Proceedings of Machine Learning Research vol 178:1–65, 2022 35th Annual Conference on Learning Theory

Dimension-free convergence rates for gradient Langevin dynamics in RKHS

Boris Muzellec[*] BORIS.MUZELLEC@GMAIL.COM
Département d'Informatique de l'École Normale Supérieure, 45 rue d'Ulm, F-75230 Paris, France

Kanji Sato DESCARTES329@GMAIL.COM
Department of Mathematical Informatics, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan

Mathurin Massias[*] MATHURIN.MASSIAS@GMAIL.COM
Univ. Lyon, INRIA, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342 Lyon, France

Taiji Suzuki[†] TAIJI@MIST.I.U-TOKYO.AC.JP
Department of Mathematical Informatics, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Gradient Langevin dynamics (GLD) and stochastic GLD (SGLD) have attracted considerable attention lately, as a way to provide convergence guarantees in a non-convex setting. However, the known rates grow exponentially with the dimension of the space under the dissipative condition. In this work, we provide a convergence analysis of GLD and SGLD when the optimization space is an infinite-dimensional Hilbert space. More precisely, we derive non-asymptotic, dimension-free convergence rates for GLD/SGLD when performing regularized non-convex optimization in a reproducing kernel Hilbert space. Amongst others, the convergence analysis relies on the properties of a stochastic differential equation, its discrete time Galerkin approximation and the geometric ergodicity of the associated Markov chains.

1. Introduction

Convex, finite-dimensional optimization problems have been studied at length, and there exists a variety of well-understood algorithms to solve them efficiently (Nesterov, 1983, 2004; Hiriart-Urruty and Lemaréchal, 1993; Boyd and Vandenberghe, 2004; Nocedal and Wright, 2006). In a non-convex optimization setting, however, these methods are only guaranteed to converge to stationary points of the objective function. This is to be contrasted with the ubiquity of non-convex optimization in machine learning applications, e.g., deep learning (Robbins and Monro, 1951; Duchi et al., 2011; Zeiler, 2012; Kingma and Ba, 2014), tensor factorization (Signoretto et al., 2013; Suzuki et al., 2016), Bayesian optimization (Vien et al., 2018; Vellanki et al., 2019), and non-convex loss learning such as robust classification (Masnadi-Shirazi and Vasconcelos, 2009). In a different perspective, *stochastic gradient Langevin dynamics (SGLD)*, which can be seen as stochastic gradient descent

[*] This work was performed while BM and MM were interning at AIP-RIKEN.
[†] TS is also affilated to AIP-RIKEN, Tokyo, Japan.

methods with additive Gaussian noise injection at each iteration, was introduced by Welling and Teh (2011). In the case of a strongly convex objective function $\mathcal{L}$, recent studies (Dalalyan, 2017b) highlighted the connections between sampling from log-concave densities $f(x) \propto \exp(-\beta\mathcal{L}(x))$ concentrated around the minimum of $\mathcal{L}$, and minimizing $\mathcal{L}$. Such distributions can be obtained as the stationary distributions of a first order Langevin dynamics. Chiang et al. (1987); Gelfand and Mitter (1991); Roberts and Tweedie (1996) studied the convergence of the dynamics to the stationary Gibbs distribution, and the concentration of the samples around the global minimum, while more recently Dalalyan (2017a); Durmus and Moulines (2016, 2017) analyzed the convergence rates of discrete time Langevin updates for sampling from log-concave densities.

Recent studies have shown that Langevin-dynamics-based algorithms converge near a global minimum of $\mathcal{L}$, even when $\mathcal{L}$ is not convex (Raginsky et al., 2017; Xu et al., 2018; Erdogdu et al., 2018; Vempala and Wibisono, 2019; Nagapetyan et al., 2017; Duncan et al., 2017). The analysis relies on the connection between the iterates of Langevin dynamics based algorithms and the Markov chain solution of the continuous time Langevin equation, which admits the Gibbs measure as invariant distribution. Raginsky et al. (2017) provided a non-asymptotic convergence rate in expectation to an *almost minimizer* of SGLD. Xu et al. (2018) improved the convergence rate while also providing an extension to variance-reduced algorithms. In an alternative approach, Zhang et al. (2017) provided bounds on the *hitting time* of SGLD to neighborhoods of local minima. However, these results only apply to finite-dimensional optimization, with rates growing exponentially with the dimension under the dissipative condition. This is quite problematic for optimizing high dimensional models such as deep learning networks that frequently appear in machine learning.

In this paper, we resolve this problem by extending Langevin dynamics algorithms to the *infinite-dimensional* setting and study their convergence rates. Our results rely on assumptions that are classical in the GLD/SGLD literature, and in the literature of approximation of invariant laws of stochastic partial differential equations (SPDE) in infinite dimension. In particular, we leverage the weak approximation error of the discrete time scheme of SPDEs analyzed by Bréhier (2014); Bréhier and Kopec (2016) for general inverse parameter $\beta > 1$, where Debussche (2011); Wang and Gan (2013); Andersson and Larsson (2016) gave discretization error non-uniformly over the time horizon, and utilize the *geometric ergodicity*[1] of continuous time dynamics (Jacquot and Royer, 1995; Goldys and Maslowski, 2006). Results in the infinite-dimensional setting usually involve a linear operator acting as a regularizer and whose spectrum "replaces" dimension in the convergence rates. More specifically, our contributions can be summarized as follows:

- We give a non-asymptotic error bound of the infinite-dimensional GLD/SGLD implemented with a spectral Galerkin method, which has an explicit dependency on the inverse temperature β and is uniform over all time horizons.
- For that purpose, the geometric ergodicity of the time-discretized dynamics is proven, which is known to be non-trivial. Besides this, we also give a bound on the discrepancy between continuous and discrete time dynamics that is optimal with respect to the step size.
- We give an upper bound of the distance between the expected objective value under the invariant measure and the global optimal solution in the infinite-dimensional setting.

1. The term "*geometric ergodicity*" means exponential convergence to its stationary distribution Kendall (1959)

2. Notation and Framework

2.1. Notation and background on RKHS

Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a Hilbert space. We will also use the notation $\|\cdot\|_{\mathcal{H}}$ to explicitly indicate the norm $\|\cdot\|$ is of $\mathcal{H}$. C_b^2 is the set of bounded, twice continuously Fréchet differentiable functions with bounded first and second derivatives. We denote by $\mathcal{B}(\mathcal{H})$ the set of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$ and $\|\cdot\|_{\mathcal{B}(\mathcal{H})}$ denotes the operator norm. For a discrete or continuous Markov chain $\{X_t\}$, note $\mathbb{E}_x[\cdot] \triangleq \mathbb{E}[\cdot \mid X_0 = x]$. To consider a "regularization" in the space $\mathcal{H}$, we define a subspace $\mathcal{H}_K$ of $\mathcal{H}$ as

$$\mathcal{H}_K = \left\{ \textstyle\sum_{k=0}^{\infty} \alpha_k f_k \,:\, \sum_{k=0}^{\infty} \frac{\alpha_k^2}{\mu_k} < \infty \right\}, \tag{1}$$

where $(f_k)_{k=0}^{\infty}$ is a complete orthonormal system in $\mathcal{H}$ and $(\mu_k)_{k=0}^{\infty}$ is a sequence of non-negative reals in decreasing order. We equip $\mathcal{H}_K$ with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ defined as $\langle f, g\rangle_{\mathcal{H}_K} = \sum_{k=0}^{\infty} \frac{\alpha_k \beta_k}{\mu_k}$ for $f = \sum_{k\geq 0} \alpha_k f_k \in \mathcal{H}_K$ and $g = \sum_{k\geq 0} \beta_k f_k \in \mathcal{H}_K$, while the inner product in $\mathcal{H}$ can be expressed by $\langle f, g\rangle = \sum_{k=0}^{\infty} \alpha_k \beta_k$. Accordingly, we define the norm $\|\cdot\|_{\mathcal{H}_K}$ as the one induced by the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$. As an important example of $\mathcal{H}_K$, we suppose a Reproducing Kernel Hilbert Space (RKHS), with a reproducing kernel K. Suppose that $\mathcal{H}$ is the Hilbert space of L^2-integrable functions with respect to a measure ρ. Then, we can define the integral operator with the kernel K as $T_K f(x) \triangleq \int K(x,y) f(y)\, \mathrm{d}\rho(y)$ for $f \in \mathcal{H}$, and the RKHS corresponding to the kernel K can be written as $\mathcal{H}_K = T_K^{1/2} L^2(\rho)$ (Caponnetto and De Vito, 2007; Steinwart and Christmann, 2008). Actually, it is known that, if $(\mu_k, f_k)_{k=0}^{\infty}$ are the eigenvalue-eigenfunction pairs of T_K (i.e., $T_K f_k = \mu_k f_k$), then the RKHS $\mathcal{H}_K$ defined in this way is expressed as in Eq. (1). In this sense, we say "RKHS" to indicate $\mathcal{H}_K$ in this paper, but we note that our analysis covers more general situations than the usual RKHS setting.

In the following, for $L : \mathcal{H} \to \mathbb{R}$, the gradient $\nabla L(x)$ is defined as the Riesz representor of the Fréchet derivative of L, $DL(x)$ (i.e., the unique vector satisfying $\forall h, L(x+h) = L(x) + \langle \nabla L(x), h\rangle + O(\|h\|^2)$). We will identify n-order derivatives with nth-linear forms, and with vectors when there is no ambiguity (e.g., we write $D^3 L(x)\cdot(h,k)$ for the Riesz representor of $l \in \mathcal{H} \mapsto D^3 L(x)\cdot(h,k,l)$).

2.2. Algorithm: gradient Langevin dynamics

We consider the following optimization problem:

$$\min_{x\in\mathcal{H}} \mathcal{L}(x) \triangleq L(x) + \frac{\lambda}{2}\|x\|_{\mathcal{H}_K}^2, \tag{2}$$

where $\lambda > 0$ and L is potentially non-convex. Assuming L admits at least one global minimizer, we note $x^* \triangleq \arg\min_{x\in\mathcal{H}} L(x)$, $\tilde{x} \triangleq \arg\min_{x\in\mathcal{H}} L(x) + \frac{\lambda}{2}\|x\|_{\mathcal{H}_K}^2$. The difference between the expected loss of the two optimal solutions, $L(x^*)$ and $L(\tilde{x})$, has been extensively studied, for example, in least squares estimation in RKHS (Caponnetto and De Vito, 2007).

We study the gradient Langevin dynamics (GLD) iterations to solve Problem (2). To define GLD, we need to make a heavy use of the infinite-dimensional Brownian motion.

Definition 1 (Cylindrical Brownian motion/Wiener process (Da Prato and Zabczyk, 1996)) *Given*

- *a complete orthonormal system of $\mathcal{H}$, $(f_i)_{i\in I}$, where $I \subset \mathbb{N}$,*

- *a family* $(\{W^i(t)\}_{t\geq 0})_{i\in I}$ *of independent real Brownian motions,*

then $\{W(t)\}_{t\geq 0} \triangleq \{\sum_{i\in I} W^i(t) f_i\}_{t\geq 0}$ *is called a* cylindrical Brownian motion.

Then, GLD updates are defined as follows: $X_0 = x_0 \in \mathcal{H}$, and

$$X_{n+1} = S_\eta X_n - \eta S_\eta \nabla L(X_n) + \sqrt{2\tfrac{\eta}{\beta}} S_\eta \varepsilon_n, \tag{3}$$

where $\eta > 0$ is the stepsize, $\beta \geq \eta$ is the inverse temperature parameter, the variables ε_n are i.i.d. cylindrical standard Gaussian (i.e., $\varepsilon_n = \sum_{k=0}^\infty \varepsilon_{n,k} f_k$ for $\varepsilon_{n,k} \sim N(0,1)$ i.i.d.) and $S_\eta \triangleq (\mathrm{Id} + \eta\tfrac{\lambda}{2}\nabla \|\cdot\|^2_{\mathcal{H}_K})^{-1}$ (i.e., $S_\eta x = \sum_{k=0}^\infty x_k/(1+\eta\lambda\mu_k^{-1})$ for $x = \sum_{k=0}^\infty x_k f_k$). Here, note that ε_n is not included in $\mathcal{H}_K$, but by applying S_η it is pushed back to $\mathcal{H}_K$ under Assumption 1 that we will mention later. This can be seen by noticing $\|S_\eta \varepsilon_n\|^2_{\mathcal{H}_K} = \sum_{k=0}^\infty \mu_k^{-1} \varepsilon_{n,k}^2/(1+\eta\lambda\mu_k^{-1})^2 \leq \sum_{k=0}^\infty \varepsilon_{n,k}^2/[\eta\lambda(1+\eta\lambda\mu_k^{-1})] \lesssim \sum_{k=0}^\infty \varepsilon_{n,k}^2/[(\eta\lambda)^2 k^2] = O_p(1)$. A crucial analysis tool is to see Eq. (3) as a time discretization of the following SPDE (Da Prato and Zabczyk, 1996): $X(0) = x_0$, and

$$\begin{aligned} \mathrm{d}X(t) &= -\nabla \mathcal{L}(X(t)) + \sqrt{\tfrac{2}{\beta}}\,\mathrm{d}W(t) \\ &= -\nabla \left(L(X(t)) + \tfrac{\lambda}{2}\|X(t)\|^2_{\mathcal{H}_K} \right) + \sqrt{\tfrac{2}{\beta}}\,\mathrm{d}W(t), \end{aligned} \tag{4}$$

where $\{W(t)\}_{t\geq 0}$ is a cylindrical Brownian motion (Definition 1). Although the cylindrical standard Gaussian variable ε_n and Brownian motion $W(t)$ are not included in $\mathcal{H}$ a.s., the dynamics is pushed back into $\mathcal{H}$ thanks to the existence of the regularization term. We refer to Da Prato and Zabczyk (1996) for the existence of solutions, its regularity conditions and related mathematical details. Note that the scheme Eq. (3) is semi-implicit: applying $(S_\eta)^{-1}$ to both terms yields

$$X_{n+1} = X_n - \eta(\nabla L(X_n) + \tfrac{\lambda}{2}\nabla \|X_{n+1}\|^2_{\mathcal{H}_K}) + \sqrt{2\tfrac{\eta}{\beta}}\varepsilon_n.$$

Approximated computation. Strictly speaking, the infinite-dimensional GLD scheme presented above is computationally intractable. The *Galerkin approximation method* projects the dynamics to a finite-dimensional subspace to make them computationally feasible. Let $\mathcal{H}_N$ be an $N+1$-dimensional subspace of $\mathcal{H}$ that is spanned by $(f_k)_{k=0}^N$: $\mathcal{H}_N \triangleq \mathrm{Span}\{f_k \mid k = 0, \dots, N\}$. Let $P_N : \mathcal{H} \to \mathcal{H}_N$ be the orthogonal projection operator onto $\mathcal{H}_N$: $P_N(\sum_{k=0}^\infty \alpha_k f_k) = \sum_{k=0}^N \alpha_k f_k$. Then, the GLD with Galerkin approximation can be formulated as

$$X^N_{n+1} = S_\eta \left(X^N_n - \eta \nabla L_N(X^N_n) + \sqrt{2\tfrac{\eta}{\beta}} P_N \varepsilon_n \right), \tag{5}$$

where $X^N_0 = P_N x_0 \in \mathcal{H}_N$ and $\nabla L_N(x) \triangleq P_N(\nabla L(P_N x))$. Since this scheme is essentially finite-dimensional, it can be implemented in practice.

Next, we consider a stochastic gradient variant of GLD (stochastic GLD; SGLD). Let us consider a finite sum risk minimization setting where $L(x) = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \ell_i(x)$ for $\ell_i : \mathcal{H} \to \mathbb{R}$ which is Fréchet differentiable[2]. SGLD makes use of a mini-batch of stochastic gradients (Welling and Teh,

2. We may generalize the setting to a situation where $\nabla L(x) = \mathbb{E}_\xi[g(x,\xi)]$ with a stochastic gradient $g(\cdot,\xi)$ in a straightforward way.

2011) instead of the full gradient $\nabla L(x)$: $g_n(x) = \frac{1}{n_b}\sum_{i\in I_n}\nabla \ell_i(x)$ where I_n is a random subset of $\{1,\dots,N\}$ chosen uniformly at random and $n_b = |I_n|$. Then, its update rule is given by

$$Y_{n+1}^N = S_\eta\left(Y_n^N - \eta g_{n,N}(Y_n^N) + \sqrt{2\tfrac{\eta}{\beta}}P_N\varepsilon_n\right), \tag{6}$$

where $g_{n,N}(x) \triangleq P_N(g_n(P_N x))$ and $Y_0^N = P_N x_0 \in \mathcal{H}_N$. These approximation techniques significantly reduce the computational cost.

2.3. Assumptions

Our goal is to study the convergence of the iterations Eq. (3), i.e., to bound $L(X_n) - L(x^*)$ with high probability. For this, we need to make assumptions on the RKHS $\mathcal{H}_K$ and on L. We first make the following assumption on $\mathcal{H}_K$, independently of the objective L:

Assumption 1 *There exists a constant $C_K > 0$ such that $\mu_k \le C_K/(k+1)^2$ $(\forall k)$.*

We note that a finite-dimensional situation is also allowed, i.e., $\mu_k = 0$ $(\forall k \ge k_0)$ for some $k_0 \in \mathbb{N}$, as long as Assumption 1 is satisfied for $k \le k_0$. The weaker assumption $\mu_k \sim k^{-p}$ with $p > 1$ is sometimes made in the literature (Caponnetto and De Vito (2007, Definition 1. *iii*)), and Steinwart and Christmann (2008)). While we focus on the specific $p = 2$ setting for technical simplicity, our analysis also applies to the more general setting of Andersson et al. (2016). To deal with more general settings, one can consider the case where $\mathcal{H}$ itself is an RKHS for a kernel K', with Mercer decomposition $K'(x,y) = \sum_k \nu_k g_k(x) g_k(y)$. Then, the "rescaled" kernel $K(x,y) = \sum_k \mu_k \nu_k g_k(x) g_k(y)$ with $\mu_k \sim \frac{1}{k^2}$ satisfies Assumption 1.

Next, we put assumptions on the objective function L. The first one is classical for gradient-based optimization (Nesterov, 2004).

Assumption 2 (Smoothness) *L is M-smooth: $\forall x, y \in \mathcal{H}$, $\|\nabla L(x) - \nabla L(y)\| \le M\|x-y\|$.*

In view of Eq. (1), we have that $A \triangleq -\frac{\lambda}{2}\nabla\|\cdot\|_{\mathcal{H}_K}^2$ is a diagonal operator, characterized by $Af_k = -\frac{\lambda}{\mu_k}f_k$. The following assumptions enforce more smoothness on L w.r.t. a norm induced by A through its second and third order derivatives.

Assumption 3 *There exists $\alpha \in (1/4, 1)$ and $\lambda_0, C_{\alpha,2} \in (0,\infty)$ such that $\forall x, h, k \in \mathcal{H}$, $|D^2L(x)\cdot(h,k)| \le C_{\alpha,2}\|h\|_{\mathcal{H}}\|k\|_\alpha$, where $\|x\|_\varepsilon \triangleq \left(\sum_{k\ge 0}(\mu_k)^{2\varepsilon}|\langle x, f_k\rangle|^2\right)^{1/2}$.*

This assumption is not standard in the previous works. However, we put this assumption so that the time-discretized dynamics satisfies geometric ergodicity. Fortunately, this assumption is not restrictive in machine learning applications (see the discussion just after Assumption 1 and Section 2.4 for details). The next one is common in the SPDE discretization literature (Bréhier and Kopec (2016, Assumption 2.7), Debussche (2011, Assumption (2.3))). It is used in Section 3.1.2 to obtain the convergence of the stationary distribution μ^η of the discrete time dynamics (3) to that of the continuous time one (4) as η goes to zero.

Assumption 4 (Bréhier and Kopec (2016, Assumption 2.7)) *Let $L_N : \mathcal{H}_N \to \mathbb{R}$, $L_N = L(P_N x)$. L is three times differentiable, and there exists $\alpha' \in [0,1)$, $C_{\alpha'} \in (0,\infty)$ such that for all $N \in \mathbb{N}$ and $\forall x, h, k \in \mathcal{H}_N$, $\left\|D^3L_N(x)\cdot(h,k)\right\|_{\alpha'} \le C_{\alpha'}\|h\|_0\|k\|_0$ and $\left\|D^3L_N(x)\cdot(h,k)\right\|_0 \le C_{\alpha'}\|h\|_{-\alpha'}\|k\|_0$ hold.*

As an example, Assumption 4 is satisfied with $\alpha = 0$ when L is C^3 with bounded second and third-order derivatives. Next, we assume the following condition to ensure the dissipativity (Proposition 2) which is essential to show geometric ergodicity.

Assumption 5 *One of the following two conditions holds:*

i) (Strict Dissipativity) $\lambda > M\mu_0$, *or*

ii) (Bounded gradients) $\|\nabla L(\cdot)\| \leq B$ *for a constant* $B > 0$.

Under Assumption 5 (i), the objective function becomes convex because the regularization term is sufficiently strong, which induces faster convergence. On the other hand, under Assumption 5 (ii), the lower bound of λ is no longer imposed so that the objective function is not necessarily convex, but the boundedness of the gradient is assumed instead to ensure convergence. The C_0-semigroup $(S_t)_{t\geq 0}$ generated by A is the one of diagonal operators determined by $S_t f_k = e^{-\lambda t/\mu_k} f_k$. It is easy to check that this semigroup is strongly continuous. Therefore, the Langevin SDE (4) is an instance of the more general semilinear SDE:

$$\mathrm{d}X(t) = \Big(AX(t) + F(X(t))\Big)\,\mathrm{d}t + \sqrt{Q}\,\mathrm{d}W(t), \tag{7}$$

where F is globally M-Lipschitz, Q is bounded and symmetrical and A is a linear unbounded operator on $\mathcal{H}$ generating a strongly continuous semigroup (Da Prato and Zabczyk, 1996; Bréhier, 2014; Bréhier and Kopec, 2016). For the SDE (4), we have $F = -\nabla L$, $Q = 2\beta^{-1}\,\mathrm{Id}$ and $A = -\frac{\lambda}{2}\nabla\|\cdot\|^2_{\mathcal{H}_K}$. The SDE (7) has been extensively studied in finite dimension (Khasminskii, 2011); in the infinite-dimensional case, several results have been shown such as the existence and uniqueness of its invariant measure (Da Prato and Zabczyk, 1992; Maslowski, 1989; Sowers, 1992), the exponential convergence of the time t distribution to this invariant measure (Jacquot and Royer, 1995; Shardlow, 1999; Hairer, 2002) and its explicit convergence rate evaluation (Goldys and Maslowski, 2006); the invariant measure π is given by

$$\frac{\mathrm{d}\pi}{\mathrm{d}\nu_\beta}(x) \propto \exp(-\beta L(x)),$$

where ν_β is the Gaussian measure in $\mathcal{H}$ with mean 0 and covariance $(-\beta A)^{-1}$ (see Da Prato and Zabczyk (1996) for the precise definition of infinite-dimensional Gaussian measures). If these assumptions are verified, we have a weaker condition than strong convexity: dissipativity (Hale, 1988).

Proposition 2 (Dissipativity (Hale, 1988)) *Under Assumptions 1, 2 and 5 there exist constants* $m, c > 0$ *verifying*

$$\forall x \in \mathcal{H}, \langle Ax - \nabla L(x), x\rangle \leq -m\,\|x\|^2 + c. \tag{8}$$

The dissipative condition proved in this proposition is quite standard to show the existence of the invariant law. For example, Raginsky et al. (2017); Xu et al. (2018) showed the convergence to the invariant law under the dissipative condition in the finite-dimensional situation. This condition intuitively indicates that the dynamics stays inside a bounded domain with high probability. If X_n (or $X(t)$) is far away from the origin, then the dynamics is forced to get back around the origin. Thanks to this condition, the dynamics can possess finite moments, which is important to ensure the existence of an invariant law. In fact, Assumption 5 ensures existence of a invariant law.

Proposition 3 *Under Assumption 5, the processes* $\{X(t)\}_{t\geq 0}$ *and* $\{X_n\}_{n\in\mathbb{N}_+}$ *admit (at least) an invariant law.*

The proof can be found for example in Proposition 4.1 of Bréhier and Kopec (2016), which utilizes the Krylov-Bogoliubov criterion (Da Prato and Zabczyk, 1996, Section 3.1). This proposition does not indicates the *uniqueness* of an invariant law. However, Bréhier and Kopec (2016) also showed that the continuous time dynamics $X(t)$ has a unique invariant law and is geometrically ergodic. As for the discrete time dynamics X_n, the uniqueness of the invariant law is already well-known under the strict dissipative condition (Assumption 5 (i)) (see Bréhier and Kopec (2016) for example). However, the uniqueness has not been shown under the bounded gradient condition (Assumption 5 (ii)). In Section 3.1.1, we will show that the uniqueness also holds under Assumption 5 (ii) if we assume Assumption 3, which has not been assumed in previous work.

Finally, in the SGLD setting we put the following stronger assumption on each ℓ_i.

Assumption 6 *Each* ℓ_i *satisfies Assumptions 2 to 4 and Assumption 5 (ii) instead of* L*, where the constants in each assumption are uniform over all* ℓ_i $(i = 1, \dots, n_{\mathrm{tr}})$.

2.4. Motivating examples

There are several machine learning problems in which non-convex optimization on a high/infinite-dimensional Hilbert space is required. Such examples include deep learning, tensor factorization (Signoretto et al., 2013; Suzuki et al., 2016), robust classification using non-convex losses such as Savage (Masnadi-Shirazi and Vasconcelos, 2009), Bayesian optimization on function space (Vien et al., 2018; Vellanki et al., 2019), and any other kernel method with non-convex loss. For the sake of instructive exposition, let us consider a situation where we observe n_{tr} input-output pairs $(z_i, y_i)_{i=1}^{n_{\mathrm{tr}}}$, where $z_i \in \mathcal{Z}$ is an input and $y_i \in \mathcal{Y}$ is its label. Accordingly, we define a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and an empirical risk: $\tilde{L}(f) = \frac{1}{n}\sum_{i=1}^{n_{\mathrm{tr}}} \ell(f(z_i), y_i)$ for a function $f : \mathcal{Z} \to \mathbb{R}$.

(i) Neural network: Consider a neural network $f_W(z) = \sum_{m=1}^{M} a_m \sigma(w_m^\top z)$ where σ is the sigmoid function, the width M can be either finite or infinite, $W = (w_m)_{m=1}^{M} \subset \mathbb{R}^d$ is the learnable parameter, and $(a_m)_{m=1}^{M} \subset \mathbb{R}$ is a fixed parameter. Then, by considering $x = [w_1^\top, w_2^\top, \dots]^\top$ as an element of a Hilbert space, optimizing the parameter W falls into our setting if $|a_m| = o(m^{-1/2})$: $\min_W \frac{1}{n}\sum_{i=1}^{n} \ell(f_W(z_i), y_i) + \frac{\lambda}{2}\sum_m \mu_m^{-1}\|w_m\|^2$ where ℓ is a smooth loss function. Note that the width M can be arbitrary (infinite/finite), which is quite different from typical analysis of neural network optimization such as mean field theory (Sirignano and Spiliopoulos, 2018; Mei et al., 2018; Nitanda and Suzuki, 2017; Chizat and Bach, 2018). See also Suzuki (2020); Suzuki and Akiyama (2021) for statistical analyses of neural networks optimized by the infinite dimensional Langevin dynamics presented in this paper.

(ii) Tensor decomposition: Signoretto et al. (2013); Suzuki et al. (2016) considered a nonparametric low-rank tensor model which is given as $f(x) = \sum_{r=1}^{R}\prod_{k=1}^{K} f_{r,k}(x_k)$ where $f_{r,k} \in \mathcal{H}_{K_k}$ is included in an RKHS $\mathcal{H}_{K_k}$. Fitting f to a training data by minimizing an empirical risk is not a convex optimization problem but falls into our setting where $\mathcal{H}_K = \mathcal{H}_{K_1} \oplus \cdots \oplus \mathcal{H}_{K_K}$.

(iii) General formulation: Here, we let $\mathcal{H}$ be a Hilbert space of functions on $\mathcal{Z}$ (which could be an RKHS) with complete orthonormal system $(f_k)_{k=0}^{\infty}$. From the expression (1), the (sub-)RKHS $\mathcal{H}_K$ can be expressed as an image of $T_K^{1/2}$, i.e., $\mathcal{H}_K = \{f = T_K^{1/2} h \mid h \in \mathcal{H}\}$ and $\|f\|_{\mathcal{H}_K} = \inf_{h\in\mathcal{H}: f = T_K^{1/2}h} \|h\|_{\mathcal{H}}$. More generally, we define an RKHS $\mathcal{H}_{K^\gamma}$ for $0 < \gamma$ as an image of $T_K^{\frac{\gamma}{2}}$:

$\mathcal{H}_{K^\gamma} = \{f = T_K^{\gamma/2} h \mid h \in \mathcal{H}\}$. We see that $\gamma = 1$ corresponds to $\mathcal{H}_K$. We employ $\mathcal{H}_{K^\gamma}$ as a model for f and let the corresponding empirical risk be $L(x) = \tilde{L}(T_K^{\frac{\gamma}{2}} x)$ (if needed, we may add a smooth regularization term). In this situation, if we have $\max_i \sup_u |\ell_i''(u)| \leq G$ and $\sup_{z \in \mathcal{Z}} K_\gamma(z,z) \leq R_\gamma$ for $G, R_\gamma > 0$, then

$$\|\nabla L(x) - \nabla L(x')\| \leq G R_\gamma \|x - x'\|_{\mathcal{H}}, \quad |D^2 L(x) \cdot (h,k)| \leq G \sqrt{R_\gamma \sum_{k=0}^{\infty} \mu_k^{\gamma - 2\alpha}}, \tag{9}$$

for $x, h, k \in \mathcal{H}$ with $\|h\| = 1$ and $\|k\|_\alpha = 1$. The proof of these inequalities is given in Appendix A. Therefore, Assumptions 2 and 3 are satisfied as long as $R_\gamma < \infty$ for $\gamma > 1$ because the condition $\mu_k \lesssim 1/k^2$ makes the right hand of Eq. (9) finite by setting $\alpha = (\gamma - 1)/2 + 1/4 > 1/4$. Assumption 4 is also verified in the same manner. Finally, if we let $f = T_K^{\frac{\gamma}{2}} x$, then $\|x\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_{K^{1+\gamma}}}$ holds, and thus it follows that

$$L(x) + \lambda \|x\|_{\mathcal{H}_K}^2 = \tilde{L}(f) + \lambda \|f\|_{\mathcal{H}_{K^{1+\gamma}}}^2.$$

Therefore, we see that our formulation covers a wide range of kernel regularization learning by adjusting γ appropriately. We would like to remark that we may deal with a situation where $L(x)$ contains a regularization term $\frac{\lambda_0}{2}\|x\|^2$ like $L(x) = \hat{L}(x) + \frac{\lambda_0}{2}\|x\|^2$. See Section A.1 for more details about this issue.

3. Main Result

Here, we give our main result on the the non-asymptotic error bound of the GLD algorithm. Define a constant $\hat{c}_\beta$ as $\hat{c}_\beta = 1$ under Assumption 5 (i) and $\hat{c}_\beta = \sqrt{\beta}$ under Assumption 5 (ii).

Theorem 4 (Main Result, GLD convergence rate) *Let Assumptions 1, 2, 4 and 5 hold. If only the bounded gradient condition (Assumption 5 (ii)) holds in Assumption 5, then we additionally assume Assumption 3. Suppose the initial solution satisfies $\|x_0\| \leq 1$. Then, there exist $\Lambda_\eta^* > 0$ for $\eta \geq 0$ and constants $C_{x_0}, C > 0$ such that for any $0 < \kappa < 1/4$ and $\delta \in (0,1)$, it holds that,*

$$\begin{aligned}\mathbb{P}(L(X_n) - L(x^*) > \delta) \leq \delta^{-1} \Big\{ & L(\tilde{x}) - L(x^*) \\ & + C_{x_0} \exp(-\Lambda_\eta^*(\eta n - 1)) + C \Big[\frac{\hat{c}_\beta}{\Lambda_0^*} \eta^{1/2-\kappa} + \frac{1}{\beta} \left(\sqrt{\tfrac{2M}{\lambda}} + 1 \right) + \lambda \left(\tfrac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|_{\mathcal{H}_K}^2 \right) \Big] \Big\},\end{aligned} \tag{10}$$

where the precise description of the spectral gap Λ_η^* and the constant C_{x_0} is given in the statement of Proposition 8 with $M' = 4\sqrt{M/e}$.

