ERROR ESTIMATES BETWEEN SGD WITH MOMENTUM AND UNDERDAMPED LANGEVIN DIFFUSION

Arnaud Guillin, Yu Wang, Lihu Xu, Haoran Yang

ERROR ESTIMATES BETWEEN SGD WITH MOMENTUM AND UNDERDAMPED LANGEVIN DIFFUSION

ARNAUD GUILLIN, YU WANG, LIHU XU, AND HAORAN YANG

Abstract. Stochastic gradient descent with momentum is a popular variant of stochastic gradient descent, which has recently been reported to have a close relationship with the underdamped Langevin diffusion. In this paper, we establish a quantitative error estimate between them in the 1-Wasserstein and total variation distances.

Keywords: Stochastic Gradient Descent with momentum (SGDm), Underdamped Langevin Diffusion, 1-Wasserstein Distance, Total Variation Distance, Variant Time Step, Malliavin Calculus

MSC2020 subject classification:60J20, 60H35, 60H30, 60F99.

Contents

EQUAL CONTRIBUTIONS

1. Introduction

Many tasks in machine learning and statistics can be formulated as an optimization problem as follows

$$\text{minimize} \quad \mathbb{E}_\xi F(\mathbf{x}, \xi), \tag{1.1}$$

where ξ is random and its distribution is not known, and $\mathbf{x} \in \mathbb{R}^d$. In practice, one needs to replace the mean $\mathbb{E}_\xi F(\mathbf{x}, \xi)$ with its sample mean as

$$\text{minimize} \quad \frac{1}{N}\sum_{i=1}^{N} F(\mathbf{x}, \xi^i), \tag{1.2}$$

where $\xi^1, ..., \xi^N$ are i.i.d. and have the same distribution as ξ. For the further use, we denote

$$f(\mathbf{x}) = \mathbb{E}_\xi F(\mathbf{x}, \xi). \tag{1.3}$$

To solve the minimization problem (1.2), one often uses online stochastic gradient descent (SGD) or online SGD with momentum (SGDm), and the latter usually converges faster or works more efficient than the former. The relationship between SGD and stochastic differential equation (SDE) has been intensively studied recently, see e.g. [LTW17, LTW19, CSX23, CLTZ20, FDBD21]. Although there have been papers reporting the connections between stochastic gradient descent with momentum (SGDm) and underdamped Langevin diffusion, see for instance [GGZ22], their relationship has not been well understood, particularly when the time tends to infinity. The primary goal of this paper is to quantify their error bound in the 1-Wasserstein and total variation distances.

The noised online SGDm has two variables $(\mathbf{m}_k, \mathbf{x}_k)$, called the moment and position respectively, satisfying

$$\begin{cases} \mathbf{m}_{k+1} = \mathbf{m}_k - \gamma\eta_{k+1}\mathbf{m}_k - \dfrac{\eta_{k+1}}{N}\displaystyle\sum_{i=1}^{N} \nabla F(\mathbf{x}_k, \xi_{k+1}^i) + \beta\sqrt{\eta_{k+1}}\boldsymbol{\zeta}_{k+1}, & k \geqslant 0, \\ \mathbf{x}_{k+1} = \mathbf{x}_k + \eta_{k+1}\mathbf{m}_k, & k \geqslant 0, \end{cases} \tag{1.4}$$

with initial value $(\mathbf{m}_0, \mathbf{x}_0) \in \mathbb{R}^{2d}$, where the constant $\gamma > 0$ is the friction-coefficient, $\beta > 0$ can be regarded as the temperature, $\{\eta_k, k \geqslant 1\}$ denote the step size, $\{\xi_k^i, k \geqslant 1,\ 1 \leqslant i \leqslant N\}$ are i.i.d. copies of ξ, and $\{\boldsymbol{\zeta}_k, k \geqslant 1\}$ are i.i.d. with standard d-dimensional normal distribution. Note that for each $k \geqslant 1$, $\nabla^k F(\mathbf{x}, \xi)$ is the k-th order derivative of F with respect to $\mathbf{x}$ throughout this paper.

We will compare the algorithm (1.4) with the underdamped Langevin diffusion $(\mathbf{M}_t, \mathbf{X}_t)_{t\geqslant 0}$ with state space $\mathbb{R}^{2d}$

$$\begin{cases} \mathrm{d}\mathbf{M}_t = -\gamma\mathbf{M}_t\,\mathrm{d}t - \nabla f(\mathbf{X}_t)\,\mathrm{d}t + \beta\,\mathrm{d}\mathbf{B}_t, \\ \mathrm{d}\mathbf{X}_t = \mathbf{M}_t\,\mathrm{d}t, \end{cases} \tag{1.5}$$

where $(\mathbf{B}_t)_{t\geqslant 0}$ is a d-dimensional standard Brownian motion. The two processes possess the same initial point $(\mathbf{M}_0, \mathbf{X}_0) = (\mathbf{m}_0, \mathbf{x}_0)$. In order to compare the algorithm (1.4) with SDE (1.5), we introduce the following intermediate stochastic system: for

$t \in [t_k, t_{k+1})$ with $k \geq 0$,

$$\begin{cases} \mathrm{d}\widetilde{\mathbf{M}}_t = -\gamma \widetilde{\mathbf{M}}_{t_k}\,\mathrm{d}t - \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \nabla F(\widetilde{\mathbf{X}}_{t_k}, \xi_{k+1}^i)\,\mathrm{d}t + \beta\,\mathrm{d}\mathbf{B}_t, \\ \mathrm{d}\widetilde{\mathbf{X}}_t = \widetilde{\mathbf{M}}_{t_k}\,\mathrm{d}t, \end{cases} \tag{1.6}$$

where $t_k = \sum_{j=1}^{k} \eta_j$ and $t_0 = 0$. It is obvious that $\{(\mathbf{m}_k, \mathbf{x}_k), k \geqslant 0\}$ and $\{(\widetilde{\mathbf{M}}_{t_k}, \widetilde{\mathbf{X}}_{t_k}), k \geqslant 0\}$ have the same distribution as long as $(\widetilde{\mathbf{M}}_{t_0}, \widetilde{\mathbf{X}}_{t_0}) = (\mathbf{m}_0, \mathbf{x}_0)$.

1.1. **Literature Review.** Comparisons between stochastic algorithms and stochastic continuous dynamics have been intensively studied recently, e.g. [GM91, RRT17, CSX23]. In particular, the online stochastic gradient Langevin descent (SGLD) related to the minimization problem (1.2) reads as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla F(\mathbf{x}_k, \xi) + \sqrt{2\beta^{-1}\eta}\boldsymbol{\zeta}_{\boldsymbol{k+1}}\ , \quad k \geqslant 0,$$

with a constant learning rate $\eta > 0$. [RRT17] studied the connection between this SGLD and the overdamped Langevin diffusion process $\mathbf{X} = (\mathbf{X}_t)_{t \geqslant 0}$

$$\mathrm{d}\mathbf{X}_t = -\nabla f(\mathbf{X}_t)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}\mathbf{B}_t,$$

see also [Ebe16, Ebe11, CHS87, Pav14] for related distribution sampling problems. Besides, [RRT17] showed that the 2-Wasserstein distance between the SGLD with constant step size η and the continuous time Langevin diffusion can be bounded by $\mathcal{O}(k\eta(\delta^{1/4} + \eta^{1/4}))$ at the k-th iteration. It is obvious that as k has an order higher than $[\eta(\delta^{1/4} + \eta^{1/4})]^{-1}$, this bound is not useful. By the Lindeberg technique, [CSX23] obtained a uniform bound $\mathcal{O}((1 + \log|\eta|)\eta)$ in 1-Wasserstein distance with respect to the time. Furthermore, [PP23] studied the convergence of the Euler-Maruyama (EM) scheme for the SDEs driven by multiplicative noises in the 1-Wasserstein and total variation distances, the steps of the EM scheme are decreasing.

The underdamped Langevin diffusion (1.5) has been extensively studied in recent years, see for instance [EGZ19, Pav14, BGM10, GM16, Wu01, Sch24, Nea11, CCBJ18]. An important advantage of the underdamped diffusion is that it often converges to the stationary faster than the overdamped diffusion, see for instance [EGZ19, CLW23, AAMN24] either in Wasserstein distance or in L^2 when the spectral gap is small by appropriately choosing friction coefficient. [Wu01] showed that the system converges to its equilibrium measure with exponential rate by constructing appropriate Lyapunov test functions. [EGZ19, Sch24] established the contraction property in the Wasserstein type distances for two solutions of (1.5) with different initial data, and [SW24] studied the EM scheme of (1.5) and its convergence in the 1-Wasserstein distance. [CCAY+18] considered the problem of sampling from distribution $p^*(\mathbf{x}) \propto \mathrm{e}^{-f(\mathbf{x})}$ where $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth everywhere and m-strongly convex outside a ball of finite radius R. Given a tolerance error ε, they proved that the complexity of the underdamped Langevin Markov chain Monte Carlo (MCMC) is $\mathcal{O}(\sqrt{d}/\varepsilon)$ if the step size is $\eta = \mathcal{O}(\varepsilon/\sqrt{d})$. What's more, [CR22, GGZ22] considered the stochastic gradient Hamilton Monte Carlo (SGHMC) for sampling a similar distribution and found a complexity bound $\mathcal{O}((k\eta)^{3/2}\sqrt{\log k\eta}(\delta^{1/4} + \eta^{1/4}) + k\eta\sqrt{\eta})$ in the 2-Wasserstein distance when the step

size is a constant η. For other applications related to underdamped Langevin diffusion, we refer the reader to [CDC15, CR22, GGZ22] for stochastic gradient Hamilton MC and to [CCBJ18, CCAY+18] for underdamped Langevin MCMC.

For other SGD algorithms such as Mini-batch SGD and Nesterov SGD, we refer the reader to [KLRT15, LZCS14, CL20, LR20, BLB17, MJ19, KF16] and the references therein. These algorithms differ in their approaches to updating the parameters during training neural networks and can be useful in improving convergence and reducing overfitting.

1.2. **Contributions and Methods.** Our contributions and methods are summarised as the below:

(1) Although there have been several papers studying the connections between the accelerated algorithms with continuous dynamics, see for instance [CR22, GGZ22], but most of them only provide qualitative limit without a quantitative error bound, particularly when the time is large. To the best of our knowledge, this paper first provides error bounds between the SGDm and the underdamped Langevin diffusion uniformly with respect to the time, these bounds depend on the dimension d polynomially and reveal convergence rates. Our results can be formally formulated as the following:

$$d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{M}_{t_n}, \mathbf{X}_{t_n}), \mathcal{L}(\mathbf{m}_n, \mathbf{x}_n)) \leqslant C\sqrt{d}(1 + 1/N)\sqrt{\eta_n},$$

$$d_{\mathrm{TV}}(\mathcal{L}(\mathbf{M}_{t_n}, \mathbf{X}_{t_n}), \mathcal{L}(\mathbf{m}_n, \mathbf{x}_n)) \leqslant Cd^4(\sqrt{\eta_n} + 1/\sqrt{N}).$$

We clearly see a significant difference between the above two bounds, one being of an order $O(\sqrt{\eta_n} + \sqrt{\eta_n}/N)$ in the 1-Wasserstein distance and the other $O(\sqrt{\eta_n} + 1/\sqrt{N})$ in the total variation distance. We think that the rate $O(1/\sqrt{N})$ in the total variation distance is due to the singularity of this distance and that it is hard to improve the $O(1/\sqrt{N})$ without paying a price on the rate $\sqrt{\eta_n}$.

To the best of our knowledge, most of the known results related to underdamped Langevin diffusion samplings and discretization schemes have a constant step size η, while our SGDm has the non-increasing step sizes η_k which include the constant step size case.

(2) There are two systems of noises, ξ_k^i and $\boldsymbol{\zeta}_{k+1}$, in SGDm (1.4), whose interplays make the SGDm much more complex than the underdamped Langevin sampling by discretising SDE (1.5). These complexities can be clearly seen in bounding the total variation distance, where we need to estimate a Malliavin matrix related to $\boldsymbol{\zeta}_{k+1}$ by splitting the regimes of ξ_k^i. What's more, ξ_k^i can be heavy tailed. More precisely, we only need to assume that ξ_k^i has second and fourth moments to bound $d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{M}_{t_n}, \mathbf{X}_{t_n})$ and $d_{\mathrm{TV}}(\mathcal{L}(\mathbf{M}_{t_n}, \mathbf{X}_{t_n}), \mathcal{L}(\mathbf{m}_n, \mathbf{x}_n))$ respectively.

(3) The term $N^{-1}\sum_{i=1}^N \nabla F(x, \xi_{k+1}^i)$ is an unbiased estimator of $\nabla f(x)$. According to the law of large number, it converges to $\nabla f(x)$ almost surely as $N \to \infty$. This, together with the exponential ergodicity of $(\mathbf{M}_t, \mathbf{X}_t)$, immediately implies an error bound in the order $O(\sqrt{\eta_n})$ for the underdamped Langevin sampling as n is large. The rate $O(\sqrt{\eta_n})$ may be not optimal due to the heavy tail effect of ξ_k^i, we conjecture that the rate $O(\sqrt{\eta_n})$ can be improved by assuming that ξ_k^i has a high order moment. We leave this possible improvement to the future research.

(4) Although the approach of our proofs is still via the classical Lindeberg principle, in contrast to the overdamped Langevin diffusion, the underdamped Langevin diffusion is degenerate and the following difficulties naturally arise: (i) the regularity

problems are very involved and we need Malliavin calculus to handle them; (ii) there are interplays between the randomnesses of ξ_k^i and $\boldsymbol{\zeta}_{k+1}$, it is much more difficult for us to estimate the Malliavin matrix and figure out the polynomial dependence on the dimension d; (iii) the related Lyapunov function is much more subtle and not intuitive.

1.3. **Structure of The Paper.** In the next section, we will introduce our assumptions and main theorems. Section 3 will provide all auxiliary lemmas, which include estimates of p-th moments, the contraction property of the 1-Wasserstein distance, and additional estimates related to the total variation distance. The proofs for the main results and their corollaries will be provided in Section 4. In Appendix A, we give the supporting lemmas related to the Lyapunov function. At last, the lemmas in Section 3 will be proved in Appendix B.

Acknowledgements: We would like to thank Professor Feng-Yu Wang for very helpful discussion and pointing out the references [Sch24] and [SW24] to us. A. Guillin is benefited from a government grant managed by the Agence Nationale de la Recherche under the France 2030 investment plan ANR-23-EXMA-0001"n and under the grant ANR-23-CE40-0003. L. Xu is supported by the National Natural Science Foundation of China No. 12071499, the Science and Technology Development Fund (FDCT) of Macau S.A.R. FDCT 0074/2023/RIA2, and the University of Macau grants MYRG2020-00039-FST, MYRG-GRG2023-00088-FST.

2. Preliminary and Main Results

2.1. **Notations and Assumptions.** We use normal font for scalars (e.g. $a, A, \dots$) and boldface for vectors and matrices (e.g. $\mathbf{x}, \mathbf{y}, \mathbf{A}, \mathbf{B}, \dots$). For any vectors $\mathbf{u} = (u_1, \dots, u_d)$, $\mathbf{v} = (v_1, \dots, v_d)$ in $\mathbb{R}^d$, their standard inner product is denoted by $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^d u_i v_i$, and denote the corresponding Euclidean norm as $|\mathbf{u}| = \left(\sum_{i=1}^d u_i^2\right)^{1/2}$. The HilbertSchmidt inner product for matrices $\mathbf{A} = (A_{ij})_{d\times d}, \mathbf{B} = (B_{ij})_{d\times d} \in \mathbb{R}^{d\times d}$ is denoted by $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{HS}} = \sum_{i,j=1}^d A_{ij} B_{ij}$, the HilbertSchmidt norm is defined as $\|\mathbf{A}\|_{\mathrm{HS}} = (\sum_{i,j=1}^d A_{ij}^2)^{1/2}$. Besides, the operator norm of $\mathbf{A}$ is denoted by $\|\mathbf{A}\|_{\mathrm{op}} = \sup_{|\mathbf{v}|=1} |\mathbf{A}\mathbf{v}|$, which has the following relationship with $\|\mathbf{A}\|_{\mathrm{HS}}$

$$\|\mathbf{A}\|_{\mathrm{op}} \leqslant \|\mathbf{A}\|_{\mathrm{HS}} \leqslant \sqrt{d}\, \|\mathbf{A}\|_{\mathrm{op}}\,.$$

If a matrix $\mathbf{A}$ is positive semi-definite, we define $\lambda_{\max}(\mathbf{A})$ as its maximal eigenvalue and $\lambda_{\min}(\mathbf{A})$ as its minimal eigenvalue.

Let $\mathcal{C}(\mathbb{R}^d, \mathbb{R})$ denote the set of all continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$ and $\mathcal{C}_b(\mathbb{R}^d, \mathbb{R})$ denote the set of all bounded continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$. For $k \geq 0$, denote by $\mathcal{C}^k(\mathbb{R}^d, \mathbb{R})$ the set of functions from $\mathbb{R}^d$ to $\mathbb{R}$ which has continuous 0-th,...,k-th order derivatives, further denote by $\mathcal{C}_b^k(\mathbb{R}^d, \mathbb{R})$ the set of functions from $\mathbb{R}^d$ to $\mathbb{R}$ which has bounded continuous 0-th,...,k-th order derivatives and by $\mathcal{C}_p^k(\mathbb{R}^d, \mathbb{R})$ the set of functions from $\mathbb{R}^d$ to $\mathbb{R}$ whose 0-th,...,k-th order derivatives have polynomial growth. For $g \in \mathcal{C}^3(\mathbb{R}^d, \mathbb{R})$ and $\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{x} \in \mathbb{R}^d$, we denote that

$$\nabla_{\mathbf{v}} g(\mathbf{x}) = \lim_{\varepsilon \to 0} \frac{g(\mathbf{x} + \varepsilon \mathbf{v}) - g(\mathbf{x})}{\varepsilon},$$
$$\nabla_{\mathbf{v}_2} \nabla_{\mathbf{v}_1} g(\mathbf{x}) = \lim_{\varepsilon \to 0} \frac{\nabla_{\mathbf{v}_1} g(\mathbf{x} + \varepsilon \mathbf{v}_2) - \nabla_{\mathbf{v}_1} g(\mathbf{x})}{\varepsilon},$$

$$\nabla_{\mathbf{v}_3}\nabla_{\mathbf{v}_2}\nabla_{\mathbf{v}_1}g(\mathbf{x}) = \lim_{\varepsilon\to 0}\frac{\nabla_{\mathbf{v}_2}\nabla_{\mathbf{v}_1}g(\mathbf{x}+\varepsilon\mathbf{v}_3) - \nabla_{\mathbf{v}_2}\nabla_{\mathbf{v}_1}g(\mathbf{x})}{\varepsilon},$$

as the directional derivatives of g. We know $\nabla g(\mathbf{x}) \in \mathbb{R}^d$, $\nabla^2 g(\mathbf{x}) \in \mathbb{R}^{d\times d}$, $\nabla^3 g(\mathbf{x}) \in \mathbb{R}^{d\times d\times d}$. Moreover, we define the operator norm of $\nabla^k g(\mathbf{x})$, $k = 2, 3$ with

$$\left\|\nabla^k g(\mathbf{x})\right\|_{\mathrm{op}} = \sup_{|\mathbf{v}_i|=1,\, i=1,\dots,k} \left|\nabla_{\mathbf{v}_k}\dots\nabla_{\mathbf{v}_1} g(\mathbf{x})\right|.$$

If g is bounded, we denote its infinity norm by $\|g\|_\infty = \sup_{\mathbf{x}} |g(\mathbf{x})|$. If g is a Lipschitz function, its Lipschitz norm is denoted by $\|g\|_{\mathrm{Lip}} = \sup_{\mathbf{x}\neq\mathbf{y}} |g(\mathbf{x}) - g(\mathbf{y})| / |\mathbf{x}-\mathbf{y}|$. Additionally, for a function $h\colon (\mathbf{m},\mathbf{x}) \in \mathbb{R}^{2d} \mapsto h(\mathbf{m},\mathbf{x}) \in \mathbb{R}$, we denote its derivatives with respect to $\mathbf{m}$ and $\mathbf{x}$ as $\nabla_{\mathbf{m}} h$ and $\nabla_{\mathbf{x}} h$ respectively. Whenever we need to emphasize the initial value $(\mathbf{M}_0, \mathbf{X}_0) = (\mathbf{m}, \mathbf{x})$ for given $(\mathbf{m}, \mathbf{x}) \in \mathbb{R}^{2d}$, we will write $(\mathbf{M}_t^{\mathbf{m}}, \mathbf{X}_t^{\mathbf{x}})$. Similarly, we use the notation $(\widetilde{\mathbf{M}}_t^{\mathbf{m}}, \widetilde{\mathbf{X}}_t^{\mathbf{x}})$. The operator semigroup induced by $(\mathbf{M}_t^{\mathbf{m}}, \mathbf{X}_t^{\mathbf{x}})_{t\geqslant 0}$ from (1.5) is given by

$$\mathrm{P}_t\, h(\mathbf{m},\mathbf{x}) = \mathbb{E}h(\mathbf{M}_t^{\mathbf{m}}, \mathbf{X}_t^{\mathbf{x}}), \quad h \in \mathcal{C}_b(\mathbb{R}^{2d}, \mathbb{R}),\ t > 0.$$

Its infinitesimal generator $\mathscr{A}$ is defined as

$$\mathscr{A}h(\mathbf{m},\mathbf{x}) := -\langle \nabla_{\mathbf{m}} h, \gamma\mathbf{m} + \nabla f(\mathbf{x})\rangle + \langle \nabla_{\mathbf{x}} h, \mathbf{m}\rangle + \frac{1}{2}\beta^2\, \Delta_{\mathbf{m}} h,$$

for any $h \in \mathcal{C}_p^2(\mathbb{R}^{2d}, \mathbb{R})$. The exact form for the domain of $\mathscr{A}$ is not necessarily figured out in this paper.

Let $\mathscr{P}(\mathbb{R}^{2d})$ denote the space of probability distributions on $\mathbb{R}^{2d}$. For any $\mu, \nu \in \mathscr{P}(\mathbb{R}^{2d})$, denote their 1-Wasserstein distance by

$$d_{\mathcal{W}_1}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} \left|\mathbf{z}^1 - \mathbf{z}^2\right| \mathrm{d}\pi(\mathbf{z}^1, \mathbf{z}^2), \quad \mu,\nu \in \mathscr{P}(\mathbb{R}^{2d}),$$

where $\Pi(\mu,\nu)$ denotes the set of probability distributions on $\mathbb{R}^{2d}\times\mathbb{R}^{2d}$ with marginal distributions μ and ν. Kantorovich-Rubinstein Theorem tells us that

$$d_{\mathcal{W}_1}(\mu,\nu) = \sup_{\|h\|_{\mathrm{Lip}}\leqslant 1} \int_{\mathbb{R}^{2d}} h(\mathbf{z})\ \mathrm{d}(\mu-\nu)(\mathbf{z}), \quad \mu,\nu \in \mathscr{P}(\mathbb{R}^{2d}).$$

Besides, the total variation distance between μ and ν can be defined by

$$d_{\mathrm{TV}}(\mu,\nu) = \sup_{\|g\|_\infty\leqslant 1} \int_{\mathbb{R}^{2d}} g(\mathbf{z})\ \mathrm{d}(\mu-\nu)(\mathbf{z}), \quad \mu,\nu \in \mathscr{P}(\mathbb{R}^{2d}).$$

Throughout this paper, we propose the following assumptions.

Assumption I. *The function $f(\mathbf{x})$, defined by (1.3), is non-negative and satisfies the following conditions:*

(i) *There exist constants $A, B \geqslant 0$ such that*

$$|f(\mathbf{0})| \leqslant A, \quad |\nabla f(\mathbf{0})| \leqslant B. \tag{2.1}$$

(ii) *f is L-smooth for some constant $L > 0$, that is,*

$$|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})| \leqslant L\,|\mathbf{x}-\mathbf{y}|, \quad \forall\ \mathbf{x},\mathbf{y} \in \mathbb{R}^d. \tag{2.2}$$

(iii) *For any $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$, there exist positive constants a and b such that*

$$\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle \geqslant a\,|\mathbf{x}-\mathbf{y}|^2 - b. \tag{2.3}$$

Assumption II. *Let $\mathcal{U}$ be a measurable space. ξ is a $\mathcal{U}$-valued random variable such that $\nabla F(\mathbf{x},\xi)$ satisfies the following conditions:*

(i) *For any $\mathbf{x}\in\mathbb{R}^d$,*

$$\mathbb{E}_\xi[\nabla F(\mathbf{x},\xi)]=\nabla f(\mathbf{x}).$$

(ii) *There exists a constant $A_0>0$ such that for some $q\geqslant 2$,*

$$\sup_{\mathbf{x}\in\mathbb{R}^d}\left\{\mathbb{E}_\xi[|\nabla f(\mathbf{x})-\nabla F(\mathbf{x},\xi)|^q]\right\}^{1/q}\leqslant A_0. \tag{2.4}$$

Here, $F(\mathbf{x},\xi)$ and $f(\mathbf{x})$ are in (1.1) and (1.3) respectively.

In practical applications, the choice of step size often varies with each iteration k. It is necessary and important to add an additional assumption as following to restrict the behavior of η_k.

Assumption III. *Let $(\eta_k)_{k\geqslant 1}$ be a non-increasing and positive sequence and let $t_n=\sum_{k=1}^n\eta_k$, which satisfy the following conditions:*

(i) $\lim_{n\to\infty}t_n=+\infty$;

(ii) *There exists a constant $\omega\in(0,2\theta)$ such that*

$$\eta_k\leqslant\frac{2\theta-\omega}{2\theta^2}\quad\text{and}\quad\eta_{k-1}-\eta_k\leqslant\omega\eta_k^2,\quad\forall\ k\geqslant 1, \tag{2.5}$$

where the constant θ will be given in Lemma 3.6.

A typical example is $\eta_k=\eta/k^\alpha$ for some constants $\eta>0$ and $\alpha\in(0,1)$, then condition (2.5) holds for sufficiently large k.

2.2. **Main Results.** We are now in the position to state our main results, whose proofs will be given in Section 4. Define the Lyapunov function $\mathcal{V}(\mathbf{m},\mathbf{x}):\mathbb{R}^{2d}\to\mathbb{R}$ of SDE (1.5) as

$$\mathcal{V}(\mathbf{m},\mathbf{x}):=f(\mathbf{x})+\frac{\gamma^2}{4}\left(\left|\mathbf{x}+\frac{1}{\gamma}\mathbf{m}\right|^2+\left|\frac{1}{\gamma}\mathbf{m}\right|^2-\lambda|\mathbf{x}|^2\right), \tag{2.6}$$

where constant $0<\lambda\leqslant\frac{1}{4}\wedge\frac{a}{4L+\gamma^2}$. More details about this Lyapunov function can be found in Appendix A.

We denote by $\mathcal{L}(\mathbf{M}_{t_n},\mathbf{X}_{t_n})$ and $\mathcal{L}(\mathbf{m}_n,\mathbf{x}_n)$ the laws of $(\mathbf{M}_{t_n},\mathbf{X}_{t_n})$ from (1.5) and $(\mathbf{m}_n,\mathbf{x}_n)$ from (1.4) respectively.

