Some Properties on the Reversibility and the Linear Response Theory of Langevin Dynamics

Yuan Gao[1] · Jian-Guo Liu[2] · Zibu Liu[3]

Abstract

Linear response theory is a fundamental framework studying the macroscopic response of a physical system to an external perturbation. This paper focuses on the rigorous mathematical justification of linear response theory for Langevin dynamics. We give some equivalent characterizations for reversible overdamped/underdamped Langevin dynamics, which is the unperturbed reference system. Then we clarify sufficient conditions for the smoothness and exponential convergence to the invariant measure for the overdamped case. We also clarify those sufficient conditions for the underdamped case, which corresponds to hypoellipticity and hypocoercivity. Based on these, the asymptotic dependence of the response function on the small perturbation is proved in both finite and infinite time horizons. As applications, Green-Kubo relations and linear response theory for a generalized Langevin dynamics are also proved in a rigorous fashion.

Keywords Onsager's principle · Hypocoercivity · Fluctuation-dissipation relation · Non-equilibrium system · Asymptotic behaviors

1 Introduction

Linear response theory is a general framework for studying the behavior of a physical system under small external perturbations. While the microscopic fluctuations in a physical system are usually complex and sustained, linear response theory provides a way to characterize or predict the corresponding macroscopic behaviors of the system. For instance, the famous Einstein relation [7] for Brownian motion fits within the framework of linear re-

✉ Y. Gao
gao662@purdue.edu

J.-G. Liu
jliu@math.duke.edu

Z. Liu
zibu.liu@duke.edu

1 Department of Mathematics, Purdue University, West Lafayette, IN, USA

2 Department of Mathematics and Department of Physics, Duke University, Durham, NC, USA

3 Department of Mathematics, Duke University, Durham, NC, USA

🙂 Springer

sponse theory, using the correlation of Brownian particles to predict macroscopic diffusion coefficients.

Linear response theory not only bridges the microscopic first physical principles with macroscopic properties (e.g., electrical, thermal, transport, and mechanical properties), but it also helps in understanding or predicting a nonequilibrium system, which deviates from equilibrium states, using only the information of the original equilibrium system. Thus, it plays a fundamental role in theoretical physics and statistical mechanics.

Given these important applications, linear response theory has been extensively studied, and various versions of response formulas for different physical contexts have been developed, cf. [4, 5, 8, 20, 28].

In this paper, we do not attempt to propose new linear response relations. Instead, we focus on clarifying mathematical conditions and collecting further properties with quantitative estimates for the well-known linear response results for Langevin dynamics. Specifically, in a mathematically rigorous way, we study the asymptotic dependence of the solution to the perturbed Fokker-Planck equation concerning the small external force added to the original reversible Langevin dynamics.

We begin with an overdamped or underdamped Langevin dynamics, initially in a reversible form or with a detailed balanced invariant measure known as the Gibbs measure. Under mild assumptions on the conservative drift, the probability distribution will exponentially converge to the invariant measure as time tends to infinity. To study the nonequilibrium system perturbed by a small non-conservative external force in the long-time regime, the existence, smoothness, and exponential stability of the perturbed invariant measure are essential preliminaries. These involve semigroup theory, Harris' theorem, and, particularly, hypoellipticity and hypocoercivity for the Fokker-Planck equation in the underdamped case.

In Sect. 2, we first provide the reversibility conditions in five equivalent forms for overdamped and underdamped Langevin dynamics, and generalized Langevin dynamics.

In Sect. 3, we first clarify the existence and uniqueness of positive invariant measures for perturbed irreversible Langevin dynamics. The smoothness and exponential convergence of the solution to the Fokker-Planck equation for the perturbed irreversible Langevin dynamics are also summarized, with proofs provided in the Appendix. After these preparations, we give a rigorous verification of linear response theory for overdamped Langevin dynamics. We analyze the asymptotic behavior of the response function, which is defined as the difference (in weak form with respect to any observation test function) between the perturbed dynamics of the probability distribution and the initial data given by the unperturbed Gibbs measure $\rho_0(\mathbf{q})$

$$R(t, \varepsilon; \varphi) = \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho^{\varepsilon}(\mathbf{q}, t) \mathrm{d}\mathbf{q} - \int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_0(\mathbf{q}) \mathrm{d}\mathbf{q} \right).$$

In Theorem 3.2, we prove the convergence of the response function $R(t, \varepsilon; \varphi)$ for either fixed t or fixed ε, as well as the double limits for both $\lim_{\varepsilon \to 0^+} \lim_{t \to +\infty}$ and $\lim_{t \to +\infty} \lim_{\varepsilon \to 0^+}$. Moreover, we also obtain the uniform convergence of the response function for $\varepsilon \leq \varepsilon_0$ in time $t \in [0, +\infty)$. In the special case where the external perturbation is a conservative force ∇W for some potential W, we derive the Green-Kubo formula, which connects the long-time behavior of the response function to the autocorrelation function over an infinite time horizon for the original unperturbed dynamics

$$\lim_{\varepsilon \to 0^+} \lim_{t \to \infty} R(t, \varepsilon; \mathcal{L}_0 g) = -\int_0^{+\infty} K_{\mathcal{L}_0 g, \mathcal{L}_0 W}(t) \, \mathrm{d}t, \quad K_{AB}(t) := \int_{\mathbb{R}^d} (e^{t\mathcal{L}_0} A) B \rho_0 \mathrm{d}\mathbf{q},$$

where $\mathcal{L}_0$ is the unperturbed generator.

In Sect. 4, we first summarize key well-posedness results and the hypoellipticity and hypocoercivity of the Fokker-Planck equation in the underdamped case. Villani's seminal work on hypocoercivity [29] shows that exponential convergence to the invariant measure can still be achieved under mild conditions on the original potential. Given that we consider an external perturbation in a conservative form and with compact support, the perturbed Fokker-Planck equation in the underdamped case will still converge exponentially to the new invariant measure. Based on this, we prove parallel results for the response function $R(t, \varepsilon; \varphi)$ under both fixed t or ε, and for the double limits behavior in Theorem 4.1. Lastly, in Sect. 4.3, as an application of Theorem 4.1, we derive the linear response theory for generalized Langevin dynamics with an exponential kernel using the corresponding augmented underdamped Langevin dynamics.

Our results on the linear response theory for stochastic dynamics extend the case where the external force is a general vector field depending on spatial variables. These results are parallel to those in the comprehensive book by PAVLIOTIS [25], which focuses on the case where the external force is independent of spatial variables. On the other hand, the abstract theorem by HAIRER-MAJDA in [11] provides a general methodology for justifying linear response theory for abstract Markov evolutions under several assumptions. Our work falls within the general framework of [11], but we offer self-contained conditions with detailed verification for Langevin dynamics and prove the uniform convergence of the response function for small external perturbations $\varepsilon \leq \varepsilon_0$ uniformly in time $t \in [0, +\infty)$. For more on linear response theory for deterministic dynamical systems, see [27].

The remaining sections of this paper are organized as follows: In Sect. 2, we provide equivalent conditions for reversibility in overdamped, underdamped, and generalized Langevin dynamics. In Sect. 3, we prove the linear response theory and the Green-Kubo relation for overdamped Langevin dynamics. In Sect. 4, we prove the linear response theory for underdamped Langevin dynamics and generalized Langevin dynamics. All omitted proofs for reversibility, hypoellipticity, and hypocoercivity are given in Appendices A and B.

2 Equivalent Conditions of Reversibility of Langevin Dynamics

Before studying the linear response theory for Langevin dynamics in the large-time regime, we need to first establish the well-posedness and stability of the invariant measure, which are built on the reversibility of the original unperturbed system. In this section, we provide several equivalent conditions for the reversibility of both overdamped and underdamped Langevin dynamics. Although these conditions are classic, some are less well-known but useful in practice. Therefore, we rigorously clarify these reversibility conditions for the reader's convenience.

2.1 Definitions and Preliminaries

We will study the reversibility of both overdamped, underdamped and generalized Langevin dynamics. In the underdamped and generalized Langevin dynamics case, some physical variables will have even or odd parities. A typical odd variable is velocity (**p**), while typical even variables are position (**q**), force (**f**), and acceleration (**a**). To preserve more flexibility, we allow different components of **x** to have different parities.

Another essential difference is the invariant measure (see (2.19)) for the underdamped case, which involves both kinetic and potential energy, while the kinetic energy is neglected for the overdamped case. Below, we provide some basic definitions for reversibility.

Definition 2.1 (Stationary process) A stochastic process $\{\mathbf{x}(t) \in \mathbb{R}^d, t \geq 0\}$ is stationary if for all $\tau, t_1, t_2, \ldots, t_n \in \mathbb{R}^+$, $n \in \mathbb{Z}^+$, and Borel sets $\mathcal{B}_j \in \mathbb{R}^d$, $j = 1, 2, \ldots, n$,

$$\mathbb{P}(\mathbf{x}(t_j) \in \mathcal{B}_j, 1 \leq j \leq n) = \mathbb{P}(\mathbf{x}(t_j + \tau) \in \mathcal{B}_j, 1 \leq j \leq n). \tag{2.1}$$

Taking $n = 1$, we deduce that the law of $\mathbf{x}(t)$ is invariant.

Definition 2.2 (Time-reversed process) Let $\{\mathbf{x}(t) \in \mathbb{R}^d, t \geq 0\}$ be a stochastic process. Fix time $T > 0$. The time-reversed process of $\mathbf{x}(t)$ (w.r.t. time T) is defined as

$$\mathbf{x}^*(t) := (\varepsilon_i x_i(T - t))_{1 \leq i \leq n}, \tag{2.2}$$

where

$$\varepsilon_i := \begin{cases} 1, & \text{if } x_i \text{ is an even variable}; \\ -1, & \text{if } x_i \text{ is an odd variable}. \end{cases}$$

Definition 2.3 (Reversibility) A stationary process $\{X_t \in \mathbb{R}^d, t \geq 0\}$ is reversible if for any $T > 0$, the time-reversed process $\{X_t^*, 0 \leq t \leq T\}$ (w.r.t. T) has the same finite-dimensional distribution as the original process X_t, i.e., for any $t_1, t_2, \ldots, t_n \in [0, T]$, $n \in \mathbb{Z}^+$, and Borel sets $\mathcal{B}_j \in \mathbb{R}^d$, $j = 1, 2, \ldots, n$,

$$\mathbb{P}(X_{t_j} \in \mathcal{B}_j, 1 \leq j \leq n) = \mathbb{P}(X_{t_j}^* \in \mathcal{B}_j, 1 \leq j \leq n). \tag{2.3}$$

2.2 Reversibility of the Overdamped Langevin Dynamics

We use $\mathbf{q}_t$ to indicate the stochastic process at time t, and we also use $\mathbf{q}$ as the displacement vector in $\mathbb{R}^d$, following the convention in mechanics. The overdamped Langevin equation is given by the following stochastic differential equation (SDE)

$$d\mathbf{q}_t = \mathbf{b}(\mathbf{q}_t)dt + \boldsymbol{\sigma}(\mathbf{q}_t)d\mathbf{B}_t, \quad \mathbf{q}_{t=0} = \mathbf{q_0}, \tag{2.4}$$

where $\mathbf{b} \in C^\infty(\mathbb{R}^d; \mathbb{R}^d)$, $\mathbf{q}_t \in \mathbb{R}^d$, and $\boldsymbol{\sigma} \in C^\infty(\mathbb{R}^d; \mathbb{R}^{d \times d})$ is nonsingular. $\mathbf{B}_t$ is the d-dimensional standard Brownian motion, and $\mathbf{q}_{t=0} = \mathbf{q_0}$ is a random variable independent of $\mathbf{B}_t$.

To ensure that (2.4) admits a unique strong solution, we assume that coefficients $\mathbf{b}(\mathbf{q})$ and $\boldsymbol{\sigma}(\mathbf{q})$ satisfy certain conditions (see (2.9) and (2.10) below).

The corresponding Fokker-Planck equation and the Fokker-Planck operator $\mathcal{L}^*$ are given by

$$\frac{\partial \rho(\mathbf{q}, t)}{\partial t} = (\mathcal{L}^* \rho)(\mathbf{q}, t), \quad \mathcal{L}^* \rho := -\nabla \cdot (\rho \mathbf{b}) + \frac{1}{2}\nabla^2 : (\rho \boldsymbol{\sigma}\boldsymbol{\sigma}^T). \tag{2.5}$$

The corresponding Kolmogorov backward equation is

$$\frac{\partial f(\mathbf{q}, t)}{\partial t} = (\mathcal{L}f)(\mathbf{q}, t), \quad \text{with generator } \mathcal{L}f := \mathbf{b} \cdot \nabla f + \frac{1}{2}(\boldsymbol{\sigma}\boldsymbol{\sigma}^T) : \nabla^2 f. \tag{2.6}$$

Define the probability flux $\mathbf{j}$ as

$$\mathbf{j}(\rho) := -\frac{1}{2}\nabla \cdot (\rho \boldsymbol{\sigma}\boldsymbol{\sigma}^T) + \rho \mathbf{b}.$$

Then the Fokker-Planck equation can be written as the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0. \tag{2.7}$$

According to (2.3), we have the following equivalent definition of reversibility.

Lemma 2.1 *Consider* (2.4) *with initial density function* $\rho_0(\mathbf{q})$. *Then the system is reversible if and only if, for any* $\varphi_1, \varphi_2 \in C_b^\infty(\mathbb{R}^d)$

$$\mathbb{E}[\varphi_1(\mathbf{q}_t)\varphi_2(\mathbf{q}_0) \mid \mathbf{q}_0 \sim \rho_0] = \mathbb{E}[\varphi_1(\mathbf{q}_0)\varphi_2(\mathbf{q}_t) \mid \mathbf{q}_0 \sim \rho_0]. \tag{2.8}$$

(Proof of Lemma 2.1 can be found in Appendix A.) We will use (2.8) to prove reversibility.

Theorem 2.1 (Reversibility of the overdamped Langevin dynamics) *Consider* (2.4) *with initial probability density function* $\rho_0(\mathbf{q}) > 0$. *Suppose that* $\boldsymbol{\sigma}$ *is constant and nonsingular. Also, suppose that* $\boldsymbol{\sigma}$ *and* $\mathbf{b}$ *satisfy*

(i) (*Local Lipschitz continuity*) *For any* $n \in \mathbb{Z}^+$, *there exists* $K_n > 0$ *such that*

$$|\mathbf{b}(\mathbf{x}) - \mathbf{b}(\mathbf{y})| \leq K_n |\mathbf{x} - \mathbf{y}|, \tag{2.9}$$

   for any $\mathbf{x}, \mathbf{y}$ *such that* $|\mathbf{x}|, |\mathbf{y}| \leq n$;

(ii) (*Monotone condition*) *There exists a constant* $C > 0$ *such that for any* $\mathbf{x} \in \mathbb{R}^d$:

$$\mathbf{x}^T \mathbf{b}(\mathbf{x}) \leq C(1 + |\mathbf{x}|^2). \tag{2.10}$$

Then the following are equivalent

(i) (*Reversibility*) *The stochastic process determined by* (2.4) *is reversible in the sense of Definition* 2.3;

(ii) (*Symmetry*) *For arbitrary* $\varphi_1, \varphi_2 \in C_b^\infty(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} (\mathcal{L}\varphi_1)(\mathbf{q})\varphi_2(\mathbf{q})\rho_0(\mathbf{q})\mathrm{d}\mathbf{q} = \int_{\mathbb{R}^d} (\mathcal{L}\varphi_2)(\mathbf{q})\varphi_1(\mathbf{q})\rho_0(\mathbf{q})\mathrm{d}\mathbf{q}; \tag{2.11}$$

(iii) (*Zero flux*) *The probability flux is zero for all* $t \geq 0$, *i.e.,* $\mathbf{j}(\rho(\cdot, t)) = 0$;

(iv) (*Potential condition*) *There exists* $U : \mathbb{R}^d \to \mathbb{R}$ *such that*

$$-\nabla U(\mathbf{q}) = 2(\boldsymbol{\sigma}\boldsymbol{\sigma}^T)^{-1}\mathbf{b}, \tag{2.12}$$

   and $\rho_0(\mathbf{q}) = \frac{1}{Z}e^{-U(\mathbf{q})}$ *with* $\int_{\mathbb{R}^d} e^{-U(\mathbf{q})}\mathrm{d}\mathbf{q} < \infty$.