The proof is in Section C. Λ_η^* may depend on β and η, but is uniformly lower-bounded with respect to $\eta > 0$. As can be seen in Eq. (10), there is a competing effect between the regularization Λ_η^* (ensuring faster convergence of the discrete chain) and the inverse temperature β (ensuring better concentration of the Langevin stationary distribution π). We can see that, for fixed λ, by setting $\eta \leq \frac{\log(1/\lambda)}{\Lambda_\eta^* n}$, Eq. (10) excluding the optimization unrelated term $L(\tilde{x}) - L(x^*)$ is of order $O_p\big(\frac{\hat{c}_\beta}{\Lambda_0^*}\big(\frac{\log(1/\lambda)}{\Lambda_\eta^* n}\big)^{1/2-\kappa} + \frac{1}{\beta\sqrt{\lambda}} + \lambda\big)$. Hence, by setting $\beta = \lambda^{-3/2}$ and $n \geq \log(1/\lambda)/[\Lambda_\eta^*(\Lambda_0^*\lambda/\hat{c}_\beta)^{(1/2-\kappa)^{-1}}]$, we have $L(X_n) - L(x^*) = O_p(\lambda)$. Note also that contrary to the finite-dimensional setting where 1-order weak convergence is possible, the $1/2$ rate in η is optimal (Bréhier, 2014). See Remark 6 for the connection to the finite-dimensional analysis. Next, we give the convergence rate of SGLD.

Theorem 5 (Main Result, SGLD convergence rate) *Under Assumptions 1 and 6 and* $\|x_0\| \leq 1$, *SGLD has the following convergence rate:*

$$\mathbb{P}(L(Y_n^N) - L(x^*) > \delta) \lesssim \delta^{-1}\Big(\Theta_n + \frac{\hat{c}_\beta}{\Lambda_0^*}\mu_{N+1}^{1/2-\kappa} + \min\{\sqrt{r_{T^*\wedge n}} + \sqrt[4]{r_{T^*\wedge n}}, q_{n_\mathrm{b}}\}\Big),$$

where $\Theta_n = \exp(-\Lambda_\eta^*(\eta n - 1)) + \frac{\hat{c}_\beta}{\Lambda_0^*}\eta^{1/2-\kappa} + \frac{1}{\beta}\big(\sqrt{\frac{2M}{\lambda}} + 1\big) + \lambda\big(\frac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|_{\mathcal{H}_K}^2\big) + L(\tilde{x}) - L(x^*)$ *which is the convergence rate of GLD shown in Theorem 4,* $r_k = \frac{k\beta\eta(n_\mathrm{tr}-n_\mathrm{b})}{n_\mathrm{b}(n_\mathrm{tr}-1)}$, $T^* = \frac{\log_+\{n_\mathrm{b}(n_\mathrm{tr}-1)/[\beta\eta(n_\mathrm{tr}-n_\mathrm{b})]\}}{\Lambda_\eta^*\eta} + \frac{1}{\eta}$, *and* $q_{n_\mathrm{b}} := \sqrt{\frac{\eta}{\lambda/\mu_0}\frac{(n_\mathrm{tr}-n_\mathrm{b})}{n_\mathrm{b}(n_\mathrm{tr}-1)}}\Big\{1 + \frac{M\mu_0}{\lambda}\exp\Big(\frac{M\mu_0}{\lambda}\Big)\Big\}$.

The approximation error induced by the Galerkin approximation corresponds to $(\hat{c}_\beta/\Lambda_0^*)\mu_{N+1}^{1/2-\kappa}$. Since $\mu_{N+1} \lesssim N^{-2}$, the approximation error decreases in a quadratic order as the dimension N increases. The error induced by the stochastic gradient corresponds to $\sqrt{r_n} + \sqrt[4]{r_n}$. As the minibatch size n_b increases, the stochastic gradient error converges to 0. This rate is slightly better than its finite-dimensional counter part (Raginsky et al., 2017; Xu et al., 2018) by a factor of $\sqrt{n\eta}$. This is due to the regularization term $\lambda\|x\|_{\mathcal{H}_K}^2$.

Remark 6 (Connection to finite-dimensional analysis) *The existing finite-dimensional analysis contains an* $\exp(d)$ *term and thus our bound cannot be achieved by starting from these kinds of finite-dimensional analysis. In our analysis, we overcame this difficulty by imposing regularization so that the solution is included in an RKHS, which is essentially assuming that the global optimum is well approximated by an element in the RKHS. The error term* $\eta^{1/2-\kappa}$ *with arbitrary small* $\kappa > 0$ *is affected by the "complexity" of the space. This term is replaced by* η *in the finite dimension case. The complexity of the RKHS can be characterized by the decay rate of the eigenvalues* $(\mu_k)_{k=0}^\infty$ *(Assumption 1). If the eigenvalue decay behaves as* $1/k^p$ *instead of* $1/k^2$, *then the error term* $\eta^{1/2-\kappa}$ *would be modified to* $\eta^{(p-1)/p-\kappa}$ *(Andersson et al., 2016). The finite-dimensional case corresponds to the limit of* $p \to \infty$ *and the existing bound* η *is recovered.*

3.1. Proof Scheme

Applying GLD and SGLD for non-convex optimization in a finite-dimensional space has been recently extensively investigated by Raginsky et al. (2017); Xu et al. (2018); Erdogdu et al. (2018) to name a few. However, unlike in the proof of such existing analyses for the finite-dimensional case, $\mathbb{E}[L(X_n) - L(x^*)]$ cannot be directly bounded in an infinite-dimensional setting where only convergence for bounded test functions is shown (see Corollary 1.2 in Bréhier (2014) for example). Instead, the bounded function $\phi(x) = \sigma(L(x) - L(x^*))$ with $\sigma(x) = 1/(1+e^{-x}) - 1/2$ is used to bound the probability of the n-th iterate X_n of Eq. (3) being in a certain level set of $L(x) - L(x^*)$, by bounding $\mathbb{E}[\phi(X_n)]$ and applying Markov's inequality (If L is bounded, we don't need to operate σ and we can directly evaluate $\mathbb{E}[L(X_n) - L(x^*)]$).

The seminal paper (Raginsky et al., 2017) derived the finite time error bound of SGLD for non-convex learning problem utilizing the decomposition

$$\mathbb{E}[\phi(X_n) - \phi(x^*)] = \mathbb{E}[\phi(X_n) - \phi(X(n\eta))] + \mathbb{E}[\phi(X(n\eta)) - \phi(X^\pi)] + \mathbb{E}[\phi(X^\pi) - \phi(x^*)], \quad (11)$$

where π is the stationary distributions of the continuous Markov chain $\{X(t)\}_{t\geq 0}$ and we denote by X^μ a random variable obeying a probability distribution μ. On the other hand, Xu et al. (2018)

observed that this decomposition could be improved by utilizing the geometric ergodicity of discrete time dynamics and proposed to use the following decomposition:

$$\mathbb{E}[\phi(X_n) - \phi(x^*)] = \mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] + \mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^{\pi})] + \mathbb{E}[\phi(X^{\pi}) - \phi(x^*)], \quad (12)$$

where μ_η is the stationary distribution of the discrete Markov chain $\{X_n\}_{n\in\mathbb{N}}$ (the existence of which is not trivial). By using this, it is shown that some polynomial order term with respect to n can be dropped to obtain a faster rate.[3] Our analysis employs this strategy. That is, we control each term in the decomposition of Eq. (12).

Extending this strategy to an infinite-dimensional setting is not trivial. For example, the boundedness of the norm of noise $\|\epsilon_n\|_{\mathcal{H}}$ does no longer hold, and thus we need an additional regularization term $AX(t)$ to make the solution bounded in $\mathcal{H}$ and hit a compact set with high probability. The time discretization of the infinite-dimensional Langevin dynamics has been studied especially as a numerical scheme of stochastic partial differential equation (Kuksin and Shirikyan, 2001; Debussche, 2011; Bréhier, 2014; Bréhier and Kopec, 2016; Andersson et al., 2016; Chen et al., 2017, 2018). Bréhier (2014); Bréhier and Kopec (2016) derived a weak approximation error of the time discretization scheme (3) from the stationary distribution π. However, their proof strategy utilizes the decomposition Eq. (11) like Raginsky et al. (2017) in the finite-dimensional counter part. As we have pointed out above, the error bound could be improved by using the decomposition Eq. (12) instead. Unfortunately, the geometric ergodicity of the discrete time dynamics has not been established so far. Therefore, we have introduced Assumption 3 so that the geometric ergodicity holds.

3.1.1. FIRST TERM: GEOMETRIC ERGODICITY OF THE DISCRETE CHAIN

First, we need a moment bound of the chain $\{X_n\}_{n\in\mathbb{N}}$ as follows

Proposition 7 *Let Assumptions 1, 2 and 5 hold. Let $\{Z_n\}_{n\in\mathbb{N}}$ solve the dynamics with $\nabla L = 0$: $Z_0 = 0$ and $Z_{n+1} = S_\eta Z_n + \sqrt{\frac{2\eta}{\beta}} S_\eta \varepsilon_n$ with $\beta > \eta$. Then, $\forall p > 0$, it holds that $k(p) \triangleq \sup_{n\geq 0} \mathbb{E}(\|Z_n\|^p) < \infty$. Using this evaluation, we have $\mathbb{E}_{x_0}\|X_n\| \leq \rho^n \|x_0\| + b$ $(\forall n \in \mathbb{N})$ with (i) (for Strict Dissipativity) $\rho = \frac{1+\eta M}{1+\lambda\eta/\mu_0} < 1$, $b = \|x^*\| + 2k(1)$, or (ii) (for Bounded gradients) $\rho = \frac{1}{1+\lambda\eta/\mu_0} < 1$, $b = \frac{\mu_0}{\lambda} B + k(1)$.*

The proof is given in Section D. This is also called a *Lyapunov condition*. Combined with this and so called *minorization condition*, we can show the geometric ergodicity in the following proposition.

Proposition 8 (Geometric ergodicity) *Let Assumptions 1, 2, 4 and 5 hold. If only the bounded gradient condition holds in Assumption 5, then we additionally assume Assumption 3. Let $\eta > 0, \beta > \eta$ and $V(x) = \|x\| + 1$. Then, there exists a unique invariant measure μ_η and $\Lambda^*_\eta > 0$ such that for all $\phi : \mathcal{H} \to \mathbb{R}$ with $|\phi(\cdot)| \leq V(\cdot)$ and $\|\phi(x) - \phi(y)\| \leq M'\|x - y\|$ $(x, y \in \mathcal{H})$, we have*

$$|\mathbb{E}_{x_0}[\phi(X_n)] - \mathbb{E}[\phi(X^{\mu_\eta})]| \leq C_{x_0} \exp(-\Lambda^*_\eta(\eta n - 1)), \quad (13)$$

where C_{x_0} and $\Lambda^*_\eta > 0$ are given by

3. We would like to point out that we have found some incorrect analysis of the error bound in Xu et al. (2018). In particular, there are several wrong evaluations about dependency of constants (including the spectral gap) on the inverse temperature parameter β.

i) (Strict dissipativity, Assumption 5 (i)) $\Lambda_\eta^* = \frac{\frac{\lambda}{\mu_0} - M}{1+\eta\frac{\lambda}{\mu_0}}$, $C_{x_0} = M'(\|x_0\|_{\mathcal{H}} + b)$,

ii) (Bounded Gradient, Assumption 5 (ii)) $\Lambda_\eta^* = \frac{\min\left(\frac{\lambda}{2\mu_0}, \frac{1}{2}\right)}{4\log(a(\bar{V}+1)/(1-\delta))}\delta$, $C_{x_0} = a[\bar{V}+1] + \frac{\sqrt{2}(V(x_0)+b)}{\sqrt{\delta}}$ *for* $0 < \delta < 1$ *satisfying* $\delta = \Omega(\exp(-O(\beta)))$[4], $\bar{b} = \max\{b, 1\}$, $a = \bar{b} + 1$ *and* $\bar{V} = 4\bar{b}/(\sqrt{(1+\rho^{1/\eta})/2} - \rho^{1/\eta})$ *where* ρ *and* b *are given in Proposition 7.*

The proof is given in Section E. Unlike existing work, this theorem asserts the geometric ergodicity of the discrete time dynamics, whilst the geometric ergodicity for "continuous time" dynamics (Eq. (4)) has been well known, see as an example (Debussche, 2011, 2013). Transforming the continuous time argument to the discrete time setting is far from trivial because there appears a "integrability" problem. Indeed, Bréhier (2014); Bréhier and Kopec (2016) pointed out there has been no work that showed the geometric ergodicity of the time-discretized dynamics. This difficulty does not occur in the finite-dimensional setting. We resolved this problem by imposing Assumption 3. Thanks to this, we have exponential convergence $\exp(-\Lambda_\eta^* n\eta)$ improving the polynomial order rate $\frac{1}{\Lambda_0^*}(n\eta)^{-1}$ of existing work.

3.1.2. SECOND TERM: WEAK CONVERGENCE OF THE DISCRETE SCHEME

The second term is linked to the weak convergence of the numerical scheme, i.e., in our case the convergence of $\phi(X_n)$ to $\phi(X(n\eta))$ for any admissible test function $\phi \in C_b^2$. We rely directly on the results of Bréhier and Kopec (2016), who prove $1/2$ order weak convergence in time and 1 order weak convergence in space for numerical schemes that have a semi-implicit discretization in time with $\beta = 1$, and a finite elements discretization in space; that is, they showed

$$|\mathbb{E}[\phi(X^{\mu^\eta}) - \phi(X^\pi)]| \leq C \|\phi\|_{0,2} \eta^{1/2-\kappa}, \tag{14}$$

where $\|\phi\|_{0,2} \triangleq \max\{\|\phi\|_\infty, \sup_{x\in\mathcal{H}} \|\nabla\phi(x)\|_{\mathcal{H}}, \sup_{x\in\mathcal{H}} \|D^2\phi(x)\|_{\mathcal{B}(\mathcal{H})}\}$ for $\phi \in C_b^2$.

In the general setting, $\beta \neq 1$, we need to evaluate the effect of β. To that purpose, we essentially consider a re-scaling argument, that is, we observe that if we replace L with $L' \triangleq \beta L$, λ with $\lambda' \triangleq \beta\lambda$ and η with $\eta' \triangleq \frac{\eta}{\beta}$ in Eq. (4) and Eq. (3), then it holds that $S_\eta = \left(\mathrm{Id} + \eta\frac{\lambda}{2}\nabla\|\cdot\|_{\mathcal{H}_K}^2\right)^{-1} = \left(\mathrm{Id} + \frac{\eta}{\beta}\frac{\beta\lambda}{2}\nabla\|\cdot\|_{\mathcal{H}_K}^2\right)^{-1} =: \tilde{S}_{\eta'}$, and thus

$$X_{n+1} = \tilde{S}_{\eta'} X_n - \eta'\tilde{S}_{\eta'}\nabla L'(X_n) + \sqrt{2\eta'}\tilde{S}_{\eta'}\varepsilon_n,$$

i.e., $\{X_n\}_{n\in\mathbb{N}}$ is the numerical approximation of $\mathrm{d}X(t) = -\nabla\mathcal{L}'(X(t)) + \sqrt{2}\,\mathrm{d}W(t)$ with time step η'. We carefully evaluate how the constant C is Eq. (14) will be changed after rescaling. We can see that β affects the rate through the spectral gap Λ_0^*, which corresponds to the continuous dynamics ($\eta = 0$). Eventually, we get the following result:

Proposition 9 (Case $\beta \neq 1$) *Under the same setting as Proposition 8, for any* $0 < \kappa < 1/2$, $0 < \eta_0$, *there exists a constant* C *such that for any bounded test function* $\phi \in C_b^2$ *and* $0 < \eta < \eta_0$, *it holds that*

$$|\mathbb{E}[\phi(X^{\mu^\eta}) - \phi(X^\pi)]| \leq C(\Lambda_0^*)^{-1} \|\phi\|_{0,2}\, \hat{c}_\beta \eta^{1/2-\kappa}. \tag{15}$$

The proof is given in Section G. Note that due to the infinite-dimensional setting, the $1/2$ rate w.r.t the time discretization η is optimal (Bréhier, 2014). This is to be contrasted with the finite-dimensional case, where 1 order weak convergence is attainable.

4. More detailed evaluation of δ can be found in the proof.

3.1.3. THIRD TERM: CONCENTRATION OF THE GIBBS DISTRIBUTION AROUND THE GLOBAL MINIMUM

The last term corresponds to the concentration of the stationary Gibbs distribution around the global minimum of L. In this infinite-dimensional setting, the regularizing effect of operator A is necessary to ensure good convergence properties of the discrete and continuous chains. Hence, even in the limit case $\beta \to 0$ one cannot expect to have arbitrary tight concentration around the global minimum. This is to be contrasted with the finite-dimensional case (Chiang et al. (1987); Gelfand and Mitter (1991); Roberts and Tweedie (1996)). In fact, A constrains the chain to remain within the support of a Gaussian process which is compactly embedded in $\mathcal{H}$.

Proposition 10 *Under Assumptions 1 and 2, it holds that*

$$\int L \,\mathrm{d}\pi - L(\tilde{x}) \lesssim \frac{1}{\beta}\left(\sqrt{\frac{2M}{\lambda}} + 1\right) + \lambda\left(\frac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|^2_{\mathcal{H}_K}\right).$$

The proof can be found in Section F. The proposition can be shown by utilizing an analogous technique to the convergence rate analysis of Gaussian process regression (van der Vaart and van Zanten, 2011). Along with this technique, the *Gaussian correlation inequality* (Royen, 2014; Latała and Matlak, 2017) is used. This inequality gives a powerful tool to lower-bound the Gaussian probability measure of the intersection of two centered convex sets.

3.1.4. ERROR BOUND FOR THE GALERKIN APPROXIMATION AND STOCHASTIC GRADIENT

The error induced by the Galerkin approximation can be evaluated as in the following proposition.

Proposition 11 *Let Assumptions 1, 2, 4 and 5 hold and suppose* $\|x\| \leq 1$. *Then, there exists an invariant measure* $\mu_{(N,\eta)}$ *for the discrete time Galerkin approximation scheme (Eq. (5)), and for any* $0 < \kappa < 1/2$, $0 < \eta_0$, *there exists a constant* $C > 0$ *such that, for any* $N \in \mathbb{N}$ *and* $0 < \eta < \eta_0$,

$$\mathbb{E}[\phi(X^{\mu_{(N,\eta)}}) - \phi(X^{\pi})] \leq \frac{C\|\phi\|_{0,2}}{\Lambda_0^*}\hat{c}_\beta\left(\mu_{N+1}^{1/2-\kappa} + \eta^{1/2-\kappa}\right).$$

The proof is in Section G. We see that, by taking $N \to \infty$, we can replicate Proposition 9. Moreover, the geometric ergodicity of the time discretized dynamics with the Garelkin approximation holds completely in the same manner as Proposition 8. The discrepancy between GLD and SGLD with the Garelkin approximation can be bounded as follows.

Proposition 12 *Suppose* $\|x_0\| \leq 1$. *There exists a constant* $C > 0$ *such that, for any* $n, N \in \mathbb{N}$, *any* $\beta > 1$ *and sufficiently small* $\eta > 0$,

$$\mathbb{E}[\phi(X_n^N) - \phi(Y_n^N)] \leq C \min\left\{\sqrt{r_{T^*\wedge n}} + \sqrt[4]{r_{T^*\wedge n}}, q_{n_\mathrm{b}}\right\},$$

where r_k, T^* *and* q_{n_b} *are as defined in Theorem 5.*

The proof is given in Section H. From these propositions, we can see that the SGLD with the Garelkin approximation also gives a reasonably good solution for sufficiently large $N \in \mathbb{N}$, sufficiently small $\eta > 0$ and sufficiently large mini-batch size. Proposition 12 is analogous to those given for finite-dimensional situations (Raginsky et al., 2017; Xu et al., 2018). However, thanks to the regularization term (appearing as S^η), our rate is better by a factor of $\sqrt{n\eta}$.

4. Other Related Work

Here, we mention other related work that have not been exposed above. An analogous assumption to Assumption 3 has already been introduced in the analysis of infinite-dimensional dynamics with nonlinear diffusion term, that is, $\mathrm{d}W(t)$ is replaced by a nonlinear quantity $\sigma(X(t))\,\mathrm{d}W(t)$ for $\sigma(X(t)) \in \mathcal{B}(\mathcal{H})$ (Conus et al., 2019; Debussche, 2011; Bréhier and Debussche, 2018). These papers analyzed the existence of stationary distribution for continuous dynamics and discrete time approximation for finite time horizon. Chen et al. (2017, 2018) analyzed linear/nonlinear Schrödinger equations and derived geometric ergodicity, but they analyzed much more specific situations or stronger assumptions (e.g. the strong dissipativity condition). The geometric ergodicity of infinite-dimensional Markov processes for discrete time settings has been investigated by Kuksin and Shirikyan (2001) and infinite-dimensional MCMC such as preconditioned Crank–Nicolson (pCN) (Hairer et al., 2014; Eberle, 2014; Vollmer, 2015; Rudolf and Sprungk, 2018), and in particular the Metropolis-Adjusted Langevin Algorithm (MALA) (Durmus and Moulines, 2015; Beskos et al., 2017). Among them, MALA is the most related to our setting. The biggest difference is the existence of a rejection step. Since the purpose of our work is rather optimization than sampling, and since the rejection step is not compatible with stochastic gradient descent, we do not pursue this direction.

Conclusion and Future Work

In this paper, we have presented a non-asymptotic analysis of the convergence of GLD and SGLD in a RKHS and for a non-convex objective function. The bounds obtained in this infinite-dimensional setting involve the spectrum of the associated integral operator and a regularization factor instead of the dimension d, which to the best of our knowledge is the first result on applying GLD in RKHS to infinite-dimensional non-convex optimization. In future work, we hope to alleviate the somewhat strict Assumption 1 linked to current results from the numerical approximation literature. Drawing inspiration from Xu et al. (2018), we also plan to extend our analysis to variance-reduced SGLD algorithms.

Acknowledgement

This work was supported by JSPS KAKENHI (20H00576), Japan Digital Design and JST CREST.

References

A. Andersson and S. Larsson. Weak convergence for a spatial approximation of the nonlinear stochastic heat equation. *Mathematics of Computation*, 85(299):1335–1358, 2016.

A. Andersson, R. Kruse, and S. Larsson. Duality in refined Sobolev–Malliavin spaces and weak approximation of SPDE. *Stochastics and Partial Differential Equations Analysis and Computations*, 4(1):113–149, 2016.

A. Beskos, M. Girolami, S. Lan, P. E. Farrell, and A. M. Stuart. Geometric MCMC for infinite-dimensional inverse problems. *Journal of Computational Physics*, 335:327 – 351, 2017.

J. Bismut. *Large Deviations and the Malliavin Calculus (Progress in Mathematics)*. Birkhäuser, 1984.

F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Serie 6, 14(3):331–352, 2005.

C. Borell. The Brunn-Minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2): 207–216, 1975.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

C.-E. Bréhier. Approximation of the invariant measure with an Euler scheme for stochastic PDEs driven by space-time white noise. *Potential Analysis*, 40(1):1–40, 2014.

C.-E. Bréhier and A. Debussche. Kolmogorov equations and weak order analysis for SPDEs with nonlinear diffusion coefficient. *Journal de Mathématiques Pures et Appliquées*, 119:193 – 254, 2018. ISSN 0021-7824.

C.-E. Bréhier and M. Kopec. Approximation of the invariant law of SPDEs: error analysis using a Poisson equation for a full-discretization scheme. *IMA Journal of Numerical Analysis*, 37(3): 1375–1410, 07 2016.

C.-E. Bréhier and G. Vilmart. High order integrator for sampling the invariant distribution of a class of parabolic stochastic pdes with additive space-time noise. *SIAM Journal on Scientific Computing*, 38(4):A2283–A2306, 2016.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

C. Chen, J. Hong, and X. Wang. Approximation of invariant measure for damped stochastic nonlinear Schrödinger equation via an ergodic numerical scheme. *Potential Analysis*, 46(2):323–367, Feb 2017.

Z. Chen, S. Gan, and X. Wang. A full-discrete exponential Euler approximation of invariant measure for parabolic stochastic partial differential equations, 2018.

T.-S. Chiang, C.-R. Hwang, and S. Sheu. Diffusion for global optimization in Rˆn. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

L. Chizat and F. Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

D. Conus, A. Jentzen, and R. Kurniawan. Weak convergence rates of spectral Galerkin approximations for SPDEs with nonlinear diffusion coefficients. *The Annals of Applied Probability*, 29(2): 653–716, 04 2019.

G. Da Prato and J. Zabczyk. Non-explosion, boundedness and ergodicity for stochastic semilinear equations. *J. Diff. Equations*, 98:181–195, 1992.

G. Da Prato and J. Zabczyk. *Ergodicity for Infinite Dimensional Systems*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1996.

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

A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. *arXiv preprint arXiv:1704.04752*, 2017b.

A. Debussche. Weak approximation of stochastic partial differential equations: the nonlinear case. *Mathematics of Computation*, 80(273):89–117, 2011.

A. Debussche. Ergodicity results for the stochastic Navier–Stokes equations: an introduction. In *Topics in mathematical fluid mechanics*, pages 23–108. Springer, 2013.

A. Debussche, Y. Hu, and G. Tessitore. Ergodic BSDEs under weak dissipative assumptions. *Stochastic Processes and their Applications*, 121(3):407–426, 2011.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159, 2011.

A. Duncan, N. Nüsken, and G. Pavliotis. Using perturbed underdamped langevin dynamics to efficiently sample from probability distributions. *Journal of Statistical Physics*, 169(6):1098–1131, 2017.

A. Durmus and É. Moulines. Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm. *Statistics and Computing*, 25(1):5–19, 2015.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

A. Eberle. Error bounds for Metropolis-Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377, 02 2014. doi: 10.1214/13-AAP926.

K. Elworthy and X. Li. Formulae for the derivatives of heat semigroups. *Journal of Functional Analysis*, 125(1):252–286, 1994. doi: https://doi.org/10.1006/jfan.1994.1124. URL https://www.sciencedirect.com/science/article/pii/S0022123684711244.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems 31*, pages 9671–9680. 2018.

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

B. Goldys and B. Maslowski. Lower estimates of transition densities and bounds on exponential ergodicity for stochastic PDEs. *The Annals of Probability*, 34(4):1451–1496, 2006.

M. Hairer. Exponential mixing properties of stochastic PDEs through asymptotic coupling. *Probab. Theory Related Fields*, 124(3):345–380, 2002.

M. Hairer, A. M. Stuart, and S. J. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490, 12 2014.

J. K. Hale. *Asymptotic Behavior of Dissipative Systems*. American Mathematical Society, 1988.

J. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. I*, volume 305. Springer-Verlag, Berlin, 1993.

S. Jacquot and G. Royer. Ergodicité d'une classe d'équations aux dérivées partielles stochastiques. *C. R. Acad. Sci. Paris Sér. I Math.*, 320(2):231–236, 1995.

D. Kendall. Unitary dilations of markov transition probabilities and the corresponding integral representations for transitions probability matrices. In U. Grenander, editor, *Probability and Statistics: the Harald Cramér Volume*, pages 139–161. 1959.

R. Khasminskii. *Stochastic stability of differential equations*. Springer Science & Business Media, 2011.

D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

M. Kopec. *Numerical methods for stochastic equations*. Theses, Ecole normale supérieure de Rennes - ENS Rennes, June 2014.

R. Kruse. Optimal error estimates of Galerkin finite element methods for stochastic partial differential equations with multiplicative noise. *IMA Journal of Numerical Analysis*, 34(1):217–251, 2013.

J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for gaussian measures. *Journal of Functional Analysis*, 116(1):133–157, 1993.

S. Kuksin and A. Shirikyan. A coupling approach to randomly forced nonlinear PDEs. I. *Communications in Mathematical Physics*, 221(2):351–366, 2001.

R. Latała and D. Matlak. *Royen's Proof of the Gaussian Correlation Inequality*, pages 265–275. Springer International Publishing, 2017.

W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic Processes: Theory and Methods*, 19:533–597, 2001.

B. Maslowski. Strong Feller property for semilinear stochastic evolution equations and applications. In *Stochastic systems and optimization (Warsaw, 1988)*, volume 136 of *Lect. Notes Control Inf. Sci.*, pages 210–224. Springer, Berlin, 1989.

H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 21*, pages 1049–1056. 2009.

J. Mattingly, A. Stuart, and D. Higham. Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185 – 232, 2002.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

S. Mischler. An introduction to evolution PDEs, Chapter 0: On the Gronwall lemma, 2019. URL https://www.ceremade.dauphine.fr/~mischler/Enseignements/M2evol2018/chap0.pdf.

T. Nagapetyan, A. B. Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. Zygalakis. The true cost of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017.

Y. Nesterov. A method for solving a convex programming problem with rate of convergence $O(1/k^2)$. *Soviet Math. Doklady*, 269(3):543–547, 1983.

Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

A. Nitanda and T. Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

D. Nualart. *The Malliavin Calculus and Related Topics*. Springer, 2006.

Y. Polyanskiy and Y. Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

J. Printems. On the discretization in time of parabolic stochastic partial differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 35(6):1055–1078, 2001.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *arXiv e-prints*, page arXiv:1702.03849, 2017.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

G. Roberts and R. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

T. Royen. A simple proof of the gaussian correlation conjecture extended to multivariate gamma distributions. *arXiv preprint arXiv:1408.1028*, 2014.

D. Rudolf and B. Sprungk. On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm. *Foundations of Computational Mathematics*, 18(2):309–343, 2018.

M. Sanz-Solé. *Malliavin Calculus with Application to Stochastic Partial Differential Equations*. EPFL Press, 2005.

T. Shardlow. Geometric ergodicity for stochastic PDEs. *Stochastic Anal. Appl.*, 17(5):857–869, 1999.

M. Signoretto, L. De Lathauwer, and J. A. Suykens. Learning tensors in reproducing kernel hilbert spaces with multilinear spectral penalties. *arXiv preprint arXiv:1310.4977*, 2013.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks. *arXiv preprint arXiv:1805.01053*, 2018.

R. Sowers. Large deviations for the invariant measure of a reaction-diffusion equation with non-Gaussian perturbations. *Probab. Theory Related Fields*, 92(3):393–421, 1992. ISSN 0178-8051.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

T. Suzuki. Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional langevin dynamics. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19224–19237. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/df1a336b7e0b0cb186de6e66800c43a9-Paper.pdf.

T. Suzuki and S. Akiyama. Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=2m0g1wEafh.

T. Suzuki, H. Kanagawa, H. Kobayashi, N. Shimizu, and Y. Tagami. Minimax optimal alternating minimization for kernel nonparametric tensor learning. In *Advances in Neural Information Processing Systems 29*, pages 3783–3791. 2016.

A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, 3: 200–222, 2008. IMS Collections.

A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

P. Vellanki, S. Rana, S. Gupta, D. R. de Celis Leal, A. Sutti, M. Height, and S. Venkatesh. Bayesian functional optimisation with shape prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1617–1624, 2019.

S. Vempala and A. Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32*, pages 8094–8106. Curran Associates, Inc., 2019.

N. A. Vien, H. Zimmermann, and M. Toussaint. Bayesian functional optimization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, pages 4171–4178, 2018.

S. J. Vollmer. Dimension-independent MCMC sampling for inverse problems with non-Gaussian priors. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):535–561, Jan 2015. ISSN 2166-2525.

X. Wang and S. Gan. Weak convergence analysis of the linear implicit euler method for semilinear stochastic partial differential equations with additive noise. *Journal of Mathematical Analysis and Applications*, 398(1):151–169, 2013.

