (2.5), applied to the P_0 group means, give $\hat{g}_P(x;\boldsymbol{\xi})$ and $\hat{\mathcal{C}}_{P,P_0}(x;\boldsymbol{\xi})$, NC-SGLM with the grouped approximation is NC-SGLM at the batch size P_0 for the stochastic gradient $\hat{g}_{P/P_0}$, and the formal analysis of Section 2.3 applies to it.

Only special forms of $\mathcal{S}_h$ avoid the eigendecomposition and allow a direct computation of $\mathcal{S}_h\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big)\chi$. One such form is $\mathcal{S}_h^{\mathrm{CC}}$ in (2.8), whose linear noise factor acts on χ through the inner products of χ with the P vectors $g(x;\xi^{(p)}) - \hat{g}_P(x;\boldsymbol{\xi})$, at a cost of order dP. Hence CC is the most affordable choice in high dimensions, despite its possible instability.

3 Error analysis results

This section states the bound on the MSE of NC-SGLM in Theorem 1, the uniform-in-time weak error in Theorem 2 and the expansion of the numerical bias in Theorem 3. It then compares these results with those of SGLM and bounds the probability of capping in Theorem 4. Finally, it gives the explicit coefficient of h^2 under Gaussian noise in Theorem 5. The proofs are given in the appendix.

3.1 Notations and assumptions

The Euclidean norm on $\mathbb{R}^d$ is $|\cdot|$, and $\|\cdot\|$ is the operator norm of matrices and tensors, such as the derivative $\nabla^i U(x)$ of order i. We write $\|\nabla^i\varphi\|_\infty := \sup_{x\in\mathbb{R}^d}\|\nabla^i\varphi(x)\|$ for a function φ. For tensors T and T' of the same order, $T : T'$ is their full contraction, as for matrices in (2.13). For $v \in \mathbb{R}^d$ and an integer $j \geqslant 1$, the tensor power $v^{\otimes j}$ is the tensor of order j with entries $v_{i_1}\cdots v_{i_j}$, which extends $v^{\otimes 2} = vv^\top$. For symmetric matrices A and B, $A \succeq B$ means that $A - B$ is positive semidefinite. When the chain $(Y_k)_{k\geqslant 0}$ of (NC-SGLM) has a unique invariant distribution μ_h, the distribution of Z_k when Y_k is distributed as μ_h is

$$\pi_h := \mu_h * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big). \tag{3.1}$$

Although $(Z_k)_{k\geqslant 0}$ does not form a Markov chain, we refer to π_h as the numerical invariant distribution of NC-SGLM, as the MSE in (2.7) is computed with respect to $(Z_k)_{k\geqslant 0}$.

The analysis uses the following assumptions on the potential, the test function and the stochastic gradient.

Assumption 1 *$U \in C^8(\mathbb{R}^d)$ satisfies*

$$|\nabla U(0)| \leqslant M, \qquad \max_{2\leqslant i\leqslant 8}\|\nabla^i U(x)\| \leqslant M, \qquad \forall x \in \mathbb{R}^d, \tag{3.2}$$

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

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

$$\max_{1\leqslant j\leqslant 4}\|\nabla^j f(x)\| \leqslant L, \qquad \forall x \in \mathbb{R}^d. \tag{3.3}$$

Assumption 3 *For every ξ, $g(\cdot;\xi) \in C^2(\mathbb{R}^d)$, with $(x,\xi) \mapsto \nabla g(x;\xi)$ measurable, and for every $x \in \mathbb{R}^d$,*

$$\Big(\mathbb{E}_\xi\big[|g(x;\xi)|^{24}\big]\Big)^{\frac{1}{24}} \leqslant M(1+|x|), \qquad \mathbb{E}_\xi\Big[\sup_{|x'-x|\leqslant 1}\|\nabla g(x';\xi)\|^2\Big] \leqslant M^2. \tag{3.4}$$

The Laplacian $\Delta\mathcal{C}$ of the covariance in (2.3), with entries $(\Delta\mathcal{C})_{ij} = \sum_{k=1}^d \partial_k^2\mathcal{C}_{ij}$, satisfies, for every $x \in \mathbb{R}^d$,

$$\|\Delta\mathcal{C}(x)\| \leqslant M(1+|x|)^2. \tag{3.5}$$

Assumption 1 is [30, Assumption 1] with the bound on $\nabla^8 U$ added, and Assumption 2 is [30, Assumption 2]. Compared with [30, Assumption 3], Assumption 3 requires the moment of order 24 instead of 12 and the bound (3.5), and drops the bound on the variance of the gradient noise uniform in x: the gradient noise may grow linearly in $|x|$ and its covariance quadratically. In particular, Assumption 3 covers the stochastic gradient $g(x;\xi) = (1+\xi)\nabla U(x)$ of the random potential $(1+\xi)U(x)$ with $\mathbb{E}\xi = 0$ and $\mathbb{E}|\xi|^{24} < \infty$, with M replaced by a constant depending only on M and $\mathbb{E}|\xi|^{24}$.

Throughout, C denotes a positive constant depending only on m, M, R and d, which changes from line to line, and $h_0 \in (0,1]$ an upper bound on h with the same dependence. Since the polynomial bound (2.6) involves no parameters other than h and λ, C and h_0 do not depend on the noise corrector. The notation $\mathcal{O}(\cdot)$ denotes a term bounded by C times its argument. Throughout this section, $\mathcal{S}_h$ is a noise corrector satisfying the polynomial bound (2.6).

3.2 Main theorems

Theorem 1 bounds the MSE (2.7) of NC-SGLM from a given initial value Y_0.

Theorem 1 *Let Assumptions 1, 2 and 3 hold. There are $h_0 \in (0,1]$ and C, depending only on m, M, R and d, such that for every $h \in (0,h_0]$, every batch size $P \geqslant 2$, every $N \geqslant 1$ and every $Y_0 \in \mathbb{R}^d$,*

$$\mathrm{MSE}(N,h) \leqslant CL^2\biggl[\bigl(1+|Y_0|\bigr)^2\frac{1}{Nh} + \bigl(1+|Y_0|\bigr)^{12}h^4\biggr]. \tag{3.6}$$

The bound (3.6) is the MSE bound of LM with the full gradient [30, Theorem 2], with Y_0 in place of Z_0, and it holds at every batch size $P \geqslant 2$. Theorem 1 shows that NC-SGLM has second order accuracy in the MSE. With h of order $\varepsilon^{1/2}$ and N of order $\varepsilon^{-5/2}$, the MSE is $\mathcal{O}(\varepsilon^2)$ as $\varepsilon \to 0$ already with $P = 2$, at the cost of N steps with $2P$ stochastic gradients each, that is, $\mathcal{O}(\varepsilon^{-5/2})$ stochastic gradients.

For $Y_0 \in \mathbb{R}^d$, let $(Y_0(t))_{t\geqslant 0}$ be the solution of the overdamped Langevin dynamics (2.1) with the initial value $Y_0(0) = Y_0$. Theorem 2 compares the law of Z_n with the law of $Y_0(nh)$, uniformly in n.

Theorem 2 *Let Assumptions 1, 2 and 3 hold. There are $h_0 \in (0,1]$, C and $\lambda > 0$, depending only on m, M, R and d, such that for every $h \in (0,h_0]$, every batch size $P \geqslant 2$, every $n \geqslant 0$ and every $Y_0 \in \mathbb{R}^d$,*

$$\bigl|\mathbb{E}[f(Z_n)] - \mathbb{E}[f(Y_0(nh))]\bigr| \leqslant CL\Bigl[\bigl(1+|Y_0|\bigr)^6 e^{-\lambda nh}\,h + h^2\Bigr]. \tag{3.7}$$

The term of first order in (3.7) decays exponentially in the time nh, and it is at most h^2 once $nh \geqslant \lambda^{-1}\log((1+|Y_0|)^6/h)$. For LM with the full gradient, the weak error has an expansion of the same form, in which the coefficient of h decays exponentially in time [17, Theorems 3.1 and 3.2].

Theorem 3 identifies the leading term of the numerical bias $\pi_h(f) - \pi(f)$.

Theorem 3 *Let Assumptions 1, 2 and 3 hold. There are $h_0 \in (0,1]$ and C, depending only on m, M, R and d, such that for every $h \in (0,h_0]$ and every batch size $P \geqslant 2$, the chain $(Y_k)_{k\geqslant 0}$ of (NC-SGLM) has a unique invariant distribution μ_h, and π_h in (3.1) satisfies*

$$\Bigl|\pi_h(f) - \pi(f) - \frac{h^2}{24}\,\pi\bigl(\nabla^3\phi_0 : \nabla^3 U\bigr)\Bigr| \leqslant CL\Bigl(\frac{h^2}{P} + h^3\Bigr), \tag{3.8}$$

where ϕ_0 solves the Poisson equation (2.11).

The bound (3.8) also implies

$$\varlimsup_{h\to 0}\,\Bigl|\frac{\pi_h(f)-\pi(f)}{h^2} - \frac{1}{24}\,\pi\bigl(\nabla^3\phi_0 : \nabla^3 U\bigr)\Bigr| \leqslant \frac{CL}{P}. \tag{3.9}$$

The full gradient $g = \nabla U$ satisfies Assumption 3, and with it $\hat{C}_P = 0$ and $\mathcal{S}_h(0) = 1$ by the polynomial bound (2.6), so that NC-SGLM is LM for every batch size $P \geqslant 2$. Letting $P \to \infty$ in (3.9) gives, for the numerical invariant distribution π_h^{LM} of LM,

$$\lim_{h\to 0}\frac{\pi_h^{\mathrm{LM}}(f) - \pi(f)}{h^2} = \frac{1}{24}\,\pi\bigl(\nabla^3\phi_0 : \nabla^3 U\bigr). \tag{3.10}$$

Hence, the leading term $\frac{h^2}{24}\,\pi(\nabla^3\phi_0 : \nabla^3 U)$ in (3.8) comes exactly from the discretization of LM, while the remaining term of order h^2 is bounded with the coefficient CL/P.

Theorems 1, 2 and 3 do not cover CC-SGLM, since $\mathcal{S}_h^{\mathrm{CC}}$ in (2.8) does not satisfy the polynomial bound (2.6). Under Assumption 3, the covariance $\mathcal{C}(x)$ can grow quadratically in $|x|$, and so can the noise correction $-\sqrt{2h}\,\frac{h}{4}\,\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\,\chi_k$ of CC-SGLM in $|Y_k|$. The proofs of the three theorems extend to CC-SGLM under a stronger assumption on the stochastic gradient, the uniform bound $\mathbb{E}_\xi\big[|g(x;\xi)-\nabla U(x)|^{24}\big] \leqslant M^{24}$ on the gradient noise for every $x \in \mathbb{R}^d$, in addition to Assumption 3; we omit the proof.

Applying the noise correction to SGLD gives the noise-corrected SGLD (NC-SGLD) integrator

$$Z_{k+1} = Z_k - h\,\hat{g}_P(Z_k;\boldsymbol{\xi}_k) + \sqrt{2h}\,\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Z_k;\boldsymbol{\xi}_k)\big)\chi_k, \tag{NC-SGLD}$$

in which the noise factor and the batch gradient are evaluated at the same point Z_k. Under Assumptions 1, 2 and 3, we can also prove that the numerical invariant distribution $\pi_h^{\text{NC-SGLD}}$ of (NC-SGLD) satisfies

$$\lim_{h\to 0}\frac{\pi_h^{\text{NC-SGLD}}(f)-\pi(f)}{h} = \frac{1}{2}\,\pi\big(\nabla^2\phi_0 : \nabla^2 U\big). \tag{3.11}$$

The right side is the term of the Euler–Maruyama discretization in (2.12): the noise correction removes the term of the gradient noise, and NC-SGLD remains of first order. We omit the proof, which is similar to those for NC-SGLM, due to the page limit.

3.3 Comparison to the results of SGLM and capping probability

The numerical bias of NC-SGLM has an error structure different from that of SGLM, which leads to a different choice of the step size h and the batch size P. When the covariance of the gradient noise is bounded uniformly in x, [30, Theorem 4] applied to the batch gradient $\hat{g}_P$ in (2.2) gives

$$\big|\pi_h^{\text{SGLM}}(f)-\pi(f)\big| \leqslant CL\Big(\frac{h}{P}+h^2\Big), \tag{3.12}$$

where the dominant term h/P comes from the variance of the stochastic gradient. In comparison, π_h^{SGLD} satisfies [30, Theorem 5]

$$\big|\pi_h^{\text{SGLD}}(f)-\pi(f)\big| \leqslant CLh, \tag{3.13}$$

where the term with $\nabla^2 U$ in (2.12) remains at every batch size. Hence the advantage of SGLM over SGLD lies mainly at large batch sizes. Moreover, each unit of time costs P/h evaluations of the stochastic gradient in SGLM, so halving h and doubling P have the same cost and both halve the term h/P in (3.12): SGLM has no preference between a smaller step size and a larger batch size.

For NC-SGLM, each unit of time costs $2P/h$ evaluations of the stochastic gradient. Halving h divides the leading term and the bound in (3.8) by at least 4, while doubling P only halves the term CLh^2/P of the bound and leaves the leading term unchanged. Hence, to achieve a sufficiently high accuracy, NC-SGLM favors a small step size h rather than a large batch size P. Like the second order accuracy at every batch size $P \geqslant 2$ in Theorems 1 and 3, this comparison describes a limit behavior as $h \to 0$.

At a finite step size, NC-SGLM is not always more accurate than SGLM. When h is moderately large, the remainder in (3.8) can be non-negligible. NC-SGLM may also have a larger numerical bias than SGLM at the same cost.

Intuitively, NC-SGLM is more accurate than SGLM at a given step size when its noise correction restores the fluctuation–dissipation relation. For RC, (2.10) holds in the step k exactly when $\mathcal{S}_h^{\mathrm{RC}}\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)$ is not capped. With $\lambda_{\max}$ the largest eigenvalue of a symmetric matrix, $\mathcal{S}_h^{\mathrm{RC}}\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)$ is capped exactly when

$$\lambda_{\max}\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big) > \frac{2}{h}, \tag{3.14}$$

and the estimate $h^2\,\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)$ of the covariance of the batch noise then exceeds the covariance $2hI_d$ of the injected noise along the eigenvectors of $\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)$ with eigenvalues above $2/h$. The condition (3.14) applies to every noise corrector satisfying the polynomial bound, through the remainder in (2.6): $\mathcal{S}_h(\lambda)-1$ differs

from its first order term $-\frac{h\lambda}{4}$ by at most $\frac{(h\lambda)^2}{8}$, which is smaller than $\frac{h\lambda}{4}$ exactly when $h\lambda < 2$. Under (3.14), this bound no longer separates the first order term from the remainder along the eigenvectors with eigenvalues above $2/h$, and the cancellation described after (2.15) can fail. With the grouping trick, $\hat{\mathcal{C}}_P$ in (3.14) is replaced by $\hat{\mathcal{C}}_{P,P_0}$ in (2.16).

To predict whether NC-SGLM is more accurate than SGLM at a given step size, one can therefore estimate the probability of (3.14): the smaller this probability, the more often the fluctuation–dissipation relation holds. Theorem 4 bounds this probability with Y_k replaced by a sample of π, for the batch approximation $\hat{\mathcal{C}}_P$ and for the grouped approximation $\hat{\mathcal{C}}_{P,P_0}$.

Theorem 4 *Let Assumptions 1 and 3 hold, and let $\mathbb{P}_{\boldsymbol{\xi}}$ be the probability over the batch $\boldsymbol{\xi}$ at fixed x. There is C, depending only on m, M, R and d, such that for every $h \in (0,1]$ and every batch size $P \geqslant 2$,*

$$\int_{\mathbb{R}^d} \mathbb{P}_{\boldsymbol{\xi}}\left(\lambda_{\max}\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) > \frac{2}{h}\right) \pi(x)\,\mathrm{d}x \leqslant C\left(\frac{h}{\sqrt{P}\,(P-1)}\right)^{12}, \tag{3.15}$$

and, for every $P_0 \geqslant 2$ dividing P,

$$\int_{\mathbb{R}^d} \mathbb{P}_{\boldsymbol{\xi}}\left(\lambda_{\max}\big(\hat{\mathcal{C}}_{P,P_0}(x;\boldsymbol{\xi})\big) > \frac{2}{h}\right) \pi(x)\,\mathrm{d}x \leqslant C\left(\frac{h\sqrt{P_0}}{P\,(P_0-1)}\right)^{12}. \tag{3.16}$$

The bound (3.15) decreases as h^{12} in h, but as P^{-18} in P. At a fixed cost of $2P/h$ evaluations of the stochastic gradient per unit of time, $h/(\sqrt{P}\,(P-1))$ is proportional to $\sqrt{P}/(P-1)$, so the bound decreases as P increases. Hence, at a finite step size, h is not the only parameter to reduce: the batch size P also needs to be large enough for (3.14) to be rare. With the grouping trick, the bound (3.16) still decreases as P increases at fixed h and P_0, and it exceeds (3.15) by a factor of order $(P/P_0)^6$, the price of the cheaper computation with the grouped approximation. The relation between the probability of (3.14) and the accuracy of NC-SGLM is only formal, and a rigorous comparison between NC-SGLM and SGLM at a finite step size is beyond the scope of this paper.

3.4 Explicit coefficients under Gaussian noise

The bound (3.9) determines the coefficient of h^2 in the numerical bias of NC-SGLM only up to a term bounded by CL/P. When the gradient noise is Gaussian, this coefficient can be computed explicitly. Define the LM coefficient, the limit in (3.10), by

$$\mathsf{K}^{\mathrm{LM}}(f) := \frac{1}{24}\,\pi\big(\nabla^3\phi_0 : \nabla^3 U\big), \tag{3.17}$$

where ϕ_0 depends on f through the Poisson equation (2.11). When U is a quadratic potential, the coefficient is immediately 0. In terms of $\nabla^2\phi_0$, the coefficient can be equivalently expressed as

$$\mathsf{K}^{\mathrm{LM}}(f) = \frac{1}{24}\,\pi\Big(\nabla^2\phi_0 : \big((\nabla U\cdot\nabla)\nabla^2 U - \nabla^2\Delta U\big)\Big), \tag{3.18}$$

where $((\nabla U\cdot\nabla)\nabla^2 U)_{ij} = \nabla U\cdot\nabla\partial_i\partial_j U$ and $(\nabla^2\Delta U)_{ij} = \partial_i\partial_j\Delta U$.

The explicit coefficient requires the square of the noise corrector to match the reduction of the injected covariance up to terms of order $(h\lambda)^3$. In this subsection, we also require the noise corrector to satisfy the following *additional polynomial bound*.

For every $h > 0$ and $\lambda \geqslant 0$, the noise corrector $\mathcal{S}_h$ additionally satisfies

$$\left|\mathcal{S}_h(\lambda)^2 - 1 + \frac{h\lambda}{2}\right| \leqslant \frac{(h\lambda)^3}{8}. \tag{3.19}$$

Among the noise correctors introduced in (2.9), the additional polynomial bound (3.19) applies only to RC. Furthermore, we assume the stochastic gradient $g(x;\xi)$ has the Gaussian form in Assumption 4.

Assumption 4 *The stochastic gradient is*

$$g(x;\xi) = \nabla U(x) + \eta(x)\,\xi, \tag{3.20}$$

with $\xi \sim \mathcal{N}(0, I_d)$ *and* $\eta \in C^4(\mathbb{R}^d; \mathbb{R}^{d\times d})$ *satisfying*

$$\max_{0\leqslant i\leqslant 4} \|\nabla^i \eta(x)\| \leqslant M, \qquad \forall x \in \mathbb{R}^d. \tag{3.21}$$

Under Assumption 4, the covariance (2.3) is $\mathcal{C}(x) = \eta(x)\,\eta(x)^\top$, and Assumption 3 holds with M replaced by a constant depending only on M and d, so that Theorem 3 applies.

Theorem 5 *Let Assumptions 1, 2 and 4 hold. There are* $h_0 \in (0,1]$ *and* C*, depending only on* m*,* M*,* R *and* d*, such that for every* $h \in (0, h_0]$ *and every batch size* $P \geqslant 2$*, the distribution* π_h *in* (3.1) *satisfies*

$$\left|\pi_h(f) - \pi(f) - h^2\Big(\mathsf{K}^{\mathrm{LM}}(f) + \frac{1}{P}\mathsf{K}^{\mathrm{SG}}(f)\Big)\right| \leqslant CLh^{\frac{5}{2}}, \tag{3.22}$$

where $\mathsf{K}^{\mathrm{LM}}(f)$ *is the LM coefficient* (3.17)*, and the additional coefficient from noise correction is*

$$\mathsf{K}^{\mathrm{SG}}(f) := \frac{1}{8}\,\pi\Big(\nabla^2\phi_0 : \big(4\,(\nabla U\cdot\nabla)\mathcal{C} - 3\,\Delta\mathcal{C} + 2\,\mathcal{C}\,\nabla^2 U\big)\Big), \tag{3.23}$$

with $((\nabla U\cdot\nabla)\mathcal{C})_{ij} = \nabla U\cdot\nabla\mathcal{C}_{ij}$*. As a consequence,*

$$\lim_{h\to 0}\frac{\pi_h(f)-\pi(f)}{h^2} = \mathsf{K}^{\mathrm{LM}}(f) + \frac{1}{P}\mathsf{K}^{\mathrm{SG}}(f). \tag{3.24}$$

In particular, when the covariance matrix is $\mathcal{C}(x) = \sigma^2 I_d$ with a constant $\sigma > 0$ (uniform Gaussian noise), (3.23) gives $\mathsf{K}^{\mathrm{SG}}(f) = \frac{\sigma^2}{4}\,\pi(\nabla^2\phi_0 : \nabla^2 U)$, and the coefficient of h^2 in the numerical bias of NC-SGLM has the explicit expression

$$\lim_{h\to 0}\frac{\pi_h(f)-\pi(f)}{h^2} = \frac{1}{24}\,\pi\big(\nabla^3\phi_0 : \nabla^3 U\big) + \frac{\sigma^2}{4P}\,\pi\big(\nabla^2\phi_0 : \nabla^2 U\big), \tag{3.25}$$

compared to (2.13) and (2.12). In all these cases, $\nabla^2\phi_0$ is the essential structure for characterizing the numerical bias: with (3.18) for the LM coefficient in (3.25), every coefficient in (3.25), and in (2.13) and (2.12) with $\mathcal{C} = \sigma^2 I_d$, pairs $\nabla^2\phi_0$ with a matrix that depends only on U, σ^2 and P. By the smoothing estimate of the semigroup of the overdamped Langevin dynamics (2.1) [29, Lemma 10], $\|\nabla^2\phi_0\|_\infty \leqslant C\|\nabla f\|_\infty$. Under Assumption 4, the expressions (3.18) and (3.23) then give $|\mathsf{K}^{\mathrm{LM}}(f)| + |\mathsf{K}^{\mathrm{SG}}(f)| \leqslant C\|\nabla f\|_\infty$, although the definition (3.17) involves the third order derivative of ϕ_0.

4 Numerical experiments

We test the noise correction in both SGLM and SGLD, and name each NC-SGLD integrator by its noise corrector, as for NC-SGLM. Section 4.1 verifies the explicit coefficient of Theorem 5 in one dimension, Section 4.2 compares the numerical biases in four dimensions, and Section 4.3 compares the four noise correctors in a Bayesian logistic regression.

4.1 Verifying the coefficients in a 1D double-well distribution

We verify the coefficient of h^2 in Theorem 5 with $d = 1$, the double-well potential and the test function

$$U(x) = \frac{x^2}{2} + 3e^{-\frac{x^2}{2}}, \qquad f(x) = \cos(2x).$$

For the gradient noise in (3.20), we take the uniform Gaussian noise with $\eta(x) = 1$ and the nonuniform Gaussian noise with $\eta(x) = 1 + \frac{1}{2}\cos(2x)$. Since $d = 1$, the value $\pi(f)$ and the exact coefficients $\mathsf{K}^{\mathrm{LM}}(f)$ in (3.17) and $\mathsf{K}^{\mathrm{SG}}(f)$ in (3.23) are computed by numerical quadrature.

We run RC-SGLM with the batch sizes $P = 4, 8, 16$ and the full gradient LM, each on 32 chains over the time duration $Nh = 2 \times 10^8$, at the step sizes $h = 2^{-1}, 2^{-\frac{3}{2}}, \dots, 2^{-4}$. The values $\pi_h(f)$ of RC-SGLM and $\pi_h^{\mathrm{LM}}(f)$ of LM are estimated by the mean over the chains of the time averages of $f(Z_k)$. Figure 1 shows the evaluated coefficients $(\pi_h(f) - \pi(f))/h^2$ and $(\pi_h^{\mathrm{LM}}(f) - \pi(f))/h^2$, together with the exact coefficients $\mathsf{K}^{\mathrm{LM}}(f) + \mathsf{K}^{\mathrm{SG}}(f)/P$ of RC-SGLM in (3.24), and $\mathsf{K}^{\mathrm{LM}}(f)$ of the full gradient LM.

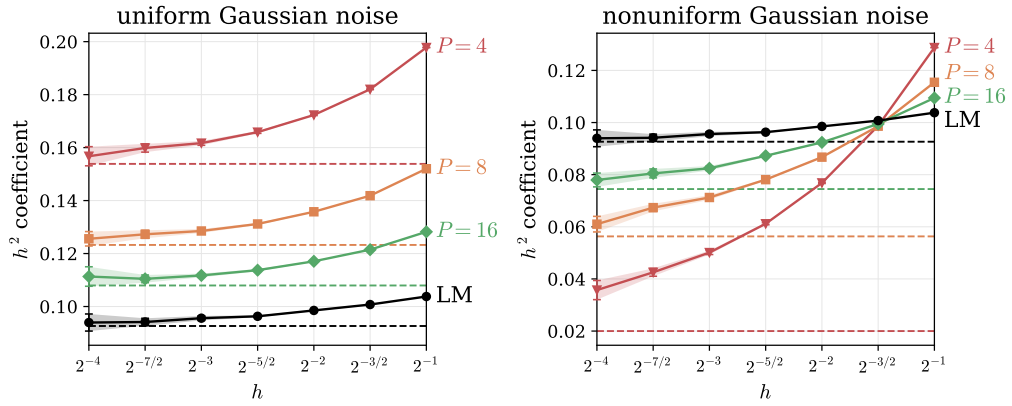


Figure 1: The evaluated coefficients $(\pi_h(f)-\pi(f))/h^2$ of RC-SGLM with $P = 4, 8, 16$ and $(\pi_h^{\mathrm{LM}}(f)-\pi(f))/h^2$ of LM against the step size h, with error bars and shaded bands of 2 standard errors, under the uniform Gaussian noise with $\eta(x) = 1$ (left) and the nonuniform Gaussian noise with $\eta(x) = 1 + \frac{1}{2}\cos(2x)$ (right). The curve of LM is the same in both panels. The dashed lines are the exact coefficients $\mathsf{K}^{\mathrm{LM}}(f) + \mathsf{K}^{\mathrm{SG}}(f)/P$ of RC-SGLM and $\mathsf{K}^{\mathrm{LM}}(f)$ of LM, in the colors of the corresponding curves.

The evaluated coefficient differs from the exact coefficient by the remainder of (3.22) divided by h^2, and in Figure 1 this difference tends to 0 roughly linearly in h, with a visible curvature at the large step sizes. We therefore fit a polynomial of degree 2 in h to the evaluated coefficients of each curve by least squares, weighted by the inverse squared standard errors, and take its value at $h = 0$ as the fitted coefficient. In Table 1, the fitted coefficients are close to the exact coefficients. Their 2 standard errors account for the fluctuation of the time averages but not for the error of the polynomial fit, and the largest difference occurs for RC-SGLM at $P = 4$ under the nonuniform noise, the curve that depends most strongly on h.

		$P = 4$	$P = 8$	$P = 16$	LM
uniform	fitted	0.1532 ± 0.0010	0.1230 ± 0.0009	0.1070 ± 0.0009	0.0917 ± 0.0008
	exact	0.1539	0.1232	0.1079	0.0926
nonuniform	fitted	0.0223 ± 0.0009	0.0569 ± 0.0009	0.0742 ± 0.0009	0.0917 ± 0.0008
	exact	0.0200	0.0563	0.0745	0.0926

Table 1: The fitted coefficients, with 2 standard errors, and the exact coefficients $\mathsf{K}^{\mathrm{LM}}(f) + \mathsf{K}^{\mathrm{SG}}(f)/P$ of RC-SGLM and $\mathsf{K}^{\mathrm{LM}}(f)$ of LM, under the uniform and the nonuniform Gaussian noise. The fitted coefficient is the value at $h = 0$ of the polynomial of degree 2 in h fitted by weighted least squares to the evaluated coefficients in Figure 1.

Under the uniform noise, $\mathsf{K}^{\mathrm{SG}}(f) > 0$, and the evaluated coefficients of RC-SGLM lie above the evaluated coefficient of LM and approach it as P increases, as the term $\mathsf{K}^{\mathrm{SG}}(f)/P$ in (3.24) indicates.

Under the nonuniform noise, the terms $4\,(\nabla U \cdot \nabla)\mathcal{C} - 3\,\Delta\mathcal{C}$ in (3.23), which involve the derivatives of $\mathcal{C} = \eta^2$, give a negative part of $\mathsf{K}^{\mathrm{SG}}(f)$, larger in absolute value than the positive part from the term $2\,\mathcal{C}\,\nabla^2 U$, so $\mathsf{K}^{\mathrm{SG}}(f) < 0$. At $h = 2^{-1}$, where the remainder of (3.22) is still large, the evaluated coefficients of RC-SGLM lie above the evaluated coefficient of LM and decrease as P increases. For $h \leqslant 2^{-2}$, they lie below the evaluated coefficient of LM, the smallest at $P = 4$: RC-SGLM with batches of 4 stochastic gradients has a smaller numerical bias than LM with the full gradient. The analysis of Section 3.4 is essential to understand this observation, since it separates the error: the coefficient of h^2 in (3.24) is the sum of $\mathsf{K}^{\mathrm{LM}}(f)$ from the

discretization of LM and $\mathsf{K}^{\mathrm{SG}}(f)/P$ from the noise correction, and the sign of $\mathsf{K}^{\mathrm{SG}}(f)$ depends on $\mathcal{C}$ through (3.23). The upper bound (3.9) controls $\mathsf{K}^{\mathrm{SG}}(f)/P$ only in size, and with such upper bounds alone, it is difficult to explain why RC-SGLM with $P = 4$ has an even smaller error than LM with the full gradient.