We point out that assumptions (2.9) and (2.10) ensure the well-posedness of the SDE, but they are irrelevant to the reversibility. We adopt (2.10) instead of the general linear growth condition as it covers more cases. For instance, $\mathbf{b}(\mathbf{x}) = (1 - |\mathbf{x}|^2)\mathbf{x}$ does not satisfy the linear growth condition but does satisfy (2.10).

2.3 Reversibility of Underdamped Langevin Equation

Now consider the underdamped Langevin equation

$$
\begin{cases}
d\mathbf{q} = \mathbf{p}dt, \\
d\mathbf{p} = -\dfrac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p}dt + \mathbf{b}(\mathbf{q})dt + \dfrac{\boldsymbol{\sigma}}{\sqrt{\beta}}d\mathbf{B}.
\end{cases}
\tag{2.13}
$$

Here $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$, $\boldsymbol{\sigma} \in \mathbb{R}^{d \times d}$ is a nonsingular constant matrix, and $\beta = 1/(k_B T)$ is the thermodynamic beta. The corresponding Fokker-Planck equation (also known as the kinetic Fokker-Planck equation) and the Fokker-Planck operator $\mathcal{L}^*$ are

$$
\frac{\partial \rho(\mathbf{q}, \mathbf{p}, t)}{\partial t} = (\mathcal{L}^*\rho)(\mathbf{q}, \mathbf{p}, t),
$$
$$
\mathcal{L}^*\rho := -\mathbf{p} \cdot \nabla_{\mathbf{q}}\rho - \mathbf{b} \cdot \nabla_{\mathbf{p}}\rho + \frac{1}{2}\nabla_{\mathbf{p}} \cdot \left(\rho\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p} + \frac{1}{\beta}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla_{\mathbf{p}}\rho\right).
\tag{2.14}
$$

The corresponding Kolmogorov backward equation is

$$
\frac{\partial f(\mathbf{q}, \mathbf{p}, t)}{\partial t} = (\mathcal{L}f)(\mathbf{q}, \mathbf{p}, t),
$$
$$
\mathcal{L}f := \mathbf{p} \cdot \nabla_{\mathbf{q}}f + \mathbf{b} \cdot \nabla_{\mathbf{p}}f - \frac{1}{2}(\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p}) \cdot \nabla_{\mathbf{p}}f + \frac{1}{2\beta}(\boldsymbol{\sigma}\boldsymbol{\sigma}^T) : \nabla_{\mathbf{p}}^2 f.
\tag{2.15}
$$

We say that (2.13) satisfies the fluctuation-dissipation theorem because the dissipative part corresponding to Ornstein–Uhlenbeck process in $\mathcal{L}^*$ can be rewritten as

$$
\frac{1}{2}\nabla_{\mathbf{p}} \cdot \left(\rho\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p} + \frac{1}{\beta}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla_{\mathbf{p}}\rho\right) = \frac{1}{2}\nabla_{\mathbf{p}} \cdot \left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T\left(\rho\nabla_{\mathbf{p}}\left(\frac{\|\mathbf{p}\|^2}{2} + \frac{\log\rho}{\beta}\right)\right)\right).
$$

As with the overdamped Langevin equation, we consider an equivalent condition for reversibility using test functions. Recall that momentum $\mathbf{p}$ is an odd variable, so the equivalent characterization below also uses parity-reversed test functions $\widetilde{\varphi}_2(\mathbf{p}) = \varphi_2(-\mathbf{p})$ and $\widetilde{\psi}_2(\mathbf{p}) = \psi_2(-\mathbf{p})$.

Lemma 2.2 *Consider* (2.13) *with the initial density function* $\rho_0(\mathbf{q})$. *Then it is reversible if and only if, for any* $\varphi_1, \varphi_2 \in C_0^\infty(\mathbb{R}^d)$,

$$
\mathbb{E}\left[\varphi_1(\mathbf{q}(0))\varphi_2(\mathbf{p}(0))\psi_1(\mathbf{q}(t))\psi_2(\mathbf{p}(t)) \,\big|\, (\mathbf{q}(0), \mathbf{p}(0)) \sim \rho_0\right]
$$
$$
= \mathbb{E}\left[\varphi_1(\mathbf{q}(t))\widetilde{\varphi}_2(\mathbf{p}(t))\psi_1(\mathbf{q}(0))\widetilde{\psi}_2(\mathbf{p}(0)) \,\big|\, (\mathbf{q}(0), \mathbf{p}(0)) \sim \rho_0\right].
\tag{2.16}
$$

The proof of Lemma 2.2 is given in Appendix A.

Next, we provide the equivalence theorem for the reversibility of the underdamped Langevin dynamics.

Theorem 2.2 (*Reversibility of the underdamped Langevin*) *Consider* (2.13) *with the initial probability density function* $\rho_0(\mathbf{q}, \mathbf{p}) > 0$. *Suppose that* $\boldsymbol{\sigma}$ *is a constant and nonsingular matrix, and that* $\boldsymbol{\sigma}$ *and* $\mathbf{b}$ *satisfy* (2.9) *and* (2.10). *Then the following are equivalent*

(i) (*Reversibility*) *The stochastic process determined by* (2.13) *is reversible.*

(ii) (*Symmetry*) $\rho_0(\mathbf{q}, \mathbf{p}) = \rho_0(\mathbf{q}, -\mathbf{p})$ *and, for arbitrary* $\varphi_1(\mathbf{q}), \varphi_2(\mathbf{p}), \psi_1(\mathbf{q}), \psi_2(\mathbf{p}) \in C_0^\infty(\mathbb{R}^d)$

$$\iint \varphi_1(\mathbf{q})\varphi_2(\mathbf{p})\mathcal{L}\left(\psi_1(\mathbf{q})\psi_2(\mathbf{p})\right)\rho_0(\mathbf{q}, \mathbf{p})\, \mathrm{d}\mathbf{q}\, \mathrm{d}\mathbf{p}$$

$$= \iint \mathcal{L}\left(\varphi_1(\mathbf{q})\widetilde{\varphi}_2(\mathbf{p})\right)\psi_1(\mathbf{q})\widetilde{\psi}_2(\mathbf{p})\rho_0(\mathbf{q}, \mathbf{p})\, \mathrm{d}\mathbf{q}\, \mathrm{d}\mathbf{p}, \qquad (2.17)$$

where $\widetilde{\varphi}_2(\mathbf{p}) = \varphi_2(-\mathbf{p})$ *and* $\widetilde{\psi}_2(\mathbf{p}) = \psi_2(-\mathbf{p})$.

(iii) (*Potential condition*) *There exists* $U : \mathbb{R}^d \to \mathbb{R}$ *such that*

$$-\nabla_{\mathbf{q}} U(\mathbf{q}) = \mathbf{b}, \qquad (2.18)$$

and

$$\rho_0(\mathbf{q}, \mathbf{p}) = \frac{1}{Z}e^{-\beta H(\mathbf{q}, \mathbf{p})}, \quad Z = \iint e^{-\beta H(\mathbf{q}, \mathbf{p})}\, \mathrm{d}\mathbf{q}\, \mathrm{d}\mathbf{p} < \infty, \qquad (2.19)$$

where

$$H(\mathbf{q}, \mathbf{p}) := \frac{|\mathbf{p}|^2}{2} + U(\mathbf{q}). \qquad (2.20)$$

(iv) (*Evenness in* $\mathbf{p}$ *variable*) $\rho_0(\mathbf{q}, \mathbf{p}) = \rho_0(\mathbf{q}, -\mathbf{p})$ *and* ρ_0 *is stationary, i.e., it solves* $\mathcal{L}^*\rho_0 = 0$.

(v) (*Separation of variables*) $\rho_0(\mathbf{q}, \mathbf{p}) = U_1(\mathbf{q})U_2(\mathbf{p})$, *where* $U_1 > 0$ *and* $U_2 > 0$, *and* ρ_0 *is stationary, i.e., it solves* $\mathcal{L}^*\rho_0 = 0$.

We remark here that the conclusions of Theorem 2.1 and Theorem 2.2 also hold for non-constant σ. However, in this case, the SDE should be written in the backward Itô's integral sense [13]. For the overdamped Langevin equation

$$\mathrm{d}\mathbf{q} = \mathbf{b}(\mathbf{q})\mathrm{d}t + \sigma(\mathbf{q})\hat{\mathrm{d}}\mathbf{B},$$

the Fokker-Planck equation is given by

$$\frac{\partial \rho(\mathbf{q}, t)}{\partial t} = (\mathcal{L}^*\rho)(\mathbf{q}, t), \quad \mathcal{L}^*\rho := \nabla \cdot \left(\frac{1}{2}\sigma\sigma^T \nabla \rho - \rho \mathbf{b}\right).$$

For the underdamped Langevin equation, the SDE should be written as

$$\begin{cases} \mathrm{d}\mathbf{q} = \mathbf{p}\mathrm{d}t, \\ \mathrm{d}\mathbf{p} = -\frac{1}{2}\sigma\sigma^T \mathbf{p}\mathrm{d}t + \mathbf{b}(\mathbf{q})\mathrm{d}t + \frac{\sigma}{\sqrt{\beta}}\hat{\mathrm{d}}\mathbf{B}, \end{cases}$$

with the Fokker-Planck equation

$$\frac{\partial \rho(\mathbf{q}, \mathbf{p}, t)}{\partial t} = (\mathcal{L}^*\rho)(\mathbf{q}, \mathbf{p}, t),$$

$$\mathcal{L}^*\rho := -\mathbf{p} \cdot \nabla_{\mathbf{q}}\rho - \mathbf{b} \cdot \nabla_{\mathbf{p}}\rho + \frac{1}{2}\nabla_{\mathbf{p}} \cdot \left(\rho\sigma\sigma^T \mathbf{p} + \frac{1}{\beta}\sigma\sigma^T \nabla_{\mathbf{p}}\rho\right).$$

2.4 The Reversibility of Generalized Langevin Dynamics with Memory

Let $n \geq 1$ be a fixed integer. Now consider the following generalized Langevin equation with memory

$$
\begin{cases}
\ddot{\mathbf{q}}(t) = \mathbf{b}(\mathbf{q}) - \sum_{i=1}^{n} \mathbf{A}_i \mathbf{A}_i^T \int_0^t e^{-\alpha_i(t-s)} \dot{\mathbf{q}}(s) \mathrm{d}s + \mathbf{A}_i \mathbf{f}_i(t), \\
\mathrm{d}\mathbf{f}_i(t) = -\alpha_i \mathbf{f}_i \mathrm{d}t + \sqrt{2\beta^{-1}\alpha_i}\,\mathrm{d}\mathbf{B}_i, \quad i = 1, 2, \ldots, n.
\end{cases}
\tag{2.21}
$$

Here, for each $i = 1, 2, \ldots, n$, $\mathbf{A}_i \in \mathbb{R}^{d \times d}$ is a constant matrix, $\alpha_i > 0$ is a constant, and $\mathbf{B}_i(t) \in \mathbb{R}^d$ is a standard Brownian motion. Additionally, $\mathbf{B}_i(t)$ for $i = 1, 2, \ldots, n$ are independent. Reversibility can also be considered for (2.21) if we reformulate it.

Let $\mathbf{z}_i(t) = -\mathbf{A}_i^T \int_0^t e^{-\alpha_i(t-s)} \dot{\mathbf{q}}(s)\mathrm{d}s + \mathbf{f}_i(t)$ and $\dot{\mathbf{q}} = \mathbf{p}$. Then (2.21) can be reformulated as

$$
\begin{cases}
\mathrm{d}\mathbf{q} = \mathbf{p}\mathrm{d}t, \\
\mathrm{d}\mathbf{p} = \left(\mathbf{b}(\mathbf{q}) + \sum_{i=1}^{n} \mathbf{A}_i \mathbf{z}_i \right) \mathrm{d}t, \\
\mathrm{d}\mathbf{z}_i = -(\alpha_i \mathbf{z}_i + \mathbf{A}_i^T \mathbf{p})\mathrm{d}t + \sqrt{2\beta^{-1}\alpha_i}\,\mathrm{d}\mathbf{B}_i, \quad i = 1, 2, \ldots, n.
\end{cases}
\tag{2.22}
$$

We will consider (2.22) instead of (2.21) for reversibility. The corresponding Fokker-Planck equation for (2.22) is given by

$$
\frac{\partial \rho}{\partial t} = \mathcal{L}^* \rho, \quad \mathcal{L}^* \rho := -\mathbf{p} \cdot \nabla_{\mathbf{q}} \rho - \mathbf{b}(\mathbf{q}) \cdot \nabla_{\mathbf{p}} \rho - \sum_{i=1}^{n} \left((\mathbf{A}_i \mathbf{z}_i) \cdot \nabla_{\mathbf{p}} \rho - \mathbf{A}_i^T \mathbf{p} \cdot \nabla_{\mathbf{z}_i} \rho \right)
$$
$$
+ \sum_{i=1}^{n} \alpha_i \nabla_{\mathbf{z}_i} \cdot \left(\mathbf{z}_i \rho + \frac{1}{\beta} \nabla_{\mathbf{z}_i} \rho \right).
\tag{2.23}
$$

The corresponding backward equation is given by

$$
\frac{\partial f}{\partial t} = \mathcal{L}f, \quad \mathcal{L}f := \mathbf{p} \cdot \nabla_{\mathbf{q}} f + \mathbf{b}(\mathbf{q}) \cdot \nabla_{\mathbf{p}} f + \sum_{i=1}^{n} \left((\mathbf{A}_i \mathbf{z}_i) \cdot \nabla_{\mathbf{p}} f - \mathbf{A}_i^T \mathbf{p} \cdot \nabla_{\mathbf{z}_i} f \right)
$$
$$
+ \sum_{i=1}^{n} \alpha_i \left(-\mathbf{z}_i \cdot \nabla_{\mathbf{z}_i} f + \frac{1}{\beta} \Delta_{\mathbf{z}_i} f \right).
\tag{2.24}
$$

We now state the equivalent theorem characterizing the reversibility of (2.22) in the case where $n = 1$ and $\mathbf{A}_1 = \mathbf{I} \in \mathbb{R}^{d \times d}$, i.e.,

$$
\begin{cases}
\mathrm{d}\mathbf{q} = \mathbf{p}\mathrm{d}t, \\
\mathrm{d}\mathbf{p} = \left(-\nabla_{\mathbf{q}} V(\mathbf{q}) + \mathbf{z} \right) \mathrm{d}t, \\
\mathrm{d}\mathbf{z} = -(\alpha \mathbf{z} + \mathbf{p})\mathrm{d}t + \sqrt{2\beta^{-1}\alpha}\,\mathrm{d}\mathbf{B}.
\end{cases}
\tag{2.25}
$$

The proof is given in Appendix A.

