



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Hypocoercivity properties of adaptive Langevin dynamics

### Citation for published version:

Leimkuhler, B, Sachs, M & Stoltz, G 2020, 'Hypocoercivity properties of adaptive Langevin dynamics', *Siam Journal on Applied Mathematics*, vol. 80, no. 3, pp. 1197–1222. <https://doi.org/10.1137/19M1291649>

### Digital Object Identifier (DOI):

[10.1137/19M1291649](https://doi.org/10.1137/19M1291649)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Publisher's PDF, also known as Version of record

### Published In:

Siam Journal on Applied Mathematics

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



## HYPOCOERCIVITY PROPERTIES OF ADAPTIVE LANGEVIN DYNAMICS\*

BENEDICT LEIMKUHLER<sup>†</sup>, MATTHIAS SACHS<sup>‡</sup>, AND GABRIEL STOLTZ<sup>§</sup>

**Abstract.** Adaptive Langevin dynamics is a method for sampling the Boltzmann–Gibbs distribution at prescribed temperature in cases where the potential gradient is subject to stochastic perturbation of unknown magnitude. The method replaces the friction in underdamped Langevin dynamics with a dynamical variable, updated according to a negative feedback loop control law as in the Nosé–Hoover thermostat. Using a hypocoercivity analysis we show that the law of Adaptive Langevin dynamics converges exponentially rapidly to the stationary distribution, with a rate that can be quantified in terms of the key parameters of the dynamics. This allows us in particular to obtain a central limit theorem with respect to the time averages computed along a stochastic path. Our theoretical findings are illustrated by numerical simulations involving classification of the MNIST data set of handwritten digits using Bayesian logistic regression.

**Key words.** Langevin dynamics, hypocoercivity, Bayesian inference, stochastic gradients, Nosé–Hoover, sampling

**AMS subject classifications.** 60J70, 35B40, 46N30, 35Q84, 65C30

**DOI.** 10.1137/19M1291649

**1. Introduction.** Langevin dynamics [28, 20, 23] is a system of stochastic differential equations which is traditionally derived as a model of a coarse-grained particle system:

$$(1.1) \quad \begin{aligned} d\mathbf{q} &= \mathbf{M}^{-1}\mathbf{p} \, dt, \\ d\mathbf{p} &= (\mathbf{F}(\mathbf{q}) - \zeta \mathbf{M}^{-1}\mathbf{p}) \, dt + \sigma \, d\mathbf{W}. \end{aligned}$$

Here  $\mathbf{q} \in \mathbb{R}^n$  represents a vector of particle positions,  $\mathbf{p}$  is the corresponding vector of momenta, the mass matrix  $\mathbf{M} \in \mathbb{R}^{n \times n}$  is symmetric positive definite,  $\mathbf{F}$  is the force field (normally the negative gradient of a potential energy function  $U$ ),  $\zeta \in \mathbb{R}$  is a (constant) friction coefficient, and  $\sigma \in \mathbb{R}$  represents the strength of coupling to the stochastic driving force defined by the Wiener process  $\mathbf{W}$ . Although conceived as a dynamical model, Langevin dynamics is among the most versatile and popular methods for computing the statistical properties in high dimension, e.g., for molecular systems or, more recently, for many problems in high-dimensional data analysis. In this approach, the dynamical properties are ignored and the stochastic differential

\*Received by the editors October 7, 2019; accepted for publication (in revised form) February 19, 2020; published electronically May 19, 2020.

<https://doi.org/10.1137/19M1291649>

**Funding:** The work of the first author was supported by the ERC project RULE through grant 320823 and by EPSRC grant EP/P006175/1. The work of the second author was supported by the National Science Foundation under grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute (SAMSI), North Carolina. The work of the third author was supported by the Agence Nationale de la Recherche, under grant ANR-14-CE23-0012 (COSMOS), and by the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013)/ERC grant agreement 614492.

<sup>†</sup>School of Mathematics, University of Edinburgh, Edinburgh EH9 3JZ, UK (b.leimkuhler@ed.ac.uk, <http://kac.maths.ed.ac.uk/~bl/>).

<sup>‡</sup>Department of Mathematics, Duke University, Durham, NC 27708 (msachs@math.duke.edu, <https://math.duke.edu/people/matthias-ernst-sachs>).

<sup>§</sup>Université Paris-Est, CERMICS (ENPC), Inria, F-77455 Marne-la-Vallée, France (gabriel.stoltz@enpc.fr, <http://cermics.enpc.fr/~stoltz>).

equations are discretized to produce “sampling paths” with weights approximating those associated to the (prescribed) Boltzmann–Gibbs stationary distribution with density  $\rho_\beta \propto e^{-\beta U}$ , where, in physical settings,  $\beta$  is the reciprocal of the temperature scaled by Boltzmann’s constant.

The key benefit of Langevin dynamics for sampling, compared to simpler methods such as random walk Monte Carlo, is the use it makes of the gradient of the energy function (or, in the case of data analysis, the “log posterior”; see section 4.3 for an example of Bayesian data analysis) which can effectively guide the collection of sampling paths, resulting in less wasted computation. The use of Langevin dynamics as a sampling scheme is further supported by its well-understood ergodic properties (see [24, 34, 4, 32] and references therein), which ensure exponential convergence of averages to their stationary values, a property which under certain technical conditions on the potential function  $U$  can be shown to carry to numerical discretization [24, 34, 21, 3, 18].

Despite these advantages of Langevin dynamics, in many applications (e.g., mixed quantum and classical molecular dynamics [35, 20] or “big data” [6]) the computation of the force is itself a very challenging task; thus the gradient may be effectively corrupted (due to approximation error), which leads to severe biasing of the invariant distribution. It was for precisely such cases that the Adaptive Langevin dynamics method [16, 7, 33, 22] was created. Note that in the machine learning literature, this method goes by the name “stochastic gradient Nosé–Hoover thermostat” (SGNHT). In this method, the friction  $\zeta$  in (1.1) is reinterpreted as a dynamical variable, defined by a negative feedback loop control law (as in the Nosé–Hoover method [27]). For concreteness, we suppose the gradient noise to be modeled by an additional stochastic process. As discussed in [22], this can, in many cases, be interpreted as an additional (unknown) Itô perturbation  $\sigma_G d\mathbf{W}_G$ , where  $\sigma_G^2$  is unknown and scales linearly with the stepsize used in the discretization of the respective continuous formulation. The system of equations now becomes<sup>1</sup>

$$(1.2) \quad \begin{aligned} d\mathbf{q} &= \mathbf{M}^{-1} \mathbf{p} dt, \\ d\mathbf{p} &= (-\nabla U(\mathbf{q}) - \zeta \mathbf{M}^{-1} \mathbf{p}) dt + \sigma_G d\mathbf{W}_G + \sigma_A d\mathbf{W}_A, \\ d\zeta &= \frac{1}{\nu} \left( \mathbf{p}^T \mathbf{M}^{-2} \mathbf{p} - \frac{1}{\beta} \text{Tr}(\mathbf{M}^{-1}) \right) dt, \end{aligned}$$

where  $\beta$ ,  $\sigma_G$ ,  $\sigma_A$ , and  $\nu$  are positive scalars, and  $\mathbf{W}_A, \mathbf{W}_G$  are two independent Wiener processes in  $\mathbb{R}^n$  with independent components (“A” stands for “applied,” and “G” stands for “gradient”). The auxiliary variable  $\zeta$  now acts as a variable friction which restores the canonical distribution associated with the prescribed inverse temperature  $\beta$ . The system (1.2) admits the invariant probability measure (see section 2)

$$(1.3) \quad \pi(d\mathbf{q} d\mathbf{p} d\zeta) = Z^{-1} \exp \left( -\beta \left[ \frac{\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}}{2} + U(\mathbf{q}) + \frac{\nu}{2} (\zeta - \gamma)^2 \right] \right) d\mathbf{q} d\mathbf{p} d\zeta,$$

where  $Z$  is a normalization constant and

$$(1.4) \quad \gamma = \frac{\beta(\sigma_G^2 + \sigma_A^2)}{2}.$$

<sup>1</sup>The formulation in [22] is slightly different in the form of the control law as a consequence of a linear transformation of the momenta in the presentation of the frictional force.

Assuming ergodicity, the system (1.2) allows sampling of the Boltzmann–Gibbs probability measure with density proportional to  $e^{-\beta[\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}/2 + U(\mathbf{q})]}$ , by marginalization, and proportional to  $\rho_\beta$  if the momenta are ignored.

The practical value of (1.2) is that it allows simulations to be performed for complicated systems in which the potential energy function  $U$  and its gradient are the consequence of substantial calculations and thus entail computational errors. The original motivation of the article of Jones and Leimkuhler [16] was in the context of multiscale models of molecular systems where the force laws were computed using a separate numerical method and the error in this process was assumed to have the character of white noise. More recently, (1.2) has been adopted in the setting of sampling of Bayesian posterior distributions in large scale data science applications [5], where the gradient noise is the consequence of incomplete calculation of the log-likelihood function based on subsampling data points from a large data set, as in the stochastic gradient Langevin dynamics method [37]. In this setup the potential function  $U$  corresponds to the negative log posterior density of a statistical model; i.e., for independent observations  $x^1, x^2, \dots, x^{\tilde{N}}$ , the negative gradient of  $U$  is of the form

$$(1.5) \quad -\nabla U(\mathbf{q}) = \nabla \log p_0(\mathbf{q}) + \sum_{j=1}^{\tilde{N}} \nabla \log p(x^j | \mathbf{q})$$

where  $p_0$  is a prior density and  $p(x^j | \mathbf{q})$  is the likelihood of the  $j$ th observation. In order to avoid the linear scaling in  $\tilde{N}$  of the computational cost per evaluation of the force (1.5), the gradient force  $-\nabla U(\mathbf{q})$  is commonly replaced by an unbiased estimator  $-\widehat{\nabla} U(\mathbf{q})$  in discretizations of (1.2). That is,

$$(1.6) \quad -\widehat{\nabla} U(\mathbf{q}) = \nabla \log \pi(\mathbf{q}) + \frac{\tilde{N}}{m} \sum_{j \in B} \nabla \log p(x^j | \mathbf{q}),$$

where  $B = \{J_l\}_{l=1}^m$ ,  $m \ll \tilde{N}$ , is a subset of the complete data index set—commonly referred to as a *minibatch*—which is comprised of uniformly and independently sampled data point indices  $J_l \in \{1, \dots, \tilde{N}\}$ ,  $l = 1, \dots, m$ , which are resampled with replacement at the beginning of every time step of a discretization of (1.2).

*Remark 1.1.* We point out that modeling the gradient noise by a white noise of the form  $\sigma_G d\mathbf{W}_G$  assumes that the covariance of the estimation error is constant in  $\mathbf{q}$  and isotropic. It is possible to extend the analysis to more general covariance matrices as long as they do not depend on  $\mathbf{q}$ . However, the latter assumption is not satisfied in most likelihood models. While in practice the Adaptive Langevin formalism nonetheless usually results in good approximations of the target measure, there are in fact no theoretical guarantees and/or bounds for the incurred error. This is a well-known limitation of the approach and a subject of current work by the authors (amongst others).

Although the presence of noise in the Adaptive Langevin model in contact with all momenta suggests hypoellipticity (as for Langevin dynamics [24]), the way in which convergence is achieved in the Adaptive Langevin system is not straightforward. Given a stochastic differential equation system with generator  $\mathcal{L}$ , let us recall that there are several well-studied frameworks which can be used to derive exponential convergence rates for the semigroup  $e^{t\mathcal{L}}$  (or equivalently for the respective adjoint semigroup) in certain functional spaces.

First, there are probabilistic techniques, which allow the derivation of exponential convergence rates of  $e^{t\mathcal{L}}$  when considered as a family of operators on weighted  $L^\infty$  spaces (see, e.g., [25, 26, 24]), or exponential convergence rates of the formally adjoint semigroup acting on Wasserstein metric spaces (see, e.g., [10, 11]). Second, there are also functional analytic proofs for exponential convergence for the case of weighted  $L^\infty$  spaces; see [30, 12]. The naive application of these methods fails in the case of (1.2) due to a lack of direct stochastic control of the auxiliary variable  $\zeta$ . It was only very recently shown in [13] that a suitable Lyapunov function can be constructed for this system which allows us to conclude exponential convergence in a weighted  $L^\infty$  space.

The approach taken here is based on a third method, the alternative hypocoercivity framework of Villani [36], as further developed by Dolbeault, Mouhot, and Schmeiser [8, 9], which can be used to derive exponential convergence rates of the semigroup when considered as a family of operators acting on subspaces of  $L^2(\mu)$ , where  $\mu$  denotes the (unique) invariant measure of the stochastic process under consideration. This technique can be applied to derive geometric convergence estimates for the underdamped Langevin equation [9, 31, 15]. We show that this framework can also be applied directly to the system (1.2), thus demonstrating the rapid convergence in law of the Adaptive Langevin system.

The exponential convergence shown here has important consequences for the statistics of the samples obtained using the Adaptive Langevin method. In particular it allows us to establish a central limit theorem (CLT). Our approach also allows us to characterize the asymptotic scaling of the spectral gap of the generator associated with (1.2) when considered as an operator on the respective weighted  $L^2$  space as  $O(\min(\gamma, \gamma^{-1}, \gamma\nu, \gamma^{-1}\nu^{-1}))$ , a qualitative characterization of the spectral gap which is missing in the analysis in [13]. The derived asymptotic scaling on the lower bounds of the spectral gap allows us in turn to conclude an asymptotic scaling of the asymptotic variance in the above-mentioned CLT as  $O(\max(\gamma, \gamma^{-1}, \gamma\nu, \gamma^{-1}\nu^{-1}))$ ; see the discussion in Remark 2.5 for an informal motivation of some terms in this asymptotic scaling.

The remainder of this paper is structured as follows. In section 2 we begin by rewriting the generators of the dynamics (1.2), where we also check the invariance of the probability measure (1.3). In subsection 2.1 we normalize the dynamics (1.2) in order to study limiting regimes associated with vanishing or diverging key parameters of the dynamics (namely the thermal mass  $\nu$  and the magnitude of the fluctuation). We can then discuss requirements of the potential energy function (subsection 2.2) and state the exponential convergence of the evolution semigroup in subsection 2.3. The CLT is derived in section 3, with upper bounds on the asymptotic variance made precise in terms of the key parameters of the dynamics. Finally, we show in subsection 3.1 that the asymptotic variance converges in the large thermal mass limit to the asymptotic variance of standard Langevin dynamics. Section 4 contains numerical experiments assessing the relevance of parameter scalings used and demonstrating the CLT in an application to Bayesian sampling.

**2. Hypocoercivity of Adaptive Langevin dynamics.** We assume that the potential energy function  $U$  is smooth, i.e.,  $U \in C^\infty(\mathbb{R}^n, \mathbb{R})$ , and such that  $e^{-\beta U(\mathbf{q})}$  is integrable. In particular, (1.3) is a well-defined probability measure. We first show that the probability measure (1.3) is indeed invariant under the dynamics (1.2).

The generator of (1.2) acts on functions  $\varphi = \varphi(\mathbf{q}, \mathbf{p}, \zeta)$  with  $(\mathbf{q}, \mathbf{p}, \zeta) \in \mathbb{R}^{2n+1}$ . It can be written as  $\mathcal{L}_{\text{AdL}} = \mathcal{L}_H + \gamma\mathcal{L}_O + \nu^{-1}\mathcal{L}_{\text{NH}}$ , where the action of these generators

can be defined on the core of  $C^\infty(\mathbb{R}^{2n+1}, \mathbb{R})$  functions with compact support as

$$\begin{aligned} \mathcal{L}_H &= \mathbf{p}^T \mathbf{M}^{-1} \nabla_q - \nabla U(\mathbf{q})^T \nabla_p = \frac{1}{\beta} (\nabla_p^* \nabla_q - \nabla_q^* \nabla_p) = \frac{1}{\beta} \sum_{i=1}^n \partial_{p_i}^* \partial_{q_i} - \partial_{q_i}^* \partial_{p_i}, \\ \mathcal{L}_O &= -\mathbf{p}^T \mathbf{M}^{-1} \nabla_p + \frac{1}{\beta} \Delta_p = -\frac{1}{\beta} \nabla_p^* \nabla_p = -\frac{1}{\beta} \sum_{i=1}^n \partial_{p_i}^* \partial_{p_i}, \end{aligned} \quad (2.1)$$

and

$$\begin{aligned} \mathcal{L}_{NH} &= -\nu(\zeta - \gamma) \mathbf{p}^T \mathbf{M}^{-1} \nabla_p + \left( \mathbf{p}^T \mathbf{M}^{-2} \mathbf{p} - \frac{1}{\beta} \text{Tr}(\mathbf{M}^{-1}) \right) \partial_\zeta \\ &= \frac{1}{\beta^2} ((\partial_\zeta - \partial_\zeta^*) \nabla_p^* \nabla_p + \Delta_p^* \partial_\zeta - \Delta_p \partial_\zeta^*), \end{aligned} \quad (2.2)$$

where adjoints are taken on  $L^2(\pi)$ . A simple computation indeed shows that  $\partial_{q_i}^* = -\partial_{q_i} + \beta \partial_{q_i} U(q)$ ,  $\partial_{p_i}^* = -\partial_{p_i} + \beta(\mathbf{M}^{-1} \mathbf{p})_i$ ,  $\partial_\zeta^* = -\partial_\zeta + \beta \nu(\zeta - \gamma)$ , and

$$\Delta_p^* = \Delta_p - 2\beta \mathbf{p}^T \mathbf{M}^{-1} \nabla_p + \beta^2 \left( \mathbf{p}^T \mathbf{M}^{-2} \mathbf{p} - \frac{1}{\beta} \text{Tr}(\mathbf{M}^{-1}) \right).$$

The above rewriting in terms of the elementary operators  $\partial_{q_i}, \partial_{p_i}, \partial_\zeta$  and their adjoints immediately shows that  $\mathcal{L}_O$  is symmetric, while  $\mathcal{L}_H$  and  $\mathcal{L}_{NH}$  are antisymmetric. Let us however emphasize that this decomposition is only used for mathematical convenience: the parameter  $\gamma$  is in fact unknown since  $\sigma_G$  is not known in practice.