4.2 Comparing the bias in a 4D Gaussian mixture distribution

The target π is a mixture of the Gaussian distributions $\mathcal{N}(\vartheta_j, \Sigma_j)$ on $\mathbb{R}^4$, $j = 1, \dots, 10$, with weights proportional to $(3/4)^{j-1}$, means ϑ_j drawn at random in the ball of radius 4.5 centered at the origin, and covariance matrices Σ_j drawn at random. With the coordinates $x_1, \dots, x_4$ of x, the stochastic gradient is

$$g(x;\xi) = \nabla U(x) + \xi \left(x_1 \sin x_1, \dots, x_4 \sin x_4\right), \qquad \xi \sim \mathcal{N}(0,1),$$

and the test function is

$$f(x) = e^{-\frac{1}{2}|x - \frac{1}{2}(\vartheta_1 + \vartheta_5)|^2}.$$

The integrators are compared at 4, 8 and 16 evaluations of the stochastic gradient per step, with $P = 4, 8, 16$ for SGLD, RC-SGLD and SGLM and $P = 2, 4, 8$ for RC-SGLM. At each step size $h = 2^{-1}, 2^{-\frac{3}{2}}, \dots, 2^{-4}$, every integrator runs 32 chains over the time duration $Nh = 4 \times 10^8$, and its numerical bias for f is estimated using the mean over the chains of the time averages of $f(Z_k)$.

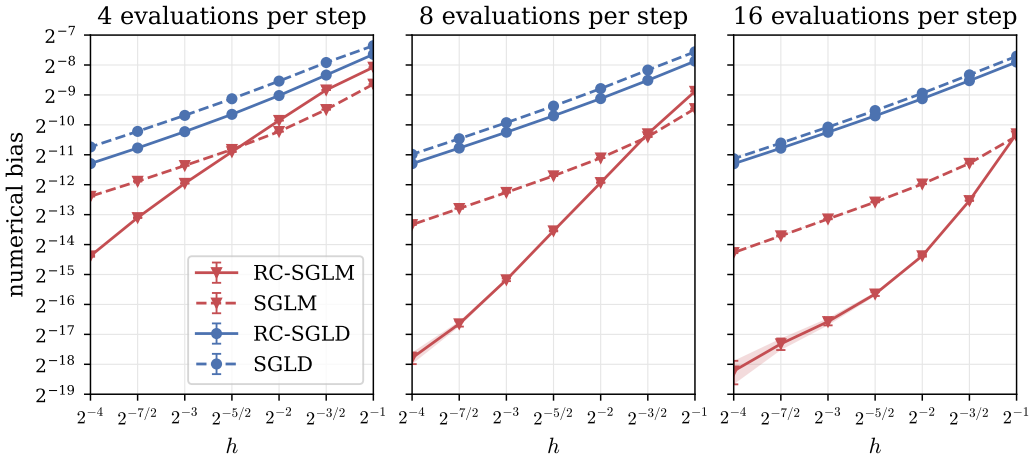


Figure 2: The absolute value of the numerical bias for f of SGLD, RC-SGLD, SGLM and RC-SGLM against the step size h, with error bars and shaded bands of 2 standard errors, at 4 (left), 8 (middle) and 16 (right) evaluations of the stochastic gradient per step.

	RC-SGLM	SGLM	RC-SGLD	SGLD
4 evals	**2.11**	1.23	1.22	1.14
8 evals	**3.05**	1.25	1.14	1.14
16 evals	**2.54**	1.26	1.13	1.14

Table 2: The fitted orders of the numerical bias of RC-SGLM, SGLM, RC-SGLD and SGLD at 4, 8 and 16 evaluations of the stochastic gradient per step. The fitted orders above 2 are in bold.

In every panel of Figure 2, the numerical bias of SGLM is below those of SGLD and RC-SGLD and decreases at about first order, as (2.13) suggests, while SGLD and RC-SGLD keep the discretization error of first order of the Euler–Maruyama step. The numerical bias of RC-SGLM is the smallest at the small step sizes in every panel. Table 2 lists the fitted order of each curve, the slope of the least squares line of the logarithm of the absolute value of the numerical bias against the logarithm of h over all step sizes: the fitted order is above 2 for RC-SGLM in every row, and between 1 and 1.3 for the other integrators.

Table 3 relates the smaller bias of RC-SGLM to the capping condition (3.14). The table lists the capping probability, the probability of (3.14) with Y_k replaced by an exact sample of π, and the bias ratio, the numerical bias of SGLM divided by that of RC-SGLM at the same number of evaluations per step. Given $Y_k = y$, the probability of (3.14) is explicit for this stochastic gradient, and the capping probability is its mean over 10^8 samples y of π. The 2 standard errors of the bias ratio are below 10% of its value, except at $h = 2^{-4}$ with 8 evaluations per step and at $h \leqslant 2^{-\frac{7}{2}}$ with 16 evaluations per step, where they are at most 27%. The capping probability decreases as h decreases, and at each step size it is smaller at a larger batch size, by orders of magnitude at the small step sizes, in line with the bound (3.15). Doubling P and halving h both double the cost per unit of time, and wherever both are possible in the table, doubling P reduces the capping probability at least as much as halving h, and much more at the small step sizes, for example by a factor of about 3×10^3 against 37 at $h = 2^{-3}$ with 8 evaluations per step. In every row, the bias ratio exceeds 1 from some step size on and keeps increasing as h decreases, and this step size is larger where the capping probability is smaller. At $h \leqslant 2^{-\frac{7}{2}}$ with 8 and 16 evaluations per step, the numerical bias of RC-SGLM is more than an order of magnitude smaller than that of SGLM. At $h = 2^{-1}$, where the remainder in (3.8) is non-negligible, the bias ratio is below 1 in every row, even at 16 evaluations per step, where the capping probability is a few percent.

		$h = 2^{-1}$	$h = 2^{-\frac{3}{2}}$	$h = 2^{-2}$	$h = 2^{-\frac{5}{2}}$	$h = 2^{-3}$	$h = 2^{-\frac{7}{2}}$	$h = 2^{-4}$
4 evals	capping	3.0×10^{-1}	2.3×10^{-1}	1.6×10^{-1}	1.1×10^{-1}	6.7×10^{-2}	3.8×10^{-2}	1.9×10^{-2}
	ratio	0.675	0.635	0.773	**1.06**	**1.51**	**2.32**	**3.97**
8 evals	capping	1.6×10^{-1}	9.2×10^{-2}	4.7×10^{-2}	2.0×10^{-2}	6.7×10^{-3}	1.5×10^{-3}	1.8×10^{-4}
	ratio	0.670	0.932	**1.77**	**3.58**	**7.57**	**14.5**	**21.9**
16 evals	capping	3.4×10^{-2}	9.6×10^{-3}	1.4×10^{-3}	9.0×10^{-5}	2.1×10^{-6}	3.2×10^{-8}	5.1×10^{-10}
	ratio	0.948	**2.37**	**5.27**	**8.41**	**10.8**	**12.2**	**15.5**

Table 3: The capping probability of RC-SGLM with $P = 2, 4, 8$, the probability of (3.14) with Y_k replaced by an exact sample of π, and the bias ratio of SGLM to RC-SGLM at 4, 8 and 16 evaluations of the stochastic gradient per step. The ratios greater than 1 are in bold.

4.3 Comparing the noise correctors in a 5D Bayesian logistic regression

The banknote authentication data set [18] contains $n = 1372$ genuine and forged banknotes, each with a label $y_i \in \{0, 1\}$ and 4 features extracted from its image. The target π is the posterior of the Bayesian logistic regression with the prior $\mathcal{N}(0, I_5)$ and the inputs $a_i \in \mathbb{R}^5$, formed by the features, centered and scaled to unit variance, and a constant 1. Its potential is

$$U(x) = \frac{|x|^2}{2} + \sum_{i=1}^{n} \ell_i(x), \qquad \ell_i(x) = \log\bigl(1 + e^{a_i \cdot x}\bigr) - y_i\, a_i \cdot x,$$

and the stochastic gradient is $g(x; \xi) = x + n\nabla\ell_\xi(x)$ with ξ uniform on $\{1, \dots, n\}$.

All integrators use the step size $h = 2^{-9}$ and 384 evaluations of the stochastic gradient per step: $P = 192$ for RC-SGLM, SC-SGLM, IC-SGLM and CC-SGLM, and $P = 384$ for SGLM, SGLD, RC-SGLD, SC-SGLD, IC-SGLD and CC-SGLD. Every integrator runs 32 chains over the time duration $Nh = 10^5$. The posterior means and variances of the 5 coordinates of x are estimated from the time averages of Z_k and of the squares of its coordinates over all chains, and compared with the means and variances of 6.4×10^7 draws of the No-U-Turn sampler (NUTS) [13]. The maximal bias is the largest absolute error of the posterior means, and the variance error is the largest absolute relative error of the posterior variances.

In the right panel of Figure 3 and in Table 4, RC-SGLD, SC-SGLD, IC-SGLD and CC-SGLD reach a common plateau, at about half the maximal bias and variance error of SGLD. As (3.11) predicts, the noise correction leaves the term of the Euler–Maruyama discretization, which depends neither on the noise corrector nor on the batch size. This term is about as large as the term with $\mathcal{C}$ in (2.12): the maximal bias of SGLD is about the sum of those of SGLM and of the NC-SGLD integrators.

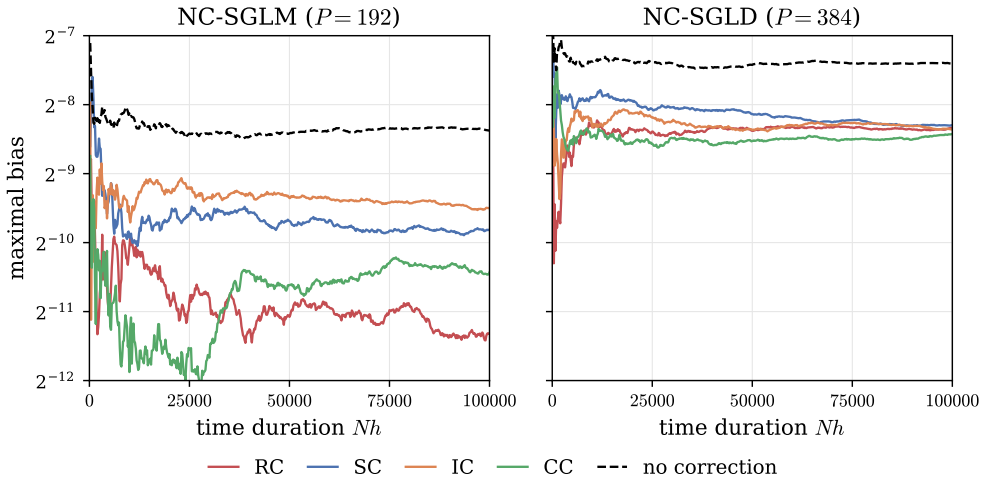


Figure 3: The maximal bias of RC-SGLM, SC-SGLM, IC-SGLM and CC-SGLM with $P = 192$ and SGLM with $P = 384$ (left), and of RC-SGLD, SC-SGLD, IC-SGLD, CC-SGLD and SGLD with $P = 384$ (right), against the time duration Nh, at $h = 2^{-9}$ with 384 evaluations of the stochastic gradient per step. The colors indicate the noise correctors, and the dashed lines are SGLM (left) and SGLD (right).

$P = 192$ (SGLM: $P = 384$)	RC-SGLM	SC-SGLM	IC-SGLM	CC-SGLM	SGLM
maximal bias ($\times 10^{-3}$)	$\mathbf{0.39 \pm 0.32}$	1.11 ± 0.30	1.38 ± 0.26	0.72 ± 0.29	3.01 ± 0.25
variance error ($\times 10^{-2}$)	$\mathbf{0.27 \pm 0.06}$	0.57 ± 0.06	0.91 ± 0.04	0.53 ± 0.05	3.12 ± 0.06
running time (min)	11.1	11.4	11.2	8.0	6.8
$P = 384$	RC-SGLD	SC-SGLD	IC-SGLD	CC-SGLD	SGLD
maximal bias ($\times 10^{-3}$)	3.09 ± 0.32	3.18 ± 0.26	3.04 ± 0.34	2.91 ± 0.29	5.90 ± 0.25
variance error ($\times 10^{-2}$)	2.49 ± 0.05	2.49 ± 0.04	2.54 ± 0.03	2.51 ± 0.04	5.27 ± 0.05
running time (min)	14.8	14.8	14.7	9.6	6.8

Table 4: The maximal bias and the variance error at $Nh = 10^5$, with 2 standard errors over the 32 chains, and the running time of the integrators of Figure 3, NC-SGLM and SGLM (top) and NC-SGLD and SGLD (bottom). The running time is measured on an AMD Ryzen 9 9950X3D with the 32 chains run in parallel on 32 threads. The smallest maximal bias and variance error are in bold.

In the left panel and in Table 4, every NC-SGLM integrator has a smaller maximal bias and variance error than SGLM and than every NC-SGLD integrator, and RC-SGLM has the smallest, by a clear margin in the variance error. The capping probability of RC-SGLM, estimated from 10^5 draws of NUTS, is about 10^{-3}, so (2.10) fails only in a fraction of about 10^{-3} of the steps, while CC, SC and IC, which do not satisfy (3.19), leave a term of order h^3 in $2h\,\mathcal{S}_h\big(\hat{\mathcal{C}}_P\big)^2$ in every step.

In Table 4, every noise-corrected integrator runs slower than SGLM and SGLD at the same number of evaluations of the stochastic gradient per step, even with $d = 5$, since it also computes the noise factor. By Section 2.4, the linear noise factor of CC costs order dP and adds the least time, and the noise factors of RC, SC and IC, computed by the eigendecomposition of $\hat{\mathcal{C}}_P$, cost order $d^2P + d^3$ and add the most.

5 Conclusion

To the best of our knowledge, NC-SGLM is the first class of stochastic gradient integrators with second order accuracy for nonconvex potentials and a general stochastic gradient. Its construction is explicit: each step multiplies the injected noise of SGLM by the noise factor, which applies a noise corrector satisfying the polynomial bound (2.6) to the batch approximation (2.5). For every such noise corrector and every

batch size $P \geqslant 2$, Theorems 1, 2 and 3 bound the MSE by a constant times $(Nh)^{-1} + h^4$, the weak error uniformly in time, and the numerical bias by a constant times h^2. Under Gaussian gradient noise and the additional polynomial bound (3.19), Theorem 5 separates the coefficient of h^2 in the numerical bias into the LM coefficient $\mathsf{K}^{\mathrm{LM}}(f)$ of the discretization and the coefficient $\mathsf{K}^{\mathrm{SG}}(f)/P$ of the noise correction.

NC-SGLM evaluates each batch at both Y_k and Z_k, and thus uses twice as many stochastic gradients per step as SGLM with the same batch size. In high dimensions, both the cost and the accuracy of NC-SGLM depend on the noise factor: its exact computation costs order $d^2P + d^3$ per step, or $dP^2 + P^3$ when $P < d$, the grouped approximation raises the bound (3.16) on the capping probability, and CC does not satisfy the polynomial bound. At a finite step size, NC-SGLM can be less accurate than SGLM, and its comparison with SGLM rests on the capping probability, whose relation to the accuracy is only formal.

Future work includes the noise correction of stochastic gradient integrators for the underdamped Langevin dynamics, more efficient implementations of the noise factor, the combination with variance reduction approaches, and realistic applications in high-dimensional Bayesian inference.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: a non-asymptotic analysis. In *Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323, 2018.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Xunpeng Huang, Difan Zou, Hanze Dong, Yi-An Ma, and Tong Zhang. Faster sampling via stochastic gradient proximal sampler. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20559–20596, 2024.

[15] Ryogo Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966. doi: 10.1088/0034-4885/29/1/306.

[16] 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.

[17] 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.

[18] Volker Lohweg. Banknote authentication. UCI Machine Learning Repository, 2012.

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

[20] 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.

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

[22] 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.

[23] Abhishek Roy, Lingqing Shen, Krishnakumar Balasubramanian, and Saeed Ghadimi. Stochastic zeroth-order discretizations of Langevin diffusions for Bayesian inference. *Bernoulli*, 28(3):1810–1834, 2022. doi: 10.3150/21-BEJ1400.

[24] 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.

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

[26] 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.

[27] 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.

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

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

[30] Xuda Ye, Yuliang Wang, Guang Lin, and Cindy Xiangrui Kong. Stochastic gradient Leimkuhler–Matthews integrator with second order discretization error. arXiv:2610.24781, 2026.

A Proof of Theorem 1: MSE estimate

The proof of Theorem 1 follows the proof of the MSE bound of LM in [30, Theorem 2]. Appendix A.1 sets up the probability space, the splitting of a step of NC-SGLM and the bounds on the stochastic gradient and the noise factor used below. Appendix A.2 recalls the modified class of [30, Section 4], with which the MSE is bounded by the mean square of a time average error along the chain $(Y_k)_{k\geqslant 0}$ and a term of order $\frac{1}{Nh}+h^4$. Lemma 1 in Appendix A.3 bounds the moments of the chain and gives its invariant distribution μ_h, Lemma 2 in Appendix A.4 shows that the local weak error of NC-SGLM against the modified dynamics is of order h^3, and Appendix A.5 combines them into Theorem 1. Throughout this appendix, the noise corrector $\mathcal{S}_h$ satisfies the polynomial bound (2.6), and it enters the proofs only through the bounds (A.10) on the noise factor. The thresholds on the step size in the results of [30] used below are included in h_0.

A.1 Setting

The proofs of Lemmas 1 and 2 control one step of NC-SGLM through moment bounds on the batch gradient, the batch approximation $\hat{\mathcal{C}}_P$ and the noise correction, uniform in the batch size P.

The Gaussian vectors are the increments $\chi_k = h^{-1/2}(B_{(k+1)h} - B_{kh})$ of the Brownian motion $(B_t)_{t\geqslant 0}$, independent of the batches $(\boldsymbol{\xi}_k)_{k\geqslant 0}$, and the initial value Y_0 is independent of both. With the filtration

$$\mathcal{F}_k := \sigma\big(Y_0,\ B_s,\ \boldsymbol{\xi}_j : 0\leqslant s\leqslant kh,\ 0\leqslant j\leqslant k-1\big), \qquad \mathbb{E}_k[\,\cdot\,] := \mathbb{E}\big[\,\cdot\,\big|\mathcal{F}_k\big], \tag{A.1}$$

the state Y_k is $\mathcal{F}_k$-measurable, and $\mathcal{F}_k$, $(B_{kh+s}-B_{kh})_{s\geqslant 0}$ and $\boldsymbol{\xi}_k$ are independent.

The moment bound in (3.4), Jensen's inequality over the batch and $|\nabla U(x)|\leqslant M(1+|x|)$ by (3.2) give, for $x\in\mathbb{R}^d$ and $1\leqslant q\leqslant 24$,

$$\mathbb{E}_{\boldsymbol{\xi}}\big[|\hat{g}_P(x;\boldsymbol{\xi})|^q\big]\leqslant M^q(1+|x|)^q, \qquad \mathbb{E}_{\boldsymbol{\xi}}\big[|\hat{g}_P(x;\boldsymbol{\xi})-\nabla U(x)|^q\big]\leqslant C(1+|x|)^q. \tag{A.2}$$

Since the variance of $g(x;\xi)$ is at most its moment of order 2, (3.4) bounds the trace of the covariance matrix,

$$\mathrm{Tr}\big[\mathcal{C}(x)\big] = \mathbb{E}_\xi\big[|g(x;\xi)-\nabla U(x)|^2\big]\leqslant M^2(1+|x|)^2, \qquad x\in\mathbb{R}^d, \tag{A.3}$$

and the trace of (2.4) gives, for $x\in\mathbb{R}^d$,

$$\mathbb{E}_{\boldsymbol{\xi}}\big[|\hat{g}_P(x;\boldsymbol{\xi})-\nabla U(x)|^2\big]\leqslant \frac{M^2(1+|x|)^2}{P}. \tag{A.4}$$

The supremum of the Jacobian of the batch gradient over the unit ball around x,

$$\bar{\Lambda}(x;\boldsymbol{\xi}) := \sup_{|x'-x|\leqslant 1}\|\nabla\hat{g}_P(x';\boldsymbol{\xi})\|, \tag{A.5}$$

satisfies, by Minkowski's inequality over the batch and (3.4),

$$\mathbb{E}_{\boldsymbol{\xi}}\big[\bar{\Lambda}(x;\boldsymbol{\xi})^2\big]\leqslant M^2, \tag{A.6}$$

and, by the definition of $\bar{\Lambda}$ for $|x-x'|\leqslant 1$ and by $|\nabla U(x)-\nabla U(x')|\leqslant M|x-x'|$ for $|x-x'|>1$,

$$|\hat{g}_P(x;\boldsymbol{\xi})-\hat{g}_P(x';\boldsymbol{\xi})|\leqslant\Big(\bar{\Lambda}(x';\boldsymbol{\xi})+M+|\hat{g}_P(x;\boldsymbol{\xi})-\nabla U(x)|+|\hat{g}_P(x';\boldsymbol{\xi})-\nabla U(x')|\Big)|x-x'|, \quad x,x'\in\mathbb{R}^d. \tag{A.7}$$

For $|x-x'| \leqslant 1$, the bound on ∇g in (3.4) and $\|\nabla^2 U\| \leqslant M$ give

$$\mathbb{E}_\xi\big[|g(x;\xi) - \nabla U(x) - g(x';\xi) + \nabla U(x')|^2\big] \leqslant 4M^2|x-x'|^2.$$

Since $\mathcal{C}(x)$ is positive semidefinite, (A.3), this bound and the Cauchy–Schwarz inequality give, for $A \in \mathbb{R}^{d\times d}$

$$|A : \mathcal{C}(x)| \leqslant M^2(1+|x|)^2\,\|A\|, \quad \|\mathcal{C}(x) - \mathcal{C}(x')\| \leqslant 4M^2(1+|x|+|x'|)^2\,|x-x'|, \quad x, x' \in \mathbb{R}^d. \tag{A.8}$$

For a batch size $P \geqslant 2$, the batch approximation $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$ in (2.5) is positive semidefinite, so that $\|\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\| \leqslant \operatorname{Tr}\big[\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big]$, and replacing $\hat{g}_P(x;\boldsymbol{\xi})$ by 0 does not decrease $\operatorname{Tr}\big[\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big]$. Hence

$$\|\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\| \leqslant \frac{1}{P(P-1)}\sum_{p=1}^{P} |g(x;\xi^{(p)})|^2,$$

and Jensen's inequality with the moment bound in (3.4) gives

$$\mathbb{E}_{\boldsymbol{\xi}}\big[\|\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\|^r\big] \leqslant \frac{M^{2r}(1+|x|)^{2r}}{(P-1)^r}, \qquad \frac{1}{2} \leqslant r \leqslant 12, \qquad x \in \mathbb{R}^d. \tag{A.9}$$

Since the noise factor has the eigenvectors of $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$, whose eigenvalues lie in $[0, \|\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\|]$, the polynomial bound (2.6) gives the bounds on the noise factor

$$\begin{aligned} \Big\|\mathcal{S}_h\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) - I_d\Big\| &\leqslant \min\Big\{1, \frac{h}{2}\,\|\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\|\Big\}, \\ \Big\|\mathcal{S}_h\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) - I_d + \frac{h}{4}\,\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\Big\| &\leqslant \frac{h^2}{8}\,\|\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\|^2. \end{aligned} \tag{A.10}$$

Since $\mathcal{S}_h$ is measurable, the noise factor $\mathcal{S}_h\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big)$ is a measurable function of $(x,\boldsymbol{\xi})$.

The step of (NC-SGLM) from Y_k to Y_{k+1} splits as

$$Y_{k+1} = Y_k + D_k + \zeta_k, \qquad D_k := \sqrt{2h}\,\chi_k - h\,\hat{g}_P(Z_k;\boldsymbol{\xi}_k), \qquad \zeta_k := \sqrt{2h}\,\Big(\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big) - I_d\Big)\chi_k, \tag{A.11}$$

where D_k is the SGLM displacement, the displacement of one step of (SGLM) from Y_k, and ζ_k is the noise correction. By (A.10), the noise correction is bounded by

$$|\zeta_k| \leqslant \sqrt{2h}\,|\chi_k| \min\Big\{1, \frac{h}{2}\,\|\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\|\Big\} \leqslant \min\Big\{h\,\|\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\|^{\frac{1}{2}},\ \frac{h}{2}\sqrt{2h}\,\|\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\|\Big\}|\chi_k|. \tag{A.12}$$

Since $\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)$ is independent of χ_k given $\mathcal{F}_k$, (A.12) and (A.9) bound the moments of ζ_k. Given $\mathcal{F}_k$ and χ_k, (A.2) at $Z_k = Y_k + \sqrt{h/2}\,\chi_k$ bounds the moments of $\hat{g}_P(Z_k;\boldsymbol{\xi}_k)$ and of D_k. Therefore, for $1 \leqslant q \leqslant 12$,

$$\begin{aligned} \mathbb{E}_k|D_k|^q &\leqslant Ch^{\frac{q}{2}} + Ch^q\big(1+|Y_k|\big)^q, \\ \mathbb{E}_k|\zeta_k|^q &\leqslant C\min\Big\{h^q\big(1+|Y_k|\big)^q,\ h^{\frac{3q}{2}}\big(1+|Y_k|\big)^{2q}\Big\}. \end{aligned} \tag{A.13}$$

A.2 Modified discrete Poisson equation

The analysis runs on the chain $(Y_k)_{k\geqslant 0}$, and the modified class absorbs the Gaussian shift $Z_k - Y_k = \sqrt{h/2}\,\chi_k$ of (NC-SGLM) through a modified test function, potential, generator and discrete Poisson solution. The modified test function is the heat flow of f,

$$\tilde{f}_h(y) := \mathbb{E}\Big[f\Big(y + \sqrt{\frac{h}{2}}\,\chi\Big)\Big], \qquad \chi \sim \mathcal{N}(0, I_d), \tag{A.14}$$

so that, since χ_k is independent of $\mathcal{F}_k$,

$$\mathbb{E}_k\big[f(Z_k)\big] = \tilde{f}_h(Y_k). \tag{A.15}$$

The modified potential of [30, Section 4.2] is

$$\tilde{U}_h(x) := U(x) + \frac{h}{4}\big(|\nabla U(x)|^2 - \Delta U(x)\big)\,\theta\Big(\frac{hx}{\delta}\Big), \qquad x \in \mathbb{R}^d, \tag{A.16}$$

with a smooth cutoff θ, $0 \leqslant \theta \leqslant 1$, equal to 1 on $\{|x| \leqslant 1\}$ and to 0 on $\{|x| \geqslant 2\}$, and the constant $\delta \in (0, 1]$ of [30, Lemma 1], chosen in its proof sufficiently small, independent of h, such that

$$\max_{1 \leqslant i \leqslant 5} \|\nabla^i \tilde{U}_h(x) - \nabla^i U(x)\| \leqslant \min\Big\{\frac{m}{2},\, M\Big\}, \qquad \forall x \in \mathbb{R}^d, \quad h \in (0, \delta]. \tag{A.17}$$

The modified target and the generator of the modified dynamics, the overdamped Langevin dynamics (2.1) with the potential $\tilde{U}_h$, are

$$\tilde{\pi}_h(x) \propto e^{-\tilde{U}_h(x)}, \qquad \tilde{\mathcal{L}}_h := \Delta - \nabla \tilde{U}_h \cdot \nabla. \tag{A.18}$$

For each k, the solution $(\tilde{Y}_k(s))_{s \geqslant 0}$ of the modified dynamics from Y_k, driven by $(B_{kh+s} - B_{kh})_{s \geqslant 0}$, satisfies

$$\mathrm{d}\tilde{Y}_k(s) = -\nabla \tilde{U}_h\big(\tilde{Y}_k(s)\big)\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}B_{kh+s}, \qquad \tilde{Y}_k(0) = Y_k, \tag{A.19}$$

so that, with $e^{t\tilde{\mathcal{L}}_h}\psi(y)$ the mean of ψ at the time $t \geqslant 0$ of the modified dynamics from y, for every measurable ψ of polynomial growth,

$$\mathbb{E}_k\big[\psi(\tilde{Y}_k(h))\big] = \big(e^{h\tilde{\mathcal{L}}_h}\psi\big)(Y_k). \tag{A.20}$$

The modified discrete Poisson solution $\tilde{\phi}_h$ and the modified discrete Poisson equation it solves are

$$\tilde{\phi}_h := h\sum_{j=0}^{\infty} e^{jh\tilde{\mathcal{L}}_h}\big(\tilde{f}_h - \tilde{\pi}_h(\tilde{f}_h)\big), \qquad \tilde{\phi}_h - e^{h\tilde{\mathcal{L}}_h}\tilde{\phi}_h = h\big(\tilde{f}_h - \tilde{\pi}_h(\tilde{f}_h)\big). \tag{A.21}$$

The modified class involves U and f only, not the stochastic gradient. Under Assumptions 1 and 2 and (A.17), [30, Lemma 2 and Corollary 1] give the convergence of the series in (A.21), and the bounds on the modified bias $\tilde{\pi}_h(\tilde{f}_h) - \pi(f)$ and on the derivatives of $\tilde{\phi}_h$

$$\big|\tilde{\pi}_h(\tilde{f}_h) - \pi(f)\big| \leqslant CLh^2, \qquad \max_{1 \leqslant j \leqslant 4} \|\nabla^j \tilde{\phi}_h(x)\| \leqslant CL, \qquad \forall x \in \mathbb{R}^d. \tag{A.22}$$

By (A.15), the differences $f(Z_k) - \tilde{f}_h(Y_k)$ are martingale differences, and, since f is L-Lipschitz,

$$\mathbb{E}_k\big[\big(f(Z_k) - \tilde{f}_h(Y_k)\big)^2\big] \leqslant \mathbb{E}_k\big[\big(f(Z_k) - f(Y_k)\big)^2\big] \leqslant L^2\,\mathbb{E}_k|Z_k - Y_k|^2 = \frac{dh}{2}\,L^2.$$

With $f(Z_k) = \tilde{f}_h(Y_k) + (f(Z_k) - \tilde{f}_h(Y_k))$ in (2.7), the Cauchy–Schwarz inequality, the orthogonality of these martingale differences and the bound on the modified bias in (A.22) bound the MSE by the mean square of the time average error $\frac{1}{N}\sum_{k=0}^{N-1}(\tilde{f}_h(Y_k) - \tilde{\pi}_h(\tilde{f}_h))$ and a term of order $\frac{1}{Nh} + h^4$:

$$\mathrm{MSE}(N, h) \leqslant 3\,\mathbb{E}\left[\left|\frac{1}{N}\sum_{k=0}^{N-1}\big(\tilde{f}_h(Y_k) - \tilde{\pi}_h(\tilde{f}_h)\big)\right|^2\right] + CL^2\Big(\frac{1}{Nh} + h^4\Big). \tag{A.23}$$

Telescoping $\tilde{\phi}_h(Y_{k+1}) - \tilde{\phi}_h(Y_k)$ along the chain gives the time-averaged error formula

$$\frac{1}{N}\sum_{k=0}^{N-1}\big(\tilde{f}_h(Y_k) - \tilde{\pi}_h(\tilde{f}_h)\big) = \frac{\tilde{\phi}_h(Y_0) - \tilde{\phi}_h(Y_N)}{Nh} + \frac{1}{N}\sum_{k=0}^{N-1}\Upsilon_k + \frac{1}{N}\sum_{k=0}^{N-1}\Psi_k, \tag{A.24}$$

with the discretization term and the martingale difference

$$\Upsilon_k := \frac{\tilde{\phi}_h(Y_{k+1}) - \tilde{\phi}_h(\tilde{Y}_k(h))}{h}, \qquad \Psi_k := \frac{\tilde{\phi}_h(\tilde{Y}_k(h)) - \tilde{\phi}_h(Y_k)}{h} + \tilde{f}_h(Y_k) - \tilde{\pi}_h(\tilde{f}_h), \tag{A.25}$$

where Υ_k and Ψ_k are $\mathcal{F}_{k+1}$-measurable and $\mathbb{E}_k\Psi_k = 0$ by (A.20) and (A.21). The conditional mean $\mathbb{E}_k\Upsilon_k$ is the local weak error of one step of NC-SGLM against the modified dynamics (A.19), divided by h.