Theorem 2.3 (*Reversibility of the generalized Langevin with memory*) *Consider* (2.22) *with initial probability density function* $\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) > 0$ *satisfying* $\rho_0 > 0$. *Suppose that* $\boldsymbol{\sigma}$ *is a constant and nonsingular matrix. Then the following are equivalent*

(i) (*Reversibility*) *The stochastic process determined by* (2.22) *is reversible.*

(ii) (*Symmetry*) $\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) = \rho_0(\mathbf{q}, -\mathbf{p}, \mathbf{z})$, *and for arbitrary* $\varphi_1(\mathbf{q}), \varphi_2(\mathbf{p}), \varphi_3(\mathbf{z}), \psi_1(\mathbf{q}), \psi_2(\mathbf{p}), \psi_3(\mathbf{z}) \in C_0^\infty(\mathbb{R}^d)$

$$
\iiint \varphi_1(\mathbf{q}) \varphi_2(\mathbf{p}) \varphi_3(\mathbf{z}) \mathcal{L}\left(\psi_1(\mathbf{q}) \psi_2(\mathbf{p}) \psi_3(\mathbf{z})\right) \rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) \, \mathrm{d}\mathbf{q} \, \mathrm{d}\mathbf{p} \, \mathrm{d}\mathbf{z}
$$
$$
= \iiint \mathcal{L}\left(\varphi_1(\mathbf{q}) \widetilde{\varphi}_2(\mathbf{p}) \varphi_3(\mathbf{z})\right) \psi_1(\mathbf{q}) \widetilde{\psi}_2(\mathbf{p}) \psi_3(\mathbf{z}) \rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) \, \mathrm{d}\mathbf{q} \, \mathrm{d}\mathbf{p} \, \mathrm{d}\mathbf{z},
\tag{2.26}
$$

where $\widetilde{\varphi}_2(\mathbf{p}) = \varphi_2(-\mathbf{p})$ *and* $\widetilde{\psi}_2(\mathbf{p}) = \psi_2(-\mathbf{p})$.

(iii) (*Potential condition*) *There exists* $U : \mathbb{R}^d \to \mathbb{R}$ *such that*

$$
-\nabla_{\mathbf{q}} U(\mathbf{q}) = \mathbf{b},
\tag{2.27}
$$

and

$$
\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) = \frac{1}{Z} e^{-\beta H(\mathbf{q}, \mathbf{p}, \mathbf{z})}, \quad Z = \iiint e^{-\beta H(\mathbf{q}, \mathbf{p}, \mathbf{z})} \, \mathrm{d}\mathbf{q} \, \mathrm{d}\mathbf{p} \, \mathrm{d}\mathbf{z} < \infty,
\tag{2.28}
$$

where

$$
H(\mathbf{q}, \mathbf{p}, \mathbf{z}) := U(\mathbf{q}) + \frac{|\mathbf{p}|^2}{2} + \frac{|\mathbf{z}|^2}{2}.
\tag{2.29}
$$

(iv) (*Evenness in* $\mathbf{p}$ *variable*) $\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) = \rho_0(\mathbf{q}, -\mathbf{p}, \mathbf{z})$ *and* ρ_0 *is stationary, i.e., it solves* $\mathcal{L}^* \rho_0 = 0$.

(v) (*Separation of variables*) $\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) = U_1(\mathbf{q}) U_2(\mathbf{p}) U_3(\mathbf{z})$, *where* $U_1, U_2, U_3 > 0$, *and* ρ_0 *is stationary, i.e., it solves* $\mathcal{L}^* \rho_0 = 0$.

3 Rigorous Verification of Linear Response Theory (LRT) and the Green-Kubo Formula for the Overdamped Langevin

In this section, we rigorously verify the linear response theory (LRT) and the Green-Kubo relation for overdamped reversible Langevin dynamics with constant diffusion coefficients. The linear response theory aims to study the asymptotic dependence of the solution ρ^ε of the perturbed Fokker-Planck equation (3.9) in terms of the small external force. Recall the Gibbs measure $\rho_0(\mathbf{q})$ for the original overdamped Langevin dynamics. In the weak formulation, for any $\varphi \in C_c^\infty(\mathbb{R}^d)$, we define the response function as follows:

$$
R(t, \varepsilon; \varphi) := \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho^\varepsilon(\mathbf{q}, t) \mathrm{d}\mathbf{q} - \int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_0(\mathbf{q}) \mathrm{d}\mathbf{q} \right).
\tag{3.1}
$$

The term $\varepsilon R(t, \varepsilon; \varphi)$ represents the leading-order (i.e., $O(\varepsilon)$) change of an observable φ at time t under the external perturbation $\varepsilon \mathbf{M}$. Mathematically, LRT focuses on the behavior of $R(t, \varepsilon; \varphi)$ as $\varepsilon \to 0$ and $t \to \infty$. In Theorem 3.2, we will prove the convergence of $R(t, \varepsilon; \varphi)$ for fixed t or fixed ε, as well as the double limits for both $\lim_{\varepsilon \to 0^+} \lim_{t \to +\infty}$

and $\lim_{t \to +\infty} \lim_{\varepsilon \to 0^+}$. The main conclusion is the convergence of the response function $R(t, \varepsilon; \varphi)$ in terms of the small parameter for the external perturbation $\varepsilon \leq \varepsilon_0$, uniformly for $t \in [0, +\infty)$. In Sect. 3.1, we first give some preparations on the perturbed invariant measure including the hypoellipticity and the exponential convergence of the Fokker-Planck equation. Then we study the linear response theory for general external force in Sect. 3.2 and the Green-Kubo relation for conservative external force in Sect. 3.3.

3.1 Invariant Measure and the Exponential Convergence for Irreversible Perturbation

In this subsection, we consider an irreversible perturbation in a form of general external force $\varepsilon\mathbf{M}$. Before studying the linear response theory with respect to the external force $\varepsilon\mathbf{M}$ in Sect. 3.2, we first prepare some preliminary results, including estimates on the perturbed invariant measure and the well-posedness, hypoellipticity and exponential convergence of the Fokker-Planck equation.

Consider the following reversible overdamped Langevin system at equilibrium

$$d\mathbf{q}_t = -\boldsymbol{\sigma}\boldsymbol{\sigma}^T \nabla V(\mathbf{q})dt + \sqrt{2}\boldsymbol{\sigma} d\mathbf{B}, \quad \mathbf{q}_{t=0} \sim \rho_0(\mathbf{q}) := \frac{1}{Z} e^{-V(\mathbf{q})}, \quad Z := \int_{\mathbb{R}^d} e^{-V(\mathbf{q})} d\mathbf{q}. \tag{3.2}$$

Here, $\boldsymbol{\sigma} \in \mathbb{R}^{d \times d}$ is constant and nonsingular.

Let $\mathbf{M} \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$. For any $\varepsilon > 0$, suppose that at time $t = 0$, an external force $\varepsilon\mathbf{M}$ is added to the system, which yields the following perturbed SDE

$$d\mathbf{q}_t^\varepsilon = \boldsymbol{\sigma}\boldsymbol{\sigma}^T(-\nabla V(\mathbf{q}^\varepsilon) + \varepsilon\mathbf{M}(\mathbf{q}^\varepsilon))dt + \sqrt{2}\boldsymbol{\sigma} d\mathbf{B}, \tag{3.3}$$

with the initial distribution $\mathbf{q}_{t=0}^\varepsilon \sim \rho_0$. Notice that the initial data for the perturbed SDE is taken as the equilibrium ρ_0 for the unperturbed SDE.

3.1.1 Invariant Measure: Existence, Uniqueness, and Positivity

In this subsection, we clarify some known results on the existence, uniqueness, and positivity of the invariant measure, i.e., the stationary solution to the corresponding Fokker-Planck equation. For simplicity in notation related to the Fokker-Planck equation, we will use $2\boldsymbol{\sigma}\boldsymbol{\sigma}^T$ as the variance.

From this point on, we impose the following assumptions on the potential $V(\mathbf{x})$

Assumption (I) There exists $\alpha > 0$ such that

$$|\nabla V(\mathbf{x})| \leq C_1 |\mathbf{x}|^\alpha + C_2, \quad \limsup_{|\mathbf{x}| \to \infty} \frac{-\mathbf{x} \cdot \nabla V(\mathbf{x})}{|\mathbf{x}|^{\alpha+1}} =: \gamma_1 < 0. \tag{3.4}$$

Assumption (II) There exists $\lambda > 0$ such that

$$\liminf_{|\mathbf{x}| \to \infty} \left(|\boldsymbol{\sigma}^T \nabla V(\mathbf{x})|^2 - 2\text{tr}(\boldsymbol{\sigma}\boldsymbol{\sigma}^T \nabla^2 V(\mathbf{x}))\right) =: \lambda > 0. \tag{3.5}$$

Assumption (I) ensures the uniqueness of the invariant measure and provides a decay estimate at infinity [2, Theorem 3.4.3]. According to [6], Assumption (II) ensures that the measure $e^{-V(\mathbf{x})}$ satisfies Poincaré's inequality.

Notice that (3.3) is irreversible, and the Gibbs measure $\rho_0(\mathbf{q})$ is no longer the invariant measure. In fact, even the existence of an invariant measure for an irreversible system is not trivial.

Under Assumption (I), the potential is a trapping potential with superlinear growth, which guarantees that both equations (3.2) and (3.3) admit a unique invariant measure with density functions ρ_0 and ρ_∞^ε, respectively. In Lemma 3.1, we establish the existence, uniqueness, and far-field decay estimates for the invariant measure under Assumption (I). Furthermore, if exponential convergence to the invariant measure over long times is considered, we will also require Assumption (II). In Proposition 3.1, we recall the result of exponential convergence for the invariant measure ρ_∞^ε in $L^1(\mathbb{R}^d)$, which is crucial for obtaining the linear response theory.

For the underdamped case, we restrict our study to perturbations in a potential form. Assumptions (I) and (II) are also essential for the well-posedness of the invariant measure and the exponential convergence of the perturbed Fokker-Planck equation in $L^2(1/\rho_\infty^\varepsilon)$; see Sect. 4.

Lemma 3.1 (*Existence, uniqueness, and positivity of the invariant measure*) *Let* $V(\mathbf{q}) \in C^\infty(\mathbb{R}^d; \mathbb{R})$ *satisfy Assumption* (I), *and let* $\mathbf{M} \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$. *For all* $\varepsilon \in [0, 1)$, *consider the SDE* (3.3). *Then it admits a unique invariant measure with density* ρ_∞^ε *that satisfies*

$$e^{-K_1(|\mathbf{q}|^{\alpha+1}+1)} \le \rho_\infty^\varepsilon \le e^{-K_2(|\mathbf{q}|^{\alpha+1}+1)}, \tag{3.6}$$

where constants $K_1, K_2 > 0$ *are uniform in* ε.

The existence and estimate (3.6) can be found in [2, Theorem 3.4.3], and uniqueness is ensured by the following lemma from [2, Theorem 4.1.6].

Lemma 3.2 (Theorem 4.1.6 in [2]) *Suppose that*

$$\mathbf{b} \in C^\infty(\mathbb{R}^d; \mathbb{R}^d), \quad \boldsymbol{\sigma} \in C^\infty(\mathbb{R}^d; \mathbb{R}^{d \times d}), \quad \boldsymbol{\sigma}\boldsymbol{\sigma}^T \text{ is strictly elliptic}. \tag{3.7}$$

Let $\rho \in C^\infty(\mathbb{R}^d)$ *solve* $\mathcal{L}^*\rho = 0$, *where* $\mathcal{L}^*$ *is defined in* (2.5). *Assume that* $\rho > 0$ *and* $\rho \in L^1(\mathbb{R}^d)$. *If*

$$\frac{\|\boldsymbol{\sigma}\boldsymbol{\sigma}^T\|}{1 + \|\mathbf{x}\|^2} \in L^1(\rho d\mathbf{x}), \quad \frac{\|\mathbf{b}\|}{1 + \|\mathbf{x}\|} \in L^1(\rho d\mathbf{x}) \tag{3.8}$$

holds, then ρ *is the unique non-zero solution of* $\mathcal{L}^*u = 0$ *such that* $u \in L^1(\mathbb{R}^d)$ *and* $u \ge 0$.

3.1.2 Well-Posedness of the Fokker-Planck Equation

Suppose the density of $\mathbf{q}_t^\varepsilon$ is given by $\rho^\varepsilon(\mathbf{q}, t)$. Then $\rho^\varepsilon(\mathbf{q}, t)$ satisfies the following Fokker-Planck equation

$$\frac{\partial \rho^\varepsilon(\mathbf{q}, t)}{\partial t} = \mathcal{L}_\varepsilon^* \rho^\varepsilon(\mathbf{q}, t) := \nabla \cdot \left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T \left((\nabla V(\mathbf{q}) - \varepsilon \mathbf{M}(\mathbf{q}))\rho^\varepsilon(\mathbf{q}, t) + \nabla \rho^\varepsilon(\mathbf{q}, t) \right) \right),$$
$$\rho^\varepsilon(\mathbf{q}, 0) = \rho_0(\mathbf{q}). \tag{3.9}$$

In fact, $\mathcal{L}_\varepsilon^*$ generates a strongly continuous semigroup of contractions, which ensures the well-posedness of (2.5).

Lemma 3.3 *Let $V(\mathbf{q}) \in C^\infty(\mathbb{R}^d; \mathbb{R})$ satisfy Assumption (I), and let $\mathbf{M} \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$. Consider (3.9). Then*

(i) $\mathcal{L}_\varepsilon^*$ *in (3.9) generates a strongly continuous semigroup of contractions in $L^2(1/\rho_\infty^\varepsilon)$;*

(ii) *(3.9) admits a unique solution $\rho^\varepsilon(\mathbf{q}, t) \in C^1([0, T], L^2(1/\rho_\infty^\varepsilon))$ for any $T > 0$.*

The proof of Lemma 3.3 is provided in Appendix A. We denote the semigroup generated by $\mathcal{L}_\varepsilon^*$ as $\left\{ e^{t\mathcal{L}_\varepsilon^*} \right\}_{t \geq 0}$.

3.1.3 Hypoellipticity

Hypoellipticity implies the smoothness of the solution to the Fokker-Planck equation, allowing us to perform integration by parts in subsequent proofs without concern. We prove hypoellipticity by applying Hörmander's celebrated result [14, Theorem 1.1].

Recall that a linear differential operator P with C^∞ coefficients in $\mathbb{R}^d$ (or an open subset of $\mathbb{R}^d$) is called hypoelliptic if for every distribution u in $\mathcal{D}'(\mathbb{R}^d)$, we have

$$\text{sing supp } u = \text{sing supp } Pu.$$

Here, $\text{sing supp } u$ denotes the singular support of u

$$\text{sing supp } u = \mathbb{R}^d \setminus \left\{ x \in \mathbb{R}^d : u \text{ is smooth near } x \right\}.$$

Consider a linear differential operator P with C^∞ coefficients, which can be written as:

$$P = \sum_{j=1}^{r} \mathbf{X}_j^2 + \mathbf{X}_0 + c,$$

where $\mathbf{X}_0, \dots, \mathbf{X}_r$ denote first-order homogeneous differential operators in $\mathbb{R}^d \times (0, \infty)$ with smooth coefficients, and $c \in C^\infty(\mathbb{R}^d \times (0, \infty))$. For example, the heat operator $\Delta - \dfrac{\partial}{\partial t}$ in $\mathbb{R}^d \times (0, \infty)$ can be recast as

$$\Delta - \frac{\partial}{\partial t} = \sum_{i=1}^{d} (\mathbf{X}_i)^2 + \mathbf{X}_0,$$

where

$$\mathbf{X}_i = \frac{\partial}{\partial x_i}, \quad i = 1, 2, \dots, d, \quad \mathbf{X}_0 = -\frac{\partial}{\partial t}.$$

Hörmander's theorem [14, Theorem 1.1] relates the Lie algebra generated by $\mathbf{X}_i$, $i = 0, 1, 2, \dots, r$ to the hypoellipticity of P.