Another benefit of the rewriting (2.1)–(2.2) is that the actions of the operators  $\mathcal{L}_H, \mathcal{L}_O, \mathcal{L}_{NH}$  make it clear that the measure with density (1.3) is indeed invariant since  $\mathcal{A}\mathbf{1} = 0$  for  $\mathcal{A} \in \{\mathcal{L}_H, \mathcal{L}_O, \mathcal{L}_{NH}\}$ , so that (denoting by  $C_0^\infty(\mathbb{R}^{2n+1}, \mathbb{R})$  the space of  $C^\infty$  functions with compact support in  $\mathbb{R}^{2n+1}$ )

$$\forall \varphi \in C_0^\infty(\mathbb{R}^{2n+1}, \mathbb{R}), \quad \int_{\mathbb{R}^{2n+1}} \mathcal{A} \varphi \, d\pi = \sigma \int_{\mathbb{R}^{2n+1}} \varphi \mathcal{A}\mathbf{1} \, d\pi = 0,$$

with  $\sigma = 1$  for  $\mathcal{A} = \mathcal{L}_O$  and  $\sigma = -1$  for  $\mathcal{A} \in \{\mathcal{L}_H, \mathcal{L}_{NH}\}$ , and where we relied for  $\mathcal{L}_{NH}$  on the fact that elementary operators acting on different variables commute. Therefore,

$$\forall \varphi \in C_0^\infty(\mathbb{R}^{2n+1}, \mathbb{R}), \quad \int_{\mathbb{R}^{2n+1}} \mathcal{L}_{AdL} \varphi \, d\pi = 0,$$

which proves the invariance of  $\pi$  under the dynamics (1.2) (see, for instance, [23]).

**2.1. Normalization of the dynamics.** To simplify the notation we let  $\mathbf{M} = \mathbf{I}$ . Let us however emphasize that our proofs and results can be adapted in a straightforward way to accommodate general mass matrices. As one of our interests in this work is to understand the limiting regimes  $\gamma \rightarrow 0$  or  $+\infty$  and/or  $\nu \rightarrow 0$  or  $+\infty$  of the Adaptive Langevin dynamics, we also need to rescale the friction variable  $\zeta$  in order for the invariant measure to be independent of the parameter  $\nu$ . More precisely, we set  $\varepsilon = \sqrt{\nu}$  and consider  $\xi = \sqrt{\nu}(\zeta - \gamma)$ , i.e.,

$$\zeta = \gamma + \frac{\xi}{\varepsilon}.$$

The latter change of variables is motivated by the fact that the invariant measure (1.3) now becomes (slightly abusing the notation  $\pi$ )

$$(2.3) \quad \pi(d\mathbf{q} \, d\mathbf{p} \, d\xi) = Z^{-1} \exp \left( -\beta \left[ \frac{\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}}{2} + U(\mathbf{q}) + \frac{\xi^2}{2} \right] \right) d\mathbf{q} \, d\mathbf{p} \, d\xi.$$

Let us emphasize that this invariant probability measure does not depend on the parameters  $\gamma, \varepsilon$ . The dynamics (1.2) then becomes

$$(2.4) \quad \begin{aligned} d\mathbf{q} &= \mathbf{p} \, dt, \\ d\mathbf{p} &= \left( -\nabla U(\mathbf{q}) - \frac{\xi}{\varepsilon} \mathbf{p} - \gamma \mathbf{p} \right) dt + \sqrt{\frac{2\gamma}{\beta}} d\mathbf{W}, \\ d\xi &= \frac{1}{\varepsilon} \left( |\mathbf{p}|^2 - \frac{n}{\beta} \right) dt, \end{aligned}$$

where  $|\mathbf{p}| = \sqrt{p_1^2 + \cdots + p_n^2}$  is the Euclidean norm of  $\mathbf{p} \in \mathbb{R}^n$ . The generator of this SDE is

$$(2.5) \quad \mathcal{L}_{\text{AdL}} = \mathcal{L}_H + \gamma \mathcal{L}_O + \varepsilon^{-1} \mathcal{L}_{\text{NH}},$$

with the above definitions (2.1) for  $\mathcal{L}_H$  and  $\mathcal{L}_O$  (upon replacing  $\mathbf{M}$  with  $\mathbf{I}$ ) and

$$(2.6) \quad \mathcal{L}_{\text{NH}} = \left( |\mathbf{p}|^2 - \frac{n}{\beta} \right) \partial_\xi - \xi \mathbf{p}^T \nabla_p = \frac{1}{\beta^2} \left( (\partial_\xi - \partial_\xi^*) \nabla_p^* \nabla_p + \Delta_p^* \partial_\xi - \Delta_p \partial_\xi^* \right).$$

**2.2. Assumptions and notation.** We denote by  $\pi_q, \pi_p, \pi_\xi$  the marginals of the probability measure (2.3) in the variables  $\mathbf{q}, \mathbf{p}$ , and  $\xi$ , respectively, so that  $\pi(d\mathbf{q} d\mathbf{p} d\xi) = \pi_q(d\mathbf{q}) \pi_p(d\mathbf{p}) \pi_\xi(d\xi)$ . Further let  $\|\cdot\|_{L^2(\pi)}$  be the norm on the Hilbert space  $L^2(\pi)$  induced by the canonical scalar product, and denote by  $L_0^2(\pi)$  the subspace of  $L^2(\pi)$  of functions with vanishing mean,

$$(2.7) \quad L_0^2(\pi) = \left\{ \varphi \in L^2(\pi) \left| \int_{\mathbb{R}^{2n+1}} \varphi \, d\pi = 0 \right. \right\},$$

and by  $\Pi_0 : L^2(\pi) \rightarrow L_0^2(\pi)$  the orthogonal projection operator onto this subspace, i.e.,

$$(2.8) \quad \Pi_0 \varphi = \varphi - \int_{\mathbb{R}^{2n+1}} \varphi \, d\pi.$$

In the remainder of this article we consider all operators as being defined on  $L^2(\pi)$  unless explicitly specified otherwise. The associated operator norm for bounded operators on  $L^2(\pi)$  is

$$\|\mathcal{T}\| = \sup_{\varphi \in L^2(\pi) \setminus \{0\}} \frac{\|\mathcal{T}\varphi\|_{L^2(\pi)}}{\|\varphi\|_{L^2(\pi)}}.$$

For an operator  $\mathcal{T}$  on  $L^2(\pi)$  with dense domain, we denote by  $\mathcal{T}^*$  its  $L^2(\pi)$ -adjoint. Throughout the remainder of this article we assume that the potential function  $U$  satisfies the following assumption.

**ASSUMPTION 1.** *The potential function is smooth, i.e.,  $U \in \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$ , and the associated probability measure  $\pi_q(d\mathbf{q}) = Z_q^{-1} e^{-\beta U(\mathbf{q})} d\mathbf{q}$  satisfies a Poincaré inequality: there exists  $\kappa_q > 0$  such that*

$$(2.9) \quad \forall \varphi \in H^1(\pi_q), \quad \left\| \varphi - \int_{\mathbb{R}^n} \varphi \, d\pi_q \right\|_{L^2(\pi_q)} \leq \frac{1}{\kappa_q} \|\nabla \varphi\|_{L^2(\pi_q)}.$$

Moreover, there exist  $c_1 > 0$ ,  $c_2 \in [0, 1)$ , and  $c_3 > 0$  such that

$$(2.10) \quad \Delta U \leq c_1 + \frac{c_2 \beta}{2} |\nabla U|^2, \quad |\nabla^2 U| \leq c_3 (1 + |\nabla U|).$$

The second condition, taken from [9, section 3], ensures that the operator  $(1 + \nabla_q^* \nabla_q)^{-1}$  is bounded from  $L^2(\pi_q)$  to  $H^2(\pi_q)$ . It will be used in technical estimates related to the proof of exponential convergence of the semigroup (see Lemma 2.11). A sufficient condition on  $U$  for  $\pi_q$  to satisfy a Poincaré inequality is for example the following (see [1, Corollary 1.6]): there exist  $a \in (0, 1)$ ,  $c > 0$ , and  $R \geq 0$  such that

$$\forall q \in \mathbb{R}^n \text{ such that } |q| \geq R, \quad a\beta |\nabla U(q)|^2 - \Delta U(q) \geq c.$$

The latter condition and (2.10) hold for instance for potentials which behave asymptotically as  $|q|^\alpha$  with  $\alpha > 1$  as  $|q| \rightarrow \infty$ .

*Remark 2.1* (singular potentials). Note that our set of assumptions excludes singular potentials. Exponential convergence of Adaptive Langevin dynamics for such potentials follows from the results in [13], which are based on Lyapunov estimates.

**2.3. Exponential convergence of the law and invertibility of the generator.** The following result states the exponential convergence in  $L^2(\pi)$  of the semigroup  $e^{t\mathcal{L}_{\text{AdL}}}$  associated with the dynamics (2.4).

**THEOREM 2.2.** *There exist  $C, \bar{\lambda}$  such that, for any  $\varepsilon, \gamma > 0$ , there is  $\lambda_{\varepsilon, \gamma} > 0$  for which*

$$(2.11) \quad \forall t \geq 0, \quad \forall \varphi \in L^2(\pi), \quad \left\| e^{t\mathcal{L}_{\text{AdL}}} \varphi - \int \varphi d\pi \right\|_{L^2(\pi)} \leq C e^{-t\lambda_{\varepsilon, \gamma}} \left\| \varphi - \int \varphi d\pi \right\|_{L^2(\pi)},$$

with the lower bound

$$(2.12) \quad \lambda_{\varepsilon, \gamma} \geq \bar{\lambda} \min \left( \gamma, \frac{1}{\gamma}, \gamma\varepsilon^2, \frac{1}{\gamma\varepsilon^2} \right).$$

Theorem 2.2 immediately implies the existence of the inverse of  $\mathcal{L}_{\text{AdL}}$  on  $L_0^2(\pi)$  and allows us to obtain bounds on the norm of the inverse in terms of the parameters  $\gamma, \varepsilon$  (see [23, Proposition 2.1]).

**COROLLARY 2.3.** *The operator  $\mathcal{L}_{\text{AdL}}$  considered on  $L_0^2(\pi)$  is invertible and*

$$\mathcal{L}_{\text{AdL}}^{-1} = - \int_0^\infty e^{t\mathcal{L}_{\text{AdL}}} dt, \quad \|\mathcal{L}_{\text{AdL}}^{-1}\|_{\mathcal{B}(L_0^2(\pi))} \leq \frac{C}{\bar{\lambda}} \max \left( \gamma, \frac{1}{\gamma}, \gamma\varepsilon^2, \frac{1}{\gamma\varepsilon^2} \right).$$

Simple computations show that some of these bounds on the resolvent are sharp. Indeed,

$$\mathcal{L}_{\text{AdL}} \left( U + \frac{|\mathbf{p}|^2}{2} \right) = \gamma \left[ \frac{n}{\beta} - \left( 1 + \frac{\xi}{\gamma\varepsilon} \right) |\mathbf{p}|^2 \right],$$

which implies that there exists  $a > 0$  such that  $\|\mathcal{L}_{\text{AdL}}^{-1}\|_{\mathcal{B}(L_0^2(\pi))} \geq a\gamma^{-1}$  by choosing  $\gamma$  small and  $\gamma\varepsilon$  large. Next,

$$\mathcal{L}_{\text{AdL}} (\gamma U + \mathbf{p}^T \nabla U) = \mathbf{p}^T (\nabla^2 U) \mathbf{p} - |\nabla U|^2 - \frac{1}{\varepsilon} \xi \mathbf{p}^T \nabla U,$$

which shows that there exists  $b > 0$  such that  $\|\mathcal{L}_{\text{AdL}}^{-1}\|_{\mathcal{B}(L_0^2(\pi))} \geq b\gamma$  by choosing  $\gamma$  large and  $\varepsilon = 1$ . Finally,

$$\mathcal{L}_{\text{AdL}} \left( \gamma\varepsilon\xi + \frac{|p|^2}{2} - \frac{1}{\gamma} \mathbf{p}^T \nabla U \right) = -\frac{1}{\varepsilon} \xi |p|^2 + \frac{1}{\gamma\varepsilon} \mathbf{p}^T \nabla U - \frac{1}{\gamma} (\mathbf{p}^T (\nabla^2 U) \mathbf{p} - |\nabla U|^2),$$



which shows that there exists  $c > 0$  such that  $\|\mathcal{L}_{\text{AdL}}^{-1}\|_{\mathcal{B}(L_0^2(\pi))} \geq c\gamma\varepsilon^2$  by choosing  $\gamma \gg \varepsilon \gg 1$ . It is on the other hand not so easy to find functions which saturate the upper bound  $1/(\gamma\varepsilon^2)$  of the resolvent since this requires a careful analysis in the regime  $\varepsilon \rightarrow 0$ , which corresponds to a singular limit where the dominant part of the dynamics is the deterministic Nosé–Hoover feedback; see Remark 3.3 below. We however demonstrate numerically the sharpness of the upper bound in section 4.1.

The proof of Theorem 2.2 relies on the hypocoercive framework of [8, 9]. The exponential decay is obtained by a Grönwall inequality in a modified norm on  $L^2(\pi)$ . The choice of the modified norm is motivated by the fact that  $\mathcal{L}_{\text{AdL}}$  is coercive in the corresponding scalar product. More precisely, we consider

$$(2.13) \quad \mathcal{H}(\varphi) = \frac{1}{2} \|\varphi\|_{L^2(\pi)}^2 + a_{\varepsilon, \gamma} \langle A_\varepsilon \varphi, \varphi \rangle_{L^2(\pi)},$$

where  $A_\varepsilon$  is a bounded operator constructed from the antisymmetric part  $\mathcal{A}_\varepsilon := \mathcal{L}_H + \varepsilon^{-1} \mathcal{L}_{\text{NH}}$  of the generator, and  $a_{\varepsilon, \gamma} \in (0, 1)$  is a constant. Formally, for a generator  $\mathcal{L}$  which can be decomposed as the sum of a symmetric part  $\mathcal{L}_{\text{sym}}$ , coercive on the range of a projector  $\mathcal{P}$ , and an antisymmetric part  $\mathcal{L}_{\text{anti}}$ , the approach in [8, 9] suggests considering the regularization operator  $-(1 - \mathcal{P}\mathcal{L}_{\text{anti}}^2\mathcal{P})^{-1}\mathcal{P}\mathcal{L}_{\text{anti}}$ . In order for the coercivity properties to behave well with  $\varepsilon$ , we however need to distinguish the “dominant” part of  $\mathcal{L}_{\text{anti}}$  and possibly renormalize this operator by a factor  $\varepsilon$  so that the antisymmetric part remains of order 1 as  $\varepsilon$  goes either to 0 or to  $+\infty$ .