M. Welling and Y.-W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, pages 681–688, 2011.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of Langevin Dynamics based algorithms for nonconvex optimization. In *NeurIPS*, pages 3122–3133, 2018.

M. D. Zeiler. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. *arXiv preprint arXiv:1702.05575*, 2017.

Appendix A. Proof of Eq. (9)

Note that, for $x = \sum_{k=0}^{\infty} \alpha_k f_k \in \mathcal{H}$, $T_K^{\frac{\gamma}{2}} x(z) = \sum_{k=0}^{\infty} \mu_k^{\frac{\gamma}{2}} \alpha_k f_k(z)$, and thus we can obtain a reproducing formula $T_K^{\frac{\gamma}{2}} x(z) = \langle x, \psi_\gamma(z) \rangle_{\mathcal{H}}$ where $\psi_\gamma(z) \triangleq \sum_{k=0}^{\infty} \mu_k^{\frac{\gamma}{2}} f_k(z) f_k$. ψ_γ defines the kernel function of $\mathcal{H}_{K^\gamma}$ as $K_\gamma(z, z') = \langle \psi_\gamma(z), \psi_\gamma(z') \rangle_{\mathcal{H}} = \sum_{k=0}^{K} \mu_k^\gamma f_k(z) f_k(z')$. Using this, we see that $\|\psi_\gamma(z)\|_{\mathcal{H}}^2 = \sum_{k=0}^{\infty} \mu_k^\gamma f_k^2(z) = K_\gamma(z, z)$ and $\|\psi_\gamma(z)\|_\epsilon^2 = \sum_{k=0}^{\infty} \mu_k^{\gamma+2\epsilon} f_k^2(z) = K_{\gamma+2\epsilon}(z, z)$.

If $\|\ell_i''\|_\infty \leq G$, then it is G-Lipschitz continuous. Therefore, it holds that

$$
\begin{aligned}
&\|\nabla L(x) - \nabla L(x')\| \\
&\leq \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} |\ell_i'(\langle x, \psi_\gamma(z) \rangle_{\mathcal{H}}) - \ell_i'(\langle x', \psi_\gamma(z') \rangle_{\mathcal{H}})| \|\psi_\gamma(z_i)\|_{\mathcal{H}} + \lambda_0 \|x - x'\|_{\mathcal{H}} \\
&\leq \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} |\ell_i'(\langle x, \psi_\gamma(z) \rangle_{\mathcal{H}}) - \ell_i'(\langle x', \psi_\gamma(z') \rangle_{\mathcal{H}})| \sqrt{K_\gamma(z_i, z_i)} + \lambda_0 \|x - x'\|_{\mathcal{H}} \\
&\leq \sup_z \sqrt{K_\gamma(z, z)} G \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \|\langle x - x', \psi_\gamma(z_i) \rangle_{\mathcal{H}}\| + \lambda_0 \|x - x'\|_{\mathcal{H}} \\
&\leq G \sup_z K_\gamma(z, z) \|x - x'\|_{\mathcal{H}} + \lambda_0 \|x - x'\|_{\mathcal{H}} \leq (G R_\gamma + \lambda_0) \|x - x'\|_{\mathcal{H}}.
\end{aligned}
$$

This yields the first inequality in Eq. (9). As for the second order derivative (the second term in Eq. (9)), first note that

$$
D^2 L(x) \cdot (h, k) = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \ell''(\langle x, \psi_\gamma(z_i) \rangle_{\mathcal{H}}) \langle \psi_\gamma(z_i), h \rangle_{\mathcal{H}} \langle \psi_\gamma(z_i), k \rangle_{\mathcal{H}} + \lambda_0 \langle h, k \rangle_{\mathcal{H}}
$$

for $h, k \in \mathcal{H}$. Therefore, we have that

$$
\begin{aligned}
&|D^2 L(x) \cdot (h, k) - \lambda_0 \langle h, k \rangle_{\mathcal{H}}| \\
&\leq \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} |\ell''(\langle x, \psi_\gamma(z_i) \rangle_{\mathcal{H}})| |\langle \psi_\gamma(z_i), h \rangle_{\mathcal{H}}| \|\psi_\gamma(z_i)\|_{-\alpha} \|k\|_\alpha \\
&\leq G \max_i \|\psi_\gamma(z_i)\|_{\mathcal{H}} \|h\|_{\mathcal{H}} \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \|\psi_\gamma(z_i)\|_{-\alpha} \|k\|_\alpha \\
&= G \max_i \sqrt{K_\gamma(z_i, z_i)} \|h\|_{\mathcal{H}} \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \sqrt{K_{\gamma - 2\alpha}(z_i, z_i)} \|k\|_\alpha \\
&\leq G \max_i \sqrt{K_\gamma(z_i, z_i)} \|h\|_{\mathcal{H}} \sqrt{\frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} K_{\gamma - 2\alpha}(z_i, z_i)} \|k\|_\alpha \\
&\leq G \max_i \sqrt{K_\gamma(z_i, z_i)} \|h\|_{\mathcal{H}} \sqrt{\sum_{k=0}^{\infty} \mu_k^{\gamma - 2\alpha}} \|k\|_\alpha \leq G \sqrt{R_\gamma} \|h\|_{\mathcal{H}} \sqrt{\sum_{k=0}^{\infty} \mu_k^{\gamma - 2\alpha}} \|k\|_\alpha.
\end{aligned}
$$

□

A.1. Remark on existence of regularization term

As an example of $L(x)$, it is useful to consider a setting where $L(x)$ can be expressed as $L(x) = \tilde{L}(x) + \frac{\lambda_0}{2}\|x\|^2$ for $\tilde{L}(x)$ that satisfies the assumptions listed in the main text and $\lambda_0 \geq 0$. In this case, $L(x)$ does not satisfy the bounded gradient condition Assumption 5 (ii). However, by considering the following update rule, we can show the same error bound for $L(x)$:

$$\begin{cases} X_0 = x_0 \in \mathcal{H}, \\ X_{n+1} = S'_\eta(X_n - \nabla \tilde{L}(X_n) + \sqrt{2\frac{\eta}{\beta}}\varepsilon_n), \end{cases} \tag{16}$$

where $S'_\eta = \left[\mathrm{Id} + \eta(\frac{\lambda_0}{2}\nabla \|\cdot\|_{\mathcal{H}_K} + \frac{\lambda}{2}\nabla \|\cdot\|_{\mathcal{H}})\right]^{-1}$.

Appendix B. Proof of Proposition 2

Proof Let us assume $\lambda > M\mu_0$ (Strict Dissipativity). Assumption 1 implies, for $x = \sum_{k=0}^{\infty} \alpha_k f_k$,

$$\begin{aligned} \langle Ax, x\rangle &= -\lambda \Big\langle \sum_{k=0}^{\infty} \frac{\alpha_k}{\mu_k} f_k, \sum_{k=0}^{\infty} \alpha_k f_k \Big\rangle \\ &= -\lambda \sum_{k=0}^{\infty} \frac{\alpha_k^2}{\mu_k} \\ &\leq -\frac{\lambda}{\mu_0} \sum_{k=0}^{\infty} \alpha_k^2 = -\frac{\lambda}{\mu_0} \|x\|^2, \end{aligned} \tag{17}$$

and Assumption 2 implies

$$\begin{aligned} \langle -\nabla L(x), x\rangle &\leq M \|x - x^*\| \|x\| \\ &\leq M \|x\|^2 + M \|x\| \|x^*\|. \end{aligned} \tag{18}$$

Hence,

$$\langle Ax - \nabla L(x), x\rangle \leq -(\tfrac{\lambda}{\mu_0} - M)\|x\|^2 + M\|x\|\|x^*\|.$$

Therefore, if $M < \frac{\lambda}{\mu_0}$, there exists $m, c > 0$ such that Eq. (8) holds. The proof when Assumption 5 (ii) holds is similar. ■

Appendix C. Proof of main result: Theorem 4 and Theorem 5

In light of Sections 3.1.1 to 3.1.3, we can now state our final result. We introduce the following bounded test function:

$$\phi(x) = \sigma(L(x) - L(x^*)) \quad (x \in \mathcal{H}), \tag{19}$$

where $\sigma(u) = \frac{1}{1+e^{-u}} - \frac{1}{2}$ $(u \in [0, \infty))$ is concave and takes values in $[0, 1)$ (note that $L(x) - L(x^*) \geq 0$ for any $x \in \mathcal{H}$). We can show that $\phi(x)$ is $4\sqrt{M/e}$-Lipschitz continuous. By the M-smoothness of L, it holds that

$$L(y) \leq L(x) + \langle y - x, \nabla L(x)\rangle + \frac{M}{2}\|y - x\|^2.$$

Therefore, by the optimality of x^*, we have

$$\begin{aligned} L(x^*) &\leq \inf_{\epsilon>0} L(x - \epsilon \nabla L(x)) \\ &\leq \inf_{\epsilon>0} \left\{ L(x) - \langle \epsilon \nabla L(x), \nabla L(x) \rangle + \frac{M\epsilon^2}{2} \|\nabla L(x)\|^2 \right\} = L(x) - \frac{1}{2M} \|\nabla L(x)\|^2. \end{aligned}$$

Hence,

$$\begin{aligned} \|\nabla \phi(x)\| &= \|\sigma'(L(x) - L(x^*)) \nabla L(x)\| \leq \sigma'(L(x) - L(x^*)) \\ &\leq |\sigma'(L(x) - L(x^*))| \sqrt{2M(L(x) - L(x^*))} \\ &\leq \sqrt{2M} \sup_{u \geq 0} \left\{ \sigma'(u) u^{1/2} \right\} = \sqrt{2M} \sup_{u \geq 0} \left\{ \frac{1}{(1+e^{-u})^2} e^{-u} u^{1/2} \right\} \\ &\leq 4\sqrt{2M} \sup_{u \geq 0} \left\{ e^{-u} u^{1/2} \right\} = 4\sqrt{2M} \left(\frac{1/2}{e} \right)^{1/2} = 4\sqrt{M/e}. \end{aligned}$$

This yields M'-Lipschitz continuity of ϕ where $M' = 4\sqrt{M/e}$. Moreover, we can check that $|\phi(\cdot)| \leq V(\cdot)$ for $V(x) = \|x\| + 1$ (because $\phi(x) \leq 1$ $(\forall x \in \mathcal{H})$), and $\phi \in C_b^2(\mathcal{H})$, hence ϕ falls within the scope of Propositions 8 and 9.

First, we note that there exists a unique invariant measure μ_η for the discrete time dynamics $\{X_n\}_n$ and there also exists a unique invariant measure $\mu_{(N,\eta)}$ for the discrete time Garelkin approximated dynamics $\{X_n^N\}_n$ by Proposition 10. To obtain the result, we make use of Markov's inequality: for any $0 < \delta < 1$,

$$\begin{aligned} &P(L(X_n) - L(x^*) > \delta) \\ &\leq P(\phi(X_n) > \sigma(\delta)) \\ &\leq \frac{\mathbb{E}[\phi(X_n)]}{\sigma(\delta)} \quad (\because \text{Markov's inequality}) \\ &= \frac{1}{\sigma(\delta)} \left(\mathbb{E}[\phi(X_n) - \phi(X^\eta)] + \mathbb{E}[\phi(X^\eta) - \phi(X^\pi)] + \mathbb{E}[\phi(X^\pi)] \right). \end{aligned}$$

The first term $(\mathbb{E}[\phi(X_n) - \phi(X^\eta)])$ can be bounded by Proposition 8. The second term $(\mathbb{E}[\phi(X^\eta) - \phi(X^\pi)])$ can be bounded by Proposition 9. Next, we bound the third term. Since $\sigma(u) \leq u$ for all $u \in [0, \infty)$ and $L(x) - L(x^*) \geq 0$ for all $x \in \mathcal{H}$, it holds that

$$\mathbb{E}[\phi(X^\pi)] \leq \mathbb{E}[L(X^\pi) - L(x^*)] = (\mathbb{E}[L(X^\pi)] - L(\tilde{x})) + (L(\tilde{x}) - L(x^*)). \tag{20}$$

Then, the first term $(\mathbb{E}[L(X^\pi)] - L(\tilde{x}))$ in the right hand side is bounded by Proposition 10. Finally, we observe that $1/\sigma(\delta) \leq 5/\delta$ for all $\delta \in (0, 1)$. Combining all results, we obtain Theorem 4.

As for the Theorem 5, we use the following decomposition

$$\begin{aligned} \mathbb{E}[\phi(X_n)] =& \mathbb{E}[\phi(Y_n^N) - \phi(X_n^N)] + \mathbb{E}[\phi(X_n^N) - \phi(X^{\mu_{(N,\eta)}})] \\ &+ \mathbb{E}[\phi(X^{\mu_{(N,\eta)}}) - \phi(X^\pi)] + \mathbb{E}[\phi(X^\pi)]. \end{aligned}$$

We apply Proposition 12 to the first term $(\mathbb{E}[\phi(Y_n^N) - \phi(X_n^N)])$ and apply Proposition 11 to the third term $(\mathbb{E}[\phi(X^{\mu_{(N,\eta)}}) - \phi(X^\pi)])$. As for the remaining terms, the same bound as Proposition 8 can be applied to the second term $(\mathbb{E}[\phi(X_n^N) - \phi(X^{\mu_{(N,\eta)}})])$, and the last term $\mathbb{E}[\phi(X^\pi)]$ can be bounded by Eq. (20) with Proposition 10. This yields Theorem 5.

Appendix D. Proof of Proposition 7

Proof

First, we show the first assertion about $\{Z_n\}_{n\in\mathbb{N}}$. This is proved in Bréhier (2014) for $\beta = 1$. The $\beta > \eta$ assumption is necessary to ensure that $k(p)$ can be treated as a constant w.r.t β and η in the following. We recall the main arguments of the proof. $\{Z_n\}_{n\in\mathbb{N}}$ is the semi-implicit approximation of the continuous Markov chain defined by:

$$\begin{cases} \mathrm{d}Z(t) = AZ(t)\,\mathrm{d}t + \sqrt{\frac{2}{\beta}}\,\mathrm{d}W(t), \\ Z(0) = 0. \end{cases} \tag{21}$$

Under Assumption 1, it can be shown that $\sup_{t\geq 0}\mathbb{E}(\|Z(t)\|^p) < \infty,\ \forall p \geq 1$. Finally, $\{Z_n\}$ is a numerical scheme with strong order $\frac{1}{4}$ ((Printems, 2001, Theorem 3.2)), which implies the result.

Next, we show the assertion on $\{X_n\}_{n\in\mathbb{N}}$. The discrete chain $Y_n \triangleq X_n - Z_n, n \geq 0$ satisfies

$$Y_{n+1} = S_\eta Y_n - \eta S_\eta \nabla L(X_n).$$

Hence, using Assumption 2 and the fact that $X_n = Y_n + Z_n$, we get

$$\begin{aligned} \|Y_{n+1}\| &\leq \|S_\eta\|_{\mathrm{op}} \|Y_n - \eta\nabla L(X_n)\| \\ &\leq \tfrac{1}{1+\lambda\eta/\mu_0}((1+\eta M)\|Y_n\| + \eta M(\|x^*\| + \|Z_n\|)). \end{aligned}$$

Taking the expectation and using $\mathbb{E}\|Z_n\| \leq k(1)$, this yields

$$\mathbb{E}\|Y_{n+1}\| \leq \tfrac{1}{1+\lambda\eta/\mu_0}((1+\eta M)\mathbb{E}\|Y_n\| + \eta M(\|x^*\| + k(1))),$$

from which we deduce

$$\mathbb{E}\|Y_n\| \leq \rho^n \|x_0\| + \tfrac{\eta(1-\rho^n)M}{(1-\rho)(1+\lambda\eta/\mu_0)}(\|x^*\| + k(1)). \tag{22}$$

Therefore,

$$\mathbb{E}_{x_0}\|X_n\| \leq \rho^n\|x_0\| + \frac{\eta M(\|x^*\| + k(1))}{(1-\rho)(1+\lambda\eta/\mu_0)} + k(1). \tag{23}$$

Finally, we conclude by observing that $\frac{\eta}{1-\rho}\frac{M}{1+\eta\lambda/\mu_0} = 1$.

The proof with bounded gradients is similar. Since

$$\begin{aligned} \|Y_{n+1}\| &\leq \|S_\eta\|_{\mathrm{op}}\|Y_n - \eta\nabla L(X_n)\| \\ &\leq \tfrac{1}{1+\lambda\eta/\mu_0}(\|Y_n\| + \eta B), \end{aligned}$$

we have

$$\|Y_n\| \leq \rho^n\|x_0\| + \frac{(1-\rho^n)}{1-\rho}\frac{\eta B}{1+\lambda\eta/\mu_0} \leq \rho^n\|x_0\| + \frac{\mu_0}{\lambda}B,$$

where $\rho = \frac{1}{1+\lambda\eta/\mu_0}$. Hence, noting $\|X_n\| \leq \|Y_n\| + \|Z_n\|$, we have that $\mathbb{E}[\|X_n\|] \leq \rho^n\|x_0\| + \frac{\mu_0}{\lambda}B + k(1)$ ■

Appendix E. Proof of Proposition 8

Proof under the Strict Dissipativity Condition (Assumption 5 (i)) First we prove the geometric ergodicity under Assumption 5 (i). To show that we first prove the exponential contraction:

$$\|X_n - Y_n\|_{\mathcal{H}} \leq \left(1 - \eta \frac{\frac{\lambda}{\mu_0} - M}{1 + \eta \frac{\lambda}{\mu_0}}\right)^n \|X_0 - Y_0\|_{\mathcal{H}}. \tag{24}$$

Once we have shown this inequality, it is easy to show the geometric ergodicity.

According to the update rule, we have that

$$X_{n+1} = S_\eta \left(X_n - \eta \nabla L(X_n) + \sqrt{\frac{2\eta}{\beta}} \epsilon_n\right),$$
$$Y_{n+1} = S_\eta \left(Y_n - \eta \nabla L(Y_n) + \sqrt{\frac{2\eta}{\beta}} \epsilon_n\right).$$

Therefore, by taking difference, we obtain

$$X_{n+1} - Y_{n+1} = S_\eta \left[(X_n - Y_n) - \eta(L(x_n) - L(Y_n))\right].$$

Then, by the triangular inequality, this yields

$$\begin{aligned}
\|X_{n+1} - Y_{n+1}\|_{\mathcal{H}} &\leq \frac{1}{1 + \eta \frac{\lambda}{\mu_0}} (\|X_n - Y_n\|_{\mathcal{H}} + \eta \|L(X_n) - L(Y_n)\|_{\mathcal{H}}) \\
&\leq \frac{1}{1 + \eta \frac{\lambda}{\mu_0}} (\|X_n - Y_n\|_{\mathcal{H}} + \eta M \|X_n - Y_n\|_{\mathcal{H}}) \\
&\leq \frac{1 + \eta M}{1 + \eta \frac{\lambda}{\mu_0}} \|X_n - Y_n\|_{\mathcal{H}} \\
&\leq \left(1 - \eta \frac{\frac{\lambda}{\mu_0} - M}{1 + \eta \frac{\lambda}{\mu_0}}\right) \|X_n - Y_n\|_{\mathcal{H}} \leq \left(1 - \eta \frac{\frac{\lambda}{\mu_0} - M}{1 + \eta \frac{\lambda}{\mu_0}}\right)^n \|X_0 - Y_0\|_{\mathcal{H}}.
\end{aligned}$$

Now, we already know that there exists an invariant low μ_η under the strong dissipativity condition. By assuming $Y_0 \sim \mu_\eta$ and $X_0 = x_0 \in \mathcal{H}$, we can show the following geometric convergence:

$$\begin{aligned}
\mathbb{E}[\phi(X_n)] - \mathbb{E}_{X \sim \mu_\eta}[\phi(X)] = \mathbb{E}[\phi(X_n)] - \mathbb{E}[\phi(Y_n)] &\leq M' \mathbb{E}[\|X_n - Y_n\|_{\mathcal{H}}] \\
&\leq M' \left(1 - \eta \frac{\frac{\lambda}{\mu_0} - M}{1 + \eta \frac{\lambda}{\mu_0}}\right)^n \mathbb{E}[\|X_0 - Y_0\|_{\mathcal{H}}].
\end{aligned}$$

Now, we see that

$$\mathbb{E}[\|X_0 - Y_0\|_{\mathcal{H}}] \leq \|x_0\|_{\mathcal{H}} + \mathbb{E}[\|Y_0\|_{\mathcal{H}}] \leq \|x_0\|_{\mathcal{H}} + b.$$

In the last inequality, we used that $\mathbb{E}[\|Y_0\|_{\mathcal{H}}] = \mathbb{E}[\|Y_n\|_{\mathcal{H}}] \leq \rho^n \mathbb{E}[\|Y_0\|_{\mathcal{H}}] + b$ for all $n = 1, 2, \ldots$ by Proposition 7 and we took $n \to \infty$. As a consequence, we obtain

$$\mathbb{E}[\phi(X_n)] - \mathbb{E}_{X \sim \mu_\eta}[\phi(X)] \leq M' \exp\left(-n\eta \frac{\frac{\lambda}{\mu_0} - M}{1 + \eta \frac{\lambda}{\mu_0}}\right) (\|x_0\|_{\mathcal{H}} + b),$$

where we used the relation $1 - a \leq \exp(-a)$ for $a > 0$. This yields the assertion.

Proof under the Bounded Gradient Condition (Assumption 5 (ii)) Next, we prove the theorem under the bounded gradient case (Assumption 5 (ii)). Under the strict dissipative condition, the statement can be immediately shown and thus we omit the proof.

We adopt the technique of Theorems 5.2 & 5.3 from Goldys and Maslowski (2006), and show the geometric ergodicity via Theorem 2.5 of Mattingly et al. (2002). We note that Theorem 2.5 of Mattingly et al. (2002) is shown for a finite-dimensional setting, but it can be adopted for an infinite-dimensional setting if the "minorization condition" (Lemma 2.3 of Mattingly et al. (2002)) and "Lyapunov condition" (Assumption 2.2 of Mattingly et al. (2002)) are satisfied.

Since the Lyapunov condition is already shown by Proposition 7, we only need to show the minorization condition. Let $\mu^x_{k,\eta}$ be the law of

$$Z^x_{k,\eta} = S^k_\eta x + \sqrt{\frac{2\eta}{\beta}} \sum_{l=0}^{k-1} S^{k-l}_\eta \varepsilon_l, \tag{25}$$

and $\mu_{k,\eta}$ be the law of

$$Z_{k,\eta} = \sqrt{\frac{2\eta}{\beta}} \sum_{l=0}^{k-1} S^{k-l}_\eta \varepsilon_l. \tag{26}$$

Let $Q \triangleq \frac{2\eta}{\beta} \mathrm{Id}$, and

$$Q_k \triangleq \sum_{l=0}^{k-1} Q S^{2(k-l)}_\eta,$$

for $k = 1, 2, \ldots$, and $Q_0 = 0$. Then, $\mu^x_{k,\eta}$ is the Gaussian process on $\mathcal{H}$ with mean $S^k_\eta x$ and covariance operator Q_k, and $\mu_{k,\eta}$ is the centered Gaussian process on $\mathcal{H}$ with the same covariance operator. By the Cameron-Martin formula, $\mu^x_{k,\eta}$ and $\mu_{k,\eta}$ are equivalent with density given by

$$\frac{\mathrm{d}\mu^x_{k,\eta}}{\mathrm{d}\mu_{k,\eta}}(y) = \exp\left\{ \langle Q_k^{-1} S^k_\eta x, y \rangle - \frac{1}{2} \left\| Q_k^{-1/2} S^k_\eta x \right\|^2 \right\}, \tag{27}$$

(see Da Prato and Zabczyk (1996) for example). We can easily check that $Q_k \succeq k Q S^{2k}_\eta$. Then, we have that

$$\langle x, S^k_\eta Q_k^{-1} y \rangle - \frac{1}{2} \left\| Q_k^{-1/2} S^k_\eta x \right\|^2 \geq -\frac{\beta}{2} \|x\|^2 - \frac{1}{2\beta} \|S^k_\eta Q_k^{-1} y\|^2 - \frac{\beta}{4\eta k} \|x\|^2.$$

and thus we have the following lower bound of the density:

$$\frac{\mathrm{d}\mu^x_{k,\eta}}{\mathrm{d}\mu_{k,\eta}}(y) \geq \exp\left\{ -\frac{\beta}{2} \left(1 + \frac{1}{2k\eta} \right) \|x\|^2 - \frac{1}{2\beta} \|S^k_\eta Q_k^{-1} y\|^2 \right\}. \tag{28}$$

For a given N (where N will be determined later on), let

$$K_k \triangleq Q_k S^{N-k}_\eta Q_N^{-1/2},$$

for $k = 0, \ldots, N$. Here, we define

$$\widehat{Z}^{x,y}_{k,\eta} \triangleq Z^x_{k,\eta} - K_k Q_N^{-1/2} (Z^x_{N,\eta} - y),$$

for $x, y \in \mathcal{H}$, and denote $\widehat{Z}_{k,\eta} \triangleq \widehat{Z}^{0,0}_{k,\eta}$. In particular, we notice that

$$\widehat{Z}_{k,\eta} = Z_{k,\eta} - K_k Q_N^{-1/2} Z_{N,\eta},$$

by definition. Let

$$Y_k \triangleq \sum_{l=k}^{N-1} S_\eta^{N-l} Q^{1/2} \epsilon_l,$$

$$H_k \triangleq Q_{N-k}^{-1/2} S_\eta^{N-k} Q^{1/2}.$$

By a simple calculation, we can show that

$$Y_k = Z_{N,\eta} - S_\eta^{N-k} Z_{k,\eta} = Q_{N-k} Q_N^{-1} Z_{N,\eta} - S_\eta^{N-k} \widehat{Z}_{k,\eta}.$$

Finally, let

$$\alpha_k \triangleq Q_{N-k}^{-1/2} H_k Y_k = \underbrace{Q_{N-k}^{1/2} H_k Q_N^{-1}}_{\triangleq B_1(k)} Z_{N,\eta} - \underbrace{Q_{N-k}^{-1/2} H_k S_\eta^{N-k}}_{\triangleq B_2(k)} \widehat{Z}_{k,\eta},$$

and accordingly, define

$$\zeta_k \triangleq \epsilon_k - \alpha_k.$$

Then, we can show that $(\widehat{Z}_{k,\eta})_k$ and $(\zeta_k)_k$ are independent of $Z_{N,\eta}$ by the same reasoning as Goldys and Maslowski (2006). To see this, we only have to show that their correlation is 0 because they are Gaussian process. First, we can show that[5]

$$\mathbb{E}\left[\epsilon_k \alpha_{k'}^*\right] = Q_{N-k'}^{-1/2} H_{k'} \mathbb{E}\left[\epsilon_k Y_{k'}^*\right] = \begin{cases} Q_{N-k'}^{-1/2} H_{k'} (Q^{1/2} S_\eta^{N-k} - S_\eta^{N-k'} Q^{1/2} S_\eta^{k'-k}) & (k' < k) \\ Q_{N-k'}^{-1/2} H_{k'} (Q^{1/2} S_\eta^{N-k}) & (k' \geq k) \end{cases}$$
$$= \begin{cases} 0 & (k' < k) \\ Q_{N-k'}^{-1/2} H_{k'} Q_{N-k}^{1/2} H_k & (k' \geq k) \end{cases}.$$

For $k \leq k'$,

$$\mathbb{E}\left[\alpha_k \alpha_{k'}^*\right] = Q_{N-k}^{-1/2} H_k \mathbb{E}\left[Y_k Y_{k'}^*\right] H_{k'} Q_{N-k'}^{-1/2} = Q_{N-k}^{-1/2} H_k \left(\sum_{l=k'}^{N-1} S_\eta^{2(N-l)} Q \right) H_{k'} Q_{N-k'}^{-1/2}$$
$$= Q_{N-k}^{-1/2} H_k \left(\sum_{l=0}^{N-k'-1} S_\eta^{2(N-k'-l)} Q \right) H_{k'} Q_{N-k'}^{-1/2}$$
$$= Q_{N-k}^{-1/2} H_k Q_{N-k'} H_{k'} Q_{N-k'}^{-1/2} = Q_{N-k}^{-1/2} H_k Q_{N-k'}^{1/2} H_{k'}.$$

5. Here, for $x, y \in \mathcal{H}$, the bounded linear operator $z \mapsto x\langle y, z\rangle$ is denoted by xy^* for simplicity.

Hence, when $k < k'$, it holds that

$$\mathbb{E}[(\epsilon_k - \alpha_k)(\epsilon_{k'} - \alpha_{k'})^*] = 0,$$

and when $k = k'$, we have that

$$\mathbb{E}[(\epsilon_k - \alpha_k)(\epsilon_k - \alpha_k)^*] = \mathrm{Id} - H_k^2.$$

Finally, we can see that

$$\begin{aligned}\mathbb{E}[(\epsilon_k - \alpha_k)Z_{N,\eta}^*] &= Q^{1/2}S_\eta^{N-k} - \left\{Q_{N-k}^{1/2}H_kQ_N^{-1}Q_N - Q_{N-k}^{-1/2}H_kS_\eta^{N-k}(Q_kS_\eta^{N-k} - K_kQ_N^{-1/2}Q_N)\right\} \\ &= Q^{1/2}S_\eta^{N-k} - Q^{1/2}S_\eta^{N-k} = 0,\end{aligned}$$

which indicates $\zeta_k = \epsilon_k - \alpha_k$ is independent of $Z_{N,\eta}$. Furthermore, we have that

$$\begin{aligned}&\mathbb{E}[Z_{N,\eta}(\widehat{Z}_{k,\eta}^{x,y} - \mathbb{E}[\widehat{Z}_{k,\eta}^{x,y}])^*] = \mathbb{E}[Z_{N,\eta}(\widehat{Z}_{k,\eta}^{x,y})^*] = \mathbb{E}[Z_{N,\eta}Z_{k,\eta}^*] - \mathbb{E}[Z_{N,\eta}Z_{N,\eta}^*Q_N^{-1/2}K_k] \\ &= Q\sum_{l=0}^{k-1}S_\eta^{k-l}S_\eta^{N-l} - Q_NQ_N^{-1/2}K_k = Q_kS_\eta^{N-k} - Q_kS_\eta^{N-k} = 0.\end{aligned}$$

This also yields that $Z_{N,\eta}$ and $\widehat{Z}_{k,\eta}^{x,y}$ $(k = 1, \ldots, N-1)$ are independent.

