HIGH-ORDER LANGEVIN MONTE CARLO ALGORITHMS

THANH DANG, MERT GÜRBÜZBALABAN, MOHAMMAD RAFIQUL ISLAM, NIAN YAO AND LINGJIONG ZHU

arXiv:2508.17545v1 [stat.ML] 24 Aug 2025

Abstract. Langevin algorithms are popular Markov chain Monte Carlo (MCMC) methods for large-scale sampling problems that often arise in data science. We propose Monte Carlo algorithms based on the discretizations of P-th order Langevin dynamics for any $P \geq 3$. Our design of P-th order Langevin Monte Carlo (LMC) algorithms is by combining splitting and accurate integration methods. We obtain Wasserstein convergence guarantees for sampling from distributions with log-concave and smooth densities. Specifically, the mixing time of the P-th order LMC algorithm scales as $O\left(d^{\frac{1}{\mathcal{R}}}/\epsilon^{\frac{1}{2\mathcal{R}}}\right)$ for $\mathcal{R} = 4 \cdot \mathbb{1}_{\{P=3\}} + (2P-1) \cdot \mathbb{1}_{\{P\geq 4\}}$, which has a better dependence on the dimension d and the accuracy level ϵ as P grows. Numerical experiments illustrate the efficiency of our proposed algorithms.

1. Introduction

Langevin algorithms are popular Markov chain Monte Carlo (MCMC) methods to sample from a given density $\mu(\theta) \propto e^{-U(\theta)}$ of interest where $\theta \in \mathbb{R}^d$, and these sampling problems appear in many applications such as Bayesian statistical inference, Bayesian formulations of inverse problems, and Bayesian classification and regression tasks in machine learning [GCSR95, Stu10, ADFDJ03, TTV16, GGHZ21, GIWZ24]. The classical Langevin Monte Carlo algorithm is based on the discretization of *overdamped (or first-order) Langevin dynamics* [Dal17, DM17, DK19, RRT17, BCM+21, CMR+21, EH21, ZADS23, BCE+22] that follows the stochastic differential equation (SDE):

$$d\theta_t = -\nabla U(\theta_t)dt + \sqrt{2}dB_t, \tag{1}$$

where $U : \mathbb{R}^d \to \mathbb{R}$ is often known as the *potential function*, and B_t is a standard d-dimensional Brownian motion with $\theta_0 \in \mathbb{R}^d$. Under some mild assumptions on $U(\cdot)$, the diffusion (1) admits a unique stationary distribution with the density $\mu(\theta) \propto e^{-U(\theta)}$, also known as the *Gibbs distribution* [CHS87, HKS89]. In computing practice, this diffusion is simulated by considering its discretization, and one of the most commonly used discretization schemes is the Euler–Maruyama discretization of (1), often known as the *unadjusted Langevin algorithm* in the literature; see e.g. [DM17]:

$$\theta_{k+1} = \theta_k - \eta\nabla U(\theta_k) + \sqrt{2\eta}\xi_{k+1}, \tag{2}$$

where ξ_k are i.i.d. $\mathcal{N}(0, I_d)$ Gaussian vectors.

In a seminal paper, [Dal17] obtained the first non-asymptotic result of the discretized Langevin dynamics (2); later, [DM17] improved the dependence on the dimension d. Both works consider the total variation (TV) as the distance to measure the convergence. In

Date: August 24, 2025.

contrast, [DM19] studied the convergence in the 2-Wasserstein distance, and [DMP18] studied variants of (2) when U is not smooth. [CB18] studied the convergence in the Kullback-Leibler distance. [EHZ22] obtained the convergence in chi-squared and Rényi divergence. [DK19, RRT17, BCM$^+$21, CMR$^+$21, ZADS23] studied the convergence when only stochastic gradients are available.

In the literature, many variants of the overdamped Langevin dynamics and the discretization schemes have been studied. One popular Langevin dynamics is the *underdamped Langevin dynamics*, also known as the *second-order* or kinetic Langevin dynamics, see e.g. [MSH02, Vil09, CCBJ18, CCA$^+$18, CLW21, CLW23, DRD20, GGZ20, MCC$^+$21, GGZ22]:

$$\begin{cases} dr_t = -\gamma r_t dt - \nabla U(\theta_t)dt + \sqrt{2\gamma}dB_t, \\ d\theta_t = r_t dt, \end{cases} \tag{3}$$

where B_t is a standard d-dimensional Brownian motion with $r_0, \theta_0 \in \mathbb{R}^d$. Under some mild assumptions on U, the SDE (3) admits a unique stationary distribution with the density $\mu(\theta, r) \propto e^{-U(\theta)-\frac{1}{2}|r|^2}$ [EGZ19], whose θ-marginal distribution coincides with the stationary distribution of (1). It is known that the second-order (underdamped) Langevin dynamics (3) might converge to the Gibbs distribution faster than the first-order (overdamped) Langevin dynamics [EGZ19, CLW23], and the discretization based on the second-order Langevin dynamics might have better iteration complexity, in particular, with a better dependence on the dimension d and the accuracy level ϵ [CCBJ18, GGZ22].

In the recent literature, higher-order, in particular, the *third-order* Langevin dynamics and its discretization have been proposed and studied in [MMW$^+$21]:

$$\begin{cases} d\theta_t = p_t \, dt, \\ dp_t = -\frac{1}{L}U(\theta_t) \, dt + \gamma r_t \, dt, \\ dr_t = -\gamma p_t \, dt - 2\gamma r_t \, dt + \sqrt{\frac{4\gamma}{L}} \, dB_t, \end{cases} \tag{4}$$

where $\gamma > 0$ is the friction parameter, L is the smoothness parameter of U and B_t is a standard Brownian motion in $\mathbb{R}^d$. Under some mild assumptions on U, the SDE (4) admits a unique stationary distribution with the density $\mu(\theta, r) \propto e^{-U(\theta)-\frac{L}{2}|p|^2-\frac{L}{2}|r|^2}$ [MMW$^+$21]. They showed that a Langevin Monte Carlo algorithm based on the discretization of the third-order Langevin SDE (4) can have even better iteration complexity in terms of dependence on the dimension d and the accuracy level ϵ compared to the algorithm based on the second-order Langevin SDE (3) [MMW$^+$21].

It is thus very natural to ask if one can propose and study a more general P-th order Langevin dynamics, and whether its discretization can lead to better iteration complexity. In the very recent probability literature, a generalized Langevin dynamics is studied in [Mon23]. Their result is in continuous time only. The focus of our paper is to propose and study the iteration complexity of an algorithm based on the discretization of the continuous-time P-th order Langevin dynamics, which we name P-th order Langevin Monte Carlo (LMC) algorithm. In the context of log-concave sampling via the Langevin equation and its variants, Table 1 compares the mixing time of our P-th order LMC

References	Assumptions on potential U	Mixing time in Wass_2
[CCBJ18, DRD20, MCC+21]	convex-smooth	$O\left(\frac{d^{1/2}}{\epsilon}\right)$
[SL19]	convex-smooth	$O\left(\frac{d^{1/3}}{\epsilon^{2/3}}\right)$
[MMW+21]	ridge-separable, convex-smooth	$O\left(\frac{d^{1/4}}{\epsilon^{1/2}}\right)$
[MMW+21]	strongly convex and smooth up to order α	$O\left(\frac{d^{1/4}}{\epsilon^{1/2}}\right)+O\left(\frac{d^{1/2}}{\epsilon^{1/\alpha-1}}\right)$
Our Theorem 2.19	convex-smooth and Condition H2	$O\left(d^{\frac{1}{\mathcal{R}}}/\epsilon^{\frac{1}{2\mathcal{R}}}\right)$, where $\mathcal{R}=4\cdot\mathbb{1}_{\{P=3\}}+(2P-1)\cdot\mathbb{1}_{\{P\geq 4\}}$

Table 1. Summary of assumptions and iteration complexities in our paper compared with the literature.

algorithm in Theorem 2.19 with the mixing time of other algorithms from the references in the literature[1]. Also note that the convex-smooth condition is our Condition H1.

Our contributions can be summarized as follows.

- We construct P-th order LMC algorithms that are based on discretizations of P-th order Langevin dynamics for $P\geq 3$. Under the condition that the potential function U is convex, sufficiently smooth and the operator norm of the derivatives of U do not grow too quickly, we show that the iteration complexity of our P-th order LMC algorithm scales as $O\left(d^{\frac{1}{\mathcal{R}}}/\epsilon^{\frac{1}{2\mathcal{R}}}\right)$ for $\mathcal{R}=4\cdot\mathbb{1}_{\{P=3\}}+(2P-1)\cdot\mathbb{1}_{\{P\geq 4\}}$. Our iteration complexity result therefore has a better dependence on the dimension d and the accuracy level ϵ as P grows. We therefore provide a positive answer to a conjecture in [MMW+21, Section 5] that one can construct LMC algorithms based on high-order Langevin dynamics that reduce the dependence of the iteration complexity on the dimension and the accuracy level.
- Inspired by existing work on second- and third-order Langevin Monte Carlo algorithms [CCBJ18, MMW+21], we propose and rigorously study novel discretization schemes for high-order Langevin dynamics that contain several stages of refinement, and each stage adopts a splitting scheme to ensure that the conditional expectation of the vector formed by the variables in each stage (conditioned on the last stage) follows a multivariate normal distribution. A natural question is what the maximum number of refinement stages one can design, which affects how much improvement one can obtain in iteration complexity. We discover that the maximum number of stages is $P-1$ (see Remark 2.15 and Remark 2.22).

[1]For comparison purpose, we focus solely on the dependence on d and ϵ of the rates in the cited references in the table. These references improve other aspects of log-concave sampling which we do not cover here.

- We perform numerical experiments and compare the performance of the third- and fourth-order LMC algorithms. In particular, we study sampling from the posterior distribution of the model parameters in Bayesian regression using real data, where the loss function is quadratic. Our numerical results show better performance for the fourth-order LMC algorithm. In addition, we consider a sigmoid loss function for sampling from the posterior distribution of the model parameters in Bayesian classification problems, using real data, which demonstrates the efficiency of our proposed algorithm.

The rest of the paper can be summarized as follows. In Section 2, main results are stated. We first provide some preliminaries for the continuous-time P-th order Langevin dynamics and state our main assumptions. In Section 2.1, for pedagogical purpose, we introduce and study the fourth-order LMC algorithm, and then in Section 2.2, we extend our results to any P-th order LMC algorithm for $P \geq 3$. We conduct numerical experiments to show the efficiency of our algorithms in Section 3. In Section 4, we conclude. Further technical details will be provided in the Appendix.

2. Main Results

In this section, we first present an important result regarding convergence toward equilibrium of (continuous-time) P-th order Langevin dynamics, that is established by Monmarché in [Mon23]. Let us start with some definitions.

Let $P, d \geq 1$. A P-th order Langevin dynamics has the form

$$\begin{aligned} dX_t &= AY_t dt, \\ dY_t &= -A^\top \nabla U(X_t) dt - \gamma B Y_t dt + \sqrt{\gamma} dW_t, \end{aligned} \tag{5}$$

where W is a standard $(P-1)d$-dimensional Brownian motion; $U \in \mathcal{C}^2(\mathbb{R}^d)$; while the $d \times (P-1)d$ matrix A and the $(P-1)d \times (P-1)d$ matrix B are given by:

$$A = \begin{pmatrix} I_d & 0 & \dots & 0 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 0 & -I_d & 0 & \dots & 0 \\ I_d & 0 & -I_d & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & I_d & 0 & -I_d \\ 0 & \dots & 0 & I_d & I_d \end{pmatrix}.$$

Regarding notations, set b as the drift coefficient of (5), that is

$$b(x, y) = \begin{pmatrix} Ay \\ -A^\top \nabla U(x) - \gamma B y \end{pmatrix}, \tag{6}$$

and denote J_b as its Jacobian matrix.

Next, we set

$$\hat{\lambda} = \min\{Re(\lambda), \lambda \text{ is an eigenvalue of } B_{\text{sim}}\}, \quad \text{where } B_{\text{sim}} := \begin{pmatrix} 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 0 & -1 \\ 0 & \dots & 0 & 1 & 1 \end{pmatrix},$$

noting that $B = B_{\text{sim}} \otimes I_d$ where $\otimes$ denotes the Kronecker product. Also,

$$\kappa = \begin{cases} \hat{\lambda}, & \text{when } B \text{ is diagonalizable}, \\ \hat{\lambda} - \epsilon, & \epsilon \in \left(0, \hat{\lambda}\right) \text{ when } B \text{ is not diagonalizable}. \end{cases}$$

We show in the proof of Corollary A.5 that $\hat{\lambda} > 0$, which implies $\kappa > 0$.

Denote the Jordan blocks of B_{sim} by $B_n, 1 \leq n \leq N$. Each block B_n of length ℓ_n is associated with the eigenvalue λ_n and the set of generalized eigenvectors $v_n^{(k)}; 1 \leq k \leq \ell_n$. In particular, $v_n^{(1)}$ is the (standard) eigenvector of B_n. Notice here we slightly abuse notations as we are using the same notations $B_n, \ell_n, v_n^{(k)}$ for the matrix B in Appendix A. For a Jordan block B_n with $\text{Re}(\lambda_n) > \kappa$, we set

$$H_n = \sum_{i=1}^{\ell_n} b_n^i v_n^{(i)} \left(\bar{v}_n^{(i)}\right)^\top,$$

where $\bar{v}^\top$ denotes the conjugate transpose of a vector v and

$$\begin{aligned} & b_n^1 = 1; \quad b_n^j = c_j (t_n)^{2(1-j)}, \quad 2 \leq j \leq \ell_n; \\ & c_1 = 1; \quad c_{j+1} = 1 + c_j^2, \quad 2 \leq j \leq \ell_n; \qquad t_n = 2\Big(\text{Re}(\lambda_n) - \hat{\lambda}\Big). \end{aligned} \tag{7}$$

Meanwhile, for a Jordan block B_m with $\text{Re}(\lambda_m) = \kappa$, we define

$$\widetilde{H}_m(\epsilon) = \sum_{i=1}^{\ell_m} b_m^i(\epsilon) v_m^{(i)} \left(\bar{v}_m^{(i)}\right)^\top,$$

where b_m^i's are the same as in (7), except that we replace the above t_n with $t_m = 2(\text{Re}(\lambda_m) - \hat{\lambda} + \epsilon)$ for any $\epsilon \in (0, \hat{\lambda})$ and write $b_m^i = b_m^i(\epsilon)$ to emphasize the dependence on ϵ. Now, assume $I = \{n \in \{1, \cdots, N\} : \ell_n \geq 2, \text{Re}(\lambda_n) = \hat{\lambda}\}$ and set

$$H(\epsilon) = \sum_{n \in \{1,\dots,N\} \setminus I} H_n + \sum_{m \in I} \widetilde{H}_m(\epsilon).$$

Next, we define

$$h_1 = \left\| H(\epsilon) \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \right\|_{\text{op}}, \qquad h_2 = h_3 = 1,$$

$$h_4 = \left(1 + \frac{P-1}{2}\right) \left\| H(\epsilon)^{-1} \right\|_{\text{op}}, \quad h_5 = (1 + P) \left\| H(\epsilon)^{-1} \right\|_{\text{op}}.$$

Regarding the potential function U, we assume that

Condition H1. U is m-strongly convex and L-smooth: $mI_d \le \nabla^2 U(x) \le LI_d$ for any $x \in \mathbb{R}^d$.

The following important result is established by Monmarché in [Mon23]. Further details are provided in Theorem A.4 and Corollary A.5 in Appendix A.

Theorem 2.1. *(a shortened version of Theorem A.4 and Corollary A.5) Assume the setup above for the P-th order Langevin dynamics* (5)*, including Condition H1 on the potential function U. If the friction γ is sufficiently large:*

$$\gamma \ge \gamma_0 := 2\sqrt{\frac{h_1 L}{\kappa}} \max\left\{ \sqrt{h_2 h_5}, \sqrt{\frac{h_4}{\kappa}} \right\},$$

then the $Pd \times Pd$ matrix $M := \begin{pmatrix} 1 & \frac{1}{\gamma}\begin{pmatrix}1 & \dots & 1\end{pmatrix} \\ \frac{1}{\gamma}\begin{pmatrix}1 & \dots & 1\end{pmatrix}^\top & \frac{\kappa}{Lh_1} H(\epsilon) \end{pmatrix} \otimes I_d$ is symmetric, positive definite and satisfies

$$MJ_b + J_b^\top M \le -2\rho M, \qquad \rho = \min\left\{ \frac{m}{3h_3\gamma}, \frac{\gamma\kappa}{6} \right\}. \tag{8}$$

In particular, $\rho = \rho(\gamma, L, P)$ and $\gamma_0 = \gamma_0(\gamma, L, P)$ depend on γ, L, P but not on the dimension parameter d. Furthermore, $\lambda_{\min,M} = \lambda_{\min,M}(P)$ and $\lambda_{\max,M} = \lambda_{\max,M}(P)$ are respectively the smallest and largest eigenvalues of the positive definite matrix M, and they depend on P but not on d.

Remark 2.2. Condition H1 is exactly Condition F in Appendix A specified for the P-th order Langevin dynamics (5).

Example 2.3. Here we demonstrate how to find γ_0 and M in Theorem 2.1 in the case $P = 4$. The matrix $B = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix}$ is diagonalizable. It has eigenvector $v_1 \approx (-0.877 - 0.745i, -0.785 + 1.307i, 1)$ corresponding to eigenvalue $0.215 + 1.307i$, $v_2 \approx (-0.877 + 0.745i, -0.785 - 1.307i, 1)$ corresponding to eigenvalue $0.215 - 1.307i$ and $v_3 \approx (0.755, -0.430, 1)$ corresponding to eigenvalue 0.570. Then the matrix H is approximately $\begin{pmatrix} 3.341 & -1.004 & -0.999 \\ -1.004 & 4.850 & -2.000 \\ -0.999 & -2.000 & 3.000 \end{pmatrix}$ with eigenvalues (approximately) $6.168, 4.098$ and 0.924. From there, one deduces $h_1 \approx 3.341, h_4 = 2.705, h_5 = 5.410, \kappa = 0.924$ and we know beforehand that $h_2 = h_3 = 1$. Thus, we have $\gamma_0 \approx 2\sqrt{\frac{3.341L}{0.215}} \cdot (3.547)$ and

$$M \approx \begin{pmatrix} 1 & 1/\gamma & 1/\gamma & 1/\gamma \\ 1/\gamma & (1/L)0.924 & -(1/L)0.278 & -(1/L)0.276 \\ 1/\gamma & -(1/L)0.278 & (1/L)1.341 & -(1/L)0.553 \\ 1/\gamma & -(1/L)0.276 & -(1/L)0.553 & (1/L)0.830 \end{pmatrix} \otimes I_d.$$

In the upcoming part, we will assume a strengthened version of Assumption 2 in [MMW$^+$21] about the potential function U. This strengthened assumption will ensure that we can approximate the nested integrals in Lemma B.1 with reasonable accuracy, and ultimately allow us to construct an MCMC algorithm with a better discretization error (with respect to the dimension d and the accuracy level ϵ) than [MMW$^+$21]. We note that in the case where U is not a polynomial or a piece-wise polynomial function, the upcoming condition basically asks that $\sup_{x\in\mathbb{R}^d} \|\nabla^\alpha U(x)\|_{\mathrm{op}}$ does not grow too fast as α increases.

Condition H2. Let the stepsize η and the dimension d be fixed. There exists a positive real number c that does not depend on the dimension d and a positive integer α large enough such that U is in $\mathcal{C}^\alpha$ and

$$\left(\frac{L_\alpha}{\alpha!}\right)^2 \left(\widetilde{C}_1\right)^\alpha (d+2\alpha)^\alpha \le c \cdot d \cdot \left(\mathbb{1}_{\{P=3\}}\eta^4 + \mathbb{1}_{\{P\ge 4\}}\eta^{2P-1}\right),$$

where $L_\alpha := \sup_{x\in\mathbb{R}^d} \|\nabla^\alpha U(x)\|_{\mathrm{op}}$. $\widetilde{C}_1$ given in Lemma C.4 is a positive constant that depends only on the friction parameter γ and the smoothness parameter L, but not on the dimension d or the stepsize η.

Remark 2.4. We observe that Condition H2 is satisfied whenever U is a polynomial of some degree k, since we can take $\alpha = k+1$ so that $\nabla^\alpha U \equiv 0$. This is the case with quadratic loss function in our numerical experiments for Bayesian linear regression (our Section 3.1). More generally, one can consider a polynomial regression problem [Jun22, Section 3.2].

Remark 2.5. In the case where U is not a polynomial, an example is the regularized Huber loss function that is $U(x) := U_0(x) + \frac{\lambda}{2}|x|^2$ for some $\lambda > 0$, where

$$U_0(x) := \begin{cases} \frac{|x|^2}{2} & \text{if } |x| \le \alpha, \\ \alpha\,|x| - \frac{\alpha^2}{2} & \text{otherwise,} \end{cases}$$

for some positive parameter α ([SC08, Page 44]). In fact, for this example, we do not need to verify Condition H2 since the latter is to ensure we can approximate the nested integrals in Lemma B.1 (a fact pointed out in the paragraph before Condition H2).

Remark 2.6. In the case where U is not a polynomial or a piece-wise polynomial function, Condition H2 basically asks that $\sup_{x\in\mathbb{R}^d} \|\nabla^\alpha U(x)\|_{\mathrm{op}}$ does not grow too fast as α increases. This Condition as stated is quite hard to verify however. Hence, an example of a condition that implies Condition H2 and is easier to check than the latter is: there exists an integer $K \in \mathbb{N}$ and real numbers $c, \beta > 1$ such that for every $k \ge K$,

$$\sup_{x\in\mathbb{R}^d} \left\|\nabla^k U(x)\right\|_{\mathrm{op}} \le \sqrt{c\Gamma(k/\beta+1)}d^k, \tag{9}$$

where $\Gamma(\cdot)$ is the gamma function. Then, since

$$\lim_{k\to\infty} \frac{c\Gamma(k/\beta+1)d^{2k}\widetilde{C}_1^k(d+2k)^k}{(k!)^2} = \lim_{k\to\infty} \frac{c\sqrt{\frac{2\pi k}{\beta}}(\frac{k}{\beta e})^{k/\beta}d^{2k}\widetilde{C}_1^k(d+2k)^k}{2\pi k(\frac{k}{e})^{2k}} = 0,$$

for any fixed $\beta > 1$ and d, where we applied the Stirling's formula $\Gamma(x+1) \sim \sqrt{2\pi x}(\frac{x}{e})^x$ as $x \to \infty$, the parameter α in Condition H2 is guaranteed to exist. Finally, we note that (9) is similar to the assumption in [WWJ16, Theorem 3.3] in the context of accelerated gradient methods in optimization.

Remark 2.7. Our Condition H2 is much stronger than Assumption 2 in [MMW+21], even though both are roughly about the smoothness of the loss function U. The reason is as follows. The mixing time of our P-th order LMC algorithm is determined by the error in our discretization scheme of a P-th order Langevin dynamics. As it will be clear from our proofs, the discretization error is a sum of two parts: the first part being the error of a splitting scheme, and the second part being the error of a polynomial approximation. As P increases, we can show that the former gets smaller; however, we cannot do the same for the latter. Thus, in order to obtain an improvement of the discretization error as P increases, one must assume some condition for the polynomial approximation error to be dominated by the splitting scheme error. Condition H2 ensures this outcome.

Remark 2.8. In practice, even when condition (9) or Condition H2 is not satisfied, our P-th order LMC algorithm might still work well; see, for example, our numerical experiments for Bayesian logistic regression (Section 3.2).

2.1. **Fourth-order Langevin Monte Carlo Algorithm.**

2.1.1. *Fourth-order Langevin Monte Carlo algorithm.* Given the iterate $x^{(k)}$, the next iterate $x^{(k+1)}$ is obtained by drawing from a multivariate normal distribution with mean $\mathbf{M}(x^{(k)})$ and covariance $\mathbf{\Sigma}$, both of which are stated in Lemma B.2.

The proof of the next result is presented at the end of Section 2.1.3.

Theorem 2.9. *Assume Equation* (10) *satisfies Conditions H1 and H2. Let a be any positive constant satisfying*

$$a \leq \min\left\{\frac{m}{3\gamma}, \frac{\gamma\kappa}{6}\right\}\lambda_{\min,M},$$

where the positive definite matrix M and the constants γ, m, L, κ are from Section 2. Denote μ the invariant measure associated with the fourth-order Langevin dynamics (10)*.*

Choose a 2*-Wasserstein accuracy of ϵ small enough such that $\eta_0 := \left(\frac{\epsilon^2}{2C_1 d}\right)^{1/7} < \min\{\eta^*, \frac{1}{h}\}$ where h is defined in Proposition 2.18 and η^*, C_1 are from Lemma B.4. Suppose we run our fourth-order Langevin Monte Carlo algorithm with stepsize η_0, then $\mathrm{Wass}_2\big(\mathrm{Law}(x^{(k^*)}), \mu\big) \leq \epsilon$, where k^* is the mixing time of the fourth-order Langevin Monte Carlo algorithm with respect to μ that is given by*

$$k^* = \log\left(\frac{2C_4\mathbb{E}_{Z\sim\mu}\left[\left|Z - x^{(0)}\right|^2\right]}{\epsilon^2}\right)\frac{(2C_3)^{1/7}}{h}\frac{d^{1/7}}{\epsilon^{2/7}} - 1,$$

where C_3, C_4 are positive constants that depend on γ, L, c but do not depend on the dimension parameter d.

Remark 2.10. Our mixing time rate of $O\left(\frac{d^{1/7}}{\epsilon^{2/7}}\right)$ improves upon the rates in [MMW$^+$21] in terms of both d and ϵ dependencies. For instance, [MMW$^+$21, Theorem 1] has a mixing time rate of $O\left(\frac{d^{1/4}}{\epsilon^{1/2}}\right)$.

2.1.2. *Derivation of the discretization scheme.* Consider the fourth-order Langevin dynamics:

$$\begin{aligned} d\theta(t) &= v_1(t)dt, \\ dv_1(t) &= \left(-\nabla U(\theta_t) + \gamma v_2(t)\right) dt, \\ dv_2(t) &= \left(-\gamma v_1(t) + \gamma v_3(t)\right) dt, \\ dv_3(t) &= (-\gamma v_2(t) - \gamma v_3(t))dt + \sqrt{2\gamma} dB_t. \end{aligned} \tag{10}$$

Remark 2.11. The equation (4) in our introduction (Section 1) is studied in [MMW$^+$21] and contains two parameters (namely γ and L in their paper) compared to our equation (5) that contains only a single parameter γ. We make such an assumption out of convenience and our paper is able to handle extra parameters as in [MMW$^+$21], and this is explained in Appendix A. Specifically in Equation (46) in Appendix A, we can take $A = -\frac{1}{L}(I_d, 0, \ldots, 0)$ and $\Sigma = \sqrt{\frac{4}{L}} I_p$.

Below we will write $|\cdot|$ for the Euclidean norm and $|\cdot|_M$ for the M-norm $|x|_M = \sqrt{x^\top M x}$. The numerical scheme for fourth-order Langevin dynamics consists of three stages: updating $x^{(k)}$ to $\hat{x}(t)$, then updating $\hat{x}(t)$ to $\tilde{x}(t)$, then updating $\tilde{x}(t)$ to $\bar{x}(t)$ (for $t \in [k\eta, (k+1)\eta]$). Each stage adopts a splitting scheme.

Stage 1: Set the initial value

$$\hat{x}(k\eta) := \left(\hat{\theta}(k\eta), \hat{v}_1(k\eta), \hat{v}_2(k\eta), \hat{v}_3(k\eta)\right) = x^{(k)}.$$

For $t \in (k\eta, (k+1)\eta]$, let

$$\hat{v}_1(t) = v_1^{(k)},$$

and

$$\begin{aligned} d\hat{\theta}(t) &= \hat{v}_1(t)dt, \\ d\hat{v}_2(t) &= \left(-\gamma \hat{v}_1(t) + \gamma v_3^{(k)}\right) dt, \\ d\hat{v}_3(t) &= (-\gamma \hat{v}_2(t) - \gamma \hat{v}_3(t))dt + \sqrt{2\gamma} dB_t. \end{aligned}$$

Stage 2: Set the initial value $\tilde{x}(k\eta) = x^{(k)}$. For $t \in (k\eta, (k+1)\eta]$, let

$$d\tilde{v}_1(t) = \left(-\tilde{g}(t) + \gamma \hat{v}_2(t)\right) dt,$$

and

$$\begin{aligned} d\tilde{\theta}(t) &= \tilde{v}_1(t)dt, \\ d\tilde{v}_2(t) &= (-\gamma \tilde{v}_1(t) + \gamma \hat{v}_3(t))dt, \end{aligned}$$

$$d\tilde{v}_3(t) = (-\gamma\tilde{v}_2(t) - \gamma\tilde{v}_3(t))dt + \sqrt{2\gamma}dB_t,$$

where $\tilde{g}(t)$ is a polynomial (in t) of degree $\alpha - 1$ and approximates $\nabla U(\hat{\theta}(t))$, and $\tilde{g}(t)$ will be defined in (13) below.

Stage 3: Set $\bar{x}(k\eta) = x^{(k)}$. For $t \in (k\eta, (k+1)\eta]$, let

$$d\bar{v}_1(t) = (-\bar{g}(t) + \gamma\tilde{v}_2(t))\,dt,$$

and

$$\begin{aligned} d\bar{\theta}(t) &= \bar{v}_1(t)dt, \\ d\bar{v}_2(t) &= (-\gamma\bar{v}_1(t) + \gamma\tilde{v}_3(t))dt, \\ d\bar{v}_3(t) &= (-\gamma\bar{v}_2(t) - \gamma\bar{v}_3(t))dt + \sqrt{2\gamma}dB_t, \end{aligned}$$

where $\bar{g}(t)$ is a polynomial (in t) of degree $\alpha - 1$ and approximates $\nabla U(\tilde{\theta}(t))$, and $\bar{g}(t)$ will be defined in (13) below.

Finally, set

$$x^{(k+1)} = \bar{x}((k+1)\eta). \tag{11}$$

Definitions of $\tilde{g}(t)$ and $\bar{g}(t)$: Recall U is a map from $\mathbb{R}^d$ to $\mathbb{R}$, so that ∇U is a map from $\mathbb{R}^d$ to $L(\mathbb{R}^d, \mathbb{R}^d)$ where $L(\mathbb{R}^d, \mathbb{R}^d)$ is the space consisting of bounded linear maps from $\mathbb{R}^d$ to $\mathbb{R}^d$. Per [Car71, Page 70], the Taylor polynomial of degree $\alpha - 1$ which is associated with ∇U and centers at the origin is

$$P_{\alpha-1}(x) = \sum_{k=0}^{\alpha-1} \frac{\nabla^k U(0)}{(k-1)!} x^{k-1}, \tag{12}$$

where per [FA89], $\frac{\nabla^k U(0)}{(k-1)!} x^{k-1} = \sum_{i_1+\ldots+i_d=k-1} \frac{1}{i_1!\ldots i_d!} \frac{\partial^k U}{\partial x_1^{i_1}\cdots\partial x_d^{i_d}}(0) x_1^{i_1}\ldots x_d^{i_d}$.

This allows us to define

$$\tilde{g}(t) := P_{\alpha-1}(\hat{\theta}(t)); \qquad \bar{g}(t) := P_{\alpha-1}(\tilde{\theta}(t)), \tag{13}$$

where

$$\hat{\theta}(t) = \theta^{(k)} + (t - k\eta)v_1^{(k)},$$

and

$$\begin{aligned} \tilde{\theta}(t) = \theta^{(k)} &+ v_1^{(k)}(t - k\eta) - \int_{k\eta}^{t}\int_{k\eta}^{s} \tilde{g}(r)drds \\ &+ \gamma v_2^{(k)} \frac{(t-k\eta)^2}{2!} + \gamma^2\left(v_3^{(k)} - v_1^{(k)}\right)\frac{(t-k\eta)^3}{3!}. \end{aligned}$$

Remark 2.12. The definitions of $\bar{g}$ and $\tilde{g}$ in (13) require finding multivariate Taylor polynomials, which is a challenging task in itself. One can use numerical software to help with this, for example, by using Maple™ ([Red12]) or the calculus package in R ([Gui22]).

The next result is a consequence of Lemma B.1 and Lemma B.2 from Appendix B.

Proposition 2.13. $\mathbb{E}\big[x^{(k+1)}|x^{(k)}\big] = \mathbb{E}[\bar{x}((k+1)\eta)|\bar{x}(k\eta)]$ *follows a multivariate normal distribution with mean* $\boldsymbol{M}(x^{(k)}) \in \mathbb{R}^4$ *and covariance* $\boldsymbol{\Sigma} \in \mathbb{R}^{4\times4}$. *The explicit forms of* $\boldsymbol{M}(x^{(k)})$ *and* $\boldsymbol{\Sigma}$ *are stated in Lemma B.2.*

Remark 2.14. The authors of [MMW+21] propose an MCMC algorithm based on third-order Langevin dynamics. In the case where U is a general potential function and not ridge separable, an important step in their algorithm is the Lagrange polynomial interpolation step ([MMW+21, Section 3.3]) to approximate the path $s \mapsto \nabla U\big(\theta^{(k)} + (s-k\eta)p^{(k)}\big)$ for given vectors $\theta^{(k)}, p^{(k)}$ in $\mathbb{R}^d$ and $s \in [k\eta, (k+1)\eta]$. There seems to be some major difficulty in applying this Lagrange polynomial interpolation step to our MCMC algorithm based on fourth-order Langevin dynamics, which pushes us to use Taylor approximation of ∇U instead. We further explain the difficulty of using Lagrange polynomial interpolation for our algorithm in Appendix D.

Remark 2.15. One cannot add another stage to the above discretization procedure of the fourth-order Langevin dynamics, since it is unclear how to implement the resulting algorithm in that case. The reason the current algorithm which is based on a three-stage discretization procedure can be easily implemented is that per Proposition 2.13, $\mathbb{E}\big[x^{(k+1)}|x^{(k)}\big] = \mathbb{E}\big[\bar{x}((k+1)\eta)|x^{(k)}\big]$ is a multivariate normal distribution. Now suppose that we add another stage of the discretization procedure:

Stage 4: Set $\check{x}(k\eta) = x^{(k)}$. For $t \in (k\eta, (k+1)\eta]$, let

$$d\check{v}_1(t) = (-\check{g}(t) + \gamma\bar{v}_2(t))\,dt,$$

and

$$\begin{aligned} d\check{\theta}(t) &= \check{v}_1(t)dt, \\ d\check{v}_2(t) &= (-\gamma\check{v}_1(t) + \gamma\bar{v}_3(t))dt, \\ d\check{v}_3(t) &= (-\gamma\check{v}_2(t) - \gamma\check{v}_3(t))dt + \sqrt{2\gamma}dB_t, \end{aligned}$$

where $\check{g}(t) := P_{\alpha-1}(\bar{\theta}_t)$ is a polynomial (in t) of degree $\alpha-1$ and approximates $\nabla U(\bar{\theta}(t))$, noting that $P_{\alpha-1}$ is the multivariate Taylor polynomial given in (12). Per Lemma B.1, $\bar{\theta}(t)$ has the general form $F(k,\eta,\gamma,t) + \int_{k\eta}^{t} G(k,\eta,\gamma,s)dB_s$, so that $\check{\theta}(t)$ is approximately

$$\theta^{(k)} - \int_{k\eta}^{t} \nabla U\bigg(F(k,\eta,\gamma,s) + \int_{k\eta}^{s} G(k,\eta,\gamma,r)dB_r\bigg)ds + \gamma\int_{k\eta}^{t} \bar{v}_2(s)ds.$$

In the case where U is not a quadratic potential function, the presence of the term $\nabla U\Big(F(k,\eta,\gamma,s) + \int_{k\eta}^{s} G(k,\eta,\gamma,r)dB_r\Big)$ on the right hand side suggests that $\mathbb{E}\big[\check{\theta}(t)|x^{(k)}\big]$ may not be multivariate normal, which makes the algorithm difficult to implement. Consequently, we do not have more than three stages in our discretization procedure.