Throughout this paper, we shall use the letters c,C,C_1,C_2,C_3 to denote positive numbers which may depend on the parameters $A_0,A,B,L,a,b,q,\theta,\omega$ in Assumptions I, II, and III above, and A_0',q',B_1,B_2 in Assumption IV below. Their values may vary from line to line, but do NOT depend on the dimension d and the sample size N. In some cases, they may depend on other parameters such as p and we will stress this in a way like 'C also depends on p' if necessary.

Theorem 1. *Under Assumptions I, II, and III, we assume that $\eta_1\leqslant c$ for some positive constant c, and $\gamma>\sqrt{2}(2L+a)/\sqrt{a}$ additionally. Then we have*

$$d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{M}_{t_n},\mathbf{X}_{t_n}),\mathcal{L}(\mathbf{m}_n,\mathbf{x}_n))\leqslant C\sqrt{\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+d}\left(1+\frac{1}{N}\right)\sqrt{\eta_n},\quad\forall n\geqslant 0.$$

To estimate the total variation distance, we rely on the following assumptions regarding the derivatives of the function $F(\mathbf{x},\xi)$, which are stronger than Assumption II.

Assumption IV. *Functions $F(\mathbf{x},\xi)$ and $f(\mathbf{x})$ are in (1.1) and (1.3) respectively. We assume that*

(i) *For any $\mathbf{x}\in\mathbb{R}^d$,*

$$\mathbb{E}_\xi\left[\nabla F(\mathbf{x},\xi)\right]=\nabla f(\mathbf{x}) \quad \text{and} \quad \mathbb{E}_\xi\left[\nabla^2 F(\mathbf{x},\xi)\right]=\nabla^2 f(\mathbf{x}).$$

(ii) *There exists a constant $A_0'>0$ such that for some $q'\geqslant 4$,*

$$\sup_{\mathbf{x}\in\mathbb{R}^d}\left\{\mathbb{E}_\xi\left[\,|\nabla f(\mathbf{x})-\nabla F(\mathbf{x},\xi)|^{q'}\,\right]\right\}^{1/q'}\leqslant A_0'. \tag{2.7}$$

(iii) *There exist some positive constants B_1 and B_2 such that*

$$\left\{\mathbb{E}\left[\sup_{\mathbf{x}\in\mathbb{R}^d}\left\|\nabla^2 F(\mathbf{x},\xi)\right\|_{\mathrm{op}}^8\right]\right\}^{1/8}\leqslant B_1,$$

$$\max\left\{\sup_{\mathbf{x}\in\mathbb{R}^d}\left\|\nabla^3 f(\mathbf{x})\right\|_{\mathrm{op}},\ \sup_{\mathbf{x}\in\mathbb{R}^d}\left(\mathbb{E}\left\|\nabla^3 F(\mathbf{x},\xi)\right\|_{\mathrm{op}}^4\right)^{1/4}\right\}\leqslant B_2.$$

Note that (2.7) covers the (2.4). We have the following result about total variation distance.

Theorem 2. *Under Assumptions I, II, III and IV, we assume that $\eta_1\leqslant cd^{-2}$ for some constant c, and $\gamma>\sqrt{2}(2L+a)/\sqrt{a}$ additionally. Then we have*

$$d_{\mathrm{TV}}(\mathcal{L}(\mathbf{M}_{t_n},\mathbf{X}_{t_n}),\mathcal{L}(\mathbf{m}_n,\mathbf{x}_n))\leqslant Cd^{7/2}\sqrt{\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+d}\left(\sqrt{\eta_n}+\frac{1}{\sqrt{N}}\right),\quad\forall\ n\geqslant 0.$$

Above two results yield the following corollaries for the case of $\eta_k=\eta/k^\alpha$ for some constant $\eta>0$ and $\alpha\in(0,1)$.

Corollary 3. *Under the same conditions in Theorem 1, let η be a (small) positive constant and $\eta_k=\eta/k^\alpha$ with $\alpha\in(0,1)$. Then, there exists a constant C such that for all $n\geqslant 0$*

$$d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{M}_{t_n},\mathbf{X}_{t_n}),\mathcal{L}(\mathbf{m}_n,\mathbf{x}_n))\leqslant C\sqrt{\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+d}\left(1+\frac{1}{N}\right)\sqrt{\frac{\eta}{n^\alpha}}.$$

Corollary 4. *Under the same conditions in Theorem 2, let η be a (small) positive constant and $\eta_k=\eta/k^\alpha$ with $\alpha\in(0,1)$. Then, there exists a constant C such that for all $n\geqslant 0$*

$$d_{\mathrm{TV}}(\mathcal{L}(\mathbf{M}_{t_n},\mathbf{X}_{t_n}),\mathcal{L}(\mathbf{m}_n,\mathbf{x}_n))\leqslant Cd^{7/2}\sqrt{\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+d}\left(\sqrt{\frac{\eta}{n^\alpha}}+\frac{1}{\sqrt{N}}\right).$$

With regard to the original minimization problem, we have the following generalization error bound.

Corollary 5. *Under the same conditions in Theorem 1, let $\mathbf{x}^*=\arg\min_{\mathbf{x}\in\mathbb{R}^d}f(\mathbf{x})$ be the minimizer of $f(\mathbf{x})$ in (1.3). It holds*

$$\mathbb{E}F(\mathbf{x}_n,\xi)-f(\mathbf{x}^*)\leqslant\left\{C\left[\left(1+\frac{1}{N}\right)\sqrt{\eta_n}\right]^{\frac{q-2}{q-1}}+B_e\mathrm{e}^{-\kappa t_n}\right\}\left(\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+d\right)+Jd,$$

where B_e and κ are the constants in (A.3), q is in Assumption II, and J has the form of

$$J=\frac{\beta^2}{4}\log\left[\frac{2\mathrm{e}L}{a}\left(\frac{2ab+B^2}{da\beta^2}+1\right)\right].$$

That is, the SGDm described by (1.4) is expected to approach the minimum point of the function $f(\mathbf{x})$ under appropriate parameters by choosing sufficiently large time n and large sample size N, and sufficiently small β. Moreover, the larger q in Assumption II is, the faster convergence we can obtain.

Remark 6. (i). Note that these results above hold for constant learning rate $\eta_n \equiv \eta$ for all $n \in \mathbb{N}$ as long as the η is a sufficiently small number.

(ii). In the algorithm (1.4), the term $N^{-1}\sum_{i=1}^{N}\nabla F(x,\xi_{k+1}^i)$ is an unbiased estimator of $\nabla f(x)$. According to the law of large number, it converges to $\nabla f(x)$ almost surely as $N\to\infty$. This, together with the exponential ergodicity of $(\mathbf{M}_t,\mathbf{X}_t)$, immediately implies an error bound in the order $O(\sqrt{\eta_n})$ in the 1-Wasserstein and total variation distance for the underdamped Langevin sampling as n is large. The rate $O(\sqrt{\eta_n})$ may be not optimal due to the heavy tail effect of ξ_k^i, we conjecture that the rate $O(\sqrt{\eta_n})$ can be improved by assuming that ξ_k^i has a high order moment. We leave this possible improvement to the future research.

3. Auxiliary Lemmas

In this section, we list auxiliary lemmas which will be used to prove our main results, and their proofs will be given in Appendix B.

Combining (2.1) and (2.3) implies the following dissipation property of f

$$\langle \mathbf{x}, \nabla f(\mathbf{x})\rangle \geqslant \frac{a}{2}|\mathbf{x}|^2 - K, \quad \forall\, \mathbf{x}\in\mathbb{R}^d,$$

where $K = b + B^2/(2a)$. By (2.2), the following linear growth condition holds

$$|\nabla f(\mathbf{x})| \leqslant L|\mathbf{x}| + B, \quad \forall\, \mathbf{x}\in\mathbb{R}^d. \tag{3.1}$$

Recall the Lyapunov function (2.6). Since $\lambda \leqslant 1/4$ and f is nonnegative, it is easy to see

$$0 \leqslant \max\left\{\frac{1-2\lambda}{4(1-\lambda)}|\mathbf{m}|^2\ ,\ \frac{\gamma^2}{8}(1-2\lambda)|\mathbf{x}|^2\right\} \leqslant \mathcal{V}(\mathbf{m},\mathbf{x}), \quad \forall\, \mathbf{m},\mathbf{x}\in\mathbb{R}^d. \tag{3.2}$$

One can verify that

$$\mathscr{A}\mathcal{V}(\mathbf{m},\mathbf{x}) \leqslant -\lambda\gamma\mathcal{V}(\mathbf{m},\mathbf{x}) + \frac{1}{2}\left(\gamma\mathring{A} + d\beta^2\right), \tag{3.3}$$

where constants λ and $\mathring{A}$ satisfy

$$0 < \lambda \leqslant \min\left\{\frac{1}{4}, \frac{a}{4L+\gamma^2}\right\}, \quad \mathring{A} \geqslant K + 2\lambda\left(\frac{B^2}{2L} + A\right).$$

More details about the Lyapunov function can be found in Appendix A.

In order to estimate SDE (1.6), we need to introduce the following auxiliary one step SDE: for $t\in[0,\eta]$ with η being a step size,

$$\begin{cases} \mathrm{d}\widetilde{\mathbf{M}}_t = -\gamma\widetilde{\mathbf{M}}_0\,\mathrm{d}t - \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\nabla F(\widetilde{\mathbf{X}}_0,\xi^i)\,\mathrm{d}t + \beta\,\mathrm{d}\mathbf{B}_t, \\ \mathrm{d}\widetilde{\mathbf{X}}_t = \widetilde{\mathbf{M}}_0\,\mathrm{d}t, \end{cases} \tag{3.4}$$

where $(\widetilde{\mathbf{M}}_0,\widetilde{\mathbf{X}}_0) = (\mathbf{m},\mathbf{x})$, and $\xi^1,...,\xi^N$ are i.i.d. and satisfy Assumption II.

3.1. **Moments Estimates.** We give in this subsection the moments estimates for $\mathbf{M}_t$, $\mathbf{X}_t$, $\widetilde{\mathbf{M}}_{t_n}$ and $\widetilde{\mathbf{X}}_{t_n}$.

Lemma 3.1. *Under Assumption I, for each $p \geqslant 1$, there exists a positive number C, which also depends on p, such that for any $t \geqslant 0$, and initial value $(\mathbf{m}, \mathbf{x})$,*

$$\mathbb{E}\left[\mathcal{V}(\mathbf{M}_t^{\mathbf{m}}, \mathbf{X}_t^{\mathbf{x}})^p\right] \leqslant \mathrm{e}^{-\lambda\gamma t}\mathcal{V}(\mathbf{m}, \mathbf{x})^p + Cd^p.$$

Combining this lemma with (3.2), for each $p \geqslant 1$, there exists a positive number C, which also depends on p, such that for all $t \geqslant 0$ and initial value $(\mathbf{m}, \mathbf{x})$

$$\mathbb{E}\left|\mathbf{M}_t^{\mathbf{m}}\right|^{2p} + \mathbb{E}\left|\mathbf{X}_t^{\mathbf{x}}\right|^{2p} \leqslant C\left(\mathrm{e}^{-\lambda\gamma t}\mathcal{V}(\mathbf{m}, \mathbf{x})^p + d^p\right). \tag{3.5}$$

Then, using (3.1) and (3.5), we can obtain the following estimates.

Lemma 3.2. *Under Assumption I, for each $p \geqslant 1$, there exists a positive number C, which also depends on p, such that for all $t \in [0, 1]$ and initial value $(\mathbf{m}, \mathbf{x})$:*

$$\mathbb{E}\left|\mathbf{M}_t^{\mathbf{m}} - \mathbf{m}\right|^{2p} + \mathbb{E}\left|\mathbf{X}_t^{\mathbf{x}} - \mathbf{x}\right|^{2p} \leqslant Ct^p(\mathcal{V}(\mathbf{m}, \mathbf{x})^p + d^p).$$

What's more, we have the similar consequences for SDE (3.4),

Lemma 3.3. *Consider SDE (3.4). Under Assumptions I and II, for each $1 \leqslant p \leqslant q/2$, there exists some positive number C, which also depends on p, such that,*

$$\mathbb{E}\left|\widetilde{\mathbf{M}}_t^{\mathbf{m}} - \mathbf{m}\right|^{2p} + \mathbb{E}\left|\widetilde{\mathbf{X}}_t^{\mathbf{x}} - \mathbf{x}\right|^{2p} \leqslant Ct^p(\mathcal{V}(\mathbf{m}, \mathbf{x})^p + d^p), \tag{3.6}$$

for all $t \in [0, \eta]$ with η being a step size and initial value $(\mathbf{m}, \mathbf{x})$. If the condition (2.7) in Assumption IV holds additionally, (3.6) holds for any $1 \leqslant p \leqslant q'/2$.

On the other hand, using the estimate for SDE (3.4) in an inductive way, SDE (1.6) has moments estimates similar to Lemma 3.1.

Lemma 3.4. *Under Assumptions I, II and III, let $\eta_1 \leqslant c$ for some positive constant c, and $t_n = \sum_{k=1}^n \eta_k$. For each $1 \leqslant p \leqslant q/2$, there exists a positive number C, which also depends on p, such that*

$$\mathbb{E}\left[\mathcal{V}\big(\widetilde{\mathbf{M}}_{t_n}^{\mathbf{m}}, \widetilde{\mathbf{X}}_{t_n}^{\mathbf{x}}\big)^p\right] \leqslant \mathrm{e}^{-\frac{\lambda\gamma}{2}t_n}\mathcal{V}(\mathbf{m}, \mathbf{x})^p + Cd^p, \tag{3.7}$$

for all $n \in \mathbb{N}$ and initial value $(\mathbf{m}, \mathbf{x})$. If the condition (2.7) in Assumption IV holds additionally, (3.7) holds for any $1 \leqslant p \leqslant q'/2$.

3.2. **Auxiliary Lemmas for 1-Wasserstein Distance.** The following lemma provides a bound for the difference between $(\mathbf{M}_\eta^{\mathbf{m}}, \mathbf{X}_\eta^{\mathbf{x}})$ and $(\widetilde{\mathbf{M}}_\eta^{\mathbf{m}}, \widetilde{\mathbf{X}}_\eta^{\mathbf{x}})$.

Lemma 3.5. *Consider SDE (3.4). Under Assumptions I and II, there exists a positive number $C > 0$ such that for any $\eta \in (0, 1)$,*

$$d_{\mathcal{W}_1}\big(\mathcal{L}(\mathbf{M}_\eta^{\mathbf{m}}, \mathbf{X}_\eta^{\mathbf{x}}), \mathcal{L}(\widetilde{\mathbf{M}}_\eta^{\mathbf{m}}, \widetilde{\mathbf{X}}_\eta^{\mathbf{x}})\big) \leqslant C\eta^{3/2}\left(1 + \frac{1}{N}\right)\sqrt{\mathcal{V}(\mathbf{m}, \mathbf{x}) + d}. \tag{3.8}$$

Let the condition (2.7) in Assumption IV hold additionally, we have

$$\mathbb{E}\left|\mathbf{M}_\eta^{\mathbf{m}} - \widetilde{\mathbf{M}}_\eta^{\mathbf{m}}\right|^4 + \mathbb{E}\left|\mathbf{X}_\eta^{\mathbf{x}} - \widetilde{\mathbf{X}}_\eta^{\mathbf{x}}\right|^4 \leqslant C\eta^6(\mathcal{V}(\mathbf{m}, \mathbf{x})^2 + d^2) + C\frac{\eta^4}{N^2}. \tag{3.9}$$

[Sch24] obtained a global contractivity for Langevin dynamics which will help us to prove the main result. Conveniently, we briefly introduce it here.

Recall the condition (2.3)

$$\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle \geqslant a\,|\mathbf{x}-\mathbf{y}|^2 - b, \quad \forall\, \mathbf{x},\mathbf{y}\in\mathbb{R}^d,$$

then it holds

$$\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle \geqslant \frac{a}{2}\,|\mathbf{x}-\mathbf{y}|^2, \quad \forall\, \mathbf{x},\mathbf{y}\in\mathbb{R}^d \text{ such that } |\mathbf{x}-\mathbf{y}|^2 \geqslant \frac{2b}{a}.$$

Together with $|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})| \leqslant L$ for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, f is a potential function with a L-Lipschitz continuous gradient and that is $a/2$-strongly convex outside a Euclidean ball of radius $\sqrt{2b/a}$. Then, [Sch24, Theorem 5] immediately implies the following lemma when γ is sufficiently large such that

$$\gamma^2 > \frac{2(2L+a)^2}{a}.$$

Lemma 3.6. *Suppose that $(\mathbf{M}_t^1, \mathbf{X}_t^1)_{t\geqslant 0}$ and $(\mathbf{M}_t^2, \mathbf{X}_t^2)_{t\geqslant 0}$ satisfy SDE (1.5) with different initial values $(\mathbf{m}^1, \mathbf{x}^1)$ and $(\mathbf{m}^2, \mathbf{x}^2)$ respectively. Let Assumption I hold and assume $\gamma > \sqrt{2}(2L+a)/\sqrt{a}$ additionally. Then there exist positive constants C and θ (θ does not depend on d) such that*

$$d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{M}_t^1, \mathbf{X}_t^1), \mathcal{L}(\mathbf{M}_t^2, \mathbf{X}_t^2)) \leqslant C\mathrm{e}^{-\theta t}\big(|\mathbf{m}^1 - \mathbf{x}^1| + |\mathbf{m}^2 - \mathbf{x}^2|\big),$$

for all $t \geqslant 0$.

3.3. **Auxiliary Lemmas for Total Variation Distance.** The following lemma is a direct application of Zhang et al. [Zha10], and we will give the details in Appendix B.3.

Lemma 3.7. *For any fixed $T > 0$, Let process $\mathbf{Y}_t = (\mathbf{M}_t, \mathbf{X}_t)_{t\in[0,T]}$ come from SDE (1.5) and let Assumption I hold. Then, for any function $\phi \in C_b^1(\mathbb{R}^{2d}, \mathbb{R})$, there exists a positive number $C > 0$ such that*

$$|\nabla \mathbb{E}\phi(\mathbf{Y}_T)| \leqslant C\,\|\phi\|_\infty \left(T^{3/2} \vee T^{-3/2}\right).$$

To estimate the total variation distance, we need to use Malliavin calculus. Let us briefly introduce its preliminary in our setting. More details can be found in [Nua06].

Denote $\mathcal{H} = L^2([0,T];\mathbb{R}^d)$. Let $\mathbf{W} = (W_t^1, \ldots, W_t^d)_{t\geqslant 0}$ be a d-dimensional Wiener process (a.k.a. Brownian motion) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is the σ-field generated by $\mathbf{W}$. For any $\mathbf{h} = (h_1, \ldots, h_d) \in \mathcal{H}$, define the Wiener integral as

$$\mathbf{W}(\mathbf{h}) = \sum_{i=1}^d \int_0^T h_i(t)\,\mathrm{d}W_t^i.$$

We denote by $\mathcal{C}_p^\infty(\mathbb{R}^m, \mathbb{R})$ the set of infinitely differentiable functions $g : \mathbb{R}^m \to \mathbb{R}$ such that g and all of its partial derivatives have polynomial growth. Denote by $\mathcal{S}$ the set of random variables in $L^2(\Omega)$ with the form:

$$G = g(\mathbf{W}(\mathbf{h}_1), \ldots, \mathbf{W}(\mathbf{h}_m)),$$

where $\mathbf{h}_i \in \mathcal{H}$, $i = 1, \dots, m$ for $m \in \mathbb{N}$. Then the first order Malliavin derivative of G is the $\mathcal{H}$-valued random variable given by

$$\mathrm{D}_t G = \sum_{j=1}^{m} \partial_j g(\mathbf{W}(\mathbf{h}_1), \dots, \mathbf{W}(\mathbf{h}_m))\mathbf{h}_j(t), \quad 0 \leqslant t \leqslant T.$$

So $\mathrm{D}_t G \in L^2(\Omega, \mathcal{H})$. We can further define the second order Malliavin derivative of G as the following

$$\mathrm{D}_{t_1}\mathrm{D}_{t_2} G = \sum_{j_1=1}^{m}\sum_{j_2=1}^{m} \partial_{j_1}\partial_{j_2} g(\mathbf{W}(\mathbf{h}_1), \dots, \mathbf{W}(\mathbf{h}_m))\mathbf{h}_{j_1}(t_1)\mathbf{h}_{j_2}(t_2), \quad 0 \leqslant t_1, t_2 \leqslant T.$$

So $\mathrm{D}_{t_1}\mathrm{D}_{t_2} G \in L^2(\Omega, \mathcal{H} \otimes \mathcal{H})$. Inductively, we can define the k-th order Malliavin derivative $D_{t_1} ... D_{t_k} G$, which is located in $L^2(\Omega, \mathcal{H}^{\otimes k})$. Define the following norm

$$\|G\|_{k,p} := \left[\mathbb{E}\,|G|^p + \sum_{j=1}^{k} \mathbb{E}\left\|\mathrm{D}_{t_1}\cdots\mathrm{D}_{t_j} G\right\|^p_{\mathcal{H}^{\otimes j}}\right]^{1/p}.$$

Under this norm, the operator D can extended from $\mathcal{S}$ to its domain, denoted by $\mathbb{D}^{k,p}$. Specially, $\mathbb{D}^{1,2}$ is also a Hilbert space with product

$$\langle G, H\rangle_{1,2} = \mathbb{E}(GH) + \mathbb{E}\,\langle \mathrm{D}G, \mathrm{D}H\rangle_{\mathcal{H}}, \quad \forall\, G, H \in \mathbb{D}^{1,2}.$$

The following relation is called integration by parts in Malliavin calculus:

$$\mathbb{E}\,[G\delta(\mathbf{h})] = \mathbb{E}\,[\langle \mathrm{D}G, \mathbf{h}\rangle_{\mathcal{H}}], \quad \forall\, G \in \mathbb{D}^{1,2}, \quad \mathbf{h} \in \mathcal{H}, \tag{3.10}$$

where $\delta(\mathbf{h})$ is called Skorohod integral.

Given $\mathbf{h} \in \mathcal{H}$, we can define the Malliavin derivative along the direction $\mathbf{h}$, denoted by $\mathrm{D}^{\mathbf{h}} G$, as the following:

$$\mathrm{D}^{\mathbf{h}} G = \langle \mathrm{D}G, \mathbf{h}\rangle_{\mathcal{H}} = \int_0^T \langle \mathbf{h}(s), \mathrm{D}_s G\rangle\, \mathrm{d}s.$$

For $\mathbf{G} = (G^1, \dots, G^d)^\top$ with each $G^i \in \mathbb{D}^{1,2}$, we define $\mathrm{D}_t\mathbf{G} = (\mathrm{D}_t G^1, \dots, \mathrm{D}_t G^d)$. And, the norm

$$\|\mathrm{D}_{t_1}\cdots\mathrm{D}_{t_k}\mathbf{G}\|^2_{\mathcal{H}^{\otimes k}} = \sum_{i=1}^{d}\left\|\mathrm{D}_{t_1}\cdots\mathrm{D}_{t_k} G^i\right\|^2_{\mathcal{H}^{\otimes k}}, \quad \|\mathbf{G}\|^p_{k,p} = \sum_{i=1}^{d}\left\|G^i\right\|^p_{k,p}.$$

The associated Malliavin matrix of $\mathbf{G}$ is defined as the following random semi-definite symmetric matrix

$$\mathbf{\Gamma}(\mathbf{G}) = \left(\left\langle \mathrm{D}G^i, \mathrm{D}G^j\right\rangle_{\mathcal{H}}\right)_{1\leqslant i,j\leqslant d}.$$

The following abstract lemma will play an important role in the proof of Theorem 2. We leave its proof in Appendix B.3.

Lemma 3.8. *Let $\mathbf{F} = (F^1, \dots, F^d)$ be a random vector such that all of its components $F^i \in \mathbb{D}^{2,8}$, $i = 1, \dots, d$, and its Malliavin matrix $\mathbf{\Gamma}(\mathbf{F})$ is invertible a.s. with $\left(\det \mathbf{\Gamma}(\mathbf{F})\right)^{-1} \in \bigcap_{p\geqslant 1} L^p(\Omega)$. Let $\mathbf{G} = (G^1, \dots, G^d)$ be another random vector with $G^i \in \mathbb{D}^{1,4}$ for all $1 \leqslant i \leqslant d$, and $g : \mathbb{R}^d \to \mathbb{R}$ be a function in $\mathcal{C}^1_b(\mathbb{R}^d, \mathbb{R})$. Then there exists a positive number $C > 0$ such that*

$$\left|\mathbb{E}\,\langle \nabla g(\mathbf{F}), \mathbf{G}\rangle\right| \leqslant C\,\|g\|_\infty\,\|\mathbf{G}\|_{1,4}\left\{\mathbb{E}[\mathcal{K}_1\mathcal{K}_2\mathcal{K}_3]\right\}^{1/4}, \tag{3.11}$$

where

$$\mathcal{K}_1 = 1 + \|\mathrm{D}\mathbf{F}\|_{\mathcal{H}}^8, \qquad \mathcal{K}_2 = 1 + \left\|\mathrm{D}^2\mathbf{F}\right\|_{\mathcal{H}\otimes\mathcal{H}}^4, \qquad \mathcal{K}_3 = 1 + \left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{HS}}^8.$$

In order to apply Lemma 3.8, we need to introduce a series of events to estimate $\mathbb{E}[\mathcal{K}_1\mathcal{K}_2\mathcal{K}_3]$. To this end, let us first recall that for each step $i \in \mathbb{N}$ of SGDm (1.4), ξ_i^r, $r = 1, \ldots, N$ are independent copies of random variable ξ satisfying Assumption II, to make notations simple, we denote $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^N)$, $i \in \mathbb{N}$, and define

$$\varphi(\boldsymbol{\xi}_i) := \frac{1}{N}\sum_{r=1}^{N} \sup_{\mathbf{x}\in\mathbb{R}^d} \left\|\nabla^2 F(\mathbf{x}, \xi_i^r)\right\|_{\mathrm{op}}. \tag{3.12}$$

Define the following events: for $j \leqslant k$ and $\tau = 1, 2$,

$$E_{j,k}^{\tau} := \left\{ \sum_{i=j+1}^{k} \eta_i(\varphi(\boldsymbol{\xi}_i)^{\tau} - \mathbb{E}[\varphi(\boldsymbol{\xi}_i)^{\tau}]) > t_k - t_j \right\}. \tag{3.13}$$

Lemma 3.9. *Given Assumptions III and IV, let $t_k = \sum_{i=1}^k \eta_i$. There exist positive constants c and C such that*

$$\mathbb{P}(E_{j,k}^{\tau}) \leqslant \frac{C\mathrm{e}^{c(t_k - t_j)}}{(t_k - t_j)^2}\eta_k^2, \quad \tau = 1, 2.$$

For notations simplicity, we denote

$$\mathbf{Y}_t = (\mathbf{M}_t, \mathbf{X}_t), \quad \widetilde{\mathbf{Y}}_t = (\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t), \quad \forall\, t \geqslant 0, \tag{3.14}$$

where $(\mathbf{M}_t, \mathbf{X}_t)$ and $(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t)$ are defined by SDEs (1.5) and (1.6) respectively. For any $0 \leqslant s \leqslant t$, denote $\mathbf{Y}_{s,t}^{\mathbf{y}}$ or $\mathbf{Y}_{s,t}(\mathbf{y})$ as the value of the process $(\mathbf{Y}_t)_{t\geqslant 0}$ at the time t given $\mathbf{Y}_s = \mathbf{y} \in \mathbb{R}^{2d}$. Similarly for $\widetilde{\mathbf{Y}}_{s,t}^{\mathbf{y}}$. When $s = 0$, we will drop the subscript s if there is no ambiguity.