As we have stated, we now show the minorization condition. Let $P_n^\eta(x, \cdot)$ be the probability measure of the law of X_n with $X_0 = x$, then by the Girsanov's theorem, $P_N^\eta(x, \cdot)$ is absolutely continuous with respect to $\mu_{N,\eta}^x$ and the Radon-Nikodym density is given by

$$\frac{\mathrm{d}P_N^\eta(x,\cdot)}{\mathrm{d}\mu_{N,\eta}^x}(y) = \mathbb{E}\left[\exp\left\{\frac{\beta}{2\eta}\sum_{k=0}^{N-1}\left(\langle -\eta\nabla L(Z_{k,\eta}^x), \epsilon_k\rangle\sqrt{2\eta/\beta} - \frac{\eta^2}{2}\|\nabla L(Z_{k,\eta}^x)\|^2\right)\right\}\Big|Z_{N,\eta}^x = y\right].$$

The right hand side can be evaluated as

$$\begin{aligned}&\mathbb{E}\left[\exp\left\{\frac{\beta}{2\eta}\sum_{k=0}^{N-1}\left(\langle -\eta\nabla L(Z_{k,\eta}^x), \epsilon_k\rangle\sqrt{2\eta/\beta} - \frac{\eta^2}{2}\|\nabla L(Z_{k,\eta}^x)\|^2\right)\right\}\Big|Z_{N,\eta} = y - S_\eta^N x\right] \\ =&\mathbb{E}\left[\exp\left\{\frac{\beta}{2\eta}\sum_{k=0}^{N-1}\left(\langle -\eta\nabla L(Z_{k,\eta}^x), \zeta_k\rangle\sqrt{2\frac{\eta}{\beta}} + \langle -\eta\nabla L(Z_{k,\eta}^x), (B_1(k)Z_{N,\eta} - B_2(k)\widehat{Z}_{k,\eta})\rangle\sqrt{2\frac{\eta}{\beta}}\right.\right. \\ &\left.\left. -\frac{\eta^2}{2}\|\nabla L(Z_{k,\eta}^x)\|^2\right)\right\}\Big|Z_{N,\eta} = y - S_\eta^N x\Bigg] \\ =&\mathbb{E}\left[\exp\left\{\frac{\beta}{2\eta}\sum_{k=0}^{N-1}\left(\langle -\eta\nabla L(\widehat{Z}_{k,\eta}^{x,y}), \zeta_k\rangle\sqrt{2\frac{\eta}{\beta}}\right.\right. \\ &\left.\left. + \langle -\eta\nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k)(y - S_\eta^N x) - B_2(k)\widehat{Z}_{k,\eta}\rangle\sqrt{\frac{2\eta}{\beta}} - \frac{\eta^2}{2}\|\nabla L(\widehat{Z}_{k,\eta}^{x,y})\|^2\right)\right\}\right],\end{aligned}$$

where we used the fact that $(\widehat{Z}_k)_k$ and $(\zeta_k)_k$ are independent of $Z_{N,\eta}$. Therefore, by Jensen's inequality, the right hand side is lower-bounded by

$$\exp\left\{\frac{\beta}{2\eta}\sum_{k=0}^{N-1}\left(\mathbb{E}\left[\langle -\eta\nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k)(y - S_\eta^N x) - B_2(k)\widehat{Z}_{k,\eta}\rangle\right]\sqrt{\frac{2\eta}{\beta}} - \frac{\eta^2}{2}\mathbb{E}[\|\nabla L(\widehat{Z}_{k,\eta}^{x,y})\|^2]\right)\right\}.$$

Thus, by the assumption that $\|\nabla L(\cdot)\| \leq B$, the right hand side is lower-bounded by

$$\begin{aligned}
&\exp\left\{-\sqrt{\frac{\beta}{2\eta}}\sum_{k=0}^{N-1}\left(\mathbb{E}\left[\langle -\eta\nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k)(y-S_\eta^N x)\right] + \eta B\mathbb{E}[\|B_2(k)\widehat{Z}_{k,\eta}\|]\right) - \frac{\beta\eta N}{2}B^2\right\} \\
&\geq \exp\left\{\sqrt{\frac{\beta\eta}{2}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle \nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k)(y-S_\eta^N x)\right] - \frac{\beta\eta N}{2}B^2 - \sum_{k=0}^{N-1}\mathbb{E}[\|B_2(k)\widehat{Z}_{k,\eta}\|^2] - \frac{\beta\eta N}{2}B^2\right\} \\
&\geq \exp\left\{\sqrt{\frac{\beta\eta}{2}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle \nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k)(y-S_\eta^N x)\right] - \beta\eta N B^2 - \sum_{k=0}^{N-1}\mathbb{E}[\|B_2(k)\widehat{Z}_{k,\eta}\|^2]\right\}.
\end{aligned} \tag{29}$$

For $z \in \mathcal{H}$, we have

$$\begin{aligned}
&\sqrt{\frac{\beta}{2\eta}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle \eta\nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k)z\rangle\right] \\
&= \sqrt{\frac{\beta}{2\eta}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle \eta\nabla L(0), B_1(k)z\rangle\right] + \sqrt{\frac{\beta}{2\eta}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle \eta(\nabla L(\widehat{Z}_{k,\eta}^{x,y}) - \nabla L(0)), B_1(k)z\rangle\right] \\
&= \sqrt{\frac{\beta\eta}{2}}\left\langle\left(\sum_{k=0}^{N-1}B_1(k)\right)\nabla L(0), z\right\rangle + \sqrt{\frac{\beta\eta}{2}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle(\nabla L(\widehat{Z}_{k,\eta}^{x,y}) - \nabla L(0)), B_1(k)z\rangle\right].
\end{aligned}$$

The first term of the right hand side can be lower-bounded by

$$-\frac{\beta\eta N}{4} - \frac{1}{2}\sum_{k=0}^{N-1}\langle B_1(k)\nabla L(0), z\rangle^2 = -\frac{\beta\eta N}{4} - \frac{1}{2}\sum_{k=0}^{N-1}\left\langle QS_\eta^{N-k}Q_N^{-1}\nabla L(0), z\right\rangle^2.$$

The second term can be evaluated as

$$\sqrt{\frac{\beta\eta}{2}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle(\nabla L(\widehat{Z}_{k,\eta}^{x,y}) - \nabla L(0)), B_1(k)z\rangle\right] = \sqrt{\frac{\beta\eta}{2}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle(D^2L(\widetilde{Z}_{k,\eta}^{x,y})\cdot\widehat{Z}_{k,\eta}^{x,y}), B_1(k)z\rangle\right],$$

where $\widetilde{Z}_{k,\eta}^{x,y}$ is an intermediate point between $\widehat{Z}_{k,\eta}^{x,y}$ and 0, i.e., there exists $\theta \in [0,1]$ such that $\widetilde{Z}_{k,\eta}^{x,y} = \theta\widehat{Z}_{k,\eta}^{x,y}$. By Assumption 3, this can be further evaluated as

$$\begin{aligned}
&\sqrt{\frac{\beta\eta}{2}}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle(D^2L(\widetilde{Z}_{k,\eta}^{x,y})\cdot\widehat{Z}_{k,\eta}^{x,y}), B_1(k)z\rangle\right] \geq -\sqrt{\frac{\beta\eta}{2}}\sum_{k=0}^{N-1}C_{\alpha,2}\mathbb{E}\left[\|\widehat{Z}_{k,\eta}^{x,y}\|_{\mathcal{H}}\|B_1(k)z\|_\alpha\right] \\
&\geq -\frac{\beta\eta}{4}C_{\alpha,2}^2\sum_{k=0}^{N-1}\mathbb{E}\left[\|\widehat{Z}_{k,\eta}^{x,y}\|_{\mathcal{H}}^2\right] - \frac{1}{2}\sum_{k=0}^{N-1}\|B_1(k)z\|_\alpha^2 \\
&= -\frac{\beta\eta}{4}C_{\alpha,2}^2\sum_{k=0}^{N-1}\mathbb{E}\left[\|\widehat{Z}_{k,\eta}^{x,y}\|_{\mathcal{H}}^2\right] - \frac{1}{2}\sum_{k=0}^{N-1}\|QS_\eta^{N-k}Q_N^{-1}z\|_\alpha^2.
\end{aligned}$$

Here, we have

$$\mathbb{E}\left[\|\widehat{Z}_{k,\eta}^{x,y}\|_{\mathcal{H}}^2\right] = \mathrm{Tr}[Q_k Q_{N-k} Q_N^{-1}] + \|(S_\eta^k - K_k Q_N^{-1/2} S_\eta^N)x + K_k Q_N^{-1/2} y\|_{\mathcal{H}}^2$$
$$\leq \mathrm{Tr}[Q S_\eta^2 (\mathrm{Id} - S_\eta^{2N})^{-1}] + 2\|x\|_{\mathcal{H}}^2 + 2\|y\|_{\mathcal{H}}^2,$$

where we used $\|S_\eta^k - K_k Q_N^{-1/2} S_\eta^N\| \leq 1$ and $\|K_k Q_N^{-1/2}\| \leq 1$. Therefore, we obtain

$$\sqrt{\frac{\beta}{2\eta}} \sum_{k=0}^{N-1} \mathbb{E}\left[\langle \eta \nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k) z\rangle\right]$$
$$\geq -\frac{\beta\eta N}{4} - \frac{\beta\eta}{4} C_{\alpha,2}^2 \sum_{k=0}^{N-1} (\mathrm{Tr}[Q S_\eta^2 (\mathrm{Id} - S_\eta^{2N})^{-1}] + 2\|x\|_{\mathcal{H}}^2 + 2\|y\|_{\mathcal{H}}^2)$$
$$- \frac{1}{2} \sum_{k=0}^{N-1} (\left\langle Q S_\eta^{N-k} Q_N^{-1} \nabla L(0), z \right\rangle^2 + \|Q S_\eta^{N-k} Q_N^{-1} z\|_\alpha^2).$$

Next we give another bound for $z = S_\eta^N x$. In this situation, thanks to the factor S_η^N, we have a simpler bound:

$$\sqrt{\frac{\beta}{2\eta}} \sum_{k=0}^{N-1} \mathbb{E}\left[\langle \eta \nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k) S_\eta^N x\rangle\right]$$
$$\geq -\sqrt{\frac{\beta\eta}{2}} \sum_{k=0}^{N-1} B \|B_1(k) S_\eta^N x\| \geq -\frac{\beta\eta N}{4} B^2 + \frac{1}{2} \sum_{k=0}^{N-1} \|B_1(k) S_\eta^N x\|^2.$$

Notice that

$$\sum_{k=0}^{N-1} B_1(k)^2 = \sum_{k=0}^{N-1} (Q_{N-k}^{1/2} H_k Q_N^{-1})^2 = \sum_{k=1}^{N-1} Q_{N-k} H_k^2 Q_N^{-2} = \sum_{k=0}^{N-1} Q_{N-k} Q_{N-k}^{-1} S_\eta^{2(N-k)} Q Q_N^{-2}$$
$$= \sum_{k=0}^{N-1} S_\eta^{2(N-k)} Q Q_N^{-2} = Q_N^{-1}.$$

Therefore, $\sum_{k=0}^{N-1} \|B_1(k) S_\eta^N x\|^2$ can be bounded as

$$\sum_{k=0}^{N-1} \|B_1(k) S_\eta^N x\|^2 = \|Q_N^{-1/2} S_\eta^N x\|^2 \leq \frac{1}{N} Q^{-1} \|x\|^2 = \frac{\beta}{2N\eta} \|x\|^2,$$

where we used $Q_N \succeq N Q S_\eta^{2N}$ and $Q = \frac{2\eta}{\beta}\,\mathrm{Id}$. Therefore, we have

$$\sqrt{\frac{\beta}{2\eta}} \sum_{k=0}^{N-1} \mathbb{E}\left[\langle \eta \nabla L(\widehat{Z}_{k,\eta}^{x,y}), B_1(k) S_\eta^N x\rangle\right] \geq -\frac{\beta\eta N}{4} B^2 - \frac{\beta}{4N\eta} \|x\|^2.$$

$$\sum_{k=0}^{N-1} \mathbb{E}[\|B_2(k) \widehat{Z}_{k,\eta}\|^2] = \sum_{k=0}^{N-1} \mathrm{Tr}[(Q_{N-k}^{-1} S_\eta^{2(N-k)} Q^{1/2})^2 (Q_k - 2Q_k^2 Q_N^{-1} S_\eta^{2(N-k)} + Q_N)]$$

$$
\begin{aligned}
&\leq \sum_{k=0}^{N-1} \mathrm{Tr}[(Q_{N-k}^{-1} S_\eta^{2(N-k)} Q^{1/2})^2 (Q_k - 2Q_k Q_N^{-1} Q_N + Q_N)] \\
&= \sum_{k=0}^{N-1} \mathrm{Tr}[(Q_{N-k}^{-1} S_\eta^{2(N-k)} Q^{1/2})^2 (Q_N - Q_k)] \\
&\leq \sum_{k=0}^{N-1} \mathrm{Tr}\left[Q_{N-k}^{-2} S_\eta^{4(N-k)} Q \left(\sum_{l=0}^{N-k-1} S_\eta^{N-l} \right) \right] = \sum_{k=0}^{N-1} \mathrm{Tr}[Q_{N-k}^{-2} S_\eta^{4(N-k)} Q Q_{N-k} S_\eta^{2k}] \\
&= \sum_{k=0}^{N-1} \mathrm{Tr}[Q_{N-k}^{-1} S_\eta^{2N} S_\eta^{2(N-k)} Q] = \sum_{k=0}^{N-1} \mathrm{Tr}\{(S_\eta^{-2} - \mathrm{Id})[Q(\mathrm{Id} - S_\eta^{2(N-k)})]^{-1} S_\eta^{2N} S_\eta^{2(N-k)} Q\} \\
&= \mathrm{Tr}\left[(S_\eta^{-2} - \mathrm{Id}) S_\eta^{2N} \sum_{k=0}^{N-1} (S_\eta^{-2(N-k)} - \mathrm{Id})^{-1} \right] \leq \mathrm{Tr}\left[(S_\eta^{-2} - \mathrm{Id}) S_\eta^{2N} (S_\eta^{-2} - \mathrm{Id})^{-1} \sum_{k=0}^{N-1} S_\eta^{2k} \right] \\
&= \mathrm{Tr}\left[(S_\eta^{-2} - \mathrm{Id}) S_\eta^{2N} (S_\eta^{-2} - \mathrm{Id})^{-1} (S_\eta^{2N} - \mathrm{Id}) (S_\eta^2 - \mathrm{Id})^{-1} \right] = \mathrm{Tr}\left[S_\eta^{2N} (S_\eta^{2N} - \mathrm{Id}) (S_\eta^2 - \mathrm{Id})^{-1} \right] \\
&\leq \mathrm{Tr}\left[(S_\eta^{4N} - S_\eta^{2N}) (S_\eta^2 - \mathrm{Id})^{-1} \right] \leq \mathrm{Tr}\left[(S_\eta^{2N+2} - S_\eta^{2N}) (S_\eta^2 - \mathrm{Id})^{-1} \right] \quad (\because N \geq 1) \\
&\leq \mathrm{Tr}\left[S_\eta^{2N} \right] \leq \mathrm{Tr}\left[(\mathrm{Id} + 2N\eta A)^{-1} \right].
\end{aligned}
$$

Therefore, we obtain, for all $y \in \mathrm{Im}(Q_N^{1/2})$,

$$
\begin{aligned}
&\frac{\mathrm{d}P_N^\eta(x,\cdot)}{\mathrm{d}\mu_{N,\eta}^x}(y) \\
&\quad \geq \exp\Bigg\{ -\frac{\beta\eta N}{4} - \frac{\beta\eta}{4} C_{\alpha,2}^2 \sum_{k=0}^{N-1} (\mathrm{Tr}[Q S_\eta^2 (\mathrm{Id} - S_\eta^{2N})^{-1}] + 2\|x\|_{\mathcal{H}}^2 + 2\|y\|_{\mathcal{H}}^2) \\
&\qquad\quad - \frac{1}{2} \sum_{k=0}^{N-1} \left(\left\langle Q S_\eta^{N-k} Q_N^{-1} \nabla L(0), y \right\rangle^2 + \|Q S_\eta^{N-k} Q_N^{-1} y\|_\alpha^2 \right) \\
&\qquad\quad - \frac{\beta\eta N}{4} B^2 - \frac{\beta}{4N\eta} \|x\|^2 \\
&\qquad\quad - \beta\eta N B^2 - \mathrm{Tr}\left[(\mathrm{Id} + 2N\eta A)^{-1} \right] \Bigg\} \\
&\quad \geq \exp\Bigg\{ \underbrace{ -\frac{\beta\eta N(1+5B^2)}{4} - \frac{\beta\eta N}{4} C_{\alpha,2}^2 \mathrm{Tr}[Q S_\eta^2 (\mathrm{Id} - S_\eta^{2N})^{-1}] - \mathrm{Tr}\left[(\mathrm{Id} + 2N\eta A)^{-1} \right] }_{=: -C_{\eta,N,\beta}} \\
&\qquad \underbrace{ -\left(\frac{\beta\eta N}{2} C_{\alpha,2}^2 + \frac{\beta}{4N\eta} \right) \|x\|^2 }_{=: -\tilde{\Lambda}_{\mathrm{x}}(x)} \\
&\qquad \underbrace{ -\frac{\beta\eta N}{2} C_{\alpha,2}^2 \|y\|^2 - \frac{1}{2} \sum_{k=0}^{N-1} (\left\langle Q S_\eta^{N-k} Q_N^{-1} \nabla L(0), y \right\rangle^2 + \|Q S_\eta^{N-k} Q_N^{-1} y\|_\alpha^2) }_{=: -\tilde{\Lambda}_{\mathrm{y}}(y)} \Bigg\}.
\end{aligned}
$$

Combining the inequalities (28) and (29), we finally obtain that

$$\begin{aligned}\frac{\mathrm{d}P_N^\eta(x,\cdot)}{\mathrm{d}\mu_{N,\eta}}(y) &= \frac{\mathrm{d}P_N^\eta(x,\cdot)}{\mathrm{d}\mu_{N,\eta}^x}(y)\frac{\mathrm{d}\mu_{N,\eta}^x}{\mathrm{d}\mu_{N,\eta}}(y)\\ &\geq \exp\left\{-\frac{\beta}{2}\left(1+\frac{1}{2\eta N}\right)\|x\|^2 - \frac{1}{4\beta}\|S_\eta^N Q_N^{-1} y\|^2 - C_{\eta,N,\beta} - \tilde{\Lambda}_{\mathrm{x}}(x) - \tilde{\Lambda}_{\mathrm{y}}(y)\right\}. \end{aligned} \tag{30}$$

Fro now on, we give a lower bound of the right hand side. To do so, we set $N = 1/\eta$. Under this setting, let $\Lambda_{\mathrm{x}}(x) := \frac{\beta}{4}\|x\|^2 + \tilde{\Lambda}_{\mathrm{x}}(x)$ and $\Lambda_{\mathrm{y}}(y) := \frac{1}{4\beta}\|S_\eta^N Q_N^{-1} y\|^2 + \tilde{\Lambda}_{\mathrm{y}}(y)$, i.e.,

$$\frac{\mathrm{d}P_N^\eta(x,\cdot)}{\mathrm{d}\mu_{N,\eta}}(y) \geq \exp\left\{-C_{\eta,N,\beta} - \Lambda_{\mathrm{x}}(x) - \Lambda_{\mathrm{y}}(y)\right\}. \tag{31}$$

We evaluate the terms in the exponent in the right hand side one by one.

(i) (Bound of $C_{\eta,N,\beta}$): Note that

$$\begin{aligned}\|(\mathrm{Id} - S_\eta^{2N})^{-1}\|_{\mathcal{B}(\mathcal{H})} &\leq [1-(1+\eta\lambda/\mu_0)^{-2N}]^{-1} \leq (1+\eta\lambda/\mu_0)^{2N}[(1+\eta\lambda/\mu_0)^{2N}-1]^{-1}\\ &\leq \frac{\exp(2N\eta\lambda/\mu_0)}{2N\eta\lambda/\mu_0} = \frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0}, \end{aligned} \tag{32}$$

and thus

$$\begin{aligned}\mathrm{Tr}[QS_\eta^2(\mathrm{Id}-S_\eta^{2N})^{-1}] &= \frac{2\eta}{\beta}\mathrm{Tr}[S_\eta^2(\mathrm{Id}-S_\eta^{2N})^{-1}] \leq \frac{2\eta}{\beta}\mathrm{Tr}[S_\eta^2]\|S_\eta^2(\mathrm{Id}-S_\eta^{2N})^{-1}\|_{\mathcal{B}(\mathcal{H})}\\ &\leq \frac{2\eta}{\beta}\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0}\sum_{k=0}^{\infty}(1+\eta\lambda/\mu_k)^{-2} \leq C_\mu\frac{2\eta}{\beta}\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0}\sqrt{\frac{1}{\eta\lambda}}\\ &= C_\mu\frac{\sqrt{\eta}}{\beta}\frac{\mu_0\exp(2\lambda/\mu_0)}{\lambda^{3/2}}, \end{aligned}$$

where C_μ is a constant depending on $(\mu_k)_{k=1}^\infty$ and we used $\mu_k \lesssim 1/k^2$ in the last inequality. This converges to 0 as $\eta \to 0$ and $\beta \to \infty$, thus $\mathrm{Tr}[QS_\eta^2(\mathrm{Id}-S_\eta^{2N})^{-1}] = O(1)$. Consequently, we have

$$C_{\eta,N,\beta} = \frac{\beta(1+5B^2)}{4} + \frac{1}{4}C_{\alpha,2}^2 C_\mu\sqrt{\eta}\frac{\mu_0\exp(2\lambda/\mu_0)}{\lambda^{3/2}} + \mathrm{Tr}\left[(\mathrm{Id}+2A)^{-1}\right] = O(\beta).$$

(ii) (Bound of $\Lambda_{\mathrm{x}}(x)$): By the definition of $\Lambda_{\mathrm{x}}(x)$, it holds that

$$\Lambda_{\mathrm{x}}(x) = \left(\frac{\beta}{2}\left(1+\frac{1}{2\eta N}\right) + \frac{\beta\eta N}{2}C_{\alpha,2}^2 + \frac{\beta}{4N\eta}\right)\|x\|^2 = \left(\beta + \frac{\beta}{2}C_{\alpha,2}^2\right)\|x\|^2 = O(\beta\|x\|^2).$$

(ii) (Bound of $\Lambda_{\mathrm{y}}(y)$): Finally, we evaluate $\Lambda_{\mathrm{y}}(y)$. When $\eta = 1/N$,

$$\Lambda_{\mathrm{y}}(y) = \frac{1}{4\beta}\|S_\eta^N Q_N^{-1} y\|^2 + \frac{\beta}{2}C_{\alpha,2}^2\|y\|^2 + \frac{1}{2}\sum_{k=0}^{N-1}(\left\langle QS_\eta^{N-k}Q_N^{-1}\nabla L(0), y\right\rangle^2 + \|QS_\eta^{N-k}Q_N^{-1}y\|_\alpha^2).$$

We can show that $\Lambda_{\mathrm{y}}(Z) < \infty$ for $Z \sim \mu_{N,\eta}$ almost surely, as follows. Since $0 \leq \Lambda_{\mathrm{y}}(y)$, we only have to evaluate $\mathbb{E}_{Z\sim\mu_{N,\eta}}[\Lambda_{\mathrm{y}}(Z)]$. To do so, we note that $\mu_{N,\eta}$ is a Gaussian process in $\mathcal{H}$

with mean 0 and covariance Q_N, which can be easily checked by its definition. By using this, we evaluate the expectation of each term as follows.

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mu_{N,\eta}}\left[\frac{1}{4\beta}\|S_\eta^N Q_N^{-1} y\|^2\right] = \frac{1}{4\beta}\mathrm{Tr}[S_\eta^{2N} Q_N^{-2} Q_N] = \frac{1}{4\beta}\mathrm{Tr}[S_\eta^{2N} Q_N^{-1}] \\
&= \frac{1}{4\beta}\mathrm{Tr}[S_\eta^{2N}(S_\eta^2 + \cdots + S_\eta^{2N})^{-1} Q^{-1}] \\
&= \frac{1}{4\beta}\mathrm{Tr}[Q^{-1} S_\eta^{2N}(\mathrm{Id} - S_\eta^2)(S_\eta^2(\mathrm{Id} - S_\eta^{2N}))^{-1}] \\
&\leq \frac{1}{4\beta}\mathrm{Tr}[Q^{-1} S_\eta^{2N}(\mathrm{Id} - S_\eta^2)(S_\eta^2)^{-1}]\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0} \quad (\because \text{Eq. (32)}) \\
&= \frac{1}{4\beta}\mathrm{Tr}[Q^{-1} S_\eta^{2N}(S_\eta^{-2} - \mathrm{Id})]\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0} \\
&\leq \frac{1}{8\eta}\mathrm{Tr}[S_\eta^{2N}(2\eta A + \eta^2 A^2)]\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0} \\
&= \frac{1}{8}\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0}\sum_{k=0}^{\infty}\frac{2\lambda/\mu_k + \eta(\lambda/\mu_k)^2}{(1+\eta\lambda/\mu_k)^{2N}} \\
&= \frac{1}{8}\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0}\sum_{k=0}^{\infty}\frac{1}{(1+\eta\lambda/\mu_k)^N}\left(\frac{2\lambda/\mu_k}{(1+\eta\lambda/\mu_k)^N} + \frac{\eta(\lambda/\mu_k)^2}{(1+\eta\lambda/\mu_k)^{\frac{N}{2}2}}\right) \\
&\leq \frac{1}{8}\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0}\sum_{k=0}^{\infty}\frac{1}{(1+\lambda/\mu_k)}\left(\frac{2\lambda/\mu_k}{(1+\lambda/\mu_k)} + \frac{\eta(\lambda/\mu_k)^2}{(1+\frac{1}{2}\lambda/\mu_k)^2}\right) \\
&\leq \frac{1}{8}\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0}\sum_{k=0}^{\infty}\frac{1}{(1+\lambda/\mu_k)}(2+4\eta) \\
&\leq \frac{1+2\eta}{4}\frac{\exp(2\lambda/\mu_0)}{2\lambda/\mu_0}\frac{C'_\mu}{\sqrt{\lambda}} = O(1),
\end{aligned}$$

where C'_μ is a constant depending only on $(\mu_k)_k$ and we again used $\mu_k \lesssim 1/k^2$ in the last inequality.

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mu_{N,\eta}}\left[\frac{\beta}{2}C_{\alpha,2}^2\|Z\|^2\right] = \frac{\beta}{2}C_{\alpha,2}^2\mathrm{Tr}[Q_N] = \frac{\beta}{2}C_{\alpha,2}^2\mathrm{Tr}[Q(S_\eta^2 - S_\eta^{2(N+1)})(\mathrm{Id} - S_\eta^2)^{-1}] \\
&= \eta C_{\alpha,2}^2\mathrm{Tr}[(\mathrm{Id} - S_\eta^{2N})(S_\eta^{-2} - \mathrm{Id})^{-1}] \\
&= \eta C_{\alpha,2}^2\sum_{k=0}^{\infty}(1 - (1+\eta\lambda/\mu_k)^{-2N})(2\eta\lambda/\mu_k + \eta^2(\lambda/\mu_k)^2)^{-1} \\
&\leq \eta C_{\alpha,2}^2\sum_{k=0}^{\infty}(2\eta\lambda/\mu_k)^{-1} = \frac{C_{\alpha,2}^2}{2\lambda}\sum_{k=0}^{\infty}\mu_k \leq \frac{C_{\alpha,2}^2}{2\lambda}C''_\mu = O(1),
\end{aligned}$$

where C''_μ is a constant depending only on $(\mu_k)_k$ and we again used $\mu_k \lesssim 1/k^2$ in the last inequality.

$$\mathbb{E}_{Z\sim\mu_{N,\eta}}\left[\sum_{k=0}^{N-1}\left(\left\langle QS_\eta^{N-k}Q_N^{-1}\nabla L(0), Z\right\rangle^2 + \|QS_\eta^{N-k}Q_N^{-1}Z\|_\alpha^2\right)\right]$$

$$= \sum_{k=0}^{N-1} \left\{ \left\langle Q S_\eta^{N-k} Q_N^{-1} \nabla L(0), Q_N Q S_\eta^{N-k} Q_N^{-1} \nabla L(0) \right\rangle + \mathrm{Tr}[Q S_\eta^{N-k} Q_N^{-1} \sqrt{Q_N}\, \mathrm{Id}_\alpha\, \sqrt{Q_N} Q S_\eta^{N-k} Q_N^{-1}] \right\}$$

where $\mathrm{Id}_\alpha : \mathcal{H} \to \mathcal{H}$ is a linear operator defined by $\langle x, \mathrm{Id}_\alpha\, y \rangle = \sum_{k=0}^\infty (\mu_k)^{2\alpha} x_k y_k$ for $x = (x_k)_k, y = (y_k)_k \in \mathcal{H}$. The first term in the right hand side can be evaluated as

$$\left\langle Q S_\eta^{N-k} Q_N^{-1} \nabla L(0), Q_N Q S_\eta^{N-k} Q_N^{-1} \nabla L(0) \right\rangle = \left\langle Q^2 S_\eta^{2(N-k)} Q_N^{-1} \nabla L(0), \nabla L(0) \right\rangle,$$

and its summation becomes

$$\begin{aligned} &\sum_{k=0}^\infty \left\langle Q^2 S_\eta^{2(N-k)} Q_N^{-1} \nabla L(0), \nabla L(0) \right\rangle = \langle Q Q_N Q_N^{-1} \nabla L(0), \nabla L(0) \rangle \\ &= \langle Q \nabla L(0), \nabla L(0) \rangle = \frac{2\eta}{\beta} \|L(0)\|^2 = O(\eta/\beta). \end{aligned}$$

The second term can be evaluated as

$$\begin{aligned} &\sum_{k=0}^\infty \mathrm{Tr}[Q S_\eta^{N-k} Q_N^{-1} \sqrt{Q_N}\, \mathrm{Id}_\alpha\, \sqrt{Q_N} Q S_\eta^{N-k} Q_N^{-1}] = \sum_{k=0}^\infty \mathrm{Tr}[Q^2 S_\eta^{2(N-k)} Q_N^{-1}\, \mathrm{Id}_\alpha] \\ &= \sum_{k=0}^\infty \mathrm{Tr}[Q Q_N Q_N^{-1}\, \mathrm{Id}_\alpha] = \sum_{k=0}^\infty \frac{2\eta}{\beta} \mathrm{Tr}[\mathrm{Id}_\alpha] = \frac{2\eta}{\beta} \sum_{k=0}^\infty \mu_k^{2\alpha} = \frac{2\eta}{\beta} C_{\mu,\alpha} = O(\eta/\beta), \end{aligned}$$

where we used the assumption $\alpha > 1/4$ and $\mu_k \lesssim 1/k^2$. Summarizing the above arguments, we obtain that

$$\mathbb{E}_{Z \sim \mu_{N,\eta}}[\Lambda_{\mathrm{y}}(Z)] \le O(1). \tag{33}$$

(iv) (Combining all bounds (i), (ii), (iii)). Combining these bounds for $C_{\eta,N,\beta}, \Lambda_{\mathrm{x}}(x), \Lambda_{\mathrm{y}}(y)$, we may give a lower bound of $P_N^\eta(x, \Gamma)$ for a measurable set $\Gamma \subset \mathcal{H}$ uniformly for all x with norm smaller than a given R, which is required to show the minorization condition. Let

$$c_R \triangleq \exp\left(-C_{\eta,N,\beta} - \frac{\beta}{2}(2 + C_{\alpha,2}^2) R^2 \right)$$

for $R \ge \frac{3}{2} k(1)$ which will be determined later, then we have shown that for all $x \in \mathcal{H}$ with $\|x\| \le R$,

$$\exp(-C_{\eta,N,\beta} - \Lambda_{\mathrm{x}}(x)) \ge c_R.$$

By Eq. (30), this gives that

$$P_N^\eta(x, \Gamma) \ge c_R \int_\Gamma e^{-\Lambda_{\mathrm{y}}(z)} \mu_{N,\eta}(dz),$$

for all $x \in \mathcal{B}_R$ and a measurable set $\Gamma \subset \mathcal{H}$. In particular, if we define

$$\bar{\mu}(\Gamma) \triangleq \frac{1}{\bar{Z}} \int_{\Gamma \cap \mathcal{B}_R} e^{-\Lambda_{\mathrm{y}}(z)} \mu_{N,\eta}(dz)$$

where $\bar{Z} = \int_{\mathcal{B}_R} e^{-\Lambda_y(z)}\mu_{N,\eta}(dz)$ so that $\bar{\mu}$ is a probability measure, then

$$P_N^\eta(x,\Gamma) \geq c_R \int_{\Gamma\cap\mathcal{B}_R} e^{-\Lambda_y(z)}\mu_{N,\eta}(dz) \geq \delta\bar{\mu}(\Gamma), \tag{34}$$

where

$$\delta \triangleq c_R\bar{Z}.$$

Here, we give a lower bound of δ. By Proposition 7,

$$\mu_{N,\eta}(\mathcal{B}_R) \geq 1 - \frac{\mathbb{E}_{Z\sim\mu_{N,\eta}}[\|Z\|]]}{R} \geq 1 - \frac{1}{R}k(1) \geq \frac{1}{3},$$

where we used $R \geq \frac{3}{2}k(1)$ and thus δ can be lower-bounded as

$$\begin{aligned}
\delta &= c_R\int_{\mathcal{B}_R} e^{-\Lambda_y(z)}\mu_{N,\eta}(dz) = c_R\mu_{N,\delta}(\mathcal{B}_R)\frac{1}{\mu_{N,\delta}(\mathcal{B}_R)}\int_{\mathcal{B}_R} e^{-\Lambda_y(z)}\mu_{N,\eta}(dz)\\
&\geq c_R\mu_{N,\delta}(\mathcal{B}_R)\exp\left(-\frac{1}{\mu_{N,\delta}(\mathcal{B}_R)}\int_{\mathcal{B}_R}\Lambda_y(z)\mu_{N,\eta}(dz)\right)\\
&\geq \frac{1}{3}c_R\exp\left(-2\int_{\mathcal{H}}\Lambda_y(z)\mu_{N,\eta}(dz)\right)\\
&\geq \frac{1}{3}c_R\exp\left(-O(1)\right),
\end{aligned}$$

where we used Eq. (33) in the final inequality. Therefore, we have shown that there exists a probability measure $\bar{\mu}$, with $\bar{\mu}(\mathcal{B}_R) = 1$ and $\bar{\mu}(\mathcal{B}_R^c) = 0$, such that Eq. (34) is satisfied for any $x \in \mathcal{B}_R$ and a measurable set $\Gamma \in \mathbb{B}(\mathcal{H})$, where $\delta \geq \frac{1}{3}c_R\exp(-O(1)) \geq \frac{1}{3}\exp(-C_{\eta,N,\beta} - \Lambda_x(x) - O(1)) \gtrsim \exp(-O(\beta))$.