2.1.3. *Proofs.* We need a few technical lemmas whose proofs are placed near the end of Appendix B. First, we quantify how well $\tilde{g}(t)$ and $\bar{g}(t)$ respectively approximate $\nabla U(\hat{\theta}(t))$ and $\nabla U(\tilde{\theta}(t))$.

Lemma 2.16. *Under Conditions H1, it holds that*

$$\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\nabla U(\hat{\theta}(t)) - \tilde{g}(t)\right|^2\right] \le \left(\frac{L_\alpha}{\alpha!}\right)^2 \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\hat{\theta}(t)^{2\alpha}\right|\right],$$

$$\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\nabla U(\tilde{\theta}(t)) - \bar{g}(t)\right|^2\right] \le \left(\frac{L_\alpha}{\alpha!}\right)^2 \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\tilde{\theta}(t)^{2\alpha}\right|\right].$$

where $L_\alpha := \sup_{x\in\mathbb{R}^d} \|\nabla^\alpha U(x)\|_{\mathrm{op}}$.

Next, we bound the differences in L^2-norm of variables of two consecutive stages.

Lemma 2.17. *Under Conditions H1 and H2, it holds for* $t \in (k\eta, (k+1)\eta]$ *that*

$$\mathbb{E}\left[|\bar{v}_1(t) - \tilde{v}_1(t)|^2\right] \le C_2(d+1)\Big(\gamma^2\Big((\gamma+1)^2 + (2\gamma+\sqrt{2\gamma})^2\Big) + c\Big)\eta^5;$$

$$\mathbb{E}\left[\left|\tilde{\theta}(t) - \bar{\theta}(t)\right|^2\right] \le C_2(d+1)\Big(\gamma^2\Big((\gamma+1)^2 + (2\gamma+\sqrt{2\gamma})^2\Big) + c\Big)\eta^7;$$

$$\begin{aligned}\mathbb{E}\left[|\bar{v}_2(t) - \tilde{v}_2(t)|^2\right] \le\ & C_2(d+1)\gamma^2\Big(\gamma^2\Big((\gamma+1)^2 + (2\gamma+\sqrt{2\gamma})^2\Big) + c\Big)\eta^7 \\ & + C_2(d+1)\gamma^2\Big((\gamma+1)^2 + (2\gamma+\sqrt{2\gamma})^2\Big)\eta^7;\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[|\bar{v}_3(t) - \tilde{v}_3(t)|^2\right] \le\ & C_2(d+1)\gamma^4\Big(\gamma^2\Big((\gamma+1)^2 + (2\gamma+\sqrt{2\gamma})^2\Big) + c\Big)\eta^9 \\ & + C_2(d+1)\gamma^4\Big((\gamma+1)^2 + (2\gamma+\sqrt{2\gamma})^2\Big)\eta^9.\end{aligned}$$

The upcoming result bounds the discretization error of the numerical scheme (11).

Proposition 2.18. *Assume Equation* (10) *satisfies Conditions H1 and H2. Let a be any positive constant satisfying* $a \le \min\left\{\frac{m}{3\gamma}, \frac{\gamma\kappa}{6}\right\}\lambda_{\min,M}$*, where the positive definite matrix* M *and the constants* γ, m, L, κ *are from Section 2. Denote* μ *the invariant measure of the fourth-order Langevin dynamics* (10)*.*

Then regarding the discretization error, it holds when $\eta < \min\{\eta^*, \frac{1}{h}\}$ *that*

$$\mathbb{E}\left[\left|x((k+1)\eta) - x^{(k+1)}\right|^2\right] \le C_3 d\eta^8 + C_4 e^{-(k+1)h\eta}\mathbb{E}\left[\left|Z - x^{(0)}\right|^2\right], \qquad Z \sim \mu.$$

In particular, η^* *is defined at* (55)*, and* $h := 2\rho - \frac{2a}{\lambda_{\min,M}}$ *where* ρ, M *are from Theorem 2.1 and a is any positive constant equal to or less than* $\min\left\{\frac{m}{3\gamma}, \frac{\gamma\kappa}{6}\right\}\lambda_{\min,M}$*. Moreover,* C_3, C_4 *are positive constants that depend only on* γ, L, c *but do not depend on the dimension parameter* d*.*

Proof. **Step 1:** Assume $t \in [k\eta, (k+1)\eta]$ and recall that $\bar{x}(t) = \big(\bar{\theta}(t), \bar{v}_1, \bar{v}_2(t), \bar{v}_3(t)\big)$. Based on (11), we have

$$d\bar{x}(t) = \bar{b}(t)dt + \sqrt{2\gamma}DdB_t,$$

where

$$D := \begin{pmatrix} 0_d & 0_d & 0_d & 0_d \\ 0_d & 0_d & 0_d & 0_d \\ 0_d & 0_d & 0_d & 0_d \\ 0_d & 0_d & 0_d & I_d \end{pmatrix}; \qquad \bar{b}(t) := \begin{pmatrix} \bar{v}_1(t) \\ -\bar{g}(t) + \gamma\tilde{v}_2(t) \\ -\gamma\bar{v}_1(t) + \gamma\tilde{v}_3(t) \\ -\gamma\bar{v}_2(t) - \gamma\bar{v}_3(t) \end{pmatrix}.$$

Meanwhile, the fourth-order Langevin dynamics in (10) can be written as

$$dx(t) = b(x(t))dt + \sqrt{2\gamma} D dB_t, \qquad b(x) = \begin{pmatrix} v_1 \\ -\nabla U(\theta) + \gamma v_2 \\ -\gamma v_1 + \gamma v_3 \\ -\gamma v_2 - \gamma v_3 \end{pmatrix}.$$

Then

$$d(x(t) - \bar{x}(t)) = (b(x(t)) - b(\bar{x}(t)))dt + \big(b(\bar{x}(t)) - \bar{b}(t)\big)dt.$$

This leads to

$$\begin{aligned} &\frac{d}{dt}\mathbb{E}\Big[(x(t) - \bar{x}(t))^\top M(x(t) - \bar{x}(t))\Big] \\ &= \mathbb{E}\Big[(x(t) - \bar{x}(t))^\top M(b(x(t)) - b(\bar{x}(t)))\Big] + \mathbb{E}\Big[(x(t) - \bar{x}(t))^\top M\big(b(\bar{x}(t)) - \bar{b}(t)\big)\Big] \\ &\quad + \mathbb{E}\Big[(b(x(t)) - b(\bar{x}(t)))^\top M(x(t) - \bar{x}(t))\Big] + \mathbb{E}\Big[\big(b(\bar{x}(t)) - \bar{b}(t)\big)^\top M(x(t) - \bar{x}(t))\Big], \end{aligned} \tag{14}$$

where

$$b(\bar{x}(t)) - \bar{b}(t) = \begin{pmatrix} 0 \\ \bar{g}(t) - \nabla U(\bar{\theta}(t)) - \gamma(\tilde{v}_2(t) - \bar{v}_2(t)) \\ \gamma(\bar{v}_3(t) - \tilde{v}_3(t)) \\ 0 \end{pmatrix}. \tag{15}$$

Recall the M-norm $|x|_M = \sqrt{x^\top M x}$ and notice that

$$\begin{aligned} &(x(t) - \bar{x}(t))^\top M(b(x(t)) - b(\bar{x}(t))) + (b(x(t)) - b(\bar{x}(t)))^\top M(x(t) - \bar{x}(t)) \\ &\le (x(t) - \bar{x}(t))^\top M \int_0^1 J_b(wx(t) + (1-w)\bar{x}(t))(x(t) - \bar{x}(t))dw \\ &\quad + \int_0^1 (x(t) - \bar{x}(t))^\top J_b(wx(t) + (1-w)\bar{x}(t))^\top dw M(x(t) - \bar{x}(t)) \\ &\le (x(t) - \bar{x}(t))^\top(-2\rho)M(x(t) - \bar{x}(t)) = -2\rho\,|x(t) - \bar{x}(t)|_M^2, \end{aligned} \tag{16}$$

where the last line is due to the contraction property (8) in Theorem 2.1.

Moreover, choose any positive $a \le \min\left\{\frac{m}{3\gamma}, \frac{\gamma\kappa}{6}\right\}\lambda_{\min,M}$ and notice that per Theorem 2.1,

$$-\rho + \frac{a}{\lambda_{\min,M}} < 0. \tag{17}$$

From (14), (16), (17) and Cauchy-Schwarz inequality, we can deduce that

$$\frac{d}{dt}\mathbb{E}\big[|x(t) - \bar{x}(t)|_M^2\big] \tag{18}$$

$$\leq -2\rho\mathbb{E}\left[|x(t)-\bar{x}(t)|_M^2 + 2a\,|x(t)-\bar{x}(t)|^2\right] + \frac{2}{a}\,\|M\|_{\mathrm{op}}^2\,(\gamma^2+1)\mathbb{E}\left[\left|b(\bar{x}(t))-\bar{b}(t)\right|^2\right]. \tag{19}$$

By combining the bound in (19) with

$$\sqrt{\lambda_{\min,M}}\,|x| \leq |x|_M \leq \sqrt{\lambda_{\min,M}}\,|x|\,, \tag{20}$$

for any x, we get

$$\begin{aligned}&\frac{d}{dt}\mathbb{E}\left[|x(t)-\bar{x}(t)|_M^2\right]\\ &\leq \left(-2\rho+\frac{2a}{\lambda_{\min,M}}\right)\mathbb{E}\left[|x(t)-\bar{x}(t)|_M^2\right] + \frac{2}{a}\,\|M\|_{\mathrm{op}}^2\,(\gamma^2+1)\mathbb{E}\left[\left|b(\bar{x}(t))-\bar{b}(t)\right|^2\right].\end{aligned} \tag{21}$$

Step 2: We will bound $\mathbb{E}\left[\left|b(\bar{x}(t))-\bar{b}(t)\right|^2\right]$ as the second term on the right hand side of (21). Based on (15), we will need to bound the L^2 norm of

$$\bar{g}(t)-\nabla U(\bar{\theta}(t)),\quad \gamma(\tilde{v}_2(t)-\bar{v}_2(t)),\quad \text{and}\quad \gamma(\bar{v}_3(t)-\tilde{v}_3(t)). \tag{22}$$

Let us start with

$$\mathbb{E}\left[\left|\bar{g}(t)-\nabla U(\bar{\theta}(t))\right|^2\right] \leq 2\mathbb{E}\left[\left|\bar{g}(t)-\nabla U(\tilde{\theta}(t))\right|^2\right] + 2\mathbb{E}\left[\left|\nabla U(\tilde{\theta}(t))-\nabla U(\bar{\theta}(t))\right|^2\right]. \tag{23}$$

The first term on the right hand side in (23) is bounded in (59) as $\mathbb{E}\left[\left|\bar{g}(t)-\nabla U(\tilde{\theta}(t))\right|^2\right] \leq cd\eta^7$. The second term on the right hand side in (23) can be bounded by L-smoothness of U in Condition H1 and Lemma 2.17 as

$$\mathbb{E}\left[\left|\nabla U(\tilde{\theta}(t))-\nabla U(\bar{\theta}(t))\right|^2\right] \leq L^2\mathbb{E}\left[\left|\bar{\theta}(t)-\tilde{\theta}(t)\right|^2\right] \leq C_1 d\eta^7,$$

where C_1 denotes a generic constant that depends only on γ, L and can change from line to line. Thus,

$$\mathbb{E}\left[\left|\bar{g}(t)-\nabla U(\bar{\theta}(t))\right|^2\right] \leq (C_1+c)d\eta^7. \tag{24}$$

We also know from Lemma 2.17 that

$$\mathbb{E}\left[|\tilde{v}_2(t)-\bar{v}_2(t)|^2\right] + \mathbb{E}\left[|\tilde{v}_3(t)-\bar{v}_3(t)|^2\right] \leq C_1 d\eta^7. \tag{25}$$

Per (22) and (24), (25), we arrive at $\mathbb{E}\left[\left|b(\bar{x}(t))-\bar{b}(t)\right|^2\right] \leq C_3 d\eta^7$. Then per (21), we have

$$\frac{d}{dt}\mathbb{E}\left[|x(t)-\bar{x}(t)|_M^2\right] \leq \left(-2\rho+\frac{2a}{\lambda_{\min,M}}\right)\mathbb{E}\left[|x(t)-\bar{x}(t)|_M^2\right] + C_3 d\eta^7. \tag{26}$$

Step 3: Let us rewrite (26) as

$$\frac{d\Delta}{dt}(t) \leq -h\Delta(t) + C_3 d\eta^7, \tag{27}$$

where $h := 2\rho - \frac{2a}{\lambda_{\min,M}} > 0$ and $\Delta(t) := \mathbb{E}\left[|x(t) - \bar{x}(t)|_M^2\right]$. We solve (27) by the integrating factor method. We integrate from $k\eta$ to t to obtain for $t \in [k\eta, (k+1)\eta]$,

$$\begin{aligned}\Delta(t) &= C_3 d e^{-ht} \int_{k\eta}^{t} e^{hs}(s-k\eta)^7 ds + e^{h(k\eta - t)}\Delta(k\eta) \\ &\leq C_3 d \int_{k\eta}^{t} (s-k\eta)^7 ds + e^{h(k\eta-t)}\Delta(k\eta) = \frac{C_3(t-k\eta)^8}{8} + e^{h(k\eta-t)}\Delta(k\eta).\end{aligned}$$

Therefore, we get

$$\Delta((k+1)\eta) \leq \frac{C_3 d}{8}\eta^8 + e^{-h\eta}\Delta(k\eta),$$

which leads to

$$\Delta((k+1)\eta) \leq \frac{C_3 d}{8}\eta^8 \sum_{j=0}^{k-1} e^{-jh\eta} + e^{-(k+1)h\eta}\Delta(0) \leq \frac{C_3 d}{8}\eta^8 \frac{1}{1-e^{-h\eta}} + e^{-kh\eta}\Delta(0).$$

Observe that when $\eta \leq \frac{1}{h}$, we have $\frac{1}{1-\frac{h\eta}{2}} \leq 2$ and hence $\frac{1}{1-e^{-h\eta}} \leq \frac{1}{h\eta(1-\frac{h\eta}{2})} \leq \frac{2}{h\eta}$. This implies

$$\Delta((k+1)\eta) \leq \frac{4C_3 d}{8h}\eta^7 + e^{-(k+1)h\eta}\Delta(0).$$

By the equivalence of norm relation (20), we further obtain

$$\lambda_{\min,M}\mathbb{E}\left[\left|x((k+1)\eta) - x^{(k)}\right|^2\right] \leq \frac{4C_3 d}{8h}\eta^7 + e^{-(k+1)h\eta}\lambda_{\min,M}\mathbb{E}\left[\left|x(0) - x^{(0)}\right|^2\right].$$

Now assume the continuous dynamics (10) is stationary and $x(0)$ is distributed as its invariant measure μ. Then $\mathbb{E}\left[\left|x(0) - x^{(0)}\right|^2\right] = \mathbb{E}\left[\left|Z - x^{(0)}\right|_M^2\right]$ where $Z \sim \mu$. We also know $\lambda_{\min,M}, \lambda_{\max,M}$ do not depend on d per Corollary A.5. Thus, we arrive at

$$\mathbb{E}\left[\left|x((k+1)\eta) - x^{(k)}\right|^2\right] \leq C_3 d\eta^7 + C_4 e^{-(k+1)h\eta}\mathbb{E}\left[\left|Z - x^{(0)}\right|^2\right],$$

where C_3, C_4 are positive constants that depend on γ, L, c but do not depend on d. Here we abuse notations and reuse C_3, C_4.

Finally, the fact that the constant $h := 2\rho - \frac{2a}{\lambda_{\min,M}} > 0$ depends on γ, L and does not depend on d is due to Theorem 2.1. This completes the proof. □

Proof of Theorem 2.9. Recall the basic fact about the 2-Wasserstein distance that

$$\mathrm{Wass}_2(\mathrm{Law}(X), \mathrm{Law}(Y)) \leq \mathbb{E}\left[|X-Y|^2\right]^{1/2}.$$

In view of Proposition 2.18, we can then derive the mixing time with respect to Wass_2 by solving for $C_3 d\eta^7 \leq \epsilon^2/2$ and $C_4 e^{-(k+1)h\eta}\mathbb{E}_{Z\sim\mu}\left[\left|Z - x^{(0)}\right|\right]^2 \leq \epsilon^2/2$. Solving for η in the first equation gives $\eta \leq \eta^* := \left(\frac{\epsilon^2}{2C_3 d}\right)^{1/7}$ Solving for k in the second equation gives $k \geq \log\left(\frac{2C_4\mathbb{E}_{Z\sim\mu}\left[\left|Z-x^{(0)}\right|^2\right]}{\epsilon^2}\right)\frac{1}{h\eta} - 1$. Plugging in the largest possible stepsize η^* into the right hand side of the previous inequality leads to the mixing time as claimed. □

2.2. P-th order Langevin Monte Carlo Algorithm for $P \geq 3$.

2.2.1. *P-th order Langevin Monte Carlo algorithm.* Given the iterate $x^{(k)}$, the next iterate $x^{(k+1)}$ is obtained by drawing from a multivariate normal distribution. The mean vector $\mathbf{M}(x^{(k)})$ and the covariance matrix $\boldsymbol{\Sigma}$ of this multivariate normal distribution are not provided explicitly, but their derivations are explained in the proof of Lemma C.2 for any order $P \geq 3$.

Below is the main result of this section. The proof is placed at the end of Section 2.2.3.

Theorem 2.19. *Assume $P \geq 3$ and Equation (28) satisfies Conditions H1 and H2. Let a be any positive constant satisfying*

$$a \leq \min\left\{\frac{m}{3\gamma}, \frac{\gamma\kappa}{6}\right\}\lambda_{\min,M},$$

where the positive definite matrix M and the constants γ, m, L, κ are from Section 2. Denote μ the invariant measure associated with the P-th order Langevin dynamics (5).

Let

$$\mathcal{R} = 4 \cdot \mathbb{1}_{\{P=3\}} + (2P-1)\cdot \mathbb{1}_{\{P\geq 4\}}.$$

Choose a 2-Wasserstein accuracy of ϵ small enough such that $\eta_0 := \left(\frac{\epsilon^2}{2\widetilde{C}_1 d}\right)^{1/\mathcal{R}} < \min\{\eta^{**}, \frac{1}{h}\}$ where h is defined in Proposition 2.26 and $\eta^{**}, \widetilde{C}_1$ are from Lemma C.4. Then $\mathrm{Wass}_2\big(\mathrm{Law}(x^{(k^*)}), \mu\big) \leq \epsilon$, where k^* is the mixing time of the P-th order Langevin Monte Carlo algorithm with respect to μ that is given by

$$k^* = \log\left(\frac{2\widetilde{C}_4 \mathbb{E}_{Z\sim\mu}\left[\left|Z - x^{(0)}\right|^2\right]}{\epsilon^2}\right)\frac{\left(2\widetilde{C}_3\right)^{1/\mathcal{R}}}{h}\frac{d^{1/\mathcal{R}}}{\epsilon^{1/(2\mathcal{R})}} - 1,$$

where $\widetilde{C}_3, \widetilde{C}_4$ are positive constants from Proposition 2.26 that depend on c, γ, L, P and do not depend on the dimension d.

Remark 2.20. In the cases $P = 3$ and $P = 4$, the results in Theorem 2.19 match, respectively, the result in [MMW+21] and the result in our Theorem 2.9.

2.2.2. *Derivation of the discretization scheme.* We generalize what was done in Section 2.1 to the P-th order Langevin dynamics which is

$$\begin{aligned} d\theta(t) &= v_1(t)dt, \\ dv_1(t) &= -\nabla U(\theta(t))dt + \gamma v_2(t)dt, \\ dv_n(t) &= -\gamma v_{n-1}(t)dt + \gamma v_{n+1}(t)dt, \quad 2 \leq n \leq P-2, \\ dv_{P-1}(t) &= -\gamma v_{P-2}(t)dt - \gamma v_{P-1}(t)dt + \sqrt{2\gamma}dB_t. \end{aligned} \tag{28}$$

Note that we can handle similar models with extra parameters as explained in Remark 2.11.

Let us describe the splitting scheme for any $P \geq 3$. We assume that we know $x^{(k)} = \left(\theta^{(k)}, v_1^{(k)}, \dots, v_{P-1}^{(k)}\right)$ and that is performing the $(k+1)$-th iterate of our algorithm.

Stage 1**:**

Set the initial value

$$x^{\mathrm{st}_1}(k\eta) := \big(\theta^{\mathrm{st}_1}(k\eta), v_1^{\mathrm{st}_1}(k\eta), \dots, v_{P-1}^{\mathrm{st}_1}(k\eta)\big) = x^{(k)}.$$

For $t \in (k\eta, (k+1)\eta]$, let

$$v_1^{\mathrm{st}_1}(t) = v_1^{(k)},$$

and

$$\begin{aligned} d\theta^{\mathrm{st}_1}(t) &= v_1^{\mathrm{st}_1}(t)dt, \\ dv_n^{\mathrm{st}_1}(t) &= -\gamma v_{n-1}^{\mathrm{st}_1}(t)dt + \gamma v_{n+1}^{(k)}dt, \quad 2 \leq n \leq P-2, \\ dv_{P-1}^{\mathrm{st}_1}(t) &= -\gamma v_{P-2}^{\mathrm{st}_1}(t)dt - \gamma v_{P-1}^{\mathrm{st}_1}(t)dt + \sqrt{2\gamma}dB_t. \end{aligned}$$

Stage j **for** $2 \leq j \leq P-1$**:**

Set the initial value

$$x^{\mathrm{st}_j}(k\eta) := \left(\theta^{\mathrm{st}_j}(k\eta), v_1^{\mathrm{st}_j}(k\eta), \dots, v_{P-1}^{\mathrm{st}_j}(k\eta)\right) = x^{(k)}.$$

For $t \in (k\eta, (k+1)\eta]$, let

$$dv_1^{\mathrm{st}_j}(t) = -g^{\mathrm{st}_j}(t)dt + \gamma v_2^{\mathrm{st}_{j-1}}(t)dt,$$

and

$$\begin{aligned} d\theta^{\mathrm{st}_j}(t) &= v_1^{\mathrm{st}_j}(t)dt, \\ dv_n^{\mathrm{st}_j}(t) &= -\gamma v_{n-1}^{\mathrm{st}_j}(t)dt + \gamma v_{n+1}^{\mathrm{st}_{j-1}}dt, \quad 2 \leq n \leq P-2, \\ dv_{P-1}^{\mathrm{st}_j}(t) &= -\gamma v_{P-2}^{\mathrm{st}_j}(t)dt - \gamma v_{P-1}^{\mathrm{st}_j}(t)dt + \sqrt{2\gamma}dB_t. \end{aligned}$$

Note that $g^{\mathrm{st}_j}(t)$ is a polynomial (in t) of degree $\alpha - 1$ and approximates $\nabla U(\theta^{\mathrm{st}_{j-1}}(t))$, and $g^{\mathrm{st}_j}(t)$ will be defined in (30) below.

Finally, we set

$$x^{(k+1)} = x^{\mathrm{st}_{P-1}}((k+1)\eta). \tag{29}$$

Definitions of $g^{\mathrm{st}_j}(t), 2 \leq j \leq P-1$**:** Here we define the polynomials $g^{\mathrm{st}_j}(t)$ which approximate $\nabla U(\theta^{\mathrm{st}_{j-1}}(t))$.

Recall the Taylor polynomial $P_{\alpha-1}$ of degree $\alpha - 1$ centering at 0 and associated with ∇U in (12). Inductively for $2 \leq j \leq P-1$, let us set

$$g^{\mathrm{st}_j}(t) := P_{\alpha-1}\big(\theta^{\mathrm{st}_{j-1}}(t)\big). \tag{30}$$

Remark 2.21. In the case $P = 3$, the above splitting scheme is fairly similar to the splitting scheme in [MMW+21]. The only notable difference is that we employ Taylor polynomial approximation while the authors of [MMW+21] employ Lagrange polynomial interpolation. The necessity of this difference has been discussed in Remark 2.14 and in Appendix D.

Remark 2.22. The general idea of our numerical scheme is that there are several stages of refinement. At every stage, the variable v_1 which contains the non-linear term $\nabla U(\theta)$ is split from the other variables and approximated first, while the vector formed by the remaining variables is approximated by a multivariate Ornstein-Uhlenbeck process. As a result, the discretization procedure only works for $P \geq 3$.

Moreover, there are $P-1$ stages in the above discretization procedure and we cannot add another one to it. The reason is similar to the one given in Remark 2.15.

2.2.3. *Proofs.* The following result is similar to Lemma 2.16. The proof is simple and is therefore omitted.

Lemma 2.23. *Denote $L_\alpha := \sup_{x\in\mathbb{R}^d} \|\nabla^\alpha U(x)\|_{\mathrm{op}}$. Under Conditions H1, it holds for $1 \leq j \leq P-1$ that*

$$\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\nabla U(\theta^{\mathrm{st}_{j-1}}(t)) - g^{\mathrm{st}_j}(t)\right|^2\right] \leq \left(\frac{L_\alpha}{\alpha!}\right)^2 \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\theta^{\mathrm{st}_{j-1}}(t)\right|^{2\alpha}\right].$$

The next two results bound the differences in L^2-norm of variables of two consecutive stages. Lemma 2.25 is a consequence of Lemma 2.24. Their proofs are deferred to near the end of Appendix C.

Lemma 2.24. *Assume $P \geq 3$ and consider the splitting scheme at the beginning of this section with $P-1$ stages. Denote the stages by $j, 1 \leq j \leq P-1$. It holds that*

- *for $j = 1$:* $\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_n^{\mathrm{st}_1}(t) - v_n^{(k)}\right|^2\right] \leq C_n^{\mathrm{st}_1} d\eta^2, 1 \leq n \leq P-2;$
  $\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_{P-1}^{\mathrm{st}_1}(t) - v_{P-1}^{(k)}\right|^2\right] \leq C_{P-1}^{\mathrm{st}_1} d\eta$ *and*
  $\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\theta^{\mathrm{st}_1}(t) - \theta^{(k)}\right|^2\right] \leq C_P^{\mathrm{st}_1} d\eta^2.$
- *for $j = 2$*

  and $P = 3$: $\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{\mathrm{st}_2}(t) - v_1^{\mathrm{st}_1}(t)\right|^2\right] \leq C_1^{\mathrm{st}_2} d\eta^2,$
  $\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_2^{\mathrm{st}_2}(t) - v_2^{\mathrm{st}_1}(t)\right|^2\right] \leq C_2^{\mathrm{st}_2} d\eta^2$ *and*
  $\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\theta^{\mathrm{st}_2}(t) - \theta^{\mathrm{st}_1}(t)\right|^2\right] \leq C_3^{\mathrm{st}_2} d\eta^4.$

  and $P \geq 4$: $\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{\mathrm{st}_2}(t) - v_1^{\mathrm{st}_1}(t)\right|^2\right] \leq C_1^{\mathrm{st}_2} d\eta^2;$
  $\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_n^{\mathrm{st}_2}(t) - v_n^{\mathrm{st}_1}(t)\right|^2\right] \leq C_n^{\mathrm{st}_2} d\eta^4, 2 \leq n \leq P-3;$

$\sup_{t\in(k\eta,(k+1)\eta]}\mathbb{E}\left[\left|v_{P-2}^{\mathrm{st}_2}(t)-v_{P-2}^{\mathrm{st}_1}(t)\right|^2\right]\le C_{P-2}^{\mathrm{st}_2}d\eta^3,$
$\sup_{t\in(k\eta,(k+1)\eta]}\mathbb{E}\left[\left|v_{P-1}^{\mathrm{st}_2}(t)-v_{P-1}^{\mathrm{st}_1}(t)\right|^2\right]\le C_{P-1}^{\mathrm{st}_2}d\eta^5$ *and*
$\sup_{t\in(k\eta,(k+1)\eta]}\mathbb{E}\left[\left|\theta^{\mathrm{st}_2}(t)-\theta^{\mathrm{st}_1}(t)\right|^2\right]\le C_{P}^{\mathrm{st}_2}d\eta^4.$

Here, $\left\{C_n^{\mathrm{st}_j}: j=1 \text{ or } j=2, 1\le n\le P\right\}$ are constants that depend on c,γ,L,P but do not depend on the dimension d.

Lemma 2.25. *Assume $P\ge 3$ and consider the splitting scheme at the beginning of this section with $P-1$ stages. Denote the stages by $j, 1\le j\le P-1$. It holds for $j\ge 3$*

a) and $1\le n\le P-j-1$: $\sup_{t\in(k\eta,(k+1)\eta]}\mathbb{E}\left[\left|v_n^{\mathrm{st}_j}(t)-v_n^{\mathrm{st}_{j-1}}(t)\right|^2\right]\le C_n^{\mathrm{st}_j}d\eta^{2j}$;

b) and $n=P-j$: $\sup_{t\in(k\eta,(k+1)\eta]}\mathbb{E}\left[\left|v_{P-j}^{\mathrm{st}_j}(t)-v_n^{\mathrm{st}_{j-1}}(t)\right|^2\right]\le C_{P-j}^{\mathrm{st}_j}d\eta^{2j-1}$;

*c) and $P-j+1\le n\le P-1$: $\sup_{t\in(k\eta,(k+1)\eta]}\mathbb{E}\left[\left|v_n^{\mathrm{st}_j}(t)-v_n^{\mathrm{st}_{j-1}}(t)\right|^2\right]$
$\le C_n^{\mathrm{st}_j}d\eta^{4j+2n-2P-1}$;*

d) $\sup_{t\in(k\eta,(k+1)\eta]}\mathbb{E}\left[\left|\theta^{\mathrm{st}_j}(t)-\theta^{\mathrm{st}_{j-1}}(t)\right|^2\right]\le C_P^{\mathrm{st}_j}d\eta^{2j+2}$.

$\left\{C_n^{\mathrm{st}_j}: 3\le j\le P-1, 1\le n\le P\right\}$ are constants that depend on c,γ,L,P but do not depend on dimension d.

Consequently, it holds for the last Stage $P-1$ that

$$\sup_{t\in(k\eta,(k+1)\eta]}\left(\mathbb{E}\left[\left|v_2^{\mathrm{st}_2}(t)-v_2^{\mathrm{st}_1}(t)\right|^2\right]+\mathbb{E}\left[\left|\theta^{\mathrm{st}_2}(t)-\theta^{\mathrm{st}_1}(t)\right|^2\right]\right)\le\left(C_2^{\mathrm{st}_2}+C_3^{\mathrm{st}_2}\right)d\eta^4,\quad P=3; \tag{31}$$

and

$$\sup_{t\in(k\eta,(k+1)\eta]}\left(\sum_{n=2}^{P-1}\mathbb{E}\left[\left|v_n^{\mathrm{st}_{P-1}}(t)-v_n^{\mathrm{st}_{P-2}}(t)\right|^2\right]+\mathbb{E}\left[\left|\theta^{\mathrm{st}_{P-1}}(t)-\theta^{\mathrm{st}_{P-2}}(t)\right|^2\right]\right)$$
$$\le\left(\sum_{n=2}^{P}C_n^{\mathrm{st}_{P-1}}\right)d\eta^{2P-1},\quad P\ge 4. \tag{32}$$

Proposition 2.26. *Assume Equation* (28) *satisfies Conditions H1 and H2. Let a be any positive constant satisfying*

$$a\le\min\left\{\frac{m}{3\gamma},\frac{\gamma\kappa}{6}\right\}\lambda_{\min,M},$$

where the positive definite matrix M and the constants γ,m,L,κ are from Section 2. Denote μ the invariant measure of the P-th order Langevin dynamics (28)*.*

Assume further that $\eta < \{\eta^{**}, \frac{1}{h}\}$. Then regarding the discretization error, it holds when and $P \geq 4$ that

$$\mathbb{E}\Big[\big|x((k+1)\eta) - x^{(k+1)}\big|^2\Big] \leq \widetilde{C}_3 d\eta^{2P-1} + \widetilde{C}_4 e^{-(k+1)h\eta}\mathbb{E}\Big[\big|Z - x^{(0)}\big|^2\Big], \qquad Z \sim \mu,$$

and when $P = 3$ that

$$\mathbb{E}\Big[\big|x((k+1)\eta) - x^{(k+1)}\big|^2\Big] \leq \widetilde{C}_3 d\eta^{4} + \widetilde{C}_4 e^{-(k+1)h\eta}\mathbb{E}\Big[\big|Z - x^{(0)}\big|^2\Big], \qquad Z \sim \mu.$$

In particular, η^{**} is defined at (73), $h := 2\rho - \frac{2a}{\lambda_{\min,M}}$ where ρ, M are from Theorem 2.1 and a is any positive constant equal to or less than $\min\left\{\frac{m}{3\gamma}, \frac{\gamma\kappa}{6}\right\}\lambda_{\min,M}$. Moreover, $\widetilde{C}_3$ and $\widetilde{C}_4$ are constants that depend on γ, L, c but not on the dimension d.

Proof. We will follow the argument in the proof of Proposition 2.18.