Lemma 3.10. *Let Assumptions I, II, III, and IV hold and recall $t_k = \sum_{i=1}^k \eta_i$. There exists a positive number $C > 0$ such that for any $\mathbf{z} \in \mathbb{R}^{2d}$ and any $\ell < k$ such that $t_k - t_\ell \leqslant 1/[5(B_1 + \gamma + 2)]$, the followings hold:*
(i) For all $t \in [t_{k-1}, t_k]$

$$\mathbb{E}\left\|\mathrm{D}\widetilde{\mathbf{Y}}_{t_\ell,t_k}^{\mathbf{z}}\right\|_{\mathcal{H}}^8 \leqslant Cd^4, \quad \mathbb{E}\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}_{t_\ell,t_{k-1}}^{\mathbf{z}})\right\|_{\mathcal{H}}^8 \leqslant Cd^4.$$

(ii) We have

$$\|\mathrm{D}\widetilde{\mathbf{Y}}_{t_\ell,t_k}^{\mathbf{z}}\|_{\mathcal{H}} \leqslant C\sqrt{d} \quad \text{on the event} \quad (E_{\ell,k}^1)^c.$$

Lemma 3.11. *Under the same setting as in Lemma 3.10, there exists a positive number $C > 0$ such that for any $\mathbf{z} \in \mathbb{R}^{2d}$ and any $\ell < k$ such that $t_k - t_\ell \leqslant 1/[5(B_1+\gamma+2)]$, there exists a positive number $C > 0$ such that the followings hold:*

$$\mathbb{E}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}_{t_\ell,t_k}^{\mathbf{z}}\right\|_{\mathcal{H}\otimes\mathcal{H}}^4 \leqslant Cd^4, \quad \mathbb{E}\left\|\mathrm{D}^2\mathbf{Y}_{t_{k-1},t_k}(\widetilde{\mathbf{Y}}_{t_\ell,t_{k-1}}^{\mathbf{z}})\right\|_{\mathcal{H}\otimes\mathcal{H}}^4 \leqslant Cd^4.$$

Lemma 3.12. *Under the same setting as in Lemma 3.10, there exists a positive number $C > 0$ such that for any $\mathbf{z} \in \mathbb{R}^{2d}$ and any $\ell < k$ such that $t_k - t_\ell \leqslant 1/[5(B_1+\gamma+2)]$, there exists a positive number $C > 0$ such that the followings hold:*
(i) We have

$$\mathbb{E}\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t_k}(\widetilde{\mathbf{Y}}_{t_\ell,t_{k-1}}^{\mathbf{z}}) - \mathrm{D}\widetilde{\mathbf{Y}}_{t_\ell,t_k}^{\mathbf{z}}\right\|_{\mathcal{H}}^4 \leqslant Cd^2\eta_k^6(\mathcal{V}(\mathbf{z})^2 + d^2) + C\frac{d^4\eta_k^4}{N^2}.$$

(ii) We have

$$\|\mathrm{D}\mathbf{Y}_{t_{k-1},t_k}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\|_{\mathcal{H}} \leqslant C\sqrt{d\eta_k} \quad \text{on the event} \quad (E^1_{\ell,k} \cup E^2_{\ell,k})^c.$$

Lemma 3.13. *Under the same setting as in Lemma 3.10, for any $\mathbf{z} \in \mathbb{R}^{2d}$ and any $\ell < k$ such that $t_k - t_\ell \leqslant 1/[5(B_1+\gamma+2)]$, there exists a positive number $C > 0$ such that*

$$\lambda_{\min}\left(\mathbf{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\big)\right) \geqslant C(t_k - t_\ell)^3 \quad \text{on the event} \quad (E^1_{\ell,k} \cup E^1_{m,k})^c,$$

where $m = \min\{j\colon t_k - t_j \leqslant (t_k - t_\ell)/[10(B_1+\gamma+2)]\}$ and $\lambda_{\min}\big(\mathbf{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\big)\big)$ is the smallest eigenvalue of the Malliavin matrix $\mathbf{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\big)$.

4. Proofs of the Main Results

In this section, we provide the proofs of Theorems 1, 2, and Corollaries 3, 5 by using the lemmas mentioned above. Employing the Lindeberg principle, we decompose $\mathbb{E}[g(\mathbf{Y}_{t_k})] - \mathbb{E}[g(\widetilde{\mathbf{Y}}_{t_k})]$ into k terms and subsequently analyze each term individually, where $\mathbf{Y}_t$ and $\widetilde{\mathbf{Y}}_t$ are defined by (3.14).

For any $0 \leqslant s \leqslant t$, recall the notations $\mathbf{Y}^{\mathbf{y}}_{s,t}$, $\mathbf{Y}_{s,t}(\mathbf{y})$ and $\widetilde{\mathbf{Y}}^{\mathbf{y}}_{s,t}$ immediately below (3.14).

4.1. Proof of Theorem 1.

Proof of Theorem 1. Applying the Lindeberg technique, we have the following decomposition for any Lipschitz function $h\colon \mathbb{R}^{2d} \to \mathbb{R}$,

$$\mathbb{E}h(\mathbf{Y}^{\mathbf{y}}_{t_n}) - \mathbb{E}h(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_n}) = \sum_{k=1}^{n} \mathbb{E}\left[h\left(\mathbf{Y}_{t_{k-1},t_n}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_{k-1}})\right) - h\left(\mathbf{Y}_{t_k,t_n}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_k})\right)\right] \tag{4.1}$$

with initial value $\mathbf{y} = (\mathbf{m}_0, \mathbf{x}_0)$.

Denote $\mathbf{z} = \mathbf{Y}_{t_{k-1},t_k}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_{k-1}})$ and $\tilde{\mathbf{z}} = \widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_k}$. We apply Lemma 3.6 on $[t_k, t_n]$ to get

$$\begin{aligned}
&\left|\mathbb{E}\left[h\left(\mathbf{Y}_{t_{k-1},t_n}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_{k-1}})\right) - h\left(\mathbf{Y}_{t_k,t_n}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_k})\right)\right]\right| \\
&= \left|\mathbb{E}\left[h\left(\mathbf{Y}_{t_k,t_n}(\mathbf{z})\right) - h\left(\mathbf{Y}_{t_k,t_n}(\tilde{\mathbf{z}})\right)\right]\right| \\
&= \left|\mathbb{E}\left[h\left(\mathbf{Y}_{t_k,t_n}(\mathbf{z}^*)\right) - h\left(\mathbf{Y}_{t_k,t_n}(\tilde{\mathbf{z}}^*)\right)\right]\right| \\
&\leqslant C\,\|h\|_{\mathrm{Lip}}\,\mathrm{e}^{-\theta(t_n-t_k)}\mathbb{E}|\mathbf{z}^* - \tilde{\mathbf{z}}^*|,
\end{aligned}$$

where $(\mathbf{z}^*, \tilde{\mathbf{z}}^*)$ is a random vector whose first and second marginal distributions are the same as $\mathbf{z}$ and $\tilde{\mathbf{z}}$ respectively. Because this inequality also holds for any $\mathbf{z}^*$ and $\tilde{\mathbf{z}}^*$ with this property, we can choose $\mathbf{z}^*$ and $\tilde{\mathbf{z}}^*$ which satisfies $\mathbb{E}|\mathbf{z}^* - \tilde{\mathbf{z}}^*| = d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{z}), \mathcal{L}(\tilde{\mathbf{z}}^*))$ (It is well known that the minimum of the coupling in the 1-Wasserstein can be realized). Hence,

$$\left|\mathbb{E}\left[h\left(\mathbf{Y}_{t_{k-1},t_n}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_{k-1}})\right) - h\left(\mathbf{Y}_{t_k,t_n}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_k})\right)\right]\right| \leqslant C\,\|h\|_{\mathrm{Lip}}\,\mathrm{e}^{-\theta(t_n-t_k)} d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{z}), \mathcal{L}(\tilde{\mathbf{z}})),$$

Applying Lemma 3.4, it holds

$$\mathbb{E}\left[\mathcal{V}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_{k-1}})\right] \leqslant \mathrm{e}^{-\lambda\gamma t_{k-1}}\mathcal{V}(\mathbf{y}) + Cd.$$

Furthermore, combining this with (3.8) in Lemma 3.5 immediately yields that

$$\begin{aligned} d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{z}),\mathcal{L}(\tilde{\mathbf{z}})) &\leqslant C\eta_k^{3/2}\left(1+\frac{1}{N}\right)\mathbb{E}\left[\sqrt{\mathcal{V}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_{k-1}})+d}\right] \\ &\leqslant C\eta_k^{3/2}\left(1+\frac{1}{N}\right)\sqrt{\mathcal{V}(\mathbf{y})+d}. \end{aligned}$$

Hence,

$$\begin{aligned} &\left|\mathbb{E}\left[h\left(\mathbf{Y}_{t_{k-1},t_n}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_{k-1}})\right)-h\left(\mathbf{Y}_{t_k,t_n}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_k})\right)\right]\right| \\ &\leqslant C\left\|h\right\|_{\mathrm{Lip}}\eta_k^{3/2}\mathrm{e}^{-\theta(t_n-t_k)}\left(1+\frac{1}{N}\right)\sqrt{\mathcal{V}(\mathbf{y})+d}. \end{aligned} \tag{4.2}$$

Substituting this formula in (4.1), we have

$$\left|\mathbb{E}h(\mathbf{Y}^{\mathbf{y}}_{t_n})-\mathbb{E}h(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_n})\right|\leqslant C\left\|h\right\|_{\mathrm{Lip}}\left(1+\frac{1}{N}\right)\sqrt{\mathcal{V}(\mathbf{y})+d}\sum_{k=1}^{n}\eta_k^{3/2}\mathrm{e}^{-\theta(t_n-t_k)}. \tag{4.3}$$

For the sum term, applying Lemma A.3 with $\varepsilon=1/2$ therein, we have

$$\sum_{k=1}^{n}\eta_k^{3/2}\mathrm{e}^{-\theta(t_n-t_k)}\leqslant\frac{4}{2\theta-\omega}\sqrt{\eta_n}.$$

Consequently, the desired result holds since $(\mathbf{m}_n,\mathbf{x}_n)_{n\geqslant0}$ and $(\widetilde{\mathbf{M}}_{t_n},\widetilde{\mathbf{X}}_{t_n})_{n\geqslant0}$ have the same distribution. □

4.2. **Proof of Theorem 2.** We denote the operator semigroups of $(\mathbf{Y}_t)_{t\geqslant0}$ and $(\widetilde{\mathbf{Y}}_t)_{t\geqslant0}$ by $\mathrm{P}_{s,t}$ and $\mathrm{Q}_{s,t}$ respectively, defined by

$$\mathrm{P}_{s,t}\,h(\mathbf{y})=\mathbb{E}\left[h(\mathbf{Y}_t)|\mathbf{Y}_s=\mathbf{y}\right],\quad \mathrm{Q}_{s,t}\,h(\mathbf{y})=\mathbb{E}[h(\widetilde{\mathbf{Y}}_t)|\widetilde{\mathbf{Y}}_s=\mathbf{y}],\quad 1\leqslant s<t,$$

for any $h\in\mathcal{C}_b(\mathbb{R}^{2d},\mathbb{R})$ and $\mathbf{y}\in\mathbb{R}^{2d}$. As $s=0$, we will drop the subscript s if no confusions arise.

Proof of Theorem 2. For any $h\in\mathcal{C}_b(\mathbb{R}^{2d},\mathbb{R})$, by the Lindeberg principle in the form of semigroup, we have the following decomposition

$$\mathbb{E}h(\mathbf{Y}^{\mathbf{y}}_{t_n})-\mathbb{E}h(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_n})=\left(\mathrm{P}_{0,t_n}-\mathrm{Q}_{0,t_n}\right)h(\mathbf{y})=\sum_{k=1}^{n}\mathrm{Q}_{0,t_{k-1}}\circ\left(\mathrm{P}_{t_{k-1},t_k}-\mathrm{Q}_{t_{k-1},t_k}\right)\circ\mathrm{P}_{t_k,t_n}\,h(\mathbf{y}).$$

Define index k_n as

$$k_n=\inf\{k\colon t_n-t_k<1\}.$$

Then, the above equation can be divided into the following two parts:

$$\begin{aligned} \mathcal{J}_1&=\sum_{k=1}^{k_n-1}\mathrm{Q}_{0,t_{k-1}}\circ\left(\mathrm{P}_{t_{k-1},t_k}-\mathrm{Q}_{t_{k-1},t_k}\right)\circ\mathrm{P}_{t_k,t_n}\,h(\mathbf{y}), \\ \mathcal{J}_2&=\sum_{k=k_n}^{n}\mathrm{Q}_{0,t_{k-1}}\circ\left(\mathrm{P}_{t_{k-1},t_k}-\mathrm{Q}_{t_{k-1},t_k}\right)\circ\mathrm{P}_{t_k,t_n}\,h(\mathbf{y}). \end{aligned} \tag{4.4}$$

For $\mathcal{J}_1$ and each $1\leqslant k<k_n$, we let $\mathrm{P}_{t_k,t_n}\,h(\mathbf{y})=\mathrm{P}_{t_k,t_{k_n}}\circ\mathrm{P}_{t_{k_n},t_n}\,h(\mathbf{y})$, and define function $g_1(\mathbf{y}):=\mathrm{P}_{t_{k_n},t_n}\,h(\mathbf{y})$. Then g_1 is a Lipschitz function according to Lemma 3.7 with $T=t_n-t_{k_n}\in[1/2,1]$ therein, and it holds

$$\left\|g_1\right\|_{\mathrm{Lip}}=\left|\nabla\,\mathrm{P}_{t_{k_n},t_n}\,h(\mathbf{y})\right|\leqslant C\left\|h\right\|_{\infty}.$$

Similar to obtaining (4.2) in the proof of Theorem 1, we have

$$\begin{aligned}&\left|\mathrm{Q}_{0,t_{k-1}}\circ\left(\mathrm{P}_{t_{k-1},t_k}-\mathrm{Q}_{t_{k-1},t_k}\right)\circ\mathrm{P}_{t_k,t_{k_n}}\,g(\mathbf{y})\right|\\&\leqslant C\,\|g_1\|_{\mathrm{Lip}}\,\eta_k^{3/2}\mathrm{e}^{-\theta(t_{k_n}-t_k)}\left(1+\frac{1}{N}\right)\sqrt{\mathcal{V}(\mathbf{y})+d}.\end{aligned}$$

Since $t_{k_n}>t_n-1$, we obtain the following estimate of $\mathcal{J}_1$:

$$|\mathcal{J}_1|\leqslant C\,\|h\|_\infty\sqrt{\mathcal{V}(\mathbf{y})+d}\left(1+\frac{1}{N}\right)\sum_{k=1}^{k_n-1}\eta_k^{3/2}\mathrm{e}^{-\theta(t_n-t_k)}. \tag{4.5}$$

For the term $\mathcal{J}_2$, we claim

$$|\mathcal{J}_2|\leqslant C\,\|h\|_\infty\,d^{7/2}\sqrt{\mathcal{V}(\mathbf{y})+d}\sum_{k=k_n}^{n}\left(\eta_k^{3/2}+\frac{\eta_k}{\sqrt{N}}\right)\mathrm{e}^{-\theta(t_n-t_k)}. \tag{4.6}$$

Combining (4.4), (4.5) and (4.6) implies

$$\left|\mathbb{E}h(\mathbf{Y}_{t_n}^{\mathbf{y}})-\mathbb{E}h(\widetilde{\mathbf{Y}}_{t_n}^{\mathbf{y}})\right|\leqslant C\,\|h\|_\infty\,d^{7/2}\sqrt{\mathcal{V}(\mathbf{y})+d}\sum_{k=1}^{n}\left(\eta_k^{3/2}+\frac{\eta_k}{\sqrt{N}}\right)\mathrm{e}^{-\theta(t_n-t_k)}. \tag{4.7}$$

Then, applying Lemma A.3 yields the desired result.

It remains to prove (4.6). For any fixed $k\in\{k_n,\dots,n\}$ in $\mathcal{J}_2$, we can find time t_ℓ such that

$$\frac{1}{6(B_1+\gamma+2)}\leqslant t_k-t_\ell\leqslant\frac{1}{5(B_1+\gamma+2)}. \tag{4.8}$$

Then we have the following decomposition,

$$\begin{aligned}&\mathrm{Q}_{0,t_{k-1}}\circ\left(\mathrm{P}_{t_{k-1},t_k}-\mathrm{Q}_{t_{k-1},t_k}\right)\circ\mathrm{P}_{t_k,t_n}\,h(\mathbf{y})\\&=\mathrm{Q}_{0,t_\ell}\circ\mathrm{Q}_{t_\ell,t_{k-1}}\circ\left(\mathrm{P}_{t_{k-1},t_k}-\mathrm{Q}_{t_{k-1},t_k}\right)g_2(\mathbf{y})\\&=\mathbb{E}\left\{\left[g_2\left(\mathbf{Y}_{t_{k-1},t_k}\left(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right)\right)-g_2\left(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right)\right]\left(\mathbf{1}_{E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k}}+\mathbf{1}_{(E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})^c}\right)\right\}\\&=\mathbb{E}\left\{\left[g_2\left(\mathbf{Y}_{t_{k-1},t_k}\left(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right)\right)-g_2\left(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right)\right]\mathbf{1}_{E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k}}\right\}\\&\quad+\mathbb{E}\left\{\mathbb{E}\left[g_2\left(\mathbf{Y}_{t_{k-1},t_k}\left(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right)\right)-g_2\left(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right)\middle|\mathscr{F}_{\ell,k}\right]\mathbf{1}_{(E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})^c}\right\}\\&=:I_{k,1}+I_{k,2},\end{aligned}$$

where $g_2(\mathbf{y})=\mathrm{P}_{t_k,t_n}\,h(\mathbf{y})$, $\mathbf{z}=\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_\ell}$, $\mathscr{F}_{\ell,k}=\sigma(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_\ell},(\boldsymbol{\xi}_i)_{\ell+1\leqslant i\leqslant k})$, and $E^1_{\ell,k}$, $E^1_{m,k}$, $E^2_{\ell,k}$ are defined in (3.13) with

$$m=\min\left\{j\colon t_k-t_j\leqslant\frac{t_k-t_\ell}{10(B_1+\gamma+2)}\right\}.$$

Let us estimate $I_{k,1}$ and $I_{k,2}$.

For $I_{k,1}$, since $C_1\leqslant t_k-t_m\leqslant t_k-t_\ell\leqslant C_2$ holds for some constants $C_1,C_2>0$, by Lemma 3.9, we have

$$\begin{aligned}|I_{k,1}|&\leqslant 2\,\|g_2\|_\infty\,\mathbb{P}(E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})\\&\leqslant 2\,\|h\|_\infty\,(\mathbb{P}(E^1_{\ell,k})+\mathbb{P}(E^1_{m,k})+\mathbb{P}(E^2_{\ell,k}))\leqslant C\,\|h\|_\infty\,\eta_k^2.\end{aligned} \tag{4.9}$$

For $I_{k,2}$, observe that

$$I_{k,2} = \int_0^1 \mathbb{E}\left\{\mathbb{E}\left[\left\langle \nabla g_2(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k} + r\boldsymbol{\Xi}_{t_\ell,t_k}), \boldsymbol{\Xi}_{t_\ell,t_k}\right\rangle \middle| \mathscr{F}_{\ell,k}\right] \mathbf{1}_{(E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})^c}\right\} \mathrm{d}r, \tag{4.10}$$

where $\boldsymbol{\Xi}_{t_\ell,t_k} = \mathbf{Y}_{t_{k-1},t_k}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}$. Denote $\mathbf{G} = \boldsymbol{\Xi}_{t_\ell,t_k}$ and $\mathbf{F} = \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k} + r\boldsymbol{\Xi}_{t_\ell,t_k}$, and we shall apply Lemma 3.8 to estimate $\mathbb{E}\left[\left\langle \nabla g_2(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k} + r\boldsymbol{\Xi}_{t_\ell,t_k}), \boldsymbol{\Xi}_{t_\ell,t_k}\right\rangle \middle| \mathscr{F}_{\ell,k}\right]$ on the event $(E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})^c$. To this end, let us consider SDE (1.6) on $[t_\ell, t_k]$. Note that $\mathscr{F}_{\ell,k}$ is independent of $(\mathbf{B}_t)_{t\in[t_\ell,t_k]}$ and that the Malliavin calculus in Lemma 3.8 is associated to $(\mathbf{B}_t)_{t\in[t_\ell,t_k]}$ and has nothing to do with $\mathscr{F}_{\ell,k}$. An advantage is that we can easily bound $\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3$ on the right hand of (3.11) on the event $(E^1_{\ell,k} \cup E^1_{m,k} \cup E^2_{\ell,k})^c$.

According to (3.9) in Lemma 3.5 and Lemma 3.12(i), $\|\mathbf{G}\|_{1,4}$ satisfies

$$\|\mathbf{G}\|_{1,4} = \left(\mathbb{E}\,|\mathbf{G}|^4 + \mathbb{E}\,\|\mathrm{D}\mathbf{G}\|^4_{\mathcal{H}}\right)^{1/4} \leqslant C\sqrt{d(\mathcal{V}(\mathbf{z})+d)}\left(\eta_k^{3/2} + \frac{\eta_k}{\sqrt{N}}\right).$$

Next we estimate $\mathbb{E}[\mathcal{K}_1\mathcal{K}_2\mathcal{K}_3]$. By Lemmas 3.10(ii) and 3.12(ii),

$$\mathcal{K}_1 = 1 + \|\mathrm{D}\mathbf{F}\|^8_{\mathcal{H}} \leqslant Cd^4 \quad \text{on the event} \quad (E^1_{\ell,k}\cup E^2_{\ell,k})^c.$$

By Lemma 3.11,

$$\mathbb{E}\mathcal{K}_2 = 1 + \mathbb{E}\|\mathrm{D}^2\mathbf{F}\|^4_{\mathcal{H}\otimes\mathcal{H}} \leqslant Cd^4.$$

Let us now bound $\mathcal{K}_3 = 1 + \|\boldsymbol{\Gamma}(\mathbf{F})^{-1}\|^8_{\mathrm{HS}}$. Since $\|\boldsymbol{\Gamma}(\mathbf{F})^{-1}\|_{\mathrm{HS}} \leqslant \sqrt{2d}\,[\lambda_{\min}(\boldsymbol{\Gamma}(\mathbf{F}))]^{-1}$, we first show

$$\lambda_{\min}(\boldsymbol{\Gamma}(\mathbf{F})) \geqslant C > 0 \quad \text{on the event} \quad (E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})^c.$$

Indeed, observe $\boldsymbol{\Gamma}(\mathbf{F}) = \boldsymbol{\Gamma}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}) + r^2\boldsymbol{\Gamma}(\boldsymbol{\Xi}_{t_\ell,t_k}) + r\mathbf{N}$, where the entries of matrix $\mathbf{N}$ are given by

$$N_{ij} = \left\langle \mathrm{D}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k})^i, \mathrm{D}(\boldsymbol{\Xi}_{t_\ell,t_k})^j\right\rangle_{\mathcal{H}} + \left\langle \mathrm{D}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k})^j, \mathrm{D}(\boldsymbol{\Xi}_{t_\ell,t_k})^i\right\rangle_{\mathcal{H}}, \quad i,j = 1,\dots,2d.$$

By Lemmas 3.10(ii) and 3.12(ii), on the event $(E^1_{\ell,k}\cup E^2_{\ell,k})^c$, we have $\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\|_{\mathcal{H}} \leqslant C\sqrt{d}$ and $\|\mathrm{D}\boldsymbol{\Xi}_{t_\ell,t_k}\|_{\mathcal{H}} \leqslant C\sqrt{d\eta_k}$. Thus

$$\begin{aligned}\|\mathbf{N}\|_{\mathrm{HS}} &\leqslant 2\left\{\sum_{i,j=1}^{2d}\left\|\mathrm{D}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k})^i\right\|^2_{\mathcal{H}}\left\|\mathrm{D}(\boldsymbol{\Xi}_{t_\ell,t_k})^j\right\|^2_{\mathcal{H}}\right\}^{1/2}\\ &\leqslant 2\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right\|_{\mathcal{H}}\|\mathrm{D}\boldsymbol{\Xi}_{t_\ell,t_k}\|_{\mathcal{H}} \leqslant Cd\sqrt{\eta_k}, \quad \text{on the event} \quad (E^1_{\ell,k}\cup E^2_{\ell,k})^c.\end{aligned}$$

This, together with Lemma 3.13 and (4.8), implies

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Gamma}(\mathbf{F})) &\geqslant \lambda_{\min}(\boldsymbol{\Gamma}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k})) + r^2\lambda_{\min}(\boldsymbol{\Gamma}(\boldsymbol{\Xi}_{t_\ell,t_k})) - r\,\|\mathbf{N}\|_{\mathrm{HS}}\\ &\geqslant C + 0 - Cd\sqrt{\eta_k}\\ &> C/2, \quad \text{on the event} \quad (E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})^c,\end{aligned}$$

for any $r \in [0,1]$ and η_k sufficiently small such that $\eta_k \leqslant cd^{-2}$ for some positive constant c. Then, we have

$$\mathcal{K}_3 = 1 + \left\|\boldsymbol{\Gamma}(\mathbf{F})^{-1}\right\|^8_{\mathrm{HS}} \leqslant Cd^4, \quad \text{on the event} \quad (E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})^c.$$

Hence, the following holds on event $(E^1_{\ell,k}\cup E^1_{m,k}\cup E^2_{\ell,k})^c$

(4.11)
$$\mathbb{E}\left[\left\langle \nabla g_2(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}+r\boldsymbol{\Xi}_{t_\ell,t_k}),\boldsymbol{\Xi}_{t_\ell,t_k}\right\rangle\middle|\mathscr{F}_{\ell,k}\right]\leqslant C\left\|g_2\right\|_\infty d^{7/2}\sqrt{\mathcal{V}(\mathbf{z})+d}\left(\eta_k^{3/2}+\frac{\eta_k}{\sqrt{N}}\right).$$

Recalling $\mathbf{z}=\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_\ell}$, Lemma 3.4 yields that

$$\mathbb{E}\left[\sqrt{\mathcal{V}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_\ell})+d}\right]\leqslant\sqrt{\mathbb{E}\left[\mathcal{V}(\widetilde{\mathbf{Y}}^{\mathbf{y}}_{t_\ell})\right]+d}\leqslant C\sqrt{\mathcal{V}(\mathbf{y})+d}.$$

This, together with (4.10) and (4.11), implies

$$|I_{k,2}|\leqslant C\left\|g_2\right\|_\infty d^{7/2}\sqrt{\mathcal{V}(\mathbf{y})+d}\left(\eta_k^{3/2}+\frac{\eta_k}{\sqrt{N}}\right). \tag{4.12}$$

It follows from (4.9), (4.12) and $\|g_2\|_\infty\leqslant\|h\|_\infty$ that

$$\begin{aligned}|\mathcal{J}_2|&\leqslant C\left\|h\right\|_\infty d^{7/2}\sqrt{\mathcal{V}(\mathbf{y})+d}\sum_{k=k_n}^n\left(\eta_k^{3/2}+\frac{\eta_k}{\sqrt{N}}\right)\\&\leqslant C\mathrm{e}^\theta\left\|h\right\|_\infty d^{7/2}\sqrt{\mathcal{V}(\mathbf{y})+d}\sum_{k=k_n}^n\left(\eta_k^{3/2}+\frac{\eta_k}{\sqrt{N}}\right)\mathrm{e}^{-\theta(t_n-t_k)},\end{aligned}$$

where in the second inequality we use $t_n-t_k<1$ such that $\mathrm{e}^{-\theta(t_n-t_k)}\geqslant\mathrm{e}^{-\theta}$ for each $k_n\leqslant k\leqslant n$. Hence, (4.6) is verified.