A.3 Moments and invariant distribution

The terms of (A.24) are bounded through moments of the chain that are uniform in time. Lemma 1 bounds these moments by a Lyapunov estimate on the step $Y_{k+1}-Y_k = D_k+\zeta_k$, and gives the invariant distribution μ_h in Theorem 3 by Harris's ergodic theorem [12, Theorem 1.2], whose minorization is checked on the map (A.27).

Lemma 1 *Let Assumptions 1 and 3 hold, $h \in (0, h_0]$ and $P \geqslant 2$. Then the chain $(Y_k)_{k\geqslant 0}$ of (NC-SGLM) satisfies, with $\mathbb{E}_k$ of (A.1), for $n = 1, \dots, 6$ and every $k \geqslant 0$,*

$$\mathbb{E}_k|Y_{k+1}|^{2n} \leqslant \Big(1 - \frac{nm}{16}\,h\Big)|Y_k|^{2n} + Ch, \tag{A.26}$$

and has a unique invariant distribution μ_h, with $\mu_h(|y|^{12}) \leqslant C$.

Proof of Lemma 1. First, we verify the Lyapunov estimate (A.26). The proof of [30, Lemma 3] for SGLM uses only the conditional mean and the conditional moments of the step, here $Y_{k+1}-Y_k = D_k+\zeta_k$ by (A.11). Since the noise factor $\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)$ is independent of χ_k given $\mathcal{F}_k$, $\mathbb{E}_k[\zeta_k] = 0$. Hence

$$\mathbb{E}_k[Y_{k+1} - Y_k] = \mathbb{E}_k[D_k] = -h\,\mathbb{E}_k[\nabla U(Z_k)].$$

The bound $x \cdot \nabla U(x) \geqslant \frac{m}{4}|x|^2 - C$ of [30, Lemma 17] at $x = Z_k$, the bound $|\nabla U(Z_k)| \leqslant M(1 + |Z_k|)$, $\mathbb{E}_k|Z_k|^2 = |Y_k|^2 + \frac{dh}{2}$ and Young's inequality give

$$Y_k \cdot \mathbb{E}_k[\nabla U(Z_k)] \geqslant \frac{m}{8}|Y_k|^2 - C.$$

By (A.13), the conditional moments of the step satisfy

$$\mathbb{E}_k|Y_{k+1} - Y_k|^q \leqslant Ch^{\frac{q}{2}} + Ch^q\big(1 + |Y_k|\big)^q, \qquad 1 \leqslant q \leqslant 12.$$

With these bounds, the proof of [30, Lemma 3] gives (A.26).

Next, we verify the minorization required by Harris's ergodic theorem. The minorization bounds the transition kernel from every point of a level set $\{|y|^2 \leqslant R_W\}$ from below by a constant times a fixed probability measure, where $R_W := 64C/m > 2Ch/(\frac{m}{16}h)$, with the constant C of (A.26) for $n = 1$, depends only on m, M, R and d. Fix y with $|y|^2 \leqslant R_W$. From $Y_0 = y$, since $\sqrt{2h}\,\chi_0 = 2v - y$ at $v = \frac{y}{2} + \sqrt{h/2}\,\chi_0$, the step (NC-SGLM) reads

$$Y_1 = \mathcal{T}_{\boldsymbol{\xi}_0}\Big(\frac{y}{2} + \sqrt{\frac{h}{2}}\,\chi_0\Big), \quad \mathcal{T}_{\boldsymbol{\xi}}(v) := 2v + \Big(\mathcal{S}_h\big(\hat{\mathcal{C}}_P(y;\boldsymbol{\xi})\big) - I_d\Big)(2v - y) - h\,\hat{g}_P\Big(\frac{y}{2} + v;\boldsymbol{\xi}\Big), \quad v \in \mathbb{R}^d, \tag{A.27}$$

where $\mathcal{S}_h\big(\hat{\mathcal{C}}_P(y;\boldsymbol{\xi})\big)$ does not depend on v. The minorization follows from the properties of $\mathcal{T}_{\boldsymbol{\xi}_0}$, required for $\boldsymbol{\xi}_0$ in an event of probability at least $\frac{1}{2}$, that $\mathcal{T}_{\boldsymbol{\xi}_0}$ is Lipschitz with constant 3 on the closed unit ball $\mathbb{B}$ and that $\mathcal{T}_{\boldsymbol{\xi}_0}(\mathbb{B})$ contains $\mathbb{B}$. Let h_0 also satisfy

$$72Mh_0\big(1 + R_W^{\frac{1}{2}}\big) \leqslant 1, \qquad 16h_0M^2\big(1 + R_W^{\frac{1}{2}}\big)^2\big(2 + R_W^{\frac{1}{2}}\big) \leqslant 1. \tag{A.28}$$

With $\bar{\Lambda}$ of (A.5), we define the event

$$\Omega_0 := \Big\{\bar{\Lambda}\Big(\frac{y}{2};\boldsymbol{\xi}_0\Big) \leqslant 4M, \quad \Big|\hat{g}_P\Big(\frac{y}{2};\boldsymbol{\xi}_0\Big) - \nabla U\Big(\frac{y}{2}\Big)\Big| \leqslant 4M\big(1 + R_W^{\frac{1}{2}}\big), \quad \|\hat{\mathcal{C}}_P(y;\boldsymbol{\xi}_0)\| \leqslant \frac{4M^2(1 + R_W^{\frac{1}{2}})^2}{P-1}\Big\}.$$

Since $|y| \leqslant R_W^{1/2}$, the bounds (A.6), (A.4) at $x = \frac{y}{2}$ and (A.9) with $r = 1$ give

$$\begin{aligned}\mathbb{P}(\Omega_0^c) &\leqslant \frac{\mathbb{E}_{\boldsymbol{\xi}_0}\big[\bar{\Lambda}(\frac{y}{2};\boldsymbol{\xi}_0)^2\big]}{16M^2} + \frac{\mathbb{E}_{\boldsymbol{\xi}_0}\big[|\hat{g}_P(\frac{y}{2};\boldsymbol{\xi}_0) - \nabla U(\frac{y}{2})|^2\big]}{16M^2(1 + R_W^{\frac{1}{2}})^2} + \frac{(P-1)\,\mathbb{E}_{\boldsymbol{\xi}_0}\big[\|\hat{\mathcal{C}}_P(y;\boldsymbol{\xi}_0)\|\big]}{4M^2(1 + R_W^{\frac{1}{2}})^2} \\ &\leqslant \frac{1}{16} + \frac{1}{16P} + \frac{1}{4} < \frac{1}{2}.\end{aligned}$$

On Ω_0, for $v, v' \in \mathbb{B}$, the definition of $\bar{\Lambda}$, $|\nabla U(\frac{y}{2})| \leqslant M(1 + R_W^{1/2})$, (A.10) and (A.28) give

$$h\Big|\hat{g}_P\Big(\frac{y}{2} + v; \boldsymbol{\xi}_0\Big) - \hat{g}_P\Big(\frac{y}{2} + v'; \boldsymbol{\xi}_0\Big)\Big| \leqslant 4Mh\,|v - v'| \leqslant \frac{1}{8}\,|v - v'|,$$
$$h\Big|\hat{g}_P\Big(\frac{y}{2} + v; \boldsymbol{\xi}_0\Big)\Big| \leqslant 9Mh\big(1 + R_W^{\frac{1}{2}}\big) \leqslant \frac{1}{8},$$
$$\Big\|\mathcal{S}_h\big(\hat{\mathcal{C}}_P(y; \boldsymbol{\xi}_0)\big) - I_d\Big\|\,|2v - y| \leqslant 2hM^2\big(1 + R_W^{\frac{1}{2}}\big)^2\big(2 + R_W^{\frac{1}{2}}\big) \leqslant \frac{1}{8},$$
$$\Big\|\mathcal{S}_h\big(\hat{\mathcal{C}}_P(y; \boldsymbol{\xi}_0)\big) - I_d\Big\| \leqslant 2hM^2\big(1 + R_W^{\frac{1}{2}}\big)^2 \leqslant \frac{1}{16}.$$

Hence $\mathcal{T}_{\boldsymbol{\xi}_0}$ is Lipschitz with constant $2 + \frac{1}{8} + \frac{1}{8} \leqslant 3$ on $\mathbb{B}$, and for $w \in \mathbb{B}$ the map

$$v \mapsto \frac{1}{2}\Big(w - \Big(\mathcal{S}_h\big(\hat{\mathcal{C}}_P(y; \boldsymbol{\xi}_0)\big) - I_d\Big)(2v - y) + h\,\hat{g}_P\Big(\frac{y}{2} + v; \boldsymbol{\xi}_0\Big)\Big)$$

sends $\mathbb{B}$ into the ball of radius $\frac{1}{2}(1 + \frac{1}{8} + \frac{1}{8}) = \frac{5}{8}$ and is Lipschitz with constant $\frac{1}{2}(\frac{1}{8} + \frac{1}{8}) = \frac{1}{8}$ on $\mathbb{B}$; its fixed point $v \in \mathbb{B}$ solves $\mathcal{T}_{\boldsymbol{\xi}_0}(v) = w$, so $\mathbb{B} \subset \mathcal{T}_{\boldsymbol{\xi}_0}(\mathbb{B})$. For a Borel set $\Gamma \subset \mathbb{R}^d$, the inclusion $\Gamma \cap \mathbb{B} \subset \mathcal{T}_{\boldsymbol{\xi}_0}(\mathcal{T}_{\boldsymbol{\xi}_0}^{-1}(\Gamma) \cap \mathbb{B})$ and the Lipschitz bound give, on Ω_0, with Leb the Lebesgue measure,

$$\mathrm{Leb}(\Gamma \cap \mathbb{B}) \leqslant 3^d\,\mathrm{Leb}\big(\mathcal{T}_{\boldsymbol{\xi}_0}^{-1}(\Gamma) \cap \mathbb{B}\big).$$

Since $\frac{y}{2} + \sqrt{h/2}\,\chi_0 \sim \mathcal{N}(\frac{y}{2}, \frac{h}{2} I_d)$ is independent of $\boldsymbol{\xi}_0$ and its density on $\mathbb{B}$ is at least a constant $\beta_h > 0$ depending only on h, d and R_W, (A.27) gives, for $Y_0 = y$,

$$\mathbb{P}\big(Y_1 \in \Gamma\big) \geqslant \mathbb{P}(\Omega_0)\,\frac{\beta_h}{3^d}\,\mathrm{Leb}(\Gamma \cap \mathbb{B}) \geqslant \frac{\beta_h}{2 \cdot 3^d}\,\mathrm{Leb}(\Gamma \cap \mathbb{B}),$$

the minorization with the uniform distribution on $\mathbb{B}$.

Finally, Harris's ergodic theorem gives the unique invariant distribution μ_h, and the law of Y_k from $Y_0 = 0$ converges to μ_h in total variation. Since $\mathbb{E}|Y_k|^{12} \leqslant C$ by (A.26) for $n = 6$, Fatou's lemma gives $\mu_h(|y|^{12}) \leqslant \liminf_{k\to\infty} \mathbb{E}|Y_k|^{12} \leqslant C$. □

A.4 Local weak error

The batch noise $\hat{g}_P(Z_k; \boldsymbol{\xi}_k) - \nabla U(Z_k)$ in D_k of (A.11) contributes to the local weak error a term of order h^2 linear in $\mathcal{C}/P$, the noise correction cancels this term, and Lemma 2 bounds the remaining local weak error by Ch^3. For a tensor T of order j and vectors $v_1, \dots, v_i$ with $i \leqslant j$, $T[v_1, \dots, v_i]$ is the tensor of order $j - i$ obtained by contracting T with $v_1, \dots, v_i$.

Lemma 2 *Let Assumptions 1 and 3 hold, $h \in (0, h_0]$ and $P \geqslant 2$, and let $G \in C^4(\mathbb{R}^d)$ have bounded derivatives of orders 1 to 4. Then the chain $(Y_k)_{k\geqslant 0}$ of (NC-SGLM) and $\tilde{Y}_k$ of (A.19) satisfy, with $\mathbb{E}_k$ of (A.1), for every $k \geqslant 0$,*

$$\Big|\mathbb{E}_k\big[G(Y_{k+1}) - G(\tilde{Y}_k(h))\big]\Big| \leqslant Ch^3 \sum_{j=1}^{4} \|\nabla^j G\|_\infty \big(1 + |Y_k|\big)^6. \tag{A.29}$$

Proof of Lemma 2. By (A.11), the step reads $Y_{k+1} = Y_k + D_k + \zeta_k$, so that

$$\mathbb{E}_k\big[G(Y_{k+1}) - G(\tilde{Y}_k(h))\big] = \mathbb{E}_k\big[G(Y_k + D_k) - G(\tilde{Y}_k(h))\big] + \mathbb{E}_k\big[G(Y_k + D_k + \zeta_k) - G(Y_k + D_k)\big], \tag{A.30}$$

where the two terms are the effects of the SGLM displacement D_k and of the noise correction ζ_k, respectively.

We show that the step of (SGLM) satisfies

$$\Big|\mathbb{E}_k\big[G(Y_k + D_k) - G(\tilde{Y}_k(h))\big] - \frac{h^2}{2P}\,\nabla^2 G(Y_k) : \mathbb{E}_k\big[\mathcal{C}(Z_k)\big]\Big| \leqslant Ch^3 \sum_{j=1}^{4} \|\nabla^j G\|_\infty \big(1 + |Y_k|\big)^6, \tag{A.31}$$

that the contribution of the noise correction satisfies

$$\Big|\mathbb{E}_k\big[G(Y_k+D_k+\zeta_k)-G(Y_k+D_k)\big]+\frac{h^2}{2P}\,\nabla^2 G(Y_k):\mathcal{C}(Y_k)\Big|\leqslant Ch^3\sum_{j=2}^{4}\|\nabla^j G\|_\infty\big(1+|Y_k|\big)^5, \tag{A.32}$$

and that the sum of the terms of order h^2 in (A.31) and (A.32) satisfies

$$\frac{h^2}{2P}\Big|\nabla^2 G(Y_k):\big(\mathbb{E}_k[\mathcal{C}(Z_k)]-\mathcal{C}(Y_k)\big)\Big|\leqslant Ch^3\|\nabla^2 G\|_\infty\big(1+|Y_k|\big)^2. \tag{A.33}$$

Inserting (A.31)–(A.33) into (A.30) gives (A.29).

For (A.31), the state Y_k+D_k differs from the state $Y_k+\sqrt{2h}\,\chi_k-h\,\nabla U(Z_k)$ after a step of LM from Y_k by $-h$ times the batch noise $\hat g_P(Z_k;\boldsymbol{\xi}_k)-\nabla U(Z_k)$. Given $\mathcal{F}_k$ and $(B_{kh+s}-B_{kh})_{0\leqslant s\leqslant h}$, which determine $\tilde Y_k(h)$ and the state after the step of LM and are independent of $\boldsymbol{\xi}_k$, the batch noise has mean 0 and covariance $\mathcal{C}(Z_k)/P$ by (2.4), and its moment of order 3 is at most $C(1+|Z_k|)^3$ by (A.2). With these properties of the batch noise, the proof of [30, Lemma 7] gives (A.31), with (A.8) in place of the bound on the covariance uniform in x assumed there.

For (A.32), Taylor's formula for G about Y_k+D_k and for ∇G about Y_k gives

$$\begin{aligned}G(Y_k+D_k+\zeta_k)-G(Y_k+D_k)&=\nabla G(Y_k+D_k)\cdot\zeta_k+\varrho_2^{\zeta_k},\\ \nabla G(Y_k+D_k)&=\nabla G(Y_k)+\nabla^2 G(Y_k)[D_k]+\frac12\,\nabla^3 G(Y_k)[D_k,D_k]+\varrho_3^{D_k},\end{aligned} \tag{A.34}$$

with $|\varrho_2^{\zeta_k}|\leqslant\frac12\|\nabla^2G\|_\infty|\zeta_k|^2$ and $|\varrho_3^{D_k}|\leqslant\frac16\|\nabla^4G\|_\infty|D_k|^3$. Inserting $\nabla G(Y_k+D_k)$ in (A.34) into $\nabla G(Y_k+D_k)\cdot\zeta_k$ gives the high-order expansion

$$G(Y_k+D_k+\zeta_k)-G(Y_k+D_k)=\nabla G(Y_k)\cdot\zeta_k+\nabla^2G(Y_k)[D_k,\zeta_k]+\frac12\,\nabla^3G(Y_k)[D_k,D_k,\zeta_k]+\varrho_3^{D_k}\cdot\zeta_k+\varrho_2^{\zeta_k}. \tag{A.35}$$

The moments (A.13) of ζ_k, with the bound $Ch^3(1+|Y_k|)^4$ for $q=2$, and of D_k, with $q=6$, and the Cauchy–Schwarz inequality give

$$\mathbb{E}_k|\varrho_2^{\zeta_k}|+\mathbb{E}_k\big|\varrho_3^{D_k}\cdot\zeta_k\big|\leqslant Ch^3\big(\|\nabla^2G\|_\infty+\|\nabla^4G\|_\infty\big)\big(1+|Y_k|\big)^5. \tag{A.36}$$

The noise correction ζ_k is linear in χ_k, and its coefficient $\sqrt{2h}\,\big(\mathcal{S}_h(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k))-I_d\big)$ and $\hat g_P(Y_k;\boldsymbol{\xi}_k)$ are independent of χ_k given $\mathcal{F}_k$. Hence

$$\mathbb{E}_k\big[\nabla G(Y_k)\cdot\zeta_k\big]=0,\qquad \mathbb{E}_k\big[\nabla^2G(Y_k)[\hat g_P(Y_k;\boldsymbol{\xi}_k),\zeta_k]\big]=0. \tag{A.37}$$

In the term $\nabla^2G(Y_k)[D_k,\zeta_k]$ of (A.35), (A.11) gives

$$D_k=\sqrt{2h}\,\chi_k-h\,\hat g_P(Y_k;\boldsymbol{\xi}_k)-h\,\big(\hat g_P(Z_k;\boldsymbol{\xi}_k)-\hat g_P(Y_k;\boldsymbol{\xi}_k)\big). \tag{A.38}$$

The part $\sqrt{2h}\,\chi_k$ of (A.38) gives, by $\mathbb{E}[\chi_k^{\otimes2}]=I_d$,

$$\sqrt{2h}\,\mathbb{E}_k\big[\nabla^2G(Y_k)[\chi_k,\zeta_k]\big]=2h\,\nabla^2G(Y_k):\mathbb{E}_k\Big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)-I_d\Big], \tag{A.39}$$

and (A.10), the mean in (2.5) and (A.9) with $r=2$ give

$$\Big|2h\,\nabla^2G(Y_k):\mathbb{E}_k\Big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\big)-I_d\Big]+\frac{h^2}{2P}\,\nabla^2G(Y_k):\mathcal{C}(Y_k)\Big|\leqslant Ch^3\|\nabla^2G\|_\infty\big(1+|Y_k|\big)^4, \tag{A.40}$$

which gives the term of order h^2 in (A.32). The part $-h\,\hat g_P(Y_k;\boldsymbol{\xi}_k)$ gives a term of mean 0 by (A.37). For the part $-h\,(\hat g_P(Z_k;\boldsymbol{\xi}_k)-\hat g_P(Y_k;\boldsymbol{\xi}_k))$, (A.7) with $Z_k-Y_k=\sqrt{h/2}\,\chi_k$, (A.12) and Hölder's inequality with (A.6), (A.2) and (A.9) give

$$\begin{aligned}&h\Big|\mathbb{E}_k\big[\nabla^2G(Y_k)\big[\hat g_P(Z_k;\boldsymbol{\xi}_k)-\hat g_P(Y_k;\boldsymbol{\xi}_k),\zeta_k\big]\big]\Big|\\ &\leqslant\frac{h^3}{2}\,\|\nabla^2G\|_\infty\,\mathbb{E}_k\Big[\big(\bar\Lambda(Y_k;\boldsymbol{\xi}_k)+M+|\hat g_P(Z_k;\boldsymbol{\xi}_k)-\nabla U(Z_k)|+|\hat g_P(Y_k;\boldsymbol{\xi}_k)-\nabla U(Y_k)|\big)\\ &\quad\times\|\hat{\mathcal{C}}_P(Y_k;\boldsymbol{\xi}_k)\|\,|\chi_k|^2\Big]\leqslant Ch^3\|\nabla^2G\|_\infty\big(1+|Y_k|\big)^3.\end{aligned} \tag{A.41}$$

In the term $\frac{1}{2}\nabla^3 G(Y_k)[D_k, D_k, \zeta_k]$ of (A.35), the part with $\sqrt{2h}\,\chi_k$ in both factors D_k is a cubic form in χ_k with coefficients independent of χ_k given $\mathcal{F}_k$, whose conditional expectation vanishes. The other parts contain the factor $h\,\hat{g}_P(Z_k;\boldsymbol{\xi}_k)$, whose moments given χ_k are bounded by (A.2) at $x = Z_k$, and Hölder's inequality with (A.13) gives

$$\frac{1}{2}\left|\mathbb{E}_k\big[\nabla^3 G(Y_k)[D_k, D_k, \zeta_k]\big]\right| \leqslant Ch^3 \|\nabla^3 G\|_\infty \big(1+|Y_k|\big)^4. \tag{A.42}$$

Inserting (A.36)–(A.42) into (A.35) gives (A.32).

For (A.33), $Z_k - Y_k = \sqrt{h/2}\,\chi_k$ with χ_k independent of $\mathcal{F}_k$, so $\mathbb{E}_k[\mathcal{C}(Z_k)] = (e^{\frac{h}{4}\Delta}\mathcal{C})(Y_k)$, with the heat semigroup $e^{t\Delta}$ acting entrywise. By (A.8), $e^{t\Delta}\mathcal{C} \to \mathcal{C}$ as $t \to 0$, and by (3.5), $\partial_t\, e^{t\Delta}\mathcal{C} = e^{t\Delta}\Delta\mathcal{C}$ for $t > 0$, so that, with $\chi \sim \mathcal{N}(0, I_d)$ and $h \leqslant 1$,

$$\big\|\mathbb{E}_k[\mathcal{C}(Z_k)] - \mathcal{C}(Y_k)\big\| \leqslant \int_0^{h/4} \big\|(e^{t\Delta}\Delta\mathcal{C})(Y_k)\big\|\,\mathrm{d}t \leqslant M\int_0^{h/4} \mathbb{E}\big(1+|Y_k|+\sqrt{2t}\,|\chi|\big)^2\,\mathrm{d}t \leqslant Ch\big(1+|Y_k|\big)^2, \tag{A.43}$$

which gives (A.33). □

A.5 Proof of Theorem 1

Proof of Theorem 1. Let $h \in (0, h_0]$, $P \geqslant 2$ and $N \geqslant 1$. Taking expectations in (A.26) with $n = 6$ and iterating over k gives

$$\mathbb{E}|Y_k|^{12} \leqslant C\big(1+|Y_0|\big)^{12}, \qquad k \geqslant 0, \tag{A.44}$$

and (A.13) and the bound $\mathbb{E}_k|\tilde{Y}_k(h) - Y_k|^2 \leqslant Ch(1+|Y_k|)^2$ in [30, Lemma 18] give

$$\mathbb{E}|Y_{k+1} - Y_k|^2 + \mathbb{E}|\tilde{Y}_k(h) - Y_k|^2 \leqslant Ch\big(1+|Y_0|\big)^2. \tag{A.45}$$

Let first $Nh \geqslant 1$. By (A.24) and the Cauchy–Schwarz inequality, the mean square of the time average error is at most

$$\frac{3}{N^2h^2}\,\mathbb{E}\big|\tilde{\phi}_h(Y_0) - \tilde{\phi}_h(Y_N)\big|^2 + \frac{3}{N^2}\,\mathbb{E}\left|\sum_{k=0}^{N-1}\Upsilon_k\right|^2 + \frac{3}{N^2}\,\mathbb{E}\left|\sum_{k=0}^{N-1}\Psi_k\right|^2, \tag{A.46}$$

the sum of the boundary term, the discretization term and the martingale term. By the Lipschitz bound on $\tilde{\phi}_h$ in (A.22) and (A.44), the boundary term satisfies

$$\frac{3}{N^2h^2}\,\mathbb{E}\big|\tilde{\phi}_h(Y_0) - \tilde{\phi}_h(Y_N)\big|^2 \leqslant \frac{CL^2}{(Nh)^2}\big(1+|Y_0|\big)^2 \leqslant \frac{CL^2}{Nh}\big(1+|Y_0|\big)^2.$$

In the discretization term, Υ_k is split into its conditional mean $\mathbb{E}_k\Upsilon_k$ and the fluctuation $\Upsilon_k - \mathbb{E}_k\Upsilon_k$. By (A.25), Lemma 2 with $G = \tilde{\phi}_h$ and (A.22) give

$$|\mathbb{E}_k\Upsilon_k| \leqslant CLh^2\big(1+|Y_k|\big)^6,$$

and Jensen's inequality and (A.44) give

$$\frac{1}{N^2}\,\mathbb{E}\left|\sum_{k=0}^{N-1}\mathbb{E}_k\Upsilon_k\right|^2 \leqslant CL^2\big(1+|Y_0|\big)^{12}h^4.$$

Since $\mathbb{E}_k\Psi_k = 0$ and $\Psi_k - h^{-1}\big(\tilde{\phi}_h(\tilde{Y}_k(h)) - \tilde{\phi}_h(Y_k)\big)$ is $\mathcal{F}_k$-measurable, the Lipschitz bound on $\tilde{\phi}_h$ and (A.45) give, for the martingale differences Ψ_k and $\Upsilon_k - \mathbb{E}_k\Upsilon_k$,

$$\mathbb{E}|\Psi_k|^2 + \mathbb{E}\big|\Upsilon_k - \mathbb{E}_k\Upsilon_k\big|^2 \leqslant \frac{1}{h^2}\,\mathbb{E}\big|\tilde{\phi}_h(\tilde{Y}_k(h)) - \tilde{\phi}_h(Y_k)\big|^2 + \mathbb{E}|\Upsilon_k|^2 \leqslant \frac{CL^2}{h}\big(1+|Y_0|\big)^2,$$

and their orthogonality bounds the martingale term and the fluctuation by

$$\frac{3}{N^2}\,\mathbb{E}\Big|\sum_{k=0}^{N-1}\Psi_k\Big|^2+\frac{1}{N^2}\,\mathbb{E}\Big|\sum_{k=0}^{N-1}\big(\Upsilon_k-\mathbb{E}_k\Upsilon_k\big)\Big|^2\leqslant\frac{CL^2}{Nh}\big(1+|Y_0|\big)^2.$$

Adding these bounds proves (3.6) for $Nh\geqslant 1$. For $Nh<1$, $|f(x)-\pi(f)|\leqslant CL(1+|x|)$ by Assumption 2 and the moments of π [30, Lemma 17], and $\mathbb{E}|Z_k|^2\leqslant C(1+|Y_0|)^2$, so that

$$\mathrm{MSE}(N,h)\leqslant CL^2\big(1+|Y_0|\big)^2\leqslant\frac{CL^2}{Nh}\big(1+|Y_0|\big)^2.$$

□

B Proof of Theorem 2: uniform-in-time weak error

Recall that $(Y_0(t))_{t\geqslant 0}$ is the solution of the overdamped Langevin dynamics (2.1) with $Y_0(0)=Y_0$. For $t\geqslant 0$ and a measurable function F of polynomial growth, the semigroup $e^{t\mathcal{L}}$ satisfies

$$(e^{t\mathcal{L}}F)(y)=\mathbb{E}\big[F(Y_0(t))\big],\qquad Y_0=y. \tag{B.1}$$

The proof of Theorem 2 splits the weak error $\mathbb{E}[f(Z_n)]-\mathbb{E}[f(Y_0(nh))]$ into two differences.

- The first difference is $\mathbb{E}[\tilde f_h(Y_n)]-(e^{nh\tilde{\mathcal{L}}_h}\tilde f_h)(Y_0)$ between NC-SGLM and the modified dynamics (A.19). It sums the local weak errors of Lemma 2 and is of order h^2 uniformly in time.
- The second difference is $(e^{nh\tilde{\mathcal{L}}_h}\tilde f_h)(Y_0)-(e^{nh\mathcal{L}}f)(Y_0)$ between the modified dynamics and the overdamped Langevin dynamics. By Lemma 3, it is the modified bias of order h^2 plus a term of order h that decays exponentially in nh.

Appendix B.1 proves the ergodicity estimate of Lemma 3, Appendix B.2 proves Theorem 2, and Appendix B.3 shows by an example that the term of first order in (3.7) cannot be removed.

B.1 Ergodicity estimate for functions of quadratic growth

For a function F of quadratic growth, Lemma 3 bounds $(e^{t\mathcal{L}}F)(y)-\pi(F)$ that decays exponentially in t. Appendix B.2 applies the lemma to $\tilde f_h-f$ and to F_s of (B.21), whose quadratic growth carries the factor h.

Lemma 3 *Let Assumption 1 hold. There are C and $\lambda_1>0$, depending only on m, M, R and d, such that for every $K\geqslant 0$ and every measurable F with $|F(x)|\leqslant K(1+|x|^2)$ for $x\in\mathbb{R}^d$,*

$$\big|(e^{t\mathcal{L}}F)(y)-\pi(F)\big|\leqslant CKe^{-\lambda_1 t}\big(1+|y|^2\big),\qquad t\geqslant 0,\quad y\in\mathbb{R}^d. \tag{B.2}$$

Proof of Lemma 3. Let $(Y_0(t))_{t\geqslant 0}$ be the solution of the overdamped Langevin dynamics (2.1) with $Y_0(0)=y$. Itô's formula gives the drift condition [12, Assumption 1], $\mathbb{E}|Y_0(t)|^2\leqslant e^{-\frac{m}{2}t}\,|y|^2+C_0$ for $t\geqslant 0$, with $C_0>0$ depending only on m, M, R and d.