Theorem 3.1 (Hypoellipticity, Theorem 1.1 in [14]) *Consider*

$$P = \sum_{j=1}^{r} \mathbf{X}_j^2 + \mathbf{X}_0 + c,$$

where $\mathbf{X}_0, \ldots, \mathbf{X}_r$ denote first-order homogeneous differential operators in $\mathbb{R}^d \times (0, \infty)$ with smooth coefficients, and $c \in C^\infty(\mathbb{R}^d \times (0, \infty))$. If, at any given point $(\mathbf{q}, t) \in \mathbb{R}^d \times (0, \infty)$,

$$\text{span}\left\{ \mathbf{X}_{j_1}, [\mathbf{X}_{j_1}, \mathbf{X}_{j_2}], [\mathbf{X}_{j_1}, [\mathbf{X}_{j_2}, \mathbf{X}_{j_3}]], \ldots, [\mathbf{X}_{j_1}, [\mathbf{X}_{j_2}, [\mathbf{X}_{j_3}, \ldots, \mathbf{X}_{j_k}]]] \right\} = \mathbb{R}^{d+1},$$

where $j_i = 0, 1, 2, \ldots, r$, then P is hypoelliptic.

Using Theorem 3.1, we obtain the smoothness of $\rho^\varepsilon(\mathbf{q}, t)$.

Lemma 3.4 *Let $\rho^\varepsilon(\mathbf{q}, t)$ be the unique solution of* (3.9). *Then it is smooth in $\mathbb{R}^d \times (0, \infty)$.*

The hypoellipticity for the Fokker-Planck equation corresponding to overdamped Langevin dynamics is trivial. However, we include the proof using Hörmander's hypoellipticity theorem in Appendix B to facilitate comparison with the underdamped case.

3.1.4 Exponential Convergence

The convergence of $\rho^\varepsilon(\mathbf{q}, t)$ to $\rho_\infty^\varepsilon(\mathbf{q})$ in $L^2(1/\rho_\infty^\varepsilon)$ cannot be derived by Poincare's inequality because (3.3) is irreversible. However, we still have exponential convergence in total variation, as shown in [15, Theorem B].

Proposition 3.1 (Convergence in $L^1(\mathbb{R}^d)$) *Suppose Assumptions (I) and (II) hold. For any $\varepsilon \in (0, 1)$, consider* (3.9). *Then there exist constants $\varepsilon_0, C > 0$, and $r > 0$ that depend only on V, $\mathbf{M}$, and $\boldsymbol{\sigma}$, such that for all $\varepsilon \in [0, \varepsilon_0)$, we have*

$$\|\rho^\varepsilon(\cdot, t) - \rho_\infty^\varepsilon\|_{L^1(\mathbb{R}^d)} \le Ce^{-rt}.$$

The proof of Proposition 3.1 employs a version of Harris's theorem by Hairer and Majda [10]. For completeness, the proof is provided in Appendix B.

3.2 Linear Response Theory (LRT)

Recall (3.1) and perturbed SDE (3.3). We will study the behaviors of the response function (3.1). Let $\widetilde{\rho}^\varepsilon(\mathbf{q}, t) = \rho^\varepsilon(\mathbf{q}, t) - \rho_0$. Then

$$R(t, \varepsilon; \varphi) := \frac{1}{\varepsilon} \int_{\mathbb{R}^d} \varphi(\mathbf{q}) \widetilde{\rho}^\varepsilon(\mathbf{q}, t) \mathrm{d}\mathbf{q}.$$

The function $\widetilde{\rho}^\varepsilon(\mathbf{q}, t)$ is smooth and satisfies

$$\frac{\partial \widetilde{\rho}^\varepsilon(\mathbf{q}, t)}{\partial t} = \mathcal{L}_\varepsilon^* \widetilde{\rho}^\varepsilon(\mathbf{q}, t) + \varepsilon \nabla \cdot (\rho_0 \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M}), \quad \widetilde{\rho}^\varepsilon(\mathbf{q}, 0) = 0. \tag{3.10}$$

Using the dissipative property of the semigroup, we can derive the following estimate for $\widetilde{\rho}^\varepsilon(\mathbf{q}, t)$.

Lemma 3.5 *Consider $\widetilde{\rho}^\varepsilon(\mathbf{q}, t)$ in* (3.10). *Then*

(i) *There exists a constant $C > 0$, which depends only on $\mathbf{M}$ and V, such that*

$$\|\widetilde{\rho}^\varepsilon(\cdot, t)\|_{L^1} \le C\varepsilon, \tag{3.11}$$

for all $t \ge 0$.

(ii) *There exists a constant $C > 0$, which depends only on $\mathbf{M}$ and V, such that*

$$\|\widetilde{\rho^\varepsilon}(\cdot, t)\|_{L^2(1/\rho^\varepsilon_\infty)} \leq C\varepsilon t, \tag{3.12}$$

for all $t \geq 0$. Furthermore, if $\mathbf{M} = \nabla W$ is in gradient form, we also have

$$\|\widetilde{\rho^\varepsilon}(\cdot, t)\|_{L^2(1/\rho^\varepsilon_\infty)} \leq C\varepsilon. \tag{3.13}$$

Proof Since $\mathbf{M} \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$, by Duhamel's principle, we have

$$\widetilde{\rho^\varepsilon}(\mathbf{q}, t) = \varepsilon \int_0^t e^{(t-s)\mathcal{L}^*_\varepsilon} \left(\nabla \cdot (\rho_0 \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M})\right) \mathrm{d}s. \tag{3.14}$$

Here, $e^{s\mathcal{L}^*_\varepsilon}\varphi$ represents the solution to (3.9) at time s with initial value φ.

First, since $\mu := \nabla \cdot (\rho_0 \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M})$ has zero Lebesgue integral, the positive part μ^+ and negative part μ^- satisfy $\mu = \mu^+ + \mu^-$, and $\int \mu^+(x)\mathrm{d}x = \int \mu^-(x)\mathrm{d}x =: c_0$. Therefore, the exponential convergence in L^1 norm from Proposition 3.1 implies

$$\varepsilon \int_0^t \int |e^{(t-s)\mathcal{L}^*_\varepsilon}(\nabla \cdot (\rho_0 \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M}))|\mathrm{d}x\mathrm{d}s$$

$$= \varepsilon \int_0^t \int |e^{(t-s)\mathcal{L}^*_\varepsilon}(\mu^+ - c_0\rho^\varepsilon_\infty + c_0\rho^\varepsilon_\infty - \mu^-)|\mathrm{d}x\mathrm{d}s$$

$$= \varepsilon \int_0^t \int |e^{(t-s)\mathcal{L}^*_\varepsilon}\mu^+ - c_0\rho^\varepsilon_\infty|\mathrm{d}x\mathrm{d}s + \varepsilon \int_0^t \int |e^{(t-s)\mathcal{L}^*_\varepsilon}\mu^- - c_0\rho^\varepsilon_\infty|\mathrm{d}x\mathrm{d}s$$

$$\leq c\varepsilon \int_0^t e^{-r(t-s)}\mathrm{d}s = \frac{c\varepsilon}{r}(1 - e^{-rt}).$$

Second, since $\nabla \cdot (\rho_0 \mathbf{M}) \in D(\mathcal{L}^*_\varepsilon)$, by the contraction property of $\mathcal{L}^*_\varepsilon$ from Lemma 3.3, we have

$$\|\widetilde{\rho^\varepsilon}(\cdot, t)\|_{L^2(1/\rho^\varepsilon_\infty)} = \varepsilon \int_0^t \|e^{(t-s)\mathcal{L}^*_\varepsilon}(\nabla \cdot (\rho_0 \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M}))\|_{L^2(1/\rho^\varepsilon_\infty)}\mathrm{d}s$$

$$\leq \varepsilon t \|\nabla \cdot (\rho_0 \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M})\|_{L^2(1/\rho^\varepsilon_\infty)}.$$

From (3.6), we know that $\|\nabla \cdot (\rho_0 \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M})\|_{L^2(1/\rho^\varepsilon_\infty)}$ can be uniformly bounded since $\mathbf{M}$ is compactly supported. This proves (3.12).

Furthermore, the exponential convergence in $L^2(1/\rho^\varepsilon_\infty)$ norm implies

$$\varepsilon \int_0^t \int |e^{(t-s)\mathcal{L}^*_\varepsilon}(\nabla \cdot (\rho_0 \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M}))|^2/\rho^\varepsilon_\infty\mathrm{d}x\mathrm{d}s$$

$$= \varepsilon \int_0^t \int |e^{(t-s)\mathcal{L}^*_\varepsilon}(\mu^+ - c_0\rho^\varepsilon_\infty + c_0\rho^\varepsilon_\infty - \mu^-)|^2/\rho^\varepsilon_\infty\mathrm{d}x\mathrm{d}s$$

$$\leq 2\varepsilon \int_0^t \int |e^{(t-s)\mathcal{L}^*_\varepsilon}\mu^+ - c_0\rho^\varepsilon_\infty|^2/\rho^\varepsilon_\infty\mathrm{d}x\mathrm{d}s + 2\varepsilon \int_0^t \int |e^{(t-s)\mathcal{L}^*_\varepsilon}\mu^- - c_0\rho^\varepsilon_\infty|^2/\rho^\varepsilon_\infty\mathrm{d}x\mathrm{d}s$$

$$\leq c\varepsilon \int_0^t e^{-r(t-s)}\mathrm{d}s = \frac{c\varepsilon}{r}(1 - e^{-rt}). \qquad \square$$

Denote the Fokker-Planck operator for the reversible part as

$$\mathcal{L}_0^* \rho := \nabla \cdot (\boldsymbol{\sigma}\boldsymbol{\sigma}^T(\rho\nabla V + \nabla\rho)) = \nabla \cdot (\rho_0 \boldsymbol{\sigma}\boldsymbol{\sigma}^T \nabla \frac{\rho}{\rho_0}). \tag{3.15}$$

Then Fokker-Planck equation (3.9) which could be reformulated as

$$\frac{\partial \rho^\varepsilon(\mathbf{q}, t)}{\partial t} = \mathcal{L}_\varepsilon^* \rho^\varepsilon(\mathbf{q}, t) = \mathcal{L}_0^* \rho^\varepsilon(\mathbf{q}, t) - \varepsilon \nabla \cdot (\rho^\varepsilon \boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M}).$$

We now consider the limit behavior of $R(t, \varepsilon; \varphi)$.

Theorem 3.2 *Suppose that Assumptions (I) and (II) hold. Let $\mathbf{M}(\mathbf{q}) \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$, and let $\rho^\varepsilon(\mathbf{q}, t)$ be the law of $\mathbf{q}_t^\varepsilon$ in (3.3). Let $\rho_\infty^\varepsilon(\mathbf{q})$ be the invariant measure of (3.3). For some $\varphi \in C_c^\infty(\mathbb{R}^d)$, consider $R(t, \varepsilon; \varphi)$ defined in (3.1). Then*

(i) *(Convergence as $\varepsilon \to 0^+$) For any given $t > 0$,*

$$\lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) = \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot (\nabla(e^{s\mathcal{L}_0}\varphi)) \right] \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}s. \tag{3.16}$$

Moreover, the limit in (3.16) holds uniformly for all $t > 0$, i.e., for any $\eta > 0$, there exists ε_0 such that for any $0 < \varepsilon < \varepsilon_0$ and $t > 0$

$$\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot (\nabla(e^{s\mathcal{L}_0}\varphi)) \right] \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}s \right| < \eta.$$

(ii) *(Convergence as $\varepsilon \to 0^+$, then $t \to \infty$) The following limit exists*

$$\lim_{t \to \infty} \lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) = \int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot (\nabla(e^{s\mathcal{L}_0}\varphi)) \right] \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}s, \tag{3.17}$$

and the convergence in t is exponentially fast, i.e., there exist constants $C > 0$ and $r > 0$, which depend on $\mathbf{M}$ and V, such that

$$\left| \lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) - \int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot (\nabla(e^{s\mathcal{L}_0}\varphi)) \right] \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}s \right| \le C e^{-rt}, \tag{3.18}$$

holds for all $t > 0$.

(iii) *(Convergence as $t \to \infty$) For any $\varepsilon > 0$,*

$$\lim_{t \to \infty} R(t, \varepsilon; \varphi) = \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_\infty^\varepsilon(\mathbf{q}) \mathrm{d}\mathbf{q} - \int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_0(\mathbf{q}) \mathrm{d}\mathbf{q} \right). \tag{3.19}$$

(iv) *(Convergence as $t \to \infty$, then $\varepsilon \to 0^+$) The following limit exists*

$$\lim_{\varepsilon \to 0^+} \lim_{t \to \infty} R(t, \varepsilon; \varphi) = \int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot (\nabla(e^{s\mathcal{L}_0}\varphi)) \right] \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}s, \tag{3.20}$$

or equivalently

$$\lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_\infty^\varepsilon(\mathbf{q}) \mathrm{d}\mathbf{q} - \int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_0(\mathbf{q}) \mathrm{d}\mathbf{q} \right)$$

$$= \int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot (\nabla(e^{s\mathcal{L}_0}\varphi)) \right] \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}s. \tag{3.21}$$

Before proving the theorem, we provide a necessary estimate that will be frequently used later.

Lemma 3.6 *Under Assumption (II), there exist constants $C > 0$ and $r > 0$ that depend on φ, $\mathbf{M}$, and V, such that for any $\varphi \in C_c^\infty(\mathbb{R}^d)$, we have*

$$\|\nabla(e^{t\mathcal{L}_0}\varphi)\|_{L^2(\rho_0)} \leq Ce^{-rt}, \quad \|\nabla(e^{t\mathcal{L}_0}\varphi)\|_{L^\infty(\mathrm{supp}(\mathbf{M}))} \leq Ce^{-rt}. \tag{3.22}$$

This lemma ensures that the integral

$$\int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot (\nabla(e^{s\mathcal{L}_0}\varphi)) \right] \rho_0 \, \mathrm{d}\mathbf{q} \, \mathrm{d}s$$

appearing in Theorem 3.2 converges

$$\left| \int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot (\nabla(e^{s\mathcal{L}_0}\varphi)) \right] \rho_0 \, \mathrm{d}\mathbf{q} \, \mathrm{d}s \right| \leq C \int_0^\infty \|\nabla(e^{t\mathcal{L}_0}\varphi)\|_{L^\infty(\mathrm{supp}(\mathbf{M}))} \, \mathrm{d}s$$

$$\leq C' \int_0^\infty e^{-rs} \, \mathrm{d}s < \infty,$$

where C and C' are constants.