We complete the proof. □

4.3. **Proofs of Corollaries 3 and 4.**

Proof of Corollary 3 . Actually, by (4.3), we only need to analyze

$$\sum_{k=1}^n\eta_k^{3/2}\mathrm{e}^{-\theta(t_n-t_k)},\quad \eta_k=\frac{\eta}{k^\alpha},\ \alpha\in(0,1).$$

Observe that

$$\eta_{k-1}-\eta_k=\eta\cdot\frac{k^\alpha-(k-1)^\alpha}{k^\alpha(k-1)^\alpha}.$$

Notice that $k^\alpha-(k-1)^\alpha\to0$ and $\sqrt{\eta_k}\mathrm{e}^{\theta t_k}\to+\infty$ as $k\to\infty$, so there exists $n_0\in\mathbb{N}$ such that for all $k\geqslant n_0$

$$\eta_k\leqslant\frac{2\theta-\omega}{2\theta^2},\quad \eta_{k-1}-\eta_k\leqslant\omega\eta_k^2\quad\text{and}\quad\sqrt{\eta_k}\mathrm{e}^{\theta t_k}\geqslant1.$$

It is obvious that the desired result holds for $n\leqslant n_0$. And for $n>n_0$, we split the sum term in (4.3) into two parts:

$$\begin{aligned}\sum_{k=1}^{n_0}\eta_k^{3/2}\mathrm{e}^{-\theta(t_n-t_k)}&\leqslant C\mathrm{e}^{-\theta t_n}\leqslant C\sqrt{\eta_n}=C\frac{\eta^{1/2}}{n^{\alpha/2}},\\\sum_{k=n_0+1}^{n}\eta_k^{3/2}\mathrm{e}^{-\theta(t_n-t_k)}&\leqslant C\sqrt{\eta_n}=C\frac{\eta^{1/2}}{n^{\alpha/2}},\end{aligned} \tag{4.13}$$

where we use Lemma A.3 to estimate the second part. Substituting (4.13) into (4.3) implies the desired result. □

Proof of Corollary 4. By (4.7), the proof is similar to the proof of Corollary 3. We only need to analyze the following sum additionally,

$$\sum_{k=1}^{n}\eta_k \mathrm{e}^{-\theta(t_n-t_k)}, \quad \eta_k=\frac{\eta}{k^{\alpha}},\ \alpha\in(0,1).$$

Using the same notations in the proof of Corollary 3, we have

$$\sum_{k=1}^{n}\eta_k \mathrm{e}^{-\theta(t_n-t_k)}=\sum_{k=1}^{n_0}\eta_k \mathrm{e}^{-\theta(t_n-t_k)}+\sum_{k=n_0+1}^{n}\eta_k \mathrm{e}^{-\theta(t_n-t_k)}\leqslant C. \tag{4.14}$$

Substituting (4.13) and (4.14) into (4.7) implies the desired result. □

4.4. **Proof of Corollary 5.**

Proof of Corollary 5. Recall $f(\mathbf{x})=\mathbb{E}F(\mathbf{x},\xi)$ for any $\mathbf{x}\in\mathbb{R}^d$. As ξ is independent of $(\widetilde{\mathbf{X}}_{t_n})_{n\geqslant 1}$, we have the following decomposition:

$$\begin{aligned}&\mathbb{E}F(\widetilde{\mathbf{X}}_{t_n},\xi)-\min_{\mathbf{x}\in\mathbb{R}^d}\mathbb{E}F(\mathbf{x},\xi)\\ =&[\mathbb{E}f(\widetilde{\mathbf{X}}_{t_n})-\mathbb{E}f(\mathbf{X}_{t_n})]+[\mathbb{E}f(\mathbf{X}_{t_n})-\mathbb{E}_{\boldsymbol{\pi}}f(\mathbf{X})]+[\mathbb{E}_{\boldsymbol{\pi}}f(\mathbf{X})-f(\mathbf{x}^*)],\end{aligned}$$

where $\mathbf{x}^*=\arg\min f(\mathbf{x})$. According to the conclusion of exponential ergodicity [MT93], the second term on the right-hand side can be estimated by

$$|\mathbb{E}f(\mathbf{X}_t)-\mathbb{E}_{\boldsymbol{\pi}}f(\mathbf{X})|\leqslant B_e(\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+1)\mathrm{e}^{-\kappa t},$$

for some positive constants B_e and κ. As for the last term, Proposition 11 in [RRT17] yields that

$$\int_{\mathbb{R}^d}f(\boldsymbol{\omega})\boldsymbol{\pi}_{\mathbf{x}}(\mathrm{d}\boldsymbol{\omega})-\min_{\mathbf{x}\in\mathbb{R}^d}f(\mathbf{x})\leqslant\frac{d\beta^2}{4}\log\left[\frac{2\mathrm{e}L}{a}\left(\frac{2K}{d\beta^2}+1\right)\right]$$

as $\beta^2\leqslant a/2$, where $\boldsymbol{\pi}_{\mathbf{x}}$ is the marginal distribution of $\boldsymbol{\pi}$. So it remains to estimate the first term by Theorem 1.

Pick a localization function $\psi\in C^{\infty}(\mathbb{R}^d,\mathbb{R})$ satisfying $0\leqslant\psi(\mathbf{x})\leqslant 1$ and

$$\psi(\mathbf{x})=\begin{cases}1, & |\mathbf{x}|\leqslant 1,\\ 0, & |\mathbf{x}|\geqslant 2.\end{cases}$$

For $n\geqslant 1$, denote

$$g_n(\mathbf{x})=f(\mathbf{x})\psi\left(\frac{\mathbf{x}}{R_n}\right),$$

with some constants $R_n>0$ determined later, then we have

$$\begin{aligned}|\nabla g_n(\mathbf{x})|&=\left|\psi\left(\frac{\mathbf{x}}{R_n}\right)\nabla f(\mathbf{x})+\frac{1}{R_n}f(\mathbf{x})\nabla\psi\left(\frac{\mathbf{x}}{R_n}\right)\right|\\ &\leqslant\left(|\nabla f(\mathbf{x})|+\frac{\|\nabla\psi\|_{\infty}}{R_n}|f(\mathbf{x})|\right)\mathbf{1}_{[0,2R_n)}(|\mathbf{x}|).\end{aligned}$$

Recall that $|f(\mathbf{x})|\leqslant C|\mathbf{x}|^2$ and $|\nabla f(\mathbf{x})|\leqslant C|\mathbf{x}|$ for all $\mathbf{x}\in\mathbb{R}^d$ satisfying $|\mathbf{x}|\geqslant 1$, so $|\nabla g_n(\mathbf{x})|\leqslant CR_n$ for any $\mathbf{x}\in\mathbb{R}^d$, i.e., $\|g_n\|_{\mathrm{Lip}}\leqslant CR_n$. It follows from Lemma 3.4 that

$$\left|\mathbb{E}f(\widetilde{\mathbf{X}}_{t_n})-\mathbb{E}g_n(\widetilde{\mathbf{X}}_{t_n})\right|\leqslant\mathbb{E}\left|f(\widetilde{\mathbf{X}}_{t_n})\mathbf{1}_{(R_n,+\infty)}(|\widetilde{\mathbf{X}}_{t_n}|)\right|$$

$$\begin{aligned}
&\leqslant C\mathbb{E}\left[\left|\widetilde{\mathbf{X}}_{t_n}\right|^2 \mathbf{1}_{(R_n,+\infty)}\left(\left|\widetilde{\mathbf{X}}_{t_n}\right|\right)\right] \\
&\leqslant CR_n^{2-2p}\mathbb{E}\left[\left|\widetilde{\mathbf{X}}_{t_n}\right|^{2p} \mathbf{1}_{(R_n,+\infty)}\left(\left|\widetilde{\mathbf{X}}_{t_n}\right|\right)\right] \\
&\leqslant CR_n^{2-2p}\left(\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)^p + d^p\right).
\end{aligned}$$

Similarly, the following holds by Lemma 3.1

$$|\mathbb{E}f(\mathbf{X}_{t_n}) - \mathbb{E}g_n(\mathbf{X}_{t_n})| \leqslant CR_n^{2-2p}\left(\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)^p + d^p\right).$$

Thus, by Theorem 1 we have

$$\begin{aligned}
&\left|\mathbb{E}f(\widetilde{\mathbf{X}}_{t_n}) - \mathbb{E}f(\mathbf{X}_{t_n})\right| \\
\leqslant &\left|\mathbb{E}f(\widetilde{\mathbf{X}}_{t_n}) - \mathbb{E}g_n(\widetilde{\mathbf{X}}_{t_n})\right| + \left|\mathbb{E}g_n(\widetilde{\mathbf{X}}_{t_n}) - \mathbb{E}g_n(\mathbf{X}_{t_n})\right| + |\mathbb{E}g_n(\mathbf{X}_{t_n}) - \mathbb{E}f(\mathbf{X}_{t_n})| \\
\leqslant &CR_n^{2-2p}\left(\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)^p + d^p\right) + C\left\|g_n\right\|_{\mathrm{Lip}}\left(1+\frac{1}{N}\right)\sqrt{\eta_n(\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+d)} \\
\leqslant &CR_n^{2-2p}\left(\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)^p + d^p\right) + CR_n\left(1+\frac{1}{N}\right)\sqrt{\eta_n(\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+d)}.
\end{aligned}$$

Since the above equation holds with arbitrary $R_n > 0$, we can take

$$R_n = \eta_n^{1/(2-4p)}\sqrt{\mathcal{V}(\mathbf{m}_0,\mathbf{x}_0)+d}.$$

Let $2p = q$, we obtain the desired. □

References

[AAMN24] Dallas Albritton, Scott Armstrong, Jean-Christophe Mourrat, and Matthew Novack. Variational methods for the kinetic Fokker-Planck equation. *Anal. PDE*, 17(6):1953–2010, 2024.

[BGM10] François Bolley, Arnaud Guillin, and Florent Malrieu. Trend to equilibrium and particle approximation for a weakly selfconsistent Vlasov-Fokker-Planck equation. *M2AN Math. Model. Numer. Anal.*, 44(5):867–884, 2010.

[BLB17] Aleksandar Botev, Guy Lever, and David Barber. Nesterov's accelerated gradient and momentum as approximations to regularised update descent. In *2017 International joint conference on neural networks (IJCNN)*, pages 1899–1903. IEEE, 2017.

[CCAY+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, 2018.

[CCBJ18] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In Sbastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018.

[CDC15] Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. *Advances in neural information processing systems*, 28, 2015.

[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.

[CL20] Jerry Chee and Ping Li. Understanding and detecting convergence for stochastic gradient descent with momentum. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 133–140. IEEE, 2020.

[CLTZ20] Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251–273, 2020.

[CLW23] Yu Cao, Jianfeng Lu, and Lihan Wang. On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. *Arch. Ration. Mech. Anal.*, 247(5):Paper No. 90, 34, 2023.

[CR22] Huy N Chau and Miklós Rásonyi. Stochastic gradient Hamiltonian Monte Carlo for non-convex learning. *Stochastic Processes and their Applications*, 149:341–368, 2022.

[CSX23] Peng Chen, Qi-Man Shao, and Lihu Xu. A probability approximation framework: Markov process approach. *The Annals of Applied Probability*, 33(2):1419–1459, 2023.

[DP14] Zhao Dong and Xuhui Peng. Malliavin matrix of degenerate SDE and gradient estimate. *Electron. J. Probab.*, 19:no. 73, 26, 2014.

[Ebe11] Andreas Eberle. Reflection coupling and Wasserstein contractivity without convexity. *Comptes Rendus Mathematique*, 349(19-20):1101–1104, 2011.

[Ebe16] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, 166(3-4):851–886, 2016.

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

[FDBD21] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for SGD and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

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

[GM91] Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

[GM16] Arnaud Guillin and Pierre Monmarché. Optimal linear drift for the speed of convergence of an hypoelliptic diffusion. *Electron. Commun. Probab.*, 21:Paper No. 74, 14, 2016.

[KF16] Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159:81–107, 2016.

[KLRT15] Jakub Konečnỳ, Jie Liu, Peter Richtárik, and Martin Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2015.

[LR20] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

[LTW17] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

[LTW19] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.

[LZCS14] Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 661–670, 2014.

[MJ19] Michael Muehlebach and Michael Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR, 2019.

[MT93] Sean P. Meyn and R. L. Tweedie. Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.*, 25(3):518–548, 1993.

[Nea11] Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 113–162. CRC Press, Boca Raton, FL, 2011.

[Nor86] James Norris. Simplified Malliavin calculus. In *Séminaire de Probabilités, XX, 1984/85*, volume 1204 of *Lecture Notes in Math.*, pages 101–130. Springer, Berlin, 1986.

[Nua06] David Nualart. *The Malliavin calculus and related topics*. Probability and its Applications (New York). Springer-Verlag, Berlin, second edition, 2006.

[Pav14] Grigorios A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

[PP23] Gilles Pages and Fabien Panloup. Unadjusted Langevin algorithm with multiplicative noise: Total variation and Wasserstein bounds. *The Annals of Applied Probability*, 33(1):726–779, 2023.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703. PMLR, 07–10 Jul 2017.

[Sch24] Katharina Schuh. Global contractivity for Langevin dynamics with distribution-dependent forces and uniform in time propagation of chaos. *Ann. Inst. Henri Poincaré Probab. Stat.*, 60(2):753–789, 2024.

[SW24] Katharina Schuh and Peter A. Whalley. Convergence of kinetic langevin samplers for non-convex potentials, 2024.

[Wu01] Liming Wu. Large and moderate deviations and exponential convergence for stochastic damping Hamiltonian systems. *Stochastic processes and their applications*, 91(2):205–238, 2001.

[Zha10] Xicheng Zhang. Stochastic flows and Bismut formulas for stochastic Hamiltonian systems. *Stochastic Process. Appl.*, 120(10):1929–1949, 2010.

Appendix

Appendix A. Supporting Lemmas

The first lemma shows that $f(\mathbf{x})$ admits lower and upper bounds that are quadratic functions.

Lemma A.1. [RRT17, Lemma 2] *If function $f(\mathbf{x})$ satisfies Assumption 2.1, then the following quadratic bounds hold,*

$$\frac{a}{2}\left|\mathbf{x}\right|^2 - \frac{K}{2}\log 3 \leqslant f(\mathbf{x}) \leqslant L\left|\mathbf{x}\right|^2 + \frac{B^2}{2L} + A, \quad \forall\, \mathbf{x} \in \mathbb{R}^d,$$

where $K = b + B^2/(2a)$.

Applying Assumption I (iii), it is easy to verify that

$$\langle \mathbf{x}, \nabla f(\mathbf{x})\rangle \geqslant \frac{a}{2}\left|\mathbf{x}\right|^2 - K \geqslant 2\lambda\left(f(\mathbf{x}) + \frac{1}{4}\gamma^2\left|\mathbf{x}\right|^2\right) - \mathring{A}\,, \quad \forall\, \mathbf{x} \in \mathbb{R}^d, \tag{A.1}$$

where the constants λ and $\mathring{A}$ satisfy

$$0 < \lambda \leqslant \min\left\{\frac{1}{4}, \frac{a}{4L+\gamma^2}\right\}, \quad \mathring{A} \geqslant K + 2\lambda\left(\frac{B^2}{2L} + A\right). \tag{A.2}$$

Applying (A.1), the next lemma shows that the Lyapunov function $\mathcal{V}(\mathbf{m}, \mathbf{x})$ defined in (2.6) satisfies (3.3).

Lemma A.2. *Under Assumption 2.1, and λ and $\mathring{A}$ are in (A.2), we have that*

$$\mathscr{A}\mathcal{V}(\mathbf{m}, \mathbf{x}) \leqslant -\lambda\gamma\mathcal{V}(\mathbf{m}, \mathbf{x}) + (\gamma\mathring{A} + d\beta^2)/2.$$

Proof. Recall that

$$\mathcal{V}(\mathbf{m}, \mathbf{x}) := f(\mathbf{x}) + \frac{\gamma^2}{4}\left(\left|\mathbf{x} + \frac{1}{\gamma}\mathbf{m}\right|^2 + \left|\frac{1}{\gamma}\mathbf{m}\right|^2 - \lambda\left|\mathbf{x}\right|^2\right),$$

and we have

$$\nabla_{\mathbf{m}}\mathcal{V}(\mathbf{m}, \mathbf{x}) = \mathbf{m} + \frac{\gamma}{2}\mathbf{x}, \quad \Delta_{\mathbf{m}}\mathcal{V}(\mathbf{m}, \mathbf{x}) = d,$$
$$\nabla_{\mathbf{x}}\mathcal{V}(\mathbf{m}, \mathbf{x}) = \nabla f(\mathbf{x}) + \frac{\gamma^2(1-\lambda)}{2}\mathbf{x} + \frac{\gamma}{2}\mathbf{m}.$$

Hence, by inequality (A.1), we have

$$\begin{aligned}
\mathscr{A}\mathcal{V}(\mathbf{m}, \mathbf{x}) &= -\left\langle \nabla_{\mathbf{m}}\mathcal{V}(\mathbf{m}, \mathbf{x}), \gamma\mathbf{m} + \nabla f(\mathbf{x})\right\rangle + \left\langle \nabla_{\mathbf{x}}\mathcal{V}(\mathbf{m}, \mathbf{x}), \mathbf{m}\right\rangle + \frac{1}{2}\beta^2\Delta_{\mathbf{m}}\mathcal{V}(\mathbf{m}, \mathbf{x}) \\
&\leqslant -\left[\lambda\gamma\big(f(\mathbf{x}) + \frac{\gamma^2}{4}\left|\mathbf{x}\right|^2\big)\right] + \frac{\gamma}{2}\mathring{A} - \frac{\lambda\gamma^2}{2}\left\langle \mathbf{m}, \mathbf{x}\right\rangle - \frac{\gamma}{2}\left|\mathbf{m}\right|^2 + \frac{1}{2}d\beta^2 \\
&= -\lambda\gamma\mathcal{V}(\mathbf{m}, \mathbf{x}) - \frac{\lambda^2\gamma^3}{4}\left|\mathbf{x}\right|^2 - \frac{(1-\lambda)\gamma}{2}\left|\mathbf{m}\right|^2 + \frac{1}{2}\left(\gamma\mathring{A} + d\beta^2\right),
\end{aligned}$$

where the inequality comes from (A.1). Since $\lambda \leqslant 1/4$, we complete the proof. □

The process $(\mathbf{M}_t, \mathbf{X}_t)_{t\geqslant 0}$ admits a unique stationary distribution as

$$\boldsymbol{\pi}(\mathrm{d}\mathbf{m}, \mathrm{d}\mathbf{x}) = \frac{1}{Z}\exp\left\{-\frac{\gamma}{\beta^2}\left(\left|\mathbf{m}\right|^2 + 2f(\mathbf{x})\right)\right\}\mathrm{d}\mathbf{m}\,\mathrm{d}\mathbf{x},$$

where Z is a normalized constant given by

$$Z = \left(\frac{\beta^2\pi}{\gamma}\right)^{d/2} \int_{\mathbb{R}^d} \mathrm{e}^{-2\gamma f(\mathbf{x})/\beta^2}\,\mathrm{d}\mathbf{x},$$

see e.g. [Pav14]. Thanks to [MT93], Lemma A.2 yields that process $(\mathbf{M}_t, \mathbf{X}_t)_{t\geqslant 0}$ is exponentially ergodic. That is, there exist some positive constants B_e and κ, such that for all $\mathbf{m}, \mathbf{x} \in \mathbb{R}^d$,

$$\sup_{g\leqslant \mathcal{V}+1} \left|\mathbb{E}\left[g(\mathbf{M}_t^{\mathbf{m}}, \mathbf{X}_t^{\mathbf{x}})\right] - \mathbb{E}_{\boldsymbol{\pi}}\left[g(\mathbf{M}, \mathbf{X})\right]\right| \leqslant B_e(\mathcal{V}(\mathbf{m}, \mathbf{x}) + 1)\mathrm{e}^{-\kappa t}, \quad \forall\, t \geqslant 0,$$

where $(\mathbf{M}, \mathbf{X})$ has the distribution $\boldsymbol{\pi}$. By the definition of $\mathcal{V}(\mathbf{m}, \mathbf{x})$, this immediately implies that for all $\mathbf{x} \in \mathbb{R}^d$,

$$|\mathbb{E}f(\mathbf{X}_t^{\mathbf{x}}) - \mathbb{E}_{\boldsymbol{\pi}}[f(\mathbf{X})]| \leqslant B_e(\mathcal{V}(\mathbf{m}, \mathbf{x}) + 1)\mathrm{e}^{-\kappa t}, \quad \forall t \geqslant 0. \tag{A.3}$$

Lemma A.3. *Let step size $(\eta_k)_{k\geqslant 1}$ satisfy Assumption III and $t_n = \sum_{i=1}^n \eta_i$. For any $\varepsilon \in [0, 1/2]$, the following holds*

$$\sum_{k=1}^n \eta_k^{1+\varepsilon}\mathrm{e}^{-\theta(t_n - t_k)} \leqslant \frac{4}{2\theta - (4\varepsilon - 1)\omega}\eta_n^{\varepsilon}.$$

Proof. Notice that for each $k \leqslant n$

$$\eta_k^{1+\varepsilon}\mathrm{e}^{-\theta(t_n - t_k)} = \mathrm{e}^{-\theta t_n} \cdot \eta_k^{1+\varepsilon}\frac{\mathrm{e}^{\theta t_k} - \mathrm{e}^{\theta t_{k-1}}}{1 - \mathrm{e}^{-\theta\eta_k}} \leqslant \frac{4\mathrm{e}^{-\theta t_n}}{2\theta + \omega} \cdot \eta_k^{\varepsilon}\left(\mathrm{e}^{\theta t_k} - \mathrm{e}^{\theta t_{k-1}}\right), \tag{A.4}$$

where in the last inequality, we use the fact

$$\mathrm{e}^{-x} \leqslant 1 - x + \frac{x^2}{2} \leqslant 1 - x + \frac{2\theta - \omega}{4\theta}x = 1 - \frac{2\theta + \omega}{4\theta}x, \quad \forall\, 0 \leqslant x \leqslant 1 - \frac{\omega}{2\theta},$$

to obtain that

$$1 - \mathrm{e}^{-\theta\eta_k} \geqslant \frac{2\theta + \omega}{4}\eta_k, \quad \forall\, \eta_k \leqslant \frac{2\theta - \omega}{2\theta^2}.$$

Besides, according to condition $\eta_{k-1} - \eta_k \leqslant \omega\eta_k^2$ and Bernoulli's inequality, it holds that

$$\eta_{k-1}^{\varepsilon} - \eta_k^{\varepsilon} = \eta_k^{\varepsilon}\left[\left[\frac{\eta_{k-1}}{\eta_k}\right]^{\varepsilon} - 1\right] \leqslant \varepsilon\eta_k^{\varepsilon}\left[\frac{\eta_{k-1}}{\eta_k} - 1\right] \leqslant \varepsilon\omega\eta_k^{1+\varepsilon},$$

which derives that

$$\begin{aligned}\sum_{k=1}^n \eta_k^{\varepsilon}\left(\mathrm{e}^{\theta t_k} - \mathrm{e}^{\theta t_{k-1}}\right) &= \sum_{k=1}^n \left(\eta_k^{\varepsilon}\mathrm{e}^{\theta t_k} - \eta_{k-1}^{\varepsilon}\mathrm{e}^{\theta t_{k-1}}\right) + \sum_{k=1}^n \left(\eta_{k-1}^{\varepsilon} - \eta_k^{\varepsilon}\right)\mathrm{e}^{\theta t_{k-1}} \\ &\leqslant \eta_n^{\varepsilon}\mathrm{e}^{\theta t_n} + \varepsilon\omega\sum_{k=1}^n \eta_k^{1+\varepsilon}\mathrm{e}^{\theta t_k}.\end{aligned} \tag{A.5}$$

Combining (A.4) and (A.5) shows that

$$\sum_{k=1}^n \eta_k^{1+\varepsilon}\mathrm{e}^{-\theta(t_n - t_k)} \leqslant \frac{4}{2\theta + \omega}\eta_n^{\varepsilon} + \frac{4\varepsilon\omega}{2\theta + \omega}\sum_{k=1}^n \eta_k^{1+\varepsilon}\mathrm{e}^{-\theta(t_n - t_k)},$$

and a simply calculation yields the desired. □

Appendix B. Proofs of Auxiliary Lemmas

Lemma B.1. *Let $\mathbf{U}^i$, $i = 1, \dots, N$ be i.i.d. random vectors in $\mathbb{R}^d$ satisfying $\mathbb{E}\mathbf{U}^1 = \mathbf{0}$ and $\mathbb{E}[|\mathbf{U}^1|^{2p}] < +\infty$ for some $p \in \mathbb{N}$. Then there exists a constant $C > 0$ only depending on p, such that*

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{U}^i\right|^{2p}\right] \leqslant \frac{C}{N^p}\mathbb{E}\left[\left|\mathbf{U}^1\right|^{2p}\right].$$

Proof. Notice that

$$\left|\sum_{i=1}^{N}\mathbf{U}^i\right|^{2p} = \prod_{j=1}^{p}\left\langle\sum_{i_{2j-1}=1}^{N}\mathbf{U}^{i_{2j-1}}, \sum_{i_{2j}=1}^{N}\mathbf{U}^{i_{2j}}\right\rangle = \sum_{i_1,\dots,i_{2p}=1}^{N}\prod_{j=1}^{p}\left\langle\mathbf{U}^{i_{2j-1}}, \mathbf{U}^{i_{2j}}\right\rangle,$$

which implies that

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{U}^i\right|^{2p}\right] = \frac{1}{N^{2p}}\sum_{i_1,\dots,i_{2p}=1}^{N}\mathbb{E}\left[\prod_{j=1}^{p}\left\langle\mathbf{U}^{i_{2j-1}}, \mathbf{U}^{i_{2j}}\right\rangle\right].$$

Since $\mathbf{U}^1, \dots, \mathbf{U}^N$ are independent and $\mathbb{E}\mathbf{U}^1 = \mathbf{0}$, we have $\mathbb{E}[\prod_{j=1}^{p}\langle\mathbf{U}^{i_{2j-1}}, \mathbf{U}^{i_{2j}}\rangle] = 0$ whenever some $i \in \{1, \dots, N\}$ appears only once among $i_1, \dots, i_{2p}$. For the rest terms, Hölder's inequality shows that

$$\mathbb{E}\left[\prod_{j=1}^{p}\left\langle\mathbf{U}^{i_{2j-1}}, \mathbf{U}^{i_{2j}}\right\rangle\right] \leqslant \mathbb{E}\left[\prod_{j=1}^{2p}\left|\mathbf{U}^{i_j}\right|\right] \leqslant \prod_{j=1}^{2p}\left\{\mathbb{E}\left[\left|\mathbf{U}^{i_j}\right|^{2p}\right]\right\}^{1/2p} = \mathbb{E}\left[\left|\mathbf{U}^1\right|^{2p}\right].$$

Furthermore, for $l = 1, \dots, p$, the index set $\{1, \dots, 2p\}$ can be decomposed into l parts such that each part consists of at least two indices, and c_l denotes the number of such partitions. Then we have

$$\sum_{i_1,\dots,i_{2p}=1}^{N}\mathbb{E}\left[\prod_{j=1}^{p}\left\langle\mathbf{U}^{i_{2j-1}}, \mathbf{U}^{i_{2j}}\right\rangle\right] \leqslant \sum_{l=1}^{p} c_l N^l \mathbb{E}\left[\left|\mathbf{U}^1\right|^{2p}\right].$$