By Proposition 7, the following contraction condition holds for $\alpha_N = \rho^N = (\frac{1}{1+\lambda\eta/\mu_0})^N \leq \exp(-\lambda/\mu_0) < 1$, $\bar{b} = \max\{\frac{\mu_0}{\lambda}B + k(1), 1\}$ under the bounded gradient condition:

$$\mathbb{E}_{x_0}\|X_N\| \leq \alpha_N\|x_0\| + \bar{b} \quad (\forall n \in \mathbb{N}).$$

Set $V(x) = \|x\| + 1$ and $\mathcal{C} = \left\{x \in \mathcal{H} \mid V(x) \leq \frac{2\bar{b}}{\sqrt{(1+\alpha_N)/2}-\alpha_N}\right\}$, then we have that $\mathcal{C} = \mathcal{B}_R$ for $R = \frac{2\bar{b}}{\sqrt{(1+\alpha_N)/2}-\alpha_N} - 1$. Here, we give lower and upper bounds of R. As for the lower bound, we can easily see that $R \geq \frac{5}{2}\bar{b} - 1 \geq \frac{3}{2}\bar{b} \geq \frac{3}{2}k(1)$. Next, we give an upper bound. Jensen's inequality and the fact $0 < \alpha_N < 1$ yield $\sqrt{(1+\alpha_N)/2} - \alpha_N \geq \frac{1+\sqrt{\alpha_N}}{2} - \sqrt{\alpha_N} = \frac{1-\sqrt{\alpha_N}}{2}$. Here for $a > 0$, it is easy to see $(1+a)^{N/2} \geq 1 + aN/2$ and thus we have $1 - (1+a)^{-N/2} \geq 1 - (1+aN/2)^{-1} = \frac{aN/2}{1+aN/2}$. Substituting $a = \lambda\eta/\mu_0$, $\frac{1-\sqrt{\alpha_N}}{2} \geq \frac{N\lambda\eta/(2\mu_0)}{1+N\lambda\eta/(2\mu_0)}$. Then, by using $\eta = 1/N$, we obtain that $\frac{2\bar{b}}{\sqrt{(1+\alpha_N)/2}-\alpha_N} \leq \frac{4b\mu_0(1+\lambda/(2\mu_0))}{\lambda} = 2\bar{b}(1 + 2\frac{\mu_0}{\lambda})$.

Then, Theorem 2.5 of Mattingly et al. (2002) asserts that there exits a invariant measure μ^η for the Markov chain $(X_{lN})_l$ and the chain satisfies the geometric ergodicity: for $\phi : \mathcal{H} \to \mathbb{R}$ such that $|\phi(\cdot)| \leq V(\cdot)$,

$$\mathbb{E}[\phi(X_{lN})] - \mathbb{E}_{X\sim\mu^\eta}[\phi(X)] \leq \kappa[\bar{V}+1](1-\delta)^{al} + \sqrt{2}V(x_0)\gamma^l(\kappa[\bar{V}+1])^{al}\frac{1}{\sqrt{\delta}}, \tag{35}$$

where $\kappa = \bar{b} + 1$, $\bar{V} = 2\sup_{x \in \mathcal{C}} V(x) = \frac{4\bar{b}}{\sqrt{(1+\alpha_N)/2} - \alpha_N}$, $\gamma = \sqrt{(\alpha_N + 1)/2}$ and $a \in (0, 1)$ so that $\gamma(\kappa[\bar{V} + 1])^a \leq (1 - \delta)^a$. In particular, we may choose $a \in (0, 1)$ as

$$a = \frac{\log(1/\gamma)}{\log(\kappa(\bar{V} + 1)/(1 - \delta))}.$$

Here, by noting that

$$\begin{aligned}\log(1/\gamma) &= -\frac{1}{2}\log\left(\frac{1 + \alpha_N}{2}\right) = -\frac{1}{2}\log\left(1 - \frac{1 - \alpha_N}{2}\right) \\ &\geq \frac{1}{2}\left(\frac{1 - \alpha_N}{2}\right) \geq \frac{1}{4}\min\left(\frac{\lambda}{2\mu_0}, \frac{1}{2}\right) = \Omega(\lambda/\mu_0),\end{aligned}$$

we may assume

$$a \geq \frac{\min\left(\frac{\lambda}{2\mu_0}, \frac{1}{2}\right)}{4\log(\kappa(\bar{V} + 1)/(1 - \delta))}.$$

Then Eq. (35) is simplified to

$$\mathbb{E}[\phi(X_{lN})] - \mathbb{E}_{X \sim \mu^\eta}[\phi(X)] \leq \left(\kappa[\bar{V} + 1] + \frac{\sqrt{2}V(x_0)}{\sqrt{\delta}}\right)(1 - \delta)^{al}. \tag{36}$$

This shows the geometric ergodicity of the sequence $(X_{lN})_{l=1}^\infty$. To extend this result to "unsampled" sequence $(X_n)_{n=1}^\infty$, we may apply the same argument to the sequence $(X_{lN+n})_{l=0}^\infty$ for each $n = 1, \ldots, N - 1$. Applying Eq. (35) where x_0 is replaced with X_n and taking expectation with respect to X_n, we have

$$\begin{aligned}&\mathbb{E}[\phi(X_{lN+n})] - \mathbb{E}_{X \sim \mu^\eta}[\phi(X)] \qquad (37)\\ &\leq \left(\kappa[\bar{V} + 1] + \frac{\sqrt{2}\mathbb{E}[V(X_n)]}{\sqrt{\delta}}\right)(1 - \delta)^{al} \\ &\leq \left(\kappa[\bar{V} + 1] + \frac{\sqrt{2}(\rho^n\|x_0\| + b + 1)}{\sqrt{\delta}}\right)(1 - \delta)^{al} \quad (\because \text{Proposition 7}) \\ &\leq \left(\kappa[\bar{V} + 1] + \frac{\sqrt{2}(V(x_0) + b)}{\sqrt{\delta}}\right)(1 - \delta)^{al}. \qquad (38)\end{aligned}$$

Finally, we note that for $0 \leq n < N$,

$$\begin{aligned}(1 - \delta)^{al} &\leq (1 - \delta)^{a(lN+n-N)/N} \leq (1 - \delta)^{a[(lN+n)/N - 1]} \leq \exp\left(-\delta a[(lN + n)/N - 1]\right) \\ &\leq \exp\left(-\Lambda_\eta^*[\eta(lN + n) - 1]\right),\end{aligned}$$

where we set

$$\Lambda_\eta^* \triangleq a\delta \geq \frac{\min\left(\frac{\lambda}{2\mu_0}, \frac{1}{2}\right)}{4\log(\kappa(\bar{V} + 1)/(1 - \delta))}\delta = \Omega(\exp(-O(\beta))).$$

This yields the assertion. □

Appendix F. Proof of Proposition 10

Lemma 13 (Gaussian correlation inequality) *Let ν_∞ be the Gaussian measure in $\mathcal{H}$ given by a random variable $\sum_{i=0}^\infty \xi_i\gamma_i f_i$ where $(\xi)_{i=0}^\infty$ is a sequence of i.i.d. standard normal variables and $(\gamma_i)_{i=0}^\infty$ is a sequence of real variables with $0 < \sum_{i=0}^\infty \gamma_i^2 < \infty$. For two sets $\mathcal{C}^1 = \{X = \sum_{i=0}^\infty \alpha_i f_i \in \mathcal{H} \mid \sum_{i=0}^\infty \alpha_i^2 \mu_i^{(1)} \leq 1\}$ and $\mathcal{C}^2 = \{X = \sum_{i=0}^\infty \alpha_i f_i \in \mathcal{H} \mid |\sum_{i=0}^\infty \alpha_i \mu_i^{(2)}| \leq 1\}$ where $(\mu_i^{(1)})_{i=1}^\infty$ is a fixed non-negative sequence and $(\mu_i^{(2)})_{i=1}^\infty$ is a fixed sequence of real numbers satisfying $\sum_{i=0}^\infty (\mu_i^{(2)})^2 < \infty$, we have*

$$\nu_\infty(\mathcal{C}^1 \cap \mathcal{C}^2) \geq \nu_\infty(\mathcal{C}^1)\nu_\infty(\mathcal{C}^2).$$

Proof Let $\mathcal{C}_n^1$ an $\mathcal{C}_n^2$ be the cylinder set that "truncates" $\mathcal{C}^1$ an $\mathcal{C}^2$ up to index n: $\mathcal{C}_n^1 = \{X = \sum_{i=0}^\infty \alpha_i f_i \in \mathcal{H} \mid \sum_{i=0}^n \alpha_i^2 \mu_i^{(1)} \leq 1\}$ and $\mathcal{C}_n^2 = \{X = \sum_{i=0}^\infty \alpha_i f_i \in \mathcal{H} \mid |\sum_{i=0}^n \alpha_i \mu_i^{(2)}| \leq 1\}$. By the Gaussian correlation inequality (Royen, 2014; Latała and Matlak, 2017), it holds that

$$\nu_\infty(\mathcal{C}_n^1 \cap \mathcal{C}_n^2) \geq \nu_\infty(\mathcal{C}_n^1)\nu_\infty(\mathcal{C}_n^2).$$

We note that $(\mathcal{C}_n^1)_n$ is a monotonically decreasing sequence, i.e., $\mathcal{C}_n^1 \subseteq \mathcal{C}_m^1$ for $m < n$, and we see that $\cap_{n=1}^\infty \mathcal{C}_n^1 = \mathcal{C}^1$. By the continuity of probability measure, this yields that $\lim_{n\to\infty} \nu_\infty(\mathcal{C}_n^1 \backslash \mathcal{C}^1) = 0$ and $\lim_{n\to\infty} \nu_\infty(\mathcal{C}_n^1) = \nu(\mathcal{C}^1)$. On the other hand, for any $\epsilon > 0$, there exists N such that $\sum_{i=N}^\infty (\gamma_i \mu_i^{(2)})^2 \leq \epsilon$ by the assumption ($\sum_{i=0}^\infty \gamma_i^2 < \infty$ and $\sum_{i=0}^\infty (\mu_i^{(2)})^2 < \infty$). Hence, it holds that $\mathbb{E}[(\sum_{i=N}^\infty \gamma_i \xi_i \mu_i^{(2)})^2] = \sum_{i=N}^\infty (\gamma_i \mu_i^{(2)})^2 \leq \epsilon$, which indicates that, by Markov's inequality,

$$\nu_\infty\left(\{\sum_{i=0}^\infty \alpha_i f_i \mid |\sum_{i=N}^\infty \alpha_i \mu_i^{(2)}| > \delta\}\right) \leq \epsilon/\delta^2$$

for any $\delta > 0$. If we set $\mathcal{C}_{(\epsilon)}^2 = \{\sum_{i=0}^\infty \alpha_i f_i \in \mathcal{H} \mid |\sum_{i=0}^\infty \alpha_i \mu_i^{(2)}| \leq 1 + \epsilon\}$, then this and the continuity of Gaussian measures (note that $\sum_{i=0}^\infty \xi_i \gamma_i \mu_i^{(2)}$ is a one dimensional Gaussian measure and has density with respect to the Lebesgue measure) yield that, for any $\epsilon > 0$, there exists N such that for all $n \geq N$, it holds that

$$\begin{aligned}
&\nu_\infty(\mathcal{C}_{(-\epsilon)}^2) - \epsilon \leq \nu_\infty(\mathcal{C}_n^2) \leq \nu_\infty(\mathcal{C}_{(\epsilon)}^2) + \epsilon, \\
&\nu_\infty(\mathcal{C}^1 \cap \mathcal{C}_{(-\epsilon)}^2) - \epsilon \leq \nu_\infty(\mathcal{C}^1 \cap \mathcal{C}_n^2) \leq \nu_\infty(\mathcal{C}^1 \cap \mathcal{C}_{(\epsilon)}^2) + \epsilon.
\end{aligned}$$

Since $\lim_{n\to\infty} \nu_\infty(\mathcal{C}_n^1 \backslash \mathcal{C}^1) = 0$, the second inequality also gives

$$\nu_\infty(\mathcal{C}_n^1 \cap \mathcal{C}_{(-\epsilon)}^2) - 2\epsilon \leq \nu_\infty(\mathcal{C}_n^1 \cap \mathcal{C}_n^2) \leq \nu_\infty(\mathcal{C}_n^1 \cap \mathcal{C}_{(\epsilon)}^2) + 2\epsilon.$$

for any $n \geq N'$ with sufficiently large N'. Therefore, since $\lim_{\epsilon\to 0} \nu_\infty((\mathcal{C}_{(\epsilon)}^2 \backslash \mathcal{C}^2) \cup (\mathcal{C}^2 \backslash \mathcal{C}_{(\epsilon)}^2)) = 0$ by the continuity of Gaussian measures and $\lim_{n\to\infty} \nu_\infty(\mathcal{C}_n^1 \backslash \mathcal{C}^1)$, by taking the limit of ϵ and n of this inequality, we have

$$\nu_\infty(\mathcal{C}^1 \cap \mathcal{C}^2) = \lim_{n\to\infty} \nu_\infty(\mathcal{C}_n^1 \cap \mathcal{C}_n^2).$$

Hence, applying the Gaussian correlation inequality to the right hand side yields

$$\nu_\infty(\mathcal{C}^1 \cap \mathcal{C}^2) = \lim_{n\to\infty} \nu_\infty(\mathcal{C}_n^1 \cap \mathcal{C}_n^2)$$

$$\geq \lim_{n\to\infty} \nu_\infty(\mathcal{C}_n^2)\nu_\infty(\mathcal{C}_n^2) = \nu_\infty(\mathcal{C}^2)\nu_\infty(\mathcal{C}^2).$$

■

Proof [Proof of Proposition 10] The proof relies on comparing the stationary distribution π of Eq. (4) to the Gaussian stationary distribution $\nu_\infty^{(\beta)}$ of Eq. (21) (case $F = 0$). We then conclude by using the small ball probability theorem (Kuelbs and Li, 1993; Li and Shao, 2001) and Lemma 13 on ν_∞. First note that

$$\begin{aligned}
&\int L(x)\,\mathrm{d}\pi(x) - L(\tilde{x}) \\
&= -\tfrac{1}{\beta}\int \log\{e^{-\beta(L(x)-L(\tilde{x}))}\}\,\mathrm{d}\pi(x) \\
&= -\tfrac{1}{\beta}\int \log\{\tfrac{1}{\Lambda}e^{-\beta(L(x)-L(\tilde{x}))}\}\,\mathrm{d}\pi(x) - \tfrac{1}{\beta}\log\Lambda \\
&= -\tfrac{1}{\beta}\mathrm{KL}(\pi||\nu_\infty^{(\beta)}) - \tfrac{1}{\beta}\log(\Lambda), \qquad (39)
\end{aligned}$$

where $\nu_\infty^{(\beta)}$ is the invariant distribution of Eq. (21), i.e., the centered Gaussian on $\mathcal{H}$ with covariance operator $(-\beta A)^{-1}$, $\Lambda \triangleq \int \exp[-\beta(L(x) - L(\tilde{x}))]\,\mathrm{d}\nu_\infty^{(\beta)}(x)$, and $\mathrm{KL}(\mu||\nu) \triangleq \int \log(\mathrm{d}\mu/\,\mathrm{d}\nu)\,\mathrm{d}\mu$ for probability measures μ and ν that are mutually absolutely continuous. Since the KL-divergence $\mathrm{KL}(\pi||\nu_\infty^{(\beta)})$ is non-negative, the right hand side is upper bounded by $-\frac{1}{\beta}\log(\Lambda)$. By definition of $\tilde{x}$, it holds that

$$\nabla L(\tilde{x}) = -\frac{\lambda}{2}\nabla\|\tilde{x}\|_{\mathcal{H}_K}^2 = -\lambda\sum_{k\geq 0}\tfrac{\langle \tilde{x}, f_k\rangle}{\mu_k} f_k.$$

Hence, using the M-smoothness of L, we obtain

$$\begin{aligned}
&L(x) - L(\tilde{x}) \\
&\leq \tfrac{1}{2}M\|x-\tilde{x}\|^2 + \lambda\langle\tilde{x}, x-\tilde{x}\rangle_{\mathcal{H}_K} \\
&\leq \tfrac{1}{2}M\|x-\tilde{x}\|^2 + \lambda\|\tilde{x}\|_{\mathcal{H}_K}\left\langle \frac{\tilde{x}}{\|\tilde{x}\|_{\mathcal{H}_K}}, x-\tilde{x}\right\rangle_{\mathcal{H}_K}.
\end{aligned}$$

Therefore, $\log(\Lambda)$ can be lower-bounded by

$$\begin{aligned}
\log(\Lambda) &\geq \log\int \exp\Big\{-\beta\Big[\tfrac{1}{2}M\|x-\tilde{x}\|^2 \\
&\qquad + \lambda\|\tilde{x}\|_{\mathcal{H}_K}\left\langle\frac{\tilde{x}}{\|\tilde{x}\|_{\mathcal{H}_K}}, x-\tilde{x}\right\rangle_{\mathcal{H}_K}\Big]\Big\}\,\mathrm{d}\nu_\infty^{(\beta)}(x) \\
&\geq -\beta\Big[\tfrac{1}{2}M\varepsilon^2 + \lambda\|\tilde{x}\|_{\mathcal{H}_K}U\Big] \\
&\qquad + \log[\nu_\infty^{(\beta)}(\{x\in\tilde{x}+\mathcal{C}_{\varepsilon,U}\})],
\end{aligned}$$

where $\mathcal{C}_{\varepsilon,U} \triangleq \{x\in\mathcal{H} \mid \|x\| \leq \varepsilon, |\langle\frac{\tilde{x}}{\|\tilde{x}\|_{\mathcal{H}_K}}, x\rangle_{\mathcal{H}_K}| \leq U\}$ for arbitrary $\varepsilon > 0$ and $U > 0$ (if $\|\tilde{x}\|_{\mathcal{H}_K} = 0$, then we treat $\frac{\tilde{x}}{\|\tilde{x}\|_{\mathcal{H}_K}} = 0$). Then, by Borell's inequality (Borell (1975), van der Vaart and van Zanten (2008, Lemma 5.2)), we have

$$\log[\nu_\infty^{(\beta)}(\{x\in\tilde{x}+\mathcal{C}_{\varepsilon,U}\})] \geq \log(\nu_\infty^{(\beta)}(\mathcal{C}_{\varepsilon,U})) - \frac{\beta\lambda}{2}\|\tilde{x}\|_{\mathcal{H}_K}^2.$$

Finally, we lower bound $\log(\nu_\infty^{(\beta)}(\mathcal{C}_{\varepsilon,U}))$. Let $\mathcal{C}_\varepsilon^{(1)} \triangleq \{x \in \mathcal{H} \mid \|x\| \leq \varepsilon\}$ and $\mathcal{C}_U^{(2)} \triangleq \{x \in \mathcal{H} \mid |\langle \frac{\tilde{x}}{\|\tilde{x}\|_{\mathcal{H}_K}}, x\rangle_{\mathcal{H}_K}| \leq U\}$ (that is, $\mathcal{C}_{\epsilon,U} = \mathcal{C}_\varepsilon^{(1)} \cap \mathcal{C}_U^{(2)}$), then by Lemma 13, it holds that

$$\log(\nu_\infty^{(\beta)}(\mathcal{C}_{\varepsilon,U})) \geq \log(\nu_\infty^{(\beta)}(\mathcal{C}_\varepsilon^{(1)})) + \log(\nu_\infty^{(\beta)}(\mathcal{C}_U^{(2)})).$$

By the small ball probability theorem (Kuelbs and Li, 1993; Li and Shao, 2001), we can lower bound the first term of the left hand side as

$$-\log(\nu_\infty^{(\beta)}(\mathcal{C}_\varepsilon^{(1)})) \lesssim (\sqrt{\beta\lambda}\varepsilon)^{-2}.$$

To evaluate $\nu_\infty^{(\beta)}(\mathcal{C}_U^{(2)})$, we note that

$$\mathbb{E}_{x\sim\nu_\infty^{(\beta)}}\left[\left\langle \frac{\tilde{x}}{\|\tilde{x}\|_{\mathcal{H}_K}}, x\right\rangle_{\mathcal{H}_K}^2\right] \leq \beta^{-1}.$$

Therefore, by the Markov's inequality,

$$\nu_\infty^{(\beta)}(\mathcal{C}_U^{(2)}) \geq 1 - \frac{1}{\beta U^2}.$$

By setting $U = \sqrt{2/\beta}$, we also have

$$-\log(\nu_\infty^{(\beta)}(\mathcal{C}_U^{(2)})) \leq \log(1/2).$$

Combining these inequalities, we finally arrive at

$$\begin{aligned}
\int L\,\mathrm{d}\pi - L(\tilde{x}) &\leq -\frac{1}{\beta}\log(\Lambda) \\
&\leq \tfrac{1}{2}M\varepsilon^2 + \lambda\|\tilde{x}\|_{\mathcal{H}_K}\sqrt{\frac{2}{\beta}} + \frac{\lambda}{2}\|\tilde{x}\|_{\mathcal{H}_K}^2 \\
&\quad + \beta^{-1}[C(\sqrt{\beta\lambda}\varepsilon)^{-2} + \log(1/2)] \\
&\lesssim \frac{M}{2}\varepsilon^2 + \lambda\left(\|\tilde{x}\|_{\mathcal{H}_K}\beta^{-1/2} + \|\tilde{x}\|_{\mathcal{H}_K}^2\right) \\
&\quad + (\beta\sqrt{\lambda}\varepsilon)^{-2} + \beta^{-1}. \qquad (40)
\end{aligned}$$

Finally, differentiating the above w.r.t. ε, we get that the optimal bound is attained for $\varepsilon = \left(\frac{2}{M\beta^2\lambda}\right)^{1/4}$, and is then equal to

$$\frac{1}{\beta}\left(\sqrt{\frac{2M}{\lambda}} + 1\right) + \lambda\left(\|\tilde{x}\|_{\mathcal{H}_K}\beta^{-1/2} + \|\tilde{x}\|_{\mathcal{H}_K}^2\right).$$

■

Appendix G. Proof of time and space approximation error (Proposition 9 and Proposition 11)

In this section, we prove Proposition 9 and Proposition 11. As we have noted, Proposition 9 is obtained as a corollary of Proposition 11 by taking the limit of $N \to \infty$. More strongly, we can show the following lemma. Let

$$\hat{c}_\beta = \begin{cases} 1 & \text{(strict dissipativity condition: Assumption 5 (i))}, \\ \sqrt{\beta} & \text{(bounded gradient condition: Assumption 5 (ii))}. \end{cases} \tag{41}$$

Lemma 14 *Suppose $\|x_0\| \leq 1$. Under the same assumptions and notations as in Propositions 9 and 11, it holds that:*

$$\begin{aligned} &\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\left[\phi\left(X_k^N\right) - \phi\left(X^\pi\right)\right] \\ &\leq \frac{C_1}{\Lambda_0^*}\hat{c}_\beta \left(1 + (n\eta')^{-1+\kappa} + (n\eta')^{-1}\right)\left(\eta^{1/2-\kappa} + \mu_{N+1}^{1/2-\kappa} + (n\eta')^{-1}\right). \end{aligned} \tag{42}$$

Proof [Proof of Proposition 9 and Proposition 11] Once we obtain this lemma (Lemma 14), then it is easy to show both propositions by taking into account that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\left[\phi\left(X_k^N\right)\right] = \mathbb{E}[\phi(X^{\mu_{(N,\eta)}})] \tag{43}$$

where $\mu_{(N,\eta)}$ is the invariant measure of $(X_k^N)_k$ whose existence and uniqueness can be shown in the same manner as Proposition 8 (see also Bréhier and Kopec (2016) for this argument). This gives Proposition 11 under the condition $\|x_0\| \leq 1$. Since the invariant measure $\mu_{(N,\eta)}$ is independent of the initial solution x_0, we may drop the condition $\|x_0\| \leq 1$. Then, we obtain Proposition 11.

We can see that the proof of Proposition 8 is valid to show the convergence of Eq. (43) uniformly over all N and its convergence is uniform over all N. Moreover, it has been already shown (see Bréhier (2014) for example) that

$$\lim_{N\to\infty} \mathbb{E}[\phi(X_k^N)] = \mathbb{E}[\phi(X_k)] \quad (\forall k \in \mathbb{N}).$$

Consequently, we can exchange the order of limit, and by applying the geometric ergodicity $\lim_{k\to\infty} \mathbb{E}[\phi(X_k)] = \mathbb{E}[\phi(X^{\mu_\eta})]$ (Proposition 8) again, we also have

$$\begin{aligned} \lim_{N\to\infty} \mathbb{E}[\phi(X^{\mu_{(N,\eta)}})] &= \lim_{N\to\infty}\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\left[\phi\left(X_k^N\right)\right] \\ &= \lim_{n\to\infty}\lim_{N\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\left[\phi\left(X_k^N\right)\right] \quad (\because \text{uniformity of convergence}) \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\left[\phi\left(X_k\right)\right] = \mathbb{E}[\phi(X^{\mu_\eta})]. \end{aligned}$$

Therefore, Lemma 14 gives the proof of Proposition 9 by taking the limit of $n \to \infty$ and $N \to \infty$. Here again, we would like to note that the assumption $\|x\| \leq 1$ can be dropped in the limit because the invariant measure μ_η is independent of the initial solution. ■

In the following, we prove Lemma 14. Our proof follows the line of Bréhier and Kopec (2016). For lighter notation, our constants may differ from line to line.

G.1. Preliminaries

In this subsection, we prepare some notations and state lemmas necessary to prove the statement. Here, we introduce the continuous time dynamics with the Galerkin approximation as

$$\begin{cases} X^N(0) = P_N x_0 \in \mathcal{H}_N, \\ \mathrm{d}X^N(t) = (AX^N(t) - \nabla L_N(X^N(t)))\mathrm{d}t + \sqrt{\frac{2}{\beta}} P_N \mathrm{d}W(t). \end{cases} \tag{44}$$

Here, we denote by $X(t,x)$ to represent $X(t)$ with $X_0 = x$ and similarly we write $Y(t,x)$ for a continuous time process $\{Y(t)\}_t$ to indicate $Y(t)$ with $Y(0) = x$. Notice that our constants should not depend on β while Bréhier and Kopec (2016) sets $\beta = 1$. Our technical contribution is to extend the work of Bréhier and Kopec (2016) to general case. To this end, we apply a change of variables with $t' \triangleq 2t/\beta$. Accordingly, Eq. (4) transforms into

$$\begin{cases} X(0) & = x_0 \in \mathcal{H}, \\ \mathrm{d}X(t') & = -\nabla \mathcal{L}'(X(t'))\mathrm{d}t' + \mathrm{d}W(t') \\ & = (A'X(t') - \nabla L'(X(t')))\mathrm{d}t' + \mathrm{d}W(t'), \end{cases} \tag{45}$$

where $A' \triangleq (\beta/2)A$, $L' \triangleq (\beta/2)L$, $\mathcal{L}' \triangleq (\beta/2)\mathcal{L}$. Accordingly, let the "time re-scaled version" of the process $X(t)$ be

$$\hat{X}(t') \triangleq X\left(\tfrac{\beta}{2}t'\right), \quad \hat{X}^N(t') \triangleq X^N\left(\tfrac{\beta}{2}t'\right) \qquad (t' \geq 0).$$

Similarly, the change of variables with $\eta' = \eta/(\beta/2)$ translates the time discretized Galerkin approximation scheme (5) into

$$\begin{cases} X_0^N = P_N x_0 \in \mathcal{H}_N, \\ X_{n+1}^N = X_n^N - \eta'(A'X_{n+1}^N - \nabla L'_N(X_n^N)) + \sqrt{\eta'}\epsilon_n \\ \quad \Leftrightarrow X_{n+1}^N = \tilde{S}_{\eta'}(X_n^N - \eta' \nabla L'_N(X_n^N) + \sqrt{\eta'}\epsilon_n), \end{cases} \tag{46}$$

where $L'_N \triangleq (\beta/2)L_N$, $\tilde{S}_{\eta'} = S_\eta$. Here we used the abuse of notation to let $A', \tilde{S}_{\eta'}$ indicate the map from $\mathcal{H}_N$ to $\mathcal{H}_N$ which is naturally defined by $A', \tilde{S}_{\eta'} : \mathcal{H} \to \mathcal{H}$ through the canonical imbedding $\iota : x \in \mathcal{H}_N \hookrightarrow \mathcal{H}$: $Ax \triangleq (A \circ \iota)x$ for $x \in \mathcal{H}_N$ (the same argument is also applied to $\tilde{S}_{\eta'}$). Note that we may set $\tilde{S}_{\eta'} = S_\eta$ because $\eta' A' = \eta A$. We write the (rescaled) continuous time corresponding to the k-th step as

$$t_k \triangleq k\eta'.$$

Our approach is to follow the proofs of Bréhier and Kopec (2016); Kopec (2014) and uncover the dependency of β step by step. For completeness, we restate the results of Bréhier and Kopec (2016) in our notations.

Proposition 15

1. *We have for any* $N \in \mathbb{N}$, $\gamma \in [-1/2, 1/2]$, *and* $x \in \mathcal{H}$,

$$\left\|(-A')^{\gamma} P_N x\right\| \leq \left\|(-A')^{\gamma} x\right\|. \tag{47}$$

2. *For* P_N, *we have the following error estimate:*

$$\left\|(-A')^{s/2}(I - P_N)(-A')^{-r/2}\right\|_{\mathcal{B}(\mathcal{H})} = \left(\frac{\mu'_{N+1}}{\beta/2}\right)^{(r-s)/2} \quad \forall 0 \leq s \leq 1,\ s \leq r \leq 2, \tag{48}$$

where

$$\mu'_i \triangleq \mu_i/\lambda.$$

The corresponding result in Bréhier and Kopec (2016) is on finite element approximations; see Andersson and Larsson (2016) for more details. However, it can be naturally extended to spectral Galerkin projection as pointed out in Bréhier and Kopec (2016). In fact, these two approximations are essentially the same; see Kruse (2013). As a consequence, we get the following result.

Proposition 16 *For any* $\kappa > 0$, *the linear operator on* $\mathcal{H}$, $P_N(-A')^{-1/2-\kappa}P_N$ *is continuous, self-adjoint and positive semi-definite. Moreover, there exists* $C_\kappa > 0$ *such that for any* $\beta > 0$

$$\sup_{N\in\mathbb{N}} \mathrm{Tr}\left(P_N(-A')^{-1/2-\kappa}P_N\right) < \frac{C_\kappa}{\beta^{1/2+\kappa}}. \tag{49}$$

This is an extension of Proposition 3.4 in Kopec (2014), where β is fixed to 1. The following fundamental inequality is important for the proof of Proposition 16.

Proposition 17 *For* $M, N \in \mathcal{B}(\mathcal{H})$ *such that* M *is symmetric and positive semi-definite,*

$$|\mathrm{Tr}(MN)| \leq \|M\|_{\mathcal{B}(\mathcal{H})} |\mathrm{Tr}(N)|. \tag{50}$$

Our next step is to extend Lemma 3.7 of Kopec (2014) to our case.