Step 1: Let us write

$$dx^{\mathrm{st}_{P-1}} = \bar{b}(t)dt + \sqrt{2\gamma}DdB_t,$$

where $\bar{b}(t) \in \mathcal{M}_{P\times 1}$ and $D \in \mathcal{M}_{Pd\times Pd}$ are respectively given by:

$$\bar{b}(t) := \begin{pmatrix} v_1^{\mathrm{st}_{P-1}}(t) \\ -g^{\mathrm{st}_{P-1}}(t) + \gamma v_2^{\mathrm{st}_{P-2}}(t) \\ -\gamma v_1^{\mathrm{st}_{P-1}}(t) + \gamma v_3^{\mathrm{st}_{P-2}}(t) \\ \vdots \\ -\gamma v_{P-3}^{\mathrm{st}_{P-1}}(t) + \gamma v_{P-1}^{\mathrm{st}_{P-2}}(t) \\ -\gamma v_{P-2}^{\mathrm{st}_{P-1}}(t) - \gamma v_{P-1}^{\mathrm{st}_{P-1}}(t) \end{pmatrix}, \quad D := \begin{pmatrix} 0_d & \cdots & \cdots & 0_d \\ \vdots & \ddots & & \vdots \\ \vdots & & 0_d & 0_d \\ 0_d & \cdots & 0_d & I_d \end{pmatrix}.$$

Meanwhile, the P-th order Langevin dynamics (28) can be written as

$$dx(t) = b(x(t))dt + \sqrt{2\gamma}DdB_t, \qquad b(x) = \begin{pmatrix} v_1(t) \\ -\nabla U(\theta(t)) + \gamma v_2(t) \\ -\gamma v_1(t) + \gamma v_3(t) \\ \vdots \\ -\gamma v_{P-3}(t) + \gamma v_{P-1}(t) \\ -\gamma v_{P-2}(t) - \gamma v_{P-1}(t) \end{pmatrix}.$$

Then

$$d\big(x(t) - x^{\mathrm{st}_{P-1}}(t)\big) = \big(b(x(t)) - b\big(x^{\mathrm{st}_{P-1}}\big)\big)dt + \big(b\big(x^{\mathrm{st}_{P-1}}(t)\big) - \bar{b}(t)\big)dt,$$

where

$$b\big(x^{\mathrm{st}_{P-1}}(t)\big) - \bar{b}(t) = \begin{pmatrix} 0 \\ (g^{\mathrm{st}_{P-1}}(t) - \nabla U(\theta^{\mathrm{st}_{P-1}}(t))) + \gamma\Big(v_2^{\mathrm{st}_{P-1}}(t) - v_2^{\mathrm{st}_{P-2}}(t)\Big) \\ \gamma\Big(v_3^{\mathrm{st}_{P-1}}(t) - v_3^{\mathrm{st}_{P-2}}(t)\Big) \\ \vdots \\ \gamma\Big(v_{P-1}^{\mathrm{st}_{P-1}}(t) - v_{P-1}^{\mathrm{st}_{P-2}}(t)\Big) \\ 0 \end{pmatrix}. \tag{33}$$

This leads to

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\Big[\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)^{\top}M\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)\Big]\\
&=\mathbb{E}\Big[\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)^{\top}M\big(b(x(t))-b(x^{\mathrm{st}_{P-1}}(t))\big)\Big]\\
&\quad+\mathbb{E}\Big[\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)^{\top}M\big(b(x^{\mathrm{st}_{P-1}}(t))-\bar{b}(t)\big)\Big]\\
&\quad\quad+\mathbb{E}\Big[\big(b(x(t))-b(x^{\mathrm{st}_{P-1}}(t))\big)^{\top}M\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)\Big]\\
&\quad\quad\quad+\mathbb{E}\Big[\big(b(x^{\mathrm{st}_{P-1}}(t))-\bar{b}(t)\big)^{\top}M\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)\Big].
\end{aligned}\tag{34}$$

Notice that

$$\begin{aligned}
&\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)^{\top}M\big(b(x(t))-b(x^{\mathrm{st}_{P-1}}(t)\big)+\big(b(x(t))-b(x^{\mathrm{st}_{P-1}}(t))\big)^{\top}M(x(t)-\bar{x}(t))\\
&\le (x(t)-\bar{x}(t))^{\top}M\int_0^1 J_b(wx(t)+(1-w)\bar{x}(t))\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)dw\\
&+\int_0^1\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)^{\top}J_b(wx(t)+(1-w)\bar{x}(t))^{\top}dwM\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)\\
&\le \big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)^{\top}(-2\rho)M\big(x(t)-x^{\mathrm{st}_{P-1}}(t)\big)=-2\rho\left|x(t)-\bar{x}(t)\right|_M^2,
\end{aligned}\tag{35}$$

where the last line is due to the contraction property (8) in Theorem 2.1.

Moreover, choose any positive $a\le\min\left\{\frac{m}{3\gamma},\frac{\gamma\kappa}{6}\right\}\lambda_{\min,M}$ and notice that per Theorem 2.1,

$$-\rho+\frac{a}{\lambda_{\min,M}}<0.\tag{36}$$

From (34), (35), (36) and Cauchy-Schwarz inequality, we can deduce that

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\Big[\left|x(t)-x^{\mathrm{st}_{P-1}}(t)\right|_M^2\Big]&\le-2\rho\mathbb{E}\Big[\left|x(t)-x^{\mathrm{st}_{P-1}}(t)\right|_M^2+2a\left|x(t)-x^{\mathrm{st}_{P-1}}(t)\right|^2\Big]\\
&\quad+\frac{2}{a}\left\|M\right\|_{\mathrm{op}}^2\left(\gamma^2+1\right)\mathbb{E}\Big[\left|b(x^{\mathrm{st}_{P-1}}(t))-\bar{b}(t)\right|^2\Big].
\end{aligned}\tag{37}$$

By combining the bound in (37) with (20), we get

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\Big[\left|x(t)-x^{\mathrm{st}_{P-1}}(t)\right|_M^2\Big]\\
&\le\left(-2\rho+\frac{2a}{\lambda_{\min,M}}\right)\mathbb{E}\Big[\left|x(t)-x^{\mathrm{st}_{P-1}}(t)\right|_M^2\Big]+\frac{2}{a}\left\|M\right\|_{\mathrm{op}}^2\left(\gamma^2+1\right)\mathbb{E}\Big[\left|b(x^{\mathrm{st}_{P-1}}(t))-\bar{b}(t)\right|^2\Big].
\end{aligned}\tag{38}$$

Step 2: We will bound $\mathbb{E}\Big[\left|b(x^{\mathrm{st}_{P-1}}(t))-\bar{b}(t)\right|^2\Big]$ as the second term on the right hand side of (38). Based on (33), we will need to bound the L^2 norm of

$$g^{\mathrm{st}_{P-1}}(t)-\nabla U\big(\theta^{\mathrm{st}_{P-1}}(t)\big),\ \gamma\Big(v_2^{\mathrm{st}_{P-1}}(t)-v_2^{\mathrm{st}_{P-2}}(t)\Big),\ \gamma\Big(v_{P-1}^{\mathrm{st}_{P-1}}(t)-v_{P-1}^{\mathrm{st}_{P-2}}(t)\Big).$$

Per estimate (32) Lemma 2.25 and in the case $P \geq 4$ (the case $P = 3$ is similar and is handled at the end of this proof), we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\gamma\left(v_2^{\mathrm{st}_{P-1}}(t) - v_2^{\mathrm{st}_{P-2}}(t)\right)\right|^2\right] + \mathbb{E}\left[\left|\gamma\left(v_{P-1}^{\mathrm{st}_{P-1}}(t) - v_{P-1}^{\mathrm{st}_{P-2}}(t)\right)\right|^2\right] \\
&\leq \gamma^2\left(C_2^{\mathrm{st}_{P-1}} + C_{P-1}^{\mathrm{st}_{P-1}}\right) d\eta^{2P-1}. \qquad (39)
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
&\mathbb{E}\left[\left|g^{\mathrm{st}_{P-1}}(t) - \nabla U\big(\theta^{\mathrm{st}_{P-1}}(t)\big)\right|^2\right] \\
&\leq 2\mathbb{E}\left[\left|g^{\mathrm{st}_{P-1}}(t) - \nabla U\big(\theta^{\mathrm{st}_{P-2}}(t)\big)\right|^2\right] + 2\mathbb{E}\left[\left|\nabla U\big(\theta^{\mathrm{st}_{P-2}}(t)\big) - \nabla U\big(\theta^{\mathrm{st}_{P-1}}(t)\big)\right|^2\right].
\end{aligned}$$

The same argument as the one in (59) yields

$$\mathbb{E}\left[\left|g^{\mathrm{st}_{P-1}}(t) - \nabla U\big(\theta^{\mathrm{st}_{P-2}}(t)\big)\right|^2\right] \leq c d\eta^{2P-1},$$

for some positive constant c. Moreover, L-smoothness of U in Condition H1 and estimate (32) of Lemma 2.25 imply

$$\mathbb{E}\left[\left|\nabla U\big(\theta^{\mathrm{st}_{P-2}}(t)\big) - \nabla U\big(\theta^{\mathrm{st}_{P-1}}(t)\big)\right|^2\right] \leq L^2 D^{\mathrm{st}_{P-1}} d\eta^{2P-1}.$$

The last three bounds lead to

$$\mathbb{E}\left[\left|g^{\mathrm{st}_{P-1}}(t) - \nabla U\big(\theta^{\mathrm{st}_{P-1}}(t)\big)\right|^2\right] \leq \big(L^2 D^{\mathrm{st}_{P-1}} + 1\big) d\eta^{2P-1}. \qquad (40)$$

The combination of (33), (39) and (40) lead to

$$\mathbb{E}\left[\left|b(x^{\mathrm{st}_{P-1}}(t)) - \bar{b}(t)\right|^2\right] \leq \widetilde{C}_1 d\eta^{2P-1}, \quad P \geq 4,$$

where $\widetilde{C}_1$ is a generic positive constant that depends only on c, γ, L, P and can change from line to line. Then from (38), we get

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left[\left|x(t) - x^{\mathrm{st}_{P-1}}(t)\right|_M^2\right] \\
&\leq \left(-2\rho + \frac{2a}{\lambda_{\min,M}}\right)\mathbb{E}\left[\left|x(t) - x^{\mathrm{st}_{P-1}}(t)\right|_M^2\right] + \frac{2}{a}\left\|M\right\|_{\mathrm{op}}^2\left(\gamma^2+1\right)\widetilde{C}_1 d\eta^{2P-1}, \quad P \geq 4. \qquad (41)
\end{aligned}$$

Step 3: Solving the differential inequality (41) by integrating factors as in the proof of Proposition 2.18, we arrive at the desired discretization error in the case $P \geq 4$.

Step 4: As the last part of this proof and in the case $P = 3$, we follow a similar path and use estimate (31) in Lemma 2.25 (instead of (32)) to get an analogous inequality to (41) that is

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left[\left|x(t) - x^{\mathrm{st}_{P-1}}(t)\right|_M^2\right] \\
&\leq \left(-2\rho + \frac{2a}{\lambda_{\min,M}}\right)\mathbb{E}\left[\left|x(t) - x^{\mathrm{st}_2}(t)\right|_M^2\right] + \frac{2}{a}\left\|M\right\|_{\mathrm{op}}^2\left(\gamma^2+1\right)\widetilde{C}_1 d\eta^4. \qquad (42)
\end{aligned}$$

Then by solving the differential inequality (42) by integrating factors as in the proof of Proposition 2.18, we arrive at the desired discretization error in the case $P = 3$. The proof is complete. □

Proof of Theorem 2.19. The argument is the same as the proof of Theorem 2.9 at the end of Section 2.1.3. We have $\mathrm{Wass}_2(\mathrm{Law}(X), \mathrm{Law}(Y)) \leq \mathbb{E}\big[|X - Y|^2\big]^{1/2}$. Then per Proposition 2.26, we can solve for η in $\widetilde{C}_3 d\eta^{4\mathbb{1}_{\{P=3\}}+(2P-1)\mathbb{1}_{\{P\geq 4\}}} \leq \epsilon^2/2$ and for k in $\widetilde{C}_4 e^{-(k+1)h\eta}\mathbb{E}_{Z\sim\mu}\left[\left|Z - x^{(0)}\right|^2\right] \leq \epsilon^2/2$ to obtain the desired mixing time. This completes the proof. □

3. Numerical Experiments

In this section, we will implement both third-order and fourth-order LMC algorithms. From Section 2.1, we recall the fourth-order LMC algorithm samples a multivariate normal distribution at every step, where mean and covariance are provided in Lemma B.1 and Lemma B.2 in Appendix B. The mean in particular contains several nested integrals that need to be exactly computed when the loss function U is a polynomial, and approximated in the case where the loss function U is not a polynomial. We provide the calculations related to these nested integrals for quadratic loss and logistic loss in Appendix E, which allow us to perform the numerical experiments for our fourth-order LMC algorithm.

In addition, we will provide some calculations necessary to perform the numerical experiments for the third-order LMC algorithm in [MMW+21] for quadratic loss.

When the potential function $U(\theta)$ satisfies Condition H1, then for a small stepsize $\eta > 0$ and two arbitrary friction parameters $\gamma > 0$ and $\xi > 0$, the third-order Langevin Monte Carlo algorithm is given as follows:

Algorithm 1: Third-Order Langevin Monte Carlo Algorithm

Let $x^{(0)} = \left(\theta^{(0)}, v_1^{(0)}, v_2^{(0)}\right) = (\theta^*, 0, 0)$
for $k = 0, 1, \cdots, N-1$ **do**
  Sample $x^{(k+1)} \sim \mathcal{N}\left(\boldsymbol{\mu}(x^{(k)}), \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are defined in the following equations
end for

The update of the states x from step k to $k+1$ is obtained by drawing from the distribution with mean $\boldsymbol{\mu}\left(x^{(k)}\right)$ and covariance $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu}(x) := \begin{pmatrix} \theta - \frac{\eta}{2L}\Delta U(\theta, v_1) + \mu_{12}v_1 + \mu_{13}v_2 \\ -\frac{1}{L}\Delta U(\theta, v_1) + \mu_{22}v_1 + \mu_{23}v_2 \\ \frac{\mu_{31}}{L}\Delta U(\theta, v_1) + \mu_{32}v_1 + \mu_{33}v_2 \end{pmatrix}, \boldsymbol{\Sigma} := \begin{pmatrix} \sigma_{11}\cdot I_d & \sigma_{12}\cdot I_d & \sigma_{13}\cdot I_d \\ \sigma_{12}\cdot I_d & \sigma_{22}\cdot I_d & \sigma_{23}\cdot I_d \\ \sigma_{13}\cdot I_d & \sigma_{23}\cdot I_d & \sigma_{33}\cdot I_d \end{pmatrix}, \quad (43)$$

where all μ's and σ's are defined in the article by [MMW+21]. Now we present the fourth-order Langevin Monte Carlo algorithm as follows:

Algorithm 2: Fourth-Order Langevin Monte Carlo Algorithm

Let $x^{(0)} = \left(\theta^{(0)}, v_1^{(0)}, v_2^{(0)}, v_3^{(0)}\right) = (\theta^*, 0, 0, 0)$
for $k = 0, 1, \cdots, N-1$ **do**
  Sample $x^{(k+1)} \sim \mathcal{N}\left(\mathbf{m}(x^{(k)}), \mathbf{\Sigma}\right)$, where $\mathbf{m}$ and $\mathbf{\Sigma}$ are defined in the following equation (44)
end for

The update of the state x from step k to $k+1$ is obtained by drawing a sample from the multivariate Gaussian distribution with mean $\mathbf{m}(x)$ and covariance $\mathbf{\Sigma}$ given by:

$$\mathbf{m}(x) := \begin{pmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \end{pmatrix}, \qquad \mathbf{\Sigma} := \begin{pmatrix} \sigma_{00} \cdot I_d & \sigma_{01} \cdot I_d & \sigma_{02} \cdot I_d & \sigma_{03} \cdot I_d \\ \sigma_{01} \cdot I_d & \sigma_{11} \cdot I_d & \sigma_{12} \cdot I_d & \sigma_{13} \cdot I_d \\ \sigma_{02} \cdot I_d & \sigma_{12} \cdot I_d & \sigma_{22} \cdot I_d & \sigma_{23} \cdot I_d \\ \sigma_{03} \cdot I_d & \sigma_{13} \cdot I_d & \sigma_{23} \cdot I_d & \sigma_{33} \cdot I_d \end{pmatrix}, \tag{44}$$

where the explicit formulas to compute $m_i \in \mathbb{R}^d$ and $\sigma_{ij} \in \mathbb{R}$ are given in Lemma B.2 and the calculations for particular loss functions are given in Appendix E. Note that both algorithms require the initialization of the model parameter θ^*. [MMW+21] recommended that θ^* can be chosen from the exact solution when U is a polynomial. However, we initialize the sampling process randomly from the standard normal distribution, which leads to superior performance.

3.1. **Bayesian linear regression.** We conduct experiments using our algorithms for Bayesian linear regression-type problems using the **Air Quality** data from the *UCI Machine Learning Repository* [Vit08]. It contains sensor readings from an array of chemical sensors deployed in an Italian city to monitor air pollution. Collected between March 2004 and February 2005. The dataset includes hourly measurements of key pollutants such as carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene, nitrogen oxides (NO_x), and ozone (O_3), along with meteorological variables such as temperature and relative humidity. The dataset is often used for regression tasks to model air quality indicators, particularly predicting CO concentration based on other environmental variables. It presents challenges such as missing values and sensor drift, making it suitable for testing robust data pre-processing and modeling techniques.

In this experiment, our goal is to sample the posterior distribution of the model parameters that regress the concentration of CO present in the air. After pre-processing. The feature matrix has $d = 16$ dimensions (including the intercept term) and a total of 7,674 observations.

We consider an arbitrary prior of θ from $\mathcal{N}(0, 10I)$. The known posterior for the linear regression problem is given as follows:

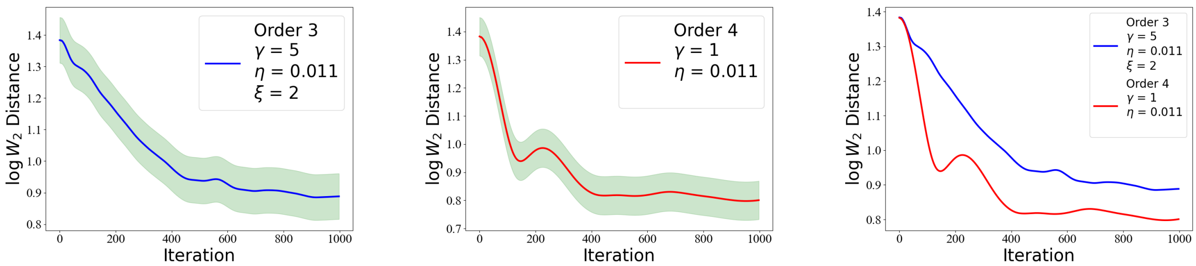
$$\mu(\theta) \sim \mathcal{N}(\mathbf{m}, \mathbf{V}); \quad \mathbf{m} := \left(\Sigma^{-1} + \frac{X^\top X}{\xi^2}\right)^{-1} \left(\frac{X^\top y}{\xi^2}\right), \quad \mathbf{V} := \left(\mathbf{\Sigma}^{-1} + \frac{X^\top X}{\xi^2}\right)^{-1}, \tag{45}$$

where X and y are input data-matrix and output vector, respectively, and $\mathbf{\Sigma} = \lambda I_d$ is the covariance matrix with the Ridge regularization (L_2) parameter λ, in this experiment, we choose a smaller penalty $\lambda = 2$.

To draw a sample from the posterior, at each iteration we perform a Cholesky decomposition to factor the covariance matrix into a lower and upper triangular matrix $\mathbf{\Sigma} = LL^\top$ and use the formula [WL06]:

$$x^{(k+1)} = \boldsymbol{\mu}\left(x^{(k)}\right) + Lu; \quad \text{or} \quad x^{(k+1)} = \mathbf{m}\left(x^{(k)}\right) + Lu,$$

where $u \in \mathbb{R}^{3d}$ (or $\mathbb{R}^{4d}$) with $u \sim \mathcal{N}(0, I)$ and $\boldsymbol{\mu}\left(x^{(k)}\right)$ (or $\mathbf{m}\left(x^{(k)}\right)$) is the mean vector at k-th iterate of the respective algorithm. Note that the covariance matrix $\mathbf{\Sigma}$ needs to be symmetric positive definite (SPD) in order to factor it using Cholesky decomposition. To ensure that we get an SPD matrix for arbitrarily chosen γ, ξ, and η values, we add a small jitter (10^{-6}) to the covariance matrix. Then we perform a grid search to find the optimal hyperparameters based on the lowest mean 2-Wasserstein distance, computed using the formula from [GS84].



a. $\mathcal{W}_2$ distance from the 3rd-order LMC

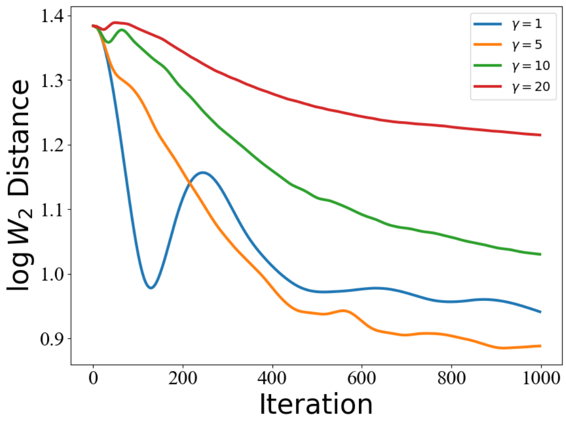
b. $\mathcal{W}_2$ distance from the 4th-order LMC

c. $\mathcal{W}_2$ distances from the 3rd- and 4th- order LMC

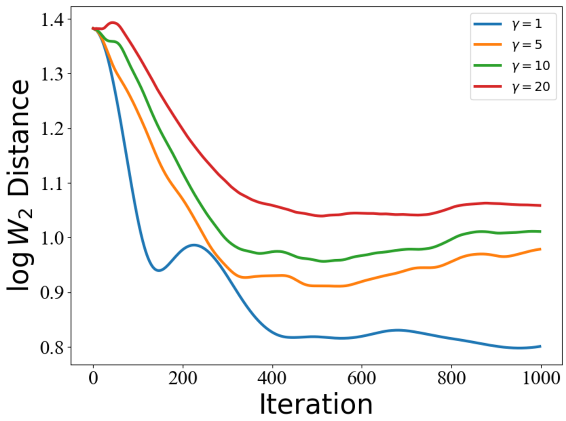
FIGURE 1. Comparative performance of the 3rd- and 4th-order Langevin Monte Carlo algorithms

The tuned hyperparameters for the third-order Langevin dynamics $\gamma = 5, \eta = 0.011$, and $\xi = 2$, and for the fourth-order Langevin dynamics $\gamma = 1$ and $\eta = 0.011$. For both dynamics, we draw $N = 1,000$ samples from the posterior distribution and compute the $\mathcal{W}_2$ (*2-Wasserstein*) distance from the known posterior defined in (45). The shaded region both in Figure 1a. and 1b. represent half of the standard deviation in *2-Wasserstein* distances. The relative performances of the third- and fourth-order LMC algorithms are presented in Figure 1c. From this set of experiments, we notice that the convergence to the posterior distribution is better for the 4th-order LMC algorithm than that of the 3rd-order LMC algorithm for a given stepsize η.

Next, we present the effect of the variation of the friction parameter γ in Figure 2. For the same stepsize η, we observe that the 4th-order LMC algorithm provides better convergence



a. Change in $\mathcal{W}_2$ distance from the 3rd-order LMC algorithm for varying γ



b. Change in $\mathcal{W}_2$ distance from the 4th-order LMC algorithm for varying γ

FIGURE 2. Comparative performance of the 3rd- and 4th-order Langevin Monte Carlo algorithms for the same stepsize η and varying the friction parameter γ

in terms of the *2-Wasserstein* distance for smaller γ values. However, this is not always the case for the 3rd-order LMC algorithm.

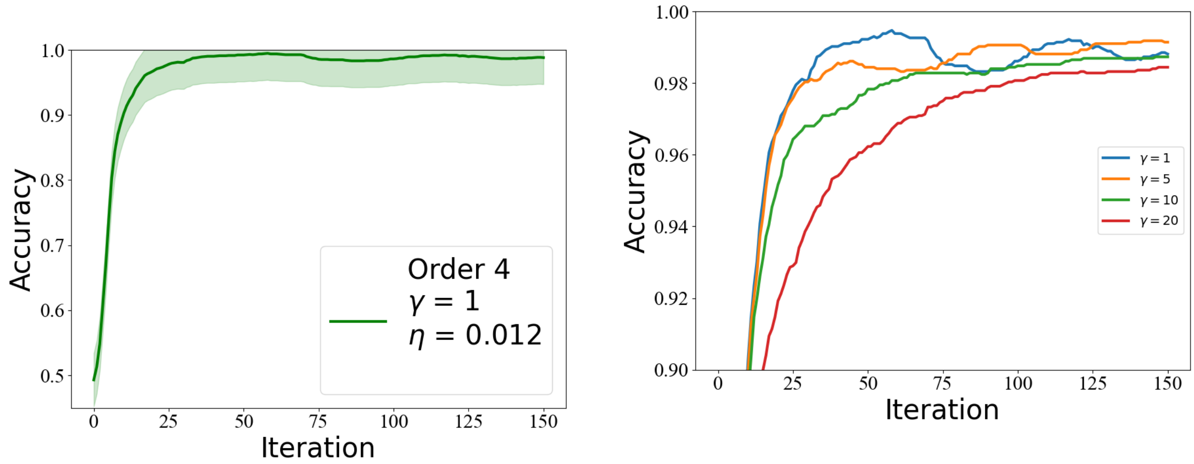
3.2. **Bayesian logistic regression.** In this section, we provide the implementation of the 4th-order LMC algorithm for sampling in a classification problem. To implement the 4th-order LMC algorithm efficiently, one needs to approximate the gradient of the potential function in higher-degree polynomials. The last step can be done via softwares per Remark 2.12; however this complicates the implementation of our algorithm. Therefore, we arbitrarily choose third-degree polynomials to approximate the gradient of the logistic loss function using a Taylor polynomial which is given in Appendix E.2.

We choose the **Mushroom** dataset from the *UCI Machine Learning Repository* [2]. The dataset contains 8,124 instances of gilled mushrooms, each described by 22 categorical features such as cap shape, odor, gill color, and habitat. After pre-processing (e.g., OneHotEncoding for categorical features). The final input dataset has a dimension $d = 118$, and a total of 8,124 observations.

The typical objective with this data is to build a machine learning model that classifies whether a mushroom is edible or poisonous based on the given attributes. However, our goal in this experiment is not to find the optimal model; rather, we sample the model parameters and see how the accuracy measure varies as we increase the number of samples from the posterior distribution of the model parameters.

Before starting the sampling process, we split the data into 70-30 training and testing ratio and check the sample quality on the test set. We generate $N = 150$ samples of

[2]Mushroom. UCI Machine Learning Repository, 1981. DOI: https://doi.org/10.24432/C5959T



a. Accuracy from the 4th-order LMC algorithm

b. Accuracy for different γ values (for the a fixed step-size $\eta = 0.012$)

FIGURE 3. Performance of the 4th-order LMC algorithm in sampling from a non-polynomial potential function

the model parameters and run a grid search for the hyperparameters η and γ. To avoid overfitting, we use a larger penalty $\lambda = 25$.

From Figure 3 (a), we see that the 4th-order LMC algorithm performs very well even for smaller degree polynomial approximation of the gradient of the potential function. We tune the model parameters $\eta = 0.012$ and $\gamma = 1$. Then we show the effect of the variation in the friction parameter for a chosen stepsize $\eta = 0.012$ in Figure 3 (b). We see that smaller γ values result in better performance in terms of higher accuracy.

4. Conclusion

In this paper, we proposed P-th order Langevin Monte Carlo algorithms based on the discretizations of P-th order Langevin dynamics for any $P \geq 3$. We designed discretization schemes based on splitting and accurate integration methods. When the density of the target distribution is log-concave and smooth, we obtained Wasserstein convergence guarantees that lead to better iteration complexities. Specifically, the mixing time of the P-th order LMC algorithm scales as $O\left(d^{\frac{1}{\mathcal{R}}}/\epsilon^{\frac{1}{2\mathcal{R}}}\right)$ for $\mathcal{R} = 4\cdot\mathbb{1}_{\{P=3\}}+(2P-1)\cdot\mathbb{1}_{\{P\geq 4\}}$, which has a better dependence on the dimension d and the accuracy level ϵ as P grows. Numerical experiments were conducted to illustrate the efficiency of our proposed algorithms.

Acknowledgments

Mert Gürbüzbalaban's research is supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485. Mohammad Rafiqul Islam is partially supported by the grant NSF DMS-2053454. Nian Yao was supported in part by the Natural Science Foundation of China under Grant 12071361, the Natural Science Foundation of Guangdong Province under Grant 2020A1515010822 and Shenzhen Natural Science Fund (the Stable Support Plan Program 20220810152104001). Lingjiong Zhu is partially supported by the grants NSF DMS-2053454 and DMS-2208303.

References

[ADFDJ03] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43, 2003.

[AE14] Anton Arnold and Jan Erb. Sharp entropy decay for hypocoercive and non-symmetric Fokker-Planck equations with linear drift. *arXiv preprint arXiv:1409.5425*, 2014.

[AJW20] Anton Arnold, Shi Jin, and Tobias Wöhrer. Sharp decay estimates in local sensitivity analysis for evolution equations with uncertainties: From ODEs to linear kinetic equations. *Journal of Differential Equations*, 268(3):1156–1204, 2020.

[BCE+22] Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: First-order stationarity guarantees for Langevin Monte Carlo. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 2896–2923. PMLR, 2022.

[BCM+21] Mathias Barkhagen, Ngoc Huy Chau, Éric Moulines, Miklós Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.

[Car71] Henri Cartan. *Differential Calculus*. Hermann, 1971.

[CB18] Xiang Cheng and Peter L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, volume 83, pages 186–211. PMLR, 2018.

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

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

[CHS87] Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

[CLW21] Yu Cao, Jianfeng Lu, and Lihan Wang. Complexity of randomized algorithms for underdamped Langevin dynamics. *Communications in Mathematical Sciences*, 19(7):1827–1853, 2021.

[CLW23] Yu Cao, Jianfeng Lu, and Lihan Wang. On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. *Archive for Rational Mechanics and Analysis*, 247(90):1–34, 2023.

[CMR+21] Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *SIAM Journal of Mathematics of Data Science*, 3(3):959–986, 2021.

[Dal17] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

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

[DM17] Alain Durmus and Eric Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.

[DM19] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[DMP18] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov Chain Monte Carlo: When Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[DRD20] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

[EGZ19] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Annals of Probability*, 47(4):1982–2010, 2019.

[EH21] Murat A. Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 1776–1822. PMLR, 2021.

[EHZ22] Murat A Erdogdu, Rasa Hosseinzadeh, and Matthew S. Zhang. Convergence of Langevin Monte Carlo in chi-squared and Rényi divergence. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 8151–8175. PMLR, 2022.

[FA89] José A Facenda Aguirre. A note on Taylor's theorem. *The American Mathematical Monthly*, 96(3):244–247, 1989.

[GCSR95] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, 1995.

[GGHZ21] Mert Gürbüzbalaban, Xuefeng Gao, Yunhan Hu, and Lingjiong Zhu. Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22(239):1–69, 2021.

[GGZ20] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Breaking reversibility accelerates Langevin dynamics for global non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[GGZ22] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of Stochastic Gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5):2931–2947, 2022.

[GIWZ24] Mert Gürbüzbalaban, Mohammad Rafiqul Islam, Xiaoyu Wang, and Lingjiong Zhu. Generalized EXTRA stochastic gradient Langevin dynamics. *arXiv preprint arXiv:2412.01993*, 2024.

[GS84] Clark R Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[Gui22] Emanuele Guidotti. calculus: High-dimensional numerical and symbolic calculus in R. *Journal of Statistical Software*, 104(5):1–37, 2022.

[HJ94] Roger A Horn and Charles R Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

[HKS89] Richard A Holley, Shigeo Kusuoka, and Daniel W Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *Journal of Functional Analysis*, 83(2):333–347, 1989.

[Jun22] Alexander Jung. *Machine learning: the basics*. Springer Nature, 2022.

[LF72] Peter Lancaster and Hanafi K Farahat. Norms on direct sums and tensor products. *Mathematics of Computation*, 26(118):401–414, 1972.

[MCC+21] Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942–1992, 2021.

[MMW+21] Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm. *Journal of Machine Learning Research*, 22(42):1–41, 2021.

[Mon23] Pierre Monmarché. Almost sure contraction for diffusions on $\mathbb{R}^d$. Application to generalized Langevin diffusions. *Stochastic Processes and their Applications*, 161:316–349, 2023.

[MSH02] Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002.

[Red12] Darren Redfern. *The Maple Handbook: Maple V Release 4*. Springer Science & Business Media, 2012.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1674–1703. PMLR, 2017.

[SBB+80] Josef Stoer, Roland Bulirsch, R Bartels, Walter Gautschi, and Christoph Witzgall. *Introduction to Numerical Analysis*, volume 1993. Springer, 1980.

[SC08] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[Stu10] Andrew M Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[TTV16] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(1):193–225, 2016.

[Vil09] Cédric Villani. Hypocoercivity. *Memoirs of the American Mathematical Society*, 202(950):iv+141, 2009.

[Vit08] Saverio Vito. Air Quality. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C59K5F.

[WL06] Jin Wang and Chunlei Liu. Generating multivariate mixture of normal distributions using a modified Cholesky decomposition. In *Proceedings of the 2006 Winter Simulation Conference*, pages 342–347. IEEE, 2006.

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[ZADS23] Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87:25, 2023.

Appendix A. Supporting Results for Section 2

Denote $\mathcal{M}_{m,n}(\mathbb{R})$ the set of real matrices of size $m\times n$. In [Mon23, Section 4.3], Monmarché considers the *generalized Langevin diffusions*:

$$\begin{aligned} dX_t &= AY_t dt, \\ dY_t &= -A^\top \nabla U(X_t)dt - \gamma BY_t dt + \sqrt{\gamma}\Sigma dW_t, \end{aligned} \tag{46}$$

with $A \in \mathcal{M}_{d,p}(\mathbb{R})$; $B, \Sigma \in \mathcal{M}_{p,p}(\mathbb{R})$; $U \in \mathcal{C}^2(\mathbb{R}^d)$; $\gamma > 0$ and W is a standard p-dimensional Brownian motion.

Set b as the drift coefficient of (46), that is

$$b(x,y) := \begin{pmatrix} Ay \\ -A^\top \nabla U(x) - \gamma By \end{pmatrix}. \tag{47}$$

We summarize here Assumptions 1, 2 and 3 in [Mon23, Section 4.3] regarding (46).

Condition F.

- There exist $m, L, \kappa > 0$ and a symmetric positive-definite matrix H of size $p \times p$ such that
$$HB + B^\top H \geq 2\kappa H, \tag{48}$$
and that U is m strongly-convex and L-smooth: $mI_d \leq \nabla^2 U(x) \leq LI_d$, for any $x \in \mathbb{R}^d$.
- $p \geq d$ and $A = (I_d, 0, \dots, 0)$.
- When $p > d$, consider the decomposition $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ where $B_{11}, B_{12}, B_{21}, B_{22}$ are respectively of size $d \times d$, $d \times (p-d)$, $(p-d) \times d$ and $(p-d) \times (p-d)$. In this case, we assume B_{22} is invertible and that
$$E := B_{11} - B_{12}B_{22}^{-1}B_{21}$$
is symmetric positive-definite. Set $D := B_{12}B_{22}^{-1}$.
- When $p = d$, we assume B is symmetric positive definite. Set $E = B$ and $D = 0$.

Remark A.1. In [Mon23, Assumption 2], the author writes $HB \geq \kappa H$. If one looks at the notation subsection right before Section 2 of the aforementioned reference, $HB \geq \kappa H$ for not necessarily symmetric matrix HB is understood as $HB + B^\top H \geq 2\kappa H$, which is what we have in our Condition F.

Remark A.2. In [Mon23], beside from Condition F, the author also assumes that $\nabla_\alpha U$ for any $|\alpha| = \sum_i \alpha_i \geq 2$ is bounded. Per private communication with the author, this is done out of convenience to avoid technical regularity issues regarding the semi-groups. In our case, we are interested in Theorem 9 in [Mon23] which only requires the boundedness of second-order derivative of U and not of the higher order derivative.

The following result is stated under Assumption 3 in [Mon23].

Lemma A.3. *Under Condition F, there exist constants $h_i > 0, 1 \le i \le 5$ such that*

$$HA^\top AH \le h_1 H, \qquad \frac{1}{h_2} I_d \le E \le h_3 I_d,$$
$$\begin{pmatrix} I_d & -D \\ 0 & 0 \end{pmatrix} \le h_4 H, \qquad \begin{pmatrix} I_d & -D \\ -D^\top & 0 \end{pmatrix} \le h_5 H.$$

Note that we follow the convention in [Mon23]: for $m \times m$ matrices M, H that are not necessarily symmetric, $M \ge H$ means $\langle x, Mx \rangle \ge \langle x, Hx \rangle$ for all $x \in \mathbb{R}^m$.