More precisely, the expression of  $A_\varepsilon$  distinguishes whether  $\varepsilon \leq 1$  or  $\varepsilon \geq 1$ . For  $\varepsilon \in (0, 1]$ , the small term in  $\mathcal{A}_\varepsilon$  is the Hamiltonian one and the expression of  $A_\varepsilon$  is the one suggested in [9], namely  $-[1 - \Pi\mathcal{A}_\varepsilon^2\Pi]^{-1}\Pi\mathcal{A}_\varepsilon$  where  $\Pi$  is the orthogonal projector on  $L^2(\pi)$  corresponding to the partial integration with respect to  $\pi_p(\text{d}\mathbf{p})$ :

$$(2.14) \quad (\Pi\varphi)(\mathbf{q}, \xi) = \int_{\mathbb{R}^n} \varphi(\mathbf{q}, \mathbf{p}, \xi) \pi_p(\text{d}\mathbf{p}).$$

For  $\varepsilon \in [1, \infty)$ , the small term in  $\mathcal{A}_\varepsilon$  is the one associated with the Nosé–Hoover-like feedback mechanism, in which case one should rescale the generator as  $\varepsilon\mathcal{L}_{\text{AdL}}$  in order to avoid degeneracies as  $\varepsilon \rightarrow +\infty$ . Up to this multiplication by  $\varepsilon$ , the regularization operator is defined as above and therefore reads  $-\varepsilon[1 - \varepsilon^2\Pi\mathcal{A}_\varepsilon^2\Pi]^{-1}\Pi\mathcal{A}_\varepsilon$ . This modification turns out to be crucial to obtain the key partial coercivity (2.19) with the appropriate rate (see the discussion following this inequality). We therefore use the following regularization operator, which reduces to the expressions discussed above upon distinguishing  $\varepsilon \leq 1$  or  $\varepsilon \geq 1$ :

$$\begin{aligned} A_\varepsilon &:= -\min\left(1, \frac{1}{\varepsilon}\right) \left[ \min\left(1, \frac{1}{\varepsilon^2}\right) - \Pi\mathcal{A}_\varepsilon^2\Pi \right]^{-1} \Pi\mathcal{A}_\varepsilon \\ &= -\min\left(1, \frac{1}{\varepsilon}\right) \left[ \min\left(1, \frac{1}{\varepsilon^2}\right) + \Pi \left( \frac{2n}{(\beta\varepsilon)^2} \partial_\xi^* \partial_\xi + \frac{1}{\beta} \nabla_q^* \nabla_q \right) \Pi \right]^{-1} \Pi\mathcal{A}_\varepsilon. \end{aligned}$$

The second expression is a consequence of the following equations:

$$\begin{aligned} \Pi\mathcal{L}_H^2\Pi &= -\frac{1}{\beta^2} \sum_{i=1}^n \Pi \partial_{p_i} \partial_{p_i}^* \Pi \partial_{q_i}^* \partial_{q_i} = -\frac{1}{\beta} \nabla_q^* \nabla_q, \\ \Pi\mathcal{L}_{\text{NH}}^2\Pi &= -\frac{1}{\beta^4} \Pi \Delta_p \Delta_p^* \Pi \partial_\xi^* \partial_\xi = -\frac{2n}{\beta^2} \partial_\xi^* \partial_\xi, \\ \Pi\mathcal{L}_{\text{NH}}\mathcal{L}_H\Pi &= \Pi\mathcal{L}_H\mathcal{L}_{\text{NH}}\Pi = 0, \end{aligned}$$

which are direct consequences of the expressions (2.1) and (2.6) for the generators in terms of the elementary operators  $\partial_{q_i}, \partial_{p_i}, \partial_\xi$ , as well as the following rules (which can be checked by direct computations):

$$(2.15) \quad \partial_{p_i} \Pi = 0, \quad \Pi \partial_{p_i}^* = 0, \quad \partial_{p_i} \partial_{p_i}^* \Pi = \beta \Pi, \quad \partial_{p_j}^2 (\partial_{p_i}^*)^2 \Pi = 2\beta^2 \Pi \delta_{ij}.$$

It can be shown that the norm of  $A_\varepsilon$  is bounded by  $1/2$  (see Lemma 2.6), so that  $\sqrt{\mathcal{H}(\cdot)}$  defines a norm equivalent to the standard norm on  $L^2(\pi)$  for any  $a_{\varepsilon, \gamma} \in (-1, 1)$ :

$$(2.16) \quad \sqrt{\frac{1 - |a_{\varepsilon, \gamma}|}{2}} \|\varphi\|_{L^2(\pi)} \leq \sqrt{\mathcal{H}(\varphi)} \leq \sqrt{\frac{1 + |a_{\varepsilon, \gamma}|}{2}} \|\varphi\|_{L^2(\pi)}.$$

By polarization we can define a real valued inner product associated with  $\sqrt{\mathcal{H}(\cdot)}$  as

$$(2.17) \quad \begin{aligned} \langle f, g \rangle_{\varepsilon, \gamma} &:= \mathcal{H}(f + g) - \mathcal{H}(f) - \mathcal{H}(g) \\ &= \langle f, g \rangle_{L^2(\pi)} + a_{\varepsilon, \gamma} \langle A_\varepsilon f, g \rangle_{L^2(\pi)} + a_{\varepsilon, \gamma} \langle A_\varepsilon g, f \rangle_{L^2(\pi)}. \end{aligned}$$

Most importantly, the construction of the operator  $A_\varepsilon$  ensures that  $\mathcal{L}_{\text{AdL}}$  is coercive for the modified scalar product (2.17), as made precise in the following key result (see section 2.4 for the proof).

**PROPOSITION 2.4.** *There exist  $\bar{a} \in (0, 1)$  and  $\tilde{\lambda} > 0$  such that, for any  $\varepsilon, \gamma > 0$  and upon choosing  $a_{\varepsilon, \gamma} = \bar{a} \min(\gamma, \gamma^{-1}, \gamma\varepsilon^2, (\gamma\varepsilon^2)^{-1})$  in (2.13),*

$$\forall \varphi \in C_0^\infty(\mathbb{R}^{2n+1}) \cap L_0^2(\pi), \quad \langle -\mathcal{L}_{\text{AdL}} \varphi, \varphi \rangle_{\varepsilon, \gamma} \geq \tilde{\lambda} \min\left(\gamma, \frac{1}{\gamma}, \gamma\varepsilon^2, \frac{1}{\gamma\varepsilon^2}\right) \|\varphi\|_{L^2(\pi)}^2.$$

Theorem 2.2 then follows from the inequality

$$\frac{d}{dt} \left[ \mathcal{H} \left( e^{t\mathcal{L}_{\text{AdL}}} \varphi \right) \right] = \langle \mathcal{L}_{\text{AdL}} e^{t\mathcal{L}_{\text{AdL}}} \varphi, e^{t\mathcal{L}_{\text{AdL}}} \varphi \rangle_{\varepsilon, \gamma} \leq -\tilde{\lambda} \min\left(\gamma, \frac{1}{\gamma}, \gamma\varepsilon^2, \frac{1}{\gamma\varepsilon^2}\right) \left\| e^{t\mathcal{L}_{\text{AdL}}} \varphi \right\|_{L^2(\pi)}^2,$$

upon using the equivalence of norms (2.16) and resorting to a Grönwall lemma.

**Remark 2.5.** We motivate why three out of the four terms are expected in the scaling (2.12) of the lower bound. First, if  $\mathcal{L}_{\text{NH}} = 0$ , the remaining part  $\mathcal{L}_{\text{H}} + \gamma\mathcal{L}_{\text{O}}$  of the generator corresponds to the underdamped Langevin equation, whose spectral gap is bounded from above by a term proportional to  $O(\min(\gamma, \gamma^{-1}))$ . Similarly, in the case  $\mathcal{L}_{\text{H}} = 0$ , it can be verified that the framework of [8] can be directly applied to  $\mathcal{L}_{\text{AdL}} = \varepsilon^{-1}(\varepsilon\gamma\mathcal{L}_{\text{O}} + \mathcal{L}_{\text{NH}})$  considered as an operator on  $L^2(\pi_p \pi_\xi)$ , meaning that the spectral gap of this operator scales as  $O(\varepsilon^{-1} \min(\gamma\varepsilon, (\gamma\varepsilon)^{-1})) = O(\min(\gamma, \gamma^{-1}\varepsilon^{-2}))$ . The only term we do not capture by this simple analysis is  $\gamma\varepsilon^2$ . This is not a surprise since the origin of this limitation on the convergence rate comes from the interaction between the Hamiltonian and Nosé–Hoover parts; see Remark 2.10 below.

**2.4. Proof of Proposition 2.4.** In the remainder of this section, we use the shorthand notation

$$\eta_\varepsilon = \min(1, \varepsilon^{-1}).$$

We first review a few properties of the operator  $A_\varepsilon$  (obtained by a straightforward adaptation of [9, Lemma 1]).

**LEMMA 2.6.** *The operators  $A_\varepsilon$  and  $\mathcal{A}_\varepsilon A_\varepsilon$  are bounded, and  $\Pi A_\varepsilon = A_\varepsilon$ . Furthermore, for any  $f \in L^2(\pi)$ ,*

$$\|A_\varepsilon f\|_{L^2(\pi)} \leq \frac{1}{2} \|(1 - \Pi)f\|_{L^2(\pi)}, \quad \|\mathcal{A}_\varepsilon A_\varepsilon f\|_{L^2(\pi)} \leq \eta_\varepsilon \|(1 - \Pi)f\|_{L^2(\pi)}.$$

*Proof.* Consider  $f \in L^2(\pi)$  and  $u = A_\varepsilon f$ . Then,  $(\eta_\varepsilon^2 - \Pi \mathcal{A}_\varepsilon^2 \Pi)u = -\eta_\varepsilon \Pi \mathcal{A}_\varepsilon f$ . This equality already shows that  $\Pi u = u$ , i.e.,  $\Pi A_\varepsilon = A_\varepsilon$ . Moreover, upon taking the scalar product with  $u$ , and noting that  $\Pi \mathcal{A}_\varepsilon \Pi = 0$ ,

$$\begin{aligned} \eta_\varepsilon^2 \|u\|_{L^2(\pi)}^2 + \|\mathcal{A}_\varepsilon \Pi u\|_{L^2(\pi)}^2 &= -\eta_\varepsilon \langle \mathcal{A}_\varepsilon \Pi u, (1 - \Pi)f \rangle_{L^2(\pi)} \\ &\leq \eta_\varepsilon \|\mathcal{A}_\varepsilon \Pi u\|_{L^2(\pi)} \|(1 - \Pi)f\|_{L^2(\pi)} \leq \frac{\eta_\varepsilon^2}{4} \|(1 - \Pi)f\|_{L^2(\pi)}^2 + \|\mathcal{A}_\varepsilon \Pi u\|_{L^2(\pi)}^2, \end{aligned}$$

which implies the claimed inequalities.  $\square$

We now fix  $\varphi \in C_0^\infty(\mathbb{R}^{2n+1}) \cap L_0^2(\pi)$  and evaluate

$$(2.18) \quad \begin{aligned} \langle -\mathcal{L}_{\text{AdL}} \varphi, \varphi \rangle_{\varepsilon, \gamma} &= -\gamma \langle \mathcal{L}_O \varphi, \varphi \rangle_{L^2(\pi)} + a_{\varepsilon, \gamma} \langle \mathcal{A}_\varepsilon A_\varepsilon \varphi, \varphi \rangle_{L^2(\pi)} \\ &\quad - a_{\varepsilon, \gamma} \langle A_\varepsilon \mathcal{A}_\varepsilon \varphi, \varphi \rangle_{L^2(\pi)} - \gamma a_{\varepsilon, \gamma} \langle A_\varepsilon \mathcal{L}_O \varphi, \varphi \rangle_{L^2(\pi)}, \end{aligned}$$

where we have used the fact that  $\langle \mathcal{L}_{\text{AdL}} \varphi, \varphi \rangle_{L^2(\pi)} = \langle \mathcal{L}_O \varphi, \varphi \rangle_{L^2(\pi)}$ , and  $\mathcal{L}_O A_\varepsilon = \mathcal{L}_O \Pi A_\varepsilon = 0$ . We next consider the four terms on the right-hand side of (2.18):

- The expression (2.1) shows that  $-\langle \mathcal{L}_O \varphi, \varphi \rangle_{L^2(\pi)} = \beta^{-1} \|\nabla_p \varphi\|_{L^2(\pi)}^2 \geq \beta^{-1} \kappa_p^2 \|(1 - \Pi)\varphi\|_{L^2(\pi)}^2$  from a Poincaré inequality for the Gaussian measure in  $\mathbf{p}$ , pointwise in  $(\mathbf{q}, \xi)$  and then integrated with respect to  $\pi_q(d\mathbf{q}) \pi_\xi(d\xi)$  (in fact,  $\kappa_p = \sqrt{\beta/m}$ ).
- The term  $\langle \mathcal{A}_\varepsilon A_\varepsilon \varphi, \varphi \rangle_{L^2(\pi)}$  is equal to  $\langle \mathcal{A}_\varepsilon A_\varepsilon \varphi, (1 - \Pi)\varphi \rangle_{L^2(\pi)}$  since  $\Pi \mathcal{A}_\varepsilon \Pi = 0$ , and is therefore larger than  $-\eta_\varepsilon \|(1 - \Pi)\varphi\|_{L^2(\pi)}^2$  in view of Lemma 2.6.
- We decompose the term  $-\langle A_\varepsilon \mathcal{A}_\varepsilon \varphi, \varphi \rangle_{L^2(\pi)}$  as  $-\langle A_\varepsilon \mathcal{A}_\varepsilon \Pi \varphi, \varphi \rangle_{L^2(\pi)} - \langle A_\varepsilon \mathcal{A}_\varepsilon (1 - \Pi)\varphi, \varphi \rangle_{L^2(\pi)}$ . We first observe that the operator  $A_\varepsilon \mathcal{A}_\varepsilon \Pi$  can be written, using spectral calculus, as

$$A_\varepsilon \mathcal{A}_\varepsilon \Pi = f_\varepsilon(\mathcal{T}), \quad \mathcal{T} = \Pi \left( \frac{2n}{(\beta\varepsilon)^2} \partial_\xi^* \partial_\xi + \frac{1}{\beta} \nabla_q^* \nabla_q \right) \Pi, \quad f_\varepsilon(x) = \frac{\eta_\varepsilon x}{\eta_\varepsilon^2 + x}.$$

Moreover, from Poincaré inequalities for  $\pi_q$  and  $\pi_\xi$  (with constants  $\kappa_q$  and  $\kappa_\xi = \sqrt{\beta}$ ),

$$\mathcal{T} \geq \alpha_\varepsilon \Pi(1 - \Pi_0), \quad \alpha_\varepsilon = \min \left( \frac{2n\kappa_\xi^2}{(\beta\varepsilon)^2}, \frac{\kappa_q^2}{\beta} \right),$$

so that

$$(2.19) \quad A_\varepsilon \mathcal{A}_\varepsilon \Pi \geq \Lambda_\varepsilon \Pi(1 - \Pi_0), \quad \Lambda_\varepsilon = \frac{\eta_\varepsilon \alpha_\varepsilon}{\eta_\varepsilon^2 + \alpha_\varepsilon}.$$

Note that  $\Lambda_\varepsilon$  is of order 1 when  $\varepsilon \leq 1$ , and of order  $\varepsilon^{-1}$  for  $\varepsilon \geq 1$ . It is precisely at this place that it is crucial to modify the definition of  $A_\varepsilon$ . Indeed, if one keeps the regularization operator  $-\left[1 - \Pi \mathcal{A}_\varepsilon^2 \Pi\right]^{-1} \Pi \mathcal{A}_\varepsilon$  as for  $\varepsilon \leq 1$ , the rate  $\Lambda_\varepsilon$  would be replaced by  $\alpha_\varepsilon/(1 + \alpha_\varepsilon)$ , which behaves as  $\alpha_\varepsilon \sim \varepsilon^{-2}$  for  $\varepsilon$  large. The quantity  $\langle A_\varepsilon \mathcal{A}_\varepsilon (1 - \Pi)\varphi, \varphi \rangle_{L^2(\pi)} = \langle A_\varepsilon \mathcal{A}_\varepsilon (1 - \Pi)\varphi, \Pi\varphi \rangle_{L^2(\pi)}$  can be shown to be larger than  $-C_1 \eta_\varepsilon / \varepsilon \|\Pi\varphi\|_{L^2(\pi)} \|(1 - \Pi)\varphi\|_{L^2(\pi)}$  upon proving that the operator  $A_\varepsilon \mathcal{A}_\varepsilon (1 - \Pi)$  is bounded by  $C_1 \eta_\varepsilon / \varepsilon$ ; see Lemma 2.9 below.

- Finally, in order to control  $\langle A_\varepsilon \mathcal{L}_O \varphi, \varphi \rangle_{L^2(\pi)} = \langle A_\varepsilon \mathcal{L}_O (1 - \Pi)\varphi, \Pi\varphi \rangle_{L^2(\pi)}$  by  $\|\Pi\varphi\|_{L^2(\pi)} \|(1 - \Pi)\varphi\|_{L^2(\pi)}$ , we prove in Lemma 2.8 that the operator  $A_\varepsilon \mathcal{L}_O$  is uniformly bounded with respect to  $\varepsilon$  by some constant  $C_2$ .