The desired result follows from

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{U}^i\right|^{2p}\right] \leqslant \frac{1}{N^{2p}}\sum_{l=1}^{p} c_l N^l \mathbb{E}\left[\left|\mathbf{U}^1\right|^{2p}\right] \leqslant \frac{C}{N^p}\mathbb{E}\left[\left|\mathbf{U}^1\right|^{2p}\right].$$

□

B.1. **Proofs of Lemmas for Moments Estimations.**

Proof of Lemma 3.1. Recalling the Lyapunov function $\mathcal{V}(\mathbf{m}, \mathbf{x})$ in (2.6), we define function F_t as

$$F_t := \mathbb{E}\mathcal{V}(\mathbf{M}_t^{\mathbf{m}}, \mathbf{X}_t^{\mathbf{x}}) = \mathbb{E}\left[f(\mathbf{X}_t^{\mathbf{x}}) + \frac{\gamma^2}{4}\left(\left|\mathbf{X}_t^{\mathbf{x}} + \frac{1}{\gamma}\mathbf{M}_t^{\mathbf{m}}\right|^2 + \left|\frac{1}{\gamma}\mathbf{M}_t^{\mathbf{m}}\right|^2 - \lambda\left|\mathbf{X}_t^{\mathbf{x}}\right|^2\right)\right].$$

Itô's formula implies that F_t satisfies the following differential inequality

$$\frac{\mathrm{d}F_t}{\mathrm{d}t} \leqslant -\lambda\gamma\, F_t + \frac{1}{2}\left(\gamma\mathring{A} + d\beta^2\right),$$

with initial condition $F_0 = \mathcal{V}(\mathbf{m}, \mathbf{x})$. Hence, we can obtain that

$$F_t \leqslant \mathrm{e}^{-\lambda\gamma t} F_0 + \frac{1}{2\lambda\gamma}\left(\gamma \mathring{A} + d\beta^2\right), \tag{B.1}$$

which yields the result with respect to of $p = 1$ immediately. As for the case of $p > 1$, we can apply similar argument. By Itô's formula, it holds

$$\begin{aligned}\mathscr{A}\left(\mathcal{V}(\mathbf{m},\mathbf{x})^p\right) &\leqslant p\mathcal{V}(\mathbf{m},\mathbf{x})^{p-1}\mathscr{A}\mathcal{V}(\mathbf{m},\mathbf{x}) + \frac{p(p-1)\beta^2}{2}\mathcal{V}(\mathbf{m},\mathbf{x})^{p-2}\left|\mathbf{m}+\frac{\gamma}{2}\mathbf{x}\right|^2 \\ &\leqslant -p\lambda\gamma\mathcal{V}(\mathbf{m},\mathbf{x})^p + C\mathcal{V}(\mathbf{m},\mathbf{x})^{p-1},\end{aligned}$$

where the last inequality is because of (3.2) and (3.3). Then, by Young's inequality, the function $G_t := \mathbb{E}\left[\mathcal{V}(\mathbf{M}_t^{\mathbf{m}}, \mathbf{X}_t^{\mathbf{x}})^p\right]$ satisfies

$$\frac{\mathrm{d}G_t}{\mathrm{d}t} \leqslant -\lambda\gamma G_t + C,$$

with initial value $G_0 = \mathcal{V}(\mathbf{m},\mathbf{x})^p$, where $C > 0$ here also depends on p. This implies the desired immediately. □

Proof of Lemma 3.2 . Under Assumption I, we have

$$\begin{aligned}\mathbb{E}\left|\mathbf{M}_t^{\mathbf{m}} - \mathbf{m}\right|^{2p} &= \mathbb{E}\left|\int_0^t \left[-\gamma\mathbf{M}_s - \nabla f(\mathbf{X}_s)\right]\mathrm{d}s + \beta\mathbf{B}_t\right|^{2p} \\ &\leqslant Ct^{2p-1}\mathbb{E}\int_0^t \left|\gamma\mathbf{M}_s + \nabla f(\mathbf{X}_s)\right|^{2p}\mathrm{d}s + C\mathbb{E}\left|\mathbf{B}_t\right|^{2p} \\ &\leqslant Ct^{2p-1}\int_0^t \left(\gamma^{2p}\mathbb{E}\left|\mathbf{M}_s\right|^{2p} + L^{2p}\mathbb{E}\left|\mathbf{X}_s\right|^{2p} + B^{2p}\right)\mathrm{d}s + Ct^p d^p, \\ &\leqslant Ct^{2p}(\mathcal{V}(\mathbf{m},\mathbf{x})^p + d^p) + Ct^p d^p,\end{aligned}$$

where the first inequality is by the Hölder's inequality, the second inequality is by (3.1) and the last one is due to (3.5). Similarly, it holds

$$\mathbb{E}\left|\mathbf{X}_t^{\mathbf{x}} - \mathbf{x}\right|^{2p} = \mathbb{E}\left|\int_0^t \mathbf{M}_s\,\mathrm{d}s\right|^{2p} \leqslant t^{2p-1}\int_0^t \mathbb{E}\left|\mathbf{M}_s\right|^{2p}\mathrm{d}s \leqslant Ct^{2p}(\mathcal{V}(\mathbf{m},\mathbf{x})^p + d^p).$$

Combining above two inequalities, we can obtain the desired.

□

Proof of Lemma 3.3 . Recall SDE (3.4) and rewrite it as

$$\begin{cases}\widetilde{\mathbf{M}}_t^{\mathbf{m}} - \mathbf{m} = -\gamma t\mathbf{m} - t\nabla f(\mathbf{x}) + t\left(\nabla f(\mathbf{x}) - \dfrac{1}{N}\displaystyle\sum_{i=1}^N \nabla F(\mathbf{x},\xi^i)\right) + \beta\mathbf{B}_t\ , \\ \widetilde{\mathbf{X}}_t^{\mathbf{x}} - \mathbf{x} = t\mathbf{m}\ .\end{cases}$$

for all $t \in [0,\eta]$ with η being a step size. To make notations simple, denote

$$\mathbf{\Lambda}(\mathbf{x},\xi^i) = \nabla f(\mathbf{x}) - \nabla F(\mathbf{x},\xi^i), \quad i = 1,\ldots,N, \tag{B.2}$$

which are i.i.d. random vectors. Then, Assumption II and Lemma B.1 imply that

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i)\right|^{2p} \leqslant \frac{C}{N^p}\mathbb{E}\left|\mathbf{\Lambda}(\mathbf{x},\xi^i)\right|^{2p} \leqslant \frac{C}{N^p}, \quad \forall\ 1 \leqslant p \leqslant \frac{q}{2}. \tag{B.3}$$

If the condition (2.7) in Assumption IV holds additionally, (B.3) holds for all $1 \leqslant p \leqslant q'/2$. Thus, we have that

$$\begin{aligned}
&\mathbb{E}\left|\widetilde{\mathbf{M}}_t^{\mathbf{m}} - \mathbf{m}\right|^{2p} \\
&\leqslant C\left[t^{2p}\mathbb{E}\left|\gamma\mathbf{m} + \nabla f(\mathbf{x})\right|^{2p} + t^{2p}\mathbb{E}\left|N^{-1}\sum\nolimits_{i=1}^{d}\mathbf{\Lambda}(\mathbf{x},\xi^i)\right|^{2p} + \beta^{2p}\mathbb{E}\left|\mathbf{B}_t\right|^{2p}\right] \\
&\leqslant C\left[|\mathbf{m}|^{2p} + |\mathbf{x}|^{2p} + 1\right]t^{2p} + Ct^p d^p.
\end{aligned}$$

Using (3.2) , we can obtain

$$\mathbb{E}\left|\widetilde{\mathbf{M}}_t^{\mathbf{m}} - \mathbf{m}\right|^{2p} \leqslant Ct^p(\mathcal{V}(\mathbf{m},\mathbf{x})^p + d^p).$$

Similarly, we have

$$\mathbb{E}\left|\widetilde{\mathbf{X}}_t^{\mathbf{x}} - \mathbf{x}\right|^{2p} = t^{2p}\mathbb{E}\left|\mathbf{m}\right|^{2p} \leqslant Ct^{2p}\mathcal{V}(\mathbf{m},\mathbf{x})^p.$$

Hence, by $0 \leqslant t \leqslant \eta < 1$, combining all above yields the desired. □

Proof of Lemma 3.4 . Consider SDE (1.6), let us prove the case of $p = 1$ firstly. By Itô's formula, $(\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t))_{0\leqslant t\leqslant\eta}$ satisfies

(B.4)

$$\begin{aligned}
\mathrm{d}\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t) &= -\frac{\gamma}{2}\left\langle\widetilde{\mathbf{M}}_t, \mathbf{m}\right\rangle\mathrm{d}t - \frac{\lambda\gamma^2}{2}\left\langle\widetilde{\mathbf{X}}_t, \mathbf{m}\right\rangle\mathrm{d}t + \left\langle\nabla f(\widetilde{\mathbf{X}}_t), \mathbf{m}\right\rangle\mathrm{d}t + \frac{d\beta^2}{2}\,\mathrm{d}t \\
&\quad - \left\langle\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t, \frac{1}{N}\sum\nolimits_{i=1}^{N}\nabla F(\mathbf{x},\xi^i)\right\rangle\mathrm{d}t + \beta\left\langle\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t, \mathrm{d}\mathbf{B}_t\right\rangle \\
&=: \left[I_1 + I_2\right]\mathrm{d}t + \beta\left\langle\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t, \mathrm{d}\mathbf{B}_t\right\rangle,
\end{aligned}$$

where I_1 and I_2 are given by

$$\begin{aligned}
I_1 &= -\frac{\gamma}{2}\left|\widetilde{\mathbf{M}}_t\right|^2 - \frac{\lambda\gamma^2}{2}\left\langle\widetilde{\mathbf{X}}_t, \widetilde{\mathbf{M}}_t\right\rangle - \frac{\gamma}{2}\left\langle\widetilde{\mathbf{X}}_t, \nabla f(\widetilde{\mathbf{X}}_t)\right\rangle + \frac{d\beta^2}{2}, \\
I_2 &= \frac{\gamma}{2}\left\langle\widetilde{\mathbf{M}}_t + \lambda\gamma\widetilde{\mathbf{X}}_t, \widetilde{\mathbf{M}}_t - \mathbf{m}\right\rangle - \left\langle\nabla f(\widetilde{\mathbf{X}}_t), \widetilde{\mathbf{M}}_t - \mathbf{m}\right\rangle \\
&\quad + \left\langle\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t, \nabla f(\widetilde{\mathbf{X}}_t) - \nabla f(\mathbf{x})\right\rangle + \left\langle\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t, \frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{\Lambda}(\mathbf{x},\xi^i)\right\rangle.
\end{aligned}$$

Here, $\mathbf{\Lambda}(\mathbf{x},\xi^i)$, $i = 1,\ldots,N$ are defined by (B.2).

For I_1, by the definition of $\mathcal{V}$ in (2.6) and (A.1) , we have

(B.5)

$$\begin{aligned}
I_1 &\leqslant -\left[\lambda\gamma\left(f(\widetilde{\mathbf{X}}_t) + \frac{\gamma^2}{4}|\widetilde{\mathbf{X}}_t|^2\right)\right] - \frac{\lambda\gamma^2}{2}\left\langle\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t\right\rangle - \frac{\gamma}{2}|\widetilde{\mathbf{M}}_t|^2 + \frac{1}{2}(\gamma\mathring{A} + d\beta^2) \\
&\leqslant -\lambda\gamma\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t) + \frac{1}{2}\left(\gamma\mathring{A} + d\beta^2\right).
\end{aligned}$$

For I_2, the Young's inequality yields that for any $\varepsilon > 0$

$$\begin{aligned}
I_2 &\leqslant \varepsilon\left[\frac{\gamma}{2}\left|\widetilde{\mathbf{M}}_t + \lambda\gamma\widetilde{\mathbf{X}}_t\right|^2 + 2\left(L^2\left|\widetilde{\mathbf{X}}_t\right|^2 + B^2\right) + 2\left|\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t\right|^2\right] \\
&\quad + \frac{1}{4\varepsilon}\left[\frac{\gamma+2}{2}\left|\widetilde{\mathbf{M}}_t - \mathbf{m}\right|^2 + L^2\left|\widetilde{\mathbf{X}}_t - \mathbf{x}\right|^2\right] + \frac{1}{4\varepsilon}\left|\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{\Lambda}(\mathbf{x},\xi^i)\right|^2.
\end{aligned}$$

Thus, by choosing a special ε and (3.2), we can obtain

(B.6)
$$I_2 \leqslant \frac{\lambda\gamma}{3}\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t) + C\left[\left|\widetilde{\mathbf{M}}_t - \mathbf{m}\right|^2 + \left|\widetilde{\mathbf{X}}_t - \mathbf{x}\right|^2\right] + C\left|\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{\Lambda}(\mathbf{x}, \xi^i)\right|^2 + C.$$

Combining (B.4), (B.5) and (B.6), we have

(B.7)
$$\begin{aligned}\mathrm{d}\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t) \leqslant & -\frac{2\lambda\gamma}{3}\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t)\,\mathrm{d}t + \beta\left\langle\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t, \mathrm{d}\mathbf{B}_t\right\rangle \\ & + C\left\{\left[\left|\widetilde{\mathbf{M}}_t - \mathbf{m}\right|^2 + \left|\widetilde{\mathbf{X}}_t - \mathbf{x}\right|^2\right] + \left|\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{\Lambda}(\mathbf{x}, \xi^i)\right|^2 + d\right\}\mathrm{d}t\end{aligned}$$

Then, by Lemmas 3.3 and B.1, we have the following differential inequality

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t) \leqslant -\frac{2\lambda\gamma}{3}\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t) + Ct\mathcal{V}(\mathbf{m}, \mathbf{x}) + Cd,$$

for all $0 \leqslant t \leqslant \eta < 1$. Then, solving above inequality, we have

$$\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_\eta, \widetilde{\mathbf{X}}_\eta) \leqslant \mathrm{e}^{-2\lambda\gamma\eta/3}\mathcal{V}(\mathbf{m}, \mathbf{x}) + C\eta^2\mathcal{V}(\mathbf{m}, \mathbf{x}) + Cd\eta.$$

Furthermore, the fact $1 - x \leqslant \mathrm{e}^{-x} \leqslant 1 - x + x^2/2$ for all $x \in \mathbb{R}$ yields that one can find a positive constant η_0 such that the following inequality holds for all $\eta \leqslant \eta_0$,

$$\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_\eta, \widetilde{\mathbf{X}}_\eta) \leqslant \mathrm{e}^{-\gamma\lambda\eta/2}\mathcal{V}(\mathbf{m}, \mathbf{x}) + Cd\eta. \tag{B.8}$$

By Markov property, (B.8) implies the following recursive inequality,

$$\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_{t_k}, \widetilde{\mathbf{X}}_{t_k}) \leqslant \mathrm{e}^{-\lambda\gamma\eta_k/2}\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_{t_{k-1}}, \widetilde{\mathbf{X}}_{t_{k-1}}) + Cd\eta_k,$$

as long as $\eta_k \leqslant \eta_0$. Then, (3.7) with $p = 1$ follows from that

(B.9)
$$\begin{aligned}&\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_{t_n}, \widetilde{\mathbf{X}}_{t_n}) - \mathrm{e}^{-\lambda\gamma t_n/2}\mathcal{V}(\mathbf{m}, \mathbf{x}) \\ &= \sum_{k=1}^{n}\mathrm{e}^{-\lambda\gamma(t_n - t_k)/2}\left[\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_{t_k}, \widetilde{\mathbf{X}}_{t_k}) - \mathrm{e}^{-\lambda\gamma\eta_k/2}\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_{t_{k-1}}, \widetilde{\mathbf{X}}_{t_{k-1}})\right] \\ &\leqslant Cd\sum_{k=1}^{n}\eta_k\mathrm{e}^{-\lambda\gamma(t_n - t_k)/2} \\ &= Cd\sum_{k=1}^{n}\frac{\eta_k}{1 - \mathrm{e}^{-\lambda\gamma\eta_k/2}}\left[\mathrm{e}^{-\lambda\gamma(t_n - t_k)/2} - \mathrm{e}^{-\lambda\gamma(t_n - t_{k-1})/2}\right] \leqslant Cd,\end{aligned}$$

where the last inequality holds as $\eta_k \leqslant 2/(\gamma\lambda)$.

As for $p > 1$, by Itô's formula and (B.4), we have that for $0 \leqslant t \leqslant \eta$,

$$\begin{aligned}\mathrm{d}\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t)^p = & p\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t)^{p-1}[I_1 + I_2]\,\mathrm{d}t + p\beta\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t)^{p-1}\left\langle\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t, \mathrm{d}\mathbf{B}_t\right\rangle \\ & + \frac{p(p-1)\beta^2}{2}\left|\widetilde{\mathbf{M}}_t + \frac{\gamma}{2}\widetilde{\mathbf{X}}_t\right|^2\mathcal{V}(\widetilde{\mathbf{M}}_t, \widetilde{\mathbf{X}}_t)^{p-2}\,\mathrm{d}t.\end{aligned}$$

This, together with (B.5), (B.6) and $|\mathbf{m}+\gamma\mathbf{x}/2|^2 \leqslant C\mathcal{V}(\mathbf{m},\mathbf{x})$ by (3.2), it holds that

$$\begin{aligned}\mathrm{d}\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^p \leqslant &\left[p\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^{p-1}\left(-\frac{2\lambda\gamma}{3}\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)+C\mathcal{R}_0\right)\right]\mathrm{d}t\\ &+p\beta\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^{p-1}\left\langle\widetilde{\mathbf{M}}_t+\frac{\gamma}{2}\widetilde{\mathbf{X}}_t,\mathrm{d}\mathbf{B}_t\right\rangle,\end{aligned} \tag{B.10}$$

where the reserved item $\mathcal{R}_0$ is of the form

$$\mathcal{R}_0=\left|\widetilde{\mathbf{M}}_t-\mathbf{m}\right|^2+\left|\widetilde{\mathbf{X}}_t-\mathbf{x}\right|^2+\left|\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{\Lambda}(\mathbf{x},\xi^i)\right|^2+d,$$

which satisfies the following by Lemmas 3.3 and B.1:

$$\mathbb{E}\left[\mathcal{R}_0^p\right]\leqslant Ct^p\mathcal{V}(\mathbf{m},\mathbf{x})^p+Cd^p. \tag{B.11}$$

By Young's inequality again, we can obtain that

$$\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^{p-1}\cdot(C\mathcal{R}_0)\leqslant\left(1-\frac{1}{p}\right)\frac{2\lambda\gamma}{3}\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^p+C\mathcal{R}_0^p.$$

Substituting this into (B.10), there is

$$\mathrm{d}\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^p\leqslant-\frac{2\lambda\gamma}{3}\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^p\,\mathrm{d}t+C\mathcal{R}_0^p\,\mathrm{d}t+p\beta\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^{p-1}\left\langle\widetilde{\mathbf{M}}_t+\frac{\gamma}{2}\widetilde{\mathbf{X}}_t,\mathrm{d}\mathbf{B}_t\right\rangle.$$

Combining (B.11), this implies the follow differential inequality

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^p\leqslant-\frac{2\lambda\gamma}{3}\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_t,\widetilde{\mathbf{X}}_t)^p+Ct^p\mathcal{V}(\mathbf{m},\mathbf{x})^p+Cd^p.$$

By Gronwall's inequality, we have

$$\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_\eta,\widetilde{\mathbf{X}}_\eta)^p\leqslant\mathrm{e}^{-2\lambda\gamma\eta/3}\mathcal{V}(\mathbf{m},\mathbf{x})^p+C\eta^p\mathcal{V}(\mathbf{m},\mathbf{x})^p+Cd^p\eta.$$

And one can also find a positive constant η_0' such that for all $\eta\leqslant\eta_0'$, the following holds,

$$\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_\eta,\widetilde{\mathbf{X}}_\eta)^p\leqslant\mathrm{e}^{-\lambda\gamma\eta/2}\mathcal{V}(\mathbf{m},\mathbf{x})^p+Cd^p\eta.$$

Hence, for any $k\geqslant 1$, as $\eta_k\leqslant\eta_0'$, by Markov property, we have

$$\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_{t_k},\widetilde{\mathbf{X}}_{t_k})^p\leqslant\mathrm{e}^{-\lambda\gamma\eta_k/2}\mathbb{E}\mathcal{V}(\widetilde{\mathbf{M}}_{t_{k-1}},\widetilde{\mathbf{X}}_{t_{k-1}})^p+Cd^p\eta_k.$$

which implies (3.7) for $1<p\leqslant q/2$ under Assumption II (or $1<p\leqslant q'/2$ if the condition (2.7) in Assumption IV holds) by recursion similar to (B.9).

The proof is complete. □

B.2. **Proofs of Auxiliary Lemmas for 1-Wasserstein Distance.**

Proof of Lemma 3.5. (i) Suppose that $(\widehat{\mathbf{M}}_t,\widehat{\mathbf{X}}_t)_{0\leqslant t\leqslant\eta}$ is the solution of SDE

$$\begin{cases}\mathrm{d}\widehat{\mathbf{M}}_t=-\gamma\mathbf{m}\,\mathrm{d}t-\nabla f(\mathbf{x})\,\mathrm{d}t+\beta\,\mathrm{d}\mathbf{B}_t,\\ \mathrm{d}\widehat{\mathbf{X}}_t=\mathbf{m}\,\mathrm{d}t,\end{cases}$$

with initial value $(\widehat{\mathbf{M}}_0,\widehat{\mathbf{X}}_0)=(\mathbf{m},\mathbf{x})$. We claim that

$$d_{\mathcal{W}_1}(\mathcal{L}(\mathbf{M}_\eta,\mathbf{X}_\eta),\mathcal{L}(\widehat{\mathbf{M}}_\eta,\widehat{\mathbf{X}}_\eta))\leqslant C\eta^{3/2}\sqrt{\mathcal{V}(\mathbf{m},\mathbf{x})+d}, \tag{B.12}$$

$$d_{\mathcal{W}_1}(\mathcal{L}(\widehat{\mathbf{M}}_\eta,\widehat{\mathbf{X}}_\eta),\mathcal{L}(\widetilde{\mathbf{M}}_\eta,\widetilde{\mathbf{X}}_\eta))\leqslant C\frac{\sqrt{d}}{N}\eta^{3/2}, \tag{B.13}$$

which immediately implies the desired result by the triangle inequality.

Let us now show (B.12). Observe that $(\mathbf{M}_t - \widehat{\mathbf{M}}_t, \mathbf{X}_t - \widehat{\mathbf{X}}_t)_{0\leqslant t\leqslant \eta}$ satisfies

$$\begin{cases} \mathrm{d}(\mathbf{M}_t - \widehat{\mathbf{M}}_t) = -\gamma(\mathbf{M}_t - \mathbf{m})\,\mathrm{d}t - (\nabla f(\mathbf{X}_t) - \nabla f(\mathbf{x}))\,\mathrm{d}t, \\ \mathrm{d}(\mathbf{X}_t - \widehat{\mathbf{X}}_t) = (\mathbf{M}_t - \mathbf{m})\,\mathrm{d}t, \end{cases}$$

with initial value $(\mathbf{M}_t - \widehat{\mathbf{M}}_t, \mathbf{X}_t - \widehat{\mathbf{X}}_t) = (\mathbf{0}, \mathbf{0})$. Then, it holds that

$$\begin{aligned} \mathbb{E}\left|\mathbf{M}_t - \widehat{\mathbf{M}}_t\right|^2 &\leqslant 2\gamma^2\mathbb{E}\left|\int_0^t (\mathbf{M}_s - \mathbf{m})\,\mathrm{d}s\right|^2 + 2\mathbb{E}\left|\int_0^t (\nabla f(\mathbf{X}_s) - \nabla f(\mathbf{x}))\,\mathrm{d}s\right|^2 \\ &\leqslant 2\gamma^2 t\mathbb{E}\int_0^t |\mathbf{M}_s - \mathbf{m}|^2\,\mathrm{d}s + 2t\mathbb{E}\int_0^t |\nabla f(\mathbf{X}_s) - \nabla f(\mathbf{x})|^2\,\mathrm{d}s, \\ \mathbb{E}\left|\mathbf{X}_t - \widehat{\mathbf{X}}_t\right|^2 &= \mathbb{E}\left|\int_0^t (\mathbf{M}_s - \mathbf{m})\,\mathrm{d}s\right|^2 \leqslant t\int_0^t \mathbb{E}\,|\mathbf{M}_s - \mathbf{m}|^2\,\mathrm{d}s. \end{aligned}$$

These, together with Lemma 3.2, we have

$$\mathbb{E}\left|\mathbf{M}_\eta - \widehat{\mathbf{M}}_\eta\right|^2 + \mathbb{E}\left|\mathbf{X}_\eta - \widehat{\mathbf{X}}_\eta\right|^2 \leqslant C\eta^3(\mathcal{V}(\mathbf{m},\mathbf{x}) + d).$$

By definition of 1-Wasserstein distance, we immediately know (B.12) holds true.