Lemma 18 *For any* $0 \leq \kappa \leq 1$, $N \in \mathbb{N}$, $\beta \geq \eta_0$, *and* $j \geq 1$,

$$\left\|(-A')^{1-\kappa}\tilde{S}^j_{\eta'}P_N\right\|_{\mathcal{B}(\mathcal{H})} \leq \frac{(\beta/2)^{1-\kappa}}{(j\eta)^{1-\kappa}}\frac{1}{(1+\eta/\mu'_0)^{j\kappa}} = \frac{1}{t_j^{1-\kappa}}\frac{1}{(1+\eta/\mu'_0)^{j\kappa}}. \tag{51}$$

Moreover,

$$\forall\, \gamma \geq 1 \quad \exists\, C_\gamma > 0 \quad \forall\, j \geq \gamma \quad \left\|(-A')^{\gamma}\tilde{S}^j_{\eta'}P_N\right\|_{\mathcal{B}(\mathcal{H})} \leq \frac{C_\gamma}{(j\eta)^\gamma}(\beta/2)^\gamma = \frac{C_\gamma}{t_j^\gamma}, \tag{52}$$

and for any $0 \leq \gamma \leq 1$,

$$\left\|(-A')^{-\gamma}(\tilde{S}_{\eta'} - I)P_N\right\|_{\mathcal{B}(\mathcal{H})} \leq 2\frac{\eta^\gamma}{(\beta/2)^\gamma} = 2\eta'^\gamma. \tag{53}$$

The proof is almost the same as in Lemma 3.7 of Kopec (2014), and thus we omit the proof. The relationship that $\eta' = \eta/(\beta/2),\ A' = (\beta/2)A$ specifies the dependence on β.

As in Bréhier and Kopec (2016), we have the following expression of X_k^N:

$$X_k^N = \tilde{S}_{\eta'}^k P_N x - \eta' \sum_{l=0}^{k-1} \tilde{S}_{\eta'}^{k-l} P_N \nabla L(X_l^N) + \sqrt{\eta'} \sum_{l=0}^{k-l} \tilde{S}_{\eta'}^{k-l} P_N \epsilon_{l+1}, \tag{54}$$

$$\sqrt{\eta'} \sum_{l=0}^{k-l} \tilde{S}_{\eta'}^{k-l} P_N \epsilon_{l+1} = \int_0^{t_k} \tilde{S}_{\eta'}^{k-l_s} P_N \mathrm{d}W(s), \tag{55}$$

where $l_s \triangleq \lfloor \frac{s}{\eta'} \rfloor$ with the notation $\lfloor \cdot \rfloor$ is the floor function. The advantage of this expression is that we can handle each term by simple estimates.

We introduce the following interpolation processes: for $0 \leq k \leq m-1$ and $t_k \leq t \leq t_{k+1}$, it holds that:

$$\tilde{X}^N(t) = X_k^N + \int_{t_k}^t \tilde{S}_{\eta'} \left[A' X_k^N - P_N \nabla L'(X_k^N) \right] \mathrm{d}s + \int_{t_k}^t \tilde{S}_{\eta'} P_N \mathrm{d}W(s). \tag{56}$$

The process $\{\tilde{X}^N(t)\}_{t \geq 0}$ is a natural interpolation of the discrete scheme $\{X_k^N\}_{k \in \mathbb{N}}$: $\{\tilde{X}^N(t_k)\}_{k \in \mathbb{N}}$ and $\{X_k^N\}_{k \in \mathbb{N}}$ have the same joint distribution.

G.2. Bounds on Moments

In this subsection, we give a few bounds on moments of $\{X(t)\}_{t \geq 0}$, $\{X^N(t)\}_{t \geq 0}$, $\{X_k^N\}_{k \in \mathbb{N}}$. Note that the constants are uniform with respect to $N \in \mathbb{N},\ 0 < \eta \leq \eta_0$ and $\beta \geq \eta_0$.

Lemma 19 *For any $p \geq 1$, there exists a constant $C_p > 0$ such that for every $N \in \mathbb{N}, t \geq 0, \beta \geq \eta_0$ and $x \in \mathcal{H}$,*

$$\mathbb{E}\left[\|X(t,x)\|^p\right], , \mathbb{E}\left[\left\|\hat{X}(t,x)\right\|^p\right], \mathbb{E}\left[\left\|X^N(t,x)\right\|^p\right], \mathbb{E}\left[\left\|\hat{X}^N(t,x)\right\|^p\right] \leq C_p(1 + \|x_0\|^p). \tag{57}$$

Lemma 20 *For any $p \geq 1, \eta_0 > 0$, there exists a constant C_p such that for every $N \in \mathbb{N},\ 0 < \eta \leq \eta_0,\ \beta \geq \eta_0,\ k \in \mathbb{N},\ t \geq 0$ and $x \in \mathcal{H}$,*

$$\mathbb{E}\left[\left\|X_k^N\right\|^p\right], \mathbb{E}\left[\left\|\tilde{X^N}(t)\right\|^p\right] \leq C_p(1 + \|x_0\|^p). \tag{58}$$

Intuitively, these lemmas hold thanks to dissipativity, a kind of boundedness of a global optimum.

Proof [Proof of Lemma 19 and Lemma 20] The proof is very similar to that of Proposition 7. We only prove the statement for the bounded gradient condition. For the strict dissipativity condition, see Proposition 3.2 of Bréhier and Vilmart (2016). We prove the statement following the same line as Lemma 4.1 and 4.2 of Bréhier (2014). There is no essentially new ingredient, but we need to take care of the effect of β. We define $Z(t) = \hat{X}(t) - W^{A'}(t)$ where $W^{A'}(t) = \int_0^t e^{(t-s)A'} \mathrm{d}W(s)$. It holds that $W^{A'}(t/(\beta/2)) = W^A(t)/\sqrt{\beta/2}$. (2.6) in Kopec (2014) implies:

$$\mathbb{E} \sup_{t' \geq 0} \left\| W^{A'}(t') \right\|^p = \mathbb{E} \sup_{t \geq 0} \left\| W^{A'}\left(\frac{t}{\beta/2}\right) \right\|^p = \mathbb{E} \sup_{t \geq 0} \left\| \frac{W^A(t)}{\sqrt{\beta/2}} \right\|^p < \frac{C_p}{(\beta/2)^{p/2}} < C_p',$$

where $C_p, C_p' > 0$ are constants independent from β.

Then, we study $\|Z(t)\|$. We have $Z(0) = \hat{X}(0) = x_0$,

$$\frac{\mathrm{d}Z(t)}{\mathrm{d}t} = \frac{\beta}{2}(AZ(t) - \nabla L(\hat{X}(t))),$$

and by Proposition 2,

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}\,\|Z(t)\|^2}{\mathrm{d}t} &= \frac{\beta}{2}\langle AZ(t) - \nabla L(\hat{X}(t)), Z(t)\rangle \\
&= \frac{\beta}{2}\langle AZ(t) - \nabla L(Z(t)), Z(t)\rangle + \frac{\beta}{2}\langle \nabla L(Z(t)) - \nabla L(\hat{X}(t)), Z(t)\rangle \\
&\leq \frac{\beta}{2}(-m\,\|Z(t)\|^2 + c + \|\nabla L\|_\infty \|Z(t)\|) \\
&\leq \frac{\beta}{2}(-m'\,\|Z(t)\|^2 + C'),
\end{aligned}$$

where m' and C' are positive constants depending only on m, c, B. Thus, we have for any $t \geq 0$

$$\begin{aligned}
&|\,\|Z(t)\|^2 - C'/m'| \leq \exp(-\beta m' t)|\,\|x_0\|^2 - C'/m'| \\
\implies\quad &\|Z(t)\|^2 \leq \exp(-\beta m' t)|\,\|x_0\|^2 - C'/m'| + C'/m' \leq C(\|x_0\|^2 + 1),
\end{aligned}$$

for a constant $C > 0$, which concludes the proof of Lemma 19, since the estimates do not depend on the dimension parameter N.

Similarly, we introduce $Z_k = X_k - w_k$, where $\{w_k\}_k$ is the numerical approximation of $W^{A'}$ defined by

$$w_{k+1} = \tilde{S}_{\eta'} w_k + \sqrt{\eta'}\tilde{S}_{\eta'}\xi_{k+1}.$$

The same argument yields

$$\mathbb{E}\,\|w_k\|^2 \leq \frac{C}{\beta} \leq C'. \tag{59}$$

Now we have $Z_0 = X_0 = x_0$,

$$Z_{k+1} = \tilde{S}_{\eta'} Z_k - \eta'\tilde{S}_{\eta'}\nabla L'(X_k),$$

since $\left\|\tilde{S}_{\eta'}\right\|_{\mathcal{B}(\mathcal{H})} \leq \frac{1}{1+\eta/\mu_0'}$, we obtain the almost sure estimates

$$\|Z_{k+1}\| \leq \frac{1}{1+\eta/\mu_0'}\,\|Z_k\| + C\eta',$$

and therefore for $\beta \geq 1$

$$\|Z_k\| \leq C(1 + \|x_0\|),$$

which concludes the proof of Lemma 20. ■

G.3. The Rate of Convergence to the Invariant Measure

Our focus in this subsection is just to state the convergence result to an invariant measure. For the existence and uniqueness of the invariant measure of the continuous time dynamics, see Debussche et al. (2011), Goldys and Maslowski (2006) and Bréhier and Kopec (2016). We have the following result thanks to a coupling argument presented in Debussche et al. (2011).

Proposition 21 *Under Assumptions 1, 2, 4 and 5. There exist the "spectral gap" λ^* and a constant $C > 0$ such that for any bounded test function $\phi : \mathcal{H} \to \mathbb{R}$, $t \geq 0$, $N \in \mathbb{N}$, $\beta \geq \eta_0$ and $x_1, x_2 \in \mathcal{H}_N$,*

$$\left|\mathbb{E}[\phi(X^N(t, x_1))] - \mathbb{E}[\phi(X^N(t, x_2))]\right| \leq C \left\|\phi\right\|_\infty (1 + \|x_1\|^2 + \|x_2\|^2) e^{-\lambda^* t}. \tag{60}$$

This also implies

$$\left|\mathbb{E}[\phi(\hat{X}^N(t, x_1))] - \mathbb{E}[\phi(\hat{X}^N(t, x_2))]\right| \leq C \left\|\phi\right\|_\infty (1 + \|x_1\|^2 + \|x_2\|^2) e^{-\beta\lambda^* t}. \tag{61}$$

A proof of this result in case $\beta = 1$ can be found in Debussche et al. (2011). We can easily see the statement holds if β is arbitrary but we have to notice the convergence rate λ^* can be varied depending on β. More concrete characterization of λ^* will be given in Remark 23. As pointed out in Raginsky et al. (2017), this spectral gap is supposed to decrease exponentially with respect to β.

Corollary 22 *For any $N \in \mathbb{N}$, the process X^N admits a unique invariant probability measure π^N and satisfies the following bound:*

$$\begin{aligned} &\exists\, c, C, \lambda^* > 0,\ \forall\, \phi : \mathcal{H} \to \mathbb{R},\ t \geq 0,\ x \in \mathcal{H}_N, \\ &\left|\mathbb{E}[\phi(X^N(t, x))] - \int_{\mathcal{H}_N} \phi \mathrm{d}\pi^N\right| \leq C \left\|\phi\right\|_\infty (1 + \|x\|^2) e^{-\lambda^* t}. \end{aligned} \tag{62}$$

These results naturally extend to an infinite-dimensional scheme by similar arguments.

Remark 23 (Characterization of λ^*) *Bréhier (2014, Theorem 1.1) showed that*

$$\lim_{\eta \to 0} \left|\mathbb{E}[\phi(X_{\lfloor t/\eta \rfloor})] - \mathbb{E}[\phi(X(t))]\right| = 0.$$

In addition to that we have shown in Proposition 8 that the discrete time dynamics satisfies the geometric ergodicity:

$$\left|\mathbb{E}[\phi(X_n)] - \mathbb{E}[\phi(X^{\mu_\eta})]\right| \leq C(1 + \|x\|) \exp(-\Lambda_\eta^*(n\eta - 1)) (\leq C'(1 + \|x\|^2) \exp(-\Lambda_\eta^*(n\eta))),$$

where we used a fact that we may set $\Lambda_\eta^* \leq 1$ (if this is not satisfied, we may set $\Lambda_\eta^* \leftarrow \min\{\Lambda_\eta^*, 1\}$). Moreover, Bréhier (2014, Corollary 1.2) gives that

$$\lim_{\eta \to 0} \left|\mathbb{E}[\phi(X^{\mu_\eta}) - \mathbb{E}[\phi(X^{\pi})]\right| = 0.$$

Combining these arguments, we see that

$$\left|\mathbb{E}[\phi(X(t))] - \mathbb{E}[\phi(X^{\pi})]\right| = \lim_{\eta \to 0} \left|\mathbb{E}[\phi(X_{\lfloor t/\eta \rfloor})] - \mathbb{E}[\phi(X^{\mu_\eta})]\right| \leq \lim_{\eta \to 0} C(1 + \|x\|^2) \exp(-\Lambda_\eta^*(n\eta)).$$

Finally, we note that Proposition 21 and Corollary 22 are used only for $\phi : \mathcal{H} \to \mathbb{R}$ satisfying $\|\phi\|_\infty \geq c$ for a positive constant $c > 0$. Hence, we may set $\lambda^* = \lim_{\eta \to 0} \Lambda_\eta^* = \Lambda_0^*$.

The same argument is also applied to $\{X_k^N\}_k$, $\{X^N(t)\}_t$ and $\{\hat{X}^N(t)\}_t$ with the same value of Λ_η^*. In the following, we use the notation λ^* to indicate Λ_0^*.

Lemma 24 *For any bounded test function $\phi : \mathcal{H} \to \mathbb{R}$, we have*

$$\lim_{N\to\infty} \bar{\phi}_N := \lim_{N\to\infty} \int_{\mathcal{H}_N} \phi \mathrm{d}\pi^N = \int_{\mathcal{H}} \phi \mathrm{d}\pi =: \bar{\phi}. \tag{63}$$

Proof For any $t \geq 0$ and any fixed initial condition $x \in \mathcal{H}$, we have

$$\begin{aligned}\bar{\phi}_N - \bar{\phi} =& \bar{\phi}_N - \mathbb{E}\phi(X^N(t)) \\ &+ \mathbb{E}\phi(X^N(t)) - \mathbb{E}\phi(X(t)) \\ &+ \mathbb{E}\phi(X(t)) - \bar{\phi}.\end{aligned}$$

Since $\lim_{N\to\infty} \mathbb{E}\phi(X^N(t)) - \mathbb{E}\phi(X(t)) = 0$ Bréhier (2014), we get that for any $t \geq 0$

$$\limsup_{N\to\infty} |\bar{\phi}_N - \bar{\phi}| \leq Ce^{-\lambda^* t},$$

and then we may take $t \to \infty$. Notice that the constants are independent of the dimensionality N. ■

G.4. Proof of Lemma 14

In this subsection, we present a technical proof procedure of Lemma 14. As in Bréhier and Kopec (2016), we will use the following decomposition:

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\phi(X_k^N) - \bar{\phi} =& \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\phi(P_{N'}X_k^N) - \bar{\phi}_{N'} \\ &+ \bar{\phi}_{N'} - \bar{\phi} + \frac{1}{n}\sum_{k=0}^{n-1} \left(\mathbb{E}\phi(X_k^N) - \mathbb{E}\phi(P_{N'}X_k^N)\right).\end{aligned}$$

Our aim is to derive a N'-free bound of each term of this decomposition and to take $N' \to \infty$. It is obvious the last two terms converges to 0 as $N' \to +\infty$ thanks to Lemma 24 and $P_{N'}X_k^N = X_k^N$ if $N' \geq N$.

It remains to bound the first term. We decompose the term by the solution of the Poisson equation defined in the following. Let $N' \in \mathbb{N}$, $\phi \in C_b^2(\mathcal{H})$. We define $\Psi^{N'}$ as the unique solution of the Poisson equation

$$\mathcal{L}^{N'}\Psi^{N'} = \phi \circ P_{N'} - \bar{\phi}_{N'} \text{ and } \int_{\mathcal{H}_{N'}} \Psi^{N'} \mathrm{d}\pi^{N'} = 0, \tag{64}$$

where $\mathcal{L}^{N'}$ is the infinitesimal generator of the SPDE[6]:

$$\begin{cases} \hat{X}^{N'}(0) = P_{N'}x_0 \in \mathcal{H}_{N'}, \\ \mathrm{d}\hat{X}^{N'}(t) = (A'\hat{X}^{N'}(t) - \nabla L'_{N'}(\hat{X}^{N'}(t)))\mathrm{d}t + P_{N'}\mathrm{d}W(t), \end{cases}$$

6. Note that from here we also use the notation t to indicate t' for notational simplicity.

defined for C^2 functions $\psi : \mathcal{H} \to \mathbb{R}$ and $x \in \mathcal{H}$ by

$$\mathcal{L}^{N'}\psi(x) = \langle A' P_{N'} x - P_{N'} \nabla L'(x), D\psi(x) \rangle + \frac{1}{2} \mathrm{Tr}(P_{N'} D^2 \psi(x)).$$

The following proposition is essential for our result. This is an extension of Proposition 6.1 in Bréhier and Kopec (2016) in that dependence on β is specified.

Proposition 25 *Let $N' \in \mathbb{N}$ and $\phi \in C_b^2(\mathcal{H})$. The function $\Psi^{N'}$ defined for any $x \in \mathcal{H}_{N'}$ by*

$$\Psi^{N'}(x) = \int_0^\infty \mathbb{E}\left[\phi(\hat{X}^{N'}(t,x)) - \bar{\phi}_{N'}\right] \mathrm{d}t,$$

is of class C_b^2 and the unique solution of Eq. (64). Moreover, we have the following estimates: for any $0 \le \epsilon, \gamma < 1/2$ there exist $C, C_\epsilon, C_{\epsilon,\gamma}$, which are independent of N' and β, such that for any $x \in \mathcal{H}_{N'}$

$$\begin{aligned}
\left\| \Psi^{N'}(x) \right\| &\le \frac{C}{\lambda^* \beta} (1 + \|x\|^2) \|\phi\|_\infty , \\
\left\| (-A')^\epsilon D\Psi^{N'}(x) \right\| &\le \frac{C_\epsilon}{\lambda^* \beta} \hat{c}_\beta \beta^\epsilon (1 + \|x\|^2) \|\phi\|_{0,1} , \\
\left\| (-A')^\epsilon D^2\Psi^{N'}(x)(-A')^\gamma \right\|_{\mathcal{B}(\mathcal{H}_M)} &\le \frac{C_{\epsilon,\gamma}}{\lambda^* \beta} \hat{c}_\beta^2 \beta^{\epsilon+\gamma} (1 + \|x\|^2) \|\phi\|_{0,2} ,
\end{aligned}$$

where $\|\phi\|_{0,i} \triangleq \max\left\{\max_{0<j\le i} \|\phi\|_{(j)}, \|\phi\|_\infty\right\}$ for $\|\phi\|_{(1)} := \sup_{x\in\mathcal{H}} \|\nabla\phi(x)\|$ and $\|\phi\|_{(2)} := \sup_{x\in\mathcal{H}} \|D^2\phi(x)\|_{\mathcal{B}(\mathcal{H})}$.

We give the proof of this proposition in Section G.6.

To show the proof, we prepare more theoretical tools. We define the function $\tilde{\Psi}^{N'}$ for $x \in \mathcal{H}$ by

$$\tilde{\Psi}^{N'}(x) = \Psi^{N'}(P_{N'} x).$$

This can be interpreted as an extension of $\Psi^{N'}$ to the entire domain $\mathcal{H}$. Then we have for any $x \in \mathcal{H}$ and $h, k \in \mathcal{H}$,

$$\begin{aligned}
\langle D\tilde{\Psi}^{N'}(x), h \rangle &= \langle D\Psi^{N'}(P_{N'}x), hP_{N'} \rangle, \\
D^2\tilde{\Psi}^{N'}(x) \cdot (h,k) &= D^2\Psi^{N'}(P_{N'}x) \cdot (P_{N'}h, P_{N'}k).
\end{aligned}$$

Proposition 25 can be also applied to $\tilde{\Psi}^{N'}$ by these equations.

Then we define the generator $\mathcal{L}^{\eta',k,N}$, discrete time version of $\mathcal{L}^{N'}$, for all $k \in \mathbb{N}$ as

$$\begin{aligned}
&\text{for } x_0 \in \mathcal{H}_N,\ \phi \in \mathcal{B}(\mathcal{H}), \\
&\mathcal{L}^{\eta',k,N}\phi(x) = \langle \tilde{S}_{\eta'}(A' X_k^N - P_N \nabla L'(X_k^N)), D\phi(x) \rangle + \frac{1}{2} \mathrm{Tr}(\tilde{S}_{\eta'} S_{\eta'}^* P_N D^2 \phi(x)).
\end{aligned}$$

Thanks to the Itô formula, we have

$$\mathbb{E}\tilde{\Psi}^{N'}(X_{k+1}^N) - \mathbb{E}\tilde{\Psi}^{N'}(X_k^N) = \int_{t_k}^{t_{k+1}} \mathbb{E}\mathcal{L}^{\eta',k,N}\tilde{\Psi}^{N'}(\tilde{X}^N(s)) \mathrm{d}s,$$

(remember the definition (56) of the interpolation process $\tilde{X}^N(s)$). Similarly, we define the generator $\mathcal{L}^N$ of X^N by

$$\mathcal{L}^N\phi(x) = \langle A'x - P_N\nabla L'(x), D\phi(x)\rangle + \frac{1}{2}\mathrm{Tr}(P_N D^2\phi(x)).$$

Putting all of the operators defined above, we have the following decomposition:

$$\begin{aligned}\mathbb{E}\tilde{\Psi}^{N'}(X^N_{k+1}) - \mathbb{E}\tilde{\Psi}^{N'}(X^N_k) = &\int_{t_k}^{t_{k+1}} \mathbb{E}\left(\mathcal{L}^{\eta',k,N} - \mathcal{L}^N\right)\tilde{\Psi}^{N'}(\tilde{X}^N(s))\mathrm{d}s \\ &+ \int_{t_k}^{t_{k+1}} \mathbb{E}\left(\mathcal{L}^N - \mathcal{L}^{N'}\right)\tilde{\Psi}^{N'}(\tilde{X}^N(s))\mathrm{d}s \\ &+ \int_{t_k}^{t_{k+1}} \mathbb{E}\mathcal{L}^{N'}\tilde{\Psi}^{N'}(\tilde{X}^N(s))\mathrm{d}s.\end{aligned} \tag{65}$$

Furthermore, the following equality for $x \in \mathcal{H}$

$$\mathcal{L}^{N'}\tilde{\Psi}^{N'}(x) = \mathcal{L}^{N'}\Psi^{N'}(x) + \langle -P_{N'}\nabla L'(x) + P_{N'}\nabla L'(P_{N'}x), D\Psi^{N'}(P_{N'}x)\rangle,$$

and the definition of $\Psi^{N'}$ yields

$$\begin{aligned}&\int_{t_k}^{t_{k+1}} \mathbb{E}\mathcal{L}^{N'}\tilde{\Psi}^{N'}(\tilde{X}^N(s))\mathrm{d}s \\ =&\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\phi(P_{N'}\tilde{X}^N(s)) - \bar{\phi}_{N'}\right]\mathrm{d}s \\ &+ \int_{t_k}^{t_{k+1}} \mathbb{E}\langle P_{N'}\left(-\nabla L'(\tilde{X}^N(s)) + \nabla L'(P_{N'}\tilde{X}^N(s))\right), D\Psi^{N'}(P_{N'}\tilde{X}^N(s))\rangle\mathrm{d}s \\ =&\eta'\left(\mathbb{E}\phi(P_{N'}X^N_k) - \bar{\phi}_{N'}\right) \\ &+ \int_{t_k}^{t_{k+1}} \mathbb{E}\left[\phi(P_{N'}\tilde{X}^N(s)) - \phi(P_{N'}X^N_k)\right]\mathrm{d}s \\ &+ \int_{t_k}^{t_{k+1}} \mathbb{E}\langle P_{N'}\left(-\nabla L'(\tilde{X}^N(s)) + \nabla L'(P_{N'}\tilde{X}^N(s))\right), D\Psi^{N'}(P_{N'}\tilde{X}^N(s))\rangle\mathrm{d}s.\end{aligned}$$

By substituting this to (65), we obtain

$$\begin{aligned}&\mathbb{E}\tilde{\Psi}^{N'}(X^N_{k+1}) - \mathbb{E}\tilde{\Psi}^{N'}(X^N_k) \\ =&\int_{t_k}^{t_{k+1}} \mathbb{E}\left(\mathcal{L}^{\eta',k,N} - \mathcal{L}^N\right)\tilde{\Psi}^{N'}(\tilde{X}^N(s))\mathrm{d}s \\ &+ \int_{t_k}^{t_{k+1}} \mathbb{E}\left(\mathcal{L}^N - \mathcal{L}^{N'}\right)\tilde{\Psi}^{N'}(\tilde{X}^N(s))\mathrm{d}s \\ &+ \eta'\left(\mathbb{E}\phi(P_{N'}X^N_k) - \bar{\phi}_{N'}\right) \\ &+ \int_{t_k}^{t_{k+1}} \mathbb{E}\left[\phi(P_{N'}\tilde{X}^N(s)) - \phi(P_{N'}X^N_k)\right]\mathrm{d}s \\ &+ \int_{t_k}^{t_{k+1}} \mathbb{E}\langle P_{N'}\left(-\nabla L'(\tilde{X}^N(s)) + \nabla L'(P_{N'}\tilde{X}^N(s))\right), D\Psi^{N'}(P_{N'}\tilde{X}^N(s))\rangle\mathrm{d}s,\end{aligned}$$

and therefore taking the sum of both terms over $k = 0, \dots, n-1$ yields

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\phi(P_{N'}X_k^N) - \bar{\phi}_{N'} \\
=&\frac{1}{n\eta'}\mathbb{E}\left[\Psi^{N'}(P_{N'}X_n^N) - \Psi^{N'}(P_{N'}X_1^N)\right] \\
&+ \frac{1}{n}\left(\phi(P_{N'}x) - \bar{\phi}_{N'}\right) \\
&+ \frac{1}{n\eta'}\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left(\mathcal{L}^{N'} - \mathcal{L}^{N}\right)\tilde{\Psi}^{N'}(\tilde{X}^N(s))\mathrm{d}s \\
&+ \frac{1}{n\eta'}\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left(\mathcal{L}^{N} - \mathcal{L}^{\eta',k,N}\right)\tilde{\Psi}^{N'}(\tilde{X}^N(s))\mathrm{d}s \\
&- \frac{1}{n\eta'}\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\phi(P_{N'}\tilde{X}^N(s)) - \phi(P_{N'}X_k^N)\right]\mathrm{d}s \\
&+ \frac{1}{n\eta'}\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\langle P_{N'}\left(\nabla L'(\tilde{X}^N(s)) - \nabla L'(P_{N'}\tilde{X}^N(s))\right), D\Psi^{N'}(P_{N'}\tilde{X}^N(s))\right\rangle\right]\mathrm{d}s \\
=:& I_1 + I_2 + I_3 + I_4 + I_5 + I_6.
\end{aligned}
$$

As in Bréhier and Kopec (2016), the fact that $\nabla L'$ is Lipschitz, Proposition 25 and Lemma 20 yield

$$\lim_{N'\to\infty} I_6 = 0,$$

and Proposition 25 and Lemma 20 yield for $0 < \eta \leq \eta_0$ and $\beta \geq \eta_0$,

$$|I_1 + I_2| \leq \frac{C}{\lambda^*\beta n\eta'}(1 + \|x_0\|^2).$$

The remaining three terms are controlled by the following lemmas, whose proofs we omit for the sake of conciseness. However, they can be shown by carefully tracing the proof line of Bréhier and Kopec (2016); Kopec (2014) (more specifically, Lemmas 6.3, 6.4 and 6.5 of Bréhier and Kopec (2016) respectively) with the estimates in Proposition 25, Lemma 18 and Section G.5.

Lemma 26 (The control of I_3; space discretization) *For any $0 < \kappa < 1/2$ and η_0, there exists a constant $C > 0$ such that for any $\phi \in C_b^2(\mathcal{H})$, $x \in \mathcal{H}$, $\beta \geq \eta_0$ and $0 < \eta \leq \eta_0$*

$$
\begin{aligned}
&\limsup_{N'\to\infty} \frac{1}{n\eta'}\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left(\mathcal{L}^{N'} - \mathcal{L}^{N}\right)\Psi^{N'}(\tilde{X}^N(s))\mathrm{d}s \\
&\leq \frac{C}{\lambda^*}(1 + \|x_0\|^3)\,\|\phi\|_{0,2}\,\hat{c}_\beta \mu_{N+1}^{1/2-\kappa}(1 + (n\eta')^{-1}).
\end{aligned} \tag{66}
$$

Lemma 27 (The control of I_4; time discretization) *For any $0 < \kappa < 1/2$ and η_0, there exists a constant $C > 0$ such that for any $\phi \in C_b^2(\mathcal{H})$, $N' \in \mathbb{N}$, $x \in \mathcal{H}$, $\beta \geq \eta_0$ and $0 < \eta \leq \eta_0$*

$$\left| \frac{1}{n\eta'} \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\left(\mathcal{L}^N - \mathcal{L}^{\eta',k,N}\right) \tilde{\Psi}^{N'}(\tilde{X}^N(s)) \mathrm{d}s \right|$$
$$\leq \frac{C}{\lambda^*} \|\phi\|_{0,2} (1 + \|x_0\|^3) \hat{c}_\beta \eta^{1/2-\kappa} (1 + (n\eta')^{-1+\kappa} + (n\eta')^{-1}).$$

Lemma 28 (The control of I_5; more time discretization) *For any $0 < \kappa < 1/4$ and η_0, there exists a constant $C, c' > 0$ such that for any $\phi \in C_b^2(\mathcal{H})$, $N' \in \mathbb{N}$, $x \in \mathcal{H}$, $\beta \geq \eta_0$ and $0 < \eta \leq \eta_0$*

$$\left| \frac{1}{n\eta'} \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\left[\phi(P_{N'}\tilde{X}^N(t)) - \phi(P_{N'}X_k^N)\right] \mathrm{d}t \right|$$
$$\leq C \|\phi\|_{0,2} \hat{c}_\beta \eta^{1/2-2\kappa} \left(1 + \frac{\|x_0\|}{(n\eta')^{1-\kappa}}\right).$$

Putting them together, we get the main result (Lemma 14).

G.5. A Malliavin Integration by Parts Formula

In the proofs of Lemmas 26 to 28, an integration by parts formula issued from Malliavin calculus is necessary to transform irregular stochastic integral terms into controllable ones; see Nualart (2006); Sanz-Solé (2005). Therefore, we restate the statement in this subsection. The notations are the same as in Bréhier and Kopec (2016); Debussche (2011).

Lemma 29 *Let $N' \in \mathbb{N}$. For any $G \in \mathbb{D}^{1,2}(\mathcal{H}_{N'})$, $u \in C_b^2(\mathcal{H}_{N'})$ and $\Psi \in L^2(\Omega \times [0,T], \mathcal{L}_2(\mathcal{H}_{N'}))$, an adapted process,*

$$\mathbb{E}\left[Du(G).\int_0^T \Psi(s)\mathrm{d}W^{N'}(s)\right] = \mathbb{E}\left[\int_0^T \mathrm{Tr}(\Psi(s)^* D^2 u(G) \mathcal{D}_s G)\mathrm{d}s\right],$$

where $\mathcal{D}_s G : x \in \mathcal{H} \mapsto \mathcal{D}_s^x G \in \mathcal{H}_{N'}$ stands for th Malliavin derivative of G, and $\mathbb{D}^{1,2}(\mathcal{H}_{N'})$ is the set of $\mathcal{H}_{N'}$-valued random variables $G = \sum_{i \leq N'} G_i f_i$, with $G_i \in \mathbb{D}^{1,2}$ the domain of the Malliavin derivative for $\mathbb{R}$-valued random variables for any i.