Gathering all estimates, we obtain, for  $\varphi \in L_0^2(\pi)$  (so that  $(1 - \Pi_0)\varphi = \varphi$ ),

$$(2.20) \quad \begin{aligned} \langle -\mathcal{L}_{\text{AdL}}\varphi, \varphi \rangle_{\varepsilon, \gamma} &\geq \left( \frac{\gamma \kappa_p^2}{\beta} - a_{\varepsilon, \gamma} \eta_\varepsilon \right) \|(1 - \Pi)\varphi\|_{L^2(\pi)}^2 + a_{\varepsilon, \gamma} \Lambda_\varepsilon \|\Pi\varphi\|_{L^2(\pi)}^2 \\ &\quad - a_{\varepsilon, \gamma} \left( C_1 \frac{\eta_\varepsilon}{\varepsilon} + \gamma C_2 \right) \|\Pi\varphi\|_{L^2(\pi)} \|(1 - \Pi)\varphi\|_{L^2(\pi)}, \end{aligned}$$

which can be rewritten as

$$(2.21) \quad \langle -\mathcal{L}_{\text{AdL}}\varphi, \varphi \rangle_{\varepsilon, \gamma} \geq X^T \mathbf{B}_{\varepsilon, \gamma} X, \quad X = \begin{pmatrix} \|\Pi\varphi\|_{L^2(\pi)} \\ \|(1 - \Pi)\varphi\|_{L^2(\pi)} \end{pmatrix}, \quad \mathbf{B}_{\varepsilon, \gamma} = \begin{pmatrix} B_{1,1} & \frac{1}{2}B_{1,2} \\ \frac{1}{2}B_{1,2} & B_{2,2} \end{pmatrix},$$

with

$$(2.22) \quad B_{1,1} = a_{\varepsilon, \gamma} \Lambda_\varepsilon, \quad B_{1,2} = -a_{\varepsilon, \gamma} \left( C_1 \frac{\eta_\varepsilon}{\varepsilon} + \gamma C_2 \right), \quad B_{2,2} = \frac{\gamma \kappa_p^2}{\beta} - a_{\varepsilon, \gamma} \eta_\varepsilon.$$

The result then follows from lower bounds on the smallest eigenvalue of  $\mathbf{B}_{\varepsilon, \gamma}$ , which reads

$$(2.23) \quad \lambda(\mathbf{B}_{\varepsilon, \gamma}) = \frac{4B_{1,1}B_{2,2} - B_{1,2}^2}{B_{1,1} + B_{2,2} + \sqrt{(B_{1,1} - B_{2,2})^2 + B_{1,2}^2}}.$$

The scaling of  $a_{\varepsilon, \gamma}$  as a function of  $\varepsilon, \gamma$  is obtained by requiring that the determinant

$$(2.24) \quad B_{1,1}B_{2,2} - \frac{B_{1,2}^2}{4} = \left( \frac{\gamma \kappa_p^2}{\beta} - a_{\varepsilon, \gamma} \right) a_{\varepsilon, \gamma} \Lambda_\varepsilon - \frac{a_{\varepsilon, \gamma}^2}{4} \left( C_1 \frac{\eta_\varepsilon}{\varepsilon} + \gamma C_2 \right)^2$$

is positive.

- For  $\varepsilon \geq 1$ , the factor  $\Lambda_\varepsilon$  is of order  $\varepsilon^{-1}$  and  $\eta_\varepsilon = \varepsilon^{-1}$ . The matrix  $\varepsilon \mathbf{B}_{\varepsilon, \gamma}$  then has the same form as for standard underdamped Langevin dynamics, upon replacing  $\gamma$  by  $\gamma\varepsilon$ . Indeed, the off-diagonal elements of  $\varepsilon \mathbf{B}_{\varepsilon, \gamma}$  are of order  $\max(1, \gamma\varepsilon)$  since  $\varepsilon \geq 1$ , while the diagonal terms are respectively of order  $\gamma\varepsilon$  and 1. This suggests choosing  $a_{\varepsilon, \gamma} = \bar{a} \min(\gamma\varepsilon, (\gamma\varepsilon)^{-1})$  with  $\bar{a} > 0$  sufficiently small, and leads to a lower bound for  $\lambda(\varepsilon \mathbf{B}_{\varepsilon, \gamma})$  of order  $\min(\gamma\varepsilon, (\gamma\varepsilon)^{-1})$ . Thus,  $\lambda(\mathbf{B}_{\varepsilon, \gamma})$  is of order  $\min(\gamma, (\gamma\varepsilon^2)^{-1})$ .
- For  $\varepsilon \leq 1$ ,  $\eta_\varepsilon = 1$  and the factor  $\Lambda_\varepsilon$  is of order 1. The scaling of  $a_{\varepsilon, \gamma}$  as a function of  $\varepsilon, \gamma$  suggested by (2.24) is

$$(2.25) \quad a_{\varepsilon, \gamma} = \bar{a} \frac{\gamma}{(\varepsilon^{-1} + \gamma)^2}$$

for  $\bar{a} > 0$  sufficiently small. We further distinguish two cases: (i) For  $\gamma\varepsilon \leq 1$ , the scaling (2.25) leads to the choice  $a_{\varepsilon, \gamma} = \bar{a}\gamma\varepsilon^2$  for  $\bar{a} > 0$  sufficiently small, in which case the smallest eigenvalue of  $\mathbf{B}_{\varepsilon, \gamma}$  is easily seen to be of order  $\gamma\varepsilon^2$  (since (2.23) is the ratio of a numerator of order  $\gamma^2\varepsilon^2$  and a denominator of order  $\gamma$ ). (ii) For  $\gamma\varepsilon \geq 1$ , the scaling (2.25) leads to the choice  $a_{\varepsilon, \gamma} = \bar{a} \min(\gamma, \gamma^{-1})$  for  $\bar{a} > 0$  sufficiently small, in which case the smallest eigenvalue of  $\mathbf{B}_{\varepsilon, \gamma}$  is easily seen to be of order  $\min(\gamma, \gamma^{-1})$ .

In conclusion, there exists  $\bar{\lambda} > 0$  such that the smallest eigenvalue of  $\mathbf{B}_{\varepsilon, \gamma}$  is lower bounded by  $\bar{\lambda} \min(\gamma, \gamma^{-1}, \gamma\varepsilon^2, (\gamma\varepsilon^2)^{-1})$ .

We conclude this section with the proofs of the two technical lemmas used above. In these proofs, we denote

$$(2.26) \quad \mathcal{G}_\varepsilon = \left( \eta_\varepsilon^2 + \Pi \left[ \frac{2n}{(\beta\varepsilon)^2} \partial_\xi^* \partial_\xi + \frac{1}{\beta} \nabla_q^* \nabla_q \right] \Pi \right)^{-1},$$

so that  $A_\varepsilon = -\eta_\varepsilon \mathcal{G}_\varepsilon \Pi \mathcal{A}_\varepsilon$ . We will repeatedly use in the proofs that  $\mathcal{G}_\varepsilon$ , when restricted to some subspace of  $L_0^2(\pi)$ , behaves as  $(1 + \Pi \partial_\xi^* \partial_\xi \Pi)^{-1}$  or  $(1 + \Pi \nabla_q^* \nabla_q \Pi)^{-1}$ . More precisely, we introduce the orthogonal projectors  $P_q$  and  $P_\xi$ , which correspond to a partial integration with respect to  $\pi_q(d\mathbf{q})$  and  $\pi_\xi(d\xi)$  (they are the counterparts for the variables  $\mathbf{q}, \xi$  of the projector  $\Pi$  defined in (2.14)):

$$(2.27) \quad (P_q \varphi)(\mathbf{p}, \xi) = \int_{\mathbb{R}^n} \varphi(\mathbf{q}, \mathbf{p}, \xi) \pi_q(d\mathbf{q}), \quad (P_\xi \varphi)(\mathbf{q}, \mathbf{p}) = \int_{\mathbb{R}} \varphi(\mathbf{q}, \mathbf{p}, \xi) \pi_\xi(d\xi).$$

Note that  $P_q, P_\xi$  both commute with  $\Pi, \nabla_q^* \nabla_q$  and  $\partial_\xi^* \partial_\xi$  (in fact  $P_q \nabla_q^* \nabla_q = \nabla_q^* \nabla_q P_q = 0$  and  $P_\xi \partial_\xi^* \partial_\xi = \partial_\xi^* \partial_\xi P_\xi = 0$ ) and therefore also with  $\mathcal{G}_\varepsilon$ , and that

$$(2.28) \quad \Pi P_q \mathcal{L}_H = 0, \quad \Pi P_\xi \mathcal{L}_{NH} = 0,$$

by the invariance of the measure  $\pi_q(d\mathbf{q})\pi_p(d\mathbf{p})$  by  $\mathcal{L}_H$ , and the invariance of  $\pi_p(d\mathbf{p})\pi_\xi(d\xi)$  by  $\mathcal{L}_{NH}$ . Moreover,  $\Pi \nabla_q^* \nabla_q \Pi \geq \kappa_q^2 \Pi(1 - P_q)$  from a Poincaré inequality for  $\pi_q$ ; and similarly,  $\Pi \partial_\xi^* \partial_\xi \Pi \geq \kappa_\xi^2 \Pi(1 - P_\xi)$  from a Gaussian Poincaré inequality for  $\pi_\xi$ . This leads to the following result.

**LEMMA 2.7.** *The operators  $\mathcal{G}_\varepsilon(1 + \Pi \nabla_q^* \nabla_q \Pi)(1 - P_q)$  and  $\varepsilon^{-2} \mathcal{G}_\varepsilon(1 + \Pi \partial_\xi^* \partial_\xi \Pi)(1 - P_\xi)$  are uniformly bounded with respect to  $\varepsilon$ . More precisely,*

$$(2.29) \quad \|\mathcal{G}_\varepsilon(1 + \Pi \nabla_q^* \nabla_q \Pi)(1 - P_q)\| \leq \beta(1 + \kappa_q^{-2}),$$

and

$$(2.30) \quad \|\mathcal{G}_\varepsilon(1 + \Pi \partial_\xi^* \partial_\xi \Pi)(1 - P_\xi)\| \leq \frac{\beta^2}{2n} (1 + \kappa_\xi^{-2}) \varepsilon^2.$$

Moreover,  $\mathcal{G}_\varepsilon^{1/2}(1 + \Pi \nabla_q^* \nabla_q \Pi)^{1/2}(1 - P_q)$  and  $\varepsilon^{-1} \mathcal{G}_\varepsilon^{1/2}(1 + \Pi \partial_\xi^* \partial_\xi \Pi)^{1/2}(1 - P_\xi)$  are also uniformly bounded with respect to  $\varepsilon$ .

*Proof.* Denoting  $A_q = (1 - P_q) \Pi \nabla_q^* \nabla_q \Pi (1 - P_q)$ ,

$$\begin{aligned} \mathcal{G}_\varepsilon(1 + \Pi \nabla_q^* \nabla_q \Pi)(1 - P_q) &= (1 - P_q) \mathcal{G}_\varepsilon(1 - P_q)(1 + \Pi \nabla_q^* \nabla_q \Pi)(1 - P_q) \\ &= (1 - P_q + A_q)^{1/2} \left[ \eta_\varepsilon^2 + 2n(\beta\varepsilon)^{-2} \Pi \partial_\xi^* \partial_\xi \Pi + \beta^{-1} \Pi \nabla_q^* \nabla_q \Pi \right]^{-1} (1 - P_q + A_q)^{1/2} \\ &= (1 - P_q + A_q)^{1/2} \left[ \eta_\varepsilon^2(1 - P_q) + \frac{2n}{(\beta\varepsilon)^2} (1 - P_q) \Pi \partial_\xi^* \partial_\xi \Pi (1 - P_q) + \beta^{-1} A_q \right]^{-1} (1 - P_q + A_q)^{1/2}, \end{aligned}$$

where all operators on the last right-hand side are considered on the subspace  $(1 - P_q)L_0^2(\pi)$ , on which  $A_q \geq \kappa_q^2$ . Therefore, in the sense of symmetric operators on  $(1 - P_q)L_0^2(\pi)$ ,

$$0 \leq \mathcal{G}_\varepsilon(1 + \Pi \nabla_q^* \nabla_q \Pi)(1 - P_q) \leq (1 + A_q)^{1/2} [\eta_\varepsilon^2 + \beta^{-1} A_q]^{-1} (1 + A_q)^{1/2} \leq g_\varepsilon(A_q),$$

with

$$g_\varepsilon(x) = \frac{1 + x}{\eta_\varepsilon^2 + \beta^{-1}x}.$$

This leads to (2.29) since  $g_\varepsilon(\kappa_q^2) \leq g_0(\kappa_q^2)$ . Similar computations lead to

$$\|\mathcal{G}_\varepsilon(1 + \Pi\partial_\xi^*\partial_\xi\Pi)(1 - P_\xi)\| \leq h_0(\kappa_\xi^2)\varepsilon^2, \quad h_\varepsilon(x) = \frac{1+x}{\min(1, \varepsilon^2) + 2n\beta^{-2}x},$$

which gives (2.30). The estimates on  $\mathcal{G}_\varepsilon^{1/2}(1 + \Pi\nabla_q^*\nabla_q\Pi)^{1/2}(1 - P_q)$  and  $\varepsilon^{-1}\mathcal{G}_\varepsilon^{1/2}(1 + \Pi\partial_\xi^*\partial_\xi\Pi)^{1/2}(1 - P_\xi)$  are obtained in a similar way.  $\square$

LEMMA 2.8. *The operator  $A_\varepsilon\mathcal{L}_O$  is uniformly bounded for  $\varepsilon > 0$ : There exists  $C_2 > 0$  such that  $\|A_\varepsilon\mathcal{L}_O\| \leq C_2$ .*

*Proof.* Since  $A_\varepsilon\mathcal{L}_O = -\eta_\varepsilon\mathcal{G}_\varepsilon\Pi\mathcal{L}_H\mathcal{L}_O - \eta_\varepsilon\varepsilon^{-1}\mathcal{G}_\varepsilon\Pi\mathcal{L}_{NH}\mathcal{L}_O$ , it suffices to prove that each operator in the right-hand side of this equality is uniformly bounded with respect to  $\varepsilon > 0$ . First, in view of (2.28), the operator

$$\mathcal{G}_\varepsilon\Pi\mathcal{L}_H\mathcal{L}_O = \mathcal{G}_\varepsilon\Pi(1 - P_q)\mathcal{L}_H\mathcal{L}_O = \mathcal{G}_\varepsilon(1 + \Pi\nabla_q^*\nabla_q\Pi)(1 - P_q)(1 + \Pi\nabla_q^*\nabla_q\Pi)^{-1}\Pi(1 - P_q)\mathcal{L}_H\mathcal{L}_O$$

is the product of the operator  $\mathcal{G}_\varepsilon(1 + \Pi\nabla_q^*\nabla_q\Pi)(1 - P_q)$  (uniformly bounded in  $\varepsilon$  from (2.29)) and the operator  $(1 + \Pi\nabla_q^*\nabla_q\Pi)^{-1}\Pi(1 - P_q)\mathcal{L}_H\mathcal{L}_O$ , which is bounded (see for instance [31, Proposition A.3]). We next consider

$$\eta_\varepsilon\varepsilon^{-1}\mathcal{G}_\varepsilon\Pi\mathcal{L}_{NH}\mathcal{L}_O = \eta_\varepsilon\varepsilon^{-1}\mathcal{G}_\varepsilon(1 + \Pi\partial_\xi^*\partial_\xi\Pi)(1 - P_\xi)(1 + \Pi\partial_\xi^*\partial_\xi\Pi)^{-1}\Pi(1 - P_\xi)\mathcal{L}_{NH}\mathcal{L}_O.$$

Note first that the norm of the operator  $\eta_\varepsilon\varepsilon^{-1}\mathcal{G}_\varepsilon(1 + \Pi\partial_\xi^*\partial_\xi\Pi)(1 - P_\xi)$  is of order  $\min(1, \varepsilon)$  by (2.30). It remains to prove that  $(1 + \Pi\partial_\xi^*\partial_\xi\Pi)^{-1}\Pi(1 - P_\xi)\mathcal{L}_{NH}\mathcal{L}_O$  is bounded. We note for this that

$$\begin{aligned} \Pi(1 - P_\xi)\mathcal{L}_{NH}\mathcal{L}_O &= -\frac{1}{\beta^3}(1 - P_\xi)\Pi((\partial_\xi - \partial_\xi^*)\nabla_p^*\nabla_p + \Delta_p^*\partial_\xi - \Delta_p\partial_\xi^*)\nabla_p^*\nabla_p \\ &= \frac{1}{\beta^3}(1 - P_\xi)\partial_\xi^*\Pi\Delta_p\nabla_p^*\nabla_p \end{aligned}$$

where we used (2.15). The conclusion then follows from the fact that  $\Pi\Delta_p\nabla_p^*\nabla_p$  is bounded (see Lemma 2.12 below) as well as  $\mathcal{T}_\xi = (1 + \Pi\partial_\xi^*\partial_\xi\Pi)^{-1}\Pi(1 - P_\xi)\partial_\xi^*$  (by computing  $\mathcal{T}_\xi\mathcal{T}_\xi^*$  and using spectral calculus together with the lower bound  $\partial_\xi^*\partial_\xi \geq \kappa_\xi^2(1 - P_\xi)$  on  $(1 - P_\xi)L_0^2(\pi)$ ).  $\square$