It remains to prove (B.13). Due to $\widehat{\mathbf{X}}_\eta = \widetilde{\mathbf{X}}_\eta = \mathbf{x} + \eta\mathbf{m}$, it is equivalent to prove

$$d_{\mathcal{W}_1}(\mathcal{L}(\widehat{\mathbf{M}}_\eta), \mathcal{L}(\widetilde{\mathbf{M}}_\eta)) \leqslant C\frac{\sqrt{d}}{N}\eta^{3/2},$$

or by Kantorovich-Rubinstein Theorem,

$$\left|\mathbb{E}h(\widehat{\mathbf{M}}_\eta) - \mathbb{E}h(\widetilde{\mathbf{M}}_\eta)\right| \leqslant C\frac{\sqrt{d}}{N}\eta^{3/2}\left\|h\right\|_{\mathrm{Lip}},$$

for any Lipschitz function h. Since h can be pointwise approximated by a sequence of $h_i \in \mathcal{C}_b^2(\mathbb{R}^d, \mathbb{R})$, $i = 1, 2, \ldots$, satisfying $\left\|\nabla h_i\right\|_\infty \leqslant 2\left\|h\right\|_{\mathrm{Lip}}$, we just need to show that

$$\left|\mathbb{E}h(\widehat{\mathbf{M}}_\eta) - \mathbb{E}h(\widetilde{\mathbf{M}}_\eta)\right| \leqslant C\frac{\sqrt{d}}{N}\eta^{3/2}\left\|\nabla h\right\|_\infty, \quad \forall\ h \in \mathcal{C}_b^2(\mathbb{R}^d). \tag{B.14}$$

For any fixed $h \in \mathcal{C}_b^2(\mathbb{R}^d)$, we define

$$h_0(\mathbf{z}) := \mathbb{E}\left[h(\mathbf{z} + \beta\mathbf{B}_\eta)\right] = \int_{\mathbb{R}^d} h(\mathbf{z}+\mathbf{y})\cdot(2\pi\beta^2\eta)^{-d/2}\exp\left\{-|\mathbf{y}|^2/(2\beta^2\eta)\right\}\mathrm{d}\mathbf{y}.$$

Straightforward calculations derive that

$$\begin{aligned} \frac{\partial^2 h_0}{\partial z_i \partial z_j}(\mathbf{z}) &= \int_{\mathbb{R}^d} \frac{\partial^2 h}{\partial z_i \partial z_j}(\mathbf{z}+\mathbf{y})\cdot(2\pi\beta^2\eta)^{-d/2}\exp\left\{-|\mathbf{y}|^2/(2\beta^2\eta)\right\}\mathrm{d}\mathbf{y} \\ &= \int_{\mathbb{R}^d} \frac{\partial h}{\partial z_i}(\mathbf{z}+\mathbf{y})\cdot\frac{y_j}{\beta^2\eta}\cdot(2\pi\beta^2\eta)^{-d/2}\exp\left\{-|\mathbf{y}|^2/(2\beta^2\eta)\right\}\mathrm{d}\mathbf{y} \\ &= \frac{1}{\beta^2\eta}\mathbb{E}\left[\frac{\partial h}{\partial z_i}(\mathbf{z}+\beta\mathbf{B}_\eta)\cdot\beta\mathbf{B}_\eta^{(j)}\right], \end{aligned}$$

where $\mathbf{B}_\eta^{(j)}$ is the j-th component of $\mathbf{B}_\eta$. That is

$$\nabla^2 h_0(\mathbf{z}) = \frac{1}{\beta\eta}\mathbb{E}\left[\nabla h(\mathbf{z}+\beta\mathbf{B}_\eta)\mathbf{B}_\eta^\top\right],$$

which implies that

$$\left\|\nabla^2 h_0\right\|_\infty \leqslant \frac{\|\nabla h\|_\infty}{\beta\eta}\mathbb{E}\left|\mathbf{B}_\eta\right| \leqslant \frac{\|\nabla h\|_\infty}{\beta}\sqrt{\frac{d}{\eta}}. \tag{B.15}$$

Taylors expansion gives that

$$\begin{aligned}
&\mathbb{E}h(\widetilde{\mathbf{M}}_\eta) - \mathbb{E}h(\widehat{\mathbf{M}}_\eta)\\
&= \mathbb{E}\left[h_0\Big((1-\eta\gamma)\mathbf{m} - \eta\frac{1}{N}\sum_{i=1}^N \nabla F(\mathbf{x},\xi^i)\Big)\right] - \mathbb{E}\left[h_0\big((1-\eta\gamma)\mathbf{m} - \eta\nabla f(\mathbf{x})\big)\right]\\
&= \eta\mathbb{E}\left[\left\langle \nabla h_0\big((1-\eta\gamma)\mathbf{m} - \eta\nabla f(\mathbf{x})\big), \frac{1}{N}\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i)\right\rangle\right]\\
&\quad + \frac{1}{2}\eta^2\mathbb{E}\left[\left\langle \frac{1}{N}\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i), \nabla^2 h_0(\boldsymbol{\zeta})\frac{1}{N}\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i)\right\rangle\right]\\
&= \frac{1}{2}\eta^2\mathbb{E}\left[\left\langle \frac{1}{N}\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i), \nabla^2 h_0(\boldsymbol{\zeta})\frac{1}{N}\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i)\right\rangle\right],
\end{aligned}$$

holds for some random vector $\boldsymbol{\zeta}$, where $\mathbf{\Lambda}(\mathbf{x},\xi^i)$, $i = 1,\dots,N$ are defined in (B.2) and we use the independence between $\mathbf{B}_\eta$ and ξ. Together with (B.15) and Lemma B.1, we have

$$\left|\mathbb{E}h(\widetilde{\mathbf{M}}_\eta) - \mathbb{E}h(\widehat{\mathbf{M}}_\eta)\right| \leqslant \frac{1}{2}\eta^2\left\|\nabla^2 h_0\right\|_\infty \mathbb{E}_\xi\left|\frac{1}{N}\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i)\right|^2 \leqslant C\frac{\sqrt{d}}{N}\eta^{3/2}\left\|\nabla h\right\|_\infty,$$

which is (B.14), and leads to (B.13).

(ii) Notice that $(\mathbf{M}_t - \widetilde{\mathbf{M}}_t, \mathbf{X}_t - \widetilde{\mathbf{X}}_t)_{0\leqslant t\leqslant\eta}$ satisfies

$$\begin{cases}
\mathrm{d}(\mathbf{M}_t - \widetilde{\mathbf{M}}_t) = -\gamma(\mathbf{M}_t - \mathbf{m})\,\mathrm{d}t - (\nabla f(\mathbf{X}_t) - \nabla f(\mathbf{x}))\,\mathrm{d}t - \dfrac{1}{N}\displaystyle\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i)\,\mathrm{d}t,\\
\mathrm{d}(\mathbf{X}_t - \widetilde{\mathbf{X}}_t) = (\mathbf{M}_t - \mathbf{m})\,\mathrm{d}t,
\end{cases}$$

with initial value $(\mathbf{M}_0 - \widetilde{\mathbf{M}}_0, \mathbf{X}_0 - \widetilde{\mathbf{X}}_0) = (\mathbf{0},\mathbf{0})$. Combining Hölder's inequality and Lemma B.1 derives that

$$\begin{aligned}
\mathbb{E}\left|\mathbf{M}_\eta - \widetilde{\mathbf{M}}_\eta\right|^4 &\leqslant C\mathbb{E}\left|\int_0^\eta (\mathbf{M}_s - \mathbf{m})\,\mathrm{d}s\right|^4 + C\mathbb{E}\left|\int_0^\eta (\nabla f(\mathbf{X}_s) - \nabla f(\mathbf{x}))\,\mathrm{d}s\right|^4\\
&\quad + C\eta^4\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N \mathbf{\Lambda}(\mathbf{x},\xi^i)\right|^4\\
&\leqslant C\eta^3\int_0^\eta \mathbb{E}\left|\mathbf{M}_s - \mathbf{m}\right|^4\mathrm{d}s + C\eta^3\int_0^\eta \mathbb{E}\left|\mathbf{X}_s - \mathbf{x}\right|^4\mathrm{d}s + C\frac{\eta^4}{N^2},\\
\mathbb{E}\left|\mathbf{X}_\eta - \widetilde{\mathbf{X}}_\eta\right|^4 &= \mathbb{E}\left|\int_0^\eta (\mathbf{M}_s - \mathbf{m})\,\mathrm{d}s\right|^4 \leqslant \eta^3\int_0^\eta \mathbb{E}\left|\mathbf{M}_s - \mathbf{m}\right|^4\mathrm{d}s.
\end{aligned}$$

Then Lemma 3.2 with $p = 2$ implies

$$\mathbb{E}\left|\mathbf{M}_\eta - \widetilde{\mathbf{M}}_\eta\right|^4 + \mathbb{E}\left|\mathbf{X}_\eta - \widetilde{\mathbf{X}}_\eta\right|^4 \leqslant C\eta^6(\mathcal{V}(\mathbf{m},\mathbf{x})^2 + d^2) + C\frac{\eta^4}{N^2}.$$

We complete the proof. □

B.3. **Proofs of Auxiliary Lemmas in the Total Variation Distance.** Before proving Lemma 3.7, we shall introduce the well known Bismut formula in Malliavin calculus. To this end, let us briefly recall the Malliavin calculus by Bismut in our setting. Let $T > 0$ be an arbitrary number, for any $t \in [0,T]$, we take $\mathbf{Y}_t$ as a functional of Brownian motion, i.e., $\mathbf{Y}_t(\mathbf{B})$, let $\mathbf{h}$ be an adaptive stochastic process in $L^2([0,T]\times\Omega,\mathbb{R}^d)$, and define a general Malliavin derivative along $\mathbf{h}$ as the following

$$\mathrm{D}^{\mathbf{h}}\mathbf{Y}_t(\mathbf{B}) = \lim_{\varepsilon\to 0}\frac{\mathbf{Y}_t(\mathbf{B}+\varepsilon\mathbf{H})-\mathbf{Y}_t(\mathbf{B})}{\varepsilon}, \tag{B.16}$$

where $\mathbf{H}(t) = \int_0^t \mathbf{h}(s)\,\mathrm{d}s$ for $t\in[0,T]$, as long as the above limit exists in $L^2(\Omega)$. For $\phi\in\mathcal{C}_b^1(\mathbb{R}^d,\mathbb{R})$, by the chain rule,

$$\mathrm{D}^{\mathbf{h}}\phi(\mathbf{Y}_t(\mathbf{B})) = \nabla\phi(\mathbf{Y}_t(\mathbf{B}))\,\mathrm{D}^{\mathbf{h}}\mathbf{Y}_t(\mathbf{B}) = \sum_{i=1}^d \partial_i\phi(\mathbf{Y}_t(\mathbf{B}))\,\mathrm{D}^{\mathbf{h}}F_t^i(\mathbf{B})$$

The following Bismut's formula, which can be taken as an integration by parts, is an important property in Malliavin calculus:

$$\mathbb{E}\left[\nabla\phi(\mathbf{Y}_t)\,\mathrm{D}^{\mathbf{h}}\mathbf{Y}_t\right] = \mathbb{E}\left[\phi(\mathbf{Y}_t)\int_0^t\langle\mathbf{h}(s),\mathrm{d}\mathbf{B}_s\rangle\right]. \tag{B.17}$$

Proof of Lemma 3.7. Let $\mathbf{Y}_t^{\mathbf{y}} = (\mathbf{M}_t^{\mathbf{m}},\mathbf{X}_t^{\mathbf{x}})$ come from SDE (1.5). Let $\mathbf{h}$ be an adaptive stochastic process in $L^2([0,T]\times\Omega,\mathbb{R}^{2d})$ to be chosen later, by [Nor86], the Malliavin derivative of $\mathbf{Y}_t$ along $\mathbf{h}$ satisfies

$$\begin{cases}\mathrm{D}^{\mathbf{h}}\mathbf{M}_t^{\mathbf{m}} = -\gamma\displaystyle\int_0^t\mathrm{D}^{\mathbf{h}}\mathbf{M}_s^{\mathbf{m}}\,\mathrm{d}s - \int_0^t\nabla^2 f(\mathbf{X}_s^{\mathbf{x}})\,\mathrm{D}^{\mathbf{h}}\mathbf{X}_s^{\mathbf{x}}\,\mathrm{d}s + \beta\int_0^t\mathbf{h}(s)\,\mathrm{d}s,\\ \mathrm{D}^{\mathbf{h}}\mathbf{X}_t^{\mathbf{x}} = \displaystyle\int_0^t\mathrm{D}^{\mathbf{h}}\mathbf{M}_s^{\mathbf{m}}\,\mathrm{d}s.\end{cases}$$

Besides, for any $\mathbf{w} = (\mathbf{u},\mathbf{v})\in\mathbb{R}^{2d}$, the Jacobian flow of $\mathbf{Y}_t$ with respect to $\mathbf{w}$ is defined as

$$\nabla\mathbf{Y}_t^{\mathbf{y}}\cdot\mathbf{w} = \lim_{\varepsilon\to 0}\frac{\mathbf{Y}_t^{\mathbf{y}+\varepsilon\mathbf{w}}-\mathbf{Y}_t^{\mathbf{y}}}{\varepsilon}.$$

More precisely, we have that

$$\begin{aligned}\nabla_{\mathbf{m}}\mathbf{Y}_t\cdot\mathbf{u} &= \begin{bmatrix}\mathbf{u}\\ \mathbf{0}\end{bmatrix} + \int_0^t\begin{bmatrix}-\gamma\mathbf{I} & \nabla^2 f(\mathbf{X}_t)\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\nabla_{\mathbf{m}}\mathbf{Y}_t\cdot\mathbf{u}\,\mathrm{d}t,\\ \nabla_{\mathbf{x}}\mathbf{Y}_t\cdot\mathbf{v} &= \begin{bmatrix}\mathbf{0}\\ \mathbf{v}\end{bmatrix} + \int_0^t\begin{bmatrix}-\gamma\mathbf{I} & \nabla^2 f(\mathbf{X}_t)\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\nabla_{\mathbf{x}}\mathbf{Y}_t\cdot\mathbf{v}\,\mathrm{d}t.\end{aligned}$$

Observe that for any $\phi\in\mathcal{C}_b^1(\mathbb{R}^{2d},\mathbb{R})$

$$\nabla\mathbb{E}\phi(\mathbf{Y}_T)\cdot\mathbf{w} = \mathbb{E}\left[\nabla\phi(\mathbf{Y}_T)\nabla_{\mathbf{m}}\mathbf{Y}_T\cdot\mathbf{u}\right] + \mathbb{E}\left[\nabla\phi(\mathbf{Y}_T)\nabla_{\mathbf{x}}\mathbf{Y}_T\cdot\mathbf{v}\right]. \tag{B.18}$$

We aim to find some special $\mathbf{h}$ and $\tilde{\mathbf{h}}$ such that

$$\nabla_{\mathbf{m}}\mathbf{Y}_T\cdot\mathbf{u} = \mathrm{D}^{\mathbf{h}}\mathbf{Y}_T \quad\text{and}\quad \nabla_{\mathbf{x}}\mathbf{Y}_T\cdot\mathbf{v} = \mathrm{D}^{\tilde{\mathbf{h}}}\mathbf{Y}_T,$$

and consequently

$$\nabla\mathbb{E}\phi(\mathbf{Y}_T)\cdot\mathbf{w} = \mathbb{E}\left[\nabla\phi(\mathbf{Y}_T)\,\mathrm{D}^{\mathbf{h}}\mathbf{Y}_T\right] + \mathbb{E}\left[\nabla\phi(\mathbf{Y}_T)\,\mathrm{D}^{\tilde{\mathbf{h}}}\mathbf{Y}_T\right],$$

which enables us to apply Bismut's formula (B.17).

Define $\mathbf{H}(t)=\int_0^t \mathbf{h}(s)\,\mathrm{d}s$ and $\widetilde{\mathbf{H}}(t)=\int_0^t \tilde{\mathbf{h}}(s)\,\mathrm{d}s$, by Zhang [Zha10, Theorem 3.3], for any $T>0$ and $\mathbf{u},\mathbf{v}\in\mathbb{R}^d$, we can choose $\mathbf{H}(t)$ and $\widetilde{\mathbf{H}}(t)$ specified as below. Firstly, for any $\mathbf{u}\in\mathbb{R}^d$, let

$$\mathbf{H}(t)=-\frac{1}{\beta}\left(\mathbf{u}-\mathbf{g}(t)-\gamma\int_0^t\mathbf{g}(s)\,\mathrm{d}s-\int_0^t\mathrm{d}s\int_0^s\nabla^2 f(\mathbf{X}_s)\mathbf{g}(r)\,\mathrm{d}r\right),$$

for $t\in[0,T]$, where $\mathbf{g}(t)$ is given by

$$\mathbf{g}(t)=\begin{cases}(T-3t)\,\mathbf{u}/T, & 0\leqslant t<T/2,\\ (t-T)\,\mathbf{u}/T, & T/2\leqslant t\leqslant T,\end{cases}$$

satisfying $\mathbf{g}(0)=\mathbf{u}$, $\mathbf{g}(T)=\mathbf{0}$ and $\int_0^T\mathbf{g}(t)\,\mathrm{d}t=\mathbf{0}$. One can verify that

$$\begin{cases}\langle\nabla_{\mathbf{m}}\mathbf{M}_t(\mathbf{m},\mathbf{x}),\mathbf{u}\rangle=\mathrm{D}^{\mathbf{h}}\mathbf{M}_t(\mathbf{m},\mathbf{x})+\mathbf{g}(t),\\ \langle\nabla_{\mathbf{m}}\mathbf{X}_t(\mathbf{m},\mathbf{x}),\mathbf{u}\rangle=\mathrm{D}^{\mathbf{h}}\mathbf{X}_t(\mathbf{m},\mathbf{x})+\displaystyle\int_0^t\mathbf{g}(s)\,\mathrm{d}s,\end{cases}$$

and $\nabla_{\mathbf{m}}\mathbf{Y}_T\cdot\mathbf{u}=\mathrm{D}^{\mathbf{h}}\mathbf{Y}_T$. Then, applying Bismut formula to $\mathbb{E}\left[\nabla\phi(\mathbf{Y}_T)\,\mathrm{D}^{\mathbf{h}}\mathbf{Y}_T\right]$ yields that

$$\langle\nabla_{\mathbf{m}}\mathbb{E}\phi(\mathbf{Y}_T),\mathbf{u}\rangle=\mathbb{E}\left[-\phi(\mathbf{Y}_T)\frac{1}{\beta}\int_0^T\left(\nabla^2 f(\mathbf{X}_s)\int_0^s\mathbf{g}(r)\,\mathrm{d}r+\gamma\mathbf{g}(s)+\dot{\mathbf{g}}(s)\right)\mathrm{d}\mathbf{B}_s\right].$$

Hence, there exists a constant C only depending on L, γ and β such that

$$|\langle\nabla_{\mathbf{m}}\mathbb{E}\phi(\mathbf{Y}_T),\mathbf{u}\rangle|\leqslant C\left\|\phi\right\|_\infty|\mathbf{u}|\left(T^{3/2}\vee T^{-1/2}\right). \tag{B.19}$$

Analogously, for any $\mathbf{v}\in\mathbb{R}^d$, let $\widetilde{\mathbf{H}}(t)$

$$\widetilde{\mathbf{H}}(t)=\frac{1}{\beta}\left[-\widetilde{\mathbf{g}}(t)-\gamma\int_0^t\widetilde{\mathbf{g}}(s)\,\mathrm{d}s-\int_0^t\nabla^2 f(\mathbf{X}_s)\left(\mathbf{v}+\int_0^s\widetilde{\mathbf{g}}(r)\,\mathrm{d}r\right)\mathrm{d}s\right],$$

for $t\in[0,T]$, where $\widetilde{\mathbf{g}}(t)$ is given by

$$\widetilde{\mathbf{g}}(t)=\begin{cases}-4t/T^2\mathbf{v} & , \quad 0\leqslant t\leqslant T/2,\\ 4(t-T)/T^2\mathbf{v}, & T/2\leqslant t\leqslant T,\end{cases}$$

satisfying $\widetilde{\mathbf{g}}(0)=\widetilde{\mathbf{g}}(T)=0$ and $\int_0^T\widetilde{\mathbf{g}}(t)\,\mathrm{d}t=-\mathbf{v}$. Then, it is easy to verify that

$$\begin{cases}\langle\nabla_{\mathbf{x}}\mathbf{M}_t(\mathbf{m},\mathbf{x}),\mathbf{v}\rangle=\mathrm{D}^{\tilde{\mathbf{h}}}\mathbf{M}_t(\mathbf{m},\mathbf{x})+\widetilde{\mathbf{g}}(t),\\ \langle\nabla_{\mathbf{x}}\mathbf{X}_t(\mathbf{m},\mathbf{x}),\mathbf{v}\rangle=\mathrm{D}^{\tilde{\mathbf{h}}}\mathbf{X}_t(\mathbf{m},\mathbf{x})+\mathbf{v}+\displaystyle\int_0^t\widetilde{\mathbf{g}}(s)\,\mathrm{d}s,\end{cases}$$

and $\nabla_{\mathbf{x}}\mathbf{Y}_T\cdot\mathbf{v}=\mathrm{D}^{\tilde{\mathbf{h}}}\mathbf{Y}_T$. Applying Bismut's formula to $\mathbb{E}\left[\nabla\phi(\mathbf{Y}_T)\,\mathrm{D}^{\tilde{\mathbf{h}}}\mathbf{Y}_T\right]$ yields

$$\langle\nabla_{\mathbf{x}}\mathbb{E}\phi(\mathbf{Y}_T),\mathbf{v}\rangle=\mathbb{E}\left\{-\phi(\mathbf{Y}_T)\frac{1}{\beta}\int_0^T\left[\nabla^2 f(\mathbf{X}_s)\left(\mathbf{v}+\int_0^s\widetilde{\mathbf{g}}(r)\,\mathrm{d}r\right)+\gamma\widetilde{\mathbf{g}}(s)+\dot{\widetilde{\mathbf{g}}}(s)\right]\mathrm{d}\mathbf{B}_s\right\},$$

which immediately implies that there exists a constant $C>0$ only depending on L, γ and β such that

$$|\langle\nabla_{\mathbf{x}}\mathbb{E}\phi(\mathbf{Y}_T),\mathbf{v}\rangle|\leqslant C\left\|\phi\right\|_\infty|\mathbf{v}|\left(T^{1/2}\vee T^{-3/2}\right). \tag{B.20}$$

Combining (B.18), (B.19) and (B.20), we have

$$|\nabla \mathbb{E}\phi(\mathbf{Y}_T)\cdot \mathbf{w}| \leqslant C\,\|\phi\|_\infty\,(|\mathbf{u}|+|\mathbf{v}|)\left(T^{3/2}\vee T^{-3/2}\right) \leqslant C\,\|\phi\|_\infty\left(T^{3/2}\vee T^{-3/2}\right)|\mathbf{w}|$$

for any $\mathbf{w}=(\mathbf{u},\mathbf{v})\in\mathbb{R}^{2d}$, and this implies the desired. □

Proof of Lemma 3.8 . Firstly, the chain rule of Malliavin derivative yields that

$$\left\langle \mathrm{D}g(\mathbf{F}),\mathrm{D}F^j\right\rangle_{\mathcal{H}} = \sum_{i=1}^d \partial_i g(\mathbf{F})\left\langle \mathrm{D}F^i,\mathrm{D}F^j\right\rangle_{\mathcal{H}} = \sum_{i=1}^d \partial_i g(\mathbf{F})(\mathbf{\Gamma}(\mathbf{F}))_{i,j},\ j=1,\dots,d,$$

which implies that

$$\partial_i g(\mathbf{F}) = \sum_{j=1}^d \left\langle \mathrm{D}g(\mathbf{F}),\mathrm{D}F^j\right\rangle_{\mathcal{H}}\left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{i,j},\ i=1,\dots,d. \tag{B.21}$$

Then, we have

$$\begin{aligned}\mathbb{E}\left\langle \nabla g(\mathbf{F}),\mathbf{G}\right\rangle &= \mathbb{E}\left\langle \mathrm{D}g(\mathbf{F}),\sum_{i=1}^d\sum_{j=1}^d \left(\mathbf{\Gamma}(\mathbf{F})\right)^{-1}_{i,j}G^i\,\mathrm{D}F^j\right\rangle_{\mathcal{H}}\\ &= \mathbb{E}\left[g(\mathbf{F})\delta\left(\sum_{j=1}^d\left(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\right)_j\mathrm{D}F^j\right)\right].\end{aligned}$$

where the first equality is by (B.21), and the last one is due to the duality (3.10) between derivative operator and divergence operator. Then, the Jensen's inequality yields that

$$|\mathbb{E}\left\langle \nabla g(\mathbf{F}),\mathbf{G}\right\rangle| \leqslant \|g\|_\infty\sqrt{\mathbb{E}\left[\left|\delta\left(\sum_{j=1}^d\left(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\right)_j\mathrm{D}F^j\right)\right|^2\right]}. \tag{B.22}$$

Applying the continuity of divergence operator δ from $\mathbb{D}^{1,2}(\mathcal{H})$ into $L^2(\Omega)$ [Nua06, Propsition 1.5.7], there exists a constant $C>0$ such that

$$\mathbb{E}\left[\left|\delta\left(\sum_{j=1}^d\left(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\right)_j\mathrm{D}F^j\right)\right|^2\right] \leqslant C(\mathcal{I}_1+\mathcal{I}_2), \tag{B.23}$$

where $\mathcal{I}_1$ and $\mathcal{I}_2$ are given by

$$\mathcal{I}_1 = \mathbb{E}\left[\left\|\sum_{j=1}^d\left(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\right)_j\mathrm{D}F^j\right\|_{\mathcal{H}}^2\right],\quad \mathcal{I}_2 = \mathbb{E}\left[\left\|\sum_{j=1}^d\mathrm{D}\left(\left(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\right)_j\mathrm{D}F^j\right)\right\|_{\mathcal{H}\otimes\mathcal{H}}^2\right]$$

For $\mathcal{I}_1$, it follows from Cauchy-Schwarz inequality that

$$\begin{aligned}\mathcal{I}_1 &\leqslant \mathbb{E}\left[\left|\sum_{j=1}^d\left|\left(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\right)_j\right|\|\mathrm{D}F_j\|_{\mathcal{H}}\right|^2\right]\\ &\leqslant \mathbb{E}\left[\left(\sum_{j=1}^d\left(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\right)_j^2\right)\left(\sum_{j=1}^d\|\mathrm{D}F^j\|_{\mathcal{H}}^2\right)\right]\\ &\leqslant \mathbb{E}\left\{\left[\left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{op}}|\mathbf{G}|\,\|\mathrm{D}\mathbf{F}\|_{\mathcal{H}}\right]^2\right\}.\end{aligned} \tag{B.24}$$

For $\mathcal{I}_2$, according to [Nua06, Lemma 2.1.6], it holds that

$$\mathrm{D}\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\big)_j = \sum_{i=1}^{d}\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\big)_{j,i}\,\mathrm{D}G^i - \sum_{i,k,l=1}^{d} G^i\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\big)_{j,k}\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\big)_{l,i}\,\mathrm{D}\big(\mathbf{\Gamma}(\mathbf{F})_{k,l}\big).$$

This, together with representation $\mathrm{D}\big(\mathbf{\Gamma}(\mathbf{F})_{k,l}\big) = \mathrm{D}\left\langle \mathrm{D}F^k, \mathrm{D}F^l\right\rangle_{\mathcal{H}}$ for any $1 \leqslant k, l \leqslant d$, we have

$$\left\| \sum_{j=1}^{d} \mathrm{D}\left(\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\big)_j \,\mathrm{D}F^j\right)\right\|_{\mathcal{H}\otimes\mathcal{H}} \leqslant \mathcal{I}_{2,1} + \mathcal{I}_{2,2} + \mathcal{I}_{2,3}, \tag{B.25}$$

where $\mathcal{I}_{2,1}$, $\mathcal{I}_{2,2}$ and $\mathcal{I}_{2,3}$ are given by

$$\mathcal{I}_{2,1} = \sum_{j=1}^{d}\left|\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\big)_j\right| \left\|\mathrm{D}^2F^j\right\|_{\mathcal{H}\otimes\mathcal{H}}$$

$$\mathcal{I}_{2,2} = \sum_{i,j=1}^{d}\left|\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\big)_{j,i}\right| \left\|\mathrm{D}G^i\right\|_{\mathcal{H}} \left\|\mathrm{D}F^j\right\|_{\mathcal{H}}$$

$$\mathcal{I}_{2,3} = \sum_{i,j,k,l=1}^{d}\left|G^i\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\big)_{j,k}\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\big)_{l,i}\right| \left\|\mathrm{D}\left\langle \mathrm{D}F^k, \mathrm{D}F^l\right\rangle_{\mathcal{H}}\right\|_{\mathcal{H}} \left\|\mathrm{D}F^j\right\|_{\mathcal{H}}.$$

It is easy to estimate $\mathcal{I}_{2,1}$ and $\mathcal{I}_{2,2}$ which satisfy

$$\mathcal{I}_{2,1} \leqslant \left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{op}} |\mathbf{G}| \left\|\mathrm{D}^2\mathbf{F}\right\|_{\mathcal{H}\otimes\mathcal{H}}, \qquad \mathcal{I}_{2,2} \leqslant \left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{HS}} \|\mathrm{D}\mathbf{G}\|_{\mathcal{H}} \|\mathrm{D}\mathbf{F}\|_{\mathcal{H}}. \tag{B.26}$$

For $\mathcal{I}_{2,3}$, we claim that

$$\mathcal{I}_{2,3} \leqslant 2\left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{HS}}^2 |\mathbf{G}| \,\|\mathrm{D}\mathbf{F}\|_{\mathcal{H}}^2 \left\|\mathrm{D}^2\mathbf{F}\right\|_{\mathcal{H}\otimes\mathcal{H}}. \tag{B.27}$$

Combining (B.25), (B.26) and (B.27), we can obtain that

$$\begin{aligned}
\mathcal{I}_2 &= \mathbb{E}\left[\left\| \sum_{j=1}^{d} \mathrm{D}\left(\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\big)_j \,\mathrm{D}F^j\right)\right\|_{\mathcal{H}\otimes\mathcal{H}}^2\right] \\
&\leqslant C\Big\{\mathbb{E}\left[\left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{op}}^2 |\mathbf{G}|^2 \left\|\mathrm{D}^2\mathbf{F}\right\|_{\mathcal{H}\otimes\mathcal{H}}^2\right] \\
&\qquad + \mathbb{E}\left[\left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{HS}}^2 \|\mathrm{D}\mathbf{G}\|_{\mathcal{H}}^2 \|\mathrm{D}\mathbf{F}\|_{\mathcal{H}}^2\right] \\
&\qquad + \mathbb{E}\left[\left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{HS}}^4 |\mathbf{G}|^2 \|\mathrm{D}\mathbf{F}\|_{\mathcal{H}}^4 \left\|\mathrm{D}^2\mathbf{F}\right\|_{\mathcal{H}\otimes\mathcal{H}}^2\right]\Big\}.
\end{aligned} \tag{B.28}$$

Combining (B.23), (B.24) and (B.28), by Hölder's inequality, we can obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left|\delta\Big(\sum_{j=1}^{d}\big(\mathbf{\Gamma}(\mathbf{F})^{-1}\mathbf{G}\big)_j \,\mathrm{D}F^j\Big)\right|^2\right] \\
&\leqslant C\,\|\mathbf{G}\|_{1,4}^2\left\{\mathbb{E}\Big[\big(1+\|\mathrm{D}\mathbf{F}\|_{\mathcal{H}}^8\big)\big(1+\left\|\mathrm{D}^2\mathbf{F}\right\|_{\mathcal{H}\otimes\mathcal{H}}^4\big)\big(1+\left\|\mathbf{\Gamma}(\mathbf{F})^{-1}\right\|_{\mathrm{HS}}^8\big)\Big]\right\}^{1/2}.
\end{aligned}$$

Substituting this into (B.22), we can obtain the desired.