Theorem A.4. *(*[Mon23, Lemma 8 and Theorem 9]*) Assume Conditions F and set*

$$\gamma_0 = 2\sqrt{\frac{h_1 L}{\kappa}} \max\left\{ \sqrt{h_2 h_5}, \sqrt{\frac{h_4}{\kappa}} \right\}.$$

Further assume that the friction coefficient γ is sufficiently large: $\gamma \ge \gamma_0$. Set $\rho = \min\left\{ \frac{m}{3h_3\gamma}, \frac{\gamma\kappa}{6} \right\}$. Recall the drift coefficient b in (47) *and denote J_b its Jacobian matrix. Write $\begin{pmatrix} I_d & -D \end{pmatrix}$ as a block matrix. Then $M := \begin{pmatrix} E & \frac{1}{\gamma}\begin{pmatrix} I_d & -D \end{pmatrix} \\ \frac{1}{\gamma}\begin{pmatrix} I_d & -D \end{pmatrix}^\top & \frac{\kappa}{Lh_1} H \end{pmatrix}$ is a symmetric positive-definite matrix of size $(d+p) \times (d+p)$ such that*

$$MJ_b + J_b^\top M \le -2\rho M.$$

Moreover, the matrix M satisfies

$$\frac{1}{2}\begin{pmatrix} E & 0 \\ 0 & \frac{\kappa}{Lh_1} H \end{pmatrix} \le M \le \frac{3}{2}\begin{pmatrix} E & 0 \\ 0 & \frac{\kappa}{Lh_1} H \end{pmatrix}. \tag{49}$$

A *P-th order Langevin dynamics* as the focus of the main paper is a special case of (46) where $p = (P-1)d$ and

$$A = A_P = \begin{pmatrix} I_d & 0 & \dots & 0 \end{pmatrix} \qquad \text{and} \qquad B = B_P = \begin{pmatrix} 0 & -I_d & 0 & \dots & 0 \\ I_d & 0 & -I_d & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & I_d & 0 & -I_d \\ 0 & \dots & 0 & I_d & I_d \end{pmatrix}. \tag{50}$$

The following Corollary is Theorem A.4 in the special case of P-th order Langevin dynamics. The proof is mostly taken from [Mon23, Section 4]. We add some details regarding dimension dependence of the parameters since this is not one of the goals of [Mon23]; however, it is a crucial concern of our paper.

Corollary A.5. *(*[Mon23, Section 4]*) Conditions F is satisfied for the P-th order Langevin dynamics* (50)*, so that the conclusion of Theorem A.4 applies to* (50)*. In particular, regarding the matrix M, we have $E = I_d, D = -\begin{pmatrix} I_d & \dots & I_d \end{pmatrix}$, while the matrix H, the constants κ and $h_i, 1 \le i \le 5$ are explicitly provided in* ***Step 3*** *of the proof.*

Moreover, $\rho = \rho(\gamma, L, P), \gamma_0 = \gamma_0(\gamma, L, P)$ depend on γ, L, P but do not depend on the dimension parameter d. Furthermore, $\lambda_{\min,M} = \lambda_{\min,M}(P)$ and $\lambda_{\max,M} = \lambda_{\max,M}(P)$ as respectively the smallest and largest eigenvalues of the positive definite matrix M depend on P and do not depend on the dimension parameter d.

Proof. The proof is divided into six steps.

Step 1: We start by observing a simplified form of the matrix $B = B_P$ in (50):

$$B = B_{\text{sim}} \otimes I_d, \qquad B_{\text{sim}} := \begin{pmatrix} 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 0 & -1 \\ 0 & \dots & 0 & 1 & 1 \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product ([HJ94]). This simplified form indicates that B and the $(P-1) \times (P-1)$ matrix B_{sim} have the same spectrum, and thus such spectrum does not depend on d. Furthermore, it indicates that if $v_i, 1 \le i \le P-1$ are eigenvectors (respectively generalized eigenvectors) of B_{sim} and $e_j, 1 \le j \le d$ are the standard basis of $\mathbb{R}^d$, then $w_i = v_i \otimes e_j, 1 \le i \le P-1, 1 \le j \le d$ are eigenvectors (respectively generalized eigenvectors) of B.

Step 2: Let us verify that $\min\{\text{Re}(\lambda) : \lambda \text{ is an eigenvalue of } B_{\text{sim}}\} > 0$, which together with B, B_{sim} having the same spectrum from the **Step 1** imply $\min\{\text{Re}(\lambda) : \lambda \text{ is an eigenvalue of } B\} > 0$.

We have the decomposition $B_{\text{sim}} = \frac{1}{2}(B_{\text{sim}} + B_{\text{sim}}^\top) + \frac{1}{2}(B_{\text{sim}} - B_{\text{sim}}^\top) := H + K$ where H is a Hermitian matrix and K is a skew-Hermitian matrix. Now assume λ is an eigenvalue of B_{sim}: $B_{\text{sim}}x = \lambda x$ for a nonzero vector $x = (x_1, \dots, x_{P-1}) \in \mathbb{C}^{P-1}$. Then $\lambda = \frac{x^* B_{\text{sim}} x}{x^* x}$ where x^* denotes the conjugate transpose of x, and hence

$$\text{Re}(\lambda) = \frac{x^* H x}{x^* x} = \frac{|x_{P-1}|^2}{|x|^2}.$$

We claim that $x_{P-1} \neq 0$ which implies $\text{Re}(\lambda) > 0$. Suppose the opposite that $x_{P-1} = 0$, then it is easy to solve for $B_{\text{sim}}x = \lambda x$ to get $x_1 = x_2 = \dots = x_{P-1} = 0$, which is a contradiction. This completes our argument for the **Step 2**.

Step 3 in the case where B is diagonalizable: Let us construct H, κ that satisfy Condition F in the simpler case where B is diagonalizable. The construction has been done in [AE14, Lemma 4.3] or [AJW20, Section 2.1], and we summarize it here for the sake of completeness.

In this case, B has $(P-1)d$ linearly independent eigenvectors $w_i, 1 \le i \le (P-1)d$ corresponding to $(P-1)d$ eigenvalues $\lambda_i, 1 \le i \le (P-1)d$. Denote w_i^* the conjugate transpose of w_i and set $H =: \sum_{i=1}^{(P-1)d} w_i w_i^*$, then

$$HB + B^\top H = \sum_{i=1}^{(P-1)d} \big(\lambda_i + \overline{\lambda_i}\big) w_i w_i^* \ge 2 \min\{\text{Re}(\lambda_i), 1 \le i \le (P-1)d\} \sum_{i=1}^{(P-1)d} w_i w_i^*$$

$$= 2\hat{\lambda}H, \quad \hat{\lambda} = \min\{\text{Re}(\lambda) : \lambda \text{ is an eigenvalue of } B\}.$$

Thus, in the case where B is diagonalizable, κ in Condition F can be taken as $\hat{\lambda}$ which is a positive number per our **Step 2** above.

Step 3 in the case where B is not diagonalizable: In contrast to the previous case, there is at least one Jordan block of B of length $\ell_n \geq 2$. In this case, and the construction of H, κ satisfying Condition F is more elaborate. Denote the Jordan blocks of B by $B_n, 1 \leq n \leq H$. Each block B_n of length ℓ_n is associated with the eigenvalue λ_n and the set of generalized eigenvectors $v_n^{(k)}; 1 \leq k \leq \ell_n$. In particular, $v_n^{(1)}$ is the (standard) eigenvector of J_n.

For a Jordan block B_n with $\text{Re}(\lambda_n) > \hat{\lambda}$, we set $H_n = \sum_{i=1}^{\ell_n} b_n^i v_n^{(i)} \left(v_n^{(i)}\right)^*$ where

$$b_n^1 = 1; b_n^j = c_j(t_n)^{2(1-j)}, 2 \leq j \leq \ell_n \qquad \text{and} \quad c_1 = 1; c_{j+1} = 1 + c_j^2,\ 2 \leq j \leq \ell_n$$
$$\text{and} \quad t_n = 2(\text{Re}(\lambda_n) - \kappa).$$

Then per [AE14, Lemma 4.3], $H_n B_n + B_n^\top H_n \geq 2\hat{\lambda} H_n$.

Meanwhile, for a Jordan block B_m with $\text{Re}(\lambda_m) = \hat{\lambda}$, we replace the above t_n with $t_m = 2(\text{Re}(\lambda_n) - \hat{\lambda} + \epsilon)$ for any $\epsilon \in (0, \hat{\lambda})$ and define $\widetilde{H}_m(\epsilon) = \sum_{i=1}^{\ell_m} b_m^i(\epsilon) v_m^{(i)} \left(v_m^{(i)}\right)^*$. Then $\widetilde{H}_m B_m + B_m^\top \widetilde{H}_m \geq 2(\hat{\lambda} - \epsilon)\widetilde{H}_m$.

Therefore, in the case where B is not diagonalizable, we denote $I = \{n \in \{1, \cdots, N\} : \ell_n \geq 2, \text{Re}(\lambda_n) = \hat{\lambda}\}$ and define $H := H(\epsilon) = \sum_{n \in \{1,\dots,N\} \setminus I} H_n + \sum_{m \in I} \widetilde{H}_m(\epsilon)$. Then we have

$$H(\epsilon)B + B^\top H(\epsilon) \geq 2(\hat{\lambda} - \epsilon)H(\epsilon).$$

Thus, in the case where B is not diagonalizable, κ in Condition F is $\hat{\lambda} - \epsilon$ for any $\epsilon \in (0, \hat{\lambda})$. Notice $\hat{\lambda} > 0$ per our **Step 2**, so that it is possible to choose $\epsilon > 0$ such that $\hat{\lambda} - \epsilon > 0$.

Step 4: Let us verify that κ, $\|H\|_{\text{op}}$ and $\|H^{-1}\|_{\text{op}}$ do not depend on d. The former is clear from the fact that κ is either $2\hat{\lambda}$ or $2(\hat{\lambda} - \epsilon)$ in the **Step 2**, and the fact that B has the same spectrum as the $(P-1) \times (P-1)$ matrix B_{sim}, per the first paragraph of this proof. Regarding $\|H\|_{\text{op}}$, we will assume B is diagonalizable to keep things simple (the case of non-diagonalizable B is almost the same). We know from the first paragraph of this proof that

$$\begin{aligned} H = \sum_{1 \leq i \leq P-1, 1 \leq j \leq d} (v_i \otimes e_j)\big(v_i^* \otimes e_j^\top\big) &= \sum_{1 \leq i \leq P-1, 1 \leq j \leq d} (v_i v_i^*) \otimes \big(e_j \otimes e_j^\top\big) \\ &= \left(\sum_{1 \leq i \leq P-1} v_i v_i^*\right) \otimes \left(\sum_{1 \leq j \leq d} e_j \otimes e_j^\top\right) \\ &= \left(\sum_{1 \leq i \leq P-1} v_i v_i^*\right) \otimes I_d. \end{aligned} \tag{51}$$

Thus, we have $\|H\|_{\text{op}} = \left\|\sum_{1\le i\le P-1} v_i v_i^*\right\|_{\text{op}} \|I_d\|_{\text{op}} = \left\|\sum_{1\le i\le P-1} v_i v_i^*\right\|_{\text{op}}$ (see [LF72, Theorem 8] regarding matrix norms and Kronecker product). Since $\sum_{1\le i\le P} v_i v_i^*$ is a $(P-1)\times(P-1)$ matrix, $\|H\|_{\text{op}}$ does not depend on d. We can reach the same conclusion for $\|H^{-1}\|_{\text{op}}$, noting that $H^{-1} = \left(\sum_{1\le i\le P-1} v_i v_i^*\right)^{-1} \otimes I_d$.

Step 5: We verify that the constants in Lemma A.3 do not depend on d. In Lemma A.3, the matrix $E = I_d$ (pointed out below Assumption 3 in [Mon23]). Thus, we can take $h_1 = \left\|HA^\top A\right\|_{\text{op}}, h_2 = 1$ and $h_3 = 1$. Then from (51) and $A = \begin{pmatrix} I_d & 0 & \dots & 0\end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0\end{pmatrix} \otimes I_d$, we deduce that $h_1 = \left\| H \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \right\|_{\text{op}}$. Thus, h_1, h_2, h_3 do not depend on d. What remain to study are h_4 and h_5. We have $D = -(I_d, \dots, I_d)$ as pointed out below Assumption 3 in [Mon23]. It is easy to verify that $(1 + (P-1)/2)I_{(P-1)d} - \begin{pmatrix} I_d & -D \\ 0 & 0 \end{pmatrix} \ge 0$, which implies

$$\begin{pmatrix} I_d & -D \\ 0 & 0 \end{pmatrix} H^{-1}H \le (1 + P/2)H^{-1}H \le (1 + (P-1)/2) \left\|H^{-1}\right\|_{\text{op}} H,$$

and hence $h_4 = (1 + (P-1)/2) \|H^{-1}\|_{\text{op}}$. The formula $h_5 = (1+P) \|H^{-1}\|_{\text{op}}$ is obtained the same way, noting that $\begin{pmatrix} I_d & -D \\ -D^\top & 0 \end{pmatrix} = \begin{pmatrix} I_d & -D \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} I_d & 0 \\ -D^\top & 0 \end{pmatrix}$. Finally, $\|H^{-1}\|_{\text{op}} = \left\|\left(\sum_{1\le i\le P-1} v_i v_i^*\right)^{-1}\right\|_{\text{op}}$ does not depend on d per the **Step 3**, so that h_4 and h_5 do not depend on d.

Step 6: Let us consider ρ, γ_0 and $\lambda_{\min,M}$ of Theorem A.4 in the context of the P-th order Langevin dynamics (50). Note that m and L are respectively the strong-convexity and smoothness constants of U does not depend on d. This, combined with the conclusions in the **Step 3** and **Step 4**, implies that $\gamma_0 = 2\sqrt{\frac{h_1 L}{\kappa}} \max\left\{\sqrt{h_2 h_5}, \sqrt{\frac{h_4}{\kappa}}\right\}$ and $\gamma > \gamma_0$ does not depend on d. Regarding $\lambda_{\min,M}$, it is pointed out below Assumption 3 in [Mon23] that in the case of P-th order Langevin dynamics (50), $E = I_d$ in Condition F. Then inequality (49) becomes

$$\frac{1}{2}\begin{pmatrix} I_d & 0 \\ 0 & \frac{\kappa}{Lh_1}H \end{pmatrix} \le M \le \frac{3}{2}\begin{pmatrix} I_d & 0 \\ 0 & \frac{\kappa}{Lh_1}H \end{pmatrix}.$$

Moreover, per (51), H and the $(P-1)\times(P-1)$ matrix $\sum_{1\le i\le P-1} v_i v_i^*$ share the same spectrum which does not depend on d. Then by a consequence of Courant–Fischer–Weyl's min-max Theorem regarding comparison of eigenvalues of positive definite matrices ([HJ94, Problem 4.2.P8, Page 238], we can conclude $\lambda_{\min,M}$ as the smallest eigenvalue of the positive definite matrix M does not depend on d. The same conclusion holds for $\lambda_{\max,M}$. The proof is complete. □

Appendix B. Details of Fourth-Order Langevin Monte Carlo Algorithm

Lemma B.1 as the first result of this Appendix contains explicit form of the components of $\bar{x}((k+1)\eta)$ in terms of the components of $x^{(k)}$ in the splitting scheme (11). Based on it, we will be able to derive in Lemma B.2 the conditional mean and covariance associated with the fourth-order LMC algorithm in Section 2.1.

Here are the components of $\hat{x}(t)$ in terms of the components of $x^{(k)}$.

$$\begin{aligned}
\hat{v}_1(t) &= v_1^{(k)}, \\
\hat{\theta}(t) &= \theta^{(k)} + (t-k\eta)v_1^{(k)}, \\
\hat{v}_2(t) &= v_2^{(k)} + \gamma\left(v_3^{(k)} - v_1^{(k)}\right)(t-k\eta), \\
\hat{v}_3(t) &= e^{-\gamma(t-k\eta)}v_3^{(k)} - \gamma\int_{k\eta}^t e^{-\gamma(t-s)}\hat{v}_2(s)ds + \sqrt{2\gamma}\int_{k\eta}^t e^{-\gamma(t-s)}dB_s \\
&= e^{-\gamma(t-k\eta)}v_3^{(k)} - \gamma v_2^{(k)}\int_{k\eta}^t e^{-\gamma(t-s)}ds \\
&\quad - \gamma^2\left(v_3^{(k)} - v_1^{(k)}\right)\int_{k\eta}^t e^{-\gamma(t-s)}(s-k\eta)ds + \sqrt{2\gamma}\int_{k\eta}^t e^{-\gamma(t-s)}dB_s.
\end{aligned} \tag{52}$$

Next are the components of $\tilde{x}(t)$ in terms of the components of $x^{(k)}$.

$$\begin{aligned}
\tilde{v}_1(t) &= v_1^{(k)} - \int_{k\eta}^t \tilde{g}(s)ds + \gamma\int_{k\eta}^t \hat{v}_2(s)ds \\
&= v_1^{(k)} - \int_{k\eta}^t \tilde{g}(s)ds + \gamma v_2^{(k)}(t-k\eta) + \gamma^2\left(v_3^{(k)} - v_1^{(k)}\right)\frac{(t-k\eta)^2}{2}, \\
\tilde{\theta}(t) &= \theta^{(k)} + \int_{k\eta}^t \tilde{v}_1(s)ds \\
&= \theta^{(k)} + v_1^{(k)}(t-k\eta) - \int_{k\eta}^t\int_{k\eta}^s \tilde{g}(r)drds + \gamma v_2^{(k)}\frac{(t-k\eta)^2}{2!} + \gamma^2\left(v_3^{(k)} - v_1^{(k)}\right)\frac{(t-k\eta)^3}{3!}, \\
\tilde{v}_2(t) &= v_2^{(k)} - \gamma\int_{k\eta}^t \tilde{v}_1(s)ds + \gamma\int_{k\eta}^t \hat{v}_3(s)ds \\
&= v_2^{(k)} - \gamma v_1^{(k)}(t-k\eta) + \gamma\int_{k\eta}^t\int_{k\eta}^s \tilde{g}(r)drds - \gamma^2 v_2^{(k)}\frac{(t-k\eta)^2}{2!} - \gamma^3\left(v_3^{(k)} - v_1^{(k)}\right)\frac{(t-k\eta)^3}{3!} \\
&\quad + \gamma v_3^{(k)}\int_{k\eta}^t e^{-\gamma(s-k\eta)}ds - \gamma^2 v_2^{(k)}\int_{k\eta}^t\int_{k\eta}^s e^{-\gamma(s-r)}drds \\
&\quad - \gamma^3\left(v_3^{(k)} - v_1^{(k)}\right)\int_{k\eta}^t\int_{k\eta}^s e^{-\gamma(s-r)}(r-k\eta)drds + \gamma\sqrt{2\gamma}\int_{k\eta}^t\int_{k\eta}^s e^{-\gamma(s-r)}dB_r ds, \\
\tilde{v}_3(t) &= v_3^{(k)}e^{-\gamma(t-k\eta)} - \gamma\int_{k\eta}^t e^{-\gamma(t-s)}\tilde{v}_2(s)ds + \sqrt{2\gamma}\int_{k\eta}^t e^{-\gamma(t-s)}dB_s
\end{aligned} \tag{53}$$

$$
\begin{aligned}
&= v_3^{(k)} e^{-\gamma(t-k\eta)} + \sqrt{2\gamma} \int_{k\eta}^t e^{-\gamma(t-s)} dB_s - \gamma v_2^{(k)} \int_{k\eta}^t e^{-\gamma(t-s)} ds + \gamma^2 v_1^{(k)} \int_{k\eta}^t e^{-\gamma(t-s)} (s-k\eta) ds \\
&\quad - \gamma^2 \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r e^{-\gamma(t-s)} \tilde{g}(w) dw dr ds \\
&\quad\quad + \gamma^3 v_2^{(k)} \int_{k\eta}^t e^{-\gamma(t-s)} \frac{(s-k\eta)^2}{2!} ds + \gamma^4 \left(v_3^{(k)} - v_1^{(k)} \right) \int_{k\eta}^t e^{-\gamma(t-s)} \frac{(s-k\eta)^3}{3!} ds \\
&\quad - \gamma^2 v_3^{(k)} \int_{k\eta}^t \int_{k\eta}^s e^{-\gamma(t-s)} e^{-\gamma(r-k\eta)} dr ds + \gamma^3 v_2^{(k)} \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r e^{-\gamma(t-s)} e^{-\gamma(r-w)} dw dr ds \\
&\quad\quad + \gamma^4 \left(v_3^{(k)} - v_1^{(k)} \right) \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r e^{-\gamma(t-s)} e^{-\gamma(r-w)} (w-k\eta) dw dr ds \\
&\quad - \gamma^2 \sqrt{2\gamma} \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r e^{-\gamma(r-w)} e^{-\gamma(t-s)} dB_w dr ds.
\end{aligned}
$$

Finally are the components of $\bar{x}(t)$ in terms of the components of $x^{(k)}$.

$$
\begin{aligned}
\bar{v}_1(t) &= v_1^{(k)} - \int_{k\eta}^t \bar{g}(s) ds + \gamma \int_{k\eta}^t \tilde{v}_2(s) ds \\
&= v_1^{(k)} - \int_{k\eta}^t \bar{g}(s) ds + \gamma v_2^{(k)} (t-k\eta) - \gamma^2 v_1^{(k)} \frac{(t-k\eta)^2}{2!} + \gamma^2 \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r \tilde{g}(w) dw dr ds \\
&\quad - \gamma^3 v_2^{(k)} \frac{(t-k\eta)^3}{3!} - \gamma^4 \left(v_3^{(k)} - v_1^{(k)} \right) \frac{(t-k\eta)^4}{4!} + \gamma^2 v_3^{(k)} \int_{k\eta}^t \int_{k\eta}^s e^{-\gamma(r-k\eta)} dr ds \\
&- \gamma^3 v_2^{(k)} \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r e^{-\gamma(r-w)} dw dr ds - \gamma^4 \left(v_3^{(k)} - v_1^{(k)} \right) \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r e^{-\gamma(r-w)} (w-k\eta) dw dr ds \\
&\quad\quad + \gamma^2 \sqrt{2\gamma} \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r e^{-\gamma(r-w)} dB_w dr ds, \\
\bar{\theta}(t) &= \theta^{(k)} + \int_{k\eta}^t \bar{v}_1(s) ds \\
&= \theta^{(k)} + v_1^{(k)} (t-k\eta) - \int_{k\eta}^t \int_{k\eta}^s \bar{g}(r) dr ds + \gamma v_2^{(k)} \frac{(t-k\eta)^2}{2!} - \gamma^2 v_1^{(k)} \frac{(t-k\eta)^3}{3!} \\
&\quad + \gamma^2 \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r \int_{k\eta}^w \tilde{g}(y) dy dw dr ds - \gamma^3 v_2^{(k)} \frac{(t-k\eta)^4}{4!} - \gamma^4 \left(v_3^{(k)} - v_1^{(k)} \right) \frac{(t-k\eta)^5}{5!} \\
&\quad\quad + \gamma^2 v_3^{(k)} \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r e^{-\gamma(w-k\eta)} dw dr ds - \gamma^3 v_2^{(k)} \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r \int_{k\eta}^w e^{-\gamma(w-y)} dy dw dr ds \\
&\quad\quad - \gamma^4 \left(v_3^{(k)} - v_1^{(k)} \right) \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r \int_{k\eta}^w e^{-\gamma(w-y)} (y-k\eta) dy dw dr ds \\
&\quad\quad + \gamma^2 \sqrt{2\gamma} \int_{k\eta}^t \int_{k\eta}^s \int_{k\eta}^r \int_{k\eta}^w e^{-\gamma(w-y)} dB_y dw dr ds,
\end{aligned}
$$

$$
\begin{aligned}
\bar{v}_2(t) &= v_2^{(k)} - \gamma \int_{k\eta}^{t} \bar{v}_1(s)ds + \gamma \int_{k\eta}^{t} \tilde{v}_3(s)ds \\
&= v_2^{(k)} + \Bigg[-\gamma v_1^{(k)}(t-k\eta) + \gamma \int_{k\eta}^{t}\int_{k\eta}^{s} \bar{g}(r)drds - \gamma^2 v_2^{(k)} \frac{(t-k\eta)^2}{2} + \gamma^3 v_1^{(k)} \frac{(t-k\eta)^3}{3!} \\
&\quad - \gamma^3 \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} \tilde{g}(y)dydwdrds + \gamma^4 v_2^{(k)} \frac{(t-k\eta)^4}{4!} + \gamma^5 \Big(v_3^{(k)} - v_1^{(k)}\Big) \frac{(t-k\eta)^5}{5!} \\
&\quad - \gamma^3 v_3^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma(w-k\eta)} dwdrds + \gamma^4 v_2^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(w-y)} dydwdrds \\
&\quad + \gamma^5 \Big(v_3^{(k)} - v_1^{(k)}\Big) \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(w-y)}(y-k\eta)dydwdrds \\
&\quad - \gamma^3 \sqrt{2\gamma} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(w-y)} dB_y dwdrds \Bigg] \\
&\quad + \Bigg[\gamma v_3^{(k)} \int_{k\eta}^{t} e^{-\gamma(s-k\eta)} ds + \gamma\sqrt{2\gamma} \int_{k\eta}^{t}\int_{k\eta}^{s} e^{-\gamma(s-r)} dB_r ds \\
&\quad - \gamma^2 v_2^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s} e^{-\gamma(s-r)} drds + \gamma^3 v_1^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s} e^{-\gamma(s-r)}(r-k\eta)drds \\
&\quad - \gamma^3 \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(s-r)} \tilde{g}(y)dydwdrds + \frac{\gamma^4}{2!} v_2^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s} e^{-\gamma(s-r)}(r-k\eta)^2 drds \\
&\quad + \frac{\gamma^5}{3!}\Big(v_3^{(k)} - v_1^{(k)}\Big) \int_{k\eta}^{t}\int_{k\eta}^{s} e^{-\gamma(s-r)}(r-k\eta)^3 drds - \gamma^3 v_3^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma(s-r)} e^{-\gamma(w-k\eta)} dwdrds \\
&\quad + \gamma^4 v_2^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(s-r)} e^{-\gamma(w-y)} dydwdrds \\
&\quad + \gamma^5 \Big(v_3^{(k)} - v_1^{(k)}\Big) \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(s-r)} e^{-\gamma(w-y)}(y-k\eta)dydwdrds \\
&\quad - \gamma^3 \sqrt{2\gamma} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(s-y)} e^{-\gamma(w-r)} dB_y dwdrds. \Bigg], \\
\bar{v}_3(t) &= \bar{v}_3^{(k)} e^{-\gamma(t-k\eta)} - \gamma \int_{k\eta}^{t} e^{-\gamma(t-s)} \bar{v}_2(s)ds + \sqrt{2\gamma} \int_{k\eta}^{t} e^{-\gamma(t-s)} dB_s \\
&= \bar{v}_3^{(k)} e^{-\gamma(t-k\eta)} + \sqrt{2\gamma} \int_{k\eta}^{t} e^{-\gamma(t-s)} dB_s \\
&\quad + \Bigg[-\gamma v_2^{(k)} \int_{k\eta}^{t} e^{-\gamma(t-s)} ds + \gamma^2 v_1^{(k)} \int_{k\eta}^{t} e^{-\gamma(t-s)}(s-k\eta)ds - \gamma^2 \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma(t-s)} \bar{g}(w)dwdrds \\
&\quad + \frac{\gamma^3}{2!} v_2^{(k)} \int_{k\eta}^{t} e^{-\gamma(t-s)}(s-k\eta)^2 ds - \frac{\gamma^4}{3!} v_1^{(k)} \int_{k\eta}^{t} e^{-\gamma(t-s)}(s-k\eta)^3 ds
\end{aligned}
$$

$$
\begin{aligned}
&+ \gamma^4 \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma(t-s)}\tilde{g}(z)dzdydwdrds \\
&- \frac{\gamma^5}{4!}v_2^{(k)}\int_{k\eta}^{t} e^{-\gamma(t-s)}(s-k\eta)^4 ds - \frac{\gamma^6}{5!}\Big(v_3^{(k)} - v_1^{(k)}\Big)\int_{k\eta}^{t} e^{-\gamma(t-s)}(s-k\eta)^5 ds \\
&+ \gamma^4 v_3^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(t-s)}e^{-\gamma(y-k\eta)}dydwdrds \\
&- \gamma^5 v_2^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-k(t-s)}e^{-\gamma(y-z)}dzdydwdrds \\
&- \gamma^6\Big(v_3^{(k)} - v_1^{(k)}\Big)\int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma(t-s)}e^{-\gamma(y-z)}(z-k\eta)dzdydwdrds \\
&+ \gamma^4\sqrt{2\gamma}\int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma(t-s)}e^{-\gamma(y-z)}dB_z dydwdrds\Bigg] \\
&+ \Bigg[-\gamma^2 v_3^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s} e^{-\gamma(t-s)}e^{-\gamma(r-k\eta)}drds - \gamma^2\sqrt{2\gamma}\int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma(t-s)}e^{-\gamma(r-w)}dB_w drds \\
&+ \gamma^3 v_2^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma(t-s)}e^{-\gamma(r-w)}dwdrds - \gamma^4 v_1^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma(t-s)}e^{-\gamma(r-w)}(w-k\eta)dwdrds \\
&+ \gamma^4 \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma(t-s)}e^{-\gamma(r-w)}\tilde{g}(z)dzdydwdrds \\
&- \frac{\gamma^5}{2!}v_2^{(k)}\int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma(t-s)}e^{-\gamma(r-w)}(w-k\eta)^2 dwdrds \\
&- \frac{\gamma^6}{3!}\Big(v_3^{(k)} - v_1^{(k)}\Big)\int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma(t-s)}e^{-\gamma(r-w)}(w-k\eta)^3 dwdrds \\
&+ \gamma^4 v_3^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w} e^{-\gamma(t-s)}e^{-\gamma(r-w)}e^{-\gamma(y-k\eta)}dydwdrds \\
&- \gamma^5 v_2^{(k)} \int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma(t-s)}e^{-\gamma(r-w)}e^{-\gamma(y-z)}dzdydwdrds \\
&- \gamma^6\Big(v_3^{(k)} - v_1^{(k)}\Big)\int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma(t-s)}e^{-\gamma(r-w)}e^{-\gamma(y-z)}(z-k\eta)dzdydwdrds \\
&+ \gamma^4\sqrt{2\gamma}\int_{k\eta}^{t}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma(t-s)}e^{-\gamma(y-z)}e^{-\gamma(r-w)}dB_z dydwdrds\Bigg].
\end{aligned}
$$

Via the above equations and software to evaluate the iterated integrals, we obtain the following result.