In the proof of Lemmas 26 to 28, we use the following estimates; see Bréhier and Kopec (2016); Bréhier (2014); Kopec (2014) for details.

Lemma 30 *For any $0 \leq \gamma < 1$ and $\eta_0 > 0$, there exists a constant $C > 0$ such that for every $h \in (0,1)$, $k \geq 1$, $0 < \eta \leq \eta_0$, $\beta > \eta_0$ and $s \in [t_k - 1/\beta, t_k]$*

$$\left\|(-A')^\gamma \mathcal{D}_s^x X_k^{N'}\right\|_{\mathcal{H}_N} \leq C(1 + M\eta)^{k-l_s} \left(\beta^\gamma + \frac{1}{(1+\eta/\mu_0')^{(1-\gamma)(k-l_s)} t_{k-l_s}^\gamma}\right) \|x\|_{\mathcal{H}_{N'}}, \quad (67)$$

for all $x \in \mathcal{H}_{N'}$. Moreover, if $t_k \leq t < t_{k+1}$, we have

$$\left\|(-A')^\gamma \mathcal{D}_s^x \tilde{X}^{N'}(t)\right\|_{\mathcal{H}_{N'}} \leq C \left\|(-A')^\gamma \mathcal{D}_s^x X_k^{N'}\right\|_{\mathcal{H}_{N'}}, \quad (68)$$

for $x \in \mathcal{H}_{N'}$.

Note that the constant $C > 0$ is uniform with respect to $N' \in \mathbb{N},\ \beta > \eta_0$.

Proof The proof is almost the same as that of Lemma 6.5 in Kopec (2014).

The second inequality is a consequence of the following equality for $s \leq t_k \leq t < t_{k+1}$, thanks to (56):

$$\mathcal{D}_s^x \tilde{X}^{N'}(t) = \mathcal{D}_s^x X_k^{N'} + (t - t_k)(A' \tilde{S}_{\eta'} \mathcal{D}_s^x X_k^{N'} - \tilde{S}_{\eta'} D(P_{N'} \nabla L')(X_k^{N'}) \cdot \mathcal{D}_s^x X_k^{N'}),$$

and the conclusion follows since

$$\sup_{N' \in \mathbb{N}} \left\| \eta' A' \tilde{S}_{\eta'} \right\|_{\mathcal{B}(\mathcal{H}_{N'})} \leq C,$$

where C is a constant that does not depend on β and the norm $\|\cdot\|_{\mathcal{B}(\mathcal{H}_{N'})}$ is taken as a linear map from $\mathcal{H}_{N'}$ to $\mathcal{H}_{N'}$.

Then we prove the first estimate. For any $k \geq 1,\ x \in \mathcal{H}_N$, and $s \in [t_k - 1/\beta, t_k]$, we have

$$\mathcal{D}_s^x X_k^{N'} = \tilde{S}_{\eta'}^{k-l_s} x - \eta' \sum_{i=l_s+1}^{k-1} \tilde{S}_{\eta'}^{k-i} D(P_{N'} \nabla L')(X_i^{N'}) . \mathcal{D}_s^x X_i^{N'}.$$

We recall that $l_s = \lfloor s/\eta' \rfloor$, so that when $i \leq l_s$ we have $\mathcal{D}_s^x X_i^{N'} = 0$.

As a consequence, the discrete Gronwall's inequality ensures that for $k \geq l_s + 1$ and a constant $C > 0$,

$$\left\| \mathcal{D}_s^x X_k^{N'} \right\|_{\mathcal{H}_{N'}} \leq (1 + M\eta)^{k-l_s} \left\| x \right\|_{\mathcal{H}_{N'}},$$

where we used $\eta' L' = \eta L$ and the Lipchitz continuity of ∇L. Now using Lemma 18, we have

$$\begin{aligned}
&\left\| (-A')^\gamma \mathcal{D}_s^x X_k^{N'} \right\|_{\mathcal{H}_{N'}} \\
&\leq \frac{1}{(1 + \eta/\mu_0')^{(1-\gamma)(k-l_s)} t_{k-l_s}^\gamma} \left\| x \right\|_{\mathcal{H}_{N'}} + M\eta \sum_{i=l_s+1}^{k-1} \frac{(1 + M\eta)^{i-l_s}}{(1 + \eta/\mu_0')^{(1-\gamma)(k-i)} t_{k-i}^\gamma} \left\| x \right\|_{\mathcal{H}_{N'}}.
\end{aligned}$$

Note that $k - l_s \leq 1/(\eta'\beta) \leq 1/\eta$ yields $(1 + M\eta)^{k-l_s} \leq C$. To conclude, we see that when $0 < \eta \leq \eta_0$, it holds that for a constant c_0 (could be dependent on η_0, μ_0'),

$$\begin{aligned}
&\eta \sum_{i=l_s+1}^{k-1} \frac{1}{(1 + \eta/\mu_0')^{(1-\gamma)(k-i)} t_{k-i}^\gamma} \leq \beta C \int_0^\infty \frac{t^{-\gamma}}{(1 + \eta/\mu_0')^{(1-\gamma)t/\eta'}} \mathrm{d}t \\
&\leq \beta C \int_0^\infty t^{-\gamma} \exp\left[-c_0 (1-\gamma)(t/\eta')(\eta/\mu_0') \right] \mathrm{d}t \\
&\leq \beta C \int_0^\infty t^{-\gamma} \exp\left[-\frac{\beta}{2} c_0 (1-\gamma) t/\mu_0' \right] \mathrm{d}t \\
&\leq C \beta^\gamma.
\end{aligned}$$

■

G.6. Proof of Proposition 25

In this subsection, we prove Proposition 25. Our argument follows the same line as Bréhier and Kopec (2016). Let $\phi \in C_b^2(\mathcal{H})$. For lighter notation, we assume $\bar{\phi} = 0$ in this section. We define the function u for any $t > 0$ and $x \in \mathcal{H}_{N'}$ by

$$u(t,x) = \mathbb{E}\left[\phi(\hat{X}^{N'}(t,x))\right], \tag{69}$$

which is the solution of a finite-dimensional Kolmogorov equation associated with (44) where $N = N'$:

$$\frac{\mathrm{d}u}{\mathrm{d}t}(t,x) = Lu(t,x) = \frac{1}{2}\mathrm{Tr}(D^2u(t,x)) + \langle A'x - \nabla L'_{N'}(x), Du(t,x)\rangle.$$

To prove Proposition 25, we only need to show that $u \in C^2$ and that u and its two first derivatives have estimates which are integrable with respect to t. Specifically we prove the following proposition.

Proposition 31 *Let $\phi \in C_b^2$ such that $\bar{\phi} = 0$ and u defined by (69). Remember that $\hat{c}_\beta$ is defined in Eq. (41) as*

$$\hat{c}_\beta = \begin{cases} 1 & (\textit{strict dissipativity condition: Assumption 5 (i)}), \\ \sqrt{\beta} & (\textit{bounded gradient condition: Assumption 5 (ii)}). \end{cases}$$

There exist constant c, $C > 0$ such that for any $0 \leq \epsilon, \gamma < 1/2$ there exist constants C_ϵ and $C_{\epsilon,\gamma}$, which is independent of β, such that for any $t > 0$ and $x \in \mathcal{H}_{N'}$,

$$\|u(t,x)\| \leq Ce^{-\beta\lambda^* t}(1 + \|x\|^2)\,\|\phi\|_\infty\,, \tag{70}$$

$$\left\|(-A')^\epsilon Du(t,x)\right\| \leq C_\epsilon \hat{c}_\beta \beta^\epsilon (1 + \frac{1}{(\beta t)^\epsilon})e^{-\beta\lambda^* t}(1 + \|x\|^2)\,\|\phi\|_{0,1}\,, \tag{71}$$

$$\left\|(-A')^\epsilon D^2u(t,x)(-A')^\gamma\right\|_{\mathcal{B}(\mathcal{H})} \leq C_{\epsilon,\gamma}\hat{c}_\beta^2\beta^{\epsilon+\gamma}\left(1 + \frac{1}{(\beta t)^{\alpha'}} + \frac{1}{(\beta t)^{\epsilon+\gamma}}\right)e^{-\beta\lambda^* t}(1 + \|x\|^2)\,\|\phi\|_{0,2}\,, \tag{72}$$

where $\lambda^* > 0$ is the spectral gap introduced in Remark 23 (see also Proposition 21) and $\alpha' \in [0,1]$ is the constant introduced in Assumption 4.

In fact the estimation (71) is true for $\alpha < 1$. The proof is a slight modification of the proof of Proposition 8.1 in Kopec (2014). Since $\phi \in C^2$, bounded and with bounded derivatives, $u \in C^2$ and the derivatives can be calculated in the following way:

- For any $h \in \mathcal{H}_{N'}$, we have

$$Du(t,x).h = \mathbb{E}\left[D\phi(\hat{X}^{N'}(t,x)).\eta^{h,x}(t)\right], \tag{73}$$

where $\eta^{h,x}(t)$ is the solution of

$$\frac{\mathrm{d}\eta^{h,x}(t)}{\mathrm{d}t} = A'\eta^{h,x}(t) - D^2L'_{N'}(\hat{X}^{N'}(t,x)).\eta^{h,x}(t),$$

$$\eta^{h,x}(0) = h.$$

- For any $h, k \in \mathcal{H}_{N'}$, we have

$$D^2u(t,x).(h,k) = \mathbb{E}\left[D^2\phi(\hat{X}^{N'}(t,x)).(\eta^{h,x}(t),\eta^{k,x}(t)) + D\phi(\hat{X}^{N'}(t,x)).\zeta^{h,k,x}(t)\right], \tag{74}$$

where $\zeta^{h,k,x}$ is the solution of

$$\frac{\mathrm{d}\zeta^{h,k,x}}{\mathrm{d}t} = A'\zeta^{h,k,x}(t) - D^2L'(\hat{X}^{N'}(t,x)).\zeta^{h,k,x}(t) - D^3L'(\hat{X}^{N'}(t,x)).(\eta^{h,x}(t),\eta^{k,x}(t)),$$

$$\zeta^{h,k,x}(0) = 0.$$

Moreover, we already have the inequality (70) thanks to Corollary 22.

The proof requires several steps. First in Lemma 32 below we prove estimates for $0 < t \leq 1/\beta$ and general $0 \leq \alpha, \gamma < 1/2$; then in Lemma 33 we study the long-time behavior in case $\alpha = \gamma = 0$; we finally conclude with the proofs of Proposition 31.

Lemma 32 *Assume Assumption 5 (ii) (bounded gradient condition). For any $0 \leq \epsilon, \gamma < 1/2$, there exist constants $C_\epsilon, C_{\epsilon,\gamma}$ such that for any $x \in \mathcal{H}_{N'}$, and any $0 < t \leq 1/\beta$,*

$$\left\|(-A')^\epsilon Du(t,x)\right\| \leq \frac{C_\epsilon}{t^\epsilon}\left\|D\phi\right\|_\infty,$$
$$\left\|(-A')^\epsilon D^2u(t,x)(-A')^\gamma\right\|_{\mathcal{B}(\mathcal{H}_{N'})} \leq C_{\epsilon,\gamma}\beta^{\epsilon+\gamma}\left(\frac{1}{(\beta t)^{\alpha'}} + \frac{1}{(\beta t)^{\epsilon+\gamma}}\right)\left(\left\|D\phi\right\|_\infty + \left\|D^2\phi\right\|_\infty\right),$$

where α' is defined in Assumption 4.

Proof Owing to (73) and (74), we only need to prove the following almost sure estimates for some constants - which may vary from line to line below: for any $0 < t \leq 1/\beta$

$$\left\|\eta^{h,x}(t)\right\| \leq \frac{C_\epsilon}{(\beta t)^\epsilon}\left\|h\right\|_\epsilon,$$
$$\left\|\zeta^{h,k,x}(t)\right\| \leq C_{\epsilon,\gamma}\beta^{\epsilon+\gamma}\left(\frac{1}{(\beta t)^{\alpha'}} + \frac{1}{(\beta t)^{\epsilon+\gamma}}\right)\left\|h\right\|_\epsilon\left\|k\right\|_\gamma.$$

To show these inequalities, first note that

$$\begin{aligned}\left\|e^{tA'}h\right\| &= \left\|t^{-\epsilon}(-tA')^\epsilon e^{tA'}(-A')^{-\epsilon}h\right\| = t^{-\epsilon}\left\|(-tA')^\epsilon e^{tA'}\right\|_{\mathcal{B}(\mathcal{H})}\left\|(-A')^{-\epsilon}h\right\| \\ &\leq t^{-\epsilon}\sup_{x\geq 0}\{x^\epsilon e^{-x}\}\left\|(-A')^{-\epsilon}h\right\| = \frac{C_\epsilon}{t^\epsilon}\left\|(-A')^{-\epsilon}h\right\|\end{aligned} \tag{75}$$

where $C_\epsilon \triangleq \sup_{x\geq 0}\{x^\epsilon e^{-x}\}$. From this, we deduce that

$$\begin{aligned}&\left\|\eta^{h,x}(t)\right\| \\ &= \left\|e^{tA'}h - \int_0^t e^{(t-s)A'}D^2L'(\hat{X}(s,x)).\eta^{h,x}(s)\mathrm{d}s\right\|\end{aligned}$$

$$\leq \frac{C_\epsilon}{t^\epsilon} \left\|(-A')^{-\epsilon} h\right\| + C \int_0^t \beta \left\|\eta^{h,x}(s)\right\| \mathrm{d}s.$$

and by the Gronwall's inequality and $t \leq 1/\beta$, we get the result.

For the second-order derivative, we moreover use the properties of L to get

$$\begin{aligned}
&\left\|\zeta^{h,k,x}(t)\right\| \\
=&\left| \int_0^t e^{(t-s)A'} D^2 L'(\hat{X}(s,x)).\zeta^{h,k,x}(s) \mathrm{d}s \right. \\
&\left. + \int_0^t e^{(t-s)A'} D^3 L'(\hat{X}(s,x)).(\eta^{h,x}(s), \eta^{k,x}(s)) \mathrm{d}s \right| \\
\leq& C \int_0^t \beta \left\|\zeta^{h,k,x}(s)\right\| \mathrm{d}s + \int_0^t \frac{C_{\alpha'} \beta^{1-\alpha'}}{(t-s)^{\alpha'}} \left\|\eta^{h,x}(s)\right\| \left\|\eta^{k,x}(s)\right\| \mathrm{d}s \\
\leq& C \int_0^t \beta \left\|\zeta^{h,k,x}(s)\right\| \mathrm{d}s \\
&+ C_{\alpha',\epsilon,\gamma} \left\|(-A')^{-\epsilon} h\right\| \left\|(-A')^{-\gamma} k\right\| \beta^{\epsilon+\gamma} (\beta t)^{1-\alpha'-\epsilon-\gamma} \int_0^1 \frac{1}{(1-s)^{\alpha'} s^{\epsilon+\gamma}} \mathrm{d}s.
\end{aligned}$$

The Gronwall's inequality yields the conclusion since for any $0 < \beta t \leq 1$ we have $(\beta t)^{1-\alpha'-\epsilon-\gamma} < (\beta t)^{-\alpha'}$ due to the assumption $\epsilon + \gamma < 1$. ■

Lemma 33 *Assume Assumption 5 (ii) (bounded gradient condition). There exist constants $C, c > 0$ such that for any $t \geq 0$, and any $x \in \mathcal{H}$,*

$$\|Du(t,x)\| \leq C\sqrt{\beta} e^{-\beta\lambda^* t} (1 + \|x\|^2) \|\phi\|_\infty ,$$

and

$$\left\|D^2 u(t,x)\right\|_{\mathcal{B}(\mathcal{H})} \leq C\beta e^{-\beta\lambda^* t} \left(1 + \frac{1}{(\beta t)^{\alpha'}}\right) (1 + \|x\|^2) \|\phi\|_\infty .$$

Proof [Proof of Lemma 33] As in Kopec (2014), we use the Bismut-Elworthy-Li formula (Bismut, 1984; Elworthy and Li, 1994) to get for $\Phi : \mathcal{H}_{N'} \to \mathbb{R}$ which belongs to class C^2 with bounded derivative and with at most quadratic growth, i.e.,

$$\exists\, M(\Phi) > 0,\ \forall\, x \in \mathcal{H}_{N'},\ \|\Phi(x)\| \leq M(\Phi)(1 + \|x\|^2),$$

and $v(t,x) \triangleq \mathbb{E}\Phi(\hat{X}^{N'}(t,x))$, we have two following formula:

$$Dv(t,x).h = \frac{1}{t} \mathbb{E}\left[\int_0^t \langle \eta^{h,x}(s), \mathrm{d}W(s) \rangle \Phi(\hat{X}^{N'}(t,x)) \right].$$

Moreover, by the Markov property $v(t,x) = \mathbb{E}v(t/2, \hat{X}^{N'}(t/2,x))$, we obtain

$$Dv(t,x).h = \frac{2}{t} \mathbb{E}\left[\int_0^{t/2} \langle \eta^{h,x}(s), \mathrm{d}W(s) \rangle v(t/2, \hat{X}^{N'}(t/2,x)) \right].$$

and thus

$$D^2v(t,x).(h,k) = \frac{2}{t}\mathbb{E}\left[\int_0^{t/2}\langle\zeta^{h,k,x}(s),\mathrm{d}W(s)\rangle v(t/2,\hat{X}^{N'}(t/2,x))\right] + \frac{2}{t}\mathbb{E}\left[\int_0^{t/2}\langle\eta^{h,x}(s),\mathrm{d}W(s)\rangle Dv(t/2,\hat{X}^{N'}(t/2,x)).\eta^{k,x}(t/2)\right].$$

We then see, using Lemma 19 and Lemma 32 with $\epsilon=\gamma=0$ that there exists $C>0$ such that for any $0<t\leq 1/\beta,\ x,h,k\in\mathcal{H}_{N'}$,

$$\begin{aligned}\|Dv(t,x).h\| &\leq \frac{C}{\sqrt{t}}M(\Phi)(1+\|x\|^2)\|h\|,\\ \left\|D^2v(t,x).(h,k)\right\| &\leq \frac{C}{t}M(\Phi)(1+\|x\|^2)\|h\|\|k\|.\end{aligned} \tag{76}$$

Indeed, to see the first inequality, the Cauchy-Schwartz inequality gives

$$\begin{aligned}Dv(t,x).h &= \frac{1}{t}\mathbb{E}\left[\int_0^t\langle\eta^{h,x}(s),\mathrm{d}W(s)\rangle\Phi(\hat{X}^{N'}(t,x))\right]\\ &\leq \frac{1}{t}\sqrt{\mathbb{E}\left[\left(\int_0^t\langle\eta^{h,x}(s),\mathrm{d}W(s)\rangle\right)^2\right]}\sqrt{\mathbb{E}[\Phi(\hat{X}^{N'}(t,x))^2]},\end{aligned}$$

and the isometry property of Ito integral and Lemma 32 give a bound of the first term as

$$\sqrt{\mathbb{E}\left[\left(\int_0^t\langle\eta^{h,x}(s),\mathrm{d}W(s)\rangle\right)^2\right]} = \sqrt{\int_0^t\|\eta^{h,x}(s)\|^2\,\mathrm{d}s} \leq C\sqrt{t}\|h\|,$$

for $t\leq 1/\beta$ and Lemma 19 gives a bound of the second term as

$$\sqrt{\mathbb{E}[\Phi(\hat{X}^{N'}(t,x))^2]} \leq CM(\Phi)(1+\|x\|^2).$$

Now when $\beta t\geq 1$ the Markov property implies that $u(t,x)=\mathbb{E}[u(t-1/\beta,\hat{X}^{N'}(1/\beta,x))]$ and by Corollary 22, we have

$$\left\|u(t-1/\beta,x)-\int_{\mathcal{H}_{N'}}\phi\mathrm{d}\bar{\mu}\right\| \leq Ce^{-\beta\lambda^*(t-1/\beta)}(1+\|x\|^2)\|\phi\|_\infty.$$

If we choose $\Phi_t(x)=u(t-1/\beta,x)-\int_{\mathcal{H}}\phi\mathrm{d}\bar{\mu}$, we have $u(t,x)=\mathbb{E}\Phi_t(\hat{X}^{N'}(1/\beta,x))+\int_{\mathcal{H}}\phi\bar{\mu}$, with $M(\Phi_t)\leq Ce^{-\beta\lambda^*(t-1/\beta)}\|\phi\|_\infty$. With (76) at $t=1/\beta$, we obtain for $t\geq 1/\beta$,

$$\begin{aligned}\|Du(t,x).h\| &\leq C\sqrt{\beta}\|\phi\|_\infty e^{-\beta\lambda^*(t-1/\beta)}(1+\|x\|^2)\|h\|,\\ \left\|D^2u(t,x).(h,k)\right\| &\leq C\beta\|\phi\|_\infty e^{-\beta\lambda^*(t-1/\beta)}(1+\|x\|^2)\|h\|\|k\|.\end{aligned}$$

We have a control when $0\leq t\leq 1/\beta$ in Lemma 32, so with a change of constants we get the result. ■

Next we show a corresponding lemma for the strict dissipativity condition in the following lemma.

Lemma 34 *Assume Assumption 5 (i) (strict dissipativity condition). For any $0 \le \epsilon, \gamma < 1/2$, there exist constants $C_\epsilon, C_{\epsilon,\gamma}$ such that for any $x \in \mathcal{H}_{N'}$, and any $0 < t$,*

$$\left\|(-A')^\epsilon Du(t,x)\right\| \le C_\epsilon \beta^\epsilon \left(1 + \frac{1}{(\beta t)^\epsilon}\right) e^{-t\beta\lambda^*} \left\|D\phi\right\|_\infty ,$$

$$\left\|(-A')^\epsilon D^2 u(t,x)(-A')^\gamma\right\|_{\mathcal{B}(\mathcal{H}_{N'})} \le C_{\epsilon,\gamma}\beta^{\epsilon+\gamma}\left(1 + \frac{1}{(\beta t)^{\alpha'}} + \frac{1}{(\beta t)^{\epsilon+\gamma}}\right) e^{-t\beta\lambda^*}\left(\left\|D\phi\right\|_\infty + \left\|D^2\phi\right\|_\infty\right),$$

where α' is defined in Assumption 4.

Proof From the definition of $\eta^{h,x}$, we have that

$$\begin{aligned}\left\|\eta^{h,x}(t)\right\| &= \left\|e^{tA'}h - \int_0^t e^{(t-s)A'} D^2 L'(\hat{X}^{N'}(s,x)).\eta^{h,x}(s)\mathrm{d}s\right\| \\ &\le \left\|e^{tA'}h\right\| + \int_0^t e^{-(t-s)\lambda/\mu_0} M\beta \left\|\eta^{h,x}(s)\right\| \mathrm{d}s.\end{aligned}$$

As in Eq. (75), for any $0 \le c_0 < 1$, the first term can be bounded by

$$\begin{aligned}\left\|e^{tA'}h\right\| &= \left\|t^{-\epsilon}(-tA')^\epsilon e^{c_0 tA'}(-A')^{-\epsilon} e^{(1-c_0)tA'}h\right\| \\ &= t^{-\epsilon}\left\|(-tA')^{c_0\epsilon}e^{tA'}\right\|_{\mathcal{B}(\mathcal{H})}\left\|(-A')^{-\epsilon}e^{(1-c_0)tA'}h\right\| \\ &\le t^{-\epsilon}\sup_{x\ge 0}\{x^\epsilon e^{-c_0 x}\}\left\|(-A')^{-\epsilon}h\right\| = \frac{C_{\epsilon,c_0}}{t^\epsilon}\left\|(-A')^{-\epsilon}e^{(1-c_0)tA'}h\right\|\end{aligned}$$

where $C_{\epsilon,c_0} \triangleq \sup_{x\ge 0}\{x^\epsilon e^{-c_0 x}\}$. Then, Gronwall's inequality gives

$$\begin{aligned}& e^{t\beta\lambda/\mu_0}\left\|\eta^{h,x}(t)\right\| \le \frac{C_{\epsilon,c_0}}{t^\epsilon} e^{c_0 t\beta\lambda/\mu_0}\left\|(-A')^{-\epsilon}h\right\| + \int_0^t \beta M e^{s\beta\lambda/\mu_0}\left\|\eta^{h,x}(s)\right\|\mathrm{d}s \\ \Rightarrow\ & e^{t\beta\lambda/\mu_0}\left\|\eta^{h,x}(t)\right\| \le \frac{C_{\epsilon,c_0}}{t^\epsilon} e^{c_0 t\beta\lambda/\mu_0}\left\|(-A')^{-\epsilon}h\right\| + \int_0^t \beta M e^{s\beta\lambda/\mu_0}\left\|\eta^{h,x}(s)\right\|\mathrm{d}s \\ \Rightarrow\ & e^{t\beta\lambda/\mu_0}\left\|\eta^{h,x}(t)\right\| \le \frac{C_{\epsilon,c_0}}{t^\epsilon} e^{c_0 t\beta\lambda/\mu_0}\left\|(-A')^{-\epsilon}h\right\| \\ &\qquad + C_{\epsilon,c_0}\int_0^t \frac{e^{c_0 s\beta\lambda/\mu_0}}{s^\epsilon}\beta M\exp((t-s)\beta M)\,\mathrm{d}s\left\|(-A')^{-\epsilon}h\right\| \\ &\qquad \le C_{\epsilon,c_0}\left\|(-A')^{-\epsilon}h\right\|\left[\frac{1}{t^\epsilon}e^{c_0 t\beta\lambda/\mu_0} + \beta^\epsilon M e^{t\beta M}\int_0^\infty \frac{e^{\tau(c_0\lambda/\mu_0 - M)}}{\tau^\epsilon}\mathrm{d}\tau\right] \\ \Rightarrow\ & \left\|\eta^{h,x}(t)\right\| \le \frac{C_{\epsilon,c_0}}{t^\epsilon}\left(e^{-(1-c_0)t\beta\lambda/\mu_0} + (\beta t M)^\epsilon\int_0^\infty \frac{e^{\tau(c_0\lambda/\mu_0 - M)}}{(M\tau)^\epsilon} M\,\mathrm{d}\tau e^{-t\beta(\lambda/\mu_0 - M)}\right).\end{aligned}$$

Therefore, if we choose c_0 as $c_0 = (\lambda/\mu_0)^{-1}M/2$, then $0 \le c_0 < 1$ by the strict dissipativity assumption and we obtain

$$\left\|\eta^{h,x}(t)\right\| \le C\frac{1 + (t\beta M)^\epsilon}{t^\epsilon}\exp[-t\beta(\lambda/\mu_0 - M)]\left\|(-A')^{-\epsilon}h\right\|$$

$$= C\frac{1+(t\beta M)^{\epsilon}}{t^{\epsilon}}\exp[-t\beta\lambda^*]\left\|(-A')^{-\epsilon}h\right\|, \tag{77}$$

where we used $\lambda^* = \lambda/\mu_0 - M\ (>0)$. Applying this to Eq. (73), we have the first inequality.

The second inequality is also shown in the same way as Lemma 32. Notice that by the Lipschitz continuity of ∇L, we have

$$\begin{aligned}
&\left\|\zeta^{h,k,x}(t)\right\| \\
\leq& \int_0^t e^{-(t-s)\beta\lambda/\mu_0}\beta M\left\|\zeta^{h,k,x}(s)\right\|\mathrm{d}s + \int_0^t \frac{C_{\alpha',c_0}\beta^{1-\alpha'}}{(t-s)^{\alpha'}}e^{-(1-c_0)(t-s)\beta\lambda/\mu_0}\left\|\eta^{h,x}(s)\right\|\left\|\eta^{k,x}(s)\right\|\mathrm{d}s \\
\leq& \int_0^t e^{-(t-s)\beta\lambda/\mu_0}\beta M\left\|\zeta^{h,k,x}(s)\right\|\mathrm{d}s \\
&+ C'_{\alpha',\epsilon,\gamma}\left\|(-A')^{-\epsilon}h\right\|\left\|(-A')^{-\gamma}k\right\|\beta^{1-\alpha'}\int_0^t \frac{(1+(M\beta s)^{\epsilon+\gamma})}{(t-s)^{\alpha'}s^{\epsilon+\gamma}}e^{-2s\beta(\lambda/\mu_0-M)}e^{-(1-c_0)(t-s)\beta\lambda/\mu_0}\mathrm{d}s.
\end{aligned}$$

From this inequality, we have

$$\begin{aligned}
&e^{t\beta\lambda/\mu_0}\left\|\zeta^{h,k,x}(t)\right\| \\
\leq& \int_0^t \beta M e^{s\beta\lambda/\mu_0}\left\|\zeta^{h,k,x}(s)\right\|\mathrm{d}s \\
&+ C_{\alpha',\epsilon,\gamma}\left\|(-A')^{-\epsilon}h\right\|\left\|(-A')^{-\gamma}k\right\|\beta^{1-\alpha'}e^{t\beta\lambda/\mu_0 - t\beta\min\{2(\lambda/\mu_0-M),(1-c_0)\lambda/\mu_0\}}\int_0^t \frac{(1+(M\beta s)^{\epsilon+\gamma})}{(t-s)^{\alpha'}s^{\epsilon+\gamma}}\mathrm{d}s \\
\leq& \int_0^t \beta M e^{s\beta\lambda/\mu_0}\left\|\zeta^{h,k,x}(s)\right\|\mathrm{d}s \\
&+ C_{\alpha',\epsilon,\gamma}\left\|(-A')^{-\epsilon}h\right\|\left\|(-A')^{-\gamma}k\right\|\beta^{1-\alpha'}e^{t\beta\lambda/\mu_0 - t\beta\min\{2(\lambda/\mu_0-M),(1-c_0)\lambda/\mu_0\}}\times \\
&(1+(M\beta t)^{\epsilon+\gamma})t^{1-\alpha'-\epsilon-\gamma}\int_0^1 \frac{1}{(1-\tilde{s})^{\alpha'}\tilde{s}^{\epsilon+\gamma}}\mathrm{d}\tilde{s} \\
\leq& \int_0^t \beta M e^{s\beta\lambda/\mu_0}\left\|\zeta^{h,k,x}(s)\right\|\mathrm{d}s \\
&+ C'_{\alpha',\epsilon,\gamma}\left\|(-A')^{-\epsilon}h\right\|\left\|(-A')^{-\gamma}k\right\|\beta^{1-\alpha'}(1+(M\beta t)^{\epsilon+\gamma})t^{1-\alpha'-\epsilon-\gamma}\times \\
&\quad e^{t\beta\lambda/\mu_0 - t\beta\min\{2(\lambda/\mu_0-M),(1-c_0)\lambda/\mu_0\}}.
\end{aligned}$$

Here, we set $c_0 = (\lambda/\mu_0)^{-1}M/2$, then we further obtain

$$\begin{aligned}
&e^{t\beta\lambda/\mu_0}\left\|\zeta^{h,k,x}(t)\right\| \\
\leq& C'_{\alpha',\epsilon,\gamma}\left\|(-A')^{-\epsilon}h\right\|\left\|(-A')^{-\gamma}k\right\|\beta^{1-\alpha'}\times \\
&\Big(e^{t\beta M}\int_0^t M\beta(1+(M\beta s)^{\epsilon+\gamma})s^{1-\alpha'-\epsilon-\gamma}e^{-s\beta\min\{\lambda/\mu_0-M,M/2\}}\mathrm{d}s + \\
&\quad (1+(M\beta t)^{\epsilon+\gamma})t^{1-\alpha'-\epsilon-\gamma}e^{t\beta M}e^{-t\beta\min\{\lambda/\mu_0-M,M/2\}}\Big) \\
=& C'_{\alpha',\epsilon,\gamma}\left\|(-A')^{-\epsilon}h\right\|\left\|(-A')^{-\gamma}k\right\|\beta^{1-\alpha'}e^{t\beta M}\times
\end{aligned}$$

$$\begin{aligned}
&\left(\int_0^t M\beta(1+(M\beta s)^{\epsilon+\gamma}) s^{1-\alpha'-\epsilon-\gamma} e^{-s\beta\min\{\lambda/\mu_0 - M, M/2\}} \, \mathrm{d}s \right. \\
&\qquad \left. + (1+(M\beta t)^{\epsilon+\gamma}) t^{1-\alpha'-\epsilon-\gamma} e^{-t\beta\min\{\lambda/\mu_0 - M, M/2\}} \right) \\
\leq & C''_{\alpha',\epsilon,\gamma} \left\|(-A')^{-\epsilon} h\right\| \left\|(-A')^{-\gamma} k\right\| \beta^{1-\alpha'} e^{t\beta M} \times \\
&\qquad [\beta^{-(1-\alpha'-\epsilon-\gamma)} + (1+(M\beta t)^{\epsilon+\gamma}) t^{1-\alpha'-\epsilon-\gamma} e^{-t\beta\min\{\lambda/\mu_0 - M, M/2\}}]
\end{aligned}$$

By multiplying both terms by $e^{-t\beta\lambda/\mu_0}$, we obtain

$$\left\|\zeta^{h,k,x}(t)\right\| \leq C'_{\alpha',\epsilon,\gamma} \left\|(-A')^{-\epsilon} h\right\| \left\|(-A')^{-\gamma} k\right\| \beta^{\epsilon+\gamma} [1 + (\beta t)^{1-\alpha'-\epsilon-\gamma}] e^{-t\beta(\lambda/\mu_0 - M)},$$

where we used that $\sup_{t>0}(1+(M\beta t)^{\epsilon+\gamma}) e^{-t\beta\min\{\lambda/\mu_0 - M, M/2\}} < C$ (bounded by a constant independent of β). Since $1-\epsilon-\gamma > 0$ and $1-\alpha' > 0$, it holds that $(\beta t)^{1-\alpha'-\epsilon-\gamma} \leq (\beta t)^{-\alpha'} + (\beta t)^{-\epsilon-\gamma}$. Then, we finally obtain

$$\left\|\zeta^{h,k,x}(t)\right\| \leq C'_{\alpha',\epsilon,\gamma} \left\|(-A')^{-\epsilon} h\right\| \left\|(-A')^{-\gamma} k\right\| \beta^{\epsilon+\gamma} \left[1 + (\beta t)^{-\alpha'} + +(\beta t)^{-\epsilon-\gamma}\right] e^{-t\beta\lambda^*},$$

where we used $\lambda^* = \lambda/\mu_0 - M \ (> 0)$. Applying this inequality and Eq. (77) to Eq. (74), we obtain the second inequality. ■

Remark 35 *Note that Lemma 34 for the strict dissipativity condition does not require the restriction $t \leq 1/\beta$ while Lemma 32 is for the bounded gradient condition. This is advantageous to show better dependency on β under the strict dissipativity condition than the bounded gradient condition.*

We can finally prove Proposition 31. The proof is again in line with Kopec (2014).