Proof of Lemma 3.6 Define $\bar{\varphi} := \int_{\mathbb{R}^d} \varphi \rho_0 \, \mathrm{d}\mathbf{q}$. Then, we know that $\psi(\mathbf{q}, t) := e^{t\mathcal{L}_0}(\varphi - \bar{\varphi})$ solves the backward equation

$$\frac{\partial \psi(\mathbf{q}, t)}{\partial t} = \mathcal{L}_0 \psi(\mathbf{q}, t), \quad \psi(\mathbf{q}, 0) = \varphi(\mathbf{q}) - \bar{\varphi}. \tag{3.23}$$

Notice that $\int_{\mathbb{R}^d} \psi(\mathbf{q}, 0)\rho_0 \, \mathrm{d}\mathbf{q} = 0$. Multiplying (3.23) by ρ_0 and integrating in $[0, t] \times \mathbb{R}^d$, we have for any $t \geq 0$,

$$\int_{\mathbb{R}^d} \psi(\mathbf{q}, t)\rho_0 \, \mathrm{d}\mathbf{q} = 0. \tag{3.24}$$

By Assumption (II), Poincare's inequality holds, and we have exponential convergence

$$\|\psi(\cdot, t)\|_{L^2(\rho_0)} \leq Ce^{-rt}. \tag{3.25}$$

Furthermore, let $\eta := \mathcal{L}_0 \psi(\mathbf{q}, t) = \mathcal{L}_0 e^{t\mathcal{L}_0}(\varphi - \bar{\varphi}) = e^{t\mathcal{L}_0}\mathcal{L}_0\varphi$. Then η solves

$$\frac{\partial \eta(\mathbf{q}, t)}{\partial t} = \mathcal{L}_0 \eta(\mathbf{q}, t), \quad \eta(\mathbf{q}, 0) = \mathcal{L}_0\varphi(\mathbf{q}).$$

Since $\int_{\mathbb{R}^d} \mathcal{L}_0\varphi\rho_0 \, \mathrm{d}\mathbf{q} = 0$, we again have exponential convergence

$$\|\mathcal{L}_0 \psi(\cdot, t)\|_{L^2(\rho_0)} \leq Ce^{-rt}.$$

Using the definition of $\mathcal{L}_0$, we compute

$$\int_{\mathbb{R}^d} \rho_0 |\boldsymbol{\sigma}^T \nabla(e^{t\mathcal{L}_0}\varphi)|^2 \, \mathrm{d}\mathbf{q} = -\int_{\mathbb{R}^d} \rho_0 \psi(\mathbf{q}, t)\mathcal{L}_0 \psi(\mathbf{q}, t) \, \mathrm{d}\mathbf{q}$$

$$\leq \|\psi(\cdot, t)\|_{L^2(\rho_0)} \|\mathcal{L}_0 \psi(\cdot, t)\|_{L^2(\rho_0)}$$

$$\leq Ce^{-rt}.$$

Since $\boldsymbol{\sigma}$ is nonsingular and constant, we have

$$\|\nabla(e^{t\mathcal{L}_0}\varphi)\|_{L^2(\rho_0)} \leq Ce^{-rt}.$$

Next, we estimate the L^∞ norm. We first observe

$$\|\psi(\cdot, t)\|_{L^\infty(\mathbb{R}^d)} \leq \|\psi(\cdot, 0)\|_{L^\infty(\mathbb{R}^d)}, \quad \|\mathcal{L}_0\psi(\cdot, t)\|_{L^\infty(\mathbb{R}^d)} \leq \|\mathcal{L}_0\psi(\cdot, 0)\|_{L^\infty(\mathbb{R}^d)}.$$

Then, by the interior estimate from [9, Theorem 9.11], for any $1 < p < \infty$, we have

$$\|\psi(\cdot, t)\|_{W^{2,p}(\text{supp}(\mathbf{M}))} \leq C\left[\|\psi(\cdot, t)\|_{L^p(B)} + \|\mathcal{L}\psi(\cdot, t)\|_{L^p(B)}\right]. \tag{3.26}$$

Here, B is a compact ball such that $\text{supp}(\mathbf{M}) \subset\subset B$. Since $\rho_0(\mathbf{q}) = e^{-V(\mathbf{q})} \geq \frac{1}{c} > 0$ for $\mathbf{q} \in B$, for some $p > d$, we have

$$\|\psi(\cdot, t)\|_{L^p(B)}^p \leq \|\psi(\cdot, t)\|_{L^\infty(\mathbb{R}^d)}^{p-2} \cdot \int_B c\rho_0|\psi(\mathbf{q}, t)|^2\,\mathrm{d}\mathbf{q} \leq c\|\psi(\cdot, 0)\|_{L^\infty(\mathbb{R}^d)}^{p-2}\|\psi(\cdot, t)\|_{L^2(\rho_0)}^2,$$

$$\|\mathcal{L}_0\psi(\cdot, t)\|_{L^p(B)}^p \leq \|\mathcal{L}_0\psi(\cdot, t)\|_{L^\infty(\mathbb{R}^d)}^{p-2} \cdot \int_B c\rho_0|\mathcal{L}_0\psi(\mathbf{q}, t)|^2\,\mathrm{d}\mathbf{q}$$

$$\leq c\|\mathcal{L}_0\psi(\cdot, 0)\|_{L^\infty(\mathbb{R}^d)}^{p-2}\|\mathcal{L}_0\psi(\cdot, t)\|_{L^2(\rho_0)}^2.$$

Combining this with (3.26), we obtain

$$\|\psi(\cdot, t)\|_{W^{2,p}(\text{supp}(\mathbf{M}))} \leq Ce^{-rt}.$$

By the embedding $W^{1,p}(B) \subset L^\infty(B)$, we conclude

$$\|\nabla(e^{t\mathcal{L}_0}\varphi)\|_{L^\infty(\text{supp}(\mathbf{M}))} \leq \|\nabla(e^{t\mathcal{L}_0}\varphi)\|_{W^{1,p}(\text{supp}(\mathbf{M}))} \leq \|\psi(\cdot, t)\|_{W^{2,p}(\text{supp}(\mathbf{M}))} \leq Ce^{-rt}.$$

This completes the proof of the lemma. $\qquad\square$

Now we can proceed to prove Theorem 3.2.

Proof of Theorem 3.2 (i) This is the key step of the proof, and parts (ii) $\sim$ (iv) will follow from it. Recall that $\rho^\varepsilon(\mathbf{q}, t)$ solves the Fokker-Planck equation (3.9), which can be reformulated as

$$\frac{\partial \rho^\varepsilon(\mathbf{q}, t)}{\partial t} = \mathcal{L}_\varepsilon^*\rho^\varepsilon(\mathbf{q}, t) = \mathcal{L}_0^*\rho^\varepsilon(\mathbf{q}, t) - \varepsilon\nabla \cdot (\rho^\varepsilon \boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M}).$$

Thus, by Duhamel's principle, we have

$$\rho^\varepsilon(\mathbf{q}, t) = e^{t\mathcal{L}_0^*}\rho_0 - \varepsilon\int_0^t e^{(t-s)\mathcal{L}_0^*}\left[\nabla \cdot \left(\rho^\varepsilon(\mathbf{q}, s)\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M}(\mathbf{q})\right)\right]\mathrm{d}s. \tag{3.27}$$

Substituting (3.27) into (3.1), we rewrite the response function as

$$R(\varepsilon, t; \varphi) = -\int_{\mathbb{R}^d}\int_0^t \varphi(\mathbf{q})e^{(t-s)\mathcal{L}_0^*}\left[\nabla \cdot \left(\rho^\varepsilon(\mathbf{q}, s)\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M}(\mathbf{q})\right)\right]\mathrm{d}s\,\mathrm{d}\mathbf{q}.$$

Since $\varphi \in C_c^\infty(\mathbb{R}^d)$, the term $\varphi(\mathbf{q})e^{(t-s)\mathcal{L}_0^*}\left[\nabla \cdot (\rho^\varepsilon(\mathbf{q}, s)\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M}(\mathbf{q}))\right]$ is smooth and has compact support in $\mathbb{R}^d \times [0, t]$, ensuring it is bounded in this domain. By Fubini's theorem, we get

$$R(\varepsilon, t; \varphi) = -\int_0^t \int_{\mathbb{R}^d} \varphi(\mathbf{q})e^{(t-s)\mathcal{L}_0^*}\left[\nabla \cdot (\rho^\varepsilon(\mathbf{q}, s)\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M}(\mathbf{q}))\right]d\mathbf{q}\,ds.$$

Integration by parts gives

$$R(\varepsilon, t; \varphi) = \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\right] \rho^\varepsilon(\mathbf{q}, s)d\mathbf{q}\,ds. \tag{3.28}$$

Now, for a fixed time $T_0 > 0$ (to be chosen later), for $t > T_0$, we have

$$\left| R(\varepsilon, t; \varphi) - \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\right] \rho_0(\mathbf{q}, s)d\mathbf{q}\,ds \right|$$

$$= \left| \int_0^{T_0} \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\right] (\rho^\varepsilon(\mathbf{q}, s) - \rho_0)d\mathbf{q}\,ds \right.$$

$$\left. + \int_{T_0}^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\right] (\rho^\varepsilon(\mathbf{q}, s) - \rho_0)d\mathbf{q}\,ds \right|.$$

This can be bounded by the sum of three terms

$$\mathrm{I} = \int_0^{T_0} \|\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\|_{L^2(\rho_0)} \left\| \frac{\widetilde{\rho^\varepsilon}(\cdot, s)}{\rho_0} \right\|_{L^2(\rho_0,\mathrm{supp}(\mathbf{M}))} ds,$$

$$\mathrm{II} = \left| \int_{T_0}^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\right] (\rho^\varepsilon(\mathbf{q}, s) - \rho_\infty^\varepsilon(\mathbf{q}))d\mathbf{q}\,ds \right|,$$

$$\mathrm{III} = \left| \int_{T_0}^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\right] (\rho_0(\mathbf{q}) - \rho_\infty^\varepsilon(\mathbf{q}))d\mathbf{q}\,ds \right|.$$

By Lemma 3.5 and Lemma 3.6, we know

$$\mathrm{I} \le C\varepsilon \int_0^{T_0} e^{(s-t)r}s\,ds = \varepsilon C_1(T_0 + 1)e^{r(T_0-t)}, \tag{3.29}$$

where C_1 is a constant depending on $\mathbf{M}$, φ, $\boldsymbol{\sigma}$, and V. By Proposition 3.1 and Lemma 3.6, we know

$$\mathrm{II} \le \int_{T_0}^t \|\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\|_{L^\infty(\mathrm{supp}(\mathbf{M}))}\|\rho^\varepsilon(\mathbf{q}, s) - \rho_\infty^\varepsilon(\mathbf{q})\|_{L^1(\mathbb{R}^d)}ds$$

$$\le C' \int_{T_0}^t e^{r(s-t)} \cdot e^{-r_1 s}ds \tag{3.30}$$

$$\le C_2 e^{-r_1 T_0}.$$

where C' and C_2 are constants and $r_1 > 0$ depends on $\mathbf{M}$, φ, $\boldsymbol{\sigma}$, and V.

Finally, by taking $W(\mathbf{q}) = |\mathbf{q}|^2$ in [2, Proposition 3.7.4], ρ_∞^ε converges to ρ_0 as $\varepsilon \to 0^+$ in $L^1(\mathbb{R}^d)$. Thus, by Lemma 3.6, as $\varepsilon \to 0^+$, we have

$$
\begin{aligned}
\text{III} &\leq \|\rho_\infty^\varepsilon - \rho_0\|_{L^1(\mathbb{R}^d)} \cdot \int_{T_0}^t \|\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi)\|_{L^\infty(\text{supp}(\mathbf{M}))} \mathrm{d}s \\
&\leq C'' \|\rho_\infty^\varepsilon - \rho_0\|_{L^1(\mathbb{R}^d)} \cdot \int_{T_0}^t e^{r(s-t)} \mathrm{d}s \\
&\leq C_3 \|\rho_\infty^\varepsilon - \rho_0\|_{L^1(\mathbb{R}^d)} \to 0.
\end{aligned}
\tag{3.31}
$$

Here, C_3 depends on $\mathbf{M}$, φ, $\boldsymbol{\sigma}$, and V.

To control the error within any $\delta > 0$, take T_0 sufficiently large such that $\text{II} \leq C_2 e^{-r_1 T_0} \leq \delta/3$. Then, choose ε_0 so that $\text{I} \leq C_1 \varepsilon_0(T_0 + 1) \leq \delta/3$ for any $\varepsilon \leq \varepsilon_0$. Finally, select $\varepsilon_1 \leq \varepsilon_0$ so that $\text{III} \leq C_3 \|\rho_\infty^\varepsilon - \rho_0\|_{L^1(\mathbb{R}^d)} \leq \delta/3$ for all $\varepsilon \leq \varepsilon_1$. Thus, for any $\varepsilon < \varepsilon_1$ and $t > T_0$, by (3.29), (3.30), and (3.31), we have

$$
\left| R(\varepsilon, t; \varphi) - \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi) \right] \rho_0(\mathbf{q}) \mathrm{d}\mathbf{q}\, \mathrm{d}s \right| \leq \text{I} + \text{II} + \text{III} \leq \delta.
\tag{3.32}
$$

For $t \in (0, T_0]$, use (3.29) to bound the difference

$$
\left| R(\varepsilon, t; \varphi) - \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi) \right] \rho_0(\mathbf{q}) \mathrm{d}\mathbf{q}\, \mathrm{d}s \right| \leq \text{I} \leq \delta/3.
\tag{3.33}
$$

Thus, the convergence is uniform in t, and by a change of variables, we obtain

$$
\int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{(t-s)\mathcal{L}_0}\varphi) \right] \rho_0(\mathbf{q}) \mathrm{d}\mathbf{q}\, \mathrm{d}s = \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi) \right] \rho_0(\mathbf{q}) \mathrm{d}\mathbf{q}\, \mathrm{d}s.
\tag{3.34}
$$

This completes the proof of (i).

Next, we prove (ii). By Lemma 3.6, we know that for $T_2 > T_1 > 0$

$$
\begin{aligned}
\int_{T_1}^{T_2} \int_{\mathbb{R}^d} [\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi)]\rho_0(\mathbf{q})\mathrm{d}\mathbf{q}\mathrm{d}s \\
\leq \int_{T_1}^{T_2} \|\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi)\|_{L^\infty(\text{supp}(\mathbf{M}))} \mathrm{d}s \leq C(e^{-rT_1} - e^{-rT_2}).
\end{aligned}
\tag{3.35}
$$

Here r and C are constants in Lemma 3.6 that depend on $\mathbf{M}$, V, $\boldsymbol{\sigma}$ and φ. Therefore, we know that the limit

$$
\lim_{t \to \infty} \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi) \right] \rho_0(\mathbf{q})\mathrm{d}\mathbf{q}\, \mathrm{d}s
\tag{3.36}
$$

exists, and by letting $T_2 \to \infty$ in (3.35), we obtain exponentially fast convergence, proving (ii).

For (iii), use Proposition 3.1

$$
\left| \int_{\mathbb{R}^d} \varphi(\mathbf{q})(\rho^\varepsilon(\mathbf{q}, t) - \rho_\infty^\varepsilon) \mathrm{d}\mathbf{q} \right| \leq \|\varphi\|_{L^\infty(\mathbb{R}^d)} \cdot \|\rho^\varepsilon(\mathbf{q}, t) - \rho_\infty^\varepsilon\|_{L^1(\mathbb{R}^d)} \leq Ce^{-rt}.
$$

Thus, for any given ε, we have

$$\lim_{t\to\infty} R(\varepsilon, t; \varphi) = \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q})(\rho_\infty^\varepsilon(\mathbf{q}) - \rho_0(\mathbf{q})) d\mathbf{q} \right).$$

This proves (iii).

Finally, we prove (iv). By the uniform convergence in (i), for any $\eta > 0$, there exists ε_0 such that for all $t > 0$ and $\varepsilon \in (0, \varepsilon_0)$

$$\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi) \right] \rho_0 d\mathbf{q} \, ds \right| < \eta.$$

Taking the limit $t \to \infty$, we get

$$\left| \lim_{t\to\infty} R(t, \varepsilon; \varphi) - \int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi) \right] \rho_0 d\mathbf{q} \, ds \right| < \eta.$$

This implies

$$\lim_{\varepsilon\to 0^+}\lim_{t\to\infty} R(\varepsilon, t; \varphi) = \int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi) \right] \rho_0 d\mathbf{q} \, ds,$$

or equivalently

$$\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q})\rho_\infty^\varepsilon(\mathbf{q}) d\mathbf{q} - \int_{\mathbb{R}^d} \varphi(\mathbf{q})\rho_0(\mathbf{q}) d\mathbf{q} \right)$$
$$= \int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi) \right] \rho_0 d\mathbf{q} \, ds. \qquad \square$$

3.3 The Green-Kubo Relation

Compared to the general linear response formula describing the behavior of the response function $R(\varepsilon, t; \varphi)$ with respect to an external force, the Green-Kubo relation is a special case where the limiting response function is explicitly computed via the stationary auto-correlation function. This auto-correlation function depends only on the correlation of the original unperturbed solution at different times, so it can be used to "predict" the averaged response for a reversible system after applying a conservative force as the perturbation.