Lemma B.1. *Recall the definition of polynomials $\tilde{g}(t)$ and $\bar{g}(t)$ in* (13)*. The following are explicit expressions of components of $\bar{x}((k+1)\eta)$ in terms of $x^{(k)}$.*

$$\bar{\theta}((k+1)\eta) = -\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\bar{g}(r)drds + \gamma^2\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\tilde{g}(y)dydwdrds$$
$$+ f_0 + \theta^{(k)}\mu_{00} + v_1^{(k)}\mu_{01} + v_2^{(k)}\mu_{02} + v_3^{(k)}\mu_{03};$$
$$\bar{v}_1((k+1)\eta) = -\int_{k\eta}^{(k+1)\eta}\bar{g}(s)ds + \gamma^2\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\tilde{g}(w)dwdrds$$
$$+ f_1 + \theta^{(k)}\mu_{10} + v_1^{(k)}\mu_{11} + v_2^{(k)}\mu_{12} + v_3^{(k)}\mu_{13};$$
$$\bar{v}_2((k+1)\eta) = \gamma\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\bar{g}(r)drds - \gamma^3\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\tilde{g}(y)dydwdrds$$
$$-\gamma^3\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}e^{-\gamma(s-r)}\tilde{g}(y)dydwdrds + f_2 + \theta^{(k)}\mu_{20} + v_1^{(k)}\mu_{21} + v_2^{(k)}\mu_{22} + v_3^{(k)}\mu_{23};$$
$$\bar{v}_3((k+1)\eta) = -\gamma^2\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}e^{-\gamma((k+1)\eta-s)}\bar{g}(w)dwdrds$$
$$+\gamma^4\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y}e^{-\gamma((k+1)\eta-s)}\tilde{g}(z)dzdydwdrds$$
$$+\gamma^4\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y}e^{-\gamma((k+1)\eta-s)}e^{-\gamma(r-w)}\tilde{g}(z)dzdydwdrds$$
$$+ f_3 + \theta^{(k)}\mu_{30} + v_1^{(k)}\mu_{31} + v_2^{(k)}\mu_{32} + v_3^{(k)}\mu_{33}.$$

Here $\mu_{ij}, 0 \le i, j \le 3$ *are the constants*

$$\mu_{00} = 1;$$
$$\mu_{01} = \eta - \gamma^2\frac{\eta^3}{3!} + \gamma^4\frac{\eta^5}{5!} + \gamma^4\left(-\frac{e^{-\gamma\eta}}{\gamma^5} + \frac{1}{\gamma^5} - \frac{\eta}{\gamma^4} + \frac{\eta^2}{2\gamma^3} - \frac{\eta^3}{6\gamma^2} + \frac{\eta^4}{24\gamma}\right);$$
$$\mu_{02} = \gamma\frac{\eta^2}{2!} - \gamma^3\frac{\eta^4}{4!} - \gamma^3\left(\frac{e^{-\gamma\eta}}{\gamma^4} - \frac{1}{\gamma^4} + \frac{\eta}{\gamma^3} - \frac{\eta^2}{2\gamma^2} + \frac{\eta^3}{6\gamma}\right);$$
$$\mu_{03} = -\gamma^4\frac{\eta^5}{5!} + \gamma^2\left(-\frac{e^{-\gamma\eta}}{\gamma^3} + \frac{1}{\gamma^3} - \frac{\eta}{\gamma^2} + \frac{\eta^2}{2\gamma}\right)$$
$$-\gamma^4\left(-\frac{e^{-\gamma\eta}}{\gamma^5} + \frac{1}{\gamma^5} - \frac{\eta}{\gamma^4} + \frac{\eta^2}{2\gamma^3} - \frac{\eta^3}{6\gamma^2} + \frac{\eta^4}{24\gamma}\right);$$
$$\mu_{10} = 0;$$
$$\mu_{11} = 1 - \gamma^2\frac{\eta^2}{2!} + \gamma^4\frac{\eta^4}{4!} + \gamma^4\left(\frac{e^{-\gamma\eta}}{\gamma^4} - \frac{1}{\gamma^4} + \frac{\eta}{\gamma^3} - \frac{\eta^2}{2\gamma^2} + \frac{\eta^3}{6\gamma}\right);$$
$$\mu_{12} = \gamma\eta - \gamma^3\frac{\eta^3}{3!} - \gamma^3\left(-\frac{e^{-\gamma\eta}}{\gamma^3} + \frac{1}{\gamma^3} - \frac{\eta}{\gamma^2} + \frac{\eta^2}{2\gamma}\right);$$
$$\mu_{13} = -\gamma^4\frac{\eta^4}{4!} + \gamma^2\left(\frac{e^{-\gamma\eta}}{\gamma^2} - \frac{1}{\gamma^2} + \frac{\eta}{\gamma}\right) - \gamma^4\left(\frac{e^{-\gamma\eta}}{\gamma^4} - \frac{1}{\gamma^4} + \frac{\eta}{\gamma^3} - \frac{\eta^2}{2\gamma^2} + \frac{\eta^3}{6\gamma}\right);$$
$$\mu_{20} = 0;$$

$$
\begin{aligned}
\mu_{21} &= -\gamma\eta + \gamma^3\frac{\eta^3}{3!} - \gamma^5\frac{\eta^5}{5!} - \gamma^5\left(-\frac{e^{-\gamma\eta}}{\gamma^5} + \frac{1}{\gamma^5} - \frac{\eta}{\gamma^4} + \frac{\eta^2}{2\gamma^3} - \frac{\eta^3}{6\gamma^2} + \frac{\eta^4}{24\gamma}\right) \\
&+ \gamma^3\left(-\frac{e^{-\gamma\eta}}{\gamma^3} + \frac{1}{\gamma^3} - \frac{\eta}{\gamma^2} + \frac{\eta^2}{2\gamma}\right) - \frac{\gamma^5}{3!}\left(-\frac{6e^{-\gamma\eta}}{\gamma^5} + \frac{6}{\gamma^5} - \frac{6\eta}{\gamma^4} + \frac{3\eta^2}{\gamma^3} - \frac{\eta^3}{\gamma^2} + \frac{\eta^4}{4\gamma}\right) \\
&- \gamma^5\left(\frac{4e^{-\gamma\eta}}{\gamma^5} - \frac{4}{\gamma^5} + \frac{\eta e^{-\gamma\eta}}{\gamma^4} + \frac{3\eta}{\gamma^4} - \frac{\eta^2}{\gamma^3} + \frac{\eta^3}{6\gamma^2}\right); \\
\mu_{22} &= 1 - \gamma^2\frac{\eta^2}{2} + \gamma^4\frac{\eta^4}{4!} + \gamma^4\left(\frac{e^{-\gamma\eta}}{\gamma^4} - \frac{1}{\gamma^4} + \frac{\eta}{\gamma^3} - \frac{\eta^2}{2\gamma^2} + \frac{\eta^3}{6\gamma}\right) - \gamma^2\left(\frac{e^{-\gamma\eta}}{\gamma^2} - \frac{1}{\gamma^2} + \frac{\eta}{\gamma}\right) \\
&+ \frac{\gamma^4}{2!}\left(\frac{2e^{-\gamma\eta}}{\gamma^4} - \frac{2}{\gamma^4} + \frac{2\eta}{\gamma^3} - \frac{\eta^2}{\gamma^2} + \frac{\eta^3}{3\gamma}\right) + \gamma^4\left(-\frac{3e^{-\gamma\eta}}{\gamma^4} + \frac{3}{\gamma^4} - \frac{\eta e^{-\gamma\eta}}{\gamma^3} - \frac{2\eta}{\gamma^3} + \frac{\eta^2}{2\gamma^2}\right); \\
\mu_{23} &= \gamma^5\frac{\eta^5}{5!} - \gamma^3\left(-\frac{e^{-\gamma\eta}}{\gamma^3} + \frac{1}{\gamma^3} - \frac{\eta}{\gamma^2} + \frac{\eta^2}{2\gamma}\right) \\
&+ \gamma^5\left(-\frac{e^{-\gamma\eta}}{\gamma^5} + \frac{1}{\gamma^5} - \frac{\eta}{\gamma^4} + \frac{\eta^2}{2\gamma^3} - \frac{\eta^3}{6\gamma^2} + \frac{\eta^4}{24\gamma}\right) + (1 - e^{-\gamma\eta}) \\
&+ \frac{\gamma^5}{3!}\left(-\frac{6e^{-\gamma\eta}}{\gamma^5} + \frac{6}{\gamma^5} - \frac{6\eta}{\gamma^4} + \frac{3\eta^2}{\gamma^3} - \frac{\eta^3}{\gamma^2} + \frac{\eta^4}{4\gamma}\right) \\
&+ \gamma^5\left(\frac{4e^{-\gamma\eta}}{\gamma^5} - \frac{4}{\gamma^5} + \frac{\eta e^{-\gamma\eta}}{\gamma^4} + \frac{3\eta}{\gamma^4} - \frac{\eta^2}{\gamma^3} + \frac{\eta^3}{6\gamma^2}\right); \\
\mu_{30} &= 0; \\
\mu_{31} &= \gamma^2\left(\frac{e^{-\gamma\eta}}{\gamma^2} - \frac{1}{\gamma^2} + \frac{\eta}{\gamma}\right) - \frac{\gamma^4}{3!}\left(\frac{6e^{-\gamma\eta}}{\gamma^4} - \frac{6}{\gamma^4} + \frac{6\eta}{\gamma^3} - \frac{3\eta^2}{\gamma^2} + \frac{\eta^3}{\gamma}\right) \\
&+ \frac{\gamma^6}{5!}\left(\frac{120e^{-\gamma\eta}}{\gamma^6} - \frac{120}{\gamma^6} + \frac{120\eta}{\gamma^5} - \frac{60\eta^2}{\gamma^4} + \frac{20\eta^3}{\gamma^3} - \frac{5\eta^4}{\gamma^2} + \frac{\eta^5}{\gamma}\right) \\
&+ \gamma^6\left(-\frac{5e^{-\gamma\eta}}{\gamma^6} + \frac{5}{\gamma^6} - \frac{\eta e^{-\gamma\eta}}{\gamma^5} - \frac{4\eta}{\gamma^5} + \frac{3\eta^2}{2\gamma^4} - \frac{\eta^3}{3\gamma^3} + \frac{\eta^4}{24\gamma^2}\right) \\
&- \gamma^4\left(-\frac{3e^{-\gamma\eta}}{\gamma^4} + \frac{3}{\gamma^4} - \frac{\eta e^{-\gamma\eta}}{\gamma^3} - \frac{2\eta}{\gamma^3} + \frac{\eta^2}{2\gamma^2}\right) \\
&+ \frac{\gamma^6}{3!}\left(-\frac{30e^{-\gamma\eta}}{\gamma^6} + \frac{30}{\gamma^6} - \frac{6\eta e^{-\gamma\eta}}{\gamma^5} - \frac{24\eta}{\gamma^5} + \frac{9\eta^2}{\gamma^4} - \frac{2\eta^3}{\gamma^3} + \frac{\eta^4}{4\gamma^2}\right) \\
&+ \gamma^6\left(\frac{10e^{-\gamma\eta}}{\gamma^6} - \frac{10}{\gamma^6} + \frac{4\eta e^{-\gamma\eta}}{\gamma^5} + \frac{6\eta}{\gamma^5} + \frac{\eta^2 e^{-\gamma\eta}}{2\gamma^4} - \frac{3\eta^2}{2\gamma^4} + \frac{\eta^3}{6\gamma^3}\right); \\
\mu_{32} &= -\gamma\left(\frac{1}{\gamma} - \frac{e^{-\gamma\eta}}{\gamma}\right) + \frac{\gamma^3}{2!}\left(-\frac{2e^{-\gamma\eta}}{\gamma^3} + \frac{2}{\gamma^3} - \frac{2\eta}{\gamma^2} + \frac{\eta^2}{\gamma}\right) \\
&- \frac{\gamma^5}{4!}\left(-\frac{24e^{-\gamma\eta}}{\gamma^5} + \frac{24}{\gamma^5} - \frac{24\eta}{\gamma^4} + \frac{12\eta^2}{\gamma^3} - \frac{4\eta^3}{\gamma^2} + \frac{\eta^4}{\gamma}\right) \\
&- \gamma^5\left(\frac{4e^{-\gamma\eta}}{\gamma^5} - \frac{4}{\gamma^5} + \frac{\eta e^{-\gamma\eta}}{\gamma^4} + \frac{3\eta}{\gamma^4} - \frac{\eta^2}{\gamma^3} + \frac{\eta^3}{6\gamma^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\gamma^3\left(\frac{2e^{-\gamma\eta}}{\gamma^3}-\frac{2}{\gamma^3}+\frac{\eta e^{-\gamma\eta}}{\gamma^2}+\frac{\eta}{\gamma^2}\right)\\
&-\frac{\gamma^5}{2!}\left(\frac{8e^{-\gamma\eta}}{\gamma^5}-\frac{8}{\gamma^5}+\frac{2\eta e^{-\gamma\eta}}{\gamma^4}+\frac{6\eta}{\gamma^4}-\frac{2\eta^2}{\gamma^3}+\frac{\eta^3}{3\gamma^2}\right)\\
&-\gamma^5\left(-\frac{6e^{-\gamma\eta}}{\gamma^5}+\frac{6}{\gamma^5}-\frac{3\eta e^{-\gamma\eta}}{\gamma^4}-\frac{3\eta}{\gamma^4}-\frac{\eta^2 e^{-\gamma\eta}}{2\gamma^3}+\frac{\eta^2}{2\gamma^3}\right);
\end{aligned}
$$

$$
\begin{aligned}
\mu_{33}=&\, e^{-\gamma\eta}-\frac{\gamma^6}{5!}\left(\frac{120e^{-\gamma\eta}}{\gamma^6}-\frac{120}{\gamma^6}+\frac{120\eta}{\gamma^5}-\frac{60\eta^2}{\gamma^4}+\frac{20\eta^3}{\gamma^3}-\frac{5\eta^4}{\gamma^2}+\frac{\eta^5}{\gamma}\right)\\
&+\gamma^4\left(-\frac{3e^{-\gamma\eta}}{\gamma^4}+\frac{3}{\gamma^4}-\frac{\eta e^{-\gamma\eta}}{\gamma^3}-\frac{2\eta}{\gamma^3}+\frac{\eta^2}{2\gamma^2}\right)\\
&-\gamma^6\left(-\frac{5e^{-\gamma\eta}}{\gamma^6}+\frac{5}{\gamma^6}-\frac{\eta e^{-\gamma\eta}}{\gamma^5}-\frac{4\eta}{\gamma^5}+\frac{3\eta^2}{2\gamma^4}-\frac{\eta^3}{3\gamma^3}+\frac{\eta^4}{24\gamma^2}\right)-\gamma^2\left(-\frac{e^{-\gamma\eta}}{\gamma^2}+\frac{1}{\gamma^2}-\frac{\eta e^{-\gamma\eta}}{\gamma}\right)\\
&-\frac{\gamma^6}{3!}\left(-\frac{30e^{-\gamma\eta}}{\gamma^6}+\frac{30}{\gamma^6}-\frac{6\eta e^{-\gamma\eta}}{\gamma^5}-\frac{24\eta}{\gamma^5}+\frac{9\eta^2}{\gamma^4}-\frac{2\eta^3}{\gamma^3}+\frac{\eta^4}{4\gamma^2}\right)\\
&+\gamma^4\left(\frac{3e^{-\gamma\eta}}{\gamma^4}-\frac{3}{\gamma^4}+\frac{2\eta e^{-\gamma\eta}}{\gamma^3}+\frac{\eta}{\gamma^3}+\frac{\eta^2 e^{-\gamma\eta}}{2\gamma^2}\right)\\
&-\gamma^6\left(\frac{10e^{-\gamma\eta}}{\gamma^6}-\frac{10}{\gamma^6}+\frac{4\eta e^{-\gamma\eta}}{\gamma^5}+\frac{6\eta}{\gamma^5}+\frac{\eta^2 e^{-\gamma\eta}}{2\gamma^4}-\frac{3\eta^2}{2\gamma^4}+\frac{\eta^3}{6\gamma^3}\right);
\end{aligned}
$$

Meanwhile, $f_i, 0\le i\le 4$ *are Itô integrals defined via*

$$
\begin{aligned}
f_0=&\,\gamma^2\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}\frac{\gamma(y-\eta(k+1))(-\gamma\eta(k+1)+\gamma y+2)-2e^{\gamma(y-\eta(k+1))}+2}{2\gamma^3}dB_y;\\
f_1=&\,\gamma^2\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}\frac{\gamma\eta+\gamma\eta k+e^{\gamma(y-\eta(k+1))}-\gamma y-1}{\gamma^2}dB_y;\\
f_2=&-\gamma^3\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}\frac{\gamma(y-\eta(k+1))(-\gamma\eta(k+1)+\gamma y+2)-2e^{\gamma(y-\eta(k+1))}+2}{2\gamma^3}dB_y\\
&-\gamma\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}\frac{e^{\gamma(y-\eta(k+1))}-1}{\gamma}dB_y\\
&-\gamma^3\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}\frac{\gamma\eta+\gamma\eta k+e^{\gamma(y-\eta(k+1))}(\gamma\eta(k+1)+\gamma(-y)+2)+\gamma(-y)-2}{\gamma^3}dB_y;\\
f_3=&\,\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}e^{-\gamma((k+1)\eta-y)}dB_y+\gamma^4\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}\frac{y-\eta k}{2\gamma^4}\Big(2e^{\gamma(y-\eta(k+1))}\\
&(-\gamma\eta(k+1)+\gamma y-3)+\gamma(y-\eta(k+1))(-\gamma\eta(k+1)+\gamma y+4)+6\Big)dB_y\\
&-\gamma^2\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}\frac{e^{\gamma(y-\eta(k+1))}(-\gamma\eta(k+1)+\gamma y-1)+1}{\gamma^2}dB_y
\end{aligned}
$$

$$+\gamma^4\sqrt{2\gamma}\int_{k\eta}^{(k+1)\eta}\frac{1}{2\gamma^4}\Big(2\gamma\eta(k+1)+e^{\gamma(y-\eta(k+1))}(\gamma(y-\eta(k+1))$$
$$\times(-\gamma\eta(k+1)+\gamma y-4)+6)-2\gamma y-6\Big)dB_y.$$

Next is the calculation for the conditional mean and covariance associated with the fourth-order LMC algorithm in Section 2.1.

Lemma B.2. *$\mathbb{E}\big[x^{(k+1)}|x^{(k)}\big]$ is a multivariate normal distribution with mean $\boldsymbol{M}(x^{(k)})=(m_i)_{0\le i\le 3}$ and symmetric covariance matrix $\boldsymbol{\Sigma}=(\sigma_{ij}\cdot I_d)_{0\le i,j\le 3}$.*

The entries $m_i, 0\le i\le 3$ and $\sigma_{ij}, 0\le i,j\le 3$ are provided below, noting that the constants $(\mu_{ij})_{0\le i,j\le 3}$ are defined in Lemma B.1.

$$m_0=-\int_{k\eta}^{(k+1)\eta}\Big((k+1)\eta-r\Big)\bar{g}(r)\,dr-\frac{\gamma^2}{6}\int_{k\eta}^{(k+1)\eta}\Big(y-(k+1)\eta\Big)^3\tilde{g}(y)\,dy$$
$$+\theta^{(k)}\mu_{00}+v_1^{(k)}\mu_{01}+v_2^{(k)}\mu_{02}+v_3^{(k)}\mu_{03};$$
$$m_1=-\int_{k\eta}^{(k+1)\eta}\bar{g}(s)\,ds+\frac{\gamma^2}{2}\int_{\eta k}^{\eta(k+1)}\Big(w-\eta(k+1)\Big)^2\tilde{g}(w)\,dw$$
$$+\theta^{(k)}\mu_{10}+v_1^{(k)}\mu_{11}+v_2^{(k)}\mu_{12}+v_3^{(k)}\mu_{13};$$
$$m_2=\gamma\int_{k\eta}^{(k+1)\eta}\Big((k+1)\eta-r\Big)\bar{g}(r)\,dr-\frac{\gamma^3}{6}\int_{k\eta}^{(k+1)\eta}\Bigg(\Big(y-(k+1)\eta\Big)^3$$
$$-\frac{6-6e^{\gamma(y-(k+1)\eta}+3\gamma(y-(k+1)\eta)\Big(2+\gamma(y-(k+1)\eta)\Big)}{\gamma^3}\Bigg)dy$$
$$+\theta^{(k)}\mu_{20}+v_1^{(k)}\mu_{21}+v_2^{(k)}\mu_{22}+v_3^{(k)}\mu_{23};$$
$$m_3=-\frac{\gamma^2}{2}\int_{k\eta}^{(k+1)\eta}\Big(w-\eta(k+1)\Big)^2\bar{g}(w)\,dw$$
$$+\gamma^4\int_{k\eta}^{(k+1)\eta}\Bigg(\frac{1}{2\gamma^4}\Big(2e^{\gamma(z-(k+1)\eta)}(-\gamma(k+1)\eta+\gamma z-3)$$
$$+\gamma(z-(k+1)\eta)(-\gamma(k+1)\eta+\gamma z+4)+6\Big)$$
$$+\frac{1}{6\gamma^4}\Big(\gamma((k+1)\eta-z)(\gamma((k+1)\eta-z)(\gamma(k+1)\eta+\gamma(-z)-3)+6)$$
$$+6e^{\gamma(z-(k+1)\eta)}-6\Big)\Bigg)\tilde{g}(z)dz$$
$$+\theta^{(k)}\mu_{30}+v_1^{(k)}\mu_{31}+v_2^{(k)}\mu_{32}+v_3^{(k)}\mu_{33}.$$

Furthermore,

$$\sigma_{00} = \frac{\gamma^3\eta^5}{10} - \frac{\gamma^2\eta^4}{2} - \frac{e^{-2\gamma\eta}}{\gamma^2} + \frac{4e^{-\gamma\eta}}{\gamma^2} - \frac{3}{\gamma^2} + \frac{4\gamma\eta^3}{3} + 2\eta^2 e^{-\gamma\eta} + \frac{2\eta}{\gamma} - 2\eta^2;$$

$$\sigma_{11} = \frac{2\gamma^3\eta^3}{3} - 2\gamma^2\eta^2 - 4\gamma\eta e^{-\gamma\eta} + 2\gamma\eta - e^{-2\gamma\eta} + 1;$$

$$\sigma_{22} = \frac{\gamma^5\eta^5}{10} - 2\gamma^3\eta^3 e^{-\gamma\eta} - \frac{4\gamma^3\eta^3}{3} - \gamma^2\eta^2 e^{-2\gamma\eta} - 10\gamma^2\eta^2 e^{-\gamma\eta} - 5\gamma\eta e^{-2\gamma\eta} - 12\gamma\eta e^{-\gamma\eta} + 8\gamma\eta - \frac{13}{2}e^{-2\gamma\eta} + 4e^{-\gamma\eta} + \frac{5}{2};$$

$$\sigma_{33} = \frac{\gamma^5\eta^7}{210} - \frac{\gamma^4\eta^6}{15} + \frac{\gamma^4\eta^5}{10} - \frac{1}{4}\gamma^4\eta^4 e^{-2\gamma\eta} + \frac{7\gamma^3\eta^5}{15} + \gamma^3\eta^4 e^{-\gamma\eta} - \frac{4\gamma^3\eta^4}{3} - \frac{7}{2}\gamma^3\eta^3 e^{-2\gamma\eta}$$
$$- 2\gamma^3\eta^3 e^{-\gamma\eta} + \frac{2\gamma^3\eta^3}{3} - 2\gamma^2\eta^4 - \frac{1}{2}\gamma^2\eta^3 e^{-2\gamma\eta} + 10\gamma^2\eta^3 e^{-\gamma\eta} + \frac{22\gamma^2\eta^3}{3} - \frac{77}{4}\gamma^2\eta^2 e^{-2\gamma\eta}$$
$$- 10\gamma^2\eta^2 e^{-\gamma\eta} - 8\gamma^2\eta^2 - \frac{21e^{-2\gamma\eta}}{2\gamma^2} + \frac{192e^{-\gamma\eta}}{\gamma^2} - \frac{363}{2\gamma^2} + 4\gamma\eta^3 e^{-\gamma\eta} + 6\gamma\eta^3 - \frac{1}{2}\eta^2 e^{-2\gamma\eta}$$
$$+ 32\eta^2 e^{-\gamma\eta} - 6\gamma\eta^2 e^{-2\gamma\eta} + 36\gamma\eta^2 e^{-\gamma\eta} - 24\gamma\eta^2 - \frac{101}{4}\eta e^{-2\gamma\eta} + 84\eta e^{-\gamma\eta} - \frac{197}{4}\gamma\eta e^{-2\gamma\eta}$$
$$+ 8\gamma\eta e^{-\gamma\eta} + 32\gamma\eta - \frac{9\eta e^{-2\gamma\eta}}{2\gamma} + \frac{96\eta e^{-\gamma\eta}}{\gamma} + \frac{159\eta}{2\gamma} - \frac{397}{8}e^{-2\gamma\eta} + 88e^{-\gamma\eta} - \frac{149e^{-2\gamma\eta}}{4\gamma}$$
$$+ \frac{204e^{-\gamma\eta}}{\gamma} - \frac{667}{4\gamma} - \frac{39\eta^2}{2} + \frac{283\eta}{4} - \frac{307}{8};$$

$$\sigma_{01} = \sigma_{10} = \frac{\gamma^3\eta^4}{4} - \gamma^2\eta^3 - \gamma\eta^2 e^{-\gamma\eta} + 2\gamma\eta^2 + 2\eta e^{-\gamma\eta} + \frac{e^{-2\gamma\eta}}{\gamma} - \frac{2e^{-\gamma\eta}}{\gamma} + \frac{1}{\gamma} - 2\eta;$$

$$\sigma_{02} = \sigma_{20} = -\frac{1}{10}\gamma^4\eta^5 + \frac{\gamma^3\eta^4}{4} + \gamma^2\eta^3 e^{-\gamma\eta} + \frac{\gamma^2\eta^3}{3} + 2\gamma\eta^2 e^{-\gamma\eta} - 2\gamma\eta^2 - \eta e^{-2\gamma\eta} + 2\eta e^{-\gamma\eta} - \frac{5e^{-2\gamma\eta}}{2\gamma} + \frac{10e^{-\gamma\eta}}{\gamma} - \frac{15}{2\gamma} + 4\eta;$$

$$\sigma_{30} = \sigma_{03} = \frac{\gamma^4\eta^6}{60} - \frac{3\gamma^3\eta^5}{20} - \frac{1}{2}\gamma^3\eta^4 e^{-\gamma\eta} + \frac{\gamma^3\eta^4}{4} + \frac{2\gamma^2\eta^4}{3} - 4\gamma^2\eta^3 e^{-\gamma\eta} - 2\gamma^2\eta^3 + \frac{2e^{-2\gamma\eta}}{\gamma^2}$$
$$- \frac{32e^{-\gamma\eta}}{\gamma^2} + \frac{30}{\gamma^2} - \gamma\eta^3 e^{-\gamma\eta} - \frac{5\gamma\eta^3}{3} - 8\eta^2 e^{-\gamma\eta} + \frac{1}{2}\gamma\eta^2 e^{-2\gamma\eta} - 12\gamma\eta^2 e^{-\gamma\eta} + 5\gamma\eta^2$$
$$+ \frac{7}{2}\eta e^{-2\gamma\eta} - 18\eta e^{-\gamma\eta} + \frac{\eta e^{-2\gamma\eta}}{2\gamma} - \frac{18\eta e^{-\gamma\eta}}{\gamma} - \frac{21\eta}{2\gamma} + \frac{27e^{-2\gamma\eta}}{4\gamma} - \frac{36e^{-\gamma\eta}}{\gamma} + \frac{117}{4\gamma} + 3\eta^2 - 8\eta;$$

$$\sigma_{12} = \sigma_{21} = -\frac{1}{4}\gamma^4\eta^4 + \frac{\gamma^3\eta^3}{3} + 3\gamma^2\eta^2 e^{-\gamma\eta} + 2\gamma^2\eta^2 + \gamma\eta\, e^{-2\gamma\eta} + 8\gamma\eta e^{-\gamma\eta} - 4\gamma\,\eta + \frac{5}{2}e^{-2\gamma\eta} - \frac{5}{2};$$

$$\sigma_{13} = \sigma_{31} = \frac{\gamma^4\eta^5}{20} - \frac{5\gamma^3\eta^4}{12} - \gamma^3\eta^3 e^{-\gamma\eta} + \frac{2\gamma^3\eta^3}{3} + \frac{5\gamma^2\eta^3}{3} - \frac{1}{2}\gamma^2\eta^2 e^{-2\gamma\eta} - 8\gamma^2\eta^2 e^{-\gamma\eta} - 5\gamma^2\eta^2$$
$$- \gamma\eta^2 e^{-\gamma\eta} - 3\gamma\eta^2 - \frac{1}{2}\eta e^{-2\gamma\eta} - 12\eta e^{-\gamma\eta} - \frac{7}{2}\gamma\eta e^{-2\gamma\eta} - 22\gamma\eta e^{-\gamma\eta} + 8\gamma\eta - \frac{27}{4}e^{-2\gamma\eta}$$
$$- 4e^{-\gamma\eta} - \frac{2e^{-2\gamma\eta}}{\gamma} - \frac{10e^{-\gamma\eta}}{\gamma} + \frac{12}{\gamma} - \frac{3\eta}{2} + \frac{43}{4};$$

$$\sigma_{23} = \sigma_{32} = -\frac{1}{60}\gamma^5\eta^6 + \frac{\gamma^4\eta^5}{10} + \frac{1}{2}\gamma^4\eta^4 e^{-\gamma\eta} - \frac{\gamma^4\eta^4}{4} - \frac{\gamma^3\eta^4}{12} + \frac{1}{2}\gamma^3\eta^3 e^{-2\gamma\eta} + 5\gamma^3\eta^3 e^{-\gamma\eta}$$
$$+ \frac{4\gamma^3\eta^3}{3} - \frac{4\gamma^2\eta^3}{3} + \frac{19}{4}\gamma^2\eta^2 e^{-2\gamma\eta} + 20\gamma^2\eta^2 e^{-\gamma\eta} + 2\gamma^2\eta^2 + \frac{1}{2}\gamma\eta^2 e^{-2\gamma\eta} + 5\gamma\eta^2 e^{-\gamma\eta}$$
$$+ 6\gamma\eta^2 + \frac{7}{2}\eta e^{-2\gamma\eta} + 18\eta e^{-\gamma\eta} + \frac{63}{4}\gamma\eta e^{-2\gamma\eta} + 24\gamma\eta e^{-\gamma\eta} - 16\gamma\eta + \frac{143}{8} e^{-2\gamma\eta}$$
$$- 12 e^{-\gamma\eta} + \frac{25 e^{-2\gamma\eta}}{4\gamma} + \frac{2 e^{-\gamma\eta}}{\gamma} - \frac{33}{4\gamma} - 7\eta - \frac{47}{8}.$$

Proof. The formula for m_i immediately follows from Lemma B.1, so that we only need to show how to compute the entries of the covariance matrix. We have

$$\mathbb{E}\Big[\big(\theta^{(k+1)} - \mathbb{E}\big[\theta^{(k+1)}|x^{(k)}\big]\big)\big(\theta^{(k+1)} - \mathbb{E}\big[\theta^{(k+1)}|x^{(k)}\big]\big)^\top \Big| x^{(k)}\Big] = \mathbb{E}\big[f_0(f_0)^\top\big] = \sigma_{00}\cdot I_d, \quad (54)$$

where f_0 is defined in Lemma B.1. Then σ_{00} on the right hand side of (54) can be computed by Itô isometry and software. The remaining covariance entries σ_{ij}'s are obtained in the same way via $\mathbb{E}\big[f_i(f_j)^\top\big] = \sigma_{ij}\cdot I_d$. ☐

Based on Lemma B.1, it possible to decompose $\bar{x}((k+1)\eta)$ into higher and lower order terms with respect to η, as the next lemma will show.

Lemma B.3. *Recall the unique minimizer θ^* of U, $\boldsymbol{M}(x^{(k)})$ from Lemma B.2 and the Jacobian matrix*
$$J_b(\theta^*, 0, \ldots, 0) = \begin{bmatrix} 0 & I_d & 0 & 0 \\ -\nabla^2 U(\theta^*) I_d & 0 & \gamma & 0 \\ 0 & -\gamma I_d & 0 & \gamma \\ 0 & 0 & -\gamma I_d & -\gamma I_d \end{bmatrix}.$$
Then it holds that

$$\boldsymbol{M}(x^{(k)}) = x^{(k)} + \eta J_b(\theta^*, 0, \ldots, 0)\big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) + R\big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big),$$

and

$$x^{(k+1)} - (\theta^*, 0, \ldots, 0) = \big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big)$$
$$+ \eta J_b(\theta^*, 0, \ldots, 0)\big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) + R\big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) + F_k,$$

where R is a $4d \times 4d$ matrix with $|R_{ij}| \le C\eta^2$, $1 \le i, j \le 4d$ and C is a constant dependent only on γ. Moreover, the entries f_i's of the $4d$-dimensional vector $F_k = (f_0 \quad f_1 \quad f_3 \quad f_4)^\top$ are defined in Lemma B.1.

Proof. Without loss of generality, let us assume that the unique minimizer $\theta^* = 0$.

First, we will rewrite $\mu_{ij}, 0 \le i, j \le 3$ that appear in the formula of $\bar{x}((k+1)\eta)$ in Lemma B.1 and make explicit lower and higher order terms with respect to η.

$$\mu_{00} = 1;$$
$$\mu_{01} = \eta - \sum_{k=2}^{\infty} \frac{\gamma^4(-\gamma\eta)^k}{k!} - \gamma^2\frac{\eta^3}{3!} + \gamma^4\frac{\eta^5}{5!} + \gamma^4\left(\frac{\eta^2}{2\gamma^3} - \frac{\eta^3}{6\gamma^2} + \frac{\eta^4}{24\gamma}\right) = \eta - O(\eta^2);$$

$$\mu_{02} = \gamma\frac{\eta^2}{2!} - \gamma^3\frac{\eta^4}{4!} - \gamma^3\left(\sum_{k=2}^{\infty}\frac{(-\gamma\eta)^k}{\gamma^4 k!} - \frac{\eta^2}{2\gamma^2} + \frac{\eta^3}{6\gamma}\right) = O(\eta^2);$$
$$\mu_{03} = -\gamma^4\frac{\eta^5}{5!} + \frac{\gamma\eta^2}{2} - \gamma^4\left(\frac{\eta^2}{2\gamma^3} - \frac{\eta^3}{6\gamma^2} + \frac{\eta^4}{24\gamma}\right) = O(\eta^2).$$

Similarly, we have

$$\mu_{10} = 0, \quad \mu_{11} = 1 + O(\eta^2), \quad \mu_{12} = -\gamma\eta + O(\eta^2), \quad \mu_{13} = O(\eta^2),$$
$$\mu_{20} = 0, \quad \mu_{21} = -\gamma\eta + O(\eta^2), \quad \mu_{22} = 1 + O(\eta^2), \quad \mu_{23} = \eta\gamma + O(\eta^2),$$
$$\mu_{30} = 0, \quad \mu_{31} = O(\eta^2), \quad \mu_{32} = -\gamma\eta + O(\eta^2), \quad \mu_{33} = 1 - \gamma\eta + O(\eta^2).$$

Next, we consider the integral terms containing $\hat{g}$ and $\bar{g}$ in the formula of $\bar{x}((k+1)\eta)$ in Lemma B.1. Since $\nabla U(\theta^*) = \nabla U(0) = 0$, we can write

$$-\int_{k\eta}^{(k+1)\eta} \bar{g}(s)ds = -\eta\nabla U(\theta^{(k)}) + O(\eta^2) = \eta\nabla^2 U(0) I_d \theta^{(k)} + O(\eta^2).$$

Meanwhile, all the remaining integral terms containing $\hat{g}$ and $\bar{g}$ in $\bar{x}((k+1)\eta)$ are of the order $O(\eta^2)$.

Consequently, we can deduce the equation in the statement of this lemma from the above calculations and Lemma B.1. □

We also need the following moment bounds.

Lemma B.4. *Assume*

$$\eta \le \eta^* := \frac{\rho}{2}\left\|M\right\|_{\mathrm{op}}^{-1}\left\|M^{-1}\right\|_{\mathrm{op}}^{-1}\frac{1}{1+5\gamma^2+L^2}, \tag{55}$$

where the matrix M is provided in Example 2.3 and γ, ρ, L are from Theorem 2.1. Then there exists a positive constant C_1 such that for all k,

$$\mathbb{E}\left[\left|x^{(k+1)}\right|^{2\alpha}\right] \le (C_1)^\alpha(d+2\alpha)^\alpha,$$

where C_1 depends on γ, L from Condition H1 and c from Condition H2, but not on d.

This further implies

$$\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[|\hat{x}(t)|^2 + |\tilde{x}(t)|^2 + |\bar{x}(t)|^2 + |\tilde{g}(t)|^2 + |\bar{g}(t)|^2\right] \le C_2(d+1),$$

for a universal constant $C_2 \ge 1$ that depends on γ, L from Condition H1 and c from Condition H2. C_2 does not depend on the dimension d.

Proof. **Part 1 of the proof:**

Without loss of generality, let us assume $\theta^* = 0$ and denote $J_b(0) = J_b(\theta^*, 0, \dots, 0)$. Let $w_k \sim \mathcal{N}(0, I_{4d})$ then per Lemma B.2 and Lemma B.3,

$$\mathbb{E}\left[\left|x^{(k+1)}\right|_M^{2\alpha}\right] = \mathbb{E}\left[\left|M^{1/2}\mathbf{M}(x^{(k)}) + (M\mathbf{\Sigma})^{1/2}w_k\right|^{2\alpha}\right]$$

$$
\begin{aligned}
&\leq \sum_{j=0}^{2\alpha} \binom{2\alpha}{j} \mathbb{E}\Big[\big|M^{1/2}\mathbf{M}(x^{(k)})\big|^{j}\Big] \mathbb{E}\Big[\Big|(M\boldsymbol{\Sigma})^{1/2} w_k\Big|^{2\alpha-j}\Big] \\
&\leq \sum_{j=0}^{2\alpha} \binom{2\alpha}{j} \mathbb{E}\Big[\big|M^{1/2}\mathbf{M}(x^{(k)})\big|^{2\alpha}\Big]^{j/2\alpha} \mathbb{E}\Big[\Big|(M\boldsymbol{\Sigma})^{1/2} w_k\Big|^{2\alpha}\Big]^{1-j/2\alpha} \\
&= \left(\mathbb{E}\Big[\big|\mathbf{M}(x^{(k)})\big|_M^{2\alpha}\Big]^{1/\alpha} + \mathbb{E}\Big[\Big|(M\boldsymbol{\Sigma})^{1/2} w_k\Big|^{2\alpha}\Big]^{1/\alpha} \right)^{2\alpha}. \qquad (56)
\end{aligned}
$$

Let us study the first term on the right hand side. Lemma B.3 and (8) imply that

$$
\begin{aligned}
\big|\mathbf{M}(x^{(k)})\big|_M^2 &= \big|x^{(k)} + \eta J_b(0) x^{(k)}\big|_M^2 \\
&= \big|x^{(k)}\big|_M^2 + \eta\big(x^{(k)}\big)^\top \big(J_b(0)^\top M + M J_b(0)\big) x^{(k)} + \eta^2 \big(x^{(k)}\big)^\top J_b(0)^\top M J_b(0) x^{(k)} \\
&\leq \big|x^{(k)}\big|_M^2 + \eta (x^{(k)})^\top (-2\rho) M x^{(k)} + \eta^2 \left\|J_b(0)^\top M J_b(0)\right\|_{\mathrm{op}} \left\|M^{-1}\right\|_{\mathrm{op}} \big|x^{(k)}\big|_M^2 \\
&\leq (1 - 2\eta\rho) \big|x^{(k)}\big|_M^2 + \eta^2 \left\|J_b(0)\right\|_{\mathrm{op}}^2 \left\|M\right\|_{\mathrm{op}} \left\|M^{-1}\right\|_{\mathrm{op}} \big|x^{(k)}\big|_M^2 .
\end{aligned}
$$

At this point, notice that the operator norm is bounded from above by the Frobenius norm, so that combining with L-smoothness of U and the explicit form of J_b in Lemma B.3, we get $\|J_b(0)\|_{\mathrm{op}}^2 \leq 1 + 5\gamma^2 + L^2$. $\|M\|_{\mathrm{op}}$ and $\|M^{-1}\|_{\mathrm{op}}$ can be computed using the explicit form of M in Example 2.3. Then assuming $\eta < \eta^*$ as defined in the statement of the lemma, we arrive at

$$
\mathbb{E}\Big[\big|\mathbf{M}(x^{(k)})\big|_M^{2\alpha}\Big]^{1/\alpha} \leq \left(1 - \frac{3}{2}\eta\rho\right)^{\alpha} \mathbb{E}\Big[\big|x^{(k)}\big|_M^{2\alpha}\Big].
$$

This can be combined with (56) and (20) to get the desired bound on $\mathbb{E}\Big[\big|x^{(k+1)}\big|^{2\alpha}\Big]$.