LEMMA 2.9. *There exists  $C_1 > 0$  such that, for any  $\varepsilon > 0$ ,  $\|A_\varepsilon\mathcal{A}_\varepsilon(1 - \Pi)\| \leq C_1\eta_\varepsilon/\varepsilon$ .*

*Proof.* We prove that the adjoint operator  $\eta_\varepsilon^{-1}(A_\varepsilon\mathcal{A}_\varepsilon(1 - \Pi))^* = \eta_\varepsilon^{-1}(1 - \Pi)\mathcal{A}_\varepsilon A_\varepsilon^* = (1 - \Pi)\mathcal{A}_\varepsilon^2\Pi\mathcal{G}_\varepsilon$  is uniformly bounded with respect to  $\varepsilon > 0$ . In fact, using  $\mathcal{L}_H\Pi P_q = 0$  and  $\mathcal{L}_{NH}\Pi P_\xi = 0$ ,

$$\begin{aligned} (2.31) \quad \eta_\varepsilon^{-1}\mathcal{A}_\varepsilon A_\varepsilon^* &= \mathcal{L}_H^2\Pi(1 - P_q)\mathcal{G}_\varepsilon + \frac{1}{\varepsilon}\mathcal{L}_{NH}\mathcal{L}_H\Pi(1 - P_q)\mathcal{G}_\varepsilon \\ &\quad + \frac{1}{\varepsilon}\mathcal{L}_H\mathcal{L}_{NH}\Pi(1 - P_\xi)\mathcal{G}_\varepsilon + \frac{1}{\varepsilon^2}\mathcal{L}_{NH}^2\Pi(1 - P_\xi)\mathcal{G}_\varepsilon. \end{aligned}$$

Let us consider successively the various terms on the right-hand side. First, in view of the rules (2.15),

$$\beta^2\mathcal{L}_H^2\Pi(1 - P_q)\mathcal{G}_\varepsilon = \sum_{i=1}^n \sum_{j=1}^n \left( \partial_{q_i}^2 \partial_{q_j} (1 - P_q)\mathcal{G}_\varepsilon \right) \left( \partial_{p_i}^* \partial_{p_j}^* \Pi \right) - \sum_{i=1}^n \sum_{j=1}^n \left( \partial_{q_i}^* \partial_{q_j} (1 - P_q)\mathcal{G}_\varepsilon \right) \left( \partial_{p_i} \partial_{p_j}^* \Pi \right),$$

which is a sum of bounded operators in view of Lemmas 2.11 and 2.12. Similarly,

$$\frac{1}{\varepsilon^2}\mathcal{L}_{NH}^2\Pi(1 - P_\xi)\mathcal{G}_\varepsilon = \frac{1}{\beta^4} \frac{1}{\varepsilon^2} [(\partial_\xi - \partial_\xi^*)\partial_\xi \nabla_p^* \nabla_p + \partial_\xi^2 \Delta_p^* - \partial_\xi^* \partial_\xi \Delta_p] \Delta_p^* \Pi(1 - P_\xi)\mathcal{G}_\varepsilon$$

is a sum of bounded operators in view of Lemmas 2.11 and 2.12. Consider now the terms involving both  $\mathcal{L}_H$  and  $\mathcal{L}_{NH}$ . We need to introduce projectors  $1 - P_q$  and  $1 - P_\xi$  in order to rely on Lemma 2.7. We note to this end that  $\mathcal{L}_{NH}\mathcal{L}_H\Pi(1 - P_q) = \mathcal{L}_{NH}(1 - P_\xi)\mathcal{L}_H\Pi(1 - P_q) + \mathcal{L}_{NH}P_\xi\mathcal{L}_H\Pi(1 - P_q)$  and  $\mathcal{L}_H\mathcal{L}_{NH}\Pi(1 - P_\xi) = \mathcal{L}_H(1 - P_q)\mathcal{L}_{NH}\Pi(1 - P_\xi) + \mathcal{L}_HP_q\mathcal{L}_{NH}\Pi(1 - P_\xi)$ . Straightforward computations show that

$$\mathcal{L}_{NH}P_\xi\mathcal{L}_H\Pi(1 - P_q)\varphi = -\xi\mathbf{p}^T\nabla_q(\Pi P_\xi(1 - P_q)\varphi),$$

which is the product of two functions depending on the variables  $\xi, \mathbf{p}$  and  $\mathbf{q}$ , respectively, with  $(\mathbf{p}, \xi) \mapsto \xi\mathbf{p}$  belonging to  $L^2(\pi_p \pi_\xi)$ . A similar reasoning shows that

$$\mathcal{L}_HP_q\mathcal{L}_{NH}\Pi(1 - P_\xi) = -2\mathbf{p}^T\nabla V\partial_\xi(\Pi P_q(1 - P_\xi)\varphi).$$

In addition,

$$\begin{aligned} & \frac{1}{\varepsilon}\mathcal{L}_{NH}(1 - P_\xi)\mathcal{L}_H\Pi(1 - P_q)\mathcal{G}_\varepsilon \\ &= \frac{1}{\beta^3} \left[ \frac{1}{\varepsilon}(\partial_\xi - \partial_\xi^*)(1 - P_\xi)\nabla_p^*\nabla_p + \frac{1}{\varepsilon}\partial_\xi(1 - P_\xi)\Delta_p^* - \frac{1}{\varepsilon}\partial_\xi^*(1 - P_\xi)\Delta_p \right] \nabla_p^*\nabla_q(1 - P_q)\mathcal{G}_\varepsilon, \end{aligned}$$

and

$$\frac{1}{\varepsilon}\mathcal{L}_H(1 - P_q)\mathcal{L}_{NH}\Pi(1 - P_\xi)\mathcal{G}_\varepsilon = \frac{1}{\beta^3} \frac{1}{\varepsilon}(\nabla_p^*\nabla_q - \nabla_q^*\nabla_p)(1 - P_q)\partial_\xi(1 - P_\xi)\mathcal{G}_\varepsilon\Delta_p^*\Pi$$

are sums of bounded operators in view of Lemma 2.11. Therefore,  $\varepsilon^{-1}\mathcal{L}_{NH}\mathcal{L}_H\Pi(1 - P_q)\mathcal{G}_\varepsilon$  and  $\varepsilon^{-1}\mathcal{L}_H\mathcal{L}_{NH}\Pi(1 - P_\xi)\mathcal{G}_\varepsilon$  both have operator norms of order  $1/\varepsilon$ . This finally gives the claimed result.  $\square$

*Remark 2.10.* Among the various terms in the decomposition of  $A_\varepsilon\mathcal{A}_\varepsilon(1 - \Pi)$  we consider in the proof of Lemma 2.9, the only ones which are not bounded as  $\varepsilon \rightarrow 0$  are  $\varepsilon^{-1}\mathcal{L}_{NH}P_\xi\mathcal{L}_H(1 - P_q)\Pi\mathcal{G}_\varepsilon$  and  $\varepsilon^{-1}\mathcal{L}_HP_q\mathcal{L}_{NH}(1 - P_\xi)\Pi\mathcal{G}_\varepsilon$ . These terms arise from the interaction between the Hamiltonian and Nosé–Hoover parts of the dynamics, and are responsible for the factor  $\eta_\varepsilon/\varepsilon$  in the expression of  $B_{1,2}$  in (2.22), which itself leads to the extra term  $\gamma\varepsilon^2$  in the scaling of the lower bound of Proposition 2.4.

Note that, crucially, operators in the  $\xi$  variable in the computations of the proof of Lemma 2.9 always appear with a prefactor  $\varepsilon^{-1}$ . The fact that this is the correct scaling for the boundedness of these operators comes from the following result.

**LEMMA 2.11.** *The operators  $\partial_{q_i, q_j}^2(1 - P_q)\mathcal{G}_\varepsilon$ ,  $\partial_{q_i}^*\partial_{q_j}(1 - P_q)\mathcal{G}_\varepsilon$ ,  $\varepsilon^{-1}\partial_\xi^*(1 - P_\xi)\partial_{q_i}(1 - P_q)\mathcal{G}_\varepsilon$ ,  $\varepsilon^{-1}\partial_{q_i}(1 - P_q)\partial_\xi(1 - P_\xi)\mathcal{G}_\varepsilon$ ,  $\varepsilon^{-1}\partial_{q_i}^*(1 - P_q)\partial_\xi(1 - P_\xi)\mathcal{G}_\varepsilon$ ,  $\varepsilon^{-2}\partial_\xi^2(1 - P_\xi)\mathcal{G}_\varepsilon$ ,  $\varepsilon^{-2}\partial_\xi^*\partial_\xi(1 - P_\xi)\mathcal{G}_\varepsilon$  are uniformly bounded with respect to  $\varepsilon > 0$ .*

*Proof.* Consider for instance  $\partial_{q_i, q_j}^2(1 - P_q)\mathcal{G}_\varepsilon$ . It is sufficient by Lemma 2.7 to prove that  $\partial_{q_i, q_j}^2(1 - P_q)(1 + \Pi\nabla_q^*\nabla_q\Pi)^{-1}$  is bounded, and in fact that operators of the form  $\mathcal{T}_i = \partial_{q_i}(1 - P_q)(1 + \Pi\partial_{q_i}^*\partial_{q_i}\Pi)^{-1/2}$  and  $\partial_{q_i}^2(1 - P_q)(1 + \Pi\nabla_q^*\nabla_q\Pi)^{-1}$  are bounded. The first statement is clear by calculating  $\mathcal{T}_i^*\mathcal{T}_i$  and using spectral calculus; for the second one we use [9, section 3]. Similar reasonings can be used to bound  $\partial_{q_i}^*\partial_{q_j}(1 - P_q)\mathcal{G}_\varepsilon$ . Bounds on  $\varepsilon^{-2}\partial_\xi^2(1 - P_\xi)\mathcal{G}_\varepsilon$ ,  $\varepsilon^{-2}\partial_\xi^*\partial_\xi(1 - P_\xi)\mathcal{G}_\varepsilon$  are obtained in a similar way, considering the specific case of quadratic potentials in  $\xi$  (so that the estimates similar to those of [9, section 3] hold in the  $\xi$  variable).

Consider next  $\varepsilon^{-1}\partial_\xi(1 - P_\xi)\partial_{q_i}^*(1 - P_q)\mathcal{G}_\varepsilon = \mathcal{T}_\xi R_{q_i} S_{q, \xi}$  with

$$S_{q, \xi} = \varepsilon^{-1}(1 - P_\xi)(1 + \Pi\partial_\xi^*\partial_\xi\Pi)^{1/2}(1 - P_q)(1 + \Pi\nabla_q^*\nabla_q\Pi)^{1/2}\mathcal{G}_\varepsilon$$

uniformly bounded in  $\varepsilon$  by Lemma 2.7,  $\mathcal{T}_\xi = \partial_\xi(1 - P_\xi)(1 + \Pi\partial_\xi^*\partial_\xi\Pi)^{-1/2}$  bounded by considering  $\mathcal{T}_\xi^*\mathcal{T}_\xi$  and resorting to spectral calculus, and  $R_{q_i} = \partial_{q_i}^*(1 - P_q)(1 + \Pi\nabla_q^*\nabla_q\Pi)^{-1/2}$ . To prove that the latter operator is bounded, we write it as the sum of  $-\partial_{q_i}(1 - P_q)(1 + \Pi\nabla_q^*\nabla_q\Pi)^{-1/2}$  (which is bounded by the same reasoning as the one used to prove that  $\mathcal{T}_\xi$  is bounded) and  $\beta\partial_{q_i}V(1 - P_q)(1 + \Pi\nabla_q^*\nabla_q\Pi)^{-1/2}$ , which is bounded in view of the inequality

$$\|\nabla V|h\|_{L^2(\pi_q)} \leq C(\|h\|_{L^2(\pi_q)} + \|\nabla h\|_{L^2(\pi_q)})$$

provided by [36, Lemma A.24]. The boundedness of  $\varepsilon^{-1}\partial_{q_i}(1 - P_q)\partial_\xi(1 - P_\xi)\mathcal{G}_\varepsilon$  and  $\varepsilon^{-1}\partial_{q_i}(1 - P_q)\partial_\xi^*(1 - P_\xi)\mathcal{G}_\varepsilon$  follows by similar arguments.  $\square$

The proof of the following lemma is obtained by straightforward computations based on integration by parts in the integral involved in the definition of  $\Pi$ .

LEMMA 2.12. *For any  $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{N}$  and  $i, j, k \in \{1, \dots, n\}$ , the operators  $\Pi\partial_{p_i}^{\alpha_1}(\partial_{p_j}^*)^{\alpha_2}\partial_{p_k}^{\alpha_3}$  are bounded (and so are their adjoints on  $L^2(\pi_p)$  and  $L^2(\pi)$ ). In particular,  $\partial_{p_i}^*\partial_{p_j}^*\Pi$  and  $\partial_{p_i}\partial_{p_j}^*\Pi$  are bounded.*

**3. Pathwise ergodicity and functional central limit theorem.** Consider, for  $\varphi \in L^1(\pi)$  given, the trajectory average of  $\varphi$  evaluated along a realization of the solution of the SDE (2.4):

$$(3.1) \quad \widehat{\varphi}_t := \frac{1}{t} \int_0^t \varphi(\mathbf{q}_s, \mathbf{p}_s, \xi_s) ds.$$

The almost-sure convergence of this estimator to  $\mathbb{E}_\pi(\varphi)$  holds by the results of [17] since the dynamics admits an invariant probability measure with a positive density, and the generator is hypoelliptic [14]. The latter property follows from the following computations on commutators:  $[\mathcal{L}_H, \partial_{p_i}] = -\partial_{q_i}$ , and  $[\mathcal{L}_{NH}, \partial_{p_i}] = -2p_i\partial_\xi + \xi\partial_{p_i}$  so that  $[[\mathcal{L}_{NH}, \partial_{p_i}], \partial_{p_i}] = 2\partial_\xi$ .

In fact, according to the results of [2], a natural CLT is a consequence of the boundedness of the inverse of the generator obtained in Corollary 2.3.

COROLLARY 3.1 (CLT for AdL). *Consider  $\varphi \in L^2(\pi)$ . Then*

$$(3.2) \quad \sqrt{t}(\widehat{\varphi}_t - \mathbb{E}_\pi\varphi) \xrightarrow[t \rightarrow +\infty]{\text{law}} \mathcal{N}(0, \sigma_{\varepsilon, \gamma}^2(\varphi)),$$

where the asymptotic variance reads

$$\sigma_{\varepsilon, \gamma}^2(\varphi) = 2 \int_{\mathbb{R}^{2n+1}} (-\mathcal{L}_{\text{AdL}}^{-1} \Pi_0 \varphi) \Pi_0 \varphi d\pi.$$

Corollary 2.3 provides the following bounds on the asymptotic variance:

$$(3.3) \quad 0 \leq \sigma_{\varepsilon, \gamma}^2(\varphi) \leq \frac{2C\|\varphi\|_{L^2(\pi)}^2}{\lambda} \max\left(\gamma, \frac{1}{\gamma}, \gamma\varepsilon^2, \frac{1}{\gamma\varepsilon^2}\right).$$

This inequality shows that integration times of order  $t = \tau \max(\gamma, \gamma^{-1}, \gamma\varepsilon^2, (\gamma\varepsilon^2)^{-1})$  should be considered in order for the estimator (3.1) to have a variance of order  $1/\tau$ .



**3.1. Langevin limit  $\varepsilon \rightarrow +\infty$ .** We consider in this section the convergence of the asymptotic variance in the limit when  $\varepsilon \rightarrow +\infty$ , which should be thought of as being somewhat similar to overdamped limits of Langevin dynamics. We do not consider the regime  $\varepsilon \rightarrow 0$  which is a mathematically a singular limit (see however Remark 3.3 below), and is also not a regime which is numerically convenient because of the stiffness of the resulting dynamics, which typically calls for integration schemes with timesteps of order  $\varepsilon$  (or the construction of dedicated numerical schemes based on averaging ideas for instance).

In the limit  $\varepsilon \rightarrow +\infty$ , for a given test function  $\varphi \in C_0^\infty(\mathbb{R}^{2n+1})$ , the function  $\mathcal{L}_{\text{AdL}}\varphi$  converges to  $\mathcal{L}_{\text{Lang}}\varphi$  where  $\mathcal{L}_{\text{Lang}} = \mathcal{L}_H + \gamma\mathcal{L}_O$  is the generator of the standard underdamped Langevin dynamics. To understand the behavior of the limiting asymptotic variance, we restrict ourselves to functions of  $(\mathbf{q}, \mathbf{p})$  only, since the variable  $\xi$  evolves very slowly and should therefore not be of interest. Since the slow convergence to equilibrium is due to the relaxation of the  $\xi$  variable in the regime  $\varepsilon \rightarrow +\infty$ , we expect that restricting the attention to such observables allows the variance to remain bounded. In fact, the following result holds (see section 3.2 for the proof).