Precisely, consider a special case of linear response theory: the perturbation is also of potential form, i.e., there exists $W \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^+)$ such that $\mathbf{M} = \nabla W$. In this case, the perturbed SDE is also of reversible form, and the invariant measure of (3.3) is given by

$$\rho_\infty^\varepsilon = \frac{1}{Z_\varepsilon} e^{-V(\mathbf{q})+\varepsilon W(\mathbf{q})}, \quad Z_\varepsilon := \int_{\mathbb{R}^d} e^{-V(\mathbf{q})+\varepsilon W(\mathbf{q})} d\mathbf{q}. \tag{3.37}$$

By Theorem 3.2, taking two special functions as $\varphi = \mathcal{L}_0 g$ and $\mathbf{M} = \nabla W$, we can rigorously verify the Green-Kubo relation (3.39). Specifically, we take a special class of test functions φ satisfying $\int \varphi\rho_0 d\mathbf{q} = 0$, and then one can uniquely solve g from $\varphi = \mathcal{L}_0 g$ up to

a constant. For this special case, we provide an intuitive proof of the Green-Kubo relation using the semigroup property. Indeed, observe that

$$\int_{\mathbb{R}^d} \left[g - e^{t\mathcal{L}_0} g \right] (\mathcal{L}_0 W) \rho_0 d\mathbf{q} = -\int_{\mathbb{R}^d} \int_0^t \frac{\partial}{\partial s} (e^{s\mathcal{L}_0} g) ds \, (\mathcal{L}_0 W) \rho_0 d\mathbf{q}$$

$$= -\int_{\mathbb{R}^d} \int_0^t e^{s\mathcal{L}_0} \mathcal{L}_0 g \, ds \, (\mathcal{L}_0 W) \rho_0 d\mathbf{q}.$$

Since $\int_{\mathbb{R}^d} (\mathcal{L}_0 W) \rho_0 d\mathbf{q} = 0$, and using the result from the proof of (3.24) with $\psi(\mathbf{q}, 0) = \mathcal{L}_0 W$, we obtain for $\bar{g} := \int_{\mathbb{R}^d} g \rho_0 d\mathbf{q}$, that $\bar{g} \int_{\mathbb{R}^d} e^{t\mathcal{L}_0} (\mathcal{L}_0 W) \rho_0 d\mathbf{q} = 0$. Thus, we have

$$\int_{\mathbb{R}^d} e^{t\mathcal{L}_0} g (\mathcal{L}_0 W) \rho_0 d\mathbf{q} = \int_{\mathbb{R}^d} e^{t\mathcal{L}_0} (g - \bar{g}) (\mathcal{L}_0 W) \rho_0 d\mathbf{q}$$

$$\leq \left(\int |e^{t\mathcal{L}_0} (g - \bar{g})|^2 \rho_0 d\mathbf{q} \right)^{\frac{1}{2}} \left(\int |\mathcal{L}_0 W|^2 \rho_0 d\mathbf{q} \right)^{\frac{1}{2}} \to 0$$

as $t \to +\infty$ due to (3.25). Therefore, the following identity holds:

$$-\int_0^\infty \int_{\mathbb{R}^d} \left[e^{s\mathcal{L}_0} (\mathcal{L}_0 g) \right] (\mathcal{L}_0 W) \rho_0 d\mathbf{q} \, ds = \int_{\mathbb{R}^d} g (\mathcal{L}_0 W) \rho_0 d\mathbf{q} = \int_{\mathbb{R}^d} (\mathcal{L}_0 g) W \rho_0 d\mathbf{q}, \quad (3.38)$$

due to the symmetry of $\mathcal{L}_0$.

In the theorem below, we provide an alternative proof starting from the linear response theory in Theorem 3.2.

Theorem 3.3 (*The Green-Kubo relation*) *Suppose that Assumptions* (*I*) *and* (*II*) *hold. Let* $W \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$ *and* $\mathbf{M} = \nabla W$ *in* (3.3). *For any* $g \in C_c^\infty$, *we have*

$$\lim_{\varepsilon \to 0^+} \lim_{t \to \infty} R(t, \varepsilon; \mathcal{L}_0 g) = \int_{\mathbb{R}^d} W(\mathbf{q})(\mathcal{L}_0 g)(\mathbf{q}) \rho_0(\mathbf{q}) d\mathbf{q}$$

$$= -\int_0^\infty \int_{\mathbb{R}^d} \left[e^{s\mathcal{L}_0} (\mathcal{L}_0 g) \right] (\mathcal{L}_0 W) \rho_0 d\mathbf{q} \, ds. \quad (3.39)$$

Using the conventional notation for the stationary auto-correlation function

$$K_{AB}(t) := \int_{\mathbb{R}^d} \left(e^{t\mathcal{L}_0} A \right) B \, \rho_0 d\mathbf{q},$$

the above Green-Kubo relation (3.39) is interpreted as

$$\lim_{\varepsilon \to 0^+} \lim_{t \to \infty} R(t, \varepsilon; \mathcal{L}_0 g) = -\int_0^{+\infty} K_{\mathcal{L}_0 g, \mathcal{L}_0 W}(t) dt. \quad (3.40)$$

Proof Recall the linear response relation (3.21). We will recast the L.H.S. and R.H.S. of (3.21) respectively, which in the special case $\varphi = \mathcal{L}_0 g$ and $\mathbf{M} = \nabla W$ will deduce the so-called Green-Kubo relation (3.39). This relation is also known as the fluctuation-dissipation theorem by Gallavotti-Cohen [8].

For a given observable $\varphi \in C_c^\infty(\mathbb{R}^d)$ and a potential $U \in C^\infty(\mathbb{R}^d)$ such that $e^{-U} \in L^1(\mathbb{R}^d)$, define

$$f(U) := \frac{\int_{\mathbb{R}^d} \varphi(\mathbf{q}) e^{-U(\mathbf{q})} d\mathbf{q}}{\int_{\mathbb{R}^d} e^{-U(\mathbf{q})} d\mathbf{q}}. \tag{3.41}$$

Then, the L.H.S. of (3.21) should be viewed as the Gateaux derivative at $U = V$ in the direction W, i.e.,

$$\lim_{\varepsilon \to 0^+} \frac{f(V + \varepsilon W) - f(V)}{\varepsilon} = \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_\infty^\varepsilon(\mathbf{q}) d\mathbf{q} - \int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_0(\mathbf{q}) d\mathbf{q} \right). \tag{3.42}$$

By the Dominated Convergence Theorem, direct computation yields

$$\lim_{\varepsilon \to 0^+} \frac{f(V + \varepsilon W) - f(V)}{\varepsilon} = \int_{\mathbb{R}^d} \varphi(\mathbf{q}) \frac{\partial}{\partial \varepsilon} \left(\frac{e^{-V(\mathbf{q}) + \varepsilon W(\mathbf{q})}}{\int_{\mathbb{R}^d} e^{-V(\mathbf{q}) + \varepsilon W(\mathbf{q})} d\mathbf{q}} \right) \Bigg|_{\varepsilon=0} d\mathbf{q}$$

$$= \int_{\mathbb{R}^d} \varphi(\mathbf{q}) W(\mathbf{q}) \rho_0 d\mathbf{q} - \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q}) \rho_0 d\mathbf{q} \right) \left(\int_{\mathbb{R}^d} W(\mathbf{q}) \rho_0 d\mathbf{q} \right). \tag{3.43}$$

For the case $\varphi = \mathcal{L}_0 g$, we have $\int_{\mathbb{R}^d} (\mathcal{L}_0 g) \rho_0 d\mathbf{q} = 0$, and thus the first equality in (3.39) holds.

Now, we reformulate the R.H.S. of (3.21). Substituting $\mathbf{M} = \nabla W$ into (3.21) yields

$$\int_0^\infty \int_{\mathbb{R}^d} \left[\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{M} \cdot \nabla(e^{s\mathcal{L}_0}\varphi) \right] \rho_0 d\mathbf{q}\, ds = -\int_0^\infty \int_{\mathbb{R}^d} \left(e^{s\mathcal{L}_0}\varphi \right) \nabla \cdot \left(\rho_0 \boldsymbol{\sigma}\boldsymbol{\sigma}^T \nabla W \right) d\mathbf{q}\, ds$$

$$= -\int_0^\infty \int_{\mathbb{R}^d} \left(e^{s\mathcal{L}_0}\varphi \right) (\mathcal{L}_0 W) \rho_0 d\mathbf{q}\, ds, \tag{3.44}$$

where the last equality is due to (3.15). By Theorem 3.2, we know that (3.44) and (3.43) are equal. Thus, for $g \in C_c^\infty(\mathbb{R}^d)$ with $\varphi = \mathcal{L}_0 g$, we derive (3.39). $\qquad\square$

We point out that literature (for instance, [25, Chap. 9.3]) also rewrites (3.39) as

$$\frac{1}{2} \int_{\mathbb{R}^d} \Gamma(g, W)(\mathbf{q}) \rho_0(\mathbf{q}) d\mathbf{q} = \int_0^\infty \mathbb{E}\left[(\mathcal{L}_0 g)(\mathbf{q}_s)(\mathcal{L}_0 W)(\mathbf{q}_0) \right] ds$$

$$= \int_0^\infty \int_{\mathbb{R}^d} \mathbb{E}^{\mathbf{q}}\left[(\mathcal{L}_0 g)(\mathbf{q}_s) \right] (\mathcal{L}_0 W) \rho_0 d\mathbf{q}\, ds$$

$$= \int_0^\infty \int_{\mathbb{R}^d} e^{s\mathcal{L}_0} (\mathcal{L}_0 g)(\mathcal{L}_0 W) \rho_0 d\mathbf{q}\, ds. \tag{3.45}$$

Here, $\mathbf{q}_s, s \geq 0$ is a trajectory of (3.2), and $\Gamma(g, h)$ is the carré du champ operator defined as

$$\Gamma(g, h) := \mathcal{L}_0(gh) - g\mathcal{L}_0 h - h\mathcal{L}_0 g. \tag{3.46}$$

Equations (3.39) and (3.45) are regarded as the Green-Kubo relation.

In the following, we provide an example to compute the diffusion coefficients via the Green-Kubo relation. In fact, (3.45) can be rewritten as

$$\int_{\mathbb{R}^d} \rho_0(\mathbf{q})(\nabla g)^\top \boldsymbol{\sigma}\boldsymbol{\sigma}^\top \nabla W d\mathbf{q} = \int_0^\infty \mathbb{E}\left[(\mathcal{L}_0 g)(\mathbf{q}_s)(\mathcal{L}_0 W)(\mathbf{q}_0) \right] ds. \tag{3.47}$$

Though we assume W and g are in $C_c^\infty(\mathbb{R}^d)$, by a density argument, (3.45) or (3.47) also holds for $W(\mathbf{q}) = q_j$, $g(\mathbf{q}) = q_i$ where $i, j = 1, 2, \ldots, d$. Thus, we derive the diffusion coefficients

$$(\boldsymbol{\sigma}\boldsymbol{\sigma}^\top)_{ij} = \int_0^\infty \mathbb{E}\left[(\boldsymbol{\sigma}_i \boldsymbol{\sigma} \nabla V)(\mathbf{q}_s)(\boldsymbol{\sigma}_j \boldsymbol{\sigma} \nabla V)(\mathbf{q}_0)\right] ds,$$

where $\boldsymbol{\sigma}_j$, $j = 1, 2, \ldots, d$ are the j-th row of $\boldsymbol{\sigma}$.

Remark 3.1 As mentioned in (3.38), the Green-Kubo relation could be verified by using the exponential convergence of $e^{s\mathcal{L}_0}\varphi$. This method is independent of linear response theory (see [25, Result 9.1]). The way we verified the Green-Kubo formula is to reveal the intrinsic relationship between it and the linear response theory. Indeed, Hairer-Majda directly termed (3.21) as the Green-Kubo relation in [11].

Remark 3.2 We also remark that the Green-Kubo relation when studying ρ_∞^ε, relaxed to ρ_0 after canceling the perturbation force, can be used to justify the Onsager regression hypothesis [22, 23]; see also the review article [20]. That is, using the following relation, one can compute (in a reversible way) the original unperturbed auto-correlation function

$$\lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^d} \varphi(\mathbf{q})\rho_0(\mathbf{q})d\mathbf{q} - \int_{\mathbb{R}^d} \varphi(\mathbf{q})\rho_\infty^\varepsilon(\mathbf{q})d\mathbf{q} \right)$$
$$= \lim_{\varepsilon \to 0^+} \int_0^\infty \int_{\mathbb{R}^d} \left[e^{s\mathcal{L}_\varepsilon}(\mathcal{L}_\varepsilon g) \right](\mathcal{L}_\varepsilon W)\rho_\infty^\varepsilon(\mathbf{q})d\mathbf{q}ds$$
$$= \int_0^\infty \int_{\mathbb{R}^d} \left[e^{s\mathcal{L}_0}(\mathcal{L}_0 g) \right](\mathcal{L}_0 W)\rho_0(\mathbf{q})d\mathbf{q}ds.$$

4 The Linear Response Theory for Underdamped Langevin and Generalized Langevin Dynamics

Unlike the overdamped case, the degenerate ellipticity of the underdamped Langevin equation prevents the direct analogy of many results from the overdamped Langevin case. In particular, few results are known for irreversible underdamped Langevin dynamics. Fortunately, Villani's remarkable work on hypocoercivity [29] has facilitated the analysis of the reversible case, showing that exponential convergence still holds under mild conditions on the potential.

In this section, we consider underdamped Langevin dynamics and study the linear response theory for the case where the external force perturbation is in a conservative form $\varepsilon \nabla W$. The main differences between the underdamped Langevin dynamics and the overdamped case are hypocoercivity and hypoellipticity. Therefore, we will first clarify hypocoercivity and hypoellipticity before proving parallel results for the behavior of the response function $R(\varepsilon, t; \varphi)$.

4.1 Hypocoercivity and Hypoellipticity of the Kinetic Fokker-Planck Equation

Suppose that $V(\mathbf{q}) \in C^\infty(\mathbb{R}^d)$ is positive and satisfies Assumptions (I) and (II). Consider the following underdamped Langevin dynamics

$$
\begin{cases}
d\mathbf{q}^\varepsilon = \mathbf{p}^\varepsilon dt, \\
d\mathbf{p}^\varepsilon = -\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{p}^\varepsilon dt - \nabla_{\mathbf{q}}[V(\mathbf{q}^\varepsilon) - \varepsilon W(\mathbf{q}^\varepsilon)]dt + \sqrt{2}\boldsymbol{\sigma}\, d\mathbf{B}, \\
(\mathbf{q}^\varepsilon(0), \mathbf{p}^\varepsilon(0)) \sim \rho_0 := \dfrac{1}{Z}e^{-|\mathbf{p}|^2/2-V(\mathbf{q})}, \quad Z := \displaystyle\int_{\mathbb{R}^{2d}} e^{-|\mathbf{p}|^2/2-V(\mathbf{q})} d\mathbf{q}d\mathbf{p},
\end{cases}
\tag{4.1}
$$

where $W(\mathbf{q}) \in C_c^\infty(\mathbb{R}^d)$ is the potential as in Sect. 3.3 such that the force perturbation is also of the potential form $\varepsilon \nabla W$.