We verify the minorization [12, Assumption 2] on the level set $\{|y|\leqslant R_0\}$, where R_0 is defined by

$$R_0^2:=\frac{4C_0}{1-e^{-m/2}}. \tag{B.3}$$

With the Brownian motion $(B_t)_{t\geqslant 0}$ of (2.1), fix $R_1>0$, depending only on d, such that

$$\mathbb{P}\Big(2\sqrt{2}\sup_{0\leqslant t\leqslant 1}|B_t|\geqslant R_1\Big)\leqslant\frac{1}{2}. \tag{B.4}$$

We define the events Ω and, for a Borel set $\Gamma\subset\mathbb{R}^d$, $\Omega_\Gamma\subset\Omega$ by

$$\Omega:=\Big\{\sup_{0\leqslant s\leqslant 1}|y+\sqrt{2}\,B_s|\leqslant R_0+R_1\Big\},\qquad \Omega_\Gamma:=\big\{y+\sqrt{2}\,B_1\in\Gamma\big\}\cap\Omega.$$

Since ∇U is bounded on the ball $\{|x| \leqslant R_0 + R_1\}$, Girsanov's theorem up to the exit time from this ball and Itô's formula for $U(y + \sqrt{2}\, B_s)$ give

$$\begin{aligned} &\mathbb{P}\big(Y_0(1) \in \Gamma\big) \geqslant \mathbb{P}\Big(Y_0(1) \in \Gamma,\ \sup_{0 \leqslant s \leqslant 1} |Y_0(s)| \leqslant R_0 + R_1\Big) \\ &= \mathbb{E}\Big[\mathbf{1}_{\Omega_\Gamma} \exp\Big(\frac{U(y) - U(y + \sqrt{2}\, B_1)}{2} - \int_0^1 \Big(\frac{1}{4}\big|\nabla U(y + \sqrt{2}\, B_s)\big|^2 - \frac{1}{2}\, \Delta U(y + \sqrt{2}\, B_s)\Big) \mathrm{d}s\Big)\Big]. \end{aligned} \tag{B.5}$$

Since $y + \sqrt{2}\, B_1$ has the law $\mathcal{N}(y, 2I_d)$, the lower bound (B.5) gives

$$\begin{aligned} &\mathbb{P}\big(Y_0(1) \in \Gamma\big) \geqslant (4\pi)^{-\frac{d}{2}} \int_\Gamma \exp\Big(\frac{U(y) - U(z)}{2} - \frac{|z - y|^2}{4}\Big) \\ &\times \mathbb{E}\Big[\mathbf{1}_\Omega \exp\Big(-\int_0^1 \Big(\frac{1}{4}\big|\nabla U(y + \sqrt{2}\, B_s)\big|^2 - \frac{1}{2}\, \Delta U(y + \sqrt{2}\, B_s)\Big) \mathrm{d}s\Big) \,\Big|\, y + \sqrt{2}\, B_1 = z\Big] \mathrm{d}z. \end{aligned} \tag{B.6}$$

Since the integrand in (B.6) is nonnegative, we restrict the integral to $z \in \Gamma$ with $|z| \leqslant R_0$, the radius of the level set. For y in the level set and such z, given $y + \sqrt{2}\, B_1 = z$,

$$\sup_{0 \leqslant s \leqslant 1} \big|y + \sqrt{2}\, B_s\big| = \sup_{0 \leqslant s \leqslant 1} \big|y + s(z - y) + \sqrt{2}\, (B_s - sB_1)\big| \leqslant R_0 + \sqrt{2} \sup_{0 \leqslant s \leqslant 1} |B_s - sB_1|,$$

where $(B_s - sB_1)_{0 \leqslant s \leqslant 1}$ is independent of B_1 and $\sqrt{2} \sup_{0 \leqslant s \leqslant 1} |B_s - sB_1| \leqslant 2\sqrt{2} \sup_{0 \leqslant t \leqslant 1} |B_t|$. Hence (B.4), (3.2) and $|z - y| \leqslant 2R_0$ give

$$\begin{aligned} &\mathbb{P}\big(\Omega \,\big|\, y + \sqrt{2}\, B_1 = z\big) \geqslant \frac{1}{2}, \qquad \frac{U(y) - U(z)}{2} - \frac{|z - y|^2}{4} \geqslant -C(1 + R_0)^2, \\ &\frac{1}{4}|\nabla U(x)|^2 - \frac{1}{2}\, \Delta U(x) \leqslant C(1 + R_0 + R_1)^2, \qquad |x| \leqslant R_0 + R_1. \end{aligned} \tag{B.7}$$

Inserting (B.7) into (B.6) gives a constant $\alpha_1 > 0$, depending only on m, M, R and d, such that for $|y| \leqslant R_0$ and every Borel set $\Gamma \subset \mathbb{R}^d$,

$$\mathbb{P}\big(Y_0(1) \in \Gamma\big) \geqslant \alpha_1 \operatorname{Leb}\big(\Gamma \cap \{|z| \leqslant R_0\}\big). \tag{B.8}$$

Hence, for $|y| \leqslant R_0$, the law of $Y_0(1)$ is at least α times the uniform distribution on $\{|z| \leqslant R_0\}$, with

$$\alpha := \min\Big\{\frac{1}{2},\ \alpha_1 \operatorname{Leb}\big(\{|z| \leqslant R_0\}\big)\Big\}. \tag{B.9}$$

Finally, Harris's ergodic theorem in the form of [12, Theorem 1.3] applies to the Markov chain $(Y_0(n))_{n \geqslant 0}$ with the Lyapunov function $|x|^2$, the drift condition at $t = 1$ and the minorization (B.8) with α of (B.9), since $R_0^2 > 2C_0/(1 - e^{-m/2})$ by (B.3). The theorem gives $\bar\beta > 0$ and $\bar\alpha \in (0, 1)$, depending only on m, M, R and d through $e^{-m/2}$, C_0, R_0 and α, such that each step of this chain contracts the total variation distance weighted by $1 + \bar\beta|x|^2$ by the factor $\bar\alpha$. Since π is invariant under (2.1), the contraction, applied n times to the point mass at y and to π, gives, for every measurable ψ with $|\psi(x)| \leqslant 1 + \bar\beta|x|^2$ for $x \in \mathbb{R}^d$,

$$\big|(e^{n\mathcal{L}}\psi)(y) - \pi(\psi)\big| \leqslant \bar\alpha^{\,n}\big(2 + \bar\beta|y|^2 + \bar\beta\, \pi(|x|^2)\big), \qquad n \geqslant 0. \tag{B.10}$$

By [30, Lemma 17],

$$\pi(|x|^2) \leqslant C. \tag{B.11}$$

Since $|F(x)| \leqslant K \max\{1, \bar\beta^{-1}\}(1 + \bar\beta|x|^2)$ for $x \in \mathbb{R}^d$, (B.10) gives

$$\big|(e^{n\mathcal{L}}F)(y) - \pi(F)\big| \leqslant CK\bar\alpha^{\,n}\big(1 + |y|^2\big), \qquad n \geqslant 0. \tag{B.12}$$

For $t \geqslant 0$, the drift condition gives

$$\big|(e^{(t - \lfloor t \rfloor)\mathcal{L}}F)(x)\big| \leqslant (1 + C_0)K\big(1 + |x|^2\big), \qquad x \in \mathbb{R}^d,$$

and the invariance of π gives $\pi(e^{(t - \lfloor t \rfloor)\mathcal{L}}F) = \pi(F)$. Hence (B.12) with $n = \lfloor t \rfloor$, applied with $e^{(t - \lfloor t \rfloor)\mathcal{L}}F$ and $(1 + C_0)K$ in place of F and K, gives (B.2) with $\lambda_1 = |\log \bar\alpha|$. □

B.2 Proof of Theorem 2

Proof of Theorem 2. Let $h \in (0, h_0]$, $P \geqslant 2$, $n \geqslant 0$ and $Y_0 \in \mathbb{R}^d$. The modified test function $\tilde{f}_h$ of (A.14) satisfies Assumption 2 with the same L, and, by (A.17), the modified potential $\tilde{U}_h$ of (A.16) satisfies Assumption 1 up to the order 5 with $2M$ and $\frac{m}{2}$ in place of M and m. Since the proof of [30, Theorem 1] uses Assumption 1 up to the order 5 only [30, Section 4.2], that theorem applies to U and f and, with constants independent of h, to $\tilde{U}_h$ and $\tilde{f}_h$. With $\tilde{\mathcal{L}}_h$ and $\tilde{\pi}_h$ of (A.18), the theorem gives $\lambda_0 > 0$, depending only on m, M, R and d, such that for $s \geqslant 0$ and $x \in \mathbb{R}^d$,

$$\max_{1 \leqslant j \leqslant 4} \left\| \nabla^j (e^{s\tilde{\mathcal{L}}_h} \tilde{f}_h)(x) \right\| \leqslant CLe^{-\lambda_0 s}. \tag{B.13}$$

The bound (B.13) with $j = 1$ and the moment bounds of $\tilde{\pi}_h$ and π in [30, Lemma 17] give

$$\left| (e^{s\tilde{\mathcal{L}}_h} \tilde{f}_h)(x) - \tilde{\pi}_h(\tilde{f}_h) \right| + \left| (e^{s\mathcal{L}} f)(x) - \pi(f) \right| \leqslant CL(1 + |x|)\, e^{-\lambda_0 s}. \tag{B.14}$$

By (A.15) and (B.1), the weak error splits as

$$\mathbb{E}[f(Z_n)] - \mathbb{E}[f(Y_0(nh))] = \underbrace{\mathbb{E}[\tilde{f}_h(Y_n)] - (e^{nh\tilde{\mathcal{L}}_h} \tilde{f}_h)(Y_0)}_{\mathcal{D}^{(1)}_{n,h}} + \underbrace{(e^{nh\tilde{\mathcal{L}}_h} \tilde{f}_h)(Y_0) - (e^{nh\mathcal{L}} f)(Y_0)}_{\mathcal{D}^{(2)}_{n,h}}. \tag{B.15}$$

For $\mathcal{D}^{(1)}_{n,h}$, telescoping over the steps and (A.20) give, with $\mathbb{E}_k$ of (A.1) and $\tilde{Y}_k$ of (A.19),

$$\begin{aligned} \mathcal{D}^{(1)}_{n,h} &= \sum_{k=0}^{n-1} \mathbb{E}\Big[(e^{(n-k-1)h\tilde{\mathcal{L}}_h} \tilde{f}_h)(Y_{k+1}) - (e^{(n-k)h\tilde{\mathcal{L}}_h} \tilde{f}_h)(Y_k) \Big] \\ &= \sum_{k=0}^{n-1} \mathbb{E}\Big[\mathbb{E}_k \big[(e^{(n-k-1)h\tilde{\mathcal{L}}_h} \tilde{f}_h)(Y_{k+1}) - (e^{(n-k-1)h\tilde{\mathcal{L}}_h} \tilde{f}_h)(\tilde{Y}_k(h)) \big] \Big]. \end{aligned} \tag{B.16}$$

Taking expectations in (A.26) for the moment of order 6 and iterating over k gives

$$\mathbb{E}\big(1 + |Y_k|\big)^6 \leqslant C\big(1 + e^{-\frac{3m}{16}kh} |Y_0|^6\big), \qquad k \geqslant 0. \tag{B.17}$$

With λ_1 of Lemma 3, we take in (3.7)

$$\lambda = \frac{1}{2} \min\Big\{ \lambda_0, \lambda_1, \frac{3m}{16} \Big\}. \tag{B.18}$$

Let h_0 also satisfy $\lambda_0 h_0 \leqslant 1$, so that

$$\sum_{k=0}^{n-1} e^{-\lambda_0 (n-k-1)h} \leqslant \frac{2}{\lambda_0 h}, \qquad \sum_{k=0}^{n-1} e^{-\lambda_0 (n-k-1)h - \frac{3m}{16}kh} \leqslant n\, e^{-2\lambda(n-1)h} \leqslant Cn\, e^{-2\lambda nh}. \tag{B.19}$$

By (B.13), (B.17), (B.19) and Lemma 2 with $G = e^{(n-k-1)h\tilde{\mathcal{L}}_h} \tilde{f}_h$ in (B.16), $\mathcal{D}^{(1)}_{n,h}$ satisfies

$$\left| \mathcal{D}^{(1)}_{n,h} \right| \leqslant CLh^3 \sum_{k=0}^{n-1} e^{-\lambda_0 (n-k-1)h} \Big(1 + e^{-\frac{3m}{16}kh} |Y_0|^6 \Big) \leqslant CLh^2 + CLh^2\, nh\, e^{-2\lambda nh}\, |Y_0|^6. \tag{B.20}$$

For $\mathcal{D}^{(2)}_{n,h}$, let F_s be the difference of the generators $\tilde{\mathcal{L}}_h$ and $\mathcal{L}$ applied to $e^{s\tilde{\mathcal{L}}_h} \tilde{f}_h$,

$$F_s := \big(\nabla U - \nabla \tilde{U}_h \big) \cdot \nabla e^{s\tilde{\mathcal{L}}_h} \tilde{f}_h = \big(\tilde{\mathcal{L}}_h - \mathcal{L} \big) e^{s\tilde{\mathcal{L}}_h} \tilde{f}_h, \qquad s \geqslant 0. \tag{B.21}$$

The bound $|\nabla \tilde{U}_h(x) - \nabla U(x)| \leqslant Ch(1 + |x|)$ of [30, Lemma 1], (B.13) and (B.11) give

$$|F_s(x)| \leqslant CLh\, e^{-\lambda_0 s}(1 + |x|), \qquad |\pi(F_s)| \leqslant CLh\, e^{-\lambda_0 s}, \qquad s \geqslant 0, \quad x \in \mathbb{R}^d. \tag{B.22}$$

For $t \geqslant 0$ and $s \in [0,t]$, the semigroup of the modified dynamics satisfies

$$\partial_s e^{(t-s)\tilde{\mathcal{L}}_h} \tilde{f}_h = -\tilde{\mathcal{L}}_h e^{(t-s)\tilde{\mathcal{L}}_h} \tilde{f}_h,$$

so that Itô's formula and (B.21) give

$$\mathrm{d}\big(e^{(t-s)\tilde{\mathcal{L}}_h} \tilde{f}_h\big)\big(Y_0(s)\big) = -F_{t-s}\big(Y_0(s)\big)\,\mathrm{d}s + \sqrt{2}\,\nabla\big(e^{(t-s)\tilde{\mathcal{L}}_h} \tilde{f}_h\big)\big(Y_0(s)\big) \cdot \mathrm{d}B_s.$$

The stochastic integral has mean 0, since $\nabla e^{(t-s)\tilde{\mathcal{L}}_h} \tilde{f}_h$ is bounded by (B.13). We integrate over $s \in [0,t]$, take expectations with (B.1), substitute $s \mapsto t-s$ and add $\big(e^{t\mathcal{L}}(\tilde{f}_h - f)\big)(Y_0)$ to both sides:

$$(e^{t\tilde{\mathcal{L}}_h} \tilde{f}_h)(Y_0) - (e^{t\mathcal{L}} f)(Y_0) = \big(e^{t\mathcal{L}}(\tilde{f}_h - f)\big)(Y_0) + \int_0^t (e^{(t-s)\mathcal{L}} F_s)(Y_0)\,\mathrm{d}s. \tag{B.23}$$

Taylor's formula for f to the order 1 in (A.14) and $\mathbb{E}\chi = 0$ give $|\tilde{f}_h(x) - f(x)| \leqslant \frac{dLh}{4}$ for $x \in \mathbb{R}^d$. Lemma 3, applied in (B.23) to $\tilde{f}_h - f$ with $K = \frac{dLh}{4}$ and to F_s with $K = CLh\,e^{-\lambda_0 s}$ by (B.22), gives, with λ of (B.18),

$$\Big|(e^{t\tilde{\mathcal{L}}_h} \tilde{f}_h)(Y_0) - (e^{t\mathcal{L}} f)(Y_0) - \pi\big(\tilde{f}_h - f\big) - \int_0^t \pi(F_s)\,\mathrm{d}s\Big| \leqslant CLh\,(1+t)\,e^{-2\lambda t}\big(1 + |Y_0|^2\big). \tag{B.24}$$

As $t \to \infty$, (B.14) turns (B.24) into

$$\pi\big(\tilde{f}_h - f\big) + \int_0^\infty \pi(F_s)\,\mathrm{d}s = \tilde{\pi}_h(\tilde{f}_h) - \pi(f). \tag{B.25}$$

At $t = nh$, (B.25) turns (B.24) into

$$\Big|\mathcal{D}_{n,h}^{(2)} - \big(\tilde{\pi}_h(\tilde{f}_h) - \pi(f)\big) + \int_{nh}^\infty \pi(F_s)\,\mathrm{d}s\Big| \leqslant CLh\,(1+nh)\,e^{-2\lambda nh}\big(1 + |Y_0|^2\big).$$

The bound on the modified bias in (A.22) and the bound on $\pi(F_s)$ in (B.22) then give

$$\big|\mathcal{D}_{n,h}^{(2)}\big| \leqslant CLh^2 + CLh\,(1+nh)\,e^{-2\lambda nh}\big(1 + |Y_0|^2\big). \tag{B.26}$$

Inserting (B.20) and (B.26) into (B.15) gives (3.7). □

B.3 Necessity of the term of first order

The following example shows that the term of first order in h in (3.7) is necessary. Let $d = 1$, $U(x) = x^2/2$, $g = \nabla U$ and $f(x) = x$, which satisfy the assumptions of Theorem 2 with $m = M = L = 1$ and $R = 0$. Then (NC-SGLM) is (LM) for every noise corrector satisfying the polynomial bound, as noted after (3.9). Hence LM gives $\mathbb{E}[Z_n] = (1-h)^n Y_0$, while $\mathbb{E}[Y_0(nh)] = e^{-nh} Y_0$. Since, uniformly in n,

$$(1-h)^n - e^{-nh} = -\frac{h}{2}\,nh\,e^{-nh} + \mathcal{O}(h^2),$$

the weak error at the time $nh = 1$, that is, for $h = 1/n$, is

$$\mathbb{E}[Z_n] - \mathbb{E}[Y_0(nh)] = -\frac{h}{2e}\,Y_0 + \mathcal{O}\big(h^2|Y_0|\big).$$

Hence, for $Y_0 \neq 0$, the weak error at the time $nh = 1$ is of first order in h. Since (B.20) bounds $|\mathcal{D}_{n,h}^{(1)}|$ by $Ch^2(1+|Y_0|)^6$ at $nh = 1$, this term of first order comes from $\mathcal{D}_{n,h}^{(2)}$.

C Proof of Theorem 3: numerical bias estimate

The proof of Theorem 3 compares NC-SGLM with the full gradient LM,

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

Since the full gradient satisfies Assumption 3 and NC-SGLM with $g = \nabla U$ is (LM), Lemmas 1 and 2 apply to (LM). The bias of NC-SGLM splits into the bias $\pi_h^{\mathrm{LM}}(f) - \pi(f)$ from LM and the difference $\pi_h(f) - \pi_h^{\mathrm{LM}}(f)$ from stochastic gradients. Appendices C.1 and C.2 give the transition density, the exponential ergodicity and the semigroup estimates of LM, Appendix C.3 compares a step of NC-SGLM with a step of LM, Appendix C.4 constructs the midpoint distribution ν_h, and Appendix C.5 combines these results into Theorem 3. Throughout this appendix, the noise corrector $\mathcal{S}_h$ satisfies the polynomial bound (2.6), and h_0 also satisfies $h_0 M \leqslant \frac{1}{4}$, $h_0 M^2 \leqslant \frac{m}{8}$ and the conditions stated in the proofs below.

The midpoint distribution ν_h plays the role of the modified target $\tilde{\pi}_h$ in [30, Section 4.2], but it is built on the midpoint potential V_h of (C.6) rather than on the modified potential, and it must meet two requirements. First, ν_h must satisfy

$$\nu_h(\tilde{f}_h) - \pi(f) = \mathcal{O}(Lh^3)$$

for the modified test function $\tilde{f}_h = e^{\frac{h}{4}\Delta} f$ of (A.14). Then, for the invariant distribution $\overline{\mu}_h$ of the chain $(Y_k)_{k\geqslant 0}$ of (LM), (3.1) and this requirement give the bias of LM:

$$\pi_h^{\mathrm{LM}}(f) - \pi(f) = \overline{\mu}_h(\tilde{f}_h) - \pi(f) = \overline{\mu}_h(\tilde{f}_h) - \nu_h(\tilde{f}_h) + \mathcal{O}(Lh^3).$$

Formally, this requirement amounts to the expansion

$$\nu_h = \pi - \frac{h}{4}\,\Delta\pi + \frac{h^2}{32}\,\Delta^2\pi + \mathcal{O}(h^3). \tag{C.1}$$

Second, the change of ν_h under one step of LM must have an explicit term of order h^3, which, divided by the step size h, gives the coefficient $\frac{1}{24}\,\pi(\nabla^3\phi_0 : \nabla^3 U)$ of h^2 in (3.8). The construction of ν_h in Appendix C.4 is therefore the core of the proof.

C.1 Markov transition density of LM

This subsection defines the one-step operators of NC-SGLM and LM, derives the transition density of LM by a change of variables in the Gaussian vector of one step, and bounds the midpoint potential that appears in this density.

With the full gradient in place of the batch gradient, the SGLM displacement D_0 of (A.11) becomes the LM step $\overline{D}_0$. For an initial state Y_0 independent of χ_0 and $\boldsymbol{\xi}_0$, with $Z_0 = Y_0 + \sqrt{h/2}\,\chi_0$ of (LM), and a measurable function G on $\mathbb{R}^d$ of polynomial growth, the LM step and the one-step operators are

$$\overline{D}_0 := \sqrt{2h}\,\chi_0 - h\,\nabla U(Z_0), \quad (\mathcal{P}_h G)(y) := \mathbb{E}\big[G(Y_1)\big], \quad (\overline{\mathcal{P}}_h G)(y) := \mathbb{E}\big[G(Y_0 + \overline{D}_0)\big], \quad Y_0 = y, \tag{C.2}$$

where Y_1 is the state after one step of (NC-SGLM) from Y_0. For a probability measure μ on $\mathbb{R}^d$, $\mathcal{P}_h^*\mu$ and $\overline{\mathcal{P}}_h^*\mu$ denote the laws of Y_1 and of $Y_0 + \overline{D}_0$ for $Y_0 \sim \mu$. For $n \geqslant 0$, $\overline{\mathcal{P}}_{nh}^*$ denotes the transition of LM on probability measures over the time nh: $\overline{\mathcal{P}}_{nh}^*\mu$ is the law after n steps of LM from $Y_0 \sim \mu$. The transition density $\overline{\mathcal{P}}_h^*(y,\cdot)$ of LM is the law of the state $y + \overline{D}_0$ after the LM step from the initial state $Y_0 = y$, and, for a probability density ν on $\mathbb{R}^d$,

$$(\overline{\mathcal{P}}_h^*\nu)(x) = \int_{\mathbb{R}^d} \nu(y)\,\overline{\mathcal{P}}_h^*(y,x)\,\mathrm{d}y, \qquad x \in \mathbb{R}^d. \tag{C.3}$$

By Lemma 1 with $g = \nabla U$, $\overline{\mu}_h$ is the unique invariant distribution of $\overline{\mathcal{P}}_h^*$ and satisfies $\overline{\mu}_h(|y|^{12}) \leqslant C$. By (3.1), the numerical invariant distribution $\overline{\pi}_h := \pi_h^{\mathrm{LM}}$ of LM in (3.10) is $\overline{\pi}_h = \overline{\mu}_h * \mathcal{N}(0, \frac{h}{2}\,I_d)$.

In this appendix, D_0, μ_h, π_h and $\mathcal{P}_h$ without a bar are defined for (NC-SGLM), and $\overline{D}_0$, $\overline{\mu}_h$, $\overline{\pi}_h$ and $\overline{\mathcal{P}}_h$ with a bar are defined for (LM). In particular, we use $\overline{\pi}_h$ instead of π_h^{LM} in the proofs for consistency of notation. Since $|\nabla U(Z_0)| \leqslant M(1+|Y_0|+|\chi_0|)$ for $h \leqslant 1$, the LM step satisfies

$$|\overline{D}_0| \leqslant \sqrt{2h}\,|\chi_0| + hM\big(1+|Y_0|+|\chi_0|\big). \tag{C.4}$$

In terms of Z_0, the state after the LM step is

$$Y_0 + \overline{D}_0 = 2\Big(Z_0 - \frac{h}{2}\,\nabla U(Z_0)\Big) - Y_0. \tag{C.5}$$

Since $\frac{h}{2}M \leqslant \frac{1}{8}$, the map $w \mapsto x + \frac{h}{2}\,\nabla U(w)$ is a contraction of $\mathbb{R}^d$ for every $x \in \mathbb{R}^d$, with a unique fixed point. Next we formally define the midpoint map and the midpoint potential.

The *midpoint map* w_h and the *midpoint potential* V_h are

$$w_h(x) - \frac{h}{2}\,\nabla U\big(w_h(x)\big) = x, \qquad V_h(x) := U\big(w_h(x)\big) - \frac{h}{4}\,\big|\nabla U\big(w_h(x)\big)\big|^2, \qquad x \in \mathbb{R}^d. \tag{C.6}$$

Hence w_h is the inverse of the map $w \mapsto w - \frac{h}{2}\,\nabla U(w)$, a diffeomorphism of $\mathbb{R}^d$ whose Jacobian satisfies

$$\frac{9}{8}\,I_d \succeq I_d - \frac{h}{2}\,\nabla^2 U(w) \succeq \frac{7}{8}\,I_d, \qquad w \in \mathbb{R}^d. \tag{C.7}$$

Differentiating (C.6) gives

$$\begin{aligned} \nabla w_h(x) &= \Big(I_d - \frac{h}{2}\,\nabla^2 U\big(w_h(x)\big)\Big)^{-1}, \qquad \nabla V_h(x) = \nabla U\big(w_h(x)\big), \\ \nabla^2 V_h(x) &= \nabla^2 U\big(w_h(x)\big)\Big(I_d - \frac{h}{2}\,\nabla^2 U\big(w_h(x)\big)\Big)^{-1}, \qquad I_d + \frac{h}{2}\,\nabla^2 V_h(x) = \Big(I_d - \frac{h}{2}\,\nabla^2 U\big(w_h(x)\big)\Big)^{-1}. \end{aligned} \tag{C.8}$$

By (C.8) and (C.7) at $w = w_h(x)$,

$$\frac{8}{7}\,I_d \succeq \nabla w_h(x) = I_d + \frac{h}{2}\,\nabla^2 V_h(x) \succeq \frac{8}{9}\,I_d, \qquad x \in \mathbb{R}^d. \tag{C.9}$$

The midpoint map w_h and the midpoint potential V_h allow an explicit expression for the transition density $\overline{\mathcal{P}}_h^*(y,y')$ of LM. By (C.6), the map $w \mapsto 2(w - \frac{h}{2}\,\nabla U(w)) - y$ in (C.5) with the initial state $Y_0 = y$ sends w to y' if and only if $w = w_h(\frac{y+y'}{2})$, and its Jacobian determinant is $2^d \det(I_d - \frac{h}{2}\,\nabla^2 U(w))$. By (C.8),

$$w_h\Big(\frac{y+y'}{2}\Big) - y = \frac{1}{2}\,(y'-y) + \frac{h}{2}\,\nabla V_h\Big(\frac{y+y'}{2}\Big), \qquad y, y' \in \mathbb{R}^d. \tag{C.10}$$

For the initial state $Y_0 = y$ and the state $y' = y + \overline{D}_0$ after the LM step, the map in (C.5) sends $Z_0 = y + \sqrt{h/2}\,\chi_0$ to y', so that $w_h(\frac{y+y'}{2}) = Z_0$ and (C.10) gives

$$y' = y - h\,\nabla V_h\Big(\frac{y+y'}{2}\Big) + \sqrt{2h}\,\chi_0, \tag{C.11}$$

that is, the LM step is a midpoint step with the midpoint potential V_h. The change of variables from w to y' in the law $\mathcal{N}(y, \frac{h}{2}\,I_d)$ of Z_0, together with (C.8) and (C.10), gives the transition density of LM:

$$\overline{\mathcal{P}}_h^*(y,y') = \Lambda_h\Big(\frac{y+y'}{2}\Big)\,(4\pi h)^{-\frac{d}{2}}\exp\Big(-\frac{|y'-y|^2}{4h} - \frac{1}{2}\,(y'-y)\cdot\nabla V_h\Big(\frac{y+y'}{2}\Big)\Big), \quad y, y' \in \mathbb{R}^d, \tag{C.12}$$

where $\Lambda_h(x)$ is defined as

$$\Lambda_h(x) := \det\Big(I_d + \frac{h}{2}\,\nabla^2 V_h(x)\Big)\exp\Big(-\frac{h}{4}\,|\nabla V_h(x)|^2\Big), \qquad x \in \mathbb{R}^d. \tag{C.13}$$

Lemma 4 below gives the estimates of V_h used to bound the transition density (C.12).

Lemma 4 *Let Assumption 1 hold and $h \in (0, h_0]$. The midpoint potential V_h of (C.6) satisfies, for $x \in \mathbb{R}^d$,*

$$\begin{gathered}\max_{2\leqslant i\leqslant 8} \|\nabla^i V_h(x)\| \leqslant C, \qquad |\nabla V_h(x)| \leqslant C(1+|x|), \qquad V_h(x) - U(0) \geqslant \frac{m}{64}\,|x|^2 - C, \\ \max_{1\leqslant i\leqslant 7} \|\nabla^i V_h(x) - \nabla^i U(x)\| \leqslant Ch(1+|x|),\end{gathered} \tag{C.14}$$

the error of the midpoint rule for the line integral $V_h(y') - V_h(y)$ of ∇V_h satisfies, for $y, y' \in \mathbb{R}^d$,

$$\Big|V_h(y') - V_h(y) - (y'-y)\cdot \nabla V_h\Big(\frac{y+y'}{2}\Big)\Big| \leqslant C|y'-y|^3, \tag{C.15}$$

and V_h satisfies, with all functions evaluated at $x \in \mathbb{R}^d$,

$$\begin{aligned}&\Big|V_h - U - \frac{h}{4}\,|\nabla U|^2 - \frac{h^2}{8}\,\nabla^2 U[\nabla U, \nabla U]\Big| + h\,\Big|\Delta V_h - \Delta U - \frac{h}{4}\,\Delta|\nabla U|^2\Big| \\ &\quad + h^2\,\Big|\nabla^2 V_h : \nabla^2 V_h + \frac{1}{2}\,\Delta^2 V_h - \nabla^2 U : \nabla^2 U - \frac{1}{2}\,\Delta^2 U\Big| \leqslant Ch^3(1+|x|)^4.\end{aligned} \tag{C.16}$$

Proof of Lemma 4. To prove (C.14), we use $\nabla V_h = \nabla U \circ w_h$, $\|\nabla w_h(x)\| \leqslant \frac{8}{7}$ and $\|\nabla w_h(x) - I_d\| \leqslant Ch$, from (C.8) and (C.9). The formula for ∇w_h, differentiated up to the order 6 with the derivatives of U of orders 3 to 8, bounds $\nabla^i w_h$ for $2 \leqslant i \leqslant 7$, and (C.6) with $|\nabla U(w)| \leqslant M(1+|w|)$ and $hM \leqslant \frac{1}{4}$ bounds $|w_h(x)|$ and $|w_h(x) - x|$ in terms of $|x|$, and $|x|$ in terms of $|w_h(x)|$:

$$\begin{gathered}\max_{2\leqslant i\leqslant 7} \|\nabla^i w_h(x)\| \leqslant Ch, \qquad |w_h(x)| \leqslant 2|x| + 1, \\ |w_h(x) - x| \leqslant Ch(1+|x|), \qquad |x| \leqslant \frac{9}{8}\,|w_h(x)| + \frac{1}{8}.\end{gathered} \tag{C.17}$$

Hence $\|\nabla^i V_h(x)\| \leqslant C$ for $2 \leqslant i \leqslant 8$. By (C.17), $|\nabla V_h(x)| \leqslant C(1+|x|)$ and

$$\|\nabla^i U(w_h(x)) - \nabla^i U(x)\| \leqslant Ch(1+|x|), \qquad 1 \leqslant i \leqslant 7, \quad x \in \mathbb{R}^d, \tag{C.18}$$

which, with $\|\nabla w_h(x) - I_d\| \leqslant Ch$ and the bound on $\nabla^i w_h$, give the bound on $\nabla^i V_h - \nabla^i U$ in (C.14). By [30, Lemma 17] and $hM^2 \leqslant \frac{m}{8}$,

$$U(w) - U(0) \geqslant \frac{m}{8}\,|w|^2 - C, \qquad \frac{h}{4}\,|\nabla U(w)|^2 \leqslant \frac{h}{2}\,M^2(1+|w|^2) \leqslant \frac{m}{16}\,|w|^2 + C, \qquad w \in \mathbb{R}^d,$$

which, at $w = w_h(x)$ and with the bound on $|x|$, give the lower bound on V_h in (C.14).