Proof [Proof of Proposition 31.] First, we show the assertion for the bounded gradient condition. By the Markov property and Lemma 33, for any $t \geq 1/\beta$, we have

$$\begin{aligned}
\|Du(t,x).h\| &\leq C\sqrt{\beta} \|\phi\|_\infty e^{-\beta\lambda^*(t-1/\beta)} \mathbb{E}\left[(1 + \left\|\hat{X}^{N'}(1/\beta, x)\right\|^2) \left\|\eta^{h,x}(1/\beta)\right\| \right] \\
&\leq C\sqrt{\beta} \|\phi\|_\infty e^{-\beta\lambda^*(t-1/\beta)} (1 + \|x\|^2) \beta^\epsilon \left\|(-A')^{-\epsilon} h\right\|,
\end{aligned}$$

where the last estimate comes from Lemma 19 and Lemma 32. Combining this estimate and Lemma 32, we obtain Eq. (71). We can easily see Eq. (72) follows from the similar argument.

As for the strict dissipativity condition, Lemma 34 directly gives the assertion. ■

Appendix H. Proof of SGLD convergence rate (Proposition 12)

In this chapter, we prove Proposition 12. Before that, we need to prepare the following lemmas to bound $\mathbb{E}[L(Y_k^N) - L(X_k^N)]$. For lighter notation, our constants may differ from line to line.

Lemma 36 *For any $x \in \mathcal{H}_N$, it holds that*

$$\mathbb{E}\,\|\nabla L(x) - g_k(x)\|^2 \leq \frac{C(n_{\mathrm{tr}} - n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}} - 1)},$$

where n_{b} is the mini-batch size and $C > 0$ is some constant.

We can prove the following bound similarly to Lemma 19 and 20 thanks to Assumption 6.

Lemma 37 *For any $p \geq 1$, there exists a constant C_p such that for every $N \in \mathbb{N}$, $\beta \geq \eta_0$, and $x \in \mathcal{H}_N$,*

$$E\left\|Y_k^N\right\|^p \leq C_p(1 + \|x_0\|^p).$$

Lemma 38 *It holds that:*

$$\exists\, C_1,\, C_2 > 0,\, \forall\, \beta > \frac{2\mu_0'}{2 + \eta/\mu_0'},$$
$$\log \mathbb{E}\left[\exp(\left\|X_k^N\right\|^2)\right] \leq \|x_0\|^2 + C_1/\beta + C_2,$$

where $C_1, C_2 > 0$ is an constant.

Remark 39 *Note that our estimate is not subject to "the curse of dimensionality" which explicitly appears in Lemma C.7 in Xu et al. (2018).*

Proof The proof is similar to that of Lemma C.7 in Xu et al. (2018). The main difference lies in the existence of regularizer in our scheme and the absence of dissipativity assumption of L_N. Instead, we assume Assumption 6.

Let $Q = \frac{2\eta}{\beta}$ and $p_j = \frac{1}{(1+\eta/\mu_j')^2}$. Let $S' := \mathrm{diag}((q_j)_{j=0}^N)$ for $q_j > 0$ $(j = 0, \ldots, N)$ and $1 > q_0 \geq q_1 \geq \cdots \geq q_N$, then we have

$$\begin{aligned}
&\mathbb{E}\left[\exp \|X_{k+1}^N\|_{S'}^2\right] \\
=&\mathbb{E}\left[\exp\left\|S_\eta(X_k^N - \eta\nabla L_N(X_k^N) + \sqrt{\frac{2\eta}{\beta}}\epsilon_k^N\right\|_{S'}^2\right],
\end{aligned}$$

where $\epsilon_k^N \sim \mathcal{N}(0, I_N)$. Let x_i and ϵ_i denote the i-th component of $X_k^N - \eta\nabla L_N(X_k^N)$ and ϵ_k^N respectively, which corresponds to the coefficient of f_i, the i-th basis in the representation (1). Under this notation, we have the following estimate:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\|S_\eta(X_k^N - \eta\nabla L_N(X_k^N) + \sqrt{\frac{2\eta}{\beta}}\epsilon_k^N)\right\|_{S'}^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left\|S_\eta(X_k^N - \eta\nabla L_N(X_k^N) + \sqrt{\frac{2\eta}{\beta}}\epsilon_k^N)\right\|_{S'}^2 \middle| X_k^N\right]\right] \\
&= \mathbb{E}\left[\prod_{i=0}^N \int \exp\left(p_i q_i\left(x_i^2 + 2\sqrt{\frac{2\eta}{\beta}}x_i\epsilon_i + \frac{2\eta}{\beta}\epsilon_i^2\right)\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\epsilon_i^2}{2}\right)\mathrm{d}\epsilon_i\right]
\end{aligned}$$

$$= \mathbb{E}\prod_{i=0}^{N}\frac{1}{\sqrt{1-p_iq_iQ}}\exp\left(\frac{x_i^2}{\frac{1}{p_iq_i}-2Q}\right)$$
$$\leq \exp\left(\sum_{j=0}^{N}\frac{Qp_jq_j}{1-2Qp_0q_0}\right)\mathbb{E}\left[\exp\left(\sum_{i=0}^{N}\frac{x_i^2}{\frac{1}{p_iq_i}-2Q}\right)\right],$$

thanks to the formula of Gaussian integral, $\mu_0' \geq \mu_i'$ and $\log(1-x) \geq -x/(1-x)$.

Then, we have

$$\exp\left(\sum_{j=0}^{N}\frac{Qp_jq_j}{1-2Qp_0q_0}\right)\mathbb{E}\left[\exp\left(\sum_{i=0}^{N}\frac{x_i^2}{\frac{1}{p_iq_i}-2Q}\right)\right]$$
$$\leq \exp\left(\sum_{j=0}^{N}\frac{Qp_jq_j}{1-2Qp_0q_0}\right)\mathbb{E}\left[\exp\left(\sum_{i=0}^{N}\frac{1}{\frac{1}{p_iq_i}-2Q}\left({X_{k,i}^N}^2-2\eta X_{k,i}^N\nabla L_{N,i}(X_k^N)+\eta^2\nabla L_{N,i}(X_k^N)^2\right)\right)\right],$$

where $X_{k,i}^N$, $L_{N,i}(X_k^N)$ denotes the i-th component of X_k^N, $L_N(X_k^N)$ respectively.

Then Assumption 6 implies

$$\exp\left(\sum_{j=0}^{N}\frac{Qp_jq_j}{1-2Qp_0q_0}\right)\mathbb{E}\left[\exp\left(\sum_{i=0}^{N}\frac{1}{\frac{1}{p_iq_i}-2Q}\left({X_{k,i}^N}^2-2\eta X_{k,i}^N\nabla L_{N,i}(X_k^N)+\eta^2\nabla L_{N,i}(X_k^N)^2\right)\right)\right],$$
$$\leq \exp\left(\sum_{j=0}^{N}\frac{Qp_jq_j}{1-2Qp_0q_0}\right)\mathbb{E}\left[\exp\left(\sum_{i=0}^{N}\frac{1}{\frac{1}{p_iq_i}-2Q}\left({X_{k,i}^N}^2+2\eta B|X_{k,i}^N|+B^2\eta^2\right)\right)\right],$$
$$\leq \exp\left(\sum_{j=0}^{N}\frac{Qp_jq_j}{1-2Qp_0q_0}\right)\mathbb{E}\left[\exp\left(\sum_{i=0}^{N}\frac{1}{\frac{1}{p_iq_i}-2Q}\left((1+\kappa){X_{k,i}^N}^2+C\left(1+\frac{1}{\kappa}\right)\eta^2\right)\right)\right]$$
$$\leq \exp\left(\sum_{j=0}^{N}\frac{Qp_jq_j}{1-2Qp_0q_0}\right)\mathbb{E}\left\{\exp\left[\sum_{i=0}^{N}\left(\frac{1+\kappa}{\frac{1}{p_iq_i}-2Q}(X_{k,i}^N)^2+\frac{Cp_iq_i(1+\frac{1}{\kappa})}{1-2Qp_0q_0}\eta^2\right)\right]\right\}$$
$$= \mathbb{E}\left\{\exp\left[\|X_k^N\|_{S^{(k)}}^2+\sum_{j=0}^{N}\left(\frac{Qp_jq_j}{1-2Qp_0q_0}+\frac{Cp_jq_j(1+\frac{1}{\kappa})}{1-2Qp_0q_0}\eta^2\right)\right]\right\},$$

where $S^{(k)} := \operatorname{diag}\left(\left(\frac{1+\kappa}{\frac{1}{p_jq_j}-2Q}\right)_{j=0}^{N}\right)$.

Since $q_0 \leq 1$, it holds that

$$\frac{Qp_jq_j}{1-2Qp_0q_0} \leq \frac{Qp_jq_j}{1-2Qp_0}.$$

If we have chosen κ so that $\frac{1+\kappa}{\frac{1}{p_0q_0}-2Q} < 1$, then

$$q_j^{(k)} := \frac{1+\kappa}{\frac{1}{p_jq_j}-2Q} \leq \frac{1+\kappa}{\frac{1}{p_0q_0}-2Q} < 1,$$

and we also have $q_0^{(k)} \geq q_1^{(k)} \geq \cdots \geq q_N^{(k)}$. Here again, since $q_j \leq 1$, it holds that

$$q_j^{(k)} = \frac{1+\kappa}{\frac{1}{p_j q_j} - 2Q} \leq \frac{1+\kappa}{(p_j^{-1} - 2Q) q_j^{-1}}.$$

Let $\kappa = \frac{1}{2}(2\eta/\mu_0' + (\eta/\mu_0')^2 - 2\eta/\beta)$, then it holds that

$$\begin{aligned}\frac{1+\kappa}{p_j^{-1} - 2Q} &= \frac{1 + \frac{1}{2}[2\eta/\mu_0' + (\eta/\mu_0')^2 - 2\eta/\beta]}{1 + 2\eta/\mu_j' + (\eta/\mu_j')^2 - 2\eta/\beta} \\ &\leq \frac{1}{1 + \frac{1}{4}[2\eta/\mu_j' + (\eta/\mu_j')^2 - 2\eta/\beta]} =: \frac{1}{1+\alpha_j} < 1.\end{aligned}$$

Therefore, we obtain the following evaluation for $q_j^{(k)}$:

$$q_j^{(k)} \leq \frac{q_j}{1+\alpha_j},$$

which implies $\|\cdot\|_{S^{(k)}} \leq \|\cdot\|_{S'}$. Hence, by noticing $p_j \leq (1+\alpha_j)^{-1}$, a recursive argument yields

$$\mathbb{E}\left[\exp\left(\|X_{k+1}^N\|_{S'}^2\right)\right] \leq \exp\left[\|x_0\|^2 + \frac{Q + C\,(1+1/\kappa)\,\eta^2}{1 - 2Qp_0q_0} \sum_{j=0}^{N} \sum_{i=0}^{k} \frac{q_j}{(1+\alpha_j)^{k+1-i}}\right].$$

The second term in the right hand side can be evaluated as

$$\sum_{i=0}^{k} \frac{1}{(1+\alpha_j)^{k+1-i}} = \frac{(1+\alpha_j)^{-1} - (1+\alpha_j)^{-(k+2)}}{1 - (1+\alpha_j)^{-1}} \leq \frac{1}{\alpha_j}.$$

Finally, if we set $q_j = 1$, then by observing that

$$\sum_{j=0}^{N} \frac{1}{\alpha_j} \lesssim \int_1^\infty \frac{1}{\eta x^2}\,\mathrm{d}x \lesssim \frac{1}{\eta},$$

we have

$$\frac{Q + C\,(1+1/\kappa)\,\eta^2}{1 - 2Qp_0q_0} \sum_{j=0}^{N} \sum_{i=0}^{k} \frac{1}{(1+\alpha_j)^{k+1-i}} q_j \lesssim \frac{Q + (1+1/\kappa)\,\eta^2}{\eta} = \frac{1}{\beta} + \left(1 + \frac{1}{\alpha_0}\right)\eta.$$

Since $\alpha_0 = O(\eta)$, the second term in the right hand side can be evaluated

$$\left(1 + \frac{1}{\alpha_0}\right)\eta \lesssim 1.$$

Combining all arguments, we obtain that

$$\mathbb{E}\left[\exp\left(\|X_{k+1}^N\|^2\right)\right] \leq \exp\left(\|x_0\|^2 + \frac{C_1}{\beta} + C_2\right).$$

■

Lemma 40 *The discrepancy between $L(X_n^N)$ and $L(Y_n^N)$ can be bounded as*

$$|\mathbb{E}[L(X_n^N) - L(Y_n^N)]| \leq B\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}}\left\{1+\frac{M\mu_0}{\lambda}\exp\left(\frac{M\mu_0}{\lambda}\right)\right\},$$

where the expectation is with respect to the choice of the mini-batches in each update.

Proof Remember that X_n^N and Y_n^N are updated as

$$X_{n+1}^N = S_\eta\left(X_n^N - \eta\nabla L_N(X_n^N) + \sqrt{2\tfrac{\eta}{\beta}}P_N\varepsilon_n\right),$$
$$Y_{n+1}^N = S_\eta\left(Y_n^N - \eta g_{n,N}(Y_n^N) + \sqrt{2\tfrac{\eta}{\beta}}P_N\varepsilon_n\right),$$

where the same noise ε_n is applied for updating both variables. Hence, by taking the difference between X_n^N and Y_n^N, $X_n^N - Y_n^N$ is updated as

$$X_{n+1}^N - Y_{n+1}^N = S_\eta\left[(X_n^N - Y_n^N) - \eta(\nabla L_N(X_n^N) - g_{n,N}(Y_n^N))\right].$$

Noticing that $X_0^N - Y_0^N = 0$, it holds that

$$\begin{aligned}
X_{n+1}^N - Y_{n+1}^N =& -\eta\sum_{k=0}^n S_\eta^{n-k+1}(\nabla L_N(X_k^N) - g_{n,N}(Y_k^N)) \\
=& -\eta\sum_{k=0}^n S_\eta^{n-k+1}(\nabla L_N(X_k^N) - \nabla L_N(Y_k^N)) \\
& -\eta\sum_{k=0}^n S_\eta^{n-k+1}(\nabla L_N(Y_k^N) - g_{n,N}(Y_k^N)).
\end{aligned}$$

Here, Lemma 36 yields that

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{k=0}^n S_\eta^{n-k+1}(\nabla L_N(Y_k^N) - g_{n,N}(Y_k^N))\right\|^2 \\
&\leq \sum_{k=0}^n \|S_\eta^{n-k+1}\|_{\mathrm{op}}^2 \frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)} \\
&\leq \sum_{k=0}^n \left(1+\eta\frac{\lambda}{\mu_0}\right)^{-2(n-k+1)} \frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)} \\
&\leq \frac{C}{\eta\lambda/\mu_0}\frac{(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)},
\end{aligned}$$

where the expectation is taken with respect to the choice of the mini-batch. By the smoothness assumption (Assumption 2), we also have $\|\nabla L_N(X_k^N) - \nabla L_N(Y_k^N)\| \leq M\|X_k^N - Y_k^N\|$. Thus, we obtain that

$$\mathbb{E}\left\|X_{n+1}^N - Y_{n+1}^N\right\| \leq \eta\sum_{k=0}^n \|S_\eta^{n-k+1}\|_{\mathrm{op}} M\mathbb{E}\left\|X_k^N - Y_k^N\right\| + \eta\sqrt{\frac{C}{\eta\lambda/\mu_0}\frac{(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}}$$

$$\leq \eta M \sum_{k=0}^{n} \left(1+\eta\frac{\lambda}{\mu_0}\right)^{-(n-k+1)} \mathbb{E}\left\|X_k^N - Y_k^N\right\| + \sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}}.$$

Applying Gronwall's inequality (see, for example, Mischler (2019, Lemma 5.2)) to this evaluation yields that

$$\begin{aligned}
&\mathbb{E}\left\|X_{n+1}^N - Y_{n+1}^N\right\| \\
\leq &\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}} \times \\
&\left\{1+\eta M \sum_{k=0}^{n}\left(1+\eta\frac{\lambda}{\mu_0}\right)^{-(n-k+1)} \prod_{i=k+1}^{n}\left[1+\eta M\left(1+\eta\frac{\lambda}{\mu_0}\right)^{-(n-i+1)}\right]\right\} \\
\leq &\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}} \times \\
&\left\{1+\eta M \sum_{k=0}^{n}\left(1+\eta\frac{\lambda}{\mu_0}\right)^{-(n-k+1)} \exp\left[\eta M \sum_{i=k+1}^{n}\left(1+\eta\frac{\lambda}{\mu_0}\right)^{-(n-i+1)}\right]\right\} \\
\leq &\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}} \left\{1+\eta M \sum_{k=0}^{n}\left(1+\eta\frac{\lambda}{\mu_0}\right)^{-(n-k+1)} \exp\left(\eta M\frac{\mu_0}{\eta\lambda}\right)\right\} \\
\leq &\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}} \left\{1+\eta M\frac{\mu_0}{\eta\lambda}\exp\left(\frac{M\mu_0}{\lambda}\right)\right\} \\
= &\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}} \left\{1+\frac{M\mu_0}{\lambda}\exp\left(\frac{M\mu_0}{\lambda}\right)\right\}.
\end{aligned}$$

Then, by the bounded gradient condition Assumption 5 (ii), the discrepancy between $L(X_n^N)$ and $L(Y_n^N)$ can be bounded as

$$\begin{aligned}
|\mathbb{E}[L(X_n^N) - L(Y_n^N)]| &\leq B\mathbb{E}\left\|X_n^N - Y_n^N\right\| \\
&\leq B\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_{\mathrm{tr}}-n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}}-1)}} \left\{1+\frac{M\mu_0}{\lambda}\exp\left(\frac{M\mu_0}{\lambda}\right)\right\},
\end{aligned}$$

where the expectation is with respect to the choice of the mini-batches in each update. ■

In addition to the bound in Lemma 40, we obtain another bound on $|\mathbb{E}[L(X_n^N) - L(Y_n^N)]|$. The following two lemmas are used to prove a version of Theorem 3.6 in Xu et al. (2018). These results can only be applied to finite-dimensional spaces. However, our schemes Y_k^N, X_k^N are no longer infinite-dimensional, which means we can follow the same argument in Xu et al. (2018).

Lemma 41 (Polyanskiy and Wu (2016); Raginsky et al. (2017); Xu et al. (2018)) *For any two probability density functions μ, ν with bounded second moments, let $g : \mathbb{R}^d \to \mathbb{R}$ be a C^1 function such that*

$$\|\nabla g(x)\|_2 \leq C_1 \|x\|_2 + C_2, \ \forall\, x \in \mathbb{R}^d,$$

for some constants $C_1, C_2 \geq 0$. Then

$$\left| \int_{\mathbb{R}^d} g \mathrm{d}\mu - \int_{\mathbb{R}^d} g \mathrm{d}\nu \right| \leq (C_1 \sigma + C_2) \mathcal{W}_2(\mu, \nu),$$

where $\mathcal{W}_2$ is the 2-Wasserstein distance and $\sigma^2 = \max\left\{ \int_{\mathbb{R}^d} \|x\|_2^2 \, \mu(\mathrm{d}x), \int_{\mathbb{R}^d} \|x\|_2^2 \, \nu(\mathrm{d}x) \right\}$.

Lemma 42 *(Corollary 2.3 in Bolley and Villani (2005)) Let ν be a probability measure on $\mathbb{R}^d$. Assume that there exist x_0 and a constant $\alpha > 0$ such that $\int \exp(\alpha \|x - x_0\|_2) \nu(\mathrm{d}x) < \infty$. Then for any probability measure μ on $\mathbb{R}^d$, it satisfies*

$$\mathcal{W}_2(\mu, \nu) \leq C_\nu (\mathrm{D}(\mu||\nu)^{1/2} + \mathrm{D}(\mu||\nu)^{1/4}),$$

where C_ν is defined as

$$C_\nu = \inf_{x_0 \in \mathbb{R}^d, \alpha > 0} \sqrt{\frac{1}{\alpha} \left(\frac{3}{2} + \log \int \exp(\alpha \|x - x_0\|_2^2) \nu(\mathrm{d}x) \right)}.$$

Proof [Proof of Proposition 12] Let P_k, Q_k denote the probability measures for GLD scheme X_k^N and SGLD scheme Y_k^N respectively. Applying Lemma 41, Lemma 37 and Lemma 20 yields

$$|\mathbb{E}\left[L(Y_k^N)\right] - \mathbb{E}\left[L(X_k^N)\right]| \leq C(1 + \|x_0\|) \mathcal{W}_2(Q_k, P_k), \tag{78}$$

where $C > 0$ are absolute constants. We further apply Lemma 42 to bound Wasserstein distance and get the following bound:

$$|\mathbb{E}\left[L(Y_k^N)\right] - \mathbb{E}\left[L(X_k^N)\right]| \leq C(1 + \|x_0\|) \Lambda (\mathrm{D}(Q_k||P_k)^{1/2} + \mathrm{D}(Q_k||P_k)^{1/4}), \tag{79}$$

where $\Lambda = \sqrt{3/2 + \log \mathbb{E}\left[\exp \left\|X_k^N\right\|^2\right]}$. Moreover, Lemma 38 yields

$$\Lambda \leq \sqrt{\frac{3}{2} + \|x_0\|^2 + \frac{C_1}{\beta} + C_2}, \tag{80}$$

where $C_1, C_2 > 0$ is some constants. To bound KL-divergence between P_k and Q_k, we use the following decomposition:

$$\begin{aligned}
\mathrm{D}(Q_k||P_k) &\leq \mathrm{D}(Q_k||P_k) + \mathrm{D}(Q_{1:k-1}|Q_k||P_{1:k-1}|P_k) = \mathrm{D}(Q_{1:k}||P_{1:k}) \\
&= \mathrm{D}(Q_1||P_1) + \sum_{i=2}^{k} \mathrm{D}(Q_i|Q_{1:i-1}||P_i|P_{1:i-1}) \\
&= \sum_{i=1}^{k} \mathrm{D}(Q_i|Q_{i-1}||P_i|P_{i-1}),
\end{aligned}$$

where $P_{1:k}, Q_{1:k}$ denotes joint distribution of $(X_1^N, \cdots, X_k^N)$ and $(Y_1^N, \cdots, Y_k^N)$ respectively and $Q_i|Q_{i-1}$ denotes the conditional distribution of X_i^N given X_{i-1}^N. The first inequality is based on

non-negativity of KL-divergence and the final equality comes from the fact that Q_0, P_0 are deterministic and that X_i^N and $(X_1^N, \cdots, X_{i-2}^N)$ are conditionally independent given X_{i-1}^N. For clarity, we write down the definition of conditional KL-divergence in the following line:

$$\mathrm{D}(F_2|F_1||G_2|G_1) = \int f(x_1, x_2) \log \frac{f(x_2|x_1)}{g(x_2|x_1)} \mathrm{d}x_1 \mathrm{d}x_2.$$

Now that $Q_i|Q_{i-1}$ and $P_i|P_{i-1}$ are both gaussian, that is,

$$X_i^N | X_{i-1}^N = x \sim \mathcal{N}(S_\eta(x - \eta \nabla L_N(x)), \frac{\eta}{\beta} S_\eta^T S_\eta),$$
$$Y_i^N | Y_{i-1}^N = x \sim \mathcal{N}(S_\eta(x - \eta g_{i-1}(x)), \frac{\eta}{\beta} S_\eta^T S_\eta),$$

we can calculate each conditional KL-divergence as below:

$$\begin{aligned}
&\mathrm{D}(Q_i|Q_{i-1}||P_i|P_{i-1}) = \mathbb{E}_Q \left[\log \frac{\mathrm{d}Q_i|Q_{i-1}}{\mathrm{d}P_i|P_{i-1}} \right] \\
&= \frac{\beta}{2\eta} \mathbb{E}_{(x,y)\sim Q_{i-1:i}} \left[\left\| S_\eta^{-1} y - (x - \eta \nabla L_N(x)) \right\|^2 - \left\| S_\eta^{-1} y - (x - \eta g_{i-1}(x)) \right\|^2 \right] \\
&= \frac{\beta}{2\eta} \mathbb{E}_{(x,y)\sim Q_{i-1:i}} \left[2\eta \langle S_\eta^{-1} y - x, \nabla L_N(x) - g_{i-1}(x) \rangle + \eta^2 (\|\nabla L_N(x)\|^2 - \|g_{i-1}(x)\|^2) \right] \\
&= \frac{\beta}{2\eta} \mathbb{E}_{x\sim Q_{i-1}} \left[2\eta \langle -\eta g_{i-1}(x), \nabla L_N(x) - g_{i-1}(x) \rangle + \eta^2 (\|\nabla L_N(x)\|^2 - \|g_{i-1}(x)\|^2) \right] \\
&= \frac{\beta\eta}{2} \mathbb{E}_{x\sim Q_{i-1}} \left[\|\nabla L_N(x) - g_{i-1}(x)\|^2 \right] \\
&\leq \frac{\beta\eta}{2} \mathbb{E}_{x\sim Q_{i-1}} \left[\frac{C(n_{\mathrm{tr}} - n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}} - 1)} \right] \\
&\leq C \frac{\beta\eta(n_{\mathrm{tr}} - n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}} - 1)},
\end{aligned}$$

thanks to Lemma 36 and 37. Therefore, we finally get the following bound:

$$\mathrm{D}(Q_k||P_k) \leq \frac{\beta\eta k(n_{\mathrm{tr}} - n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}} - 1)} = C r_k. \tag{81}$$

Combining all of the above yields the claim:

$$|\mathbb{E}[L(X_k^N) - L(Y_k^N)]| \leq C(1 + \|x_0\|)\Lambda \left(\sqrt{r_k} + \sqrt[4]{r_k} \right). \tag{82}$$

Now, for a non-negative integer K, let $X_{n|K}^N$ be the solution following the update 5 with the full gradient for $n \geq K$, but $X_{K|K}^N = Y_K^N$. That is, $X_{K|K}^N = Y_K^N$, and

$$X_{n+1|K}^N = S_\eta \left(X_{n|K}^N - \eta \nabla L_N(X_{n|K}^N) + \sqrt{2\tfrac{\eta}{\beta}} P_N \varepsilon_n \right),$$

for $n \geq K$. Then, notice that, for $n \geq K$, Proposition 8 yields that

$$|\mathbb{E}[L(Y_n^N) - L(X_n^N)]|$$

$$
\begin{aligned}
&\leq |\mathbb{E}[L(Y_n^N) - L(X_{n|K}^N)]| + |\mathbb{E}[L(X_{n|K}^N) - L(X_n^N)]| \\
&\leq |\mathbb{E}[L(Y_n^N) - L(X_{n|K}^N)]| + |\mathbb{E}[L(X_{n|K}^N) - L(X^{\mu_{(N,\eta)}})]| + |\mathbb{E}[L(X_n^N) - L(X^{\mu_{(N,\eta)}})]| \\
&\leq |\mathbb{E}[L(Y_n^N) - L(X_{n|K}^N)]| + \mathbb{E}\left[C_{Y_K^N}\right] \exp\left\{-\Lambda_\eta^*[\eta(n-K)-1]\right\} + C_{x_0} \exp\left(-\Lambda_\eta^*(\eta n - 1)\right) \\
&\leq |\mathbb{E}[L(Y_n^N) - L(X_{n|K}^N)]| + CC_{x_0} \exp\left\{-\Lambda_\eta^*[\eta(n-K)-1]\right\},
\end{aligned}
$$

where $C > 0$ is a constant and we used $\mathbb{E}[C_{Y_K^N}] \lesssim C_{x_0}$ from Lemma 37 in the last inequality. Applying the bound (82) with $x_0 = Y_K^N$ and $k = n - K$ to $|\mathbb{E}[L(Y_n^N) - L(X_{n|K}^N)]|$, we have

$$
|\mathbb{E}[L(Y_n^N) - L(X_n^N)]| \lesssim \sqrt{r_{n-K}} + \sqrt[4]{r_{n-K}} + \exp\left\{-\Lambda_\eta^*[\eta(n-K)-1]\right\},
$$

where we used Lemma 37 again. Since the choice of $K \in \{0, \dots, n\}$ is arbitrary, we may pick up the one that minimizes the right hand side. Indeed, $T^* = \frac{\log_+\{n_\mathrm{b}(n_\mathrm{tr}-1)/[\beta\eta(n_\mathrm{tr}-n_\mathrm{b})]\}}{\Lambda_\eta^*\eta} + \frac{1}{\eta}$ minimizes it as a choice of $n - K$ up to constant. If $n \geq T^*$, then we may set $K = 0$ because $X_{n|0}^N = X_n^N$.

On the other hand, Lemma 40 also asserts that

$$
|\mathbb{E}[L(Y_n^N) - L(X_n^N)]| \leq B\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{C(n_\mathrm{tr}-n_\mathrm{b})}{n_\mathrm{b}(n_\mathrm{tr}-1)}}\left\{1 + \frac{M\mu_0}{\lambda}\exp\left(\frac{M\mu_0}{\lambda}\right)\right\}.
$$

Then, by taking the smaller upper bound, we finally obtain that

$$
\begin{aligned}
&|\mathbb{E}[L(Y_n^N) - L(X_n^N)]| \\
&\lesssim \min\left\{\sqrt{r_{T^*\wedge n}} + \sqrt[4]{r_{T^*\wedge n}}, B\sqrt{\frac{\eta}{\lambda/\mu_0}\frac{(n_\mathrm{tr}-n_\mathrm{b})}{n_\mathrm{b}(n_\mathrm{tr}-1)}}\left\{1 + \frac{M\mu_0}{\lambda}\exp\left(\frac{M\mu_0}{\lambda}\right)\right\}\right\},
\end{aligned}
$$

which yields the assertion. ■