**PROPOSITION 3.2.** *Fix  $\gamma > 0$ . Assume that  $U$  satisfies Assumption 1, is semi-convex (there exist a bounded smooth function  $U_1$  with bounded derivatives and a smooth convex function  $U_2$  such that  $U = U_1 + U_2$ ), grows at most polynomially at infinity and its derivatives as well, and that there exist  $K > 0$ ,  $R \in \mathbb{R}$ , and  $a \in (0, 1)$  such that*

$$\frac{1}{2}\mathbf{q}^T \nabla U(\mathbf{q}) \geq aU(\mathbf{q}) + \gamma^2 \frac{a(2-a)}{8(1-a)} |\mathbf{q}|^2 - K, \quad U(\mathbf{q}) \geq R|\mathbf{q}|^2.$$

*Consider a smooth function  $\varphi = \varphi(\mathbf{q}, \mathbf{p})$  growing at most polynomially in  $(\mathbf{q}, \mathbf{p})$  and whose derivatives grow at most polynomially. Then there exists  $C > 0$  (depending on  $\gamma$ ,  $\varphi$ ) such that the asymptotic variance  $\sigma_{\varepsilon, \gamma}^2(\varphi)$  defined in Corollary 3.1 satisfies*

$$\forall \varepsilon \geq 1, \quad |\sigma_{\varepsilon, \gamma}^2(\varphi) - \sigma_{\infty, \gamma}^2(\varphi)| \leq \frac{C}{\varepsilon},$$

*where  $\sigma_{\infty, \gamma}^2(\varphi)$  involves only asymptotic variances of underdamped Langevin dynamics. More precisely,*

$$\sigma_{\infty, \gamma}^2(\varphi) = \frac{2}{\beta} \left( \gamma \|\nabla_p \Phi_0\|_{L^2(\pi_q \pi_p)}^2 - \gamma \frac{\langle \nabla_p \Phi_{-1}, \nabla_p \Phi_0 \rangle_{L^2(\pi_q \pi_p)}^2}{\|\nabla_p \Phi_{-1}\|_{L^2(\pi_q \pi_p)}^2} + \frac{\beta^2 \langle \Phi_{-1}, \mathcal{L}_H \Phi_0 \rangle_{L^2(\pi_q \pi_p)}^2}{\gamma \|\nabla_p \Phi_{-1}\|_{L^2(\pi_q \pi_p)}^2} \right)$$

*where  $\Phi_0 = -\mathcal{L}_{\text{Lang}}^{-1} \Pi_0 \varphi$  and  $\Phi_{-1} = -\mathcal{L}_{\text{Lang}}^{-1} \left( \mathbf{p}^2 - \frac{n}{\beta} \right)$ .*

Note that the first term on the right-hand side of the expression of  $\sigma_{\infty, \gamma}^2(\varphi)$  corresponds to the asymptotic variance of a standard underdamped Langevin dynamics. The Nosé–Hoover-like thermostat adds two terms in the large  $\varepsilon$  limit, one nonpositive and one nonnegative, so that it is not clear in general whether  $\sigma_{\infty, \gamma}^2(\varphi)$  is larger than  $2\gamma\beta^{-1} \|\nabla_p \Phi_0\|_{L^2(\pi_q \pi_p)}^2$ . Overall, it however still holds that  $\sigma_{\infty, \gamma}^2(\varphi) \geq 0$  as expected since a Cauchy–Schwarz inequality shows that the sum of the two first terms in the brackets on the right-hand side is indeed nonnegative.

The extra conditions on the potential, taken from [18], are satisfied for potentials growing at infinity as  $|\mathbf{q}|^\alpha$  with  $\alpha > 2$ . They ensure that  $\mathcal{L}_{\text{Lang}}^{-1}$  stabilizes the vector space of smooth functions of  $(\mathbf{q}, \mathbf{p})$  with mean zero with respect to  $\pi_q \pi_p$ , growing at most polynomially at infinity, and whose derivatives grow at most polynomially at infinity.

It is in fact possible to write an expansion in inverse powers of  $\varepsilon$  for the difference  $\sigma_{\varepsilon,\gamma}^2(\varphi) - \sigma_{\infty,\gamma}^2(\varphi)$ , and in particular to make precise the leading order term in this expansion. We however refrain from doing so because the expressions are cumbersome. Note also that the proof of Proposition 3.2 allows us to write the action of  $\mathcal{L}_{\text{AdL}}^{-1}$  on  $L_0^2(\pi)$  at leading order  $\varepsilon^{-2}$  (in a similar fashion to the results presented in [21, Theorem 2.5], which provides an expansion of the resolvent of the generator of the underdamped Langevin dynamics in inverse powers of  $\gamma$ ); see Remark 3.4.

*Remark 3.3.* In the limit  $\varepsilon \rightarrow 0$ , the dynamics (2.4) is described by the ODEs:  $d\mathbf{p} = -\frac{\xi}{\varepsilon}\mathbf{p}dt$ ;  $d\xi = \frac{1}{\varepsilon}(|\mathbf{p}|^2 - \frac{n}{\beta})dt$  (with  $q$  constant). The only equilibrium points correspond to  $|\mathbf{p}|^2 = n\beta^{-1}$  and  $\xi = 0$ . A simple computation shows that  $\Phi(q, \mathbf{p}, \xi) = \xi^2 + |\mathbf{p}|^2 - \frac{2n}{\beta} \ln |\mathbf{p}|^2$  is an invariant of the dynamics. It is therefore expected that (2.4) corresponds to a fast averaging on the level sets of  $\Phi$ , with a superimposed slow variation of the values of  $\Phi$  induced by the Langevin part of the dynamics. Since the dynamics is at leading order a dynamics on the two one-dimensional variables  $P = |\mathbf{p}|^2$  and  $\xi$  only, it may be possible to adapt the techniques from [29] in order to determine the dominant behavior of the asymptotic variance in the regime  $\varepsilon \rightarrow 0$ .

**3.2. Proof of Proposition 3.2.** The idea of the proof is to construct an approximate solution  $\psi_\varepsilon$  to the Poisson equation  $-\mathcal{L}_{\text{AdL}}\phi_\varepsilon = \Pi_0\varphi$ , using asymptotic analysis. The scaling of the resolvent  $-\mathcal{L}_{\text{AdL}}$  as given by Corollary 2.3 suggests that, in the limit  $\varepsilon \rightarrow +\infty$ ,

$$(3.4) \quad \psi_\varepsilon = \varepsilon^2\Psi_{-2} + \varepsilon\Psi_{-1} + \Psi_0 + \varepsilon^{-1}\Psi_1 + \dots$$

The various functions in (3.4) formally satisfy, by identifying powers of  $\varepsilon$ ,

$$\begin{aligned} -\mathcal{L}_{\text{Lang}}\Psi_{-2} &= 0, & -\mathcal{L}_{\text{Lang}}\Psi_{-1} &= \mathcal{L}_{\text{NH}}\Psi_{-2}, & -\mathcal{L}_{\text{Lang}}\Psi_0 &= \Pi_0\varphi + \mathcal{L}_{\text{NH}}\Psi_{-1}, \\ -\mathcal{L}_{\text{Lang}}\Psi_i &= \mathcal{L}_{\text{NH}}\Psi_{i-1} & \text{for } i &\geq 1. \end{aligned}$$

The strategy of the proof is to construct the leading order terms  $\Psi_{-2}, \Psi_{-1}, \dots, \Psi_2 \in L_0^2(\pi)$  in order to obtain some approximate solution  $\psi_\varepsilon$  (obtained by a truncation of (3.4)), and then to use resolvent estimates to conclude that  $\phi_\varepsilon - \psi_\varepsilon$  is small.

We will repeatedly use the fact that the unique solution  $G$  of  $-\mathcal{L}_{\text{Lang}}G = g$  for  $g$  a smooth function with average 0 with respect to  $\pi_p(d\mathbf{p})\pi_q(d\mathbf{q})$  growing at most polynomially at infinity and whose derivatives also grow at most polynomially at infinity is a well-defined smooth function, which grows at most polynomially at infinity and whose derivatives also grow at most polynomially at infinity (by the results of [18]).

*Construction of the leading order terms in the expansion.* The equation  $-\mathcal{L}_{\text{Lang}}\Psi_{-2} = 0$  shows that  $\Psi_{-2}$  only depends on  $\xi$ , i.e.,  $\Psi_{-2}(\mathbf{q}, \mathbf{p}, \xi) = f_{-2}(\xi)$  for some function  $f_{-2}$ . Next,  $-\mathcal{L}_{\text{Lang}}\Psi_{-1} = \mathcal{L}_{\text{NH}}\Psi_{-2} = (\mathbf{p}^2 - n\beta^{-1})f'_{-2}(\xi)$ , so that

$$\Psi_{-1}(\mathbf{q}, \mathbf{p}, \xi) = f'_{-2}(\xi)\Phi_{-1}(\mathbf{q}, \mathbf{p}) + f_{-1}(\xi), \quad \Phi_{-1}(\mathbf{q}, \mathbf{p}) = -\mathcal{L}_{\text{Lang}}^{-1}\left(\mathbf{p}^2 - \frac{n}{\beta}\right).$$

The equation for  $\Psi_0$  then becomes

$$-\mathcal{L}_{\text{Lang}}\Psi_0 = \Pi_0\varphi + f''_{-2}(\xi)\left(\mathbf{p}^2 - \frac{n}{\beta}\right)\Phi_{-1} + f'_{-1}(\xi)\left(\mathbf{p}^2 - \frac{n}{\beta}\right) - \xi f'_{-2}(\xi)\mathbf{p}^T \nabla_p \Phi_{-1}.$$

The solvability condition for this equation is that the right-hand side has average 0 with respect to the probability measure  $\pi_q(d\mathbf{q})\pi_p(d\mathbf{p})$ . Integration by parts shows

that, for any test function  $\phi$ ,

$$\int_{\mathbb{R}^n} \mathbf{p}^T \nabla_p \phi \, d\pi_p = \beta \int_{\mathbb{R}^n} \phi \left( \mathbf{p}^2 - \frac{n}{\beta} \right) d\pi_p,$$

so that the solvability condition is  
(3.5)

$$a \mathcal{L}_{\text{eff}, \xi} f_{-2} = - \int_{\mathbb{R}^{2n}} \Pi_0 \varphi \, d\pi_p \, d\pi_q = 0, \quad a = \int_{\mathbb{R}^{2n}} \left( \mathbf{p}^2 - \frac{n}{\beta} \right) \Phi_{-1} \, d\pi_p \, d\pi_q \geq 0,$$

where  $\mathcal{L}_{\text{eff}, \xi}$  is the generator of an effective Ornstein–Uhlenbeck process acting on functions  $u = u(\xi)$  as  $\mathcal{L}_{\text{eff}, \xi} u = u'' - \beta \xi u'$ . In fact  $a > 0$  since  $a = \gamma \beta^{-1} \|\nabla_p \Phi_1\|^2 = 0$  would imply that  $\Phi_{-1}$  is constant in  $\mathbf{p}$ , which is in contradiction to the definition of  $\Phi_{-1}$  because

$$(-\mathcal{L}_{\text{Lang}} \Phi_{-1})(\mathbf{q}, \mathbf{p}) = \mathbf{p}^T \cdot \nabla_q \Phi_{-1}(\mathbf{q}) \neq \left( \mathbf{p}^2 - \frac{n}{\beta} \right).$$

The fact that  $a$  is nonzero implies that the first equality in (3.5) holds if and only if  $f_{-2} = 0$ , so that  $\Psi_{-2} = 0$  and  $\Psi_{-1} = f_{-1}$ . Moreover,

$$\Psi_0(\mathbf{q}, \mathbf{p}, \xi) = \Phi_0(\mathbf{q}, \mathbf{p}) + f'_{-1}(\xi) \Phi_{-1}(\mathbf{q}, \mathbf{p}) + f_0(\xi), \quad \Phi_0 = -\mathcal{L}_{\text{Lang}}^{-1} \Pi_0 \varphi.$$

*Remark 3.4.* The equality (3.5) shows that the action of leading order of the resolvent for Adaptive Langevin for functions  $\varphi \in L_0^2(\pi)$  is  $a^{-1} \varepsilon^2 \mathcal{L}_{\text{eff}, \xi}^{-1} \Pi P_q \varphi$  (with  $P_q$  defined in (2.27)).

The condition at the next order is

$$-\mathcal{L}_{\text{Lang}} \Psi_1 = \mathcal{L}_{\text{NH}} \Psi_0 = -\xi \mathbf{p}^T \nabla_p \Phi_0 - \xi f'_{-1} \mathbf{p}^T \nabla_p \Phi_{-1} + \left( \mathbf{p}^2 - \frac{n}{\beta} \right) [f'_0 + f''_{-1} \Phi_{-1}].$$

The solvability condition reads  $a \mathcal{L}_{\text{eff}, \xi} f_{-1} = \xi b_0$  with  $b_0 = \Pi P_q (\mathbf{p}^T \nabla_p \Phi_0)$ , so that  $f_{-1}(\xi) = -\xi b_0 / (a\beta)$ , and

$$\Psi_1(\mathbf{q}, \mathbf{p}, \xi) = f'_0(\xi) \Phi_{-1}(\mathbf{q}, \mathbf{p}) + \xi \Phi_1(\mathbf{q}, \mathbf{p}) + f_1(\xi), \quad \Phi_1 = -\mathcal{L}_{\text{Lang}}^{-1} \left( \frac{b_0}{a\beta} \mathbf{p}^T \nabla_p \Phi_{-1} - \mathbf{p}^T \nabla_p \Phi_0 \right).$$

Next,

$$-\mathcal{L}_{\text{Lang}} \Psi_2 = \mathcal{L}_{\text{NH}} \Psi_1 = -\xi f'_0 \mathbf{p}^T \nabla_p \Phi_{-1} - \xi^2 \mathbf{p}^T \nabla_p \Phi_1 + \left( \mathbf{p}^2 - \frac{n}{\beta} \right) [\Phi_1 + f'_1 + f''_0 \Phi_{-1}],$$

for which the solvability condition reads  $a \mathcal{L}_{\text{eff}, \xi} f_0 = (\xi^2 - \beta^{-1}) b_1$  with  $b_1 = \Pi P_q (\mathbf{p}^T \nabla_p \Phi_1)$ . Therefore,  $f_0(\xi) = (\beta^{-1} - \xi^2) b_1 / (2\beta a)$ , so that

$$\begin{aligned} \Psi_2(\mathbf{q}, \mathbf{p}, \xi) &= \mathcal{L}_{\text{Lang}}^{-1} \left[ \left( \mathbf{p}^2 - \frac{n}{\beta} \right) \left( \frac{b_1}{\beta a} \Phi_{-1} - \Phi_1 \right) \right] + \xi^2 \mathcal{L}_{\text{Lang}}^{-1} \left( \mathbf{p}^T \nabla_p \Phi_1 - \frac{b_1}{\beta a} \mathbf{p}^T \nabla_p \Phi_{-1} \right) \\ &\quad + f'_1(\xi) \Phi_{-1}(\mathbf{q}, \mathbf{p}) + f_2(\xi). \end{aligned}$$

*Obtaining bounds on the difference of the variances.* We now choose  $f_1 = f_2 = 0$  and compute

$$\mathcal{L}_{\text{AdL}} \left( \varepsilon \Psi_{-1} + \Psi_0 + \frac{1}{\varepsilon} \Psi_1 + \frac{1}{\varepsilon^2} \Psi_2 - \phi_\varepsilon \right) = \frac{1}{\varepsilon^3} \mathcal{L}_{\text{NH}} \Psi_2.$$