To prove (C.15), we apply Taylor's formula for V_h about $\frac{y+y'}{2}$ to the order 2 at the points $\frac{y+y'}{2} \pm \frac{1}{2}(y'-y)$, which gives (C.15) with the constant $\frac{1}{24}\sup_{x\in\mathbb{R}^d}\|\nabla^3 V_h(x)\|$.

To prove (C.16), we apply Taylor's formula for U about $w_h(x)$ to the order 2 at $x = w_h(x) - \frac{h}{2}\,\nabla V_h(x)$, with $|\nabla V_h(x)| \leqslant C(1+|x|)$ by (C.14), and insert the result into (C.6):

$$\begin{aligned}U(x) &= U\big(w_h(x)\big) - \frac{h}{2}\,|\nabla V_h(x)|^2 + \frac{h^2}{8}\,\nabla^2 U\big(w_h(x)\big)\big[\nabla V_h(x), \nabla V_h(x)\big] + \mathcal{O}\big(h^3(1+|x|)^3\big), \\ V_h(x) &= U(x) + \frac{h}{4}\,|\nabla V_h(x)|^2 - \frac{h^2}{8}\,\nabla^2 U\big(w_h(x)\big)\big[\nabla V_h(x), \nabla V_h(x)\big] + \mathcal{O}\big(h^3(1+|x|)^3\big).\end{aligned} \tag{C.19}$$

Taylor's formula for ∇U about x to the order 1 at $w_h(x) = x + \frac{h}{2}\,\nabla V_h(x)$, with (C.8) and the bound on $\nabla V_h - \nabla U$ in (C.14), gives

$$\nabla V_h(x) = \nabla U(x) + \frac{h}{2}\,\nabla^2 U(x)\,\nabla U(x) + \mathcal{O}\big(h^2(1+|x|)^2\big),$$

and inserting this expansion and (C.18) with $i = 2$ into the expansion of $V_h(x)$ in (C.19) gives

$$V_h(x) = U(x) + \frac{h}{4}\,|\nabla U(x)|^2 + \frac{h^2}{8}\,\nabla^2 U(x)\big[\nabla U(x), \nabla U(x)\big] + \mathcal{O}\big(h^3(1+|x|)^4\big). \tag{C.20}$$

By (C.8) and Taylor's formula for ΔU about x to the order 1 at $w_h(x) = x + \frac{h}{2}\nabla V_h(x)$, with the bound on $\nabla V_h - \nabla U$,

$$\Delta V_h(x) = \Delta U\big(w_h(x)\big) + \frac{h}{2}\,\nabla^2 U\big(w_h(x)\big) : \nabla^2 U\big(w_h(x)\big) + \mathcal{O}(h^2),$$
$$\Delta U\big(w_h(x)\big) = \Delta U(x) + \frac{h}{2}\,\nabla\Delta U(x)\cdot\nabla U(x) + \mathcal{O}\big(h^2(1+|x|)^2\big),$$

and, by (C.18) with $i = 2$ and since $\Delta|\nabla U|^2 = 2\,\nabla^2 U : \nabla^2 U + 2\,\nabla U\cdot\nabla\Delta U$,

$$\Delta V_h(x) = \Delta U(x) + \frac{h}{4}\,\Delta|\nabla U|^2(x) + \mathcal{O}\big(h^2(1+|x|)^2\big). \tag{C.21}$$

Finally, the bound on $\nabla^2 V_h$ and the bound on $\nabla^i V_h - \nabla^i U$ for $i = 2$ and $i = 4$ in (C.14) give $|\nabla^2 V_h : \nabla^2 V_h + \frac{1}{2}\,\Delta^2 V_h - \nabla^2 U : \nabla^2 U - \frac{1}{2}\,\Delta^2 U| \leqslant Ch(1+|x|)$, which, with (C.20) and (C.21), gives (C.16). □

C.2 Ergodicity and semigroup estimates of LM

Lemma 5 gives the exponential ergodicity of $\overline{\mathcal{P}}_h$ at a rate per unit time independent of h and, as a consequence, a solution Φ_h of the discrete Poisson equation

$$\Phi_h - \overline{\mathcal{P}}_h\Phi_h = F - \overline{\mu}_h(F) \tag{C.22}$$

with a bound of order h^{-1}. The equation (C.22) differs from the discrete Poisson equations of [30, Sections 3.3 and 4.4] in its semigroup: it uses $\overline{\mathcal{P}}_h$ of the discrete Markov chain in (LM). We write F and Φ_h, with capital letters, for the right side and the solution of (C.22), to distinguish them from the test function f and the solutions ϕ_0 of (2.11) and $\tilde{\phi}_h$ of (A.21). For a measurable function F on $\mathbb{R}^d$, let

$$\|F\|_{\mathrm{w}} := \sup_{y\in\mathbb{R}^d} \frac{|F(y)|}{1+|y|^6}. \tag{C.23}$$

Lemma 5 *Let Assumption 1 hold. There is $\gamma > 0$, depending only on m, M, R and d, such that for every $h \in (0, h_0]$, every measurable F with $\|F\|_{\mathrm{w}} < \infty$, every $n \geqslant 0$ and every $y \in \mathbb{R}^d$,*

$$\big|(\overline{\mathcal{P}}_h^n F)(y) - \overline{\mu}_h(F)\big| \leqslant Ce^{-\gamma nh}\,\|F\|_{\mathrm{w}}\big(1+|y|^6\big). \tag{C.24}$$

Consequently, $\Phi_h := \sum_{n\geqslant 0}(\overline{\mathcal{P}}_h^n F - \overline{\mu}_h(F))$ solves (C.22) and satisfies

$$|\Phi_h(y)| \leqslant \frac{C}{h}\,\|F\|_{\mathrm{w}}\big(1+|y|^6\big), \qquad y\in\mathbb{R}^d. \tag{C.25}$$

Proof of Lemma 5. The proof applies Harris's ergodic theorem to the transition $\overline{\mathcal{P}}_{Jh}^*$ with $J := \lceil 1/h\rceil$, so that $1 \leqslant Jh \leqslant 2$, and sums the resulting bound (C.24) over n for (C.25). By (A.26) for the moment of order 6, with $g = \nabla U$, and $1 \leqslant Jh \leqslant 2$, iterating J steps gives the drift condition [12, Assumption 1]

$$\overline{\mathcal{P}}_h|y|^6 \leqslant \Big(1 - \frac{3m}{16}\,h\Big)|y|^6 + Ch \quad\Longrightarrow\quad \overline{\mathcal{P}}_h^J|y|^6 \leqslant e^{-\frac{3m}{16}}\,|y|^6 + C_0, \tag{C.26}$$

with $C_0 > 0$ depending only on m, M, R and d, not on h, and thus the bound

$$\overline{\mathcal{P}}_h^i\big(1+|y|^6\big) \leqslant (1+C_0)\big(1+|y|^6\big), \qquad 0\leqslant i\leqslant J. \tag{C.27}$$

We verify the minorization [12, Assumption 2] for $\overline{\mathcal{P}}_{Jh}^*$ on the level set $\{|y| \leqslant R_0\}$, where $R_0 > 0$ is defined by

$$R_0^6 := \frac{4C_0}{1-e^{-3m/16}}. \tag{C.28}$$

The minorization requires a lower bound on the transition density $\overline{\mathcal{P}}_{Jh}^*(y_0, y_J)$ of J steps of LM, the density at y_J of the law after J steps from the initial state $Y_0 = y_0$, uniformly for $|y_0|, |y_J| \leqslant R_0$. By the Markov

property, this transition density is the integral of the product of the transition densities (C.12) of the J steps over the states $y_1, \dots, y_{J-1}$:

$$\overline{\mathcal{P}}^*_{Jh}(y_0, y_J) = \int_{(\mathbb{R}^d)^{J-1}} \prod_{k=0}^{J-1} \overline{\mathcal{P}}^*_h(y_k, y_{k+1}) \, \mathrm{d}y_1 \cdots \mathrm{d}y_{J-1}. \tag{C.29}$$

Note that $I_d + \frac{h}{2} \nabla^2 V_h(x) \succeq (1 + \frac{h}{2} M)^{-1} I_d \succeq e^{-hM/2} I_d$ by (C.8). Then (C.13)–(C.15) give

$$\Lambda_h(x) \geqslant e^{-Ch(1+|x|^2)}, \quad \Big| \sum_{k=0}^{J-1} (y_{k+1} - y_k) \cdot \nabla V_h\Big(\frac{y_k + y_{k+1}}{2}\Big) - V_h(y_J) + V_h(y_0) \Big| \leqslant C \sum_{k=0}^{J-1} |y_{k+1} - y_k|^3. \tag{C.30}$$

With the Brownian motion $(B_t)_{t \geqslant 0}$ of (2.1), fix $R_1 > 0$, depending only on d, such that

$$\mathbb{P}\Big(2\sqrt{2} \sup_{0 \leqslant t \leqslant 2} |B_t| \geqslant R_1\Big) \leqslant \frac{1}{4}, \tag{C.31}$$

and define

$$\Omega_1 := \Big\{ (y_1, \dots, y_{J-1}) \in (\mathbb{R}^d)^{J-1} : \max_{1 \leqslant k \leqslant J-1} |y_k| \leqslant R_0 + R_1, \ \sum_{k=0}^{J-1} |y_{k+1} - y_k|^3 \leqslant 1 \Big\}.$$

The set Ω_1 is the regular domain of the intermediate states $(y_1, \dots, y_{J-1})$, on which the product of the transition densities $\overline{\mathcal{P}}^*_h(y_k, y_{k+1})$ in (C.29) is bounded below by a product of Gaussian densities.

On Ω_1, (C.12), (C.30), $|V_h(y_J) - V_h(y_0)| \leqslant C(1 + R_0^2)$ by the bound on ∇V_h in (C.14), and $Jh \leqslant 2$ give

$$\prod_{k=0}^{J-1} \overline{\mathcal{P}}^*_h(y_k, y_{k+1}) \geqslant \alpha_1 \prod_{k=0}^{J-1} (4\pi h)^{-\frac{d}{2}} e^{-\frac{|y_{k+1} - y_k|^2}{4h}}, \tag{C.32}$$

with $\alpha_1 > 0$ depending only on m, M, R and d. As a function of $(y_1, \dots, y_{J-1})$, the product of Gaussian densities on the right side of (C.32) is proportional to the joint density of the values

$$y_0 + \frac{k}{J}(y_J - y_0) + \sqrt{2}\Big(B_{kh} - \frac{k}{J} B_{Jh}\Big), \qquad 1 \leqslant k \leqslant J-1,$$

at the times kh of the Brownian motion $y_0 + \sqrt{2}\, B_t$, $0 \leqslant t \leqslant Jh$, conditioned on $y_0 + \sqrt{2}\, B_{Jh} = y_J$, since $(B_t - \frac{t}{Jh} B_{Jh})_{0 \leqslant t \leqslant Jh}$ is independent of B_{Jh}. We write $\mathbb{P}_{y_0, y_J}$ and $\mathbb{E}_{y_0, y_J}$ for the probability and the expectation under this joint density of $(y_1, \dots, y_{J-1})$. Since $|y_0|, |y_J| \leqslant R_0$ and $\sup_{t \leqslant Jh} |B_t - \frac{t}{Jh} B_{Jh}| \leqslant 2 \sup_{t \leqslant 2} |B_t|$, (C.31) gives

$$\mathbb{P}_{y_0, y_J}\Big(\max_{1 \leqslant k \leqslant J-1} |y_k| > R_0 + R_1\Big) \leqslant \frac{1}{4}. \tag{C.33}$$

Under $\mathbb{P}_{y_0, y_J}$, the increment $y_{k+1} - y_k$, $0 \leqslant k \leqslant J-1$, has the law of $\frac{1}{J}(y_J - y_0) + \sqrt{2}\,(B_{(k+1)h} - B_{kh} - \frac{1}{J} B_{Jh})$, and $\mathbb{E}[(B_{(k+1)h} - B_{kh} - \frac{1}{J} B_{Jh})^{\otimes 2}] = (1 - \frac{1}{J})\, h I_d$. Markov's inequality and this identity give

$$\begin{aligned} \mathbb{P}_{y_0, y_J}\Big(\sum_{k=0}^{J-1} |y_{k+1} - y_k|^3 > 1\Big) &\leqslant \mathbb{E}_{y_0, y_J} \sum_{k=0}^{J-1} |y_{k+1} - y_k|^3 \\ &\leqslant 4 \sum_{k=0}^{J-1} \Big(\frac{|y_J - y_0|^3}{J^3} + 2^{\frac{3}{2}}\, \mathbb{E}\Big| B_{(k+1)h} - B_{kh} - \frac{1}{J} B_{Jh} \Big|^3 \Big) \leqslant C\big(R_0^3 h^2 + h^{\frac{1}{2}}\big) \leqslant \frac{1}{4} \end{aligned} \tag{C.34}$$

by choosing h small enough. By (C.33) and (C.34), $\mathbb{P}_{y_0, y_J}(\Omega_1) \geqslant \frac{1}{2}$. The product of Gaussian densities in (C.32) integrates over Ω_1 to $(4\pi Jh)^{-\frac{d}{2}} e^{-|y_J - y_0|^2/(4Jh)}\, \mathbb{P}_{y_0, y_J}(\Omega_1)$, so that (C.29) with the integral restricted to Ω_1, $\mathbb{P}_{y_0, y_J}(\Omega_1) \geqslant \frac{1}{2}$, $1 \leqslant Jh \leqslant 2$ and $|y_J - y_0| \leqslant 2R_0$ give

$$\overline{\mathcal{P}}^*_{Jh}(y_0, y_J) \geqslant \frac{\alpha_1}{2} (4\pi Jh)^{-\frac{d}{2}} e^{-\frac{|y_J - y_0|^2}{4Jh}} \geqslant \alpha_2, \tag{C.35}$$

where α_2 is defined as

$$\alpha_2 := \frac{\alpha_1}{2}\,(8\pi)^{-\frac{d}{2}} e^{-R_0^2} > 0.$$

By (C.35), the minorization [12, Assumption 2] holds with the uniform distribution on $\{|y| \leqslant R_0\}$: for $|y_0| \leqslant R_0$ and $y_J \in \mathbb{R}^d$,

$$\overline{\mathcal{P}}^*_{Jh}(y_0, y_J) \geqslant \frac{\alpha\, \mathbf{1}_{\{|y_J| \leqslant R_0\}}}{\mathrm{Leb}(\{|y| \leqslant R_0\})}, \tag{C.36}$$

where α is defined as

$$\alpha := \min\Big\{\frac{1}{2},\ \alpha_2\,\mathrm{Leb}\big(\{|y| \leqslant R_0\}\big)\Big\}. \tag{C.37}$$

Harris's ergodic theorem in the form of [12, Theorem 1.3] applies to the transition $\overline{\mathcal{P}}^*_{Jh}$ with the Lyapunov function $|x|^6$, the drift condition (C.26) and the minorization (C.36) on the level set $\{|y| \leqslant R_0\}$, with α of (C.37), since $R_0^6 > 2C_0/(1 - e^{-3m/16})$ by (C.28). The theorem gives $\bar\beta > 0$ and $\bar\alpha \in (0,1)$, depending only on m, M, R and d through $e^{-3m/16}$, C_0, R_0 and α, such that

$$\int_{\mathbb{R}^d} \big(1 + \bar\beta|x|^6\big)\, \big|\overline{\mathcal{P}}^*_{Jh}\mu_1 - \overline{\mathcal{P}}^*_{Jh}\mu_2\big|(\mathrm{d}x) \leqslant \bar\alpha \int_{\mathbb{R}^d} \big(1 + \bar\beta|x|^6\big)\, |\mu_1 - \mu_2|(\mathrm{d}x) \tag{C.38}$$

for distributions μ_1 and μ_2 on $\mathbb{R}^d$. As in the proof of Lemma 3, (C.38) applied n times to the point mass at y and to the invariant distribution $\overline{\mu}_h$, with $\overline{\mu}_h(|x|^6) \leqslant C$, gives

$$\big|(\overline{\mathcal{P}}_h^{nJ} F)(y) - \overline{\mu}_h(F)\big| \leqslant C\bar\alpha^{\,n}\, \|F\|_{\mathrm{w}} \big(1 + |y|^6\big), \qquad n \geqslant 0. \tag{C.39}$$

Every integer $n \geqslant 0$ can be written as $n = qJ + k$ with $q \geqslant 0$ and $0 \leqslant k < J$, and then $q \geqslant \frac{nh}{2} - 1$ since $Jh \leqslant 2$. By (C.27), $\|\overline{\mathcal{P}}_h^k F\|_{\mathrm{w}} \leqslant (1 + C_0)\|F\|_{\mathrm{w}}$, and $\overline{\mu}_h(\overline{\mathcal{P}}_h^k F) = \overline{\mu}_h(F)$ by the invariance of $\overline{\mu}_h$. Hence (C.39) with q in place of n and $\overline{\mathcal{P}}_h^k F$ in place of F gives (C.24) with $\gamma := \frac{1}{2}|\log\bar\alpha|$.

To prove (C.25) and that Φ_h solves (C.22), let h_0 also satisfy $\gamma h_0 \leqslant 1$, so that $\sum_{n\geqslant 0} e^{-\gamma n h} \leqslant \frac{2}{\gamma h}$. By (C.24), the series defining Φ_h converges absolutely and satisfies (C.25), and dominated convergence with (C.27) for $i = 1$ gives (C.22):

$$\overline{\mathcal{P}}_h \Phi_h = \sum_{n\geqslant 1} \big(\overline{\mathcal{P}}_h^n F - \overline{\mu}_h(F)\big) = \Phi_h - \big(F - \overline{\mu}_h(F)\big).$$

□

Lemma 6 below derives the semigroup estimates of $\overline{\mathcal{P}}_h$ on functions F with $\|F\|_{\mathrm{w}} < \infty$.

Lemma 6 *Let Assumption 1 hold, $h \in (0, h_0]$, and let F be measurable with $\|F\|_{\mathrm{w}} < \infty$. Then*

$$\big\|\nabla^j (\overline{\mathcal{P}}_h F)(y)\big\| \leqslant C h^{-\frac{j}{2}}\, \|F\|_{\mathrm{w}} (1 + |y|)^6, \qquad j = 1, 2, 3, \qquad y \in \mathbb{R}^d. \tag{C.40}$$

Proof of Lemma 6. To prove (C.40), we bound the derivatives in y of the transition density $\overline{\mathcal{P}}^*_h(y, y')$ and differentiate

$$(\overline{\mathcal{P}}_h F)(y) = \int_{\mathbb{R}^d} F(y')\, \overline{\mathcal{P}}^*_h(y, y')\, \mathrm{d}y' \tag{C.41}$$

under the integral. Fix $y' \in \mathbb{R}^d$. By (C.12), (C.13) and (C.10),

$$\overline{\mathcal{P}}^*_h(y, y') = (4\pi h)^{-\frac{d}{2}} \det\Big(I_d + \frac{h}{2}\nabla^2 V_h\Big(\frac{y + y'}{2}\Big)\Big) \exp\Big(-\frac{1}{h}\Big|w_h\Big(\frac{y + y'}{2}\Big) - y\Big|^2\Big). \tag{C.42}$$

For the initial state $Y_0 = y$ and the state $y' = y + \overline{D}_0$ after the LM step, distributed as $\overline{\mathcal{P}}^*_h(y, \cdot)$, the midpoint is $\frac{y+y'}{2} = Z_0 - \frac{h}{2}\nabla U(Z_0)$ by (C.5), and (C.6) gives

$$\mathcal{Q}_h(y) := w_h\Big(\frac{y + y'}{2}\Big) - y = Z_0 - y = \sqrt{\frac{h}{2}}\,\chi_0, \qquad h^{-\frac{1}{2}}\,|\mathcal{Q}_h(y)| = \frac{|\chi_0|}{\sqrt{2}}. \tag{C.43}$$

Here $\mathcal{Q}_h(y)$ depends only on y for fixed y', and the other equalities hold at $y' = y + \overline{D}_0$. Hence, under $\overline{\mathcal{P}}_h^*(y,\cdot)$, the law of $h^{-1/2}\,|\mathcal{Q}_h(y)|$ does not depend on h.

By (C.9) and (C.17), the derivatives of $\mathcal{Q}_h(y)$ with respect to y satisfy

$$\big\|\nabla \mathcal{Q}_h(y)\big\| \leqslant 2, \qquad \big\|\nabla^i \mathcal{Q}_h(y)\big\| \leqslant Ch, \quad i = 2,3,4. \tag{C.44}$$

Since $(\nabla w_h)(\frac{y+y'}{2}) = 2(I_d + \nabla \mathcal{Q}_h(y))$, (C.9) gives

$$\begin{aligned} \det\Big(I_d + \frac{h}{2}\,\nabla^2 V_h\Big(\frac{y+y'}{2}\Big)\Big) &= 2^d \det\big(I_d + \nabla \mathcal{Q}_h(y)\big) \geqslant \Big(\frac{8}{9}\Big)^d, \\ \big\|\nabla^j \det\big(I_d + \nabla \mathcal{Q}_h(y)\big)\big\| &\leqslant C \det\big(I_d + \nabla \mathcal{Q}_h(y)\big), \qquad j = 1,2,3. \end{aligned} \tag{C.45}$$

The Leibniz rule for the product (C.42), with (C.45) for its determinant and the chain rule for $\exp(-\frac{1}{h}\,|\mathcal{Q}_h(y)|^2)$, gives, for $j = 1,2,3$,

$$\big\|\nabla^j \overline{\mathcal{P}}_h^*(y,y')\big\| \leqslant C\,\overline{\mathcal{P}}_h^*(y,y') \sum_{\ell=0}^{j} \sum_{i_1+\cdots+i_q=\ell} \prod_{k=1}^{q} \frac{1}{h}\,\big\|\nabla^{i_k} |\mathcal{Q}_h(y)|^2\big\|, \tag{C.46}$$

where the inner sum runs over the partitions of the ℓ directions of differentiation into blocks of sizes $i_1, \dots, i_q$, and the empty product for $\ell = 0$ is 1. By the Leibniz rule and (C.44),

$$\frac{1}{h}\,\big\|\nabla |\mathcal{Q}_h(y)|^2\big\| \leqslant \frac{4}{h}\,|\mathcal{Q}_h(y)|, \quad \frac{1}{h}\,\big\|\nabla^2 |\mathcal{Q}_h(y)|^2\big\| \leqslant \frac{8}{h} + C\,|\mathcal{Q}_h(y)|, \quad \frac{1}{h}\,\big\|\nabla^3 |\mathcal{Q}_h(y)|^2\big\| \leqslant C + C\,|\mathcal{Q}_h(y)|,$$

so that, for $h \leqslant 1$ and $i = 1,2,3$,

$$\frac{1}{h}\,\big\|\nabla^i |\mathcal{Q}_h(y)|^2\big\| \leqslant Ch^{-\frac{i}{2}}\big(1 + h^{-\frac{1}{2}}\,|\mathcal{Q}_h(y)|\big)^i. \tag{C.47}$$

In each product of (C.46), the sizes i_k add up to $\ell \leqslant j$, so that (C.47) and $h \leqslant 1$ give, for $j = 1,2,3$,

$$\big\|\nabla^j \overline{\mathcal{P}}_h^*(y,y')\big\| \leqslant Ch^{-\frac{j}{2}}\big(1 + h^{-\frac{1}{2}}\,|\mathcal{Q}_h(y)|\big)^j\,\overline{\mathcal{P}}_h^*(y,y'). \tag{C.48}$$

Since (C.10), (C.8), (C.17) and $hM \leqslant \frac{1}{4}$ give $|\mathcal{Q}_h(y)| \geqslant \frac{3}{8}\,|y' - y| - \frac{1}{4}\,(1 + |y|)$, the right side of (C.48) is bounded by a Gaussian function of y' uniformly for y in bounded sets, so that dominated convergence allows differentiating (C.41) under the integral. Since $\overline{\mathcal{P}}_h^*(y,\cdot)$ is the law of the state $y + \overline{D}_0$ after the LM step, (C.48), (C.43) and (C.4) give, for $j = 1,2,3$,

$$\int_{\mathbb{R}^d} |F(y')|\,\big\|\nabla^j \overline{\mathcal{P}}_h^*(y,y')\big\|\,\mathrm{d}y' \leqslant Ch^{-\frac{j}{2}}\,\|F\|_{\mathrm{w}}\,\mathbb{E}\Big[(1 + |\chi_0|)^3\big(1 + |y + \overline{D}_0|^6\big)\Big] \leqslant Ch^{-\frac{j}{2}}\,\|F\|_{\mathrm{w}}(1 + |y|)^6,$$

which gives (C.40). □

C.3 One step of NC-SGLM against one step of LM

We compare NC-SGLM with LM through the one-step operators $\mathcal{P}_h$ and $\overline{\mathcal{P}}_h$ of (C.2), on test functions whose derivatives of orders 2 and 3 grow at most polynomially.

Lemma 7 *Let Assumptions 1 and 3 hold, $h \in (0, h_0]$ and $P \geqslant 2$, and let $K \geqslant 0$ and $G \in C^3(\mathbb{R}^d)$ satisfy $\|\nabla^2 G(x)\| + \|\nabla^3 G(x)\| \leqslant K(1 + |x|)^6$ for $x \in \mathbb{R}^d$. Then the one-step operators $\mathcal{P}_h$ and $\overline{\mathcal{P}}_h$ of (C.2) satisfy*

$$\big|(\mathcal{P}_h G)(y) - (\overline{\mathcal{P}}_h G)(y)\big| \leqslant CK\,\frac{h^3}{P}\,(1 + |y|)^{10}, \qquad y \in \mathbb{R}^d. \tag{C.49}$$

Proof of Lemma 7. Fix the initial state $Y_0 = y$. With the SGLM displacement D_0 and the noise correction ζ_0 of (A.11) and the LM step $\overline{D}_0$ of (C.2), the state Y_1 after one step of NC-SGLM reads

$$Y_1 = Y_0 + D_0 + \zeta_0 = Y_0 + \overline{D}_0 + \hat{\zeta}_0, \qquad \hat{\zeta}_0 = -h\big(\hat{g}_P(Z_0;\boldsymbol{\xi}_0) - \nabla U(Z_0)\big) + \zeta_0, \tag{C.50}$$

where $\hat{\zeta}_0$ follows from $D_0 - \overline{D}_0 = -h\big(\hat{g}_P(Z_0;\boldsymbol{\xi}_0) - \nabla U(Z_0)\big)$. The terms of order h^2/P from ζ_0 and from the batch noise at Z_0 cancel up to $\mathcal{C}(Z_0) - \mathcal{C}(Y_0)$, and two bounds on the batch noise control the remaining terms of Taylor's formula for G about $Y_0 + \overline{D}_0$.