It remains to prove (B.27). Observe that for any $1 \leqslant k, l \leqslant d$

$$\left\|\mathrm{D}\left\langle \mathrm{D}F^k, \mathrm{D}F^l\right\rangle_{\mathcal{H}}\right\|_{\mathcal{H}} \leqslant \left\|\mathrm{D}^2F^k\right\|_{\mathcal{H}\otimes\mathcal{H}} \left\|\mathrm{D}F^l\right\|_{\mathcal{H}} + \left\|\mathrm{D}F^k\right\|_{\mathcal{H}} \left\|\mathrm{D}^2F^l\right\|_{\mathcal{H}\otimes\mathcal{H}}.$$

Then, for each $j = 1, \dots, d$, it holds that

$$\begin{aligned}
&\sum_{i,k,l=1}^{d} \left| G^i \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{j,k} \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{l,i} \right| \left\| \mathrm{D} \left\langle \mathrm{D}F^k, \mathrm{D}F^l \right\rangle_{\mathcal{H}} \right\|_{\mathcal{H}} \\
\leqslant &\sum_{k=1}^{d} \left| \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{j,k} \right| \left\| \mathrm{D}^2 F^k \right\|_{\mathcal{H}\otimes\mathcal{H}} \sum_{i,l=1}^{d} \left| G^i \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{l,i} \right| \left\| \mathrm{D}F^l \right\|_{\mathcal{H}} \\
&+ \sum_{k=1}^{d} \left| \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{j,k} \right| \left\| \mathrm{D}F^k \right\|_{\mathcal{H}} \sum_{i,l=1}^{d} \left| G^i \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{l,i} \right| \left\| \mathrm{D}^2 F^l \right\|_{\mathcal{H}\otimes\mathcal{H}}.
\end{aligned}$$

The Cauchy-Schwarz inequality implies that

$$\begin{aligned}
\left\{ \sum_{i,l=1}^{d} \left| G^i \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{l,i} \left\| \mathrm{D}F^l \right\|_{\mathcal{H}} \right| \right\}^2 &\leqslant \left[\sum_{i=1}^{d} (G^i)^2 \right] \left[\sum_{i,l=1}^{d} \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{l,i}^2 \right] \left[\sum_{l=1}^{d} \left\| \mathrm{D}F^l \right\|_{\mathcal{H}}^2 \right] \\
&\leqslant |\mathbf{G}|^2 \left\| \mathbf{\Gamma}(\mathbf{F})^{-1} \right\|_{\mathrm{HS}}^2 \left\| \mathrm{D}\mathbf{F} \right\|_{\mathcal{H}}^2,
\end{aligned}$$

and similarly

$$\left\{ \sum_{i,l=1}^{d} \left| G^i \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{l,i} \right| \left\| \mathrm{D}^2 F^l \right\|_{\mathcal{H}\otimes\mathcal{H}} \right\}^2 \leqslant |\mathbf{G}|^2 \left\| \mathbf{\Gamma}(\mathbf{F})^{-1} \right\|_{\mathrm{HS}}^2 \left\| \mathrm{D}^2 \mathbf{F} \right\|_{\mathcal{H}}^2.$$

Combining inequalities above, we can obtain

$$\begin{aligned}
\mathcal{I}_{2,3} &\leqslant |\mathbf{G}| \left\| \mathbf{\Gamma}(\mathbf{F})^{-1} \right\|_{\mathrm{HS}} \left\| \mathrm{D}\mathbf{F} \right\|_{\mathcal{H}} \sum_{j,k=1}^{d} \left| \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{j,k} \right| \left\| \mathrm{D}^2 F^k \right\|_{\mathcal{H}\otimes\mathcal{H}} \left\| \mathrm{D}F^j \right\|_{\mathcal{H}} \\
&\quad + |\mathbf{G}| \left\| \mathbf{\Gamma}(\mathbf{F})^{-1} \right\|_{\mathrm{HS}} \left\| \mathrm{D}^2 \mathbf{F} \right\|_{\mathcal{H}\otimes\mathcal{H}} \sum_{j,k=1}^{d} \left| \left(\mathbf{\Gamma}(\mathbf{F})^{-1}\right)_{j,k} \right| \left\| \mathrm{D}F^k \right\|_{\mathcal{H}} \left\| \mathrm{D}F^j \right\|_{\mathcal{H}} \\
&\leqslant 2 \left\| \mathbf{\Gamma}(\mathbf{F})^{-1} \right\|_{\mathrm{HS}}^2 |\mathbf{G}| \left\| \mathrm{D}\mathbf{F} \right\|_{\mathcal{H}}^2 \left\| \mathrm{D}^2 \mathbf{F} \right\|_{\mathcal{H}\otimes\mathcal{H}}.
\end{aligned}$$

The proof is complete. □

Proof of Lemma 3.9. According to Markov's inequality, it holds

$$\begin{aligned}
\mathbb{P}(E_{j,k}^1) &\leqslant \mathbb{P}\left(\eta_j \sum_{i=j+1}^{k} (\varphi(\boldsymbol{\xi}_i) - \mathbb{E}\varphi(\boldsymbol{\xi}_i)) > (k-j)\eta_k \right) \\
&\leqslant \left(\frac{\eta_j}{\eta_k}\right)^4 \mathbb{E} \left| \frac{1}{k-j} \sum_{i=j+1}^{k} \left(\varphi(\boldsymbol{\xi}_i) - \mathbb{E}\varphi(\boldsymbol{\xi}_i)\right) \right|^4,
\end{aligned}$$

where in the first inequality we use $t_k - t_j \geqslant (k-j)\eta_k$ due to $\eta_i \leqslant \eta_j$ for any $i > j$. Assumption III implies that

$$\frac{\eta_j}{\eta_k} = \prod_{i=j+1}^{k} \frac{\eta_{i-1}}{\eta_i} \leqslant \prod_{i=j+1}^{k} (1 + \omega\eta_i) \leqslant \prod_{i=j+1}^{k} \mathrm{e}^{\omega\eta_i} = \mathrm{e}^{\omega(t_k - t_j)},$$

where the first inequality follows from the condition $\eta_{i-1} \leqslant \eta_i(1+\omega\eta_i)$. Assumption IV guarantees that $\mathbb{E}\left|\varphi(\boldsymbol{\xi}_i) - \mathbb{E}[\varphi(\boldsymbol{\xi}_i)]\right|^4 < \infty$, then together with Lemma B.1, the

following holds for constant $C > 0$,

$$\mathbb{P}(E^1_{j,k}) \leqslant \frac{C\mathrm{e}^{4\omega(t_k-t_j)}}{(k-j)^2} \leqslant \frac{C\mathrm{e}^{4\omega(t_k-t_j)}}{(t_k-t_j)^2}\eta_j^2 \leqslant \frac{C\mathrm{e}^{6\omega(t_k-t_j)}}{(t_k-t_j)^2}\eta_k^2.$$

Since Assumption IV implies $\mathbb{E}\,|\varphi(\boldsymbol{\xi}_i)^2 - \mathbb{E}[\varphi(\boldsymbol{\xi}_i)^2]|^4 < +\infty$, $\mathbb{P}(E^2_{j,k})$ can be analogously estimated only with $\varphi(\boldsymbol{\xi}_i)$ replaced by $\varphi(\boldsymbol{\xi}_i)^2$. □

Recall SDE (1.6), we have

$$\begin{cases} \widetilde{\mathbf{M}}_{t_j} = (1-\gamma\eta_j)\widetilde{\mathbf{M}}_{t_{j-1}} - \dfrac{1}{N}\displaystyle\sum_{r=1}^{N} \nabla F(\widetilde{\mathbf{X}}_{t_{j-1}}, \xi_j^r)\eta_j + \beta(\mathbf{B}_{t_j} - \mathbf{B}_{t_{j-1}}), \\ \widetilde{\mathbf{X}}_{t_j} = \eta_j\widetilde{\mathbf{M}}_{t_{j-1}} + \widetilde{\mathbf{X}}_{t_{j-1}}, \end{cases}$$

for $j \geqslant 1$ and initial value $(\widetilde{\mathbf{M}}_0, \widetilde{\mathbf{X}}_0) = (\mathbf{m}, \mathbf{x})$. By [Nua06, Section 2.2.2], the Malliavin derivative satisfies

$$\begin{bmatrix} \mathrm{D}_s\widetilde{\mathbf{M}}_{t_j} \\ \mathrm{D}_s\widetilde{\mathbf{X}}_{t_j} \end{bmatrix} = \mathbf{E}_j^\top \begin{bmatrix} \mathrm{D}_s\widetilde{\mathbf{M}}_{t_{j-1}} \\ \mathrm{D}_s\widetilde{\mathbf{X}}_{t_{j-1}} \end{bmatrix} + \beta \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0} \end{bmatrix} \mathbf{1}_{[t_{j-1},t_j)}(s), \tag{B.29}$$

where the matrix $\mathbf{E}_j$ is defined for simplicity as

$$\mathbf{E}_j \equiv \begin{bmatrix} (1-\gamma\eta_j)\mathbf{I}_d & \eta_j\mathbf{I}_d \\ -N^{-1}\sum_{r=1}^{N} \nabla^2 F(\widetilde{\mathbf{X}}_{t_{j-1}}, \xi_j^r)\eta_j & \mathbf{I}_d \end{bmatrix}. \tag{B.30}$$

Besides, define $\mathbf{S}_j$ as

$$\mathbf{S}_k \equiv \mathbf{I}_{2d}, \qquad \mathbf{S}_j \equiv \mathbf{E}_{j+1}\ldots\mathbf{E}_k, \quad j = 1, \ldots, k-1.$$

Given $(\widetilde{\mathbf{M}}_{t_\ell}, \widetilde{\mathbf{X}}_{t_\ell}) = \mathbf{z}$, the Malliavin derivative of $\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k} = (\widetilde{\mathbf{M}}_{t_k}, \widetilde{\mathbf{X}}_{t_k})$ has the following recursion:

$$\begin{bmatrix} \mathrm{D}_s\widetilde{\mathbf{M}}_{t_k} \\ \mathrm{D}_s\widetilde{\mathbf{X}}_{t_k} \end{bmatrix} = \beta \sum_{j=\ell+1}^{k} \mathbf{S}_j^\top \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0} \end{bmatrix} \mathbf{1}_{[t_{j-1},t_j)}(s), \tag{B.31}$$

and the corresponding Malliavin matrix is given by

$$\boldsymbol{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\big) = \int_{t_\ell}^{t_k} \begin{bmatrix} \mathrm{D}_s\widetilde{\mathbf{M}}_{t_k} \\ \mathrm{D}_s\widetilde{\mathbf{X}}_{t_k} \end{bmatrix} \begin{bmatrix} \mathrm{D}_s\widetilde{\mathbf{M}}_{t_k} \\ \mathrm{D}_s\widetilde{\mathbf{X}}_{t_k} \end{bmatrix}^\top \mathrm{d}s = \beta^2 \sum_{j=\ell+1}^{k} \eta_j \mathbf{S}_j^\top \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \end{bmatrix} \mathbf{S}_j. \tag{B.32}$$

Before estimating the lower bound of eigenvalues of $\boldsymbol{\Gamma}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k})$, let us derive an estimate of $\|\mathbf{S}_j - \mathbf{I}_{2d}\|_{\mathrm{op}}$ first.

Lemma B.2. *For any fixed $k \geqslant 1$, the following holds for each $1 \leqslant j \leqslant k$,*

$$\|\mathbf{S}_j - \mathbf{I}_{2d}\|_{\mathrm{op}} \leqslant \exp\left\{\sum_{i=j+1}^{k} \eta_i\big(\varphi(\boldsymbol{\xi}_i) + \gamma + 1\big)\right\} - 1,$$

where $\varphi(\boldsymbol{\xi}_i) = N^{-1}\sum_{r=1}^{N} \sup_{\mathbf{x}\in\mathbb{R}^d} \|\nabla^2 F(\mathbf{x}, \xi_i^r)\|_{\mathrm{op}}$, $i = j+1, \ldots, k$, are in (3.12).

Proof. We prove this by a backward induction. By definition, $\mathbf{S}_k = \mathbf{I}_{2d}$, so this result holds for $j = k$. Assume that

$$\|\mathbf{S}_\ell - \mathbf{I}_{2d}\|_{\mathrm{op}} \leqslant \exp\left\{\sum_{i=\ell+1}^{k} \eta_i(\varphi(\boldsymbol{\xi}_i) + \gamma + 1)\right\} - 1,$$

holds for some $\ell \geqslant 2$, then

$$\|\mathbf{S}_{\ell-1} - \mathbf{I}_{2d}\|_{\mathrm{op}} = \|\mathbf{E}_\ell \mathbf{S}_\ell - \mathbf{E}_\ell + \mathbf{E}_\ell - \mathbf{I}_{2d}\|_{\mathrm{op}} \leqslant \|\mathbf{E}_\ell\|_{\mathrm{op}} \|\mathbf{S}_\ell - \mathbf{I}_{2d}\|_{\mathrm{op}} + \|\mathbf{E}_\ell - \mathbf{I}_{2d}\|_{\mathrm{op}}.$$

For any matrix $\mathbf{K} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{2d\times 2d}$ and $\mathbf{w} = (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^d$, it holds

$$|\mathbf{K}\mathbf{w}| = |\mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} + \mathbf{C}\mathbf{u}| \leqslant [\|\mathbf{A}\|_{\mathrm{op}} + \|\mathbf{B}\|_{\mathrm{op}} + \|C\|_{\mathrm{op}}]\, |\mathbf{w}|,$$

which implies that $\|\mathbf{K}\|_{\mathrm{op}} \leqslant \|\mathbf{A}\|_{\mathrm{op}} + \|\mathbf{B}\|_{\mathrm{op}} + \|C\|_{\mathrm{op}}$.Then, we have

$$\|\mathbf{E}_\ell - \mathbf{I}_{2d}\|_{\mathrm{op}} = \eta_\ell \left\| \begin{bmatrix} -\gamma \mathbf{I}_d & \mathbf{I}_d \\ -\frac{1}{N}\sum_{r=1}^N \nabla^2 F(\widetilde{\mathbf{X}}_{t_{\ell-1}}, \xi_\ell^r) & \mathbf{0} \end{bmatrix} \right\|_{\mathrm{op}} \leqslant \eta_\ell \big(\varphi(\boldsymbol{\xi}_\ell) + \gamma + 1\big).$$

Hence, we obtain the desired result by

$$\begin{aligned} \|\mathbf{S}_{\ell-1} - \mathbf{I}_{2d}\|_{\mathrm{op}} &\leqslant \big[\, \|\mathbf{E}_\ell - \mathbf{I}_{2d}\|_{\mathrm{op}} + 1\big]\, \|\mathbf{S}_\ell - \mathbf{I}_{2d}\|_{\mathrm{op}} - 1 \\ &\leqslant \exp\left\{ \sum_{i=\ell}^k \eta_i \big(\varphi(\boldsymbol{\xi}_i) + \gamma + 1\big) \right\} - 1. \end{aligned}$$

The proof is complete. □

Next, we prove Lemmas 3.10 – 3.13. To make notations simple, we define

$$\mathbf{b}(\mathbf{y}) = \begin{bmatrix} -\gamma \mathbf{m} - \nabla f(\mathbf{x}) \\ \mathbf{m} \end{bmatrix}, \qquad \widetilde{\mathbf{b}}(\mathbf{y}, \boldsymbol{\xi}) = \begin{bmatrix} -\gamma \mathbf{m} - N^{-1}\sum_{i=1}^N \nabla F(\mathbf{x}, \xi^i) \\ \mathbf{m} \end{bmatrix}, \tag{B.33}$$

for $\mathbf{y} = (\mathbf{m}, \mathbf{x})$ and $\boldsymbol{\xi} = (\xi^1, \dots, \xi^N)$.

Proof of Lemma 3.10. (i) By the definition and (B.32),

$$\left\| \mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell, t_k} \right\|_{\mathcal{H}}^2 = \mathrm{tr}\left(\mathbf{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell, t_k}\big) \right) = \beta^2 \sum_{j=\ell+1}^k \eta_j \,\mathrm{tr}\left(\mathbf{S}_j^\top \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \end{bmatrix} \mathbf{S}_j \right) \leqslant 2d\beta^2 \sum_{j=\ell+1}^k \eta_j \|\mathbf{S}_j\|_{\mathrm{op}}^2.$$

Together with Hölder's inequality, we have

$$\left\| \mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell, t_k} \right\|_{\mathcal{H}}^8 \leqslant C d^4 (t_k - t_\ell)^3 \sum_{j=\ell+1}^k \eta_j \|\mathbf{S}_j\|_{\mathrm{op}}^8. \tag{B.34}$$

Notice that $\mathbf{S}_j = \mathbf{E}_{j+1} \dots \mathbf{E}_k$ with $\mathbf{E}_j$ defined in (B.30) implies

$$\mathbb{E}\, \|\mathbf{S}_j\|_{\mathrm{op}}^8 \leqslant \mathbb{E}\left\{ \mathbb{E}\left[\Big(1 + \|\mathbf{E}_k - \mathbf{I}_{2d}\|_{\mathrm{op}}\Big)^8 \middle| (\widetilde{\mathbf{X}}_{t_{i-1}}, \boldsymbol{\xi}_i)_{j+1\leqslant i\leqslant k-1} \right] \|\mathbf{E}_{j+1} \dots \mathbf{E}_{k-1}\|_{\mathrm{op}}^8 \right\},$$

and

$$\begin{aligned} \Big(1 + \|\mathbf{E}_k - \mathbf{I}_{2d}\|_{\mathrm{op}}\Big)^8 &\leqslant \left[1 + \eta_k \left(\frac{1}{N}\sum_{r=1}^N \sup_{\mathbf{x}\in\mathbb{R}^d} \left\| \nabla^2 F(\mathbf{x}, \xi_k^r) \right\|_{\mathrm{op}} + \gamma + 1 \right) \right]^8 \\ &\leqslant \frac{1}{1 - 7\eta_k} \left[1 + \eta_k \left(\frac{1}{N}\sum_{r=1}^N \sup_{\mathbf{x}\in\mathbb{R}^d} \left\| \nabla^2 F(\mathbf{x}, \xi_k^r) \right\|_{\mathrm{op}} + \gamma + 1 \right)^8 \right] \\ &\leqslant \left(1 + \frac{7\eta_k}{1 - 7\eta_1} \right) \left[1 + C\eta_k \left(\frac{1}{N}\sum_{r=1}^N \sup_{\mathbf{x}\in\mathbb{R}^d} \left\| \nabla^2 F(\mathbf{x}, \xi_k^r) \right\|_{\mathrm{op}}^8 + 1 \right) \right], \end{aligned}$$

where in the second inequality we use $(1+\eta_k x)^8 \leqslant 1+\eta_k x^8 + 7\eta_k(1+\eta_k x)^8$ for $x \geqslant 0$. According to Assumption IV , we have

$$\mathbb{E}\left\|\mathbf{S}_j\right\|_{\mathrm{op}}^8 \leqslant (1+C\eta_k)\mathbb{E}\left\|\mathbf{E}_{j+1}\ldots\mathbf{E}_{k-1}\right\|_{\mathrm{op}}^8 \leqslant \mathrm{e}^{C\eta_k}\mathbb{E}\left\|\mathbf{E}_{j+1}\ldots\mathbf{E}_{k-1}\right\|_{\mathrm{op}}^8.$$

Then, applying this method recursively shows that $\mathbb{E}\left\|\mathbf{S}_j\right\|_{\mathrm{op}}^8 \leqslant \mathrm{e}^{C(t_k-t_j)}$ for $j=\ell+1,\ldots,k$, so (B.34) and condition $t_k - t_\ell \leqslant 1/[5(B_1+\gamma+2)]$ imply that

$$\mathbb{E}\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right\|_{\mathcal{H}}^8 \leqslant Cd^4(t_k-t_\ell)^3\sum_{j=\ell+1}^{k}\eta_j\mathrm{e}^{C(t_k-t_\ell)} \leqslant Cd^4.$$

Analogously, it can be shown that $\mathbb{E}\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\|_{\mathcal{H}}^8 \leqslant Cd^4$.

For any $t\in[t_{k-1},t_k]$ and $\mathbf{h}\in\mathcal{H}$ with $\|\mathbf{h}\|_{\mathcal{H}}\leqslant 1$, by [Nor86]

$$\begin{aligned}\mathrm{D}^{\mathbf{h}}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) &= \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}} + \beta\int_{t_{k-1}}^{t}\begin{bmatrix}\mathbf{I}_d\\ \mathbf{0}\end{bmatrix}\mathbf{h}_s\,\mathrm{d}s\\ &\quad+\int_{t_{k-1}}^{t}\nabla\mathbf{b}(\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}))\,\mathrm{D}^{\mathbf{h}}\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\,\mathrm{d}u.\end{aligned} \tag{B.35}$$

Since $\|\nabla\mathbf{b}(\mathbf{y})\|_{\mathrm{op}}\leqslant L+\gamma+1$ where $\mathbf{b}(\mathbf{y})$ is in (B.33), it follows that

$$\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}} \leqslant \left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} + \beta\sqrt{\eta_k} + C\int_{t_{k-1}}^{t}\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}}\mathrm{d}u.$$

Then Grönwall's inequality implies

$$\mathbb{E}\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}}^8 \leqslant C\left(\mathbb{E}\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}}^8+1\right)\leqslant Cd^4,\quad t\in[t_{k-1},t_k].$$

(ii) Given event $(E^1_{\ell,k})^c$, i.e., $\sum_{i=\ell+1}^{k}\eta_i(\varphi(\boldsymbol{\xi}_i)-\mathbb{E}\varphi(\boldsymbol{\xi}_i))\leqslant t_k-t_\ell$, Lemma B.2 and Assumption IV derive the following equation for some constant $C>0$,

$$\left\|\mathbf{S}_j\right\|_{\mathrm{op}} \leqslant \left\|\mathbf{S}_j-\mathbf{I}_{2d}\right\|_{\mathrm{op}}+1 \leqslant \mathrm{e}^{\sum_{i=j+1}^{k}\eta_i(\varphi(\boldsymbol{\xi}_i)+\gamma+1)} \leqslant \mathrm{e}^{(t_k-t_\ell)(\mathbb{E}\varphi(\boldsymbol{\xi}_i)+\gamma+2)} \leqslant C,$$

for $j=\ell+1,\ldots,k$. According to (B.34), the following holds ,

$$\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right\|_{\mathcal{H}}^8 \leqslant Cd^4(t_k-t_\ell)^3\sum_{j=\ell+1}^{k}\eta_j\left\|\mathbf{S}_j\right\|_{\mathrm{op}}^8 \leqslant Cd^4 \quad\text{on}\quad (E^1_{\ell,k})^c.$$

The proof is complete. □

Proof of Lemma 3.11. For any $\mathbf{h}_1,\mathbf{h}_2\in\mathcal{H}$ with $\|\mathbf{h}_1\|_{\mathcal{H}},\|\mathbf{h}_2\|_{\mathcal{H}}\leqslant 1$, by [Nor86], we have

$$\begin{aligned}\mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k} &= \mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}} + \eta_k\nabla\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\,\mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\\ &\quad+\eta_k\nabla^2\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\,\mathrm{D}^{\mathbf{h}_1}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\,\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\end{aligned}$$

where $\widetilde{\mathbf{b}}(\mathbf{y},\boldsymbol{\xi})$ is defined in (B.33). Hence,

$$\begin{aligned}\left|\mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right| &\leqslant \left|\mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right| + \eta_k\left\|\nabla\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|_{\mathrm{op}}\left|\mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right|\\ &\quad+\eta_k\left\|\nabla^2\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|_{\mathrm{op}}\left|\mathrm{D}^{\mathbf{h}_1}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right|\left|\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right|,\end{aligned}$$

which implies that

$$\begin{aligned}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right\|_{\mathcal{H}\otimes\mathcal{H}} \leqslant{}& \left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}\otimes\mathcal{H}} + \eta_k \left\|\nabla\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|_{\mathrm{op}} \left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}\otimes\mathcal{H}} \\ &+ \eta_k \left\|\nabla^2\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|_{\mathrm{op}} \left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^2_{\mathcal{H}}.\end{aligned}$$

Notice that $\eta_k \leqslant \eta_1$ and $(x+\eta_k y)^4 \leqslant x^4 + \eta_k y^4 + 3\eta_k(x+\eta_k y)^4$ for $x, y \geqslant 0$, so

$$\begin{aligned}\mathbb{E}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right\|^4_{\mathcal{H}\otimes\mathcal{H}} \leqslant{}& \left(1+\frac{3\eta_k}{1-3\eta_1}\right)\mathbb{E}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}\otimes\mathcal{H}} \\ &+ \frac{8\eta_k}{1-3\eta_1}\left\{\mathbb{E}\left[\left\|\nabla\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|^4_{\mathrm{op}}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}\otimes\mathcal{H}}\right]\right. \\ &\left.+\mathbb{E}\left[\left\|\nabla^2\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|^4_{\mathrm{op}}\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^8_{\mathcal{H}}\right]\right\}.\end{aligned} \tag{B.36}$$

Combining Assumption IV, Lemma 3.10, and the independence between $\boldsymbol{\xi}_k$ and other random variables on the right-hand side, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla^2\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|^4_{\mathrm{op}}\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^8_{\mathcal{H}}\right] \\ &= \mathbb{E}\left\{\mathbb{E}_{\boldsymbol{\xi}_k}\left[\left\|\nabla^2\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|^4_{\mathrm{op}}\right]\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^8_{\mathcal{H}}\right\} \\ &\leqslant C\mathbb{E}\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^8_{\mathcal{H}} \leqslant Cd^4.\end{aligned} \tag{B.37}$$

Similarly, by Assumption IV, and the independence between $\boldsymbol{\xi}_k$ and other random variables, we also have

$$\mathbb{E}\left[\left\|\nabla\widetilde{\mathbf{b}}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k\big)\right\|^4_{\mathrm{op}}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}\otimes\mathcal{H}}\right] \leqslant C\mathbb{E}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}\otimes\mathcal{H}}. \tag{B.38}$$

Combining (B.36), (B.37) and (B.38), since $\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_\ell} = \mathbf{0}$, we have by an induction

$$\begin{aligned}\mathbb{E}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k}\right\|^4_{\mathcal{H}\otimes\mathcal{H}} \leqslant{}& (1+C\eta_k)\mathbb{E}\left\|\mathrm{D}^2\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}\otimes\mathcal{H}} + Cd^4\eta_k \\ \leqslant{}& Cd^4\sum_{i=\ell+1}^{k}\eta_i\prod_{j=i+1}^{k}(1+C\eta_j) \leqslant Cd^4(t_k-t_\ell)\mathrm{e}^{C(t_k-t_\ell)}.\end{aligned}$$

Due to $t_k - t_\ell \leqslant 1/[5(B_1+\gamma+1)]$, we can obtain the desired.