We deduce, in view of Corollary 2.3, that there exists a constant  $C_\gamma \in \mathbb{R}_+$  such that, for any  $\varepsilon \geq 1$ ,

$$\left\| \varepsilon \Psi_{-1} + \Psi_0 + \frac{1}{\varepsilon} \Psi_1 + \frac{1}{\varepsilon^2} \Psi_2 - \phi_\varepsilon \right\|_{L^2(\pi)} \leq \frac{C_\gamma}{\varepsilon} \|\mathcal{L}_{\text{NH}} \Psi_2\|_{L^2(\pi)},$$

and in fact  $\|\varepsilon \Psi_{-1} + \Psi_0 - \phi_\varepsilon\|_{L^2(\pi)} \leq R_{\gamma, \varphi} / \varepsilon$ , for some constant  $R_{\gamma, \varphi} \in \mathbb{R}_+$ . The asymptotic variance  $\sigma_{\varepsilon, \gamma}^2(\varphi)$  then coincides up to an error of order  $\varepsilon^{-1}$  with

$$\tilde{\sigma}_{\varepsilon, \gamma}^2(\varphi) = 2 \int_{\mathbb{R}^{2n+1}} (\varepsilon \Psi_{-1} + \Psi_0) \Pi_0 \varphi \, d\pi = 2 \int_{\mathbb{R}^{2n+1}} \left( \Phi_0 - \frac{b_0}{a\beta} \Phi_{-1} \right) \Pi_0 \varphi \, d\pi,$$

where we used for the second equality the fact that the average with respect to  $\pi$  of the product of a function of  $\xi$  and  $\Pi_0 \varphi$  vanishes. Finally, by integrating in  $\xi$  and expressing  $a, b_0$  in terms of the generator of the Langevin dynamics, namely,

$$\begin{aligned} a &= - \int_{\mathbb{R}^{2n}} (\mathcal{L}_{\text{Lang}} \Phi_{-1}) \Phi_{-1} \, d\pi_q \, d\pi_p = \frac{\gamma}{\beta} \|\nabla_p \Phi_{-1}\|_{L^2(\pi_q \pi_p)}^2, \\ b_0 &= \beta \int_{\mathbb{R}^{2n}} \left( \mathbf{p}^2 - \frac{n}{\beta} \right) \Phi_0 \, d\pi_q \, d\pi_p = -\beta \int_{\mathbb{R}^{2n}} (\mathcal{L}_{\text{Lang}} \Phi_{-1}) \Phi_0 \, d\pi_q \, d\pi_p, \end{aligned}$$

it follows that

$$\tilde{\sigma}_{\varepsilon, \gamma}^2(\varphi) = 2 \left( \int_{\mathbb{R}^{2n}} \Phi_0 \Pi_0 \varphi \, d\pi_q \, d\pi_p - \frac{\beta \langle \mathcal{L}_{\text{Lang}} \Phi_{-1}, \Phi_0 \rangle_{L^2(\pi_q \pi_p)} \langle \mathcal{L}_{\text{Lang}} \Phi_0, \Phi_{-1} \rangle_{L^2(\pi_q \pi_p)}}{\gamma \|\nabla_p \Phi_{-1}\|_{L^2(\pi_q \pi_p)}^2} \right).$$

Now,

$$\begin{aligned} \langle \mathcal{L}_{\text{Lang}} \Phi_{-1}, \Phi_0 \rangle_{L^2(\pi_q \pi_p)} &= -\frac{\gamma}{\beta} \langle \nabla_p \Phi_{-1}, \nabla_p \Phi_0 \rangle_{L^2(\pi_q \pi_p)} - \langle \Phi_{-1}, \mathcal{L}_H \Phi_0 \rangle_{L^2(\pi_q \pi_p)}, \\ \langle \mathcal{L}_{\text{Lang}} \Phi_0, \Phi_{-1} \rangle_{L^2(\pi_q \pi_p)} &= -\frac{\gamma}{\beta} \langle \nabla_p \Phi_{-1}, \nabla_p \Phi_0 \rangle_{L^2(\pi_q \pi_p)} + \langle \Phi_{-1}, \mathcal{L}_H \Phi_0 \rangle_{L^2(\pi_q \pi_p)}, \end{aligned}$$

so that

$$\tilde{\sigma}_{\varepsilon, \gamma}^2(\varphi) = \frac{2}{\beta} \left( \gamma \|\nabla_p \Phi_0\|_{L^2(\pi_q \pi_p)}^2 - \gamma \frac{\langle \nabla_p \Phi_{-1}, \nabla_p \Phi_0 \rangle_{L^2(\pi_q \pi_p)}^2}{\|\nabla_p \Phi_{-1}\|_{L^2(\pi_q \pi_p)}^2} + \frac{\beta^2 \langle \Phi_{-1}, \mathcal{L}_H \Phi_0 \rangle_{L^2(\pi_q \pi_p)}^2}{\gamma \|\nabla_p \Phi_{-1}\|_{L^2(\pi_q \pi_p)}^2} \right),$$

which gives the claimed result.

**4. Numerical results.** In this section, we present the results of several numerical experiments. First, we consider a simple illustration to demonstrate the scaling of the spectral gap as a function of  $\gamma$  and  $\varepsilon$  as predicted in section 2. Second, we demonstrate the scaling of the asymptotic variance, as predicted in section 3. We also verify the existence of an asymptotic CLT for the case of a Bayesian data analysis problem.

**4.1. Spectral gap in Galerkin subspace.** Let  $U : \mathbb{R} \rightarrow \mathbb{R}$ ,  $U(\mathbf{q}) = \frac{1}{2} \mathbf{q}^2$ . Moreover, denote by  $h_l$  the  $l$ th Hermite Polynomial, i.e.,

$$(4.1) \quad h_l(x) = \frac{1}{\sqrt{l!}} \tilde{H}_l(\sqrt{\beta}x), \quad \tilde{H}_l(x) = (-1)^l e^{x^2/2} \frac{d^l}{dx^l} (e^{-x^2/2}),$$

and consider for prescribed integers  $L \in \mathbb{N}$  the finite dimensional Galerkin subspace  $\mathcal{G}_L$  spanned by polynomials of the form

$$(4.2) \quad \psi_{k,l,m}(\mathbf{p}, \xi, \mathbf{q}) = h_k(\mathbf{p})h_l(\xi)h_m(\mathbf{q}), \quad 0 \leq l, k, m \leq L-1,$$

and the associated projection operator

$$(4.3) \quad \Pi_{\text{Galerkin}}^L : L^2(\pi) \rightarrow \mathcal{G}^L, \quad \varphi \mapsto \sum_{k=0}^{L-1} \sum_{l=0}^{L-1} \sum_{m=0}^{L-1} u_{k,l,m} \psi_{k,l,m},$$

where  $u_{k,l,m} := \langle \varphi, \psi_{k,l,m} \rangle_{L^2(\pi)}$ . In order to simplify notation we consider a linear indexing of the coefficients  $u_{k,l,m}$  and the polynomials  $\psi_{k,l,m}$  using a hash map of the form  $I : (k, l, m) \mapsto 1 + m + Lk + L^2l$  so that we can write the action of the Galerkin operator on functions  $\varphi \in L_0^2(\pi)$  in the compact form  $\Pi_{\text{Galerkin}}^L \varphi = \mathbf{u} \cdot \boldsymbol{\psi}$ , where  $\mathbf{u} = [\tilde{u}_i]_{1 \leq i \leq L^3}$  and  $\boldsymbol{\psi} = [\tilde{\psi}_i]_{1 \leq i \leq L^3}$ , where  $\tilde{u}$  and  $\tilde{\psi}$  are such that  $\tilde{u}_i = u_{k,l,m}$  and  $\tilde{\psi}_i = \psi_{k,l,m}$  for  $i = I(k, l, m)$ .

Let  $\mathcal{G}_0^L := \mathcal{G}^L \cap L_0^2(\pi)$ . For observables  $\varphi \in \mathcal{G}_0^L$ , standard procedures allow us to derive a stiffness matrix  $\mathbf{A} \in \mathbb{R}^{L^3 \times L^3}$  in terms of which the action of the generator  $\mathcal{L}_{\text{AdL}} = \mathcal{L}_H + \gamma \mathcal{L}_O + \varepsilon^{-1} \mathcal{L}_{\text{NH}}$  can be written as  $\mathcal{L}_{\text{AdL}} \varphi = \mathcal{L}_{\text{AdL}}(\mathbf{u} \cdot \boldsymbol{\psi}) = (\mathbf{A} \mathbf{u}) \cdot \boldsymbol{\psi}$  (see section SM1 of the Supplementary Material for details). Consequently, the spectrum of  $\mathcal{L}_{\text{AdL}}$  in the respective Galerkin subspace is exactly given by the eigenvalues of  $\mathbf{A}$  and we can numerically compute the spectral gap  $\hat{\lambda}_{\varepsilon, \gamma}$  of  $-\mathcal{L}_{\text{AdL}}$  restrained to the respective Galerkin subspace by diagonalizing the matrix  $\mathbf{A}$ . Figure 1 shows the spectral gap of  $-\mathbf{A}$  for  $L = 10$ . As suggested by (2.12) we observe for all considered values of  $\gamma$  a scaling of  $\hat{\lambda}_{\varepsilon, \gamma}$  as  $O(\varepsilon^2)$  when  $\varepsilon \rightarrow 0$  and as  $O(\varepsilon^{-2})$  when  $\varepsilon \rightarrow \infty$  (see Figure 1, Panel A). Similarly, for fixed values of  $\varepsilon$  we observe a scaling of  $\hat{\lambda}_{\varepsilon, \gamma}$  as  $O(\gamma)$  when  $\gamma \rightarrow 0$  and as  $O(\gamma^{-1})$  when  $\gamma \rightarrow \infty$  (see Figure 1, Panel B). Finally, consider the scaling of the spectral gap  $\hat{\lambda}_{\alpha, \alpha}$  as a function of the single scalar  $\alpha$ . As  $\alpha \rightarrow \infty$ , we expect  $\hat{\lambda}_{\alpha, \alpha} = O(\alpha^{-3})$ , and as  $\alpha \rightarrow 0$ , we expect  $\hat{\lambda}_{\alpha, \alpha} = O(\alpha^3)$ . Indeed, this is what we observe (see Figure 1, Panel C).

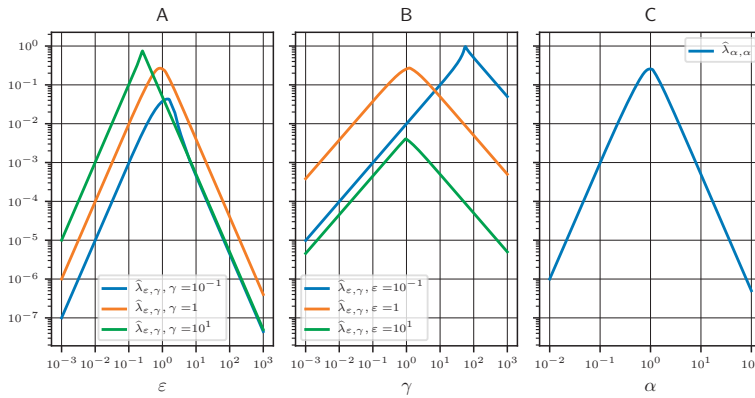


FIG. 1. Spectral gap,  $\hat{\lambda}_{\varepsilon, \gamma}$ , of  $-\mathcal{L}_{\text{AdL}}$  when considered as an operator on  $\mathcal{G}_0^L$ , with  $L = 10$ . Panel A shows  $\hat{\lambda}_{\varepsilon, \gamma}$  as a function of  $\varepsilon$  for fixed  $\gamma$ . Panel B shows  $\hat{\lambda}_{\varepsilon, \gamma}$  as a function of  $\gamma$  for fixed  $\varepsilon$ . Panel C shows  $\hat{\lambda}_{\alpha, \alpha}$  as a function of the scalar  $\alpha$ .

**4.2. Scaling of asymptotic variance and demonstration of CLT.** We next consider a simple skewed double-well potential  $U : \mathbb{R} \rightarrow \mathbb{R}$ , of the form  $U(\mathbf{q}) = \frac{b}{a}(\mathbf{q}^2 - a)^2 + c\mathbf{q}$ , which we parameterize as  $b = 1, a = 1, c = 1/2$ . We use the BADODAB symmetric splitting scheme from [22] (see also section SM3 of the supplementary material) to simulate trajectories of the SDE (2.4). In a first set of simulations we consider different parameterizations with  $\varepsilon$  taking values within the interval  $[10^{-2}, 10]$  and  $\gamma$  taking values within the interval  $[10^{-4}, 10^2]$ . For each parameterization we simulate  $N = 10,000$  independent replicas for  $K = 100,000$  time steps at unit temperature using a stepsize  $\Delta t = 2 \times 10^{-3}$ . We randomly initialized each replica according to the associated equilibrium measure  $\pi$  using a simple rejection sampling algorithm. We denote by  $\hat{\varphi}_K = \frac{1}{K} \sum_{k=0}^{K-1} \varphi(\mathbf{q}^{(k)}, \mathbf{p}^{(k)}, \xi^{(k)})$  the time average of the observable  $\varphi$  evaluated along a finite trajectory  $(\mathbf{q}^{(k)}, \mathbf{p}^{(k)}, \xi^{(k)})_{1 \leq k \leq K}$  of the discretized process which we use as a (biased, due to discretization) Monte Carlo estimate of the expectation  $\mathbb{E}_\pi(\varphi)$ . Let  $\hat{\varphi}_K^{(n)}$  denote the Monte Carlo estimate obtained from the trajectory of the  $n$ th replica, and denote by

$$(4.4) \quad \bar{\varphi}_K := \frac{1}{N} \sum_{n=1}^N \hat{\varphi}_K^{(n)}$$

the empirical mean of the respective estimates over the  $N$  independent replicas. We estimate the asymptotic variance of  $\varphi$  under the discretized dynamics using

$$\hat{\sigma}_{\varepsilon, \gamma}^2(K) = \frac{1}{N} \sum_{n=0}^{N-1} \left( \hat{\varphi}_K^{(n)} - \bar{\varphi}_K \right)^2.$$

Figure 2 shows such computed estimates of the asymptotic variance as a function of  $\varepsilon$  (Panel A), and as a function of  $\gamma$  (Panel B), respectively. We confirm the qualitative behavior predicted in section 3 for the asymptotic variance: for fixed  $\gamma = 1$ , the asymptotic variance  $\hat{\sigma}_{\varepsilon, \gamma}^2(K)$  of observables scales at most quadratically in  $\varepsilon$  as  $\varepsilon \rightarrow \infty$ . Similarly, as  $\varepsilon \rightarrow 0$ , the estimated asymptotic variance  $\hat{\sigma}_{\varepsilon, \gamma}^2(K)$  of the observables we consider remains of order 1 (while it could increase as  $\varepsilon^{-2}$  at most according to (3.3)). For fixed  $\varepsilon = 1$ , the estimated asymptotic variance of observables scales as at most linearly in  $\gamma$  as  $\gamma \rightarrow \infty$ . For the considered model system and observables the increase of the estimated asymptotic variance  $\hat{\sigma}_{\varepsilon, \gamma}^2(K)$  is sublinear in  $\gamma^{-1}$  as  $\gamma \rightarrow 0$ . We provide additional results for a slightly modified version of the model system considered here in section SM2, where the increase of the asymptotic variance of certain observables is indeed observed to be asymptotically linear in  $\gamma^{-1}$  as  $\gamma \rightarrow 0$ . We use a second set of simulations to demonstrate the CLT obtained in Corollary 3.1 for estimates  $\hat{\varphi}_K$  obtained as Monte Carlo estimates from the discretization of the SDE (2.4). That is, we show that for sufficiently large  $K \in \mathbb{N}$  the law of the estimated rescaled residual error

$$(4.5) \quad \sqrt{\frac{K \Delta t}{\sigma_{\varepsilon, \gamma}^2(\varphi)}} (\hat{\varphi}_K - \mathbb{E}_\pi(\varphi))$$

is approximately Gaussian with vanishing mean and variance  $\hat{\sigma}_{\varepsilon, \gamma}^2(\varphi)$  (we treat any systematic bias induced by the discretization as negligible). For parameter values  $\gamma = \varepsilon = 1$ , we simulate  $N = 500,000$  independent trajectories for up to  $K_{\max} = 1000$  steps using the stepsize  $\Delta t = 10^{-1}$ . For each trajectory we compute an estimate of

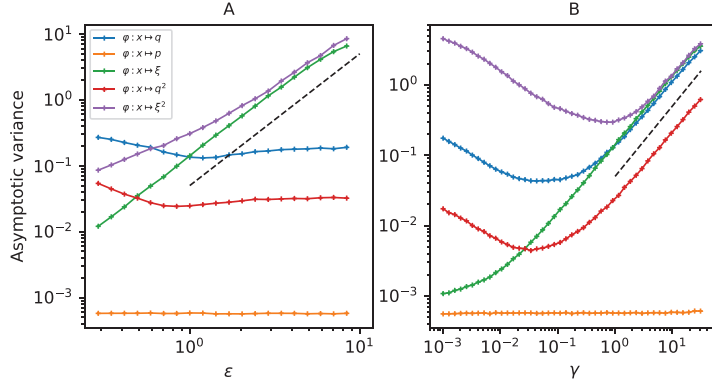


FIG. 2. Estimated asymptotic variance  $\hat{\sigma}_{\varepsilon,\gamma}^2(K)$  for various observables with fixed  $\gamma = 1$  as a function of  $\varepsilon$  (Panel A), and with fixed  $\varepsilon = 1$  as a function of  $\gamma$  (Panel B), respectively. The dashed line in Panel A corresponds to the slope of a quadratic function in  $\varepsilon$ . The dashed line in Panel B indicates the slope of a linear function in  $\gamma$ .

the rescaled residual errors by replacing  $\mathbb{E}_\pi(\varphi)$  and  $\sigma_{\varepsilon,\gamma}^2(\varphi)$  in the expression (4.5) by the Monte Carlo estimates  $\bar{\varphi}_{K_{\max}}$  and  $\hat{\sigma}_{\varepsilon,\gamma}^2(K_{\max})$ , respectively. Figure 3, Panel A, and Figure 3, Panel B, show the empirical probability density function of the rescaled residual errors of the estimated mean and the estimated variance of the position variable  $\mathbf{q}$ , respectively. The empirical probability density functions are plotted for different values of  $K$ . As  $K$  increases we observe that for sufficiently large  $K$  the computed empirical probability density functions indeed closely follow the predicted Gaussian limiting distributions.