Let $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, s)$ be the law of $(\mathbf{q}_s^\varepsilon, \mathbf{p}_s^\varepsilon)$. It satisfies the following Fokker-Planck equation

$$
\frac{\partial \rho^\varepsilon(\mathbf{q}, \mathbf{p}, t)}{\partial t} = \mathcal{L}_\varepsilon^* \rho(\mathbf{q}, \mathbf{p}, t), \quad \rho^\varepsilon(\mathbf{q}, \mathbf{p}, 0) = \rho_0,
\tag{4.2}
$$

where $\mathcal{L}_\varepsilon^*$ is the perturbed Fokker-Planck operator

$$
\mathcal{L}_\varepsilon^* \rho := -\mathbf{p} \cdot \nabla_{\mathbf{q}}\rho + \nabla_{\mathbf{q}}[V(\mathbf{q}) - \varepsilon W(\mathbf{q})] \cdot \nabla_{\mathbf{p}}\rho + \nabla_{\mathbf{p}} \cdot (\rho\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{p} + \boldsymbol{\sigma}\boldsymbol{\sigma}^T \nabla_{\mathbf{p}}\rho).
\tag{4.3}
$$

The invariant measure of (4.2) is the Gibbs measure

$$
\rho_\infty^\varepsilon := \frac{1}{Z_\varepsilon}e^{-|\mathbf{p}|^2/2-V(\mathbf{q})+\varepsilon W(\mathbf{q})}, \quad Z_\varepsilon := \int_{\mathbb{R}^{2d}} e^{-|\mathbf{p}|^2/2-V(\mathbf{q})+\varepsilon W(\mathbf{q})} d\mathbf{q}d\mathbf{p}.
\tag{4.4}
$$

LRT is interested in the limit behavior of the response function given an observable $\varphi \in C_c^\infty(\mathbb{R}^{2d})$, i.e.,

$$
R(\varepsilon, t; \varphi) := \frac{1}{\varepsilon}\left(\int_{\mathbb{R}^{2d}} \varphi(\mathbf{q}, \mathbf{p})\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t) d\mathbf{q}d\mathbf{p} - \int_{\mathbb{R}^{2d}} \varphi(\mathbf{q}, \mathbf{p})\rho_0(\mathbf{q}, \mathbf{p}) d\mathbf{q}d\mathbf{p} \right).
\tag{4.5}
$$

As in (2.15), the unperturbed backward operator $\mathcal{L}_0$ is given by

$$
\mathcal{L}_0 f := \mathbf{p} \cdot \nabla_{\mathbf{q}} f - \nabla_{\mathbf{q}} V(\mathbf{q}) \cdot \nabla_{\mathbf{p}} f - (\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{p}) \cdot \nabla_{\mathbf{p}} f + (\boldsymbol{\sigma}\boldsymbol{\sigma}^T) : \nabla_{\mathbf{p}}^2 f,
\tag{4.6}
$$

and the unperturbed Fokker-Planck operator $\mathcal{L}_0^*$ is given by

$$
\mathcal{L}_0^* \rho := -\mathbf{p} \cdot \nabla_{\mathbf{q}}\rho + \nabla_{\mathbf{q}} V(\mathbf{q}) \cdot \nabla_{\mathbf{p}}\rho + \nabla_{\mathbf{p}} \cdot (\rho\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{p} + \boldsymbol{\sigma}\boldsymbol{\sigma}^T \nabla_{\mathbf{p}}\rho).
\tag{4.7}
$$

We now consider the equation of $p^\varepsilon(\mathbf{q}, \mathbf{p}, t) = \dfrac{\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t)}{\rho_\infty^\varepsilon(\mathbf{q}, \mathbf{p})}$, which is

$$
\frac{\partial p^\varepsilon}{\partial t} = \mathcal{L}_{\varepsilon, 1} p^\varepsilon, \quad p^\varepsilon(\mathbf{q}, \mathbf{p}, 0) = \frac{\rho_0}{\rho_\infty^\varepsilon},
$$

where

$$
\mathcal{L}_{\varepsilon, 1} f := -\mathbf{p} \cdot \nabla_{\mathbf{q}} f + (\nabla_{\mathbf{q}} V(\mathbf{q}) - \varepsilon \nabla_{\mathbf{q}} W(\mathbf{q})) \cdot \nabla_{\mathbf{p}} f - (\boldsymbol{\sigma}\boldsymbol{\sigma}^T \mathbf{p}) \cdot \nabla_{\mathbf{p}} f + (\boldsymbol{\sigma}\boldsymbol{\sigma}^T) : \nabla_{\mathbf{p}}^2 f.
\tag{4.8}
$$

We point out that, unlike the overdamped Langevin dynamics, $\mathcal{L}_{\varepsilon,1}$ is not exactly the same as the generator, since the first two terms in $\mathcal{L}_0$ form the Liouville operator.

We can prove that $\mathcal{L}_{\varepsilon,1}$ generates a strongly continuous semigroup of contraction on $L^2(\rho_\infty^\varepsilon)$.

Lemma 4.1 *Let $V(\mathbf{q}) \in C^\infty(\mathbb{R}^d; \mathbb{R})$ satisfy Assumption (I) and $W(\mathbf{q}) \in C_c^\infty(\mathbb{R}^d; \mathbb{R})$. Consider equation* (4.2). *Then*

(i) $\rho_0/\rho_\infty^\varepsilon \in H^k(\rho_\infty^\varepsilon)$, $k = 1, 2, \ldots$
(ii) $\mathcal{L}_{\varepsilon,1}$ *in* (4.8) *generates a strongly continuous semigroup of contractions in $L^2(\rho_\infty^\varepsilon)$;*
(iii) *Equation* (4.2) *admits a unique solution $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t) \in C^1([0, T]; L^2(1/\rho_\infty^\varepsilon))$ for any $T > 0$.*

By Hörmander's hypoellipticity Theorem 3.1, we can prove that the unique solution of (4.2) is smooth (proof is in Appendix B).

Lemma 4.2 *Let $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t)$ be the unique solution of* (4.2). *Then it is smooth in $\mathbb{R}^{2d} \times (0, \infty)$.*

4.2 Linear Response Theory for Underdamped Langevin

According to [6], for any $\varphi \in C_c^\infty(\mathbb{R}^{2d})$, the solution to the backward equation

$$\frac{\partial \varphi(\mathbf{q}, \mathbf{p}, t)}{\partial t} = \mathcal{L}_0 \varphi(\mathbf{q}, \mathbf{p}, t), \quad \varphi(\mathbf{q}, \mathbf{p}, 0) = \varphi(\mathbf{q}, \mathbf{p}), \tag{4.9}$$

converges exponentially fast in $L^2(\rho_0)$, i.e.,

$$\|\varphi(\cdot, \cdot, t) - \bar{\varphi}\|_{L^2(\rho_0)} \le C e^{-\lambda t}. \tag{4.10}$$

Here $\bar{\varphi} := \int_{\mathbb{R}^{2d}} \varphi(\mathbf{q}, \mathbf{p}) \rho_0(\mathbf{q}, \mathbf{p}) d\mathbf{q} d\mathbf{p}$, $C > 0$ is a constant depending on V, $\boldsymbol{\sigma}$, and φ, and $r > 0$ is a constant depending on V and $\boldsymbol{\sigma}$.

Since W is of compact support, the conditions of hypocoercivity for V in Assumption (II) still apply to the perturbed Fokker-Planck equation [6]. Hence, exponential convergence also holds

$$\|\rho^\varepsilon(\cdot, \cdot, t) - \rho_\infty^\varepsilon(\cdot, \cdot)\|_{L^2(1/\rho_\infty^\varepsilon)} \le e^{-rt} \|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)}. \tag{4.11}$$

Here $r > 0$ depends only on V and $\boldsymbol{\sigma}$, but not on ε; $C_1 > 0$ depends only on V and $\boldsymbol{\sigma}$. The reason they do not depend on ε is that the perturbation from W is smooth and of compact support.

Using these exponential convergence results (as in Lemma 3.6), we can rigorously verify the linear response theory.

Theorem 4.1 *Let $V(\mathbf{q}) \in C^\infty(\mathbb{R}^d)$ be positive and satisfy Assumptions (I) and (II). Suppose $W(\mathbf{q}) \in C_c^\infty(\mathbb{R}^d)$. Let $\rho^\varepsilon(\mathbf{q}, t)$ be the law of $(\mathbf{q}_t^\varepsilon, \mathbf{p}_t^\varepsilon)$ in* (4.1), *and let $\rho_\infty^\varepsilon(\mathbf{q})$ be the invariant measure of* (4.1). *For some $\varphi \in C_c^\infty(\mathbb{R}^{2d})$, consider $R(t, \varepsilon; \varphi)$ defined in* (4.5). *Then*

(i) *(Convergence as $\varepsilon \to 0^+$) For any given $t > 0$,*

$$\lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) = \int_0^t \int_{\mathbb{R}^{2d}} [\nabla_\mathbf{q} W \cdot (\nabla_\mathbf{p}(e^{s\mathcal{L}_0} \varphi))] \rho_0 d\mathbf{q} d\mathbf{p} ds. \tag{4.12}$$

Moreover, the limit in (4.12) *holds uniformly for all $t > 0$; i.e., for any $\eta > 0$, there exists ε_0 such that for any $0 < \varepsilon < \varepsilon_0$ and $t > 0$,*

$$\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s \right| < \eta;$$

(ii) (*Convergence as $\varepsilon \to 0^+$, then $t \to \infty$*) *The following limit exists*

$$\lim_{t \to \infty} \lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) = \int_0^\infty \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s, \qquad (4.13)$$

and the convergence in t is exponentially fast. That is, there exist constants $C > 0$ and $r > 0$ (depending on $\mathbf{M}$ and V) such that

$$\left| \lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) - \int_0^\infty \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s \right| \leq Ce^{-rt} \qquad (4.14)$$

holds for all $t > 0$;

(iii) (*Convergence as $t \to \infty$*) *For any $\varepsilon > 0$,*

$$\lim_{t \to \infty} R(t, \varepsilon; \varphi) = \frac{1}{\varepsilon}\left(\int_{\mathbb{R}^{2d}} \varphi(\mathbf{q})\rho_\infty^\varepsilon(\mathbf{q})\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p} - \int_{\mathbb{R}^{2d}} \varphi(\mathbf{q})\rho_0(\mathbf{q}, t)\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p} \right); \qquad (4.15)$$

(iv) (*Convergence as $t \to \infty$, then $\varepsilon \to 0^+$*) *The following limit exists*

$$\lim_{\varepsilon \to 0^+} \lim_{t \to \infty} R(t, \varepsilon; \varphi) = \int_0^\infty \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s, \qquad (4.16)$$

or equivalently,

$$\lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}\left(\int_{\mathbb{R}^{2d}} \varphi(\mathbf{q})\rho_\infty^\varepsilon(\mathbf{q})\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p} - \int_{\mathbb{R}^{2d}} \varphi(\mathbf{q})\rho_0(\mathbf{q})\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p} \right)$$
$$= \int_0^\infty \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s. \qquad (4.17)$$

Proof We first prove (i). By Lemma 4.2, $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t)$ is smooth. By Duhamel's principle, we have

$$\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t) = e^{t\mathcal{L}_0^*}\rho_0 - \varepsilon \int_0^t e^{(t-s)\mathcal{L}_0^*}(\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}}\rho^\varepsilon(\mathbf{q}, \mathbf{p}, s))\mathrm{d}s. \qquad (4.18)$$

Remember that $e^{t\mathcal{L}_0^*}\rho_0 = \rho_0$, so

$$R(t, \varepsilon; \varphi) = \frac{1}{\varepsilon}\left(\int_{\mathbb{R}^{2d}} \varphi(\mathbf{q}, \mathbf{p})\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t)\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p} - \int_{\mathbb{R}^{2d}} \varphi(\mathbf{q}, \mathbf{p})\rho_0(\mathbf{q}, \mathbf{p})\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p} \right)$$
$$= -\int_{\mathbb{R}^{2d}} \int_0^t \varphi(\mathbf{q}, \mathbf{p})e^{(t-s)\mathcal{L}_0^*}(\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}}\rho^\varepsilon(\mathbf{q}, \mathbf{p}, s))\mathrm{d}s\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}.$$

Since $\varphi(\mathbf{q}, \mathbf{p})e^{(t-s)\mathcal{L}_0^*}(\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}}\rho^\varepsilon(\mathbf{q}, \mathbf{p}, s))$ is a smooth function with compact support in $\mathbb{R}^{2d} \times [0, t]$, Fubini's theorem gives

$$R(t, \varepsilon; \varphi) = -\int_0^t \int_{\mathbb{R}^{2d}} \varphi(\mathbf{q}, \mathbf{p})e^{(t-s)\mathcal{L}_0^*}(\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}}\rho^\varepsilon(\mathbf{q}, \mathbf{p}, s))\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s.$$

Integration by parts then yields

$$R(t, \varepsilon; \varphi) = \int_0^t \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{(t-s)\mathcal{L}_0} \varphi))] \rho^\varepsilon(\mathbf{q}, \mathbf{p}, s) \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}s.$$

By (4.11), we know

$$
\begin{aligned}
\|\rho^\varepsilon(\cdot, \cdot, s) - \rho_0(\cdot, \cdot)\|_{L^2(1/\rho_\infty^\varepsilon)} &\leq \|\rho^\varepsilon(\cdot, \cdot, s) - \rho_\infty^\varepsilon(\cdot, \cdot)\|_{L^2(1/\rho_\infty^\varepsilon)} + \|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)} \\
&\leq (1 + e^{-rs}) \|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)} \\
&\leq 2\|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)}.
\end{aligned}
$$

Thus, by Cauchy-Schwarz's inequality, we have

$$
\begin{aligned}
&\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi))] \rho_0 \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}s \right| \\
&= \left| \int_0^t \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{(t-s)\mathcal{L}_0} \varphi))] (\rho^\varepsilon(\mathbf{q}, \mathbf{p}, s) - \rho_0(\mathbf{q}, \mathbf{p})) \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}s \right| \\
&\leq \int_0^t \|\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{(t-s)\mathcal{L}_0} \varphi))\|_{L^2(\rho_\infty^\varepsilon)} \|\rho^\varepsilon(\mathbf{q}, \mathbf{p}, s) - \rho_0(\mathbf{q}, \mathbf{p})\|_{L^2(1/\rho_\infty^\varepsilon)} \mathrm{d}s \\
&\leq 2\|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)} \cdot \int_0^t \|\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi))\|_{L^2(\rho_\infty^\varepsilon)} \mathrm{d}s.
\end{aligned}
\tag{4.19}
$$

Because W has compact support, there exists a constant $C' > 0$, depending only on V, W, and $\boldsymbol{\sigma}$, such that

$$\|\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi))\|_{L^2(\rho_\infty^\varepsilon)} \leq C\|\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi)\|_{L^2(\rho_\infty^\varepsilon)} \leq C'\|\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi)\|_{L^2(\rho_0)}. \tag{4.20}$$

Notice that $\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi) = \nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}(\varphi - \bar{\varphi}))$. Thus, by (4.10), we have

$$
\begin{aligned}
\|\boldsymbol{\sigma}^T \nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi)\|_{L^2(\rho_0)}^2 &= -\int_{\mathbb{R}^{2d}} (e^{s\mathcal{L}_0}(\varphi - \bar{\varphi}))(\mathcal{L}_0 e^{s\mathcal{L}_0}(\varphi - \bar{\varphi})) \rho_0 \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \\
&\leq \|e^{s\mathcal{L}_0}(\varphi - \bar{\varphi})\|_{L^2(\rho_0)} \|\mathcal{L}_0 e^{s\mathcal{L}_0}(\varphi - \bar{\varphi})\|_{L^2(\rho_0)} \\
&\leq C'' e^{-2\lambda t}.
\end{aligned}
$$

Here $C'' > 0$ is a constant depending on φ, V, and $\boldsymbol{\sigma}$, and λ is the constant in (4.10). Since $\boldsymbol{\sigma}$ is non-singular, we have

$$\|\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi)\|_{L^2(\rho_0)}^2 \leq C_1 e^{-rt}. \tag{4.21}$$

Here C_1 is a constant depending on φ, V, and $\boldsymbol{\sigma}$. Returning to (4.19), we obtain

$$
\begin{aligned}
\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0} \varphi))] \rho_0 \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}s \right| &\leq 2C_2 \|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)} \cdot \int_0^t e^{-rs} \mathrm{d}s \\
&\leq C_3 \|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)}.
\end{aligned}
$$

Here $C_3 > 0$ is a constant depending on φ, V, W, and $\boldsymbol{\sigma}$. Since ρ_0 and ρ_∞^ε are the Gibbs measures given in (4.1) and (4.4) respectively, and W is of compact support, the dominated convergence theorem directly implies

$$\|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)} \to 0$$

as $\varepsilon \to 0^+$. Since $C_3\|\rho_0 - \rho_\infty^\varepsilon\|_{L^2(1/\rho_\infty^\varepsilon)}$ is independent of t and approaches 0 as $\varepsilon \to 0^+$, (i) is proved, and the convergence is uniform in t.