Part 2 of the proof:

In this part, we will use the result from Part 1 and the explicit formula at the beginning of Appendix B to bound $\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\big[|\hat{x}(t)|^2 + |\tilde{x}(t)|^2 + |\bar{x}(t)|^2 + |\tilde{g}(t)|^2 + |\bar{g}(t)|^2\big]$. The upcoming argument is rather tedious, so we will only demonstrate how to bound $\sup_{s\in[k\eta,(k+1)\eta]} \mathbb{E}\big[|\tilde{v}_1(t)|^2\big]$. By Equation (53), we can write

$$
\begin{aligned}
\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\big[|\tilde{v}_1(t)|^2\big] \leq 5\Bigg(&\mathbb{E}\Big[\Big|v_1^{(k)}\Big|^2\Big] + \gamma^2\eta^2 \mathbb{E}\Big[\Big|v_2^{(k)}\Big|^2\Big] + \gamma^4 \frac{\eta^4}{4}\left(\mathbb{E}\Big[\Big|v_1^{(k)}\Big|^2\Big] + \mathbb{E}\Big[\Big|v_3^{(k)}\Big|^2\Big]\right) \\
&+ \eta^2 \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}[|\tilde{g}(t)|]^2 \Bigg).
\end{aligned}
$$

Lemma 2.16 and L-smoothness of U implies

$$
\mathbb{E}[|\tilde{g}(t)|]^2 \leq \left(\frac{L_\alpha}{\alpha!}\right)^2 \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\Big[\Big|\hat{\theta}(t)\Big|^{2\alpha}\Big] + L_\alpha^2 \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\Big[\Big|\hat{\theta}(t)\Big|^2\Big],
$$

where α is from Condition H2. Moreover, we also have from (52) and Part 1 of this proof that

$$\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\hat{\theta}(t)\right|^{2\alpha}\right] \leq \mathbb{E}\left[\left(2\left|\theta^{(k)}\right| \vee 2\left|v_1^{(k)}\right|\right)^{2\alpha}\right]$$

$$\leq 4\mathbb{E}\left[\left|\theta^{(k)}\right|^{2\alpha}\right] + 4\mathbb{E}\left[\left|v_1^{(k)}\right|^{2\alpha}\right] \leq 8(C_1)^{\alpha}(d+2\alpha)^{\alpha}.$$

Condition H2 then says $\left(\frac{L_\alpha}{\alpha!}\right)^2 \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\hat{\theta}(t)\right|^{2\alpha}\right] \leq cd\eta^2$. Thus, we arrive at

$$\mathbb{E}[|\tilde{g}(t)|]^2 \leq cd\eta^2 + L^2 8C_1(d+2).$$

Hence,

$$\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[|\tilde{v}_1(t)|^2\right] \leq 5\left(1+\gamma^2\eta^2+\gamma^4\frac{\eta^4}{4}\right)4C_1(d+2) + 5\eta^2\big(cd\eta^2 + L^2 8C_1(d+2)\big).$$

This completes the proof. $\square$

Next are the proofs of Lemma 2.16 and Lemma 2.17 in the main paper.

Proof of Lemma 2.16. By Condition H1 and [Car71, Theorem 5.6.2], we have

$$|\nabla U(x) - P_{\alpha-1}(x)| \leq L_\alpha \frac{|x|^\alpha}{\alpha!}.$$

This leads to

$$\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\nabla U(\hat{\theta}(t)) - \tilde{g}(t)\right|^2\right] = \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\nabla U(\hat{\theta}(t)) - P_{\alpha-1}(\hat{\theta}(t))\right|^2\right]$$

$$\leq \left(\frac{L_\alpha}{\alpha!}\right)^2 \sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\hat{\theta}(t)^{2\alpha}\right|\right].$$

The bound on $\sup_{t\in[k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\nabla U(\tilde{\theta}(t)) - \bar{g}(t)\right|^2\right]$ is obtained in the same way. $\square$

Proof of Lemma 2.17. **First part of the proof:** we will bound the difference of the components of $\tilde{x}(t) - \hat{x}(t)$ in L^2-norm for $t \in (k\eta, (P=1)\eta]$.

We start with $\tilde{v}_1(t) - \hat{v}_1(t) = \int_{k\eta}^t (-\tilde{g}(s) + \gamma\hat{v}_2(s))\, ds$, which combined with the moment bounds in Lemma B.4 leads to

$$\mathbb{E}\left[|\tilde{v}_1(t) - \hat{v}_1(t)|^2\right] \leq C_2(d+1)(\gamma+1)^2(t-k\eta)^2.$$

Next, $\tilde{\theta}(t) - \hat{\theta}(t) = \int_{k\eta}^t (\tilde{v}_1(s) - \hat{v}_1(s))\, ds$ combined with Lemma B.4 leads to

$$\mathbb{E}\left[\left|\tilde{\theta}(t) - \hat{\theta}(t)\right|^2\right] \leq C_2(d+1)(\gamma+1)^2(t-k\eta)^4.$$

Moreover,

$$\tilde{v}_2(t) - \hat{v}_2(t) = \int_{k\eta}^{t} \Big(-\gamma(\tilde{v}_1(s) - \hat{v}_1(s)) + \gamma\Big(\hat{v}_3(s) - v_3^{(k)}\Big)\Big)\, ds.$$

We know $\hat{v}_3(t) - v_3^{(k)} = \int_{k\eta}^{t} (-\gamma\hat{v}_3(s) - \gamma\hat{v}_2(s))\, ds + \sqrt{2\gamma}(B_t - B_{k\eta})$ which along with Lemma B.4 imply the bound

$$\mathbb{E}\bigg[\Big|\hat{v}_3(t) - v_3^{(k)}\Big|^2\bigg] \le C_2(d+1)(2\gamma + \sqrt{2\gamma})^2(t - k\eta).$$

Consequently,

$$\begin{aligned}\mathbb{E}\big[|\tilde{v}_2(t) - \hat{v}_2(t)|^2\big] &\le C_2(d+1)(\gamma+1)^2(t-k\eta)^4 + C_2(d+1)(2\gamma + \sqrt{2\gamma})^2(t-k\eta)^3 \\ &\le C_2(d+1)\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big)(t-k\eta)^3. \qquad (57)\end{aligned}$$

Finally, $\tilde{v}_3(t) - \hat{v}_3(t) = \int_{k\eta}^{t} -\gamma e^{-\gamma s}(\tilde{v}_2(s) - \hat{v}_2(s))ds$ and the moment bound in Lemma B.4 imply

$$\mathbb{E}\big[|\tilde{v}_3(t) - \hat{v}_3(t)|^2\big] \le C_2(d+1)\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big)(t-k\eta)^5.$$

Second part of the proof: We will bound the difference of the components of $\bar{x}(t) - \tilde{x}(t)$ in L^2 norm for $t \in (k\eta, (k+1)\eta]$.

We start with

$$\begin{aligned}\bar{v}_1(t) - \tilde{v}_1(t) &= \int_{k\eta}^{t} -(\bar{g}(s) - \tilde{g}(s))ds + \int_{k\eta}^{t} \gamma(\tilde{v}_2(s) - \hat{v}_2(s))ds \\ &= \int_{k\eta}^{t} \bigg(- \Big(\bar{g}(s) - \nabla U(\tilde{\theta}(s))\Big) - \Big(\nabla U(\tilde{\theta}(s)) - \nabla U(\hat{\theta}(s))\Big) \\ &\qquad - \Big(\tilde{g}(s) - \nabla U(\hat{\theta}(s))\Big)\bigg)ds + \int_{k\eta}^{t} \gamma(\tilde{v}_2(s) - \hat{v}_2(s))ds. \qquad (58)\end{aligned}$$

The bound in (57) leads to

$$\mathbb{E}\bigg[\Big|\int_{k\eta}^{t} \gamma(\tilde{v}_2(s) - \hat{v}_2(s))ds\Big|^2\bigg] \le C_2(d+1)\gamma^2\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big)(t-k\eta)^5.$$

Meanwhile, Lemma 2.16, Lemma B.4 and Condition H2 say

$$\mathbb{E}\bigg[\Big|\int_{k\eta}^{t} \Big(\bar{g}(s) - \nabla U(\tilde{\theta}(s))\Big)\, ds\Big|^2\bigg] \le \Big(\frac{L_\alpha}{\alpha!}\Big)^2 (C_1)^\alpha (d + 2\alpha)^\alpha \eta^2 \le cd\eta^9. \qquad (59)$$

Similarly, we have $\mathbb{E}\bigg[\Big|\int_{k\eta}^{t} \Big(\tilde{g}(s) - \nabla U(\hat{\theta}(s))\Big)\, ds\Big|^2\bigg] \le cd\eta^9$. Also based on the first part of the proof and L-smoothness of U in Condition H1,

$$\mathbb{E}\bigg[\Big|\int_{k\eta}^{t} \Big(\nabla U(\tilde{\theta}(s)) - \nabla U(\bar{\theta}(s))\Big)\, ds\Big|^2\bigg] \le C_2(d+1)L^2(\gamma+1)^2(t-k\eta)^6.$$

By combining the previous calculations, we get for any $t \in (k\eta, (k+1)\eta]$ that

$$\mathbb{E}\big[|\bar{v}_1(t) - \tilde{v}_1(t)|^2\big] \le C_2(d+1)\Big(\gamma^2\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big) + c\Big)\eta^5.$$

Next, $\tilde{\theta}(t) - \bar{\theta}(t) = \int_{k\eta}^{t} (\tilde{v}_1(s) - \bar{v}_1(s))\, ds$ leads to

$$\mathbb{E}\left[\left|\tilde{\theta}(t) - \bar{\theta}(t)\right|^2\right] \le C_2(d+1)\Big(\gamma^2\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big) + c\Big)\eta^7.$$

Moreover, $\tilde{v}_2(t) - \bar{v}_2(t) = \int_{k\eta}^{t} (-\gamma(\tilde{v}_1(s) - \bar{v}_1(s)) + \gamma(\hat{v}_3(s) - \tilde{v}_3(s)))\, ds$ leads to

$$\begin{aligned}\mathbb{E}\big[|\bar{v}_2(t) - \tilde{v}_2(t)|^2\big] \le C_2(d+1)\gamma^2\Big(\gamma^2\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big) + c\Big)\eta^7 \\ + C_2(d+1)\gamma^2\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big)\eta^7.\end{aligned}$$

Finally, $\tilde{v}_3(t) - \bar{v}_3(t) = \int_{k\eta}^{t} -\gamma e^{-\gamma s}(\tilde{v}_2(s) - \bar{v}_2(s))ds$ implies

$$\begin{aligned}\mathbb{E}\big[|\bar{v}_3(t) - \tilde{v}_3(t)|^2\big] \le C_2(d+1)\gamma^4\Big(\gamma^2\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big) + c\Big)\eta^9 \\ + C_2(d+1)\gamma^4\Big((\gamma+1)^2 + (2\gamma + \sqrt{2\gamma})^2\Big)\eta^9.\end{aligned}$$

This completes the proof. □

Appendix C. Details of P-th Order Langevin Monte Carlo Algorithm

The following is a generalized version of Lemma B.1.

Lemma C.1. *Choose any positive integers i and j in $[1, P-1]$. Then*

1. *the auxiliary process $\{v_i^{\mathrm{st}_j}(t), t \ge 0\}$ in Section 2.2 has the form*

$$v_i^{\mathrm{st}_j}(t) = \sum_{1\le \ell \le P-1} v_\ell^{(k)} \mu_{i,\ell}^{\mathrm{st}_j}(t) + \theta^{(k)}\mu_{i,P}^{\mathrm{st}_j}(t) + \mathbb{1}_{\{i \ge P-j\}} \int_{k\eta}^{t} h_i^{\mathrm{st}_j}(s,t) dB_s + \sum_{2\le \ell\le j} \int_{k\eta}^{t} \kappa_{i,\ell}^{\mathrm{st}_j}(s,t) g^{\mathrm{st}_\ell}(s) ds, \qquad (60)$$

such that

a. *The kernel $h_i^{\mathrm{st}_j}(s,t)$ is deterministic and has the form*

$$h_i^{\mathrm{st}_j}(s,t) = \sum_{0\le m\le M_1} a_{1,m} e^{b_{1,m}(t-s)+c_{1,m}} (s - k\eta)^{d_{1,m}}, \qquad (61)$$

where M_1 is a positive integer; $d_{1,m}$'s are non-negative integers; $a_{1,m}, b_{1,m}, c_{1,m}$'s are rational functions in variables k, η, γ, t, i.e. they are ratios of multivariate polynomials in k, η, γ, t. [3]

[3]The coefficients $a_{1,m}, b_{1,m}, c_{1,m}$ and $d_{1,m}$ on the right hand side of (61) depend on i, j; however we hide this dependence to lighten the notations. We do the same thing in Equation (62) and Equation (63).

b. $\mu_{i,\ell}^{\mathrm{st}_j}(t)$ *has the form*

$$\mu_{i,\ell}^{\mathrm{st}_j}(t) = \sum_{0\le m\le M_2} a_{2,m} e^{b_{2,m}(t-k\eta)+c_{2,m}} (t-k\eta)^{d_{2,m}}, \tag{62}$$

where M_2 *is a positive integer;* $a_{2,m}, b_{2,m}, c_{2,m}$*'s are rational functions in variables* k, η, γ*; and* $d_{2,m}$*'s are non-negative integers.*

c. $\kappa_{i,\ell}^{\mathrm{st}_j}(s,t)$ *is deterministic and has the form*

$$\kappa_{i,\ell}^{\mathrm{st}_j}(s,t) = \sum_{0\le m\le M_3} a_{3,m} e^{b_{3,m}(t-s)+c_{3,m}} d_{3,m}(s), \tag{63}$$

where M_3 *is a positive integer;* $d_{3,m}$*'s are polynomial in* s*;* $a_{3,m}, b_{3,m}, c_{3,m}$*'s are rational functions in variables* k, η, γ, t*, i.e. they are ratios of multivariate polynomials in* k, η, γ, t*.*

2. the auxiliary process $\{\theta^{\mathrm{st}_j}(t), t \ge 0\}$ *in Section 2.2 has the form*

$$\theta^{\mathrm{st}_j}(t) = \sum_{1\le \ell\le P-1} v_\ell^{(k)} \mu_{P,\ell}^{\mathrm{st}_j}(t) + \theta^{(k)} \mu_{P,P}^{\mathrm{st}_j}(t) + \mathbb{1}_{\{j=P-1\}} \int_{k\eta}^t h_P^{\mathrm{st}_j}(s,t) dB_s + \sum_{2\le \ell\le j} \int_{k\eta}^t \kappa_{P,\ell}^{\mathrm{st}_j}(s,t) g^{\mathrm{st}_\ell}(s) ds, \tag{64}$$

such that $\mu_{P,\ell}(t)$ *has a similar form to* (62)*, while* $h_P^{\mathrm{st}_j}(s,t)$ *and* $\kappa_{P,\ell}^{\mathrm{st}_j}(s,t)$ *have similar forms to respectively* (61) *and* (63)*.*

Proof. We will employ an induction argument.

Step 1: Stage $j = 1$.

We will verify that $v_n^{\mathrm{st}_1}(t), 1 \le n \le P-1$ and $\theta^{\mathrm{st}_1}(t)$ have respectively the general forms (60) and (64).

We have $v_1^{\mathrm{st}_1}(t) = v_1^{(k)}$ so that

$$\theta_1^{\mathrm{st}_1}(t) = v_1^{(k)}(t-k\eta),$$

and

$$v_2^{\mathrm{st}_1}(t) = v_2^{(k)} - \gamma \int_{k\eta}^t v_1^{\mathrm{st}_1}(s) ds + \gamma v_3^{(k)}(t-k\eta) = v_2^{(k)} - \gamma(t-k\eta) v_1^{(k)} + \gamma v_3^{(k)}(t-k\eta). \tag{65}$$

Proceed similarly for increasing n to get for $3 \le n \le P-2$,

$$v_n^{\mathrm{st}_1}(t) = v_n^{(k)} - \gamma \int_{k\eta}^t v_{n-1}^{\mathrm{st}_1}(s) ds + \gamma v_{n-1}^{(k)}(t-k\eta) = \sum_{\ell=1}^{P-1} a_\ell (t-k\eta)^{d_\ell} v_\ell^{(k)}, \tag{66}$$

where a_ℓ's are polynomials in γ and d_ℓ's are non-negative integers. If we set $\mu_{n,\ell}^{\mathrm{st}_1}(t) := a_\ell (t-k\eta)^{d_\ell}$ then this coefficient is of the form described in the statement of the lemma.

Moreover, the formula (66) in the case $n = P - 2$ implies

$$\begin{aligned} v_{P-1}^{\text{st}_1}(t) &= e^{-\gamma(t-k\eta)} v_{P-1}^{(k)} - \gamma \int_{k\eta}^{t} e^{-\gamma(t-s)} v_{P-2}^{\text{st}_1}(s) ds + \sqrt{2\gamma} \int_{k\eta}^{t} e^{-\gamma(t-s)} dB_s \\ &= e^{-\gamma(t-k\eta)} v_{P-1}^{(k)} + \sqrt{2\gamma} \int_{k\eta}^{t} e^{-\gamma(t-s)} dB_s - \sum_{\ell=1}^{P-1} v_\ell^{(k)} \gamma a_\ell \int_{k\eta}^{t} e^{-\gamma(t-s)} (s - k\eta)^{d_\ell} ds. \end{aligned}$$

Via integration by parts, it is easy to see

$$\int_{k\eta}^{t} e^{-\gamma(t-s)} (s - k\eta)^{d_\ell} ds = \sum_j e^{-\gamma(t-k\eta)} a_{\ell,j} (t - k\eta)^{d_{\ell,j}}, \tag{67}$$

where $d_{\ell,j}$'s are non-negative integers and $a_{\ell,j}$'s are polynomials in γ. Setting

$$\mu_{P-1,\ell}^{\text{st}_1}(t) := \gamma a_\ell \sum_j e^{-\gamma(t-k\eta)} a_{\ell,j} (t - k\eta)^{d_{\ell,j}},$$

for $1 \le \ell \le P - 2$ and

$$\mu_{P-1,P-1}^{\text{st}_1}(t) := \gamma a_{P-1} \sum_j e^{-\gamma(t-k\eta)} a_{P-1,j} (t - k\eta)^{d_{P-1,j}} + e^{-\gamma(t-k\eta)},$$

we arrive at

$$v_{P-1}^{\text{st}_1}(t) = \sqrt{2\gamma} \int_{k\eta}^{t} e^{-\gamma(t-s)} dB_s + \sum_{\ell=1}^{P-1} v_\ell^{(k)} \mu_{P-1,\ell}^{\text{st}_1}(t). \tag{68}$$

Finally, notice that among $\theta_1^{\text{st}_1}(t)$ and $v_n^{\text{st}_1}(t), 1 \le n \le P - 1$, the Itô integral only appears in $v_{P-1}^{\text{st}_1}(t)$, which explains the indicator functions in (60) and (64) when $j = 1$.

Step 2: Stage $j = 2$.

We will verify that $v_n^{\text{st}_2}(t), 1 \le n \le P - 1$ and $\theta^{\text{st}_2}(t)$ have respectively the general forms (60) and (64).

We have based on (65) that

$$\begin{aligned} v_1^{\text{st}_2}(t) &= v_1^{(k)} - \int_{k\eta}^{t} g^{\text{st}_2}(s) ds + \gamma \int_{k\eta}^{t} v_2^{\text{st}_1}(s) ds \\ &= v_1^{(k)} - \int_{k\eta}^{t} g^{\text{st}_2}(s) ds - \gamma^2 v_1^{(k)} \frac{(t - k\eta)^2}{2} + \gamma^2 v_3^{(k)} \frac{(t - k\eta)^2}{2}. \end{aligned} \tag{69}$$

This implies

$$\begin{aligned} \theta^{\text{st}_2}(t) &= \theta^{(k)} - \int_{k\eta}^{t} v_1^{\text{st}_2}(s) ds \\ &= \theta^{(k)} - \int_{k\eta}^{t} \int_{k\eta}^{s_2} g^{\text{st}_2}(s_1) ds_1 ds_2 - \gamma^3 v_1^{(k)} \frac{(t - k\eta)^3}{6} + \gamma^2 v_3^{(k)} \frac{(t - k\eta)^3}{6}, \end{aligned}$$

noting that $\int_{k\eta}^{t}\int_{k\eta}^{s_2} g^{\mathrm{st}_2}(s_1)ds_1ds_2 = \int_{k\eta}^{t}(t-s_1)g^{\mathrm{st}_2}(s_1)ds_1$. Per the previous calculations (69) and (66) in the case $n=3$, we can further write

$$\begin{aligned}
v_2^{\mathrm{st}_2}(t) &= v_2^{(k)} - \gamma\int_{k\eta}^{t} v_1^{\mathrm{st}_2}(s)ds + \gamma\int_{k\eta}^{t} v_3^{\mathrm{st}_1}(s)ds \\
&= v_2^{(k)} + \Bigg(-\gamma\theta^{(k)} + \gamma\int_{k\eta}^{t}(t-s_1)g^{\mathrm{st}_2}(s_1)ds_1 + \gamma^4 v_1^{(k)}\frac{(t-k\eta)^4}{4!} - \gamma^3 v_3^{(k)}\frac{(t-k\eta)^4}{4!}\Bigg) \\
&\qquad + \Bigg(\gamma\sum_{i=1}^{P-1}\frac{(t-k\eta)^{a_{3,i}+1}}{a_{3,i}+1}v_i^{(k)}b_{3,i}(\gamma)\Bigg),
\end{aligned}$$

which is of the form described in the statement of the lemma. Proceed similarly for increasing $n, 3\le n\le P-3$ to get

$$\begin{aligned}
v_n^{\mathrm{st}_2}(t) &= v_n^{(k)} - \gamma\int_{k\eta}^{t} v_{n-1}^{\mathrm{st}_2}(s)ds + \gamma\int_{k\eta}^{t} v_{n+1}^{\mathrm{st}_1}(s)ds \\
&= \sum_{1\le\ell\le P-1} v_\ell^{(k)}\mu_{i,\ell}^{\mathrm{st}_2}(t) + \theta^{(k)}\mu_{i,P}^{\mathrm{st}_2}(t) + e_n\int_{k\eta}^{t}\int_{k\eta}^{s_n}\cdots\int_{k\eta}^{s_2} g^{\mathrm{st}_2}(s_1)ds_1\ldots ds_{n-1}ds_n, \quad (70)
\end{aligned}$$

where e_n are rational functions in variables k,η,γ. Moreover, the last term can be simplified as

$$\begin{aligned}
&e_n\int_{k\eta}^{t}\int_{k\eta}^{s_n}\cdots\int_{k\eta}^{s_2} g^{\mathrm{st}_2}(s_1)ds_1\ldots ds_{n-1}ds_n \\
&= e_n\int_{k\eta}^{t}\int_{s_1}^{t}\int_{s_1}^{s_n}\cdots\int_{s_1}^{s_3} g^{\mathrm{st}_2}(s_1)ds_2\ldots ds_n ds_1 = \int_{k\eta}^{t} p(s_1)g^{\mathrm{st}_2}(s_1)ds_1,
\end{aligned}$$

where p is a polynomial in s_1.

Next, we have

$$v_{P-2}^{\mathrm{st}_2}(t) = v_{P-2}^{(k)} - \gamma\int_{k\eta}^{t} v_{P-3}^{\mathrm{st}_2}(s)ds + \gamma\int_{k\eta}^{t} v_{P-1}^{\mathrm{st}_1}(s)ds.$$

We will only expand the term $\int_{k\eta}^{t} v_{P-3}^{\mathrm{st}_2}(s)ds$ using (70) when $n=P-3$. The term $\int_{k\eta}^{t} v_{P-1}^{\mathrm{st}_1}(s)ds$ can be handled in similar fashion using (68):

$$\begin{aligned}
\int_{k\eta}^{t} v_{P-3}^{\mathrm{st}_2}(s)ds &= v_{P-3}^{(k)}(t-k\eta) + \sum_{1\le\ell\le P-1} v_\ell^{(k)}\int_{k\eta}^{t}\mu_{i,\ell}^{\mathrm{st}_2}(s)ds + \theta^{(k)}\int_{k\eta}^{t}\mu_{i,P}^{\mathrm{st}_2}(s)ds \\
&\qquad + e_n\int_{s_2=k\eta}^{t}\int_{s_1=k\eta}^{s_2} p(s_1)g^{\mathrm{st}_2}(s_1)ds_1ds_2,
\end{aligned}$$

where we can further compute that

$$\begin{aligned}
\int_{s_2=k\eta}^{t}\int_{s_1=k\eta}^{s_2} p(s_1)g^{\mathrm{st}_2}(s_1)ds_1ds_2 &= e_n\int k\eta^t\int_{s_1}^{t} p(s_1)g^{\mathrm{st}_2}(s_1)ds_2ds_1 \\
&= e_n\int_{k\eta}^{t}(t-s_1)p(s_1)g^{\mathrm{st}_2}(s_1)ds_1.
\end{aligned}$$

Hence, we arrive at

$$
\begin{aligned}
&v^{\mathrm{st}_2}_{P-2}(t)\\
&= \sum_{1\le \ell\le P-1} v^{(k)}_\ell \mu^{\mathrm{st}_2}_{P-2,\ell}(t) + \theta^{(k)}\mu^{\mathrm{st}_2}_{P-2,P}(t) + \int_{k\eta}^t h^{\mathrm{st}_2}_{P-2}(s,t)dB_s + \int_{k\eta}^t \kappa^{\mathrm{st}_2}_{P-2}(s,t) g^{\mathrm{st}_2}(s)ds
\end{aligned}
$$

as described in (60).

Finally, we have

$$
v^{\mathrm{st}_2}_{P-1}(t) = e^{-\gamma(t-k\eta)} v^{(k)}_{P-1} - \gamma\int_{k\eta}^t e^{-\gamma(t-s)} v^{\mathrm{st}_2}_{P-2}(s)ds + \sqrt{2\gamma}\int_{k\eta}^t e^{-\gamma(t-s)}dB_s.
$$

The second term on the right hand side can be expanded by plugging in the formula for $v^{\mathrm{st}_2}_{P-2}(s)$, then applying (67) and the fact that $\int_{k\eta}^t\int_{k\eta}^{s_3}\int_{k\eta}^{s_2} e^{-\gamma(t-s_3)}e^{-\gamma(s_2-s_1)}dB_{s_1}ds_2ds_3 = \int_{k\eta}^t\int_{s_1}^t\int_{s_1}^{s_3} e^{-\gamma(t-s_3)}e^{-\gamma(s_2-s_1)}ds_2ds_3dB_{s_1} = \int_{k\eta}^t\Big(\frac{1}{\gamma^2} - \frac{e^{-\gamma(t-s_1)}}{\gamma^2} + \frac{s_1 e^{-\gamma(t-s_1)}}{\gamma} - \frac{te^{-\gamma(t-s_1)}}{\gamma}\Big)dB_{s_1}$, and also that $\int_{k\eta}^t\int_{k\eta}^{s_2} e^{-\gamma(t-s_2)}p(s_1)g^{\mathrm{st}_2}(s_1)ds_1ds_2 = \int_{k\eta}^t\int_{s_1}^t e^{-\gamma(t-s_2)}p(s_1)g^{\mathrm{st}_2}(s_1)ds_2ds_1 = \int_{k\eta}^t(1/\gamma)e^{-\gamma(t-s_1)}p(s_1)g^{\mathrm{st}_2}(s_1)ds_1$. Consequently, $v^{\mathrm{st}_2}_{P-1}(t)$ can be written as (60).

Finally, we note that among $\theta^{\mathrm{st}_2}_1(t)$ and $v^{\mathrm{st}_2}_n(t), 1\le n\le P-1$, Itô integrals only appear in the formulas of $v^{\mathrm{st}_2}_{P-1}(t)$ and $v^{\mathrm{st}_2}_{P-2}(t)$, which explains the indicator functions in (60) and (64) when $j=2$. This completes the proof for Stage $j=2$.

Step 3: Induction argument.

As the induction hypothesis, we assume the statement of the lemma holds for Stage j and verify Stage $j+1$. The proof is similar to **Step 2** above (proceeding from Stage 1 to Stage 2) and is therefore omitted. □

Lemma C.2. $\mathbb{E}\big[x^{(k+1)}|x^{(k)}\big]$ *follows a multivariate normal distribution in* $\mathbb{R}^{Pd}$ *whose mean vector and covariance matrix can be determined from Lemma C.1.*

Proof. Lemma C.1 provides us with the formulas for the components $x^{(k+1)} = x^{\mathrm{st}_{P-1}}((k+1)\eta)$. Based on those formulas, we can see that $\mathbb{E}\big[x^{(k+1)}|x^{(k)}\big]$ follows a multivariate normal distribution in $\mathbb{R}^{Pd}$. From there, calculating the mean and covariance is straightforward and is the same as the proof of Lemma B.2. □

The following is a general version of Lemma B.3.

Lemma C.3. *Recall the unique minimizer θ^* of U and the $Pd \times Pd$ Jacobian matrix*

$$J_b(\theta^*, 0, \ldots, 0) = \begin{pmatrix} 0_d & I_d & 0_d & \cdots & \cdots & \cdots & \cdots & \cdots & 0_d \\ -\nabla U^2(\theta^*) I_d & 0_d & \gamma I_d & 0_d & \cdots & \cdots & \cdots & \cdots & 0_d \\ 0_d & -\gamma I_d & 0_d & \gamma I_d & 0_d & \cdots & \cdots & \cdots & 0_d \\ 0_d & 0_d & -\gamma I_d & 0_d & \gamma I_d & 0_d & \cdots & \cdots & 0_d \\ 0_d & 0_d & 0_d & -\gamma I_d & 0_d & \gamma I_d & 0_d & \cdots & 0_d \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0_d & \cdots & \cdots & \cdots & \cdots & \cdots & -\gamma I_d & 0_d & \gamma I_d \\ 0_d & \cdots & \cdots & \cdots & \cdots & \cdots & 0_d & -\gamma I_d & -\gamma I_d \end{pmatrix}.$$

Then it holds for Stage $j, 1 \le j \le P-1$ that

$$\begin{aligned} x^{\mathrm{st}_j}(t) - (\theta^*, 0, \ldots, 0) &= \big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) + (t - k\eta) J_b(\theta^*, 0, \ldots, 0) \\ &\cdot \big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) + R(t)\big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) + F_k(t), \end{aligned} \tag{71}$$

where $R(t)$ is a $Pd \times Pd$ matrix with $|R_{ij}|(t) \le C(t-k\eta)^2$, $1 \le i, j \le 4d$ and C is a constant that depends only on γ, P. Moreover, $F_k(t)$ is the Pd-dimensional vector $(f_P(t) \quad f_1(t) \quad f_2(t) \quad f_3(t) \quad \cdots \quad f_{P-1}(t))^\top$ where for each i, $f_i(t) := \int_{k\eta}^t h_i^{\mathrm{st}_j}(s,t) dB_s$ is the d-dimensional Itô integral defined in Lemma C.1.

Consequently, we have

$$\begin{aligned} x^{(k+1)} - (\theta^*, 0, \ldots, 0) &= \big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) + \eta J_b(\theta^*, 0, \ldots, 0)\big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) \\ &+ R((k+1)\eta)\big(x^{(k)} - (\theta^*, 0, \ldots, 0)\big) + F_k((k+1)\eta). \end{aligned} \tag{72}$$

Proof. Since $x^{(k+1)} = x^{\mathrm{st}_{P-1}}((k+1)\eta)$, it is sufficient to prove (71). Without loss of generality, we assume the unique minimizer of U is $\theta^* = 0$. The proof follows an induction argument.

Step 1: The Base Case $j = 2$.

Per the proof of Lemma C.1, we can deduce that

$$\begin{aligned} v_1^{\mathrm{st}_2}(t) &= v_1^{(k)} - \int_{k\eta}^t g^{\mathrm{st}_2}(s) ds + \gamma \int_{k\eta}^t v_2^{\mathrm{st}_1}(s) ds \\ &= v_1^{(k)} - (t-k\eta)\nabla^2 U(0)\theta^{(k)} + (t-k\eta) v_2^{(k)} + O\big((t-k\eta)^2\big) \sum_{i \ne 1,2} v_i^{(k)} + \int_{k\eta}^t h_1^{\mathrm{st}_2}(s,t) dB_s. \end{aligned}$$

Next, we have

$$\begin{aligned} \theta^{\mathrm{st}_2}(t) &= \theta^{(k)} + \int_{k\eta}^t v_1^{\mathrm{st}_1}(s) ds \\ &= \theta^{(k)} + (t-k\eta) v_1^{(k)} + O\big((t-k\eta)^2\big) \sum_{i \ne 1} v_i^{(k)} + \int_{k\eta}^t h_P^{\mathrm{st}_2}(s,t) dB_s. \end{aligned}$$

Similarly,

$$v_n^{\mathrm{st}_2}(t) = v_n^{(k)} - \gamma \int_{k\eta}^t v_{n-1}^{\mathrm{st}_2}(s) ds + \gamma \int_{k\eta}^t v_{n+1}^{\mathrm{st}_1}(s) ds$$

$$
\begin{aligned}
&= v_n^{(k)} - (t-k\eta)\gamma v_{n-1}^{(k)} + (t-k\eta)\gamma v_{n+1}^{(k)} \\
&\qquad + O\big((t-k\eta)^2\big) \sum_{i \notin \{n-1,n,n+1\}} v_i^{(k)} + \int_{k\eta}^{t} h_n^{\mathrm{st}_2}(s,t) dB_s, \quad 2 \le n \le P-2; \\
v_{P-1}^{\mathrm{st}_2}(t) &= v_{P-1}^{(k)} - (t-k\eta)\gamma v_{P-2}^{(k)} - (t-k\eta)\gamma v_{P-1}^{(k)} \\
&\qquad + O\big((t-k\eta)^2\big) \sum_{i \notin \{P-2,P-1\}} v_i^{(k)} + \int_{k\eta}^{t} h_{P-1}^{\mathrm{st}_2}(s,t) dB_s.
\end{aligned}
$$

At this point, we can conclude (71) holds for Stage $j = 2$.

Step 2: The Induction Argument.

We assume (71) is true up to Stage j and verify Stage $j+1$. The argument is similar to the one in **Step 1** and is therefore omitted. The proof is complete. □

The upcoming moment bounds are similar to the ones in Lemma B.4.