To prove (C.51), let $x \in \mathbb{R}^d$. The batch noise $\hat{g}_P(x;\boldsymbol{\xi}) - \nabla U(x)$ is the mean of the P independent vectors $g(x;\xi^{(p)}) - \nabla U(x)$, which have mean 0 and satisfy $\mathbb{E}_\xi|g(x;\xi) - \nabla U(x)|^q \leqslant C(1+|x|)^q$ for $q = 2, 4$ by (3.4) and $|\nabla U(x)| \leqslant M(1+|x|)$. In the expansion of $|\sum_{p=1}^P (g(x;\xi^{(p)}) - \nabla U(x))|^4$, the terms in which some index p occurs in only 1 of the 4 factors have mean 0, and the other terms, at most $3P^2$ in number, give

$$\mathbb{E}_{\boldsymbol{\xi}}\big[|\hat{g}_P(x;\boldsymbol{\xi}) - \nabla U(x)|^4\big] \leqslant \frac{C}{P^2}\,(1+|x|)^4. \tag{C.51}$$

To prove (C.52), let $z, v \in \mathbb{R}^d$. By (3.4) and dominated convergence, $\mathbb{E}_\xi[\nabla g(z;\xi)] = \nabla^2 U(z)$, so that $(\nabla \hat{g}_P(z;\boldsymbol{\xi}) - \nabla^2 U(z))\,v$ is the mean of P independent vectors of mean 0 with $\mathbb{E}_\xi|(\nabla g(z;\xi) - \nabla^2 U(z))\,v|^2 \leqslant \mathbb{E}_\xi|\nabla g(z;\xi)\,v|^2 \leqslant M^2|v|^2$. For $x, x' \in \mathbb{R}^d$, the difference of the batch noise at x' and at x is

$$\big(\hat{g}_P(x';\boldsymbol{\xi}) - \nabla U(x')\big) - \big(\hat{g}_P(x;\boldsymbol{\xi}) - \nabla U(x)\big) = \int_0^1 \Big(\nabla \hat{g}_P\big(x + s(x'-x);\boldsymbol{\xi}\big) - \nabla^2 U\big(x + s(x'-x)\big)\Big)(x'-x)\,\mathrm{d}s.$$

Jensen's inequality over s, with $\mathbb{E}_{\boldsymbol{\xi}}|(\nabla \hat{g}_P(z;\boldsymbol{\xi}) - \nabla^2 U(z))\,v|^2 \leqslant \frac{M^2}{P}\,|v|^2$, gives

$$\mathbb{E}_{\boldsymbol{\xi}}\Big[\big|\big(\hat{g}_P(x';\boldsymbol{\xi}) - \nabla U(x')\big) - \big(\hat{g}_P(x;\boldsymbol{\xi}) - \nabla U(x)\big)\big|^2\Big] \leqslant \frac{M^2}{P}\,|x'-x|^2, \qquad x, x' \in \mathbb{R}^d. \tag{C.52}$$

Since $\boldsymbol{\xi}_0$ is independent of χ_0, (C.51) with $x = Z_0$ holds given χ_0. Hölder's inequality between the moments of orders 4 and 14 of the batch noise, with the moment of order 14 bounded by (A.2), and the bound $|\zeta_0| \leqslant h\,\|\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\|^{1/2}|\chi_0|$ of (A.12) with (A.9) give

$$\mathbb{E}|\hat{\zeta}_0|^4 \leqslant \frac{Ch^4}{P^2}\,(1+|Y_0|)^4, \qquad \mathbb{E}|\hat{\zeta}_0|^9 \leqslant \frac{Ch^9}{P}\,(1+|Y_0|)^9. \tag{C.53}$$

Taylor's formula for G about $Y_0 + \overline{D}_0$ to the order 2 gives

$$\mathbb{E}\big[G(Y_1) - G(Y_0 + \overline{D}_0)\big] = \mathbb{E}\big[\nabla G(Y_0 + \overline{D}_0)\cdot\hat{\zeta}_0\big] + \frac{1}{2}\,\mathbb{E}\big[\nabla^2 G(Y_0 + \overline{D}_0) : \hat{\zeta}_0^{\otimes 2}\big] + \mathbb{E}\big[\varrho_3^{Y_1}\big], \tag{C.54}$$

where the remainder $\varrho_3^{Y_1}$ satisfies, by $\|\nabla^3 G(x)\| \leqslant K(1+|x|)^6$, Hölder's inequality, (C.53) and (C.4),

$$|\varrho_3^{Y_1}| \leqslant CK\Big(\big(1 + |Y_0 + \overline{D}_0|\big)^6|\hat{\zeta}_0|^3 + |\hat{\zeta}_0|^9\Big), \qquad \mathbb{E}|\varrho_3^{Y_1}| \leqslant CK\frac{h^3}{P}(1+|Y_0|)^9. \tag{C.55}$$

Since $\overline{D}_0$ is a function of χ_0 and the batch noise has mean 0 given χ_0, (C.50) reduces the term of (C.54) linear in $\hat{\zeta}_0$ to

$$\mathbb{E}\big[\nabla G(Y_0 + \overline{D}_0)\cdot\hat{\zeta}_0\big] = \mathbb{E}\big[\nabla G(Y_0 + \overline{D}_0)\cdot\zeta_0\big]. \tag{C.56}$$

For $\psi \in C^1(\mathbb{R}^d)$ such that ψ and $\nabla\psi$ grow at most polynomially, Gaussian integration by parts in χ_0 and the Jacobian of $\overline{D}_0$ in χ_0 read

$$\mathbb{E}\big[\chi_0\,\psi(\chi_0)\big] = \mathbb{E}\big[\nabla\psi(\chi_0)\big], \qquad \nabla_{\chi_0}\overline{D}_0 = \sqrt{2h}\,I_d - h\sqrt{\frac{h}{2}}\,\nabla^2 U(Z_0). \tag{C.57}$$

Since $\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)$ is independent of χ_0, (C.57) gives

$$\begin{aligned}\mathbb{E}\big[\nabla G(Y_0 + \overline{D}_0)\cdot\zeta_0\big] &= \sqrt{2h}\,\mathbb{E}\Big[\nabla G(Y_0 + \overline{D}_0)\cdot\Big(\mathbb{E}\big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)\big] - I_d\Big)\chi_0\Big] \\ &= 2h\,\mathbb{E}\big[\nabla^2 G(Y_0 + \overline{D}_0)\big] : \Big(\mathbb{E}\big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)\big] - I_d\Big) \\ &\quad - h^2\,\mathbb{E}\Big[\mathrm{Tr}\Big[\Big(\mathbb{E}\big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)\big] - I_d\Big)\,\nabla^2 G(Y_0 + \overline{D}_0)\,\nabla^2 U(Z_0)\Big]\Big].\end{aligned} \tag{C.58}$$

By (A.10), the mean in (2.5) and (A.9),

$$\Big\|\mathbb{E}\big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)\big] - I_d + \frac{h}{4P}\,\mathcal{C}(Y_0)\Big\| \leqslant \frac{Ch^2}{P^2}\,(1+|Y_0|)^4, \qquad \Big\|\mathbb{E}\big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)\big] - I_d\Big\| \leqslant \frac{Ch}{P}\,(1+|Y_0|)^2,$$

so that (C.58), the bound on $\nabla^2 G$ and (C.4) give

$$\Big|\mathbb{E}\big[\nabla G(Y_0+\overline{D}_0)\cdot\zeta_0\big] + \frac{h^2}{2P}\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0)\big] : \mathcal{C}(Y_0)\Big| \leqslant CK\frac{h^3}{P}(1+|Y_0|)^{10}. \tag{C.59}$$

By (C.50), the term of (C.54) quadratic in $\hat{\zeta}_0$ is the expectation of

$$\frac{1}{2}\,\nabla^2 G(Y_0+\overline{D}_0) : \hat{\zeta}_0^{\otimes 2} = \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3, \tag{C.60}$$

where the three terms are defined as

$$\begin{gathered}\mathcal{T}_1 := \frac{h^2}{2}\,\nabla^2 G(Y_0+\overline{D}_0) : \big(\hat{g}_P(Z_0;\boldsymbol{\xi}_0) - \nabla U(Z_0)\big)^{\otimes 2}, \qquad \mathcal{T}_2 := \frac{1}{2}\,\nabla^2 G(Y_0+\overline{D}_0) : \zeta_0^{\otimes 2},\\ \mathcal{T}_3 := -h\,\nabla^2 G(Y_0+\overline{D}_0)\big[\hat{g}_P(Z_0;\boldsymbol{\xi}_0) - \nabla U(Z_0),\,\zeta_0\big].\end{gathered} \tag{C.61}$$

Since the batch noise has the covariance $\mathcal{C}(Z_0)/P$ given χ_0 by (2.4) at $x = Z_0$, (C.61) gives

$$\mathbb{E}[\mathcal{T}_1] = \frac{h^2}{2P}\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0) : \mathcal{C}(Z_0)\big]. \tag{C.62}$$

The bound $|\zeta_0| \leqslant \frac{h}{2}\sqrt{2h}\,\|\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\|\,|\chi_0|$ of (A.12) and (A.9) give

$$\big|\mathbb{E}[\mathcal{T}_2]\big| \leqslant CK\frac{h^3}{P^2}(1+|Y_0|)^{10}. \tag{C.63}$$

In $\mathcal{T}_3$, we split the batch noise at Z_0 into the batch noise at Y_0, which is independent of χ_0, and the difference of the batch noise at Z_0 and at Y_0, bounded in (C.52):

$$\mathcal{T}_3 = \mathcal{T}_{3,a} + \mathcal{T}_{3,b}, \tag{C.64}$$

where $\mathcal{T}_{3,a}$ and $\mathcal{T}_{3,b}$ are defined as

$$\begin{aligned}\mathcal{T}_{3,a} &:= -h\,\nabla^2 G(Y_0+\overline{D}_0)\big[\hat{g}_P(Y_0;\boldsymbol{\xi}_0) - \nabla U(Y_0),\,\zeta_0\big],\\ \mathcal{T}_{3,b} &:= -h\,\nabla^2 G(Y_0+\overline{D}_0)\big[\hat{g}_P(Z_0;\boldsymbol{\xi}_0) - \nabla U(Z_0) - \hat{g}_P(Y_0;\boldsymbol{\xi}_0) + \nabla U(Y_0),\,\zeta_0\big].\end{aligned} \tag{C.65}$$

For $\mathcal{T}_{3,a}$ of (C.65), (C.57) given $\boldsymbol{\xi}_0$ replaces the factor χ_0 of ζ_0 by the derivative of $\nabla^2 G(Y_0+\overline{D}_0)$ in χ_0, and the bound on $\nabla^3 G$, the Jacobian, (A.10), (A.4), (A.9) with $r = 2$ and (C.4) give

$$\big|\mathbb{E}[\mathcal{T}_{3,a}]\big| \leqslant CKh^3(1+|Y_0|)^6\,\mathbb{E}\Big[\big|\hat{g}_P(Y_0;\boldsymbol{\xi}_0) - \nabla U(Y_0)\big|\,\|\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\|\Big] \leqslant CK\frac{h^3}{P^{\frac{3}{2}}}(1+|Y_0|)^9. \tag{C.66}$$

For $\mathcal{T}_{3,b}$, the bound on $\nabla^2 G$, (C.52) with $x = Y_0$ and $x' = Z_0$, which holds given χ_0, the identity $|Z_0 - Y_0| = \sqrt{h/2}\,|\chi_0|$ and the bound $|\zeta_0| \leqslant \frac{h}{2}\sqrt{2h}\,\|\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\|\,|\chi_0|$ of (A.12) give

$$\big|\mathbb{E}[\mathcal{T}_{3,b}]\big| \leqslant CK\frac{h^3}{P^{\frac{3}{2}}}(1+|Y_0|)^9. \tag{C.67}$$

By (C.64), (C.66) and (C.67),

$$\big|\mathbb{E}[\mathcal{T}_3]\big| \leqslant CK\frac{h^3}{P^{\frac{3}{2}}}(1+|Y_0|)^9. \tag{C.68}$$

By (C.60), (C.62), (C.63) and (C.68),

$$\Big|\frac{1}{2}\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0) : \hat{\zeta}_0^{\otimes 2}\big] - \frac{h^2}{2P}\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0) : \mathcal{C}(Z_0)\big]\Big| \leqslant CK\frac{h^3}{P^{\frac{3}{2}}}(1+|Y_0|)^{10}. \tag{C.69}$$

Inserting (C.56), (C.59), (C.69) and (C.55) into (C.54) gives

$$\Big|\mathbb{E}\big[G(Y_1) - G(Y_0 + \overline{D}_0)\big] - \frac{h^2}{2P}\,\mathbb{E}\Big[\nabla^2 G(Y_0 + \overline{D}_0) : \big(\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\big)\Big]\Big| \leqslant CK\frac{h^3}{P}(1 + |Y_0|)^{10}. \tag{C.70}$$

It remains to bound the term with $\mathcal{C}(Z_0) - \mathcal{C}(Y_0)$ in (C.70). The bound (A.43) at $k = 0$ and (A.8) give

$$\big\|\mathbb{E}[\mathcal{C}(Z_0)] - \mathcal{C}(Y_0)\big\| \leqslant Ch(1 + |Y_0|)^2, \qquad \big\|\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\big\| \leqslant 4M^2(1 + |Y_0| + |Z_0|)^2\,|Z_0 - Y_0|.$$

The bound on $\nabla^3 G$ gives

$$\big\|\nabla^2 G(Y_0 + \overline{D}_0) - \nabla^2 G(Y_0)\big\| \leqslant K\big(1 + |Y_0| + |\overline{D}_0|\big)^6\,|\overline{D}_0|.$$

With $\nabla^2 G(Y_0 + \overline{D}_0)$ split into $\nabla^2 G(Y_0)$ and $\nabla^2 G(Y_0 + \overline{D}_0) - \nabla^2 G(Y_0)$, the bound in (C.4) gives

$$\Big|\mathbb{E}\Big[\nabla^2 G(Y_0 + \overline{D}_0) : \big(\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\big)\Big]\Big| \leqslant CKh\,(1 + |Y_0|)^9,$$

and with (C.70) we prove (C.49). □

C.4 The midpoint distribution and its change under LM

This subsection constructs the density ν_h that meets the two requirements stated in the opening of Appendix C. This subsection does not involve stochastic gradients, so the only randomness comes from the Gaussian noise in one step of LM. By (C.12), the product $e^{-V_h(y)}\,\overline{\mathcal{P}}_h^*(y, y')$ is symmetric in (y, y') up to a factor whose exponent is half the error of the midpoint rule bounded in (C.15). Hence one step of LM changes e^{-V_h} only through this error, which gives a term of order h^2 and terms of order h^3. We seek $\nu_h \propto \exp(-V_h + \Theta_h)$ with a correction Θ_h of order h. To match the expansion (C.1), (C.16) formally gives

$$\Theta_h - \frac{h}{4}\,\Delta V_h = -\frac{h^2}{16}\Big(\nabla^2 U : \nabla^2 U + \frac{1}{2}\,\Delta^2 U\Big).$$

By (C.16), $\nabla^2 V_h : \nabla^2 V_h + \frac{1}{2}\,\Delta^2 V_h = \nabla^2 U : \nabla^2 U + \frac{1}{2}\,\Delta^2 U + \mathcal{O}(h)$. Hence, formally,

$$\Theta_h = \frac{h}{4}\,\Delta V_h - \frac{h^2}{16}\Big(\nabla^2 V_h : \nabla^2 V_h + \frac{1}{2}\,\Delta^2 V_h\Big) + \mathcal{O}(h^3).$$

In the change of e^{-V_h} under one step of LM, the term $\frac{h}{4}\,\Delta V_h$ of Θ_h also cancels the term of order h^2, and Θ_h cancels the terms of order h^3 other than the explicit term. Next we drop $\mathcal{O}(h^3)$ and define the midpoint correction potential and the midpoint distribution.

The *midpoint correction potential* Θ_h and the *midpoint distribution* ν_h are

$$\Theta_h(x) := \frac{h}{4}\,\Delta V_h(x) - \frac{h^2}{16}\Big(\nabla^2 V_h(x) : \nabla^2 V_h(x) + \frac{1}{2}\,\Delta^2 V_h(x)\Big), \qquad x \in \mathbb{R}^d, \tag{C.71}$$

$$\nu_h(x) := \frac{1}{\mathcal{Z}_h}\exp\big(-V_h(x) + \Theta_h(x)\big), \qquad x \in \mathbb{R}^d, \tag{C.72}$$

where $\mathcal{Z}_h$ is the normalizing constant.

First, we show that the convolution of ν_h with $\mathcal{N}(0, \frac{h}{2}I_d)$ differs from π by $\mathcal{O}(h^3)$.

Lemma 8 *Let Assumption 1 hold. For every $h \in (0, h_0]$, the midpoint distribution ν_h satisfies*

$$\int_{\mathbb{R}^d}(1 + |x|)\,\Big|\Big(\nu_h * \mathcal{N}\Big(0, \frac{h}{2}I_d\Big)\Big)(x) - \pi(x)\Big|\,\mathrm{d}x \leqslant Ch^3. \tag{C.73}$$

Proof of Lemma 8. By (3.2) and (C.16), $V_h(x) - U(0) \leqslant C$ for $|x| \leqslant 1$. This bound, (C.71) and (C.14) give $\mathcal{Z}_h \geqslant e^{-U(0)}/C$ and

$$\max_{0 \leqslant i \leqslant 4} \|\nabla^i \Theta_h(x)\| \leqslant Ch, \qquad \nu_h(x) \leqslant C e^{-\frac{m}{64}|x|^2}, \qquad x \in \mathbb{R}^d, \tag{C.74}$$

so that $\nu_h(|x|^p) \leqslant C$ for every $p \geqslant 1$, with C depending also on p.

We next expand ν_h about π. The identities $\Delta e^{-U} = (|\nabla U|^2 - \Delta U)\, e^{-U}$ and

$$\Delta^2 e^{-U} = \Big(\big(|\nabla U|^2 - \Delta U\big)^2 + 2\,\nabla^2 U : \nabla^2 U + 4\,\nabla U \cdot \nabla \Delta U - \Delta^2 U - 4\,\nabla^2 U[\nabla U, \nabla U] \Big)\, e^{-U}, \tag{C.75}$$

together with (C.71) and (C.16), give

$$-V_h(x) + \Theta_h(x) = -U(x) + a_h(x) + \varrho_3^{V_h}(x), \qquad |\varrho_3^{V_h}(x)| \leqslant Ch^3(1+|x|)^4, \tag{C.76}$$

where $a_h(x)$, which contains the terms of first and second order in h of the expansion, is defined as

$$a_h(x) := -\frac{h}{4}\, e^U \Delta e^{-U} + \frac{h^2}{32} \Big(e^U \Delta^2 e^{-U} - \big(e^U \Delta e^{-U}\big)^2 \Big). \tag{C.77}$$

By (C.75), the term $(|\nabla U|^2 - \Delta U)^2$ of $e^U \Delta^2 e^{-U}$ cancels $(e^U \Delta e^{-U})^2$, so that (3.2) gives $|a_h| \leqslant Ch(1+|x|)^2$. We apply Taylor's formula for the exponential function about 0 to the order 2 at $a_h + \varrho_3^{V_h}$ of (C.76), and bound the factor $\max\{e^{-U}, e^{-V_h+\Theta_h}\}$ in its remainder by the lower bounds on U of [30, Lemma 17] and on V_h in (C.14), with (C.74):

$$\Big| e^{-V_h+\Theta_h} - e^{-U} \Big(1 + a_h + \frac{1}{2}\, a_h^2\Big) \Big| \leqslant Ch^3 (1+|x|)^{12} e^{-U(0) - \frac{m}{64}|x|^2},$$

and, with (C.77) and $\frac{1}{2}\, a_h^2 = \frac{h^2}{32}\, (e^U \Delta e^{-U})^2 + \mathcal{O}(h^3(1+|x|)^4)$, this bound gives

$$\Big| e^{-V_h+\Theta_h} - e^{-U} + \frac{h}{4}\, \Delta e^{-U} - \frac{h^2}{32}\, \Delta^2 e^{-U} \Big| \leqslant Ch^3 (1+|x|)^{12} e^{-U(0) - \frac{m}{64}|x|^2}. \tag{C.78}$$

Integrating (C.78) over $\mathbb{R}^d$, since Δe^{-U} and $\Delta^2 e^{-U}$ integrate to 0, gives $\mathcal{Z}_h = (1 + \mathcal{O}(h^3)) \int_{\mathbb{R}^d} e^{-U} \mathrm{d}x$ and

$$\nu_h = \pi - \frac{h}{4}\, \Delta \pi + \frac{h^2}{32}\, \Delta^2 \pi + \varrho_3^{\nu_h}, \qquad \int_{\mathbb{R}^d} (1+|x|)\, |\varrho_3^{\nu_h}(x)|\, \mathrm{d}x \leqslant Ch^3. \tag{C.79}$$

Finally, we convolve (C.79) with $\mathcal{N}(0, \frac{h}{2} I_d)$. For a function ω integrable against $1+|x|$, the heat semigroup satisfies $\omega * \mathcal{N}(0, \frac{h}{2} I_d) = e^{\frac{h}{4}\Delta} \omega$ and

$$\int_{\mathbb{R}^d} (1+|x|)\, \big| e^{t\Delta} \omega \big|\, \mathrm{d}x \leqslant \big(1 + \sqrt{2td}\big) \int_{\mathbb{R}^d} (1+|x|)\, |\omega|\, \mathrm{d}x, \qquad t \geqslant 0. \tag{C.80}$$

For $j \in \{0,1,2\}$ and $\omega \in C^{2j+2}(\mathbb{R}^d)$ whose derivatives of orders 0 to $2j+2$ are integrable against $1+|x|$, Taylor's formula in time for the heat semigroup, with (C.80) for $\Delta^{j+1}\omega$ in place of ω, gives

$$\int_{\mathbb{R}^d} (1+|x|) \Big| \omega * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big) - \sum_{i=0}^{j} \frac{h^i}{4^i\, i!}\, \Delta^i \omega \Big|\, \mathrm{d}x \leqslant Ch^{j+1} \int_{\mathbb{R}^d} (1+|x|)\, |\Delta^{j+1}\omega|\, \mathrm{d}x. \tag{C.81}$$

The bound (C.81) applies to the terms of (C.79) with $j = 2$ for π, $j = 1$ for $\Delta\pi$ and $j = 0$ for $\Delta^2\pi$. In each case $\Delta^{j+1}\omega = \Delta^3 \pi$, which involves the derivatives of U of orders up to 6. With (C.80) at $t = \frac{h}{4}$ for $\omega = \varrho_3^{\nu_h}$, the bounds (C.81) for π, $\Delta\pi$ and $\Delta^2\pi$ give

$$\int_{\mathbb{R}^d} (1+|x|) \Big| \nu_h * \mathcal{N}\Big(0, \frac{h}{2} I_d\Big) - \pi - \frac{h}{4}\, \Delta\pi - \frac{h^2}{32}\, \Delta^2 \pi + \frac{h}{4} \Big(\Delta\pi + \frac{h}{4}\, \Delta^2 \pi \Big) - \frac{h^2}{32}\, \Delta^2 \pi \Big|\, \mathrm{d}x \leqslant Ch^3,$$

in which the terms with $\Delta\pi$ and $\Delta^2\pi$ cancel, so that the bound reduces to (C.73). □

Second, we show that the change of ν_h under one step of LM has an explicit term of order h^3.

Lemma 9 *Let Assumption 1 hold. For every $h \in (0, h_0]$, the midpoint distribution ν_h satisfies*

$$\int_{\mathbb{R}^d} (1+|x|)^6 \left|(\overline{\mathcal{P}}_h^* \nu_h)(x) - \nu_h(x)\right| \mathrm{d}x \leqslant Ch^2, \tag{C.82}$$

and, for every $K \geqslant 0$ and every $\psi \in C^4(\mathbb{R}^d)$ with $\max_{1\leqslant j\leqslant 4} \|\nabla^j \psi\|_\infty \leqslant K$,

$$\left|\nu_h\big(\overline{\mathcal{P}}_h \psi - \psi\big) - \frac{h^3}{24}\,\pi\big(\nabla^3 \psi : \nabla^3 U\big)\right| \leqslant CKh^4. \tag{C.83}$$

Proof of Lemma 9. To prove (C.82) and (C.83), we write the LM step through its midpoint. Let the initial state $Y_0 \sim \nu_h$ be independent of χ_0, and let $\overline{D}_0$ be the LM step of (C.2). The step from Y_0 to $Y_0 + \overline{D}_0$ has the midpoint $Y_0 + \frac{1}{2}\overline{D}_0$. The pair $(Y_0, Y_0 + \overline{D}_0)$ has the density $\nu_h(y)\,\overline{\mathcal{P}}_h^*(y, y')$ at (y, y'). By (C.12) and (C.72), the change of variables $(y, y') = (x - \eta, x + \eta)$, with the Jacobian determinant 2^d, shows that the pair $(Y_0 + \frac{1}{2}\overline{D}_0, \frac{1}{2}\overline{D}_0)$ has the density

$$\mathfrak{p}(x, \eta) := \Lambda_h(x)\,\big(\nu_h(x+\eta)\,\nu_h(x-\eta)\big)^{\frac{1}{2}}\, e^{\Xi_h(x,\eta)}\,(\pi h)^{-\frac{d}{2}} e^{-\frac{|\eta|^2}{h}}, \qquad x, \eta \in \mathbb{R}^d, \tag{C.84}$$

where $(\pi h)^{-\frac{d}{2}} e^{-|\eta|^2/h}$ is the density of $\mathcal{N}(0, \frac{h}{2} I_d)$ at η, and the exponent Ξ_h is defined as

$$\Xi_h(x, \eta) := \frac{1}{2}\big(V_h(x+\eta) - V_h(x-\eta)\big) - \eta \cdot \nabla V_h(x) - \frac{1}{2}\big(\Theta_h(x+\eta) - \Theta_h(x-\eta)\big), \qquad x, \eta \in \mathbb{R}^d. \tag{C.85}$$

The pair $(Y_0 + \frac{1}{2}\overline{D}_0, -\frac{1}{2}\overline{D}_0)$ has the density $\mathfrak{p}(x, -\eta)$ at (x, η).

We next bound the exponent Ξ_h and the factors Λ_h and $(\nu_h(x+\eta)\,\nu_h(x-\eta))^{1/2}$ of $\mathfrak{p}$ in (C.84), and the gradient in x of the product of these three functions,

$$\mathfrak{P}_h(x, \eta) := \Lambda_h(x)\,\big(\nu_h(x+\eta)\,\nu_h(x-\eta)\big)^{\frac{1}{2}}\,\Xi_h(x, \eta), \qquad x, \eta \in \mathbb{R}^d, \tag{C.86}$$

which is odd in η since Ξ_h is. The bound (C.15) at $(y, y') = (x - \eta, x + \eta)$, the odd part of Taylor's formula for ∇V_h about x to the order 2, (C.14) and (C.74) give

$$|\Xi_h(x, \eta)| + \big|\nabla_x \Xi_h(x, \eta)\big| \leqslant C\big(|\eta|^3 + h|\eta|\big), \qquad x, \eta \in \mathbb{R}^d. \tag{C.87}$$

By (C.13), (C.9) and (C.14),

$$\Lambda_h(x) \leqslant \Big(\frac{8}{7}\Big)^d, \qquad |\Lambda_h(x) - 1| + |\nabla \log \Lambda_h(x)| \leqslant Ch(1+|x|)^2, \qquad x \in \mathbb{R}^d. \tag{C.88}$$

The bound on ν_h in (C.74) gives

$$\big(\nu_h(x+\eta)\,\nu_h(x-\eta)\big)^{\frac{1}{2}} \leqslant Ce^{-\frac{m}{64}(|x|^2+|\eta|^2)}, \qquad x, \eta \in \mathbb{R}^d. \tag{C.89}$$

By (C.14) and (C.74), $|\nabla \log \nu_h(x)| = |\nabla \Theta_h(x) - \nabla V_h(x)| \leqslant C(1+|x|)$. Then (C.87)–(C.89) give

$$|\nabla_x \mathfrak{P}_h(x, \eta)| \leqslant C(1+|x|+|\eta|)^2\big(|\eta|^3 + h|\eta|\big)e^{-\frac{m}{64}(|x|^2+|\eta|^2)}, \qquad x, \eta \in \mathbb{R}^d. \tag{C.90}$$

To prove (C.82), we write the change of the density with (C.84). The densities of $Y_0 + \overline{D}_0$ and of Y_0 at $z \in \mathbb{R}^d$ are the integrals over η of $\mathfrak{p}(z - \eta, \eta)$ and of $\mathfrak{p}(z - \eta, -\eta)$, so that

$$(\overline{\mathcal{P}}_h^* \nu_h)(z) - \nu_h(z) = \int_{\mathbb{R}^d} \big(\mathfrak{p}(z-\eta, \eta) - \mathfrak{p}(z-\eta, -\eta)\big)\mathrm{d}\eta, \qquad z \in \mathbb{R}^d. \tag{C.91}$$

Since Ξ_h is odd in η and the other factors of $\mathfrak{p}$ are even in η, (C.86) gives

$$\left|\mathfrak{p}(x, \eta) - \mathfrak{p}(x, -\eta) - 2\,\mathfrak{P}_h(x, \eta)\,(\pi h)^{-\frac{d}{2}} e^{-\frac{|\eta|^2}{h}}\right| \leqslant \frac{1}{6}\,|\Xi_h(x, \eta)|^3\big(\mathfrak{p}(x, \eta) + \mathfrak{p}(x, -\eta)\big), \qquad x, \eta \in \mathbb{R}^d. \tag{C.92}$$

In every expectation over η from here on in this proof, $\eta = \sqrt{h/2}\,\chi$ with $\chi \sim \mathcal{N}(0, I_d)$, so that the expectation is the integral against the factor $(\pi h)^{-\frac{d}{2}} e^{-|\eta|^2/h}$ of $\mathfrak{p}$, not against $\mathfrak{p}(x,\eta)$. Then (C.91), (C.92) with $x = z-\eta$, the densities $\mathfrak{p}(x, \pm\eta)$ of the pairs $(Y_0 + \frac{1}{2}\overline{D}_0, \pm\frac{1}{2}\overline{D}_0)$, (C.87), (C.4) and the moments of ν_h from (C.74) give

$$\int_{\mathbb{R}^d} (1+|z|)^6 \left| (\overline{\mathcal{P}}_h^* \nu_h)(z) - \nu_h(z) - 2\,\mathbb{E}\big[\mathfrak{P}_h(z-\eta,\eta)\big] \right| \mathrm{d}z \leqslant C\,\mathbb{E}\Big[\big(1+|Y_0|+|\overline{D}_0|\big)^6 \big(|\overline{D}_0|^3 + h|\overline{D}_0|\big)^3 \Big] \leqslant C h^{\frac{9}{2}}. \tag{C.93}$$

Since $\mathfrak{P}_h$ is odd in η, $\mathbb{E}[\mathfrak{P}_h(z,\eta)] = 0$, so that $\mathbb{E}[\mathfrak{P}_h(z-\eta,\eta)] = \mathbb{E}[\mathfrak{P}_h(z-\eta,\eta) - \mathfrak{P}_h(z,\eta)]$. The identity $\mathfrak{P}_h(z-\eta,\eta) - \mathfrak{P}_h(z,\eta) = -\int_0^1 \nabla_x \mathfrak{P}_h(z - s\eta, \eta)\cdot\eta\,\mathrm{d}s$ and (C.90) give

$$\int_{\mathbb{R}^d} (1+|z|)^6 \left| \mathbb{E}\big[\mathfrak{P}_h(z-\eta,\eta) - \mathfrak{P}_h(z,\eta)\big] \right| \mathrm{d}z \leqslant C\,\mathbb{E}\Big[|\eta|\,(1+|\eta|)^8 \big(|\eta|^3 + h|\eta|\big) \Big] \leqslant C h^2. \tag{C.94}$$

The bounds (C.94) and (C.93) give (C.82).