For $\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})$, $t \in [t_{k-1},t_k]$ and any $\mathbf{h}_i \in \mathcal{H}$ with $\|\mathbf{h}_i\|_{\mathcal{H}} \leqslant 1$, $i = 1,2$, by [Nor86], the second order Malliavin derivation satisfies

$$\begin{aligned}\mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) ={}& \mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}} \\ &+\int_{t_{k-1}}^{t}\nabla\mathbf{b}\big(\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\big)\,\mathrm{D}^{\mathbf{h}_1}\mathrm{D}^{\mathbf{h}_2}\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\,\mathrm{d}u \\ &+\int_{t_{k-1}}^{t}\nabla^2 b\big(\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\big)\,\mathrm{D}^{\mathbf{h}_1}\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\,\mathrm{D}^{\mathbf{h}_2}\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\,\mathrm{d}u.\end{aligned}$$

Since $\|\nabla \mathbf{b}(\mathbf{y})\|_{\mathrm{op}} \leqslant L+\gamma+1$ and $\|\nabla^2 b(\mathbf{y})\|_{\mathrm{op}} \leqslant B_2$ due to Assumption IV, we have

$$\left\|\mathrm{D}^2 \mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}\otimes\mathcal{H}} \leqslant \left\|\mathrm{D}^2 \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}\otimes\mathcal{H}} + C\int_{t_{k-1}}^{t} \left\|\mathrm{D}^2 \mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}\otimes\mathcal{H}} \mathrm{d}u + C\int_{t_{k-1}}^{t_k} \left\|\mathrm{D} \mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|^2_{\mathcal{H}} \mathrm{d}u.$$

By Grönwall's inequality,

$$\left\|\mathrm{D}^2 \mathbf{Y}_{t_{k-1},t_k}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}\otimes\mathcal{H}} \leqslant C\left\|\mathrm{D}^2 \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}\otimes\mathcal{H}} + C\int_{t_{k-1}}^{t_k} \left\|\mathrm{D} \mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|^2_{\mathcal{H}} \mathrm{d}u.$$

We have shown that $\mathbb{E}\|\mathrm{D}^2 \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\|^4_{\mathcal{H}\otimes\mathcal{H}} \leqslant Cd^4$, together with Lemma 3.10 and Hölder's inequality, we get

$$\mathbb{E}\left\|\mathrm{D}^2 \mathbf{Y}_{t_{k-1},t_k}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|^4_{\mathcal{H}\otimes\mathcal{H}} \leqslant Cd^4.$$

The proof is complete. □

Proof of Lemma 3.12. Recall the SDEs (1.5) and (1.6), and define

$$\boldsymbol{\Xi}_{t_\ell,t_k} := \mathbf{Y}_{t_{k-1},t_k}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k} = \int_{t_{k-1}}^{t_k} \left(\mathbf{b}(\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})) - \widetilde{\mathbf{b}}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}, \boldsymbol{\xi}_k)\right) \mathrm{d}t,$$

where functions $\mathbf{b}$ and $\widetilde{\mathbf{b}}$ are defined in (B.33). For any $\mathbf{h} \in \mathcal{H}$ with $\|\mathbf{h}\|_{\mathcal{H}} \leqslant 1$, by [Nor86], we have

(B.39)
$$\mathrm{D}^{\mathbf{h}}\boldsymbol{\Xi}_{t_\ell,t_k} = \int_{t_{k-1}}^{t_k} \left(\nabla \mathbf{b}(\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}))\, \mathrm{D}^{\mathbf{h}}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \nabla\widetilde{\mathbf{b}}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}, \boldsymbol{\xi}_k)\, \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right) \mathrm{d}t.$$

(i) Observe that $\mathrm{D}^{\mathbf{h}}\boldsymbol{\Xi}_{t_\ell,t_k} = \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3$, where

(B.40)
$$\begin{aligned}
\mathcal{I}_1 &= \int_{t_{k-1}}^{t_k} \nabla \mathbf{b}(\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})) \left(\mathrm{D}^{\mathbf{h}}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right) \mathrm{d}t,\\
\mathcal{I}_2 &= \int_{t_{k-1}}^{t_k} \left(\nabla \mathbf{b}(\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})) - \nabla \mathbf{b}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right) \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\, \mathrm{d}t,\\
\mathcal{I}_3 &= \eta_k \left(\nabla \mathbf{b}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \nabla \tilde{b}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}, \boldsymbol{\xi}_k)\right) \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}.
\end{aligned}$$

We estimate them one by one.

For $\mathcal{I}_1$, one has

$$\begin{aligned}
&\mathrm{D}^{\mathbf{h}}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\\
&= \beta\int_{t_{k-1}}^{t} \begin{bmatrix}\mathbf{I}_d\\ \mathbf{0}\end{bmatrix} \mathbf{h}_s\, \mathrm{d}s + \int_{t_{k-1}}^{t} \nabla \mathbf{b}(\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}))\, \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\, \mathrm{d}u\\
&\quad + \int_{t_{k-1}}^{t} \nabla \mathbf{b}(\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}))\left[\mathrm{D}^{\mathbf{h}}\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right] \mathrm{d}u.
\end{aligned}$$

Together with $\|\nabla \mathbf{b}(\mathbf{y})\|_{\mathrm{op}} \leqslant L+\gamma+1$, the following holds,

$$\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} \leqslant C\int_{t_{k-1}}^{t} \left\|\mathrm{D}\mathbf{Y}_{t_{k-1},u}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} \mathrm{d}u$$

$$+ C\sqrt{\eta_k}\left(\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} + 1\right),$$

for all $t \in [t_{k-1}, t_k]$. And Grönwall's inequality derives

$$\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} \leqslant C\sqrt{\eta_k}\left(\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} + 1\right). \tag{B.41}$$

which leads to

$$|\mathcal{I}_1|^4 \leqslant C\eta_k^6\left(\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}} + 1\right)\|\mathbf{h}\|^4_{\mathcal{H}}. \tag{B.42}$$

For $\mathcal{I}_2$, we let $\boldsymbol{\Delta} = \mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}$, then

$$\mathcal{I}_2 = \mathrm{D}^{\mathbf{h}}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\int_{t_{k-1}}^{t_k}\int_0^1 \nabla^2\mathbf{b}\big(r\boldsymbol{\Delta} + \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\big)\cdot\boldsymbol{\Delta}\,\mathrm{d}r\,\mathrm{d}t.$$

By $\|\nabla^2\mathbf{b}\|_{\mathrm{op}} \leqslant B_2$ due to Assumption IV, we have

$$|\mathcal{I}_2|^4 \leqslant C\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}}\eta_k^3\int_{t_{k-1}}^{t_k}\left|\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right|^4\mathrm{d}t\,\|\mathbf{h}\|^4_{\mathcal{H}}. \tag{B.43}$$

Besides, by Lemmas 3.2 and 3.4, we have for any $t \in [t_{k-1}, t_k]$,

$$\begin{aligned}\mathbb{E}\left[\left|\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}) - \widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right|^4\right] &\leqslant C(t-t_{k-1})^2\Big(\mathbb{E}\big[\mathcal{V}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})^2\big] + d^2\Big)\\ &\leqslant C\eta_k^2\big(\mathcal{V}(\mathbf{z})^2 + d^2\big).\end{aligned}$$

Then, by Markov property, it holds that

$$\mathbb{E}\,|\mathcal{I}_2|^4 \leqslant C\eta_k^6\big(\mathcal{V}(\mathbf{z})^2 + d^2\big)\,\|\mathbf{h}\|^4_{\mathcal{H}}\,\mathbb{E}\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}}. \tag{B.44}$$

For $\mathcal{I}_3$, since $\|\mathbf{A}\|_{\mathrm{op}} \leqslant \|\mathbf{A}\|_{\mathrm{HS}} \leqslant \sqrt{d}\,\|\mathbf{A}\|_{\mathrm{op}}$ for any matrix $\mathbf{A}$, it follows from Lemma B.1 that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi}}\left\|\nabla\mathbf{b}(\mathbf{y}) - \nabla\widetilde{\mathbf{b}}(\mathbf{y},\boldsymbol{\xi})\right\|^4_{\mathrm{op}} = \mathbb{E}_{\boldsymbol{\xi}}\left\|\frac{1}{N}\sum_{r=1}^N(\nabla^2F(\mathbf{x},\xi^r) - \nabla^2 f(\mathbf{x}))\right\|^4_{\mathrm{op}}\\ \leqslant\ &\frac{C}{N^2}\mathbb{E}_\xi\left\|\nabla^2F(\mathbf{x},\xi^r) - \nabla^2 f(\mathbf{x})\right\|^4_{\mathrm{HS}} \leqslant \frac{Cd^2}{N^2}.\end{aligned}$$

Due $\boldsymbol{\xi}$ is independent of other random variables, this immediately leads to

$$\mathbb{E}\,|\mathcal{I}_3|^4 \leqslant C\eta_k^4\,\|\mathbf{h}\|^4_{\mathcal{H}}\,\frac{d^2}{N^2}\mathbb{E}\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|^4_{\mathcal{H}}. \tag{B.45}$$

Since $\mathbb{E}\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\|^4_{\mathcal{H}} \leqslant Cd^2$ by Lemma 3.10, then Combining (B.40), (B.42), (B.44) and (B.45), we get

$$\mathbb{E}\,\|\mathrm{D}\boldsymbol{\Xi}_{t_\ell,t_k}\|^4_{\mathcal{H}} \leqslant Cd^2\eta_k^6(\mathcal{V}(\mathbf{z})^2 + d^2) + C\frac{d^4\eta_k^4}{N^2}.$$

(ii) According to (B.39), we have that

$$\begin{aligned}\|\mathrm{D}\boldsymbol{\Xi}_{t_\ell,t_k}\|_{\mathcal{H}} \leqslant &\int_{t_{k-1}}^{t_k}\|\nabla\mathbf{b}\|_{\mathrm{op}}\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}}\mathrm{d}t\\ &+\eta_k\left\|\nabla\widetilde{\mathbf{b}}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}},\boldsymbol{\xi}_k)\right\|_{\mathrm{op}}\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}}\end{aligned} \tag{B.46}$$

(B.41) together with a triangle inequality, it holds

$$\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}} - \left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} \leqslant C\sqrt{\eta_k}\left(\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} + 1\right)$$

for all $t \in [t_{k-1}, t_k]$. Besides, by Lemma 3.10, $\left\|\mathrm{D}\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}\right\|_{\mathcal{H}} \leqslant C\sqrt{d}$ on the event $(E^1_{\ell,k})^c$. Then, we have

$$\left\|\mathrm{D}\mathbf{Y}_{t_{k-1},t}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}})\right\|_{\mathcal{H}} \leqslant C\sqrt{d}, \quad \text{on the event} \quad (E^1_{\ell,k})^c. \tag{B.47}$$

On the other hand, given $(E^2_{\ell,k})^c$, i.e., $\sum_{i=\ell+1}^k \eta_i(\varphi(\boldsymbol{\xi}_i)^2 - \mathbb{E}[\varphi(\boldsymbol{\xi}_i)^2]) \leqslant t_k - t_\ell$ with $\varphi(\boldsymbol{\xi}_i)$ defined by (3.12), we have

$$\eta_k\varphi(\boldsymbol{\xi}_k)^2 \leqslant \sum_{i=\ell+1}^k \eta_i\varphi(\boldsymbol{\xi}_i)^2 \leqslant (t_k - t_\ell)(\mathbb{E}[\varphi(\boldsymbol{\xi}_k)^2] + 1) \leqslant C, \tag{B.48}$$

where the last inequality is due to Assumption IV.

Combining $\|\nabla b\|_{\mathrm{op}} \leqslant L + \gamma + 1$, (B.46), (B.47) and (B.48), we have

$$\begin{aligned}\|\mathrm{D}\boldsymbol{\Xi}_{t_\ell,t_k}\|_{\mathcal{H}} &\leqslant C\sqrt{d}\eta_k\left(\left\|\nabla\widetilde{\mathbf{b}}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_{k-1}}, \boldsymbol{\xi}_k)\right\|_{\mathrm{op}} + 1\right)\\ &\leqslant C\sqrt{d}\eta_k(\varphi(\boldsymbol{\xi}_k) + \gamma + 2) \leqslant C\sqrt{d\eta_k}, \quad \text{on the event} \quad \left(E^1_{\ell,k} \cup E^2_{\ell,k}\right)^c.\end{aligned}$$

The proof is complete. □

Proof of Lemma 3.13. For any vector $\mathbf{w} \in \mathbb{R}^{2d}$ with $|\mathbf{w}| = 1$, the relation (B.32) yields that

$$\mathbf{w}^\top \boldsymbol{\Gamma}(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell,t_k})\mathbf{w} = \beta^2 \sum_{j=\ell+1}^k \eta_j \left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix}\mathbf{S}_j\mathbf{w}\right|^2. \tag{B.49}$$

Inspired by [DP14], we divide $\mathbf{w}$ and $\mathbf{S}_j$ into

$$\mathbf{w} = \begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}, \qquad \mathbf{S}_j = \begin{bmatrix}\mathbf{A}_j & \mathbf{B}_j\\ \mathbf{C}_j & \mathbf{D}_j\end{bmatrix},$$

with vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ and matrices $\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j, \mathbf{D}_j \in \mathbb{R}^{d\times d}$.

We split the proof into two cases, one being $|\mathbf{u}| > (t_k - t_\ell)/4$ and the other $|\mathbf{u}| \leqslant (t_k - t_\ell)/4$.

(i) The case $|\mathbf{u}| > (t_k - t_\ell)/4$. Observe that

$$\begin{aligned}\left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix}\mathbf{S}_j\mathbf{w}\right|^2 &\geqslant \frac{1}{2}\left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix}\mathbf{w}\right|^2 - \left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix}(\mathbf{S}_j - \mathbf{I}_{2d})\mathbf{w}\right|^2\\ &\geqslant \frac{1}{2}|\mathbf{u}|^2 - \|\mathbf{S}_j - \mathbf{I}_{2d}\|^2_{\mathrm{op}}.\end{aligned} \tag{B.50}$$

Assumption IV and the definition of m yield that

$$(t_k - t_m)(\mathbb{E}\varphi(\boldsymbol{\xi}_k) + \gamma + 2) \leqslant \frac{t_k - t_\ell}{10}. \tag{B.51}$$

Given $(E^1_{m,k})^c$, i.e., $\sum_{i=m+1}^k \eta_i(\varphi(\boldsymbol{\xi}_i) - \mathbb{E}\varphi(\boldsymbol{\xi}_i)) \leqslant t_k - t_m$, we have

$$\sum_{i=m+1}^k \eta_i(\varphi(\boldsymbol{\xi}_i) + \gamma + 1) \leqslant (t_k - t_m)(\mathbb{E}\varphi(\boldsymbol{\xi}_k) + \gamma + 2). \tag{B.52}$$

Combining (B.51), (B.52) and Lemma B.2, we have that for $j = m, \dots, k$,

$$\|\mathbf{S}_j - \mathbf{I}_{2d}\|_{\mathrm{op}} \leqslant \mathrm{e}^{\sum_{i=m+1}^{k} \eta_i(\varphi(\boldsymbol{\xi}_i)+\gamma+1)} - 1 \leqslant \mathrm{e}^{(t_k - t_\ell)/10} - 1 \leqslant \frac{t_k - t_\ell}{8},$$

where the last inequality follows from $\mathrm{e}^{x/10} - 1 \leqslant x/8$ for $x \in (0, 1/10)$. Substituting this into (B.50) and $|\mathbf{u}| > (t_k - t_\ell)/4$ imply

$$\left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix} \mathbf{S}_j \mathbf{w}\right|^2 \geqslant \frac{(t_k - t_\ell)^2}{64}.$$

Then, due to $t_k - t_{m-1} \geqslant (t_k - t_\ell)/[10(\gamma + B_1 + 2)]$ by the definition of m, we obtain

$$\mathbf{w}^\top \boldsymbol{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell, t_k}\big)\mathbf{w} \geqslant \beta^2 \sum_{j=m}^{k} \eta_j \left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix} \mathbf{S}_j \mathbf{w}\right|^2 \geqslant C(t_k - t_\ell)^3, \text{ on } (E^1_{m,k})^c. \tag{B.53}$$

(ii) The case $|\mathbf{u}| \leqslant (t_k - t_\ell)/4$. In this case we clearly have $|\mathbf{v}| = (1 - |\mathbf{u}|^2)^{1/2} \geqslant 3/4$. Notice that $\mathbf{S}_{i-1} = \mathbf{E}_i \mathbf{S}_i$ implies

$$\begin{bmatrix}\mathbf{A}_{i-1} & \mathbf{B}_{i-1}\end{bmatrix} = \begin{bmatrix}(1 - \gamma\eta_i)\mathbf{I}_d & \eta_i \mathbf{I}_d\end{bmatrix} \begin{bmatrix}\mathbf{A}_i & \mathbf{B}_i \\ \mathbf{C}_i & \mathbf{D}_i\end{bmatrix},$$

i.e.

$$\begin{bmatrix}\mathbf{A}_{i-1} & \mathbf{B}_{i-1}\end{bmatrix} - \begin{bmatrix}\mathbf{A}_i & \mathbf{B}_i\end{bmatrix} = -\gamma\eta_i \begin{bmatrix}\mathbf{A}_i & \mathbf{B}_i\end{bmatrix} + \eta_i \begin{bmatrix}\mathbf{C}_i & \mathbf{D}_i\end{bmatrix}.$$

Then, summing this over i, we can obtain that

$$\begin{bmatrix}\mathbf{A}_j & \mathbf{B}_j\end{bmatrix} - \begin{bmatrix}\mathbf{A}_k & \mathbf{B}_k\end{bmatrix} = -\gamma \sum_{i=j+1}^{k} \eta_i \begin{bmatrix}\mathbf{A}_i & \mathbf{B}_i\end{bmatrix} + \sum_{i=j+1}^{k} \eta_i \begin{bmatrix}\mathbf{C}_i & \mathbf{D}_i\end{bmatrix}.$$

Due to $\mathbf{A}_k = \mathbf{I}_d$ and $\mathbf{B}_k = \mathbf{0}$, we get the following relation,

$$\mathbf{L}_j := \begin{bmatrix}\mathbf{A}_j & \mathbf{B}_j\end{bmatrix} + \gamma \sum_{i=j+1}^{k} \eta_i \begin{bmatrix}\mathbf{A}_i & \mathbf{B}_i\end{bmatrix} = \sum_{i=j+1}^{k} \eta_i \begin{bmatrix}\mathbf{C}_i & \mathbf{D}_i\end{bmatrix} + \begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix} =: \mathbf{R}_j, \tag{B.54}$$

where we denote the matrices on the left-hand side and right-hand side by $\mathbf{L}_j$ and $\mathbf{R}_j$ respectively.

For $\mathbf{L}_j$, observe that $[\mathbf{A}_j \quad \mathbf{B}_j] = [\mathbf{I}_d \quad \mathbf{0}]\mathbf{S}_j$, then we have

$$\begin{aligned}|\mathbf{L}_j \mathbf{w}|^2 &\leqslant 2\left|\begin{bmatrix}\mathbf{A}_j & \mathbf{B}_j\end{bmatrix} \mathbf{w}\right|^2 + 2\gamma^2 \left(\sum_{i=\ell+1}^{k} \eta_i \left|\begin{bmatrix}\mathbf{A}_i & \mathbf{B}_i\end{bmatrix} \mathbf{w}\right|\right)^2 \\ &\leqslant 2\left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix} \mathbf{S}_j \mathbf{w}\right|^2 + 2\gamma^2 (t_k - t_\ell) \sum_{i=\ell+1}^{k} \eta_i \left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix} \mathbf{S}_i \mathbf{w}\right|^2,\end{aligned}$$

which immediately implies that

$$\sum_{j=\ell+1}^{k} \eta_j |\mathbf{L}_j \mathbf{w}|^2 \leqslant 2[1 + \gamma^2 (t_k - t_\ell)^2] \sum_{j=\ell+1}^{k} \eta_j \left|\begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix} \mathbf{S}_j \mathbf{w}\right|^2. \tag{B.55}$$

For $\mathbf{R}_j$, observe that $[\mathbf{C}_j \quad \mathbf{D}_j] = [\mathbf{0} \quad \mathbf{I}_d]\mathbf{S}_j$, then

$$\mathbf{R}_j \mathbf{w} = \sum_{i=j+1}^{k} \eta_i \begin{bmatrix}\mathbf{0} & \mathbf{I}_d\end{bmatrix} \mathbf{w} + \sum_{i=j+1}^{k} \eta_i \left(\begin{bmatrix}\mathbf{C}_i & \mathbf{D}_i\end{bmatrix} - \begin{bmatrix}\mathbf{0} & \mathbf{I}_d\end{bmatrix}\right) \mathbf{w} + \begin{bmatrix}\mathbf{I}_d & \mathbf{0}\end{bmatrix} \mathbf{w}$$

$$= (t_k - t_j)\mathbf{v} + \sum_{i=j+1}^{k} \eta_i \left(\begin{bmatrix} \mathbf{C}_i & \mathbf{D}_i \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{I}_d \end{bmatrix} \right) \mathbf{w} + \mathbf{u}.$$

Given $(E^1_{\ell,k})^c$, together with Lemma B.2 implies

$$\begin{aligned} \left\| \begin{bmatrix} \mathbf{C}_i & \mathbf{D}_i \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{I}_d \end{bmatrix} \right\|_{\mathrm{op}} \leqslant \|\mathbf{S}_i - \mathbf{I}_{2d}\|_{\mathrm{op}} &\leqslant \mathrm{e}^{\sum_{i=\ell+1}^{k} \eta_i(\varphi(\boldsymbol{\xi}_i)+\gamma+1)} - 1 \\ &\leqslant \mathrm{e}^{(t_k - t_\ell)(\mathbb{E}\varphi(\boldsymbol{\xi}_k)+\gamma+2)} - 1 \leqslant \mathrm{e}^{1/5} - 1 \leqslant \frac{1}{4}. \end{aligned}$$

Together with $|\mathbf{u}| \leqslant (t_k - t_\ell)/4$ and $|\mathbf{v}| \geqslant 3/4$, we have

$$\begin{aligned} |\mathbf{R}_j \mathbf{w}| &\geqslant (t_k - t_j)\,|\mathbf{v}| - \sum_{i=j+1}^{k} \eta_i \left| \left(\begin{bmatrix} \mathbf{C}_i & \mathbf{D}_i \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{I}_d \end{bmatrix} \right) \mathbf{w} \right| - |\mathbf{u}| \\ &\geqslant \frac{3}{4}(t_k - t_j) - \sum_{i=j+1}^{k} \eta_i \left\| \begin{bmatrix} \mathbf{C}_i & \mathbf{D}_i \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{I}_d \end{bmatrix} \right\|_{\mathrm{op}} - \frac{1}{4}(t_k - t_\ell) \\ &\geqslant \frac{1}{2}(t_k - t_j) - \frac{1}{4}(t_k - t_\ell). \end{aligned}$$

Take m' such that $t_k - t_{m'} \geqslant 2(t_k - t_\ell)/3$ and $t_{m'} - t_\ell \geqslant (t_k - t_\ell)/4$, then

$$\begin{aligned} \sum_{j=\ell+1}^{k} \eta_j \,|\mathbf{R}_j \mathbf{w}|^2 &\geqslant \sum_{j=\ell+1}^{m'} \eta_j \,|\mathbf{R}_j \mathbf{w}|^2 \\ &\geqslant \left[\frac{1}{2}(t_k - t_{m'}) - \frac{1}{4}(t_k - t_\ell) \right]^2 \sum_{j=\ell+1}^{m'} \eta_j \geqslant \frac{(t_k - t_\ell)^3}{576}. \end{aligned} \tag{B.56}$$

Combining (B.54), (B.55) and (B.56) leads to

$$2[1 + \gamma^2 (t_k - t_\ell)^2] \sum_{j=\ell+1}^{k} \eta_j \left| \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \end{bmatrix} \mathbf{S}_j \mathbf{w} \right|^2 \geqslant \frac{(t_k - t_\ell)^3}{576}.$$

This, together with (B.49) and $t_k - t_\ell \leqslant 1/[5(B_1 + \gamma + 2)]$, immediately implies

$$\mathbf{w}^\top \boldsymbol{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell, t_k}\big)\mathbf{w} = \beta^2 \sum_{j=1}^{k} \eta_j \left| \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \end{bmatrix} \mathbf{S}_j \mathbf{w} \right|^2 \geqslant C(t_k - t_\ell)^3, \quad \text{on } (E^1_{\ell,k})^c. \tag{B.57}$$

Consequently, (B.53) and (B.57) yield the desired,

$$\lambda_{\min}\left(\boldsymbol{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell, t_k}\big) \right) = \min_{|\mathbf{w}|=1} \mathbf{w}^\top \boldsymbol{\Gamma}\big(\widetilde{\mathbf{Y}}^{\mathbf{z}}_{t_\ell, t_k}\big)\mathbf{w} \geqslant C(t_k - t_\ell)^3, \quad \text{on } \big(E^1_{\ell,k} \cup E^1_{m,k}\big)^c.$$

The proof is complete. □

(A. Guillin) Laboratoire de Mathematiques Blaise Pascal, CNRS-UMR 6620, Universite Clermont-Auvergne, avenue des landais, 63177, Aubiere, cedex
E-mail address: arnaud.guillin@uca.fr

(Y. Wang) 1. Department of Mathematics, Faculty of Science and Technology, University of Macau, Taipa, Macau, 999078, China; 2. UM Zhuhai Research Institute, Zhuhai, Guangdong, 519000, China.
E-mail address: yc17447@um.edu.mo

(L. Xu) Department of Mathematics, Faculty of Science and Technology, University of Macau, Macau S.A.R., China.
E-mail address: lihuxu@um.edu.mo

(H. Yang) 1. School of Mathematical Sciences, Peking University, Beijing, China. 2. Beijing International Center for Mathematical Research (BICMR), Peking University, Beijing, China.
E-mail address: yanghr@pku.edu.cn