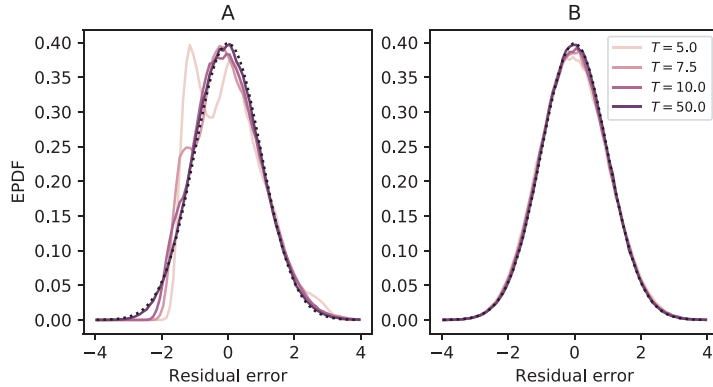


FIG. 3. Empirical probability distribution (EPDF) of the rescaled residual error at different times  $T = K\Delta t$ . Panel A shows the EPDF of the residual error of the estimated mean of the position variable  $\mathbf{q}$ , i.e.,  $\varphi : (\mathbf{q}, \mathbf{p}, \xi) \mapsto \mathbf{q}$ . Panel B shows the residual error of the estimated second moment of the position variable, i.e.,  $\varphi : (\mathbf{q}, \mathbf{p}, \xi) \mapsto \mathbf{q}^2$ . Dotted lines show the density  $\mathcal{N}(0, \sigma^2(\varphi))$ , where  $\sigma^2(\varphi)$  corresponds to the asymptotic variance of the respective observable, which is estimated using the complete trajectory data up to index  $K = 1000$ .

**4.3. Application to Bayesian logistic regression.** For the purpose of demonstrating the CLT in a Bayesian posterior sampling application we consider a Bayesian logistic regression trained on a subset of the MNIST benchmark data set [19] of hand-

written digits for binary classification of the digits 7 and 9. Let us however warn the reader that Adaptive Langevin may not completely remove the minibatching bias in this case (see Remark 1.1). We preprocess the data by means of a principal component analysis. After centering the mean of each pixel, we retain the first 100 principal components and whiten the obtained data by normalizing the variance of the corresponding loadings. The corresponding data points are denoted by  $x^j$ . Pictures corresponding to the number 7 are associated with  $y^j = 0$ , while  $y^j = 1$  corresponds to pictures of 9. Training is run on a subset of 12,251 data points and testing on a separate subset of 2000 data points. Assuming a weakly informative Gaussian prior distribution on the parameters  $\mathbf{q} \in \mathbb{R}^{100}$  to sample, with density  $p_0(\mathbf{q}) \propto \exp(-\mathbf{q}^T \mathbf{q} / (2\sigma^2))$ , where  $\sigma^2 = 100$ , and a likelihood

$$p(y^j, x^j | \mathbf{q}) = \frac{\exp(y^j (x^j)^T \mathbf{q})}{1 + \exp((x^j)^T \mathbf{q})},$$

where  $\tilde{N} = 12251$ ,  $y^j \in \{0, 1\}$ ,  $x^j \in \mathbb{R}^{100}$ , the corresponding posterior distribution is of the form

$$(4.6) \quad \pi(\mathbf{q}) d\mathbf{q} \propto p_0(\mathbf{q}) \prod_{j=1}^{\tilde{N}} p(y^j, x^j | \mathbf{q}) d\mathbf{q} =: \exp(-U(\mathbf{q})) d\mathbf{q}.$$

We use the ODABADO scheme described in the Supplementary Material (section SM3) in order to numerically discretize (1.2) in combination with an unbiased estimator  $-\hat{\nabla}U(\mathbf{q})$  of the gradient force which we obtain by subsampling data points as specified in (1.6) using minibatches of size  $m = 100$ . Besides the introduced gradient noise we do not apply additional random forces, i.e.,  $\sigma_A = 0$ .



FIG. 4. Typical images from the MNIST data set. The upper row shows the original images as obtained from the repository [19]; the lower row shows the projection of the same images onto the first 100 principal components which were used for inference.

In a first set of simulations we generate  $N = 10,000$  independent trajectories for a total number of  $K = 10,000$  steps using a stepsize of  $\Delta t = 10^{-2}$  with coupling parameter  $\nu = 1$ . We initialize the position variable of all replicas at the same location which is a point close to the mode of the target distribution, we set the initial value  $\xi(0)$  of the friction variable to 0, and for each trajectory we independently sample the initial momenta from the stationary measure, i.e.,  $\mathbf{p}(0) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ . Following the same steps as described above in the demonstration of the CLT in the previous example we compute the appropriately rescaled residual errors of the estimated mean and the estimated variance at various time points of the single coordinate variable  $\mathbf{q}_i$  whose index  $i = 65$  we randomly selected. Figure 5 shows the histograms of the empirical distribution of the residual error of these estimates after an increasing number of time steps. Again, as in the example of the previous section we observe that for a sufficiently large number of time steps, the distribution of the residual error follows closely the anticipated Gaussian distribution. We confirm that we observe



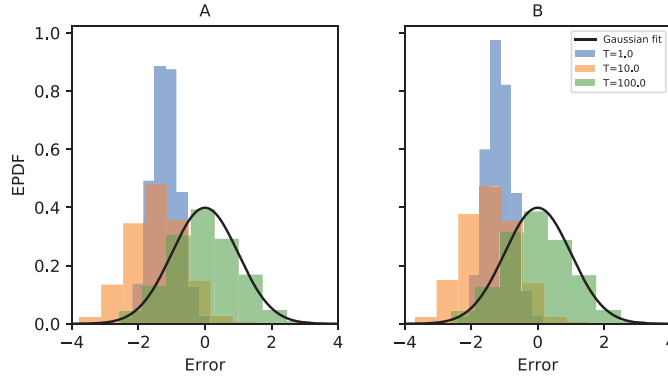


FIG. 5. Empirical probability distribution (EPDF) of the rescaled residual error at different times  $T = K\Delta t$  in the case of the Bayesian logistic regression posterior sampling problem. Panel A shows the EPDF of the rescaled residual error of the estimated mean of the 65th regression variable. Panel B shows the residual error of the estimated variance of the same regression variable.

that also for other choices of the coordinate index  $i$  the empirical law of the residual error converges to a centered Gaussian distribution.

In a second set of simulations we investigate the effect of different values of the thermal mass  $\nu$  on the convergence speed of the estimates of expectations of certain observables obtained from single trajectories. We consider the same setup as described above but generate single trajectories for different values of the coupling parameter, i.e.,  $\nu = \varepsilon^2 \in \{1, 10, 100\}$ . As observables we consider again the projection onto a single coordinate variable, and the average likelihood over the test set, a quantity commonly used for benchmarking purposes in machine learning applications, i.e.,

$$\varphi(\mathbf{q}, \mathbf{p}, \xi) = \frac{1}{\hat{N}} \sum_{i=1}^{\hat{N}} p(y^i, x^i | \mathbf{q}) = \frac{1}{\hat{N}} \sum_{i=1}^{\hat{N}} \frac{\exp(y^i (x^i)^T \mathbf{q})}{1 + \exp((x^i)^T \mathbf{q})},$$

where  $\hat{N} = 2000$ , and  $(x^i, y^i), i = 1, \dots, \hat{N}$ , are the data points of the test data set.

Figure 6 shows the time evolution of the corresponding Monte Carlo estimates of the mean of the 65th component (Panel A), and the average likelihood over the test set (Panel B). As one may have anticipated based on the asymptotic scaling of the spectral gap as  $O(\nu^{-1})$  as  $\nu \rightarrow \infty$ , the convergence of the respective cumulative averages (in time) of the observables under consideration becomes slower with increasing values of  $\nu$ . We mention that estimates appear to converge to different values in the limit  $T = \Delta t K \rightarrow \infty$ . This observation can be explained by the fact that the invariant measure of the discretized dynamics can be expected to depend on the value of the coupling parameter  $\nu$ . We refer the reader to [22] for a detailed analysis of this dependency in the case of the similar BADODAB splitting scheme. A reduction of this discrepancy can be achieved by a reduction of the stepsize  $\Delta t$ .

**Acknowledgments.** This work was initiated during the authors' stay at the Institut Henri Poincaré-Centre Emile Borel during the trimester "Stochastic Dynamics Out of Equilibrium" (April–July 2017). The authors warmly thank this institution for its hospitality. G.S. also benefited from the scientific environment of the Laboratoire International Associé between the Centre National de la Recherche Scientifique and the University of Illinois at Urbana-Champaign.

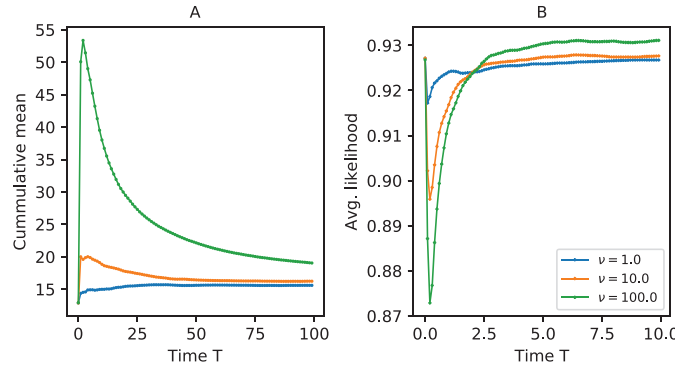


FIG. 6. Simulated time  $T = K\Delta t$  versus value of the Monte Carlo estimate of the mean of the 65th regression variable (Panel A), and the value of the estimated average likelihood over the test set (Panel B).

## REFERENCES

- [1] D. BAKRY, F. BARTHE, P. CATTIAUX, AND A. GUILLIN, *A simple proof of the Poincaré inequality for a large class of probability measures*, Electron. Commun. Probab., 13 (2008), pp. 60–66.
- [2] R. N. BHATTACHARYA, *On the functional central limit theorem and the law of the iterated logarithm for Markov processes*, Z. Wahrsch. Verw. Gebiete, 60 (1982), pp. 185–201.
- [3] N. BOU-RABEE AND H. OWHADI, *Long-run accuracy of variational integrators in the stochastic context*, SIAM J. Numer. Anal., 48 (2010), pp. 278–297, <https://doi.org/10.1137/090758842>.
- [4] E. CANCES, F. LEGOLL, AND G. STOLTZ, *Theoretical and numerical comparison of some sampling methods for molecular dynamics*, ESAIM Math. Model. Numer. Anal., 41 (2007), pp. 351–389.
- [5] T. CHEN, E. FOX, AND C. GUESTRIN, *Stochastic gradient Hamiltonian Monte Carlo*, in International Conference on Machine Learning, Beijing, China, 2014, pp. 1683–1691.
- [6] X. CHENG, N. S. CHATTERJI, P. L. BARTLETT, AND M. I. JORDAN, *Underdamped Langevin MCMC: A non-asymptotic analysis*, in Proceedings of the 31st Conference on Learning Theory, Proc. Mach. Learn. Res. 75, PMLR, 2018, pp. 300–323.
- [7] N. DING, Y. FANG, R. BABBUSH, C. CHEN, R. D. SKEEL, AND H. NEVEN, *Bayesian sampling using stochastic gradient thermostats*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2014, pp. 3203–3211.
- [8] J. DOLBEAULT, C. MOUHOT, AND C. SCHMEISER, *Hypocoercivity for kinetic equations with linear relaxation terms*, C. R. Math. Acad. Sci. Paris, 347 (2009), pp. 511–516.
- [9] J. DOLBEAULT, C. MOUHOT, AND C. SCHMEISER, *Hypocoercivity for linear kinetic equations conserving mass*, Trans. Amer. Math. Soc., 367 (2015), pp. 3807–3828.
- [10] A. EBERLE, *Reflection coupling and Wasserstein contractivity without convexity*, C. R. Math. Acad. Sci. Paris, 349 (2011), pp. 1101–1104.
- [11] A. EBERLE, A. GUILLIN, AND R. ZIMMER, *Couplings and quantitative contraction rates for Langevin dynamics*, Ann. Probab., 47 (2019), pp. 1982–2010.
- [12] M. HAIRER AND J. C. MATTINGLY, *Yet another look at Harris ergodic theorem for Markov chains*, in Seminar on Stochastic Analysis, Random Fields and Applications VI, Progr. Probab. 63, Springer, Berlin, 2011, pp. 109–117.
- [13] D. P. HERZOG, *Exponential relaxation of the Nosé-Hoover equation under Brownian heating*, Commun. Math. Sci., 16 (2018), pp. 2231–2260.
- [14] L. HÖRMANDER, *Hypoelliptic second order differential equations*, Acta Math., 119 (1967), pp. 147–171.
- [15] A. IACOBUCCI, S. OLLA, AND G. STOLTZ, *Convergence rates for nonequilibrium Langevin dynamics*, Ann. Math. Qué., 43 (2019), pp. 73–98.
- [16] A. JONES AND B. LEIMKUHLER, *Adaptive stochastic methods for sampling driven molecular systems*, J. Chem. Phys., 135 (2011), 084125.
- [17] W. KLIEMANN, *Recurrence and invariant measures for degenerate diffusions*, Ann. Probab., 15 (1987), pp. 690–707.

- [18] M. KOPEC, *Weak backward error analysis for Langevin process*, BIT, 55 (2015), pp. 1057–1103.
- [19] Y. LECUN AND C. CORTES, *MNIST handwritten digit database*, 2010, <http://yann.lecun.com/exdb/mnist/>.
- [20] B. LEIMKUHLER AND C. MATTHEWS, *Molecular Dynamics. With Deterministic and Stochastic Numerical Methods*, Interdiscip. Appl. Math. 39, Springer, Cham, 2015.
- [21] B. LEIMKUHLER, C. MATTHEWS, AND G. STOLTZ, *The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics*, IMA J. Numer. Anal., 36 (2015), pp. 13–79.
- [22] B. LEIMKUHLER AND X. SHANG, *Adaptive thermostats for noisy gradient systems*, SIAM J. Sci. Comput., 38 (2016), pp. A712–A736, <https://doi.org/10.1137/15M102318X>.
- [23] T. LELIÈVRE AND G. STOLTZ, *Partial differential equations and stochastic methods in molecular dynamics*, Acta Numer., 25 (2016), pp. 681–880.
- [24] J. C. MATTINGLY, A. M. STUART, AND D. J. HIGHAM, *Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise*, Stochastic Process. Appl., 101 (2002), pp. 185–232.
- [25] S. P. MEYN AND R. L. TWEEDIE, *Stability of Markovian processes II: Continuous-time processes and sampled chains*, Adv. Appl. Probab., 25 (1993), pp. 487–517.
- [26] S. P. MEYN AND R. L. TWEEDIE, *Markov Chains and Stochastic Stability*, Springer, New York, 2012.
- [27] S. NOSÉ, *A unified formulation of the constant temperature molecular dynamics methods*, J. Chem. Phys., 81 (1984), pp. 511–519.
- [28] G. A. PAVLIOTIS, *Stochastic Processes and Applications*, Springer, New York, 2016.
- [29] G. A. PAVLIOTIS AND A. VOGIANNOU, *Diffusive transport in periodic potentials: underdamped dynamics*, Fluct. Noise Lett., 8 (2008), pp. L155–L173.
- [30] L. REY-BELLET, *Ergodic properties of Markov processes*, in Open Quantum Systems II, S. Attal, A. Joye, and C.-A. Pillet, eds., Lecture Notes in Math. 1881, Springer, Berlin, 2006, pp. 1–39.
- [31] J. ROUSSEL AND G. STOLTZ, *Spectral methods for Langevin dynamics and associated error estimates*, ESAIM Math. Model. Numer. Anal., 52 (2018), pp. 1051–1083.
- [32] M. SACHS, B. LEIMKUHLER, AND V. DANOS, *Langevin dynamics with variable coefficients and nonconservative forces: From stationary states to numerical methods*, Entropy, 19 (2017), 647.
- [33] X. SHANG, Z. ZHU, B. LEIMKUHLER, AND A. J. STORKEY, *Covariance-controlled adaptive Langevin thermostat for large-scale Bayesian sampling*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2015, pp. 37–45.
- [34] D. TALAY, *Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme*, Markov Process. Rel. Fields, 8 (2002), pp. 163–198.
- [35] C. VÁRNAI, N. BERNSTEIN, L. MONES, AND G. CSÁNYI, *Tests of an adaptive QM/MM calculation on free energy profiles of chemical reactions in solution*, J. Phys. Chem. B, 117 (2013), pp. 12202–12211.
- [36] C. VILLANI, *Hypoocoercivity*, Mem. Amer. Math. Soc., 202 (2009), 950.
- [37] M. WELLING AND Y. W. TEH, *Bayesian learning via stochastic gradient Langevin dynamics*, in Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, 2011, pp. 681–688.