Lemma C.4. *Assume*

$$
\eta \le \eta^{**} := \frac{\rho}{2} \|M\|_{\mathrm{op}}^{-1} \left\|M^{-1}\right\|_{\mathrm{op}}^{-1} \frac{1}{1 + (1 + (P-2)2)\gamma^2 + L^2}, \tag{73}
$$

where the matrix M and γ, ρ, L are from Theorem 2.1. Then there exists a positive constant $\widetilde{C}_1$ such that for all k,

$$
\mathbb{E}\Big[\big|x^{(k+1)}\big|^{2\alpha}\Big] \le \Big(\widetilde{C}_1\Big)^{\alpha} (d + 2\alpha)^{\alpha},
$$

where $\widetilde{C}_1$ depends on P and γ, L from Condition H1 and c from Condition H2, and not on d. This further implies

$$
\begin{aligned}
\sup_{1 \le j,n \le P-1} \sup_{s \in [k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v_n^{\mathrm{st}_j}(s)\big|^2 + \big|\theta_n^{\mathrm{st}_j}(s)\big|^2\Big] + \sup_{2 \le j \le P-1} \sup_{s \in [k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|g^{\mathrm{st}_j}(s)\big|^2\Big] \\
\le \widetilde{C}_2(d+1),
\end{aligned}
$$

for a universal constant $\widetilde{C}_2 > 1$ that depends only on P, γ and L.

Proof. Similarly as how we apply Lemma B.3 to get the moment bounds in Lemma B.4 in Lemma B.4 for fourth-order LMC algorithm, we can apply Lemma C.3 to derive the moment bounds for P-th order LMC algorithm. The proof is very similar to the proof of Lemma B.4 and is therefore omitted. □

Proof of Lemma 2.24. **Step 1:** $j = 1$

We start with $v_n^{\mathrm{st}_1}(t) - v_1^{(k)} = 0, t \in (k\eta, (k+1)\eta]$. Next,

$$
v_2^{\mathrm{st}_1}(t) - v_2^{(k)} = \int_{k\eta}^{t} \Big(-\gamma v_1^{\mathrm{st}_1}(s) + \gamma v_3^{(k)}\Big) ds,
$$

which combined with the second moment bounds in Lemma C.4 lead to

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_2^{\mathrm{st}_1}(t) - v_2^{(k)}\right|^2\right] \le \eta^2\gamma^2\left(\mathbb{E}\left[\left|v_1^{(k)}\right|^2\right] + \mathbb{E}\left[\left|v_3^{(k)}\right|^2\right]\right) \le C_2^{\mathrm{st}_1} d\eta^2.$$

Proceed similarly for increasing n with $3 \le n \le P-2$ to obtain

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_n^{\mathrm{st}_1}(t) - v_n^{(k)}\right|^2\right] \le \eta^2\gamma^2\left(\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_{n-1}^{\mathrm{st}_1}(t)\right|^2\right] + \mathbb{E}\left[\left|v_{n+1}^{(k)}\right|^2\right]\right)$$
$$\le C_n^{\mathrm{st}_1} d\eta^2.$$

Next, we have $v_{P-1}^{\mathrm{st}_1}(t) = v_{P-1}^{(k)} + \int_{k\eta}^t \left(-\gamma v_{P-2}^{\mathrm{st}_1}(s) + \gamma v_{P-1}^{(k)}\right) ds + \sqrt{2\gamma}(B_t - B_{k\eta})$ which along with Lemma C.4 imply

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_{P-1}^{\mathrm{st}_1}(t) - v_{P-1}^{(k)}\right|^2\right] \le C_{P-1}^{\mathrm{st}_1} d\eta.$$

Finally, Lemma C.4 implies

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\theta^{\mathrm{st}_1}(t) - \theta^{(k)}\right|^2\right] \le \eta^2\gamma^2 \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{(k)}\right|^2\right] \le C_P^{\mathrm{st}_1} d\eta^2.$$

Step 2: $j = 2$ **and** $P = 3$

We have $v_1^{\mathrm{st}_2}(t) - v_1^{\mathrm{st}_1}(t) = v_1^{\mathrm{st}_2}(t) - v_1^{(k)} = \int_{k\eta}^t \left(-g^{\mathrm{st}_2}(s) + \gamma v_2^{\mathrm{st}_1}(s)\right) ds$ so that by Lemma C.4,

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{\mathrm{st}_2}(t) - v_1^{\mathrm{st}_1}(t)\right|^2\right]$$
$$\le \eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|g^{\mathrm{st}_2}(t)\right|^2\right] + \gamma^2\eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_2^{\mathrm{st}_1}(t)\right|^2\right] \le C_1^{\mathrm{st}_2} d\eta^2. \tag{74}$$

Moreover, $v_2^{\mathrm{st}_2}(t) - v_2^{\mathrm{st}_1}(t) = \int_{k\eta}^t \left(-\gamma\left(v_1^{\mathrm{st}_2}(s) - v_1^{\mathrm{st}_1}(s)\right) + \gamma\left(v_3^{\mathrm{st}_j}(s) - v_3^{(k)}(s)\right)\right) ds$ so that using (74) and the calculation in **Step 1**, we get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_2^{\mathrm{st}_2}(t) - v_2^{\mathrm{st}_1}(t)\right|^2\right] \le C_2^{\mathrm{st}_2} d\eta^4.$$

Finally, by (74), we get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\theta^{\mathrm{st}_2}(t) - \theta^{\mathrm{st}_1}(t)\right|^2\right] \le \eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{\mathrm{st}_2}(t) - v_1^{\mathrm{st}_1}(t)\right|^2\right] \le C_3^{\mathrm{st}_2} d\eta^4.$$

Step 3: $j = 2$ **and** $P \ge 4$

In the same way as (74), we get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{\mathrm{st}_2}(t) - v_1^{\mathrm{st}_1}(t)\right|^2\right] \le C_1^{\mathrm{st}_2} d\eta^2. \tag{75}$$

Next, (75) and the calculation in **Step 1** imply that

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_2^{\mathrm{st}_2}(t) - v_2^{\mathrm{st}_1}(t)\right|^2\right]$$

$$\leq \gamma^2\eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \left(\mathbb{E}\left[\left|v_1^{\mathrm{st}_2}(t) - v_1^{\mathrm{st}_1}(t)\right|^2\right] + \mathbb{E}\left[\left|v_3^{\mathrm{st}_2}(t) - v_3^{(k)}\right|^2\right]\right) \leq C_2^{\mathrm{st}_2} d\eta^4.$$

Proceed similarly for increasing $n, 3 \leq n \leq P-3$ to obtain

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_n^{\mathrm{st}_2}(t) - v_n^{\mathrm{st}_1}(t)\right|^2\right]$$
$$\leq \eta^2\gamma^2 \sup_{t\in(k\eta,(k+1)\eta]} \left(\mathbb{E}\left[\left|v_{n-1}^{\mathrm{st}_2}(t) - v_{n-1}^{\mathrm{st}_1}(t)\right|^2\right] + \mathbb{E}\left[\left|v_{n+1}^{\mathrm{st}_1}(t) - v_{n+1}^{(k)}\right|^2\right]\right) \leq C_n^{\mathrm{st}_2} d\eta^4. \quad (76)$$

Furthermore, (76) and the calculation in **Step 1** lead to

$$\begin{aligned}
&\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_{P-2}^{\mathrm{st}_2}(t) - v_{P-2}^{\mathrm{st}_1}(t)\right|^2\right] \\
&\leq \eta^2\gamma^2 \sup_{t\in(k\eta,(k+1)\eta]} \left(\mathbb{E}\left[\left|v_{P-3}^{\mathrm{st}_2}(t) - v_{P-3}^{\mathrm{st}_1}(t)\right|^2\right] + \mathbb{E}\left[\left|v_{P-1}^{\mathrm{st}_1}(t) - v_{P-1}^{(k)}\right|^2\right]\right) \\
&\leq \eta^2\gamma^2 \sup_{t\in(k\eta,(k+1)\eta]} \left(C_{P-3}^{\mathrm{st}_2} d\eta^2 + C_{P-1}^{\mathrm{st}_2} d\eta\right) \leq C_{P-2}^{\mathrm{st}_2} d\eta^3. \quad (77)
\end{aligned}$$

We also have from (77) that

$$\begin{aligned}
&\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_{P-1}^{\mathrm{st}_2}(t) - v_{P-1}^{\mathrm{st}_1}(t)\right|^2\right] \\
&\leq \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|-\gamma\int_{k\eta}^t e^{-\gamma s}\left(v_{P-2}^{\mathrm{st}_2}(s) - v_{P-2}^{\mathrm{st}_1}(s)\right)ds\right|^2\right] \leq C_{P-1}^{\mathrm{st}_2} d\eta^5. \quad (78)
\end{aligned}$$

Finally, (75) implies

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\theta^{\mathrm{st}_2}(t) - \theta^{\mathrm{st}_1}(t)\right|^2\right] \leq \eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{\mathrm{st}_2}(t) - v_1^{\mathrm{st}_1}(t)\right|^2\right] \leq C_P^{\mathrm{st}_j} d\eta^4.$$

This completes the proof. □

Proof of Lemma 2.25. We will prove the formulas in Parts $a), b), c)$ and $d)$ for Stage $j \geq 3$ via induction. We will assume $P \geq 4$, since there is no Stage 3 when $P = 3$.

First half of the proof: checking the base case $j = 3$

The first half of the proof will consist of four steps.

Step 1: Verifying Part $a)$ for Stage $j = 3$.

We have

$$\begin{aligned}
&v_1^{\mathrm{st}_3}(t) - v_1^{\mathrm{st}_2}(t) \\
&= \int_{k\eta}^t \left(-\left(g^{\mathrm{st}_3}(s) - g^{\mathrm{st}_2}(s)\right) + \gamma\left(v_2^{\mathrm{st}_j}(s) - v_2^{\mathrm{st}_1}(s)\right)\right) ds \\
&= \int_{k\eta}^t \Big(- \left(g^{\mathrm{st}_3}(s) - \nabla U\left(\theta^{\mathrm{st}_2}(s)\right)\right) - \left(\nabla U\left(\theta^{\mathrm{st}_2}(s)\right) - \nabla U\left(\theta^{\mathrm{st}_1}(s)\right)\right) \\
&\qquad - \left(\nabla U\left(\theta^{\mathrm{st}_1}(s)\right) - g^{\mathrm{st}_2}(s)\right)\Big) ds + \gamma \int_{k\eta}^t \left(v_2^{\mathrm{st}_2}(s) - v_2^{\mathrm{st}_1}(s)\right) ds.
\end{aligned}$$

so that by L-smoothness of U,

$$\begin{aligned}
&\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v_1^{\mathrm{st}_3}(t) - v^{\mathrm{st}_2}(t)\big|^2\Big] \\
&\le \eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \Bigg(\mathbb{E}\Big[\big|g^{\mathrm{st}_3}(t) - \nabla U\big(\theta^{\mathrm{st}_2}(t)\big)\big|^2\Big] + \mathbb{E}\Big[\big|\nabla U\big(\theta^{\mathrm{st}_1}(t)\big) - g^{\mathrm{st}_2}(t)\big|^2\Big] \\
&\qquad + L^2\mathbb{E}\Big[\big|\theta^{\mathrm{st}_2}(s) - \theta^{\mathrm{st}_1}(t)\big|^2\Big] + \gamma^2\mathbb{E}\Big[\big|v_2^{\mathrm{st}_2}(t) - v_2^{\mathrm{st}_1}(t)\big|^2\Big]\Bigg). \qquad (79)
\end{aligned}$$

Note the third and last terms on the right hand side in (79) are bounded in Lemma C.4 as

$$\mathbb{E}\Big[\big|\theta^{\mathrm{st}_2}(s) - \theta^{\mathrm{st}_1}(t)\big|^2\Big] \le C_P^{\mathrm{st}_2} d\eta^4 \qquad \text{and} \qquad \mathbb{E}\Big[\big|v_2^{\mathrm{st}_2}(t) - v_2^{\mathrm{st}_1}(t)\big|^2\Big] \le C_2^{\mathrm{st}_2} d\eta^4.$$

Regarding the first two terms on the right hand side in (79), similar to the argument at (59), Condition H2 indicates there exists a positive constant c such that

$$\sup_{t\in(k\eta,(k+1)\eta]} \Big(\mathbb{E}\Big[\big|g^{\mathrm{st}_3}(t) - \nabla U\big(\theta^{\mathrm{st}_2}(t)\big)\big|^2\Big] + \mathbb{E}\Big[\big|\nabla U\big(\theta^{\mathrm{st}_1}(t)\big) - g^{\mathrm{st}_2}(t)\big|^2\Big]\Big) \le cd\eta^{2P-1}.$$

Thus, we get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v_1^{\mathrm{st}_3}(t) - v^{\mathrm{st}_2}(t)\big|^2\Big] \le C_1^{\mathrm{st}_{j+1}} d\eta^6. \qquad (80)$$

Next, we have $v_2^{\mathrm{st}_3}(t) - v_2^{\mathrm{st}_2}(t) = \int_{k\eta}^{t} \big(-\gamma\big(v_1^{\mathrm{st}_3}(s) - v_1^{\mathrm{st}_2}(s)\big) + \gamma\big(v_3^{\mathrm{st}_2}(s) - v_3^{\mathrm{st}_1}(s)\big)\big)\, ds$, so that

$$\begin{aligned}
&\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v_2^{\mathrm{st}_3}(t) - v_2^{\mathrm{st}_2}(t)\big|^2\Big] \\
&\le \gamma^2\eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \Big(\mathbb{E}\Big[\big|v_1^{\mathrm{st}_3}(t) - v_1^{\mathrm{st}_2}(t)\big|^2\Big] + \mathbb{E}\Big[\big|v_3^{\mathrm{st}_2}(t) - v_3^{\mathrm{st}_1}(t)\big|^2\Big]\Big). \qquad (81)
\end{aligned}$$

The first term on the right hand side in (81) is bounded at (80), while the second term in (81) is bounded in Lemma 2.24 as $\mathbb{E}\Big[\big|v_3^{\mathrm{st}_2}(t) - v_3^{\mathrm{st}_1}(t)\big|^2\Big] \le C_3^{\mathrm{st}_2} d\eta^4$. Then

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v_2^{\mathrm{st}_3}(t) - v_2^{\mathrm{st}_2}(t)\big|^2\Big] \le C_2^{\mathrm{st}_3} d\eta^6. \qquad (82)$$

Now proceed similarly for increasing n with $3 \le n \le P-4$ and we get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v_n^{\mathrm{st}_3}(t) - v_2^{\mathrm{st}_2}(t)\big|^2\Big] \le C_n^{\mathrm{st}_3} d\eta^6, \quad 3 \le n \le P-4. \qquad (83)$$

Step 2: Verifying Part b) for Stage $j = 3$.

We use (83) in the case $n = P-4$ and Lemma 2.24 to get

$$\begin{aligned}
&\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v_{P-3}^{\mathrm{st}_3}(t) - v_{P-3}^{\mathrm{st}_2}(t)\big|^2\Big] \\
&\le \gamma^2\eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \Big(\mathbb{E}\Big[\big|v_{P-4}^{\mathrm{st}_3}(t) - v_{P-4}^{\mathrm{st}_2}(t)\big|^2\Big] + \mathbb{E}\Big[\big|v_{P-2}^{\mathrm{st}_2}(t) - v_{P-2}^{\mathrm{st}_1}(t)\big|^2\Big]\Big)
\end{aligned}$$

$$\leq \gamma^2\eta^2\big(C^{\mathrm{st}_3}_{P-4}d\eta^6 + C^{\mathrm{st}_2}_{P-2}\eta^3\big) \leq C^{\mathrm{st}_3}_{P-3}d\eta^5. \tag{84}$$

Step 3: Verifying Part c) for Stage $j = 3$.

Using (84) and Lemma 2.24, we can write

$$\begin{aligned}
&\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v^{\mathrm{st}_3}_{P-2}(t) - v^{\mathrm{st}_2}_{P-2}(t)\big|^2\Big] \\
&\leq \gamma^2\eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \bigg(\mathbb{E}\Big[\big|v^{\mathrm{st}_3}_{P-3}(t) - v^{\mathrm{st}_2}_{P-3}(t)\big|^2\Big] + \mathbb{E}\Big[\Big|v^{\mathrm{st}_j}_{P-1}(t) - v^{\mathrm{st}_1}_{P-1}(t)\Big|^2\Big]\bigg) \\
&\leq \gamma^2\eta^2\big(C^{\mathrm{st}_3}_{P-4}d\eta^5 + C^{\mathrm{st}_2}_{P-2}\eta^{8+2(P-1)-2P-1}\big) \leq C^{\mathrm{st}_3}_{P-2}d\eta^7.
\end{aligned}$$

Now proceed similarly for increasing $n, P-j \leq n \leq P-2$ and we get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v^{\mathrm{st}_3}_{P-2}(t) - v^{\mathrm{st}_2}_{P-2}(t)\big|^2\Big] \leq C^{\mathrm{st}_3}_n d\eta^{4+3+2(n-(P-2))} = C^{\mathrm{st}_3}_n d\eta^{11+2n-2P}. \tag{85}$$

Finally, the bound in (85) in the case $n = P-2$ implies

$$\begin{aligned}
&\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v^{\mathrm{st}_3}_{P-1}(t) - v^{\mathrm{st}_2}_{P-1}(t)\big|^2\Big] \\
&\leq \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|-\gamma\int_{k\eta}^t e^{-\gamma s}\big(v^{\mathrm{st}_3}_{P-2}(s) - v^{\mathrm{st}_2}_{P-2}(s)\big)ds\right|^2\right] \leq C^{\mathrm{st}_3}_{P-1}d\eta^9. \qquad (86)
\end{aligned}$$

Step 4: Verifying Part d) for Stage $j = 3$.

By (80), we have

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|\theta^{\mathrm{st}_3}(t) - \theta^{\mathrm{st}_2}(t)\big|^2\Big] \leq \eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\Big[\big|v^{\mathrm{st}_3}_1(t) - v^{\mathrm{st}_2}_1(t)\big|^2\Big] \leq C^{\mathrm{st}_j}_P d\eta^8.$$

Second half of the proof: the induction argument

The second half will also consist of four steps. As the induction hypothesis, we assume Part $a), b), c)$ and $d)$ of the current Proposition are true up to Stage j.

Step 1: Verifying Part a) for Stage $j+1$.

We have

$$\begin{aligned}
&v^{\mathrm{st}_{j+1}}_1(t) - v^{\mathrm{st}_j}_1(t) \\
&= \int_{k\eta}^t \Big(-\big(g^{\mathrm{st}_{j+1}}(s) - g^{\mathrm{st}_j}(s)\big) + \gamma\Big(v^{\mathrm{st}_j}_2(s) - v^{\mathrm{st}_{j-1}}_2(s)\Big)\Big)\, ds \\
&= \int_{k\eta}^t \bigg(- \Big(g^{\mathrm{st}_{j+1}}(s) - \nabla U\big(\theta^{\mathrm{st}_j}(s)\big)\Big) - \Big(\nabla U\big(\theta^{\mathrm{st}_j}(t)\big) - \nabla U\big(\theta^{\mathrm{st}_{j-1}}(s)\big)\Big) \\
&\qquad - \Big(\nabla U\big(\theta^{\mathrm{st}_{j-1}}(s)\big) - g^{\mathrm{st}_j}(s)\Big)ds + \gamma\int_{k\eta}^t v^{\mathrm{st}_j}_2(s) - v^{\mathrm{st}_{j-1}}_2(s)\bigg)ds,
\end{aligned}$$

so that by L-smoothness of U, we obtain:

$$\begin{aligned} &\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{\mathrm{st}_{j+1}}(t) - v^{\mathrm{st}_j}(t)\right|^2\right] \\ &\leq \eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \Bigg(\mathbb{E}\left[\left|g^{\mathrm{st}_{j+1}}(t) - \nabla U\big(\theta^{\mathrm{st}_j}(t)\big)\right|^2\right] + \mathbb{E}\left[\left|\nabla U\big(\theta^{\mathrm{st}_j}(t)\big) - g^{\mathrm{st}_{j-1}}(t)\right|^2\right] \\ &\qquad + L^2\mathbb{E}\left[\left|\theta^{\mathrm{st}_j}(s) - \theta^{\mathrm{st}_{j-1}}(t)\right|^2\right] + \gamma^2\mathbb{E}\left[\left|v_2^{\mathrm{st}_j}(t) - v_2^{\mathrm{st}_{j-1}}(t)\right|^2\right]\Bigg). \end{aligned} \tag{87}$$

The third and last terms on the right hand side in (87) are bounded respectively by Part d) and Part a) of the induction hypothesis:

$$\mathbb{E}\left[\left|\theta^{\mathrm{st}_j}(s) - \theta^{\mathrm{st}_{j-1}}(t)\right|^2\right] \leq C_P^{\mathrm{st}_j} d\eta^{2j+2} \qquad \text{and} \qquad \mathbb{E}\left[\left|v_2^{\mathrm{st}_j}(t) - v_2^{\mathrm{st}_{j-1}}(t)\right|^2\right] \leq C_2^{\mathrm{st}_j} d\eta^{2j}.$$

Regarding the first two terms on the right hand side in (87), similar to the argument at (59), Condition H2 indicates there is a positive constant c such that

$$\sup_{t\in(k\eta,(k+1)\eta]} \left(\mathbb{E}\left[\left|g^{\mathrm{st}_{j+1}}(t) - \nabla U\big(\theta^{\mathrm{st}_j}(t)\big)\right|^2\right] + \mathbb{E}\left[\left|\nabla U\big(\theta^{\mathrm{st}_j}(t)\big) - g^{\mathrm{st}_{j-1}}(t)\right|^2\right]\right) \leq cd\eta^{2P-1}.$$

Since for $1 \leq j \leq P-1$, we have $2j \leq 2P-1$, the above calculations lead to

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_1^{\mathrm{st}_{j+1}}(t) - v^{\mathrm{st}_j}(t)\right|^2\right] \leq C_1^{\mathrm{st}_{j+1}} d\eta^{2j+2}. \tag{88}$$

Next, we have $v_2^{\mathrm{st}_{j+1}}(t) - v_2^{\mathrm{st}_j}(t) = \int_{k\eta}^t \Big(-\gamma\Big(v_1^{\mathrm{st}_{j+1}}(s) - v_1^{\mathrm{st}_j}(s)\Big) + \gamma\Big(v_3^{\mathrm{st}_j}(s) - v_3^{\mathrm{st}_{j-1}}(s)\Big)\Big)\, ds$, so that

$$\begin{aligned} &\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_2^{\mathrm{st}_{j+1}}(t) - v_2^{\mathrm{st}_j}(t)\right|^2\right] \\ &\leq \gamma^2\eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \left(\mathbb{E}\left[\left|v_1^{\mathrm{st}_{j+1}}(t) - v_1^{\mathrm{st}_j}(t)\right|^2\right] + \mathbb{E}\left[\left|v_3^{\mathrm{st}_j}(t) - v_3^{\mathrm{st}_{j-1}}(t)\right|^2\right]\right). \end{aligned}$$

The first term on the right hand side is bounded at (88), while the second term is bounded per Part a) of the induction hypothesis as $\mathbb{E}\left[\left|v_3^{\mathrm{st}_j}(t) - v_3^{\mathrm{st}_{j-1}}(t)\right|^2\right] \leq C_3^{\mathrm{st}_j} d\eta^{2j}$. Then

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_2^{\mathrm{st}_{j+1}}(t) - v_2^{\mathrm{st}_j}(t)\right|^2\right] \leq C_2^{\mathrm{st}_{j+1}} d\eta^{2j+2}. \tag{89}$$

Now proceed similarly for increasing n with $3 \leq n \leq P-j-2$ and we get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v_n^{\mathrm{st}_{j+1}}(t) - v_2^{\mathrm{st}_j}(t)\right|^2\right] \leq C_n^{\mathrm{st}_{j+1}} d\eta^{2j+2}, \quad 3 \leq n \leq P-j-2. \tag{90}$$

Via (88), (89) and (90), we confirm via induction that Part a) of this Proposition is true.

Step 2: Verifying Part b) for Stage $j+1$.

We use (90) in the case $n = P - j - 2$ and Part $b)$ of the induction hypothesis to get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v^{\mathrm{st}_{j+1}}_{P-j-1}(t) - v^{\mathrm{st}_j}_{P-j-1}(t)\right|^2\right]$$
$$\leq \gamma^2\eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \left(\mathbb{E}\left[\left|v^{\mathrm{st}_{j+1}}_{P-j-2}(t) - v^{\mathrm{st}_j}_{P-j-2}(t)\right|^2\right] + \mathbb{E}\left[\left|v^{\mathrm{st}_j}_{P-j}(t) - v^{\mathrm{st}_{j-1}}_{P-j}(t)\right|^2\right]\right)$$
$$\leq \gamma^2\eta^2\left(C^{\mathrm{st}_{j+1}}_{P-j-2} d\eta^{2j+2} + C^{\mathrm{st}_j}_{P-j}\eta^{2j-1}\right) \leq C^{\mathrm{st}_{j+1}}_{P-j-1} d\eta^{2j+1}. \tag{91}$$

Thus, Part $b)$ of this Proposition is true.

Step 3: Verifying Part $c)$ for Stage $j + 1$.

Using (91) and Part $c)$ of the induction hypothesis, we can write

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v^{\mathrm{st}_{j+1}}_{P-j}(t) - v^{\mathrm{st}_j}_{P-j}(t)\right|^2\right]$$
$$\leq \gamma^2\eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \left(\mathbb{E}\left[\left|v^{\mathrm{st}_{j+1}}_{P-j-1}(t) - v^{\mathrm{st}_j}_{P-j-1}(t)\right|^2\right] + \mathbb{E}\left[\left|v^{\mathrm{st}_j}_{P-j+1}(t) - v^{\mathrm{st}_{j-1}}_{P-j+1}(t)\right|^2\right]\right)$$
$$\leq \gamma^2\eta^2\left(C^{\mathrm{st}_{j+1}}_{P-j-2} d\eta^{2j+1} + C^{\mathrm{st}_j}_{P-j}\eta^{4j+2(P-j+1)-2P-1}\right) \leq C^{\mathrm{st}_{j+1}}_{P-j} d\eta^{2j+3}.$$

Now proceed similarly for increasing $n, P - j \leq n \leq P - 2$ and we get

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v^{\mathrm{st}_{j+1}}_{P-j}(t) - v^{\mathrm{st}_j}_{P-j}(t)\right|^2\right] \leq C^{\mathrm{st}_{j+1}}_{n} d\eta^{2j+3+2(n-(P-j))} = C^{\mathrm{st}_{j+1}}_{n} d\eta^{4(j+1)+2n-2P-1}. \tag{92}$$

Finally, the bound in (92) in the case $n = P - 2$ implies

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v^{\mathrm{st}_{j+1}}_{P-1}(t) - v^{\mathrm{st}_j}_{P-1}(t)\right|^2\right]$$
$$\leq \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|-\gamma\int_{k\eta}^{t} e^{-\gamma s}\left(v^{\mathrm{st}_{j+1}}_{P-2}(s) - v^{\mathrm{st}_j}_{P-2}(s)\right)\right|^2\right] \leq C^{\mathrm{st}_{j+1}}_{P-1} d\eta^{4j+1}. \tag{93}$$

By (92) and (93), we conclude via induction that Part $c)$ of this Proposition is true.

Step 4: Verifying Part $d)$ for Stage $j + 1$.

By (88), we have

$$\sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|\theta^{\mathrm{st}_{j+1}}(t) - \theta^{\mathrm{st}_j}(t)\right|^2\right] \leq \eta^2 \sup_{t\in(k\eta,(k+1)\eta]} \mathbb{E}\left[\left|v^{\mathrm{st}_{j+1}}_1(t) - v^{\mathrm{st}_j}_1(t)\right|^2\right] \leq C^{\mathrm{st}_j}_P d\eta^{2j+4},$$

so that Part $d)$ of this Proposition is true. This also completes our induction argument.

The estimate for Stage $P - 1$ of the Proposition is straightforward given the previous results and the results in Lemma 2.24. This completes the proof. □

Proof of Theorem 2.19. □

Appendix D. Choice of Polynomial Approximation

In this appendix, we expand on Remark 2.14 regarding the difficulty in applying Lagrange polynomial interpolation to our MCMC algorithm based on fourth-order Langevin dynamics.

Recall from [MMW$^+$21, Section 3.3] and also from [SBB$^+$80], the Chebyshev nodes on the interval $[k\eta,(k+1)\eta]$ are $s_i = k\eta + \frac{\eta}{2}\big(1+\cos\big(\frac{2i-1}{2\alpha}\pi\big)\big), i=1,2,\dots,\alpha$. Then the $(\alpha-1)$-degree Lagrange polynomial associated with a $\mathbb{R}^d$-valued path $t\in[k\eta,(k+1)\eta]\mapsto z(t)$ is $\phi_z(t) := \sum_{i=1}^{\alpha} z(s_i)\prod_{j\neq i}\frac{t-s_i}{s_j-s_i}$. The error estimate when z has up to α-th order derivatives is ([SBB$^+$80, Section 3.1])

$$\sup_{t\in[k\eta,(k+1)\eta]}|z(t)-\phi_z(t)| \leq \frac{\eta^\alpha}{2^{\alpha-1}\alpha!}\sup_{t\in[k\eta,(k+1)\eta]}\left|\frac{d^\alpha}{dt^\alpha}z(t)\right|. \tag{94}$$

Coming back to our MCMC algorithm based on fourth-order Langevin dynamics, we need to approximate the path

$$p_1(t): t\mapsto \nabla U\Big(\theta^{(k)}+(t-k\eta)v_1^{(k)}\Big), \tag{95}$$

and also the path

$$p_2(t): s\mapsto \nabla U\Big(\tilde{\theta}(t)\Big), \tag{96}$$

where

$$\begin{aligned}\tilde{\theta}(t) = \theta^{(k)} + v_1^{(k)}(t-k\eta) - \int_{k\eta}^{t}\int_{k\eta}^{s}\nabla U\Big(\theta^{(k)}+(r-k\eta)v_1^{(k)}\Big)drds \\ + \gamma v_2^{(k)}\frac{(t-k\eta)^2}{2!} + \gamma^2\Big(v_3^{(k)}-v_1^{(k)}\Big)\frac{(t-k\eta)^3}{3!}.\end{aligned}$$

Lagrange polynomial interpolation of the path p_1 in (95) has been done in [MMW$^+$21] by defining $g_1(t) := \sum_{i=1}^{\alpha}\nabla U\Big(\theta^{(k)}+(s_i-k\eta)v_1^{(k)}\Big)\prod_{j\neq i}\frac{t-s_i}{s_j-s_i}$. Note that $g_1(t)$ is a polynomial of degree $\alpha-1$ in t, and the error $\sup_{t\in[k\eta,(k+1)\eta]}|p_1(t)-g_1(t)|$ is bounded in [MMW$^+$21, Section 4.3.2] using (94) as

$$\begin{aligned}\sup_{t\in[k\eta,(k+1)\eta]}|p_1(t)-g_1(t)| &\leq \frac{\eta^\alpha}{2^{\alpha-1}\alpha!}\sup_{t\in[k\eta,(k+1)\eta]}\left|\frac{d^\alpha}{dt^\alpha}\nabla U\Big(\theta^{(k)}+(t-k\eta)v_1^{(k)}\Big)\right| \\ &\leq \frac{\eta^\alpha}{2^{\alpha-1}\alpha!}\sup_{t\in[k\eta,(k+1)\eta]}\left|\nabla^\alpha U\Big(\theta^{(k)}+(t-k\eta)v_1^{(k)}\Big)v_1^{(k)}\right|.\end{aligned}$$

We observe that this bound is simple since $t\mapsto\theta^{(k)}+(t-k\eta)v_1^{(k)}$ is a linear function in t, so that second and higher derivatives of $t\mapsto\theta^{(k)}+(t-k\eta)v_1^{(k)}$ immediately vanish.

Meanwhile, we can approximate the path p_2 by defining

$$g_2(t) := \sum_{i=1}^{\alpha}\nabla U(T(s_i))\prod_{j\neq i}\frac{t-s_i}{s_j-s_i},$$

where

$$T(t) = \theta^{(k)} + v_1^{(k)}(t-k\eta) - \int_{k\eta}^{t}\int_{k\eta}^{s} g_1(r)drds \\ + \gamma v_2^{(k)}\frac{(t-k\eta)^2}{2!} + \gamma^2\left(v_3^{(k)} - v_1^{(k)}\right)\frac{(t-k\eta)^3}{3!}.$$

From (94), we get

$$\sup_{t\in[k\eta,(k+1)\eta]} |g_2(t) - \nabla U(T(t))| \le \frac{\eta^{\alpha-1}}{2^{\alpha-2}(\alpha-1)!}\sup_{t\in[k\eta,(k+1)\eta]}\left|\frac{d^{\alpha-1}}{dt^{\alpha-1}}\nabla U(T(t))\right|.$$

In particular, the fact that $g_1(t)$ is a polynomial of degree $\alpha-1$ suggests $T(t)$ is a polynomial of degree $\alpha+1$. We use Faà Di Bruno's formula to get

$$\frac{d^{\alpha-1}}{dt^{\alpha-1}}\nabla U(T(t)) = \sum_{M_{\alpha-1}} \frac{(\alpha-1)!}{\prod_{i=1}^{\alpha-1} m_i i^{m_i}} \nabla^{1+\sum_{i=1}^{\alpha-1} m_i} U(T(t)) \prod_{i=1}^{\alpha-1}\left(\frac{d^i}{dt^i}T(t)\right)^{m_i}, \tag{97}$$

where $M_{\alpha-1} := \{(m_1,\dots,m_{\alpha-1}) : m_i \ge 0 \text{ and } \sum_{i=1}^{\alpha-1} i m_i = \alpha-1\}$. Since $T(t)$ is not a linear function and is a polynomial of potentially high degree, most terms in (97) does not vanish, which makes the error bound quite challenging.

Appendix E. Extra Calculations for the Numerical Experiments

E.1. **Quadratic loss function.** In this section, we consider the case where the loss function $U(\theta)$ is quadratic. Consider the mean-squared error in a regression problem with Ridge regularization.If we have a training dataset $Z = \{z_1, z_2, \cdots, z_n\}$, with $z_i = (X_i, y_i), i = 1, 2, \cdots, n$. Here, $X_i \in \mathbb{R}^d$ is a d-dimensional input and $y_i \in \mathbb{R}$ is the one-dimensional output. Then the potential (loss) function is defined as

$$U(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \theta^\top X_i\right)^2 + \frac{\lambda}{2}|\theta|^2 = \frac{1}{2n}|y - X\theta|^2 + \frac{\lambda}{2}|\theta|^2. \tag{98}$$

The gradient of the loss is then given by

$$\nabla U(\theta) = -\frac{1}{n}X^\top(y - X\theta) + \lambda\theta = \frac{1}{n}X^\top X\theta - \frac{1}{n}X^\top y + \lambda\theta = \left(\frac{1}{n}X^\top X + \lambda I\right)\theta - \frac{1}{n}X^\top y. \tag{99}$$

Define $A = \left(\frac{1}{n}X^\top X + \lambda I\right)$ and $b = \frac{1}{n}X^\top y$. Then $\nabla U(\theta) = A\theta - b$ is linear in θ.