To prove (C.83), let ψ have bounded derivatives as in Lemma 9. Since $\overline{\mathcal{P}}_h^* \nu_h$ and ν_h are probability densities, $\nu_h(\overline{\mathcal{P}}_h\psi - \psi)$ is the integral of $\psi(z) - \psi(0)$ against $(\overline{\mathcal{P}}_h^*\nu_h)(z) - \nu_h(z)$ in (C.91). Then $|\psi(z) - \psi(0)| \leqslant K|z|$, (C.93), the substitution $x = z - \eta$ and the oddness of $\mathfrak{P}_h$ in η give

$$\Big| \nu_h\big(\overline{\mathcal{P}}_h\psi - \psi\big) - \int_{\mathbb{R}^d} \mathbb{E}\Big[\mathfrak{P}_h(x,\eta)\,\big(\psi(x+\eta) - \psi(x-\eta)\big) \Big] \mathrm{d}x \Big| \leqslant C K h^{\frac{9}{2}}. \tag{C.95}$$

We next expand the integrand of (C.95) about x, which involves $\mathfrak{P}_h(x,\eta)$ itself and the difference $\psi(x+\eta) - \psi(x-\eta)$. For the factor $(\nu_h(x+\eta)\,\nu_h(x-\eta))^{1/2}$ of $\mathfrak{P}_h(x,\eta)$ in (C.86), we apply Taylor's formula for $V_h - \Theta_h$ to the order 3 at $x+\eta$ and $x-\eta$, with (C.14) and (C.74), and then Taylor's formula for the exponential function to the order 1. The remainder of Taylor's formula for the exponential function carries the factor $\max\{\nu_h(x), (\nu_h(x+\eta)\,\nu_h(x-\eta))^{1/2}\} \leqslant C e^{-\frac{m}{64}|x|^2}$ by (C.74) and (C.89), so that, for $x, \eta \in \mathbb{R}^d$,

$$\Big| \big(\nu_h(x+\eta)\,\nu_h(x-\eta)\big)^{\frac{1}{2}} - \nu_h(x) \Big(1 - \frac{1}{2}\,\nabla^2 V_h(x) : \eta^{\otimes 2}\Big) \Big| \leqslant C\big(|\eta|^4 + h|\eta|^2\big)\big(1+|\eta|^4\big) e^{-\frac{m}{64}|x|^2}. \tag{C.96}$$

Taylor's formula about x for V_h to the order 6, for ΔV_h to the order 4 and for $\nabla^2 V_h : \nabla^2 V_h + \frac{1}{2}\,\Delta^2 V_h$ to the order 2 at $x+\eta$ and $x-\eta$ in (C.85), with (C.71), gives

$$\Xi_h(x,\eta) = \Xi_h^{(3)}(x,\eta) + \Xi_h^{(5)}(x,\eta) + \Xi_h^{(7)}(x,\eta), \tag{C.97}$$

where, with all derivatives evaluated at x from here on in this proof, $\Xi_h^{(3)}$ and $\Xi_h^{(5)}$ are defined, and the remainder $\Xi_h^{(7)}$ is bounded by (C.14), as

$$\begin{aligned}
\Xi_h^{(3)}(x,\eta) &:= \frac{1}{6}\,\nabla^3 V_h : \eta^{\otimes 3} - \frac{h}{4}\,\nabla\Delta V_h \cdot \eta, \\
\Xi_h^{(5)}(x,\eta) &:= \frac{1}{120}\,\nabla^5 V_h : \eta^{\otimes 5} - \frac{h}{24}\,\nabla^3\Delta V_h : \eta^{\otimes 3} + \frac{h^2}{16}\,\nabla\Big(\nabla^2 V_h : \nabla^2 V_h + \frac{1}{2}\,\Delta^2 V_h\Big)\cdot\eta, \\
|\Xi_h^{(7)}(x,\eta)| &\leqslant C\big(|\eta|^7 + h|\eta|^5 + h^2|\eta|^3\big).
\end{aligned} \tag{C.98}$$

Taylor's formula for ψ about x to the order 3 gives

$$\psi(x+\eta) - \psi(x-\eta) = 2\,\nabla\psi(x)\cdot\eta + \frac{1}{3}\,\nabla^3\psi(x) : \eta^{\otimes 3} + \varrho_4^{\psi}(x,\eta), \tag{C.99}$$

where, by the integral form of Taylor's remainder and the bound on $\nabla^4\psi$, the remainder ϱ_4^{ψ} satisfies

$$\varrho_4^{\psi}(x,\eta) = \int_0^1 \frac{(1-s)^3}{6}\big(\nabla^4\psi(x+s\eta) - \nabla^4\psi(x-s\eta)\big) : \eta^{\otimes 4}\,\mathrm{d}s, \qquad |\varrho_4^{\psi}(x,\eta)| \leqslant \frac{K}{12}\,|\eta|^4. \tag{C.100}$$

We insert (C.97) and (C.99) into (C.95), and, in the term with the product of $\Xi_h^{(3)} + \Xi_h^{(5)}$ and $2\,\nabla\psi\cdot\eta + \frac{1}{3}\,\nabla^3\psi : \eta^{\otimes 3}$, we replace $(\nu_h(x+\eta)\,\nu_h(x-\eta))^{1/2}$ by $\nu_h(x)(1 - \frac{1}{2}\,\nabla^2 V_h(x) : \eta^{\otimes 2})$ from (C.96). The coefficient

$\frac{h^3}{24}$ in (C.83) comes from this product. Counting the pairings of the factors of χ gives

$$\begin{aligned} 2\,\mathbb{E}\big[\Xi_h^{(3)}\,(\nabla\psi\cdot\eta)\big] = 0, \qquad \frac{1}{3}\,\mathbb{E}\big[\Xi_h^{(3)}\,(\nabla^3\psi:\eta^{\otimes 3})\big] = \frac{h^3}{24}\,\nabla^3\psi:\nabla^3 V_h, \\ 2\,\mathbb{E}\big[\Xi_h^{(5)}\,(\nabla\psi\cdot\eta)\big] = \frac{h^3}{16}\,\nabla\big(\nabla^2 V_h:\nabla^2 V_h\big)\cdot\nabla\psi = \mathbb{E}\big[(\nabla^2 V_h:\eta^{\otimes 2})\,\Xi_h^{(3)}\,(\nabla\psi\cdot\eta)\big]. \end{aligned} \tag{C.101}$$

By (C.101), the terms of order h^2 and h^3 of the product of the Taylor approximations of $\mathfrak{P}_h(x,\eta)/(\Lambda_h(x)\,\nu_h(x))$ and of $\psi(x+\eta)-\psi(x-\eta)$ have the total expectation $\frac{h^3}{24}\,\nabla^3\psi:\nabla^3 V_h$. Elementary inequalities with (C.14), the bound K on the derivatives of ψ and $\mathbb{E}|\eta|^b \leqslant Ch^{b/2}$ absorb all terms of higher order into CKh^4, so that

$$\Big|\mathbb{E}\Big[\Big(1-\frac{1}{2}\,\nabla^2 V_h:\eta^{\otimes 2}\Big)\,(\Xi_h^{(3)}+\Xi_h^{(5)})\,\Big(2\,\nabla\psi\cdot\eta+\frac{1}{3}\,\nabla^3\psi:\eta^{\otimes 3}\Big)\Big]-\frac{h^3}{24}\,\nabla^3\psi:\nabla^3 V_h\Big| \leqslant CKh^4, \quad x\in\mathbb{R}^d. \tag{C.102}$$

It remains to bound the three remainders in the comparison of (C.95) with (C.102): $\mathcal{R}_1$ from the approximation of $(\nu_h(x+\eta)\,\nu_h(x-\eta))^{1/2}$, $\mathcal{R}_2$ from $\Xi_h^{(7)}$, and $\mathcal{R}_3$ from ϱ_4^{ψ}. Inserting these three approximations into $\mathfrak{P}_h(x,\eta)\,(\psi(x+\eta)-\psi(x-\eta))$ gives

$$\begin{aligned} &\mathfrak{P}_h(x,\eta)\,\big(\psi(x+\eta)-\psi(x-\eta)\big)-\mathcal{R}_1(x,\eta)-\mathcal{R}_2(x,\eta)-\mathcal{R}_3(x,\eta) \\ &\quad=\Lambda_h(x)\,\nu_h(x)\Big(1-\frac{1}{2}\,\nabla^2 V_h:\eta^{\otimes 2}\Big)\,(\Xi_h^{(3)}+\Xi_h^{(5)})\,\Big(2\,\nabla\psi\cdot\eta+\frac{1}{3}\,\nabla^3\psi:\eta^{\otimes 3}\Big), \end{aligned} \tag{C.103}$$

where the remainders are defined as

$$\begin{aligned} \mathcal{R}_1(x,\eta) &:= \Lambda_h(x)\Big(\big(\nu_h(x+\eta)\,\nu_h(x-\eta)\big)^{\frac{1}{2}}-\nu_h(x)\Big(1-\frac{1}{2}\,\nabla^2 V_h:\eta^{\otimes 2}\Big)\Big) \\ &\qquad\times(\Xi_h^{(3)}+\Xi_h^{(5)})\,\Big(2\,\nabla\psi\cdot\eta+\frac{1}{3}\,\nabla^3\psi:\eta^{\otimes 3}\Big), \\ \mathcal{R}_2(x,\eta) &:= \Lambda_h(x)\big(\nu_h(x+\eta)\,\nu_h(x-\eta)\big)^{\frac{1}{2}}\,\Xi_h^{(7)}(x,\eta)\,\Big(2\,\nabla\psi\cdot\eta+\frac{1}{3}\,\nabla^3\psi:\eta^{\otimes 3}\Big), \\ \mathcal{R}_3(x,\eta) &:= \mathfrak{P}_h(x,\eta)\,\varrho_4^{\psi}(x,\eta). \end{aligned} \tag{C.104}$$

For $\mathcal{R}_1$ and $\mathcal{R}_2$, (C.88), (C.98) with (C.14), the bound K on the derivatives of ψ, (C.96) for $\mathcal{R}_1$ and (C.89) for $\mathcal{R}_2$ give, for $x,\eta\in\mathbb{R}^d$,

$$|\mathcal{R}_1(x,\eta)| \leqslant CK(|\eta|^2+h)^4(1+|\eta|^8)\,e^{-\frac{m}{64}|x|^2}, \qquad |\mathcal{R}_2(x,\eta)| \leqslant CK(|\eta|^2+h)^4(1+|\eta|^2)\,e^{-\frac{m}{64}|x|^2}. \tag{C.105}$$

Since $\mathbb{E}[(|\eta|^2+h)^4(1+|\eta|^8)] \leqslant Ch^4$, (C.105) gives $\mathbb{E}|\mathcal{R}_1(x,\eta)|+\mathbb{E}|\mathcal{R}_2(x,\eta)| \leqslant CKh^4 e^{-\frac{m}{64}|x|^2}$. For $\mathcal{R}_3$, (C.87) and (C.100) give only $|\int_{\mathbb{R}^d}\mathbb{E}[\mathcal{R}_3(x,\eta)]\,\mathrm{d}x| \leqslant CKh^{7/2}$. By (C.104),

$$\int_{\mathbb{R}^d}\mathbb{E}\big[\mathcal{R}_3(x,\eta)\big]\,\mathrm{d}x=\int_0^1\frac{(1-s)^3}{6}\,\mathbb{E}\Big[\int_{\mathbb{R}^d}\mathfrak{P}_h(x,\eta)\big(\nabla^4\psi(x+s\eta)-\nabla^4\psi(x-s\eta)\big):\eta^{\otimes 4}\,\mathrm{d}x\Big]\mathrm{d}s,$$

and, for $s\in[0,1]$, the substitutions $x=z\mp s\eta$ move the difference of $\nabla^4\psi$ to $\mathfrak{P}_h$. The identity $\mathfrak{P}_h(z-s\eta,\eta)-\mathfrak{P}_h(z+s\eta,\eta)=-\int_{-s}^{s}\nabla_x\mathfrak{P}_h(z+r\eta,\eta)\cdot\eta\,\mathrm{d}r$ and (C.90) then give

$$\Big|\int_{\mathbb{R}^d}\mathfrak{P}_h(x,\eta)\big(\nabla^4\psi(x+s\eta)-\nabla^4\psi(x-s\eta)\big):\eta^{\otimes 4}\,\mathrm{d}x\Big| \leqslant CK|\eta|^5(1+|\eta|)^2\big(|\eta|^3+h|\eta|\big),$$

and $\mathbb{E}\big[|\eta|^5(1+|\eta|)^2(|\eta|^3+h|\eta|)\big] \leqslant Ch^4$. Hence

$$\Big|\int_{\mathbb{R}^d}\mathbb{E}\big[\mathcal{R}_1(x,\eta)\big]\mathrm{d}x\Big|+\Big|\int_{\mathbb{R}^d}\mathbb{E}\big[\mathcal{R}_2(x,\eta)\big]\mathrm{d}x\Big|+\Big|\int_{\mathbb{R}^d}\mathbb{E}\big[\mathcal{R}_3(x,\eta)\big]\mathrm{d}x\Big| \leqslant CKh^4. \tag{C.106}$$

Combining (C.95), (C.103), (C.102) and the remainder estimates (C.106) gives

$$\Big|\nu_h\big(\overline{\mathcal{P}}_h\psi-\psi\big)-\frac{h^3}{24}\int_{\mathbb{R}^d}\Lambda_h(x)\,\nu_h(x)\,\nabla^3\psi(x):\nabla^3 V_h(x)\,\mathrm{d}x\Big| \leqslant CKh^4. \tag{C.107}$$

Finally, we pass from (C.107) to (C.83). By (C.88), $\int_{\mathbb{R}^d} |\nu_h(x) - \pi(x)|\,\mathrm{d}x \leqslant Ch$ from (C.79), and $\|\nabla^3 V_h(x) - \nabla^3 U(x)\| \leqslant Ch(1+|x|)$ from (C.14),

$$\Big| \int_{\mathbb{R}^d} \Lambda_h(x)\, \nu_h(x)\, \nabla^3\psi(x) : \nabla^3 V_h(x)\,\mathrm{d}x - \pi\big(\nabla^3\psi : \nabla^3 U\big) \Big| \leqslant CKh.$$

This bound and (C.107) give (C.83). □

C.5 Proof of Theorem 3

Proof of Theorem 3. Let $h \in (0, h_0]$ and $P \geqslant 2$. Lemma 1 gives the unique invariant distribution μ_h of NC-SGLM, with $\mu_h(|y|^{12}) \leqslant C$. By (3.1) and (A.14), $\pi_h(f) = \mu_h(\tilde{f}_h)$ and $\overline{\pi}_h(f) = \overline{\mu}_h(\tilde{f}_h)$.

We split the bias $\pi_h(f) - \pi(f)$ into the bias $\overline{\pi}_h(f) - \pi(f)$ of the LM discretization and the error $\pi_h(f) - \overline{\pi}_h(f)$ of the random batch, and show

$$\Big| \overline{\pi}_h(f) - \pi(f) - \frac{h^2}{24}\,\pi\big(\nabla^3\phi_0 : \nabla^3 U\big) \Big| \leqslant CLh^3 \tag{C.108}$$

and

$$\big| \pi_h(f) - \overline{\pi}_h(f) \big| \leqslant CL\,\frac{h^2}{P}, \tag{C.109}$$

which together give (3.8). Both proofs use a solution of the discrete Poisson equation (C.22) of LM, constructed with Lemma 5.

We first solve the discrete Poisson equation of LM with the right side $h(\tilde{f}_h - \overline{\mu}_h(\tilde{f}_h))$, starting from the modified discrete Poisson solution $\tilde{\phi}_h$ of (A.21). Let

$$\tilde{F}^\star := \tilde{\phi}_h - \overline{\mathcal{P}}_h\tilde{\phi}_h - h\big(\tilde{f}_h - \tilde{\pi}_h(\tilde{f}_h)\big). \tag{C.110}$$

For the initial state $Y_0 = y$, (A.21), (C.2) and (A.20) give

$$\tilde{F}^\star(y) = \big(e^{h\tilde{\mathcal{L}}_h}\tilde{\phi}_h\big)(y) - (\overline{\mathcal{P}}_h\tilde{\phi}_h)(y) = \mathbb{E}\big[\tilde{\phi}_h(\tilde{Y}_0(h)) - \tilde{\phi}_h(Y_0 + \overline{D}_0)\big].$$

Hence $\tilde{F}^\star(y)$ is the local weak error of one step of LM against the modified dynamics. By (A.22), $\tilde{\phi}_h$ satisfies the hypothesis of Lemma 2. This lemma with $g = \nabla U$, $G = \tilde{\phi}_h$ and $k = 0$ gives, for the norm $\|\cdot\|_{\mathrm{w}}$ of (C.23),

$$|\tilde{F}^\star(y)| \leqslant CLh^3(1+|y|)^6, \qquad \|\tilde{F}^\star\|_{\mathrm{w}} \leqslant CLh^3, \qquad y \in \mathbb{R}^d. \tag{C.111}$$

By (C.111), Lemma 5 with $F = \tilde{F}^\star$ gives a solution $\tilde{\Phi}_h^\star$ with

$$|\tilde{\Phi}_h^\star(y)| \leqslant CLh^2(1+|y|^6), \qquad \tilde{\Phi}_h^\star - \overline{\mathcal{P}}_h\tilde{\Phi}_h^\star = \tilde{F}^\star - \overline{\mu}_h(\tilde{F}^\star), \qquad y \in \mathbb{R}^d. \tag{C.112}$$

Integrating (C.110) against the invariant distribution $\overline{\mu}_h$ of LM gives

$$\overline{\mu}_h(\tilde{F}^\star) - h\,\tilde{\pi}_h(\tilde{f}_h) = -h\,\overline{\mu}_h(\tilde{f}_h), \tag{C.113}$$

so that (C.112) and (C.113) give

$$(\tilde{\phi}_h - \tilde{\Phi}_h^\star) - \overline{\mathcal{P}}_h(\tilde{\phi}_h - \tilde{\Phi}_h^\star) = h\big(\tilde{f}_h - \overline{\mu}_h(\tilde{f}_h)\big). \tag{C.114}$$

We then compare $\tilde{\phi}_h$ with the solution ϕ_0 of (2.11). By [30, Section 3.3], ϕ_0 is unique up to an additive constant among the functions with a bounded gradient and satisfies

$$\max_{1\leqslant j\leqslant 4} \|\nabla^j\phi_0\|_\infty \leqslant CL. \tag{C.115}$$

For the initial state $Y_0 = y$, (C.57) gives $\mathbb{E}[\chi_0\,\nabla U(Z_0)^\top] = \sqrt{h/2}\,\mathbb{E}[\nabla^2 U(Z_0)]$. This identity, the symmetry of χ_0, Taylor's formula for ∇U about y to the order 1 and (C.4) give, for the LM step $\overline{D}_0$,

$$\begin{gathered} \big|\mathbb{E}[\overline{D}_0] + h\,\nabla U(y)\big| \leqslant Ch^2, \qquad \big\|\mathbb{E}\big[(\overline{D}_0)^{\otimes 2}\big] - 2hI_d\big\| \leqslant Ch^2(1+|y|)^2, \\ \big\|\mathbb{E}\big[(\overline{D}_0)^{\otimes 3}\big]\big\| \leqslant Ch^2(1+|y|)^3, \qquad \mathbb{E}|\overline{D}_0|^4 \leqslant Ch^2(1+|y|)^4. \end{gathered} \tag{C.116}$$

By (C.115) and (C.116), Taylor's formula for ϕ_0 about y to the order 3 gives

$$\left|(\overline{\mathcal{P}}_h\phi_0)(y) - \phi_0(y) - h\,(\mathcal{L}\phi_0)(y)\right| \leqslant CLh^2(1+|y|)^4. \tag{C.117}$$

By (C.110) and the Poisson equation (2.11),

$$(\tilde{\phi}_h - \phi_0) - \overline{\mathcal{P}}_h(\tilde{\phi}_h - \phi_0) - h\big(\pi(f) - \tilde{\pi}_h(\tilde{f}_h)\big) = \tilde{F}^\star + \big(\overline{\mathcal{P}}_h\phi_0 - \phi_0 - h\,\mathcal{L}\phi_0\big) + h(\tilde{f}_h - f). \tag{C.118}$$

By (C.111), (C.117) and (3.3), the right side F of (C.118) satisfies $\|F\|_{\mathrm{w}} \leqslant CLh^2$. Lemma 5 with this F gives a solution $\Phi_h^\star$ with

$$|\Phi_h^\star(y)| \leqslant CLh(1+|y|^6), \qquad \Phi_h^\star - \overline{\mathcal{P}}_h\Phi_h^\star = F - \overline{\mu}_h(F), \qquad y \in \mathbb{R}^d. \tag{C.119}$$

Integrating (C.118) against the invariant distribution $\overline{\mu}_h$ shows that the constant $h(\pi(f) - \tilde{\pi}_h(\tilde{f}_h))$ equals $-\overline{\mu}_h(F)$. Hence $\tilde{\phi}_h - \phi_0 - \Phi_h^\star$ is invariant under $\overline{\mathcal{P}}_h$. Since $\tilde{\phi}_h$ and ϕ_0 have bounded gradients, (C.119) gives $\|\tilde{\phi}_h - \phi_0 - \Phi_h^\star\|_{\mathrm{w}} < \infty$, and (C.24) as $n \to \infty$ shows that $\tilde{\phi}_h - \phi_0 - \Phi_h^\star$ is a constant $c_h \in \mathbb{R}$, that is,

$$\tilde{\phi}_h - \phi_0 = \Phi_h^\star + c_h. \tag{C.120}$$

The specific value of c_h does not affect the derivatives despite the dependence of h.

To prove (C.108), we compare $\overline{\mu}_h(\tilde{f}_h)$ with $\nu_h(\tilde{f}_h)$ by Lemma 9 and $\nu_h(\tilde{f}_h)$ with $\pi(f)$ by Lemma 8, and then replace $\tilde{\phi}_h$ by ϕ_0 in the resulting coefficient of h^2. Integrating (C.114) against ν_h gives

$$\overline{\mu}_h(\tilde{f}_h) - \nu_h(\tilde{f}_h) = \frac{1}{h}\,\nu_h\big(\overline{\mathcal{P}}_h\tilde{\phi}_h - \tilde{\phi}_h\big) - \frac{1}{h}\int_{\mathbb{R}^d}\big((\overline{\mathcal{P}}_h^*\nu_h)(x) - \nu_h(x)\big)\,\tilde{\Phi}_h^\star(x)\,\mathrm{d}x. \tag{C.121}$$

Since $\nu_h(\tilde{f}_h) = (\nu_h * \mathcal{N}(0, \frac{h}{2}I_d))(f)$ and $|f(x) - f(0)| \leqslant L|x|$, Lemma 8 gives $|\nu_h(\tilde{f}_h) - \pi(f)| \leqslant CLh^3$. By (A.22), $\tilde{\phi}_h$ satisfies the hypothesis of Lemma 9 with $K = CL$. The lemma with $\psi = \tilde{\phi}_h$, the bound on $\tilde{\Phi}_h^\star$ in (C.112) and the bound on $\nu_h(\tilde{f}_h) - \pi(f)$ turn (C.121) into

$$\Big|\overline{\pi}_h(f) - \pi(f) - \frac{h^2}{24}\,\pi\big(\nabla^3\tilde{\phi}_h : \nabla^3 U\big)\Big| \leqslant CLh^3. \tag{C.122}$$

To replace $\tilde{\phi}_h$ by ϕ_0 in the coefficient $\pi(\nabla^3\tilde{\phi}_h : \nabla^3 U)$, we write $\pi(\nabla^3\psi : \nabla^3 U)$ as $\pi(W_3\,\psi)$ for $\psi = \tilde{\phi}_h$ and $\psi = \phi_0$, where W_3 is defined as

$$W_3 := -e^U \sum_{i,j,k=1}^{d} \partial_i\partial_j\partial_k\big(e^{-U}\,\partial_i\partial_j\partial_k U\big). \tag{C.123}$$

Since $e^U\partial_i(e^{-U}\,\cdot\,) = \partial_i - \partial_i U$, the function W_3 in (C.123) is a polynomial of degree at most 3 in ∇U with coefficients formed by the derivatives of U of orders 2 to 6. Hence (3.2) gives

$$|W_3(x)| \leqslant C(1+|x|)^3, \qquad x \in \mathbb{R}^d. \tag{C.124}$$

Let $\psi \in C^3(\mathbb{R}^d)$ have derivatives of orders 0 to 3 that grow at most polynomially. Since $\pi(x) \leqslant Ce^{-m|x|^2/8}$ [30, Lemma 17], integrating by parts 3 times in $\int_{\mathbb{R}^d}\psi\,W_3\,e^{-U}\,\mathrm{d}x$ gives

$$\pi(W_3\,\psi) = \pi\big(\nabla^3\psi : \nabla^3 U\big). \tag{C.125}$$

In particular, $\pi(W_3) = 0$. Then (C.120), $\pi(W_3) = 0$ and (C.125) with $\psi = \tilde{\phi}_h$ and $\psi = \phi_0$ give

$$\pi\big(\nabla^3\tilde{\phi}_h : \nabla^3 U\big) - \pi\big(\nabla^3\phi_0 : \nabla^3 U\big) = \pi\big(W_3\,(\tilde{\phi}_h - \phi_0)\big) = \pi\big(W_3\,\Phi_h^\star\big). \tag{C.126}$$

By (C.119) and (C.124), $|\pi(W_3\,\Phi_h^\star)| \leqslant CLh$. Hence (C.126) and (C.122) give (C.108).

To prove (C.109), we first check the hypothesis of Lemma 7 with $K = CL$ for the solution $\tilde{\phi}_h - \tilde{\Phi}_h^\star$ of (C.114). By (3.2), the state $y + \overline{D}_0$ after the LM step from the initial state $Y_0 = y$ satisfies $\|\nabla_y^j(y + \overline{D}_0)\| \leqslant C$ for $j = 1, 2, 3$. Then (A.22), (3.3) and the chain rule under the expectation in (C.2) give

$$\|\nabla^j(\overline{\mathcal{P}}_h\tilde{\phi}_h)\|_\infty + \|\nabla^j\tilde{F}^\star\|_\infty \leqslant CL, \qquad j = 2, 3. \tag{C.127}$$

By (C.112), $\tilde{\Phi}_h^\star = \tilde{F}^\star - \overline{\mu}_h(\tilde{F}^\star) + \overline{\mathcal{P}}_h\tilde{\Phi}_h^\star$. Hence (C.127) and Lemma 6 with $F = \tilde{\Phi}_h^\star$ give, for $j = 2, 3$,

$$\left\|\nabla^j(\tilde{\phi}_h - \tilde{\Phi}_h^\star)(y)\right\| \leqslant \|\nabla^j\tilde{\phi}_h\|_\infty + \|\nabla^j\tilde{F}^\star\|_\infty + \|\nabla^j(\overline{\mathcal{P}}_h\tilde{\Phi}_h^\star)(y)\| \leqslant CL + CLh^{2-\frac{j}{2}}(1+|y|)^6 \leqslant CL(1+|y|)^6.$$

Integrating (C.114) against the invariant distribution μ_h of NC-SGLM and applying Lemma 7 with $G = \tilde{\phi}_h - \tilde{\Phi}_h^\star$ give

$$h\left|\mu_h(\tilde{f}_h) - \overline{\mu}_h(\tilde{f}_h)\right| = \left|\mu_h\Big(\mathcal{P}_h(\tilde{\phi}_h - \tilde{\Phi}_h^\star) - \overline{\mathcal{P}}_h(\tilde{\phi}_h - \tilde{\Phi}_h^\star)\Big)\right| \leqslant CL\,\frac{h^3}{P}\,\mu_h\big((1+|y|)^{10}\big) \leqslant CL\,\frac{h^3}{P}.$$

Dividing by h gives (C.109). □

D Proof of Theorem 4: capping probability

The proof bounds the largest eigenvalue of $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$ by a sum of P independent terms, bounds the probability of (3.14) at a fixed x through the moments of this sum, and integrates the result against π. The same argument with the P_0 group means in place of the P stochastic gradients gives (3.16).

Proof of Theorem 4. Fix $x \in \mathbb{R}^d$. Since $\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})$ is positive semidefinite and $\hat{g}_P(x;\boldsymbol{\xi})$ minimizes $v \mapsto \sum_{p=1}^P |g(x;\xi^{(p)}) - v|^2$, (2.5) gives

$$\lambda_{\max}\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) \leqslant \frac{1}{P(P-1)}\sum_{p=1}^P |g(x;\xi^{(p)}) - \nabla U(x)|^2, \tag{D.1}$$

where the P stochastic gradients of the sum are independent, each with mean $\mathrm{Tr}[\mathcal{C}(x)] \leqslant M^2(1+|x|)^2$.

For (3.15), let first $\mathrm{Tr}[\mathcal{C}(x)] \leqslant (P-1)/h$. Then, by (D.1), $\lambda_{\max}(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})) > \frac{2}{h}$ implies

$$\sum_{p=1}^P\Big(|g(x;\xi^{(p)}) - \nabla U(x)|^2 - \mathrm{Tr}[\mathcal{C}(x)]\Big) > \frac{2P(P-1)}{h} - P\,\mathrm{Tr}[\mathcal{C}(x)] \geqslant \frac{P(P-1)}{h}.$$

The P terms of this sum are independent with mean 0, and by (3.4) and (3.2) each has a moment of order 12 at most $C(1+|x|)^{24}$, so Markov's inequality with the moment of order 12 of the sum gives

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\lambda_{\max}\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) > \frac{2}{h}\right) \leqslant \left(\frac{h}{P(P-1)}\right)^{12} CP^6(1+|x|)^{24} = C\left(\frac{h}{\sqrt{P}\,(P-1)}\right)^{12}(1+|x|)^{24}.$$

If instead $\mathrm{Tr}[\mathcal{C}(x)] > (P-1)/h$, then $M^2h(1+|x|)^2/(P-1) > 1$, so that

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\lambda_{\max}\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) > \frac{2}{h}\right) \leqslant \left(\frac{M^2h(1+|x|)^2}{P-1}\right)^{18} \leqslant C\left(\frac{h}{\sqrt{P}\,(P-1)}\right)^{12}(1+|x|)^{36}.$$

In both cases,

$$\mathbb{P}_{\boldsymbol{\xi}}\left(\lambda_{\max}\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) > \frac{2}{h}\right) \leqslant C\left(\frac{h}{\sqrt{P}\,(P-1)}\right)^{12}(1+|x|)^{36},$$

and integrating against π, whose moments are bounded [30, Lemma 17], gives (3.15).

For (3.16), the P_0 group means $\hat{g}_{P/P_0}(x;\boldsymbol{\xi}^{[j]})$ of (2.16) replace the P stochastic gradients. Since $\hat{g}_P(x;\boldsymbol{\xi})$ is the mean of the group means, the argument for (D.1) shows

$$\lambda_{\max}\big(\hat{\mathcal{C}}_{P,P_0}(x;\boldsymbol{\xi})\big) \leqslant \frac{1}{P_0(P_0-1)}\sum_{j=1}^{P_0}\big|\hat{g}_{P/P_0}(x;\boldsymbol{\xi}^{[j]}) - \nabla U(x)\big|^2,$$

where the P_0 terms of the sum are independent, each with mean $\frac{P_0}{P}\operatorname{Tr}[\mathcal{C}(x)]$ by (2.4) with P/P_0 in place of P. Each $\hat{g}_{P/P_0}(x;\boldsymbol{\xi}^{[j]}) - \nabla U(x)$ is the mean of P/P_0 independent vectors with mean 0 and moment of order 24 at most $C(1+|x|)^{24}$, so that

$$\mathbb{E}_{\boldsymbol{\xi}}\Big[\big|\hat{g}_{P/P_0}(x;\boldsymbol{\xi}^{[j]}) - \nabla U(x)\big|^{24}\Big] \leqslant C\Big(\frac{P_0}{P}\Big)^{12}(1+|x|)^{24}.$$

Let first $\frac{P_0}{P}\operatorname{Tr}[\mathcal{C}(x)] \leqslant (P_0-1)/h$. The argument for (3.15), with this moment bound, gives

$$\mathbb{P}_{\boldsymbol{\xi}}\Big(\lambda_{\max}\big(\hat{\mathcal{C}}_{P,P_0}(x;\boldsymbol{\xi})\big) > \frac{2}{h}\Big) \leqslant \Big(\frac{h}{P_0(P_0-1)}\Big)^{12} C P_0^6 \Big(\frac{P_0}{P}\Big)^{12}(1+|x|)^{24} = C\Big(\frac{h\sqrt{P_0}}{P\,(P_0-1)}\Big)^{12}(1+|x|)^{24}.$$

If instead $\frac{P_0}{P}\operatorname{Tr}[\mathcal{C}(x)] > (P_0-1)/h$, then $2M^2h(1+|x|)^2/P > 1$, so that

$$\mathbb{P}_{\boldsymbol{\xi}}\Big(\lambda_{\max}\big(\hat{\mathcal{C}}_{P,P_0}(x;\boldsymbol{\xi})\big) > \frac{2}{h}\Big) \leqslant \Big(\frac{2M^2h(1+|x|)^2}{P}\Big)^{18} \leqslant C\Big(\frac{h\sqrt{P_0}}{P\,(P_0-1)}\Big)^{12}(1+|x|)^{36}.$$

Integrating against π, as for (3.15), gives (3.16). □

E Proof of Theorem 5: explicit coefficient under Gaussian noise

The proof of Theorem 5 refines the proof of Theorem 3 under Assumption 4. The bound (C.108) already gives the LM coefficient $\mathsf{K}^{\mathrm{LM}}(f)$, and it remains to refine (C.109) to $|\pi_h(f) - \pi_h^{\mathrm{LM}}(f) - \frac{h^2}{P}\mathsf{K}^{\mathrm{SG}}(f)| \leqslant CLh^{5/2}/P$. Appendix E.1 checks that Theorem 3, its proof and the lemmas of Appendices A and C apply under Assumption 4 and extends the semigroup estimates of Lemma 6 to the order 4, Appendix E.2 expands one step of NC-SGLM against one step of LM to the order h^3, and Appendix E.3 compares μ_h with π and combines these results. Throughout this appendix, the noise corrector $\mathcal{S}_h$ also satisfies the additional polynomial bound (3.19), and h_0 satisfies the conditions of Appendix C.

E.1 Setting under Gaussian noise

This subsection writes the batch gradient and the batch approximation in terms of the mean of the batch and a matrix independent of this mean, bounds the moments of the mean, of the batch approximation and of the difference between the noise factor and I_d, bounds the derivatives of the covariance $\mathcal{C}$, checks that Assumption 3 holds under Assumption 4, and extends the semigroup estimates of Lemma 6 to the order 4.

Under Assumption 4, the batch gradient (2.2) and the batch approximation (2.5) depend on a batch $\boldsymbol{\xi} = (\xi^{(1)},\dots,\xi^{(P)})$ of P independent vectors drawn from $\mathcal{N}(0,I_d)$ through its mean $\hat{\xi}$ and the matrix $\hat{\Sigma}$,

$$\hat{\xi} := \frac{1}{P}\sum_{p=1}^{P}\xi^{(p)}, \qquad \hat{\Sigma} := \frac{1}{P-1}\sum_{p=1}^{P}\big(\xi^{(p)}-\hat{\xi}\big)^{\otimes 2}, \tag{E.1}$$

since (3.20) gives

$$\hat{g}_P(x;\boldsymbol{\xi}) = \nabla U(x) + \eta(x)\,\hat{\xi}, \qquad \hat{\mathcal{C}}_P(x;\boldsymbol{\xi}) = \frac{1}{P}\,\eta(x)\,\hat{\Sigma}\,\eta(x)^\top, \qquad x\in\mathbb{R}^d. \tag{E.2}$$

The vectors $\hat{\xi}$ and $\xi^{(p)}-\hat{\xi}$, $p=1,\dots,P$, form a Gaussian vector, and $\mathbb{E}[\hat{\xi}\,(\xi^{(p)}-\hat{\xi})^\top] = 0$ for every p, so $\hat{\xi}$ and $\hat{\Sigma}$ are independent. Since $\hat{\xi}\sim\mathcal{N}(0,\frac{1}{P}I_d)$,

$$\mathbb{E}\big[\hat{\xi}\big] = 0, \qquad \mathbb{E}\big[\hat{\xi}^{\otimes 3}\big] = 0, \qquad \mathbb{E}\big[|\hat{\xi}|^q\big] \leqslant \frac{C}{P^{\frac{q}{2}}}, \quad 1\leqslant q\leqslant 12. \tag{E.3}$$

Using $\hat{\Sigma}$ is positive semidefinite and $\sum_{p=1}^P|\xi^{(p)}-\hat{\xi}|^2 \leqslant \sum_{p=1}^P|\xi^{(p)}|^2$, Jensen's inequality and $P\leqslant 2(P-1)$ give $\mathbb{E}[\|\hat{\Sigma}\|^r] \leqslant \mathbb{E}[\operatorname{Tr}[\hat{\Sigma}]^r] \leqslant 2^r\,\mathbb{E}[|\xi^{(1)}|^{2r}] \leqslant C$ for $1\leqslant r\leqslant 12$. Hence (E.2), (3.21) and (A.10) give, for every $x\in\mathbb{R}^d$ and $1\leqslant r\leqslant 12$,

$$\mathbb{E}\big[\|\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\|^r\big] + h^{-r}\,\mathbb{E}\Big[\Big\|\mathcal{S}_h\big(\hat{\mathcal{C}}_P(x;\boldsymbol{\xi})\big) - I_d\Big\|^r\Big] \leqslant \frac{C}{P^r}. \tag{E.4}$$

Since $\mathcal{C} = \eta\,\eta^\top$ under Assumption 4, the Leibniz rule gives

$$\max_{0\leqslant i\leqslant 4} \|\nabla^i \mathcal{C}(x)\| \leqslant 16M^2, \qquad x \in \mathbb{R}^d. \tag{E.5}$$

Assumptions 1 and 3 hold, with M replaced by a larger constant depending only on M and d: by (3.2) and (3.21), $|g(x;\xi)| \leqslant M(1+|x|) + M|\xi|$ and $\|\nabla g(x;\xi)\| \leqslant M + M|\xi|$, the moments of $|\xi|$ of orders up to 24 depend only on d, and $\|\Delta\mathcal{C}(x)\| \leqslant 16dM^2$ by (E.5). Hence Theorem 3, its proof and the lemmas of Appendices A and C apply under Assumptions 1, 2 and 4.

Lemma 10 extends Lemma 6 to the order 4. Appendix E.3 applies this extension with $F = \tilde{\Phi}_h^\star$ to bound the derivative of order 4 of the solution $\tilde{\phi}_h - \tilde{\Phi}_h^\star$ of (C.114).

Lemma 10 *Let Assumption 1 hold, $h \in (0, h_0]$, and let F be measurable with $\|F\|_{\mathrm{w}} < \infty$. Then $\overline{\mathcal{P}}_h F \in C^4(\mathbb{R}^d)$ and*

$$\big\|\nabla^4(\overline{\mathcal{P}}_h F)(y)\big\| \leqslant Ch^{-2}\,\|F\|_{\mathrm{w}}(1+|y|)^6, \qquad y \in \mathbb{R}^d. \tag{E.6}$$

Proof of Lemma 10. To prove (E.6), we follow the proof of Lemma 6 with $j = 4$. By (C.17) with $i = 5$, the function $\mathcal{Q}_h(y)$ of (C.43) also satisfies $\|\nabla^5 \mathcal{Q}_h(y)\| \leqslant Ch$. Hence (C.45) holds for $j = 4$, and the Leibniz rule with (C.44) gives (C.47) for $i = 4$. Then (C.46) and (C.48) hold for $j = 4$, and the rest of the proof of Lemma 6, with the factor $(1+|\chi_0|)^4$ in place of $(1+|\chi_0|)^3$, gives $\overline{\mathcal{P}}_h F \in C^4(\mathbb{R}^d)$ and (E.6). □

E.2 Expansion of one step of NC-SGLM against one step of LM

Lemma 11 below identifies the term of order h^3/P of the difference $\mathcal{P}_h G - \overline{\mathcal{P}}_h G$ bounded in Lemma 7 as $\frac{h^3}{P}\,\mathcal{A}G$, where the operator $\mathcal{A}$ is defined, for $G \in C^3(\mathbb{R}^d)$, as

$$\mathcal{A}G := \frac{1}{8}\,\Delta\mathcal{C} : \nabla^2 G + \frac{1}{2}\,\nabla\mathcal{C} : \nabla^3 G + \frac{1}{4}\,\nabla^2 G : \big(\mathcal{C}\,\nabla^2 U\big), \tag{E.7}$$

where $\nabla\mathcal{C} : \nabla^3 G = \sum_{i,j,k=1}^d \partial_k \mathcal{C}_{ij}\,\partial_i\partial_j\partial_k G$. By (E.5) and $\|\nabla^2 U\| \leqslant M$,

$$|(\mathcal{A}G)(y)| \leqslant C\big(\|\nabla^2 G(y)\| + \|\nabla^3 G(y)\|\big), \qquad y \in \mathbb{R}^d. \tag{E.8}$$

Lemma 11 *Let Assumptions 1 and 4 hold, $h \in (0, h_0]$ and $P \geqslant 2$, and let $K \geqslant 0$ and $G \in C^4(\mathbb{R}^d)$ satisfy $\|\nabla^2 G(x)\| + \|\nabla^3 G(x)\| + \|\nabla^4 G(x)\| \leqslant K(1+|x|)^6$ for $x \in \mathbb{R}^d$. Then the one-step operators $\mathcal{P}_h$ and $\overline{\mathcal{P}}_h$ of (C.2) and the operator $\mathcal{A}$ of (E.7) satisfy*

$$\Big|(\mathcal{P}_h G)(y) - (\overline{\mathcal{P}}_h G)(y) - \frac{h^3}{P}\,(\mathcal{A}G)(y)\Big| \leqslant CK\,\frac{h^{\frac{7}{2}}}{P}\,(1+|y|)^8, \qquad y \in \mathbb{R}^d. \tag{E.9}$$

Proof of Lemma 11. Fix the initial state $Y_0 = y$, and let $\hat{\xi}$ and $\hat{\Sigma}$ be given by (E.1), so that χ_0, $\hat{\xi}$ and $\hat{\Sigma}$ are independent. By (C.50) and (E.2), the state after one step of NC-SGLM is $Y_1 = Y_0 + \overline{D}_0 + \hat{\zeta}_0$ with

$$\hat{\zeta}_0 = -h\,\eta(Z_0)\,\hat{\xi} + \zeta_0, \tag{E.10}$$

where ζ_0 is the noise correction of (A.11). By (3.21), (E.3) and (E.4), for $1 \leqslant q \leqslant 12$,

$$\mathbb{E}\big[|h\,\eta(Z_0)\,\hat{\xi}|^q\big] \leqslant C\,\frac{h^q}{P^{\frac{q}{2}}}, \qquad \mathbb{E}\big[|\zeta_0|^q\big] \leqslant C\,\frac{h^{\frac{3q}{2}}}{P^q}, \tag{E.11}$$

and, by (C.4),

$$|\overline{D}_0| + |Z_0 - Y_0| \leqslant Ch^{\frac{1}{2}}(1+|Y_0|)(1+|\chi_0|). \tag{E.12}$$

Taylor's formula for G about $Y_0 + \overline{D}_0$ to the order 3 gives

$$G(Y_1) - G(Y_0 + \overline{D}_0) = \sum_{i=1}^3 \frac{1}{i!}\,\nabla^i G(Y_0 + \overline{D}_0) : \hat{\zeta}_0^{\otimes i} + \varrho_4^{Y_1}, \qquad |\varrho_4^{Y_1}| \leqslant \frac{K}{24}\big(1 + |Y_0 + \overline{D}_0| + |\hat{\zeta}_0|\big)^6\,|\hat{\zeta}_0|^4. \tag{E.13}$$

(E.11), (E.12) and Hölder's inequality give $\mathbb{E}|\varrho_4^{Y_1}| \leqslant CKh^4P^{-2}(1+|Y_0|)^6$. Inserting (E.10) into the terms of orders 1 to 3 of (E.13) gives products of degrees 0 to 3 in $\hat{\xi}$, and the products of degrees 1 and 3 have mean 0 by (E.3). The product of degree 2 from the term with $\nabla^2 G$ is $\mathcal{T}_1$ of (C.61), so that (C.62) gives

$$\frac{1}{2}\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0) : \big(h\,\eta(Z_0)\,\hat{\xi}\big)^{\otimes 2}\big] = \mathbb{E}[\mathcal{T}_1] = \frac{h^2}{2P}\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0) : \mathcal{C}(Z_0)\big]. \tag{E.14}$$

By (E.11), (E.12) and Hölder's inequality, the products of degrees 2 and 0 from the term with $\nabla^3 G$ satisfy

$$\Big|\mathbb{E}\Big[\frac{1}{2}\,\nabla^3 G(Y_0+\overline{D}_0)\big[h\,\eta(Z_0)\,\hat{\xi},\, h\,\eta(Z_0)\,\hat{\xi},\, \zeta_0\big] + \frac{1}{6}\,\nabla^3 G(Y_0+\overline{D}_0) : \zeta_0^{\otimes 3}\Big]\Big| \leqslant CK\,\frac{h^{\frac{7}{2}}}{P^2}\,(1+|Y_0|)^6. \tag{E.15}$$

Inserting the bound on $\mathbb{E}|\varrho_4^{Y_1}|$, (E.14) and (E.15) into (E.13) gives

$$\begin{aligned}\Big|\mathbb{E}\Big[&G(Y_1) - G(Y_0+\overline{D}_0) - \nabla G(Y_0+\overline{D}_0)\cdot\zeta_0 \\ &- \frac{1}{2}\,\nabla^2 G(Y_0+\overline{D}_0) : \zeta_0^{\otimes 2} - \frac{h^2}{2P}\,\nabla^2 G(Y_0+\overline{D}_0) : \mathcal{C}(Z_0)\Big]\Big| \leqslant CK\,\frac{h^{\frac{7}{2}}}{P^2}\,(1+|Y_0|)^6.\end{aligned} \tag{E.16}$$

The products in (E.16) that contain ζ_0 are $\nabla G(Y_0+\overline{D}_0)\cdot\zeta_0$ and $\frac{1}{2}\,\nabla^2 G(Y_0+\overline{D}_0) : \zeta_0^{\otimes 2}$, and we estimate their means. For $\nabla G(Y_0+\overline{D}_0)\cdot\zeta_0$, (A.10), (E.4), the mean in (2.5) and (E.12) give

$$\begin{aligned}\Big\|\mathbb{E}\big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)\big] - I_d + \frac{h}{4P}\,\mathcal{C}(Y_0)\Big\| &\leqslant C\,\frac{h^2}{P^2}, \\ \Big\|\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0)\,\nabla^2 U(Z_0)\big] - \nabla^2 G(Y_0)\,\nabla^2 U(Y_0)\Big\| &\leqslant CKh^{\frac{1}{2}}(1+|Y_0|)^7,\end{aligned} \tag{E.17}$$

and inserting (E.17) into (C.58) gives

$$\begin{aligned}\Big|\mathbb{E}\big[\nabla G(Y_0+\overline{D}_0)\cdot\zeta_0\big] - 2h\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0)\big] : \Big(\mathbb{E}\big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)\big] - I_d\Big) \\ - \frac{h^3}{4P}\,\nabla^2 G(Y_0) : \big(\mathcal{C}(Y_0)\,\nabla^2 U(Y_0)\big)\Big| \leqslant CK\,\frac{h^{\frac{7}{2}}}{P}\,(1+|Y_0|)^7.\end{aligned} \tag{E.18}$$

For $\frac{1}{2}\,\nabla^2 G(Y_0+\overline{D}_0) : \zeta_0^{\otimes 2}$, since $\mathbb{E}[\chi_0^{\otimes 2}] = I_d$,

$$\frac{1}{2}\,\mathbb{E}\big[\nabla^2 G(Y_0) : \zeta_0^{\otimes 2}\big] = h\,\nabla^2 G(Y_0) : \mathbb{E}\Big[\Big(\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big) - I_d\Big)^2\Big], \tag{E.19}$$

and (E.12) and (E.4) allow replacing $\nabla^2 G(Y_0)$ by $\nabla^2 G(Y_0+\overline{D}_0)$ on the left side and by $\mathbb{E}[\nabla^2 G(Y_0+\overline{D}_0)]$ on the right side of (E.19):

$$\Big|\frac{1}{2}\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0) : \zeta_0^{\otimes 2}\big] - h\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0)\big] : \mathbb{E}\Big[\Big(\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big) - I_d\Big)^2\Big]\Big| \leqslant CK\,\frac{h^{\frac{7}{2}}}{P^2}\,(1+|Y_0|)^7. \tag{E.20}$$

The terms with the factor $\mathbb{E}[\nabla^2 G(Y_0+\overline{D}_0)]$ in (E.18) and (E.20) add up to

$$h\,\mathbb{E}\big[\nabla^2 G(Y_0+\overline{D}_0)\big] : \Big(\mathbb{E}\Big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)^2\Big] - I_d\Big). \tag{E.21}$$

The additional polynomial bound (3.19) on the eigenvalues of $\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)$ gives

$$\Big\|\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)^2 - I_d + \frac{h}{2}\,\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\Big\| \leqslant \frac{h^3}{8}\,\big\|\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big\|^3,$$

and (E.4) with $r=3$ and the mean in (2.5) give

$$\Big\|\mathbb{E}\Big[\mathcal{S}_h\big(\hat{\mathcal{C}}_P(Y_0;\boldsymbol{\xi}_0)\big)^2\Big] - I_d + \frac{h}{2P}\,\mathcal{C}(Y_0)\Big\| \leqslant C\,\frac{h^3}{P^3}. \tag{E.22}$$

Inserting (E.18)–(E.22) into (E.16) gives

$$\begin{aligned}\Big|\mathbb{E}\big[G(Y_1) - G(Y_0 + \overline{D}_0)\big] - \frac{h^2}{2P}\,\mathbb{E}\Big[\nabla^2 G(Y_0 + \overline{D}_0) : \big(\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\big)\Big] \\ - \frac{h^3}{4P}\,\nabla^2 G(Y_0) : \big(\mathcal{C}(Y_0)\,\nabla^2 U(Y_0)\big)\Big| \leqslant CK\,\frac{h^{\frac{7}{2}}}{P}\,(1 + |Y_0|)^7.\end{aligned} \tag{E.23}$$

It remains to expand the mean with the factor $\mathcal{C}(Z_0) - \mathcal{C}(Y_0)$ in (E.23). With $Z_0 - Y_0 = \sqrt{h/2}\,\chi_0$ and $\overline{D}_0 = \sqrt{2h}\,\chi_0 - h\,\nabla U(Z_0)$ of (C.2), Taylor's formula for $\mathcal{C}$ about Y_0 to the order 2 and for $\nabla^2 G$ about Y_0 to the order 1, with (E.5), gives

$$\begin{aligned}\mathcal{C}(Z_0) - \mathcal{C}(Y_0) &= \sqrt{\frac{h}{2}}\,\nabla\mathcal{C}(Y_0)[\chi_0] + \frac{h}{4}\,\nabla^2\mathcal{C}(Y_0)[\chi_0, \chi_0] + \varrho_3^{\chi_0}, \qquad \|\varrho_3^{\chi_0}\| \leqslant M^2 h^{\frac{3}{2}}\,|\chi_0|^3, \\ \nabla^2 G(Y_0 + \overline{D}_0) &= \nabla^2 G(Y_0) + \nabla^3 G(Y_0)[\overline{D}_0] + \varrho_2^{\chi_0}, \qquad \|\varrho_2^{\chi_0}\| \leqslant \frac{K}{2}\big(1 + |Y_0| + |\overline{D}_0|\big)^6\,|\overline{D}_0|^2.\end{aligned} \tag{E.24}$$

In the expansion of $\nabla^2 G(Y_0 + \overline{D}_0) : (\mathcal{C}(Z_0) - \mathcal{C}(Y_0))$ by (E.24), the means of the contractions

$$-h\,\nabla^3 G(Y_0)[\nabla U(Z_0)] : \big(\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\big), \quad \sqrt{2h}\,\nabla^3 G(Y_0)[\chi_0] : \Big(\mathcal{C}(Z_0) - \mathcal{C}(Y_0) - \sqrt{\frac{h}{2}}\,\nabla\mathcal{C}(Y_0)[\chi_0]\Big) \tag{E.25}$$

are bounded by (E.5) and (E.12). Hence (E.24) and (E.25) give

$$\Big|\mathbb{E}\Big[\nabla^2 G(Y_0 + \overline{D}_0) : \big(\mathcal{C}(Z_0) - \mathcal{C}(Y_0)\big)\Big] - \frac{h}{4}\,\nabla^2 G(Y_0) : \Delta\mathcal{C}(Y_0) - h\,\nabla\mathcal{C}(Y_0) : \nabla^3 G(Y_0)\Big| \leqslant CKh^{\frac{3}{2}}(1 + |Y_0|)^8. \tag{E.26}$$

By (C.2) with $Y_0 = y$ and (E.7), inserting (E.26) into (E.23) gives (E.9). □

E.3 Proof of Theorem 5

The proof of Theorem 5 replaces the mean of a function under μ_h by its mean under π, and Lemma 12 bounds the error of this replacement by applying Theorem 3 to the heat flow of the function.

Lemma 12 *Let Assumptions 1 and 4 hold, $h \in (0, h_0]$ and $P \geqslant 2$. Then the invariant distribution μ_h of NC-SGLM satisfies, for every $\psi \in C^1(\mathbb{R}^d)$ with $\|\nabla\psi\|_\infty < \infty$,*

$$|\mu_h(\psi) - \pi(\psi)| \leqslant Ch^{\frac{1}{2}}\,\|\nabla\psi\|_\infty. \tag{E.27}$$

Proof of Lemma 12. For $\tau \in (0, 1]$, we define the heat flow ψ_τ of ψ by

$$\psi_\tau(x) := \int_{\mathbb{R}^d} \psi(x + \sqrt{\tau}\,z)\,(2\pi)^{-\frac{d}{2}} e^{-\frac{|z|^2}{2}}\,\mathrm{d}z, \qquad x \in \mathbb{R}^d. \tag{E.28}$$

Then $\|\psi_\tau - \psi\|_\infty \leqslant (\tau d)^{1/2}\,\|\nabla\psi\|_\infty$, and $j - 1$ integrations by parts in z in (E.28) give $\|\nabla^j\psi_\tau\|_\infty \leqslant C\tau^{-(j-1)/2}\,\|\nabla\psi\|_\infty$ for $j = 1, \dots, 4$. Hence Assumption 2 holds with ψ_τ in place of f and $C\tau^{-3/2}\,\|\nabla\psi\|_\infty$ in place of L, and Theorem 3, the bound on $\nabla^3\phi_0$ in (C.115) and $\|\nabla^3 U\| \leqslant M$ give $|\pi_h(\psi_\tau) - \pi(\psi_\tau)| \leqslant Ch^2\tau^{-3/2}\,\|\nabla\psi\|_\infty$. By (3.1) and (E.28), $\pi_h(\psi_{h/2}) = \mu_h(\psi_h)$. Hence the bounds on $\psi_\tau - \psi$ with $\tau = h$ and $\tau = h/2$, and on $\pi_h(\psi_\tau) - \pi(\psi_\tau)$ with $\tau = h/2$, give

$$|\mu_h(\psi) - \pi(\psi)| \leqslant \|\psi - \psi_h\|_\infty + |\pi_h(\psi_{h/2}) - \pi(\psi_{h/2})| + \|\psi_{h/2} - \psi\|_\infty \leqslant Ch^{\frac{1}{2}}\,\|\nabla\psi\|_\infty,$$

which is (E.27). □

Proof of Theorem 5. Let $h \in (0, h_0]$ and $P \geqslant 2$. By Appendix E.1, Theorem 3 and its proof apply with $\overline{\pi}_h$ in place of π_h^{LM} as in Appendix C, and give the solutions $\tilde{\Phi}_h^\star$ of (C.112) and $\Phi_h^\star$ of (C.119). By (3.1) and (A.14), $\pi_h(f) - \overline{\pi}_h(f) = \mu_h(\tilde{f}_h) - \overline{\mu}_h(\tilde{f}_h)$. We show

$$\Big|\mu_h(\tilde{f}_h) - \overline{\mu}_h(\tilde{f}_h) - \frac{h^2}{P}\,\mathsf{K}^{\mathrm{SG}}(f)\Big| \leqslant CL\,\frac{h^{\frac{5}{2}}}{P}, \tag{E.29}$$

which, with (C.108) and (3.17), gives (3.22) and hence (3.24).

To prove (E.29), we first check the hypothesis of Lemma 11 with $K = CL$ for the solution $\tilde\phi_h - \tilde\Phi^\star_h$ of (C.114). By (C.114),

$$\tilde\phi_h - \tilde\Phi^\star_h = \overline{\mathcal{P}}_h\tilde\phi_h - \overline{\mathcal{P}}_h\tilde\Phi^\star_h + h\big(\tilde f_h - \overline\mu_h(\tilde f_h)\big). \tag{E.30}$$

By (3.2), the state $y + \overline{D}_0$ after the LM step from the initial state $Y_0 = y$ satisfies $\|\nabla^j_y(y + \overline{D}_0)\| \leqslant C$ for $j = 1, \dots, 4$. Then (A.22), (3.3) and the chain rule under the expectations in (C.2) and (A.14) give

$$\|\nabla^j(\overline{\mathcal{P}}_h\tilde\phi_h)\|_\infty \leqslant CL, \qquad \|\nabla^j \tilde f_h\|_\infty \leqslant L, \qquad j = 2, 3, 4. \tag{E.31}$$

By (C.112), Lemmas 6 and 10 with $F = \tilde\Phi^\star_h$ give

$$\big\|\nabla^j(\overline{\mathcal{P}}_h\tilde\Phi^\star_h)(y)\big\| \leqslant CLh^{2-\frac{j}{2}}(1 + |y|)^6, \qquad j = 2, 3, 4, \qquad y \in \mathbb{R}^d. \tag{E.32}$$

Hence (E.30), (E.31) and (E.32) give $\tilde\phi_h - \tilde\Phi^\star_h \in C^4(\mathbb{R}^d)$ and

$$\big\|\nabla^j(\tilde\phi_h - \tilde\Phi^\star_h)(y)\big\| \leqslant CL(1 + |y|)^6, \qquad j = 2, 3, 4, \qquad y \in \mathbb{R}^d. \tag{E.33}$$

Integrating (C.114) against the invariant distribution μ_h of NC-SGLM gives

$$h\big(\mu_h(\tilde f_h) - \overline\mu_h(\tilde f_h)\big) = \mu_h\Big(\mathcal{P}_h(\tilde\phi_h - \tilde\Phi^\star_h) - \overline{\mathcal{P}}_h(\tilde\phi_h - \tilde\Phi^\star_h)\Big),$$

and Lemma 11 with $G = \tilde\phi_h - \tilde\Phi^\star_h$ and $K = CL$ of (E.33), together with $\mu_h(|y|^{12}) \leqslant C$ of Lemma 1, gives

$$\Big|h\big(\mu_h(\tilde f_h) - \overline\mu_h(\tilde f_h)\big) - \frac{h^3}{P}\,\mu_h\big(\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h)\big)\Big| \leqslant CL\,\frac{h^{\frac72}}{P}. \tag{E.34}$$

By (E.34), (E.29) follows from

$$\Big|\mu_h\big(\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h)\big) - \mathsf{K}^{\mathrm{SG}}(f)\Big| \leqslant CLh^{\frac12}. \tag{E.35}$$

To prove (E.35), we first replace μ_h by π. By (E.30), $\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h) = \mathcal{A}(\overline{\mathcal{P}}_h\tilde\phi_h) - \mathcal{A}(\overline{\mathcal{P}}_h\tilde\Phi^\star_h) + h\,\mathcal{A}\tilde f_h$, and (E.8), (E.32) and (E.31) give

$$\Big|\big(\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h)\big)(y) - \big(\mathcal{A}(\overline{\mathcal{P}}_h\tilde\phi_h)\big)(y)\Big| \leqslant CLh^{\frac12}(1 + |y|)^6, \qquad y \in \mathbb{R}^d. \tag{E.36}$$

By (E.7), (E.31), (E.5) and (3.2), $\|\nabla\mathcal{A}(\overline{\mathcal{P}}_h\tilde\phi_h)\|_\infty \leqslant CL$. Hence (E.36), integrated against μ_h and against π, and Lemma 12 with $\psi = \mathcal{A}(\overline{\mathcal{P}}_h\tilde\phi_h)$ give

$$\Big|\mu_h\big(\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h)\big) - \pi\big(\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h)\big)\Big| \leqslant CLh^{\frac12}. \tag{E.37}$$

It remains to compare $\pi(\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h))$ with $\mathsf{K}^{\mathrm{SG}}(f)$. By (C.120),

$$\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h) = \mathcal{A}\phi_0 + \mathcal{A}(\Phi^\star_h - \tilde\Phi^\star_h). \tag{E.38}$$

Since $\pi(x) \leqslant Ce^{-m|x|^2/8}$ [30, Lemma 17], integrating by parts gives $\pi(\nabla\mathcal{C} : \nabla^3\phi_0) = \pi(\nabla^2\phi_0 : ((\nabla U\cdot\nabla)\mathcal{C} - \Delta\mathcal{C}))$. Hence (E.7) and (3.23) give

$$\pi(\mathcal{A}\phi_0) = \frac18\,\pi\Big(\nabla^2\phi_0 : \big(4\,(\nabla U\cdot\nabla)\mathcal{C} - 3\,\Delta\mathcal{C} + 2\,\mathcal{C}\,\nabla^2 U\big)\Big) = \mathsf{K}^{\mathrm{SG}}(f). \tag{E.39}$$

By (C.120), (C.115) and (E.33), $\Phi^\star_h - \tilde\Phi^\star_h$ is in $C^4(\mathbb{R}^d)$ with derivatives of polynomial growth. Integrating by parts gives $\pi(\mathcal{A}(\Phi^\star_h - \tilde\Phi^\star_h)) = \int_{\mathbb{R}^d}(\Phi^\star_h - \tilde\Phi^\star_h)\,\mathcal{A}^*\pi\,\mathrm{d}x$, where $\mathcal{A}^*$ is the adjoint operator in $L^2(\mathbb{R}^d)$, and

$$\mathcal{A}^*\pi = \frac18\sum_{i,j=1}^d \partial_i\partial_j\big((\Delta\mathcal{C})_{ij}\,\pi\big) - \frac12\sum_{i,j,k=1}^d \partial_i\partial_j\partial_k\big(\partial_k\mathcal{C}_{ij}\,\pi\big) + \frac14\sum_{i,j=1}^d \partial_i\partial_j\big((\mathcal{C}\,\nabla^2U)_{ij}\,\pi\big). \tag{E.40}$$

Since $\nabla\pi = -\pi\,\nabla U$, (E.40), (E.5) and (3.2) give $|(\mathcal{A}^*\pi)(x)| \leqslant C(1 + |x|)^3\,\pi(x)$. Hence (E.38), (E.39) and the bounds on $\Phi^\star_h$ in (C.119) and on $\tilde\Phi^\star_h$ in (C.112) give

$$\Big|\pi\big(\mathcal{A}(\tilde\phi_h - \tilde\Phi^\star_h)\big) - \mathsf{K}^{\mathrm{SG}}(f)\Big| = \Big|\pi\big(\mathcal{A}(\Phi^\star_h - \tilde\Phi^\star_h)\big)\Big| \leqslant CLh,$$

which, with (E.37), gives (E.35). □