Third-order computation: Now we are ready to get an explicit form of the vector $\Delta U(\theta, v_1)$ used in equation (43) for the third-order dynamics

$$\Delta U(\theta, v_1) := \int_0^\eta \nabla U(\theta + tv_1)dt = \int_0^\eta \left[A(\theta + tv_1) - b\right]dt = A\left(\eta\theta + \frac{\eta^2}{2}v_1\right) - b\eta.$$

Fourth-order computation: We use the following version of the formulas to compute the mean vector components m_is' which are the expanded forms of the formulas presented in Lemma B.2. Note that, except for θ, v_1, v_2, and v_3, all other variables used in the computation processes are dummy variables.

$$
\begin{aligned}
m_0 = &-\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\nabla U\Bigg(\theta+(r-k\eta)v_1-\int_{k\eta}^{r}\int_{k\eta}^{w}\nabla U(\theta+(y-k\eta)v_1)dydw \\
&+\gamma v_2\frac{(r-k\eta)^2}{2!}+\gamma^2(-v_1+v_3)\frac{(r-k\eta)^3}{3!}\Bigg)drds \\
&+\gamma^2\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\nabla U(\theta+(y-k\eta)v_1)dydwdrds \\
&+\theta\mu_{00}+v_1\mu_{01}+v_2\mu_{02}+v_3\mu_{03}.
\end{aligned}
$$

Let us split the integral into small parts, we have

$$
m_0=-\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\nabla U(\theta+(r-k\eta)v_1-T_1+T_2)drds+T_3+T_4,
$$

where

$$
\begin{aligned}
T_1 &= \int_{k\eta}^{r}\int_{k\eta}^{w}\nabla U(\theta+(y-k\eta)v_1)dydw, \\
T_2 &= \gamma v_2\frac{(r-k\eta)^2}{2!}+\gamma^2(-v_1+v_3)\frac{(r-k\eta)^3}{3!}, \\
T_3 &= \gamma^2\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\nabla U(\theta+(y-k\eta)v_1)dydwdrds, \\
T_4 &= \theta\mu_{00}+v_1\mu_{01}+v_2\mu_{02}+v_3\mu_{03}.
\end{aligned}
$$

This implies

$$
\begin{aligned}
m_0 = &\ \theta+\mu_{01}v_1+\mu_{02}v_2+\mu_{03}v_3 \\
&+\left(\frac{\eta^4\gamma^2}{24}-\frac{\eta^2}{2}\right)(A\theta-b)+\frac{\eta^4}{24}(A(A\theta-b)) \\
&+\left(\frac{\eta^5\gamma^2}{60}-\frac{\eta^3}{6}\right)Av_1+\frac{\eta^5}{120}A(Av_1)-\frac{\eta^4\gamma}{24}Av_2-\frac{\eta^5\gamma^2}{120}Av_3,
\end{aligned}
$$

where we used $\mu_{00}=1$. Next,

$$
\begin{aligned}
m_1 = &-\int_{k\eta}^{(k+1)\eta}\nabla U\Bigg(\theta+v_1(s-k\eta)-\int_{k\eta}^{s}\int_{k\eta}^{r}\nabla U\left(\theta+(w-k\eta)v_1\right)dwdr+\gamma v_2\frac{(s-k\eta)^2}{2!} \\
&+\gamma^2\left(-v_1+v_3\right)\frac{(s-k\eta)^3}{3!}\Bigg)ds+\gamma^2\int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\nabla U\left(\theta+(w-k\eta)v_1\right)dwdrds \\
&+\theta\mu_{10}+v_1\mu_{11}+v_2\mu_{12}+v_3\mu_{13}.
\end{aligned}
$$

Split the integral into smaller parts, we have

$$
m_1=-\int_{k\eta}^{(k+1)\eta}\nabla U\left(\theta+(s-k\eta)v_1-T_1+T_2\right)ds+T_3+T_4,
$$

where

$$
\begin{aligned}
T_1 &= \int_{k\eta}^{s} \int_{k\eta}^{r} \nabla U \left(\theta + (w - k\eta)v_1\right) dwdr, \\
T_2 &= \gamma v_2 \frac{(s-k\eta)^2}{2!} + \gamma^2 \left(-v_1 + v_3\right) \frac{(s-k\eta)^3}{3!}, \\
T_3 &= \gamma^2 \int_{k\eta}^{(k+1)\eta} \int_{k\eta}^{s} \int_{k\eta}^{r} \nabla U \left(\theta + (w - k\eta)v_1\right) dwdrds, \\
T_4 &= \theta \mu_{10} + v_1 \mu_{11} + v_2 \mu_{12} + v_3 \mu_{13},
\end{aligned}
$$

which implies

$$
\begin{aligned}
m_1 = \mu_{11} v_1 + \mu_{12} v_2 + \mu_{13} v_3 &+ \left(\frac{\eta^3\gamma^2}{6} - \eta\right)(A\theta - b) + \frac{\eta^3}{6} A(A\theta - b) \\
&+ \left(\frac{\eta^4\gamma^2}{12} - \frac{\eta^2}{2}\right) Av_1 + \frac{\eta^4}{24} A(Av_1) - \frac{\eta^3\gamma}{6} Av_2 - \frac{\eta^4\gamma^2}{24} Av_3,
\end{aligned}
$$

where we used $\mu_{10} = 0$. Next, we compute m_2 as follows:

$$
\begin{aligned}
m_2 = \gamma \int_{k\eta}^{(k+1)\eta} &\int_{k\eta}^{s} \nabla U \Bigg(\theta + v_1(r - k\eta) - \int_{k\eta}^{r} \int_{k\eta}^{w} \nabla U \Big(\theta + (y - k\eta)v_1\Big) dydw \\
&+ \gamma v_2 \frac{(r-k\eta)^2}{2!} + \gamma^2 \Big(-v_1 + v_3\Big) \frac{(r-k\eta)^3}{3!}\Bigg) drds \\
&- \gamma^3 \int_{k\eta}^{(k+1)\eta} \int_{k\eta}^{s} \int_{k\eta}^{r} \int_{k\eta}^{w} \nabla U \Big(\theta + (y - k\eta)v_1\Big) dydwdrds \\
&- \gamma^3 \int_{k\eta}^{(k+1)\eta} \int_{k\eta}^{s} \int_{k\eta}^{r} \int_{k\eta}^{w} e^{-\gamma(s-r)} \nabla U \Big(\theta + (y - k\eta)v_1\Big) dydwdrds \\
&+ \theta \mu_{20} + v_1 \mu_{21} + v_2 \mu_{22} + v_3 \mu_{23}.
\end{aligned}
$$

Split the integral into smaller parts. Define

$$
\begin{aligned}
T_1 &= \int_{k\eta}^{r} \int_{k\eta}^{w} \nabla U \Big(\theta + (y - k\eta)v_1\Big) dydw \\
T_2 &= \gamma v_2 \frac{(r-k\eta)^2}{2!} + \gamma^2 \Big(-v_1 + v_3\Big) \frac{(r-k\eta)^3}{3!} \\
T_3 &= -\gamma^3 \int_{k\eta}^{(k+1)\eta} \int_{k\eta}^{s} \int_{k\eta}^{r} \int_{k\eta}^{w} \nabla U \Big(\theta + (y - k\eta)v_1\Big) dydwdrds \\
T_4 &= -\gamma^3 \int_{k\eta}^{(k+1)\eta} \int_{k\eta}^{s} \int_{k\eta}^{r} \int_{k\eta}^{w} e^{-\gamma(s-r)} \nabla U \Big(\theta + (y - k\eta)v_1\Big) dydwdrds \\
T_5 &= \theta \mu_{20} + v_1 \mu_{21} + v_2 \mu_{22} + v_3 \mu_{23}.
\end{aligned}
$$

Then the integral becomes

$$
m_2 = \gamma \int_{k\eta}^{(k+1)\eta} \int_{k\eta}^{s} \nabla U(\theta + (r - k\eta)v_1 - T_1 + T_2) drds + T_3 + T_4 + T_5
$$

$$
\begin{aligned}
&= \mu_{21}v_1 + \mu_{22}v_2 + \mu_{23}v_3 + \Big(\frac{1-e^{-\eta\gamma}}{\gamma} - \frac{\eta^4\gamma^3}{24} - \frac{\eta^3\gamma^2}{6} + \eta^2\gamma - \eta\Big)(A\theta - b) \\
&- \frac{\eta^4\gamma}{24}(A(A\theta-b)) + \Big(-\frac{\eta^5\gamma^3}{60} - \frac{\eta^4\gamma^2}{24} + \frac{\eta^3\gamma}{3} - \frac{\eta^2}{2} + \frac{\eta}{\gamma} - \frac{1-e^{-\eta\gamma}}{\gamma^2}\Big)(Av_1) \\
&- \frac{\eta^5\gamma}{120}(A(Av_1)) + \frac{\eta^4\gamma^3}{24}(Av_2) + \frac{\eta^5\gamma^3}{120}(Av_3),
\end{aligned}
$$

where we used $\mu_{20} = 0$. Finally,

$$
\begin{aligned}
m_3 &= -\gamma^2 \int_{k\eta}^{(k+1)\eta} \int_{k\eta}^{s}\int_{k\eta}^{r} e^{-\gamma((k+1)\eta - s)} \\
&\cdot \nabla U\Bigg(\theta + (w-k\eta)v_1 - \int_{k\eta}^{w}\int_{k\eta}^{y} \nabla U\Big(\theta + (z-k\eta)v_1\Big)dzdy + \gamma v_2 \frac{(w-k\eta)^2}{2!} \\
&+ \gamma^2\Big(-v_1 + v_3\Big)\frac{(w-k\eta)^3}{3!}\Bigg)dwdrds \\
&+ \gamma^4 \int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma((k+1)\eta-s)}\nabla U\Big(\theta + (z-k\eta)v_1\Big)dzdydwdrds \\
&+ \gamma^4 \int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma((k+1)\eta-s)}e^{-\gamma(r-w)}\nabla U\Big(\theta + (z-k\eta)v_1\Big)dzdydwdrds \\
&+ \theta\mu_{30} + v_1\mu_{31} + v_2\mu_{32} + v_3\mu_{33}.
\end{aligned}
$$

Denote

$$
\begin{aligned}
T_1 &= \int_{k\eta}^{w}\int_{k\eta}^{y}\nabla U\Big(\theta + (z-k\eta)v_1\Big)dzdy, \\
T_2 &= \gamma v_2 \frac{(w-k\eta)^2}{2!} + \gamma^2\Big(-v_1+v_3\Big)\frac{(w-k\eta)^3}{3!}, \\
T_3 &= \gamma^4 \int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma((k+1)\eta-s)}\nabla U\Big(\theta + (z-k\eta)v_1\Big)dzdydwdrds, \\
T_4 &= \gamma^4 \int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r}\int_{k\eta}^{w}\int_{k\eta}^{y} e^{-\gamma((k+1)\eta-s)}e^{-\gamma(r-w)}\nabla U\Big(\theta + (z-k\eta)v_1\Big)dzdydwdrds, \\
T_5 &= \theta\mu_{30} + v_1\mu_{31} + v_2\mu_{32} + v_3\mu_{33}.
\end{aligned}
$$

Then the integral becomes

$$
\begin{aligned}
m_3 = -\gamma^2 \int_{k\eta}^{(k+1)\eta}\int_{k\eta}^{s}\int_{k\eta}^{r} & e^{-\gamma((k+1)\eta-s)}\nabla U(\theta + (w-k\eta)v_1 - T_1 + T_2)dwdrds \\
&+ T_3 + T_4 + T_5.
\end{aligned}
$$

Re-arranging, we have

$$
\begin{aligned}
m_3 &= \mu_{31}v_1 + \mu_{32}v_2 + \mu_{33}v_3 \\
&+ \left(\frac{\eta^4\gamma^3}{24} - \eta^2\gamma + \eta\left(3 + e^{-\eta\gamma}\right) - \frac{4\left(1-e^{-\eta\gamma}\right)}{\gamma}\right)(A\theta - b)
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\frac{\eta^4\gamma}{24}-\frac{\eta^3}{6}+\frac{\eta^2}{2\gamma}-\frac{\eta}{\gamma^2}+\frac{1-e^{-\eta\gamma}}{\gamma^3}\right)A(A\theta-b)\\
&\quad+\left(\frac{\eta^5\gamma^3}{60}-\frac{\eta^4\gamma^2}{24}-\frac{\eta^3\gamma}{6}+\eta^2-\frac{4e^{-\eta\gamma}}{\gamma^2}-\frac{\eta e^{-\eta\gamma}}{\gamma}-\frac{3\eta}{\gamma}+\frac{4}{\gamma^2}\right)(Av_1)\\
&\quad\quad+\left(\frac{\eta^5\gamma}{120}-\frac{\eta^4}{24}+\frac{\eta^3}{6\gamma}-\frac{\eta^2}{2\gamma^2}+\frac{e^{-\eta\gamma}}{\gamma^4}+\frac{\eta}{\gamma^3}-\frac{1}{\gamma^4}\right)(A(Av_1))\\
&\quad\quad\quad+\left(-\frac{1}{24}\eta^4\gamma^2+\frac{\eta^3\gamma}{6}-\frac{\eta^2}{2}+\frac{e^{-\eta\gamma}}{\gamma^2}+\frac{\eta}{\gamma}-\frac{1}{\gamma^2}\right)(Av_2)\\
&\quad\quad\quad\quad+\Big(-\frac{1}{120}\eta^5\gamma^3+\frac{\eta^4\gamma^2}{24}-\frac{\eta^3\gamma}{6}+\frac{\eta^2}{2}-\frac{e^{-\eta\gamma}}{\gamma^2}-\frac{\eta}{\gamma}+\frac{1}{\gamma^2}\Big)(Av_3),
\end{aligned}
$$

where we used $\mu_{30}=0$.

E.2. **Logistic loss function.** Similar to the quadratic case, let us assume that we have an input data set $X\in\mathbb{R}^{n\times d}$, an output dataset $\mathbf{y}\in\{0,1\}^n$, and $\theta\in\mathbb{R}^d$ being the model parameters or weights. Then the predicted probability for the i-th sample $y_i=1$ is

$$\sigma(z_i)=\mathbb{P}\left(y_i=1|X_i;\theta\right)=\frac{1}{1+e^{-z_i}}=\hat{y}_i; \tag{100}$$

where $z_i=X_i^\top\theta\in\mathbb{R}$ and $\sigma(z)$ is a real-valued function. However, if $z=X\theta\in\mathbb{R}^n$ then we define the vector-valued sigmoid function as

$$\vec{\sigma}(z):=\frac{1}{1+e^{-z}}:=\left(\frac{1}{1+e^{-z_1}},\ldots,\frac{1}{1+e^{-z_n}}\right). \tag{101}$$

Therefore, for a two-class classification problem, we define

$$
\begin{aligned}
&\mathbb{P}(y_i=1|X_i;\theta)=\hat{y}_i=\sigma(x_i^\top\theta),\\
&\mathbb{P}(y_i=0|X_i;\theta)=1-\hat{y}_i=1-\sigma(X_i^\top\theta).
\end{aligned} \tag{102}
$$

Then for a given y_i, the probability of taking one of the classes, we combine the above equations (102) into a single equation

$$\mathbb{P}(y_i|X_i;\theta)=\hat{y}_i^{\,y_i}(1-\hat{y}_i)^{1-y_i}. \tag{103}$$

For the independent and identically distributed ($i.i.d.$) data we define the loss

$$\mathcal{L}(\theta)=\prod_{i=1}^{n}\mathbb{P}(y_i|X_i;\theta)=\prod_{i=1}^{n}\hat{y}_i^{\,y_i}(1-\hat{y}_i)^{1-y_i},$$

which implies

$$\log\mathcal{L}(\theta)=\sum_{i=0}^{n}\left[y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)\right].$$

We take the negative of log likelihood as the probabilities are often smaller numbers. Then we define the potential function with a penalty term (i.e., L_2 or Ridge regularization),

$$U(\theta)=-\sum_{i=0}^{n}\left[y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)\right]+\frac{\lambda}{2}|\theta|^2$$

$$= -\sum_{i=0}^{n} [y_i \log(\sigma(z_i)) + (1 - y_i)\log(1 - \sigma(z_i))] + \frac{\lambda}{2}|\theta|^2.$$

Therefore, the gradient of the regularized loss function in vector form would be

$$\nabla U(\theta) = X^\top(\hat{y} - y) + \lambda\theta = X^\top(\vec{\sigma}(X\theta) - y) + \lambda\theta, \tag{104}$$

where $\vec{\sigma}$ is defined in (101). The detailed derivation of $\nabla U(\theta)$ in (104) will be given in Lemma E.1.

The fourth-order sampling process requires the computation of the integral of the gradient $\int_0^t \nabla U(\theta + tv_1)$ for any $t \in [0, \eta]$; however, for a non-polynomial or black-box potential, it is quite hard or sometimes impossible to compute the exact integrals. Thus, in this case, we approximate the integrals using **Taylor Series** expansion.

For the Taylor expansion, let us define,

$$z(t) := X(\theta + tv_1) = X\theta + tXv_1 \in \mathbb{R}^n, \qquad s(t) := \vec{\sigma}(z(t)) \in \mathbb{R}^n,$$

where $\vec{\sigma}$ is defined in (101). Furthermore, we define

$$\omega(t) = \nabla U(\theta + tv_1) = \lambda(\theta + tv_1) + X^\top \left(\vec{\sigma}(X(\theta + tv_1)) - y\right) = \lambda(\theta + tv_1) + X^\top(s(t) - y).$$

Now we expand $\omega(t)$ in the Taylor series for $t = 0$ up to a 3rd-degree polynomial to approximate the integrals in the sampling process.

$$\omega(t) = \omega(0) + \omega'(0)t + \omega''(0)\frac{t^2}{2} + \omega'''(0)\frac{t^3}{6} + \mathcal{O}(t^4). \tag{105}$$

The next steps are the computation of the derivatives. First, the constant term in (105) is given by

$$\omega(0) = \lambda\theta + X^\top(\vec{\sigma}(X\theta) - y). \tag{106}$$

We can compute that the first derivative is given by

$$\omega'(t) = \lambda v_1 + X^\top \left(\frac{ds(t)}{dt}\right).$$

Moreover,

$$\frac{ds(t)}{dt} = \vec{\sigma}(z(t)) \odot (1 - \vec{\sigma}(z(t))) \odot (Xv_1) = s(t) \odot (1 - s(t)) \odot (Xv_1), \tag{107}$$

which implies

$$\omega'(t) = \lambda v_1 + X^\top \Big[s(t) \odot (1 - s(t)) \odot (Xv_1)\Big],$$

and in particular,

$$\omega'(0) = \lambda v_1 + X^\top \Big[s \odot (1 - s) \odot (Xv_1)\Big]. \tag{108}$$

We can compute that the second derivative is given by

$$\omega''(t) = X^\top \frac{d}{dt}\left[\vec{\sigma}(z(t))(1 - \vec{\sigma}(z(t))) \odot (Xv_1)\right]. \tag{109}$$

Since $s(t) = \vec{\sigma}(z(t))$, we can compute that $s'(t) = s(t)(1-s(t))(Xv_1)$. Then we have

$$\frac{d}{dt}[s(t)(1-s(t))] = s'(t)(1-s(t)) - s(t)s'(t) = s(t)(1-s(t))(1-2s(t))(Xv_1). \quad (110)$$

Plugging (110) into (109), we obtain,

$$\omega''(t) = X^\top \Big[s(t) \odot (1-s(t)) \odot (1-2s(t)) \odot (Xv_1) \odot (Xv_1) \Big],$$

which implies

$$\omega''(0) = X^\top \Big[s \odot (1-s) \odot (1-2s) \odot (Xv_1) \odot (Xv_1) \Big]. \quad (111)$$

We can compute the third derivative is given by

$$\omega'''(t) = X^\top \frac{d}{dt} \left[s(t)(1-s(t))(1-2s(t)) \odot (Xv_1) \odot (Xv_1) \right].$$

Using the results in equation (107) and equation (110), we obtain:

$$\begin{aligned}
&\frac{d}{dt} \left[s(t)(1-s(t))(1-2s(t)) \right] \\
&= \left(s'(t)(1-s(t))(1-2s(t)) + s(t) \frac{d}{dt} \big[(1-s(t))(1-2s(t)) \big] \right) (Xv_1) \\
&= \big(s(t)(1-s(t)) \left(1 - 6s(t) + 6[s(t)]^2 \right) \big) (Xv_1),
\end{aligned}$$

which implies

$$\omega'''(0) = X^\top \Big[s \odot (1-s) \odot (1-6s+6s^2) \odot (Xv_1) \odot (Xv_1) \odot (Xv_1) \Big]. \quad (112)$$

Substituting (106), (108), (111) and (112) into (105), we get the Taylor expansion of the gradient function,

$$\begin{aligned}
\omega(t) =& \lambda\theta + X^\top (s-y) \\
&+ \left[\lambda v_1 + X^\top \left(s \odot (1-s) \odot (Xv_1) \right) \right] t \\
&+ \left[X^\top \Big(s \odot (1-s) \odot (1-2s) \odot (Xv_1) \odot (Xv_1) \Big) \right] \frac{t^2}{2} \\
&+ \left[X^\top \Big(s \odot (1-s) \odot (1-6s+6s^2) \odot (Xv_1) \odot (Xv_1) \odot (Xv_1) \Big) \right] \frac{t^3}{6} + O(t^4),
\end{aligned}$$

where $s = \sigma(X\theta) \in \mathbb{R}^n$ and $\odot$ is the elementwise (Hadamard) product. We can rewrite $\omega(t)$ as

$$\begin{aligned}
\omega(t) = \nabla U(\theta + tv_1) = \lambda\theta + M_0 + (M_1 \odot (Xv_1) + \lambda v_1)t + \frac{1}{2}(M_2 \odot Xv_1 \odot Xv_1)t^2 \\
+ \frac{1}{6}\Big(M_3 \odot Xv_1 \odot Xv_1 \odot Xv_1 \Big) t^3,
\end{aligned}$$

where

$$\begin{aligned}
&M_0 = X^\top (s-y), \qquad M_1 = X^\top \left(s \odot (1-s) \right), \\
&M_2 = X^\top \Big(s \odot (1-s) \odot (1-2s) \Big), \qquad M_3 = X^\top \Big(s \odot (1-s) \odot (1-6s+6s^2) \Big),
\end{aligned}$$

and all $\{M_i\}_{i=0}^3 \in \mathbb{R}^d$ and $s = \vec{\sigma}(X\theta) \in \mathbb{R}^n$.

Fourth-order computations: Once we have the Taylor expanded form of $\nabla U(\theta + tv_1)$ (i.e., $\omega(t)$), the calculation processes are the same as the quadratic function. We use standard mathematical software *Mathematica 12.0* to compute those nested integrals and obtain the following results.

$$\begin{aligned}
m_0 = &\left(\frac{\gamma^2\eta^4\lambda}{24} + \frac{\eta^4\lambda^2}{24} - \frac{\eta^2\lambda}{2} + \mu_{00}\right)\theta + \left(\frac{\gamma^2\eta^5\lambda}{60} + \frac{\eta^5\lambda^2}{120} - \frac{\eta^3\lambda}{6} + \mu_{01}\right)v_1 \\
&+ \left(\mu_{02} - \frac{\gamma\eta^4\lambda}{24}\right)v_2 + \left(\mu_{03} - \frac{\gamma^2\eta^5\lambda}{120}\right)v_3 + \left(\frac{\gamma^2\eta^4}{24} + \frac{\eta^4\lambda}{24} - \frac{\eta^2}{2}\right)M_0 \\
&+ \left(\frac{\gamma^2\eta^5}{120} + \frac{\eta^5\lambda}{120} - \frac{\eta^3}{6}\right)M_1 \odot Xv_1 + \left(\frac{\gamma^2\eta^6}{720} + \frac{\eta^6\lambda}{720} - \frac{\eta^4}{24}\right)M_2 \odot Xv_1 \odot Xv_1 \\
&+ \left(\frac{\gamma^2\eta^7}{5040} + \frac{\eta^7\lambda}{5040} - \frac{\eta^5}{120}\right)M_3 \odot Xv_1 \odot Xv_1 \odot Xv_1,
\end{aligned}$$

$$\begin{aligned}
m_1 = &\left(\frac{\gamma^2\eta^3\lambda}{6} + \frac{\eta^3\lambda^2}{6} - \eta\lambda\right)\theta + \left(\frac{\gamma^2\eta^4\lambda}{12} + \frac{\eta^4\lambda^2}{24} - \frac{\eta^2\lambda}{2} + \mu_{11}\right)v_1 \\
&+ \left(\mu_{12} - \frac{\gamma\eta^3\lambda}{6}\right)v_2 + \left(\mu_{13} - \frac{\gamma^2\eta^4\lambda}{24}\right)v_3 + \left(\frac{\gamma^2\eta^3}{6} + \frac{\eta^3\lambda}{6} - \eta\right)M_0 \\
&+ \left(\frac{\gamma^2\eta^4}{24} + \frac{\eta^4\lambda}{24} - \frac{\eta^2}{2}\right)M_1 \odot Xv_1 + \left(\frac{\gamma^2\eta^5}{120} + \frac{\eta^5\lambda}{120} - \frac{\eta^3}{6}\right)M_2 \odot Xv_1 \odot Xv_1 \\
&+ \left(\frac{\gamma^2\eta^6}{720} + \frac{\eta^6\lambda}{720} - \frac{\eta^4}{24}\right)M_3 \odot Xv_1 \odot Xv_1 \odot Xv_1,
\end{aligned}$$

$$\begin{aligned}
m_2 = &\left(-\frac{\gamma\eta^2\lambda}{24}\left(\gamma^2\eta^2 + \eta^2\lambda - 12\right) + \frac{\lambda\left(-\gamma^3\eta^3 + 3\gamma^2\eta^2 - 6\gamma\eta - 6e^{-\gamma\eta} + 6\right)}{6\gamma}\right)\theta \\
&+ \left(\mu_{21} - \frac{\gamma\eta^3\lambda\left(2\gamma^2\eta^2 + \eta^2\lambda - 20\right)}{120} + \frac{\lambda\left(-\gamma^4\eta^4 + 4\gamma^3\eta^3 - 12\gamma^2\eta^2 + 24\gamma\eta + 24e^{-\gamma\eta} - 24\right)}{24\gamma^2}\right)v_1 \\
&+ \left(\frac{\gamma^2\eta^4\lambda}{24} + \mu_{22}\right)v_2 + \left(\frac{\gamma^3\eta^5\lambda}{120} + \mu_{23}\right)v_3 \\
&+ \left(-\frac{\gamma\eta^2\left(\gamma^2\eta^2 + \eta^2\lambda - 12\right)}{24} - \frac{\gamma^3\eta^3 - 3\gamma^2\eta^2 + 6\gamma\eta + 6e^{-\gamma\eta} - 6}{6\gamma}\right)M_0 \\
&+ \left(\frac{\gamma^4\eta^4 - 4\gamma^3\eta^3 + 12\gamma^2\eta^2 - 24\gamma\eta - 24e^{-\gamma\eta} + 24}{24\gamma^2} - \frac{\gamma\eta^3\left(\gamma^2\eta^2 + \eta^2\lambda - 20\right)}{120}\right)M_1 \odot Xv_1 \\
&+ \left(-\frac{\gamma\eta^4\left(\gamma^2\eta^2 + \eta^2\lambda - 30\right)}{720} + \frac{1 - e^{-\gamma\eta}}{\gamma^3} - \frac{1}{120}\gamma^2\eta^5 - \frac{\eta}{\gamma^2} + \frac{\gamma\eta^4}{24} + \frac{\eta^2}{2\gamma} - \frac{\eta^3}{6}\right)M_2 \odot Xv_1 \odot Xv_1 \\
&+ \Bigg(- \frac{\gamma\eta^5\left(\gamma^2\eta^2 + \eta^2\lambda - 42\right)}{5040}
\end{aligned}$$

$$
- \frac{\gamma^6\eta^6 - 6\gamma^5\eta^5 + 30\gamma^4\eta^4 - 120\gamma^3\eta^3 + 360\gamma^2\eta^2 - 720\gamma\eta - 720e^{-\gamma\eta} + 720}{720\gamma^4} \Bigg) \\
M_3 \odot Xv_1 \odot Xv_1 \odot Xv_1,
$$

$$
\begin{aligned}
m_3 = &\Bigg(\frac{\lambda e^{-\gamma\eta}}{24\gamma^3} \Big(\gamma^6\eta^4 e^{\gamma\eta} + \gamma^4\eta^2 e^{\gamma\eta} \left(\eta^2\lambda - 24\right) - 4\gamma^3\eta \left(e^{\gamma\eta}\left(\eta^2\lambda - 18\right) - 6\right) \\
&+ 12\gamma^2 \left(e^{\gamma\eta}\left(\eta^2\lambda - 8\right) + 8\right) - 24\gamma\eta\lambda e^{\gamma\eta} + 24\lambda\left(e^{\gamma\eta} - 1\right) \Big) \Bigg) \theta \\
&+ \Bigg(\mu_{31} + \frac{\lambda e^{-\gamma\eta}}{120\gamma^4} \Big(2\gamma^7\eta^5 e^{\gamma\eta} - 5\gamma^6\eta^4 e^{\gamma\eta} + \gamma^5\eta^3 e^{\gamma\eta}\left(\eta^2\lambda - 20\right) - 5\gamma^4\eta^2 e^{\gamma\eta}\left(\eta^2\lambda - 24\right) \\
&+ 20\gamma^3\eta\left(e^{\gamma\eta}\left(\eta^2\lambda - 18\right) - 6\right) - 60\gamma^2\left(e^{\gamma\eta}\left(\eta^2\lambda - 8\right) + 8\right) + 120\gamma\eta\lambda e^{\gamma\eta} \\
&- 120\lambda\left(e^{\gamma\eta} - 1\right) \Big) \Bigg) v_1 \\
&+ \left(\mu_{32} + \frac{\lambda\left(-\gamma^4\eta^4 + 4\gamma^3\eta^3 - 12\gamma^2\eta^2 + 24\gamma\eta + 24e^{-\gamma\eta} - 24\right)}{24\gamma^2} \right) v_2 \\
&+ \left(\mu_{33} + \frac{\lambda\left(-\gamma^5\eta^5 + 5\gamma^4\eta^4 - 20\gamma^3\eta^3 + 60\gamma^2\eta^2 - 120\gamma\eta - 120e^{-\gamma\eta} + 120\right)}{120\gamma^2} \right) v_3 \\
&+ \left(\frac{\gamma^3\eta^4}{24} + \frac{\lambda - \lambda e^{-\gamma\eta}}{\gamma^3} - \frac{\eta\lambda}{\gamma^2} + \frac{4e^{-\gamma\eta} + \frac{\eta^2\lambda}{2} - 4}{\gamma} + \gamma\left(\frac{\eta^4\lambda}{24} - \eta^2\right) + \eta\left(e^{-\gamma\eta} + 3\right) - \frac{\eta^3\lambda}{6} \right) M_0 \\
&+ \Bigg(\frac{\lambda\left(e^{-\gamma\eta} - 1\right)}{\gamma^4} + \frac{\gamma^3\eta^5}{120} + \frac{\eta\lambda}{\gamma^3} + \frac{-5e^{-\gamma\eta} - \frac{\eta^2\lambda}{2} + 5}{\gamma^2} + \frac{\eta\left(-e^{-\gamma\eta} - 4\right) + \frac{\eta^3\lambda}{6}}{\gamma} \\
&+ \frac{1}{120}\gamma\eta^3\left(\eta^2\lambda - 40\right) - \frac{1}{24}\eta^2\left(\eta^2\lambda - 36\right) \Bigg) M_1 \odot Xv_1 \\
&+ \Bigg(\frac{\lambda - \lambda e^{-\gamma\eta}}{\gamma^5} - \frac{\eta\lambda}{\gamma^4} + \frac{\gamma^3\eta^6}{720} + \frac{6e^{-\gamma\eta} + \frac{\eta^2\lambda}{2} - 6}{\gamma^3} + \frac{\eta\left(e^{-\gamma\eta} + 5\right) - \frac{\eta^3\lambda}{6}}{\gamma^2} \\
&+ \frac{1}{720}\gamma\eta^4\left(\eta^2\lambda - 60\right) + \frac{\frac{\eta^4\lambda}{24} - 2\eta^2}{\gamma} - \frac{1}{120}\eta^3\left(\eta^2\lambda - 60\right) \Bigg) M_2 \odot Xv_1 \odot Xv_1 \\
&+ \Bigg(\frac{\lambda\left(e^{-\gamma\eta} - 1\right)}{\gamma^6} + \frac{\eta\lambda}{\gamma^5} + \frac{-7e^{-\gamma\eta} - \frac{\eta^2\lambda}{2} + 7}{\gamma^4} + \frac{\gamma^3\eta^7}{5040} + \frac{\eta\left(-e^{-\gamma\eta} - 6\right) + \frac{\eta^3\lambda}{6}}{\gamma^3}. + \frac{60\eta^2 - \eta^4\lambda}{24\gamma^2} \\
&+ \frac{\gamma\eta^5\left(\eta^2\lambda - 84\right)}{5040} + \frac{\eta^3\left(\eta^2\lambda - 80\right)}{120\gamma} - \frac{1}{720}\eta^4\left(\eta^2\lambda - 90\right) \Bigg) M_3 \odot Xv_1 \odot Xv_1 \odot Xv_1.
\end{aligned}
$$

Lemma E.1. *Define the potential (loss) function with a penalty term (i.e., L_2 or Ridge regularization):*

$$U(\theta) = -\sum_{i=0}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] + \frac{\lambda}{2}|\theta|^2$$
$$= -\sum_{i=0}^{n} [y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i))] + \frac{\lambda}{2}|\theta|^2.$$

Then the gradient of the regularized loss function is given as

$$\nabla U(\theta) = X^\top(\hat{y} - y) + \lambda\theta = X^\top(\vec{\sigma}(X\theta) - y) + \lambda\theta.$$

Proof. To make the calculation easier, we consider the non-regularized elementwise gradient of the loss. Additionally, we use the recursive property of the *sigmoid* function $\frac{d\sigma(z_i)}{d\theta} = \sigma(z_i)(1 - \sigma(z_i))X_i \in \mathbb{R}^d$. Thus,

$$\nabla U(\theta) = -\sum_{i=1}^{n} [y_i \nabla_\theta \log(\sigma(z_i)) + (1 - y_i)\nabla_\theta \log(1 - \sigma(z_i))]$$
$$= -\sum_{i=1}^{n} \left[y_i \frac{1}{\sigma(z_i)} \frac{d}{d\theta}\sigma(z_i) - (1 - y_i) \cdot \frac{1}{\sigma(z_i)} \frac{d}{d\theta}\sigma(z_i)\right]$$
$$= -\sum_{i=1}^{n} \left[y_i \frac{1}{\sigma(z_i)}\sigma(z_i)(1 - \sigma(z_i))X_i - (1 - y_i)\frac{1}{\sigma(z_i)}\sigma(z_i)(1 - \sigma(z_i))X_i\right]$$
$$= -\sum_{i=1}^{n} [y_i(1 - \sigma(z_i))X_i - (1 - y_i) \cdot (1 - \sigma(z_i))X_i]$$
$$= \sum_{i=1}^{n} [-y_i(1 - \sigma(z_i))X_i + (1 - y_i) \cdot (1 - \sigma(z_i))X_i]$$
$$= \sum_{i=1}^{n} [(\sigma(z_i) - y_i)X_i]$$
$$= (\sigma(z_1) - y_1)X_1 + (\sigma(z_2) - y_2)X_2 + \cdots + (\sigma(z_n) - y_n)X_n.$$

Therefore, the gradient of the regularized loss function in vector form would be

$$\nabla U(\theta) = X^\top(\hat{y} - y) + \lambda\theta = X^\top(\vec{\sigma}(X\theta) - y) + \lambda\theta.$$

□