Now we proceed to prove (ii). By (4.21) and Cauchy-Schwarz's inequality, for any $T_1, T_2 > 0$, we have

$$\left| \int_{T_1}^{T_2} \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s \right| \leq \|\nabla_{\mathbf{q}} W\|_{L^2(\rho_0)} \cdot \int_{T_1}^{T_2} e^{-sr}\mathrm{d}s$$

$$\leq C(e^{-T_1 r} - e^{-T_2 r}).$$

Here C is a constant depending on V, W, and $\boldsymbol{\sigma}$. Thus, the limit in (4.13) exists and converges exponentially fast in t.

Next, we prove (iii). This follows immediately from (4.11). For any given $\varepsilon > 0$, we have

$$\left| \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^{2d}} \varphi(\mathbf{q}, \mathbf{p})\rho_\infty^\varepsilon(\mathbf{q}, \mathbf{p})\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p} - \int_{\mathbb{R}^{2d}} \varphi(\mathbf{q}, \mathbf{p})\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t)\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p} \right) \right|$$

$$\leq \frac{1}{\varepsilon} \|\varphi\|_{L^2(1/\rho_\infty^\varepsilon)} \|\rho^\varepsilon(\cdot, \cdot, t) - \rho_\infty^\varepsilon(\cdot, \cdot)\|_{L^2(1/\rho_\infty^\varepsilon)}$$

$$\leq C_1 e^{-r_1 t}/\varepsilon.$$

Here $C_1 > 0$ is a constant depending on φ, V, W, and $\boldsymbol{\sigma}$, and $r_1 > 0$ is in (4.11). Thus, the limit as $t \to \infty$ exists for each $\varepsilon > 0$.

Finally, we prove (iv). By (i), we know that for any $\eta > 0$, there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$, we have

$$\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s \right| < \eta.$$

Passing to the limit as $t \to \infty$, we obtain

$$\left| \lim_{t \to \infty} R(t, \varepsilon; \varphi) - \int_0^\infty \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s \right| \leq \eta.$$

Thus, by definition,

$$\lim_{\varepsilon \to 0^+} \lim_{t \to \infty} R(t, \varepsilon; \varphi) = \int_0^\infty \int_{\mathbb{R}^{2d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}s.$$

By (iii), this also implies (4.17). This completes the proof. $\qquad\square$

4.3 Linear Response for the Generalized Langevin with Memory

The generalized Langevin equation with a general algebraic memory kernel was introduced in [18, 21]. Recent works [17, 19] have extended the fluctuation-dissipation relation for both overdamped and underdamped generalized Langevin dynamics.

When the memory kernel takes an exponential form, the generalized Langevin dynamics can be transformed into classical underdamped Langevin dynamics by introducing an additional variable [24, 26]. As an application of Theorem 4.1, we consider the generalized Langevin equation with an exponential memory kernel, i.e., (2.21)

$$
\begin{cases}
\ddot{\mathbf{q}}_t = -\nabla_{\mathbf{q}} V(\mathbf{q}) - \sum_{i=1}^{n} \mathbf{A}_i \mathbf{A}_i^T \int_0^t e^{-\alpha_i(t-s)} \dot{\mathbf{q}}_s \, \mathrm{d}s + \mathbf{A}_i \mathbf{f}_t^i, \\
\mathrm{d}\mathbf{f}_t^i = -\alpha_i \mathbf{f}_t^i \mathrm{d}t + \sqrt{2\beta^{-1}\alpha_i} \, \mathrm{d}\mathbf{B}_t^i, \quad i = 1, 2, \ldots, n.
\end{cases}
$$

Here, $V \in C^\infty(\mathbb{R}^d)$ satisfies Assumptions (I) and (II) with $\boldsymbol{\sigma} = \mathbf{I}_d$, and for each $i = 1, 2, \ldots, n$, $\mathbf{A}_i \in \mathbb{R}^{d \times d}$ is a constant matrix, $\alpha_i > 0$ is a constant, and $\mathbf{B}_t^i \in \mathbb{R}^d$ is a standard Brownian motion. The $\mathbf{B}_t^i$'s are independent. Furthermore, we assume that the Hessian of V is uniformly bounded, i.e., there exists a constant $C > 0$ such that for all $\mathbf{q} \in \mathbb{R}^d$,

$$
\|\nabla^2 V(\mathbf{q})\| \leq C. \tag{4.22}
$$

Here, the norm refers to the Frobenius norm.

Let $\mathbf{z}_t^i = -\mathbf{A}_i^T \int_0^t e^{-\alpha_i(t-s)} \dot{\mathbf{q}}_s \, \mathrm{d}s + \mathbf{f}_t^i$ and $\dot{\mathbf{q}} = \mathbf{p}$. Then (2.21) can be reformulated as (2.22), i.e.,

$$
\begin{cases}
\mathrm{d}\mathbf{q} = \mathbf{p}\mathrm{d}t, \\
\mathrm{d}\mathbf{p} = \left(-\nabla_{\mathbf{q}} V(\mathbf{q}) + \sum_{i=1}^{n} \mathbf{A}_i \mathbf{z}_i\right) \mathrm{d}t, \\
\mathrm{d}\mathbf{z}_i = -(\alpha_i \mathbf{z}_i + \mathbf{A}_i^T \mathbf{p})\mathrm{d}t + \sqrt{2\beta^{-1}\alpha_i} \, \mathrm{d}\mathbf{B}_i, \quad i = 1, 2, \ldots, n.
\end{cases}
$$

For simplicity and without loss of generality, we consider the case $n = 1$ and $\mathbf{A}_1 = \mathbf{I}_d$, i.e. (2.25). The conclusions in this section can be easily extended to the general case.

Now based on (2.25), we impose a compact and smooth perturbation on the potential term, i.e., for $W \in C_c^\infty(\mathbb{R}^d)$, and consider the following perturbed SDE

$$
\begin{cases}
\mathrm{d}\mathbf{q}^\varepsilon = \mathbf{p}^\varepsilon \mathrm{d}t, \\
\mathrm{d}\mathbf{p}^\varepsilon = \left(-\nabla_{\mathbf{q}} V(\mathbf{q}^\varepsilon) + \varepsilon \nabla_{\mathbf{q}} W(\mathbf{q}^\varepsilon) + \mathbf{z}^\varepsilon\right) \mathrm{d}t, \\
\mathrm{d}\mathbf{z}^\varepsilon = -(\alpha \mathbf{z}^\varepsilon + \mathbf{p}^\varepsilon)\mathrm{d}t + \sqrt{2\beta^{-1}\alpha} \, \mathrm{d}\mathbf{B}, \\
(\mathbf{q}^\varepsilon, \mathbf{p}^\varepsilon, \mathbf{z}^\varepsilon) \sim \rho_0.
\end{cases} \tag{4.23}
$$

The invariant measure of (4.23) is given by

$$
\rho_\infty^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}) := \frac{1}{Z^\varepsilon} \exp(-\beta H(\mathbf{q}, \mathbf{p}, \mathbf{z}) + \varepsilon \beta W(\mathbf{q})),
$$
$$
Z^\varepsilon := \int_{\mathbb{R}^{3d}} e^{-\beta H(\mathbf{q}, \mathbf{p}, \mathbf{z}) + \varepsilon \beta W(\mathbf{q})} \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{z} < \infty. \tag{4.24}
$$

Let $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}, t)$ be the law of $(\mathbf{q}^\varepsilon(t), \mathbf{p}^\varepsilon(t), \mathbf{z}^\varepsilon(t))$ in (2.25). Then $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}, t)$ satisfies the following perturbed Fokker-Planck equation with initial value $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}, 0) =$

$\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z})$

$$\frac{\partial \rho^\varepsilon}{\partial t} = \mathcal{L}_\varepsilon^* \rho, \quad \mathcal{L}_\varepsilon^* \rho := - \mathbf{p} \cdot \nabla_\mathbf{q} \rho + \nabla_\mathbf{q}(V(\mathbf{q}) - \varepsilon W(\mathbf{q})) \cdot \nabla_\mathbf{p} \rho + (\mathbf{p} \cdot \nabla_\mathbf{z} \rho - \mathbf{z} \cdot \nabla_\mathbf{p} \rho)$$

$$+ \alpha \nabla_\mathbf{z} \cdot \left(\mathbf{z} \rho + \frac{1}{\beta} \nabla_\mathbf{z} \rho \right). \tag{4.25}$$

Consider $\varphi \in C_c^\infty(\mathbb{R}^{3d})$. The linear response theory focuses on the limit behavior of

$$R(\varepsilon, t; \varphi) := \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}) \rho^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}, t) \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{z} - \int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}) \rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{z} \right). \tag{4.26}$$

The unperturbed backward operator $\mathcal{L}_0$ is given by

$$\mathcal{L}_0 f := \mathbf{p} \cdot \nabla_\mathbf{q} f - \nabla_\mathbf{q} V(\mathbf{q}) \cdot \nabla_\mathbf{p} f - (\mathbf{p} \cdot \nabla_\mathbf{z} f - \mathbf{z} \cdot \nabla_\mathbf{p} f) - \alpha \mathbf{z}^T \nabla_\mathbf{z} f + \frac{\alpha}{\beta} \Delta_\mathbf{z} f. \tag{4.27}$$

The backward equation with initial value φ is then written as

$$\frac{\partial \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}, t)}{\partial t} = \mathcal{L}_0 \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}, t), \quad \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}, 0) = \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}). \tag{4.28}$$

According to the hypocoercivity result in [12], under Assumptions (I), (II), and (4.22) on the smooth potential V, the solution to (4.28) converges exponentially in the weighted Sobolev space $H^1(\rho_0)$

$$\left\| \varphi(\cdot, \cdot, \cdot, t) - \int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}) \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{z} \right\|_{H^1(\rho_0)}$$

$$\leq e^{-\lambda t} \left\| \varphi(\cdot, \cdot, \cdot, 0) - \int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}) \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{z} \right\|_{H^1(\rho_0)}. \tag{4.29}$$

We emphasize that (4.29) is stronger than the result used in the underdamped Langevin case, which focuses on exponential convergence in $L^2(\rho_0)$, while (4.29) concerns $H^1(\rho_0)$. With (4.29), we can rigorously verify the LRT for the generalized Langevin dynamics.

Theorem 4.2 *Let $V(\mathbf{q}) \in C^\infty(\mathbb{R}^d)$ be positive, satisfying Assumptions (I), (II), and (4.22). Suppose $W(\mathbf{q}) \in C_c^\infty(\mathbb{R}^d)$. Let $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}, t)$ be the law of $(\mathbf{q}^\varepsilon(t), \mathbf{p}^\varepsilon(t), \mathbf{z}^\varepsilon(t))$ in (4.23), and let $\rho_\infty^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z})$ be the invariant measure of (4.23). For some $\varphi \in C_c^\infty(\mathbb{R}^{3d})$, consider $R(t, \varepsilon; \varphi)$ defined in (4.26). Then*

(i) (*Convergence as $\varepsilon \to 0^+$*) *For any given $t > 0$,*

$$\lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) = \int_0^t \int_{\mathbb{R}^{3d}} [\nabla_\mathbf{q} W \cdot (\nabla_\mathbf{p}(e^{s\mathcal{L}_0} \varphi))] \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{z}\mathrm{d}s. \tag{4.30}$$

Moreover, the limit in (4.30) holds uniformly for all $t > 0$; i.e., for any $\eta > 0$, there exists ε_0 such that for any $0 < \varepsilon < \varepsilon_0$ and $t > 0$,

$$\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^{3d}} [\nabla_\mathbf{q} W \cdot (\nabla_\mathbf{p}(e^{s\mathcal{L}_0} \varphi))] \rho_0 \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{z}\mathrm{d}s \right| < \eta.$$

(ii) (*Convergence as $\varepsilon \to 0^+$ then $t \to \infty$*) *The following limit exists*

$$\lim_{t \to \infty} \lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) = \int_0^\infty \int_{\mathbb{R}^{3d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 d\mathbf{q}d\mathbf{p}d\mathbf{z}ds, \tag{4.31}$$

and the convergence in t is exponentially fast; i.e., there exist constants $C > 0$ and $r > 0$ that depend on W and V such that

$$\left| \lim_{\varepsilon \to 0^+} R(t, \varepsilon; \varphi) - \int_0^\infty \int_{\mathbb{R}^{3d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 d\mathbf{q}d\mathbf{p}d\mathbf{z}ds \right| \leq Ce^{-rt} \tag{4.32}$$

holds for all $t > 0$.

(iii) (*Convergence as $t \to \infty$*) *For any $\varepsilon > 0$,*

$$\lim_{t \to \infty} R(t, \varepsilon; \varphi) = \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z})\rho_\infty^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z})d\mathbf{q}d\mathbf{p}d\mathbf{z} \right.$$

$$\left. - \int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z})\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z})d\mathbf{q}d\mathbf{p}d\mathbf{z} \right). \tag{4.33}$$

(iv) (*Convergence as $t \to \infty$ then $\varepsilon \to 0^+$*) *The following limit exists*

$$\lim_{\varepsilon \to 0^+} \lim_{t \to \infty} R(t, \varepsilon; \varphi) = \int_0^\infty \int_{\mathbb{R}^{3d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 d\mathbf{q}d\mathbf{p}d\mathbf{z}ds, \tag{4.34}$$

or equivalently,

$$\lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z})\rho_\infty^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z})d\mathbf{q}d\mathbf{p}d\mathbf{z} - \int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z})\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z})d\mathbf{q}d\mathbf{p}d\mathbf{z} \right)$$

$$= \int_0^\infty \int_{\mathbb{R}^{3d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 d\mathbf{q}d\mathbf{p}d\mathbf{z}ds.$$
$$\tag{4.35}$$

Proof The proof follows closely from the case of the underdamped Langevin dynamics, so we provide only a sketch.

For (i), as in the proof of Theorem 4.1, by Duhamel's principle, we have

$$R(t, \varepsilon; \varphi) = -\int_{\mathbb{R}^{3d}} \int_0^t \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z})e^{(t-s)\mathcal{L}_0^*} \left(\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}}\rho^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}) \right) d\mathbf{q}d\mathbf{p}d\mathbf{z}ds. \tag{4.36}$$

Using integration by parts and Fubini's theorem, this becomes

$$R(t, \varepsilon; \varphi) = \int_0^t \int_{\mathbb{R}^{3d}} \left[\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}} \left(e^{(t-s)\mathcal{L}_0}\varphi \right) \right] \rho^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}, s)d\mathbf{q}d\mathbf{p}d\mathbf{z}ds. \tag{4.37}$$

The variational structure of the Fokker-Planck equation gives the bound

$$\|\rho^\varepsilon(\cdot, \cdot, \cdot, s) - \rho_0(\cdot, \cdot, \cdot)\|_{L^2(1/\rho_\infty^\varepsilon)} \leq \|\rho^\varepsilon(\cdot, \cdot, \cdot, s) - \rho_\infty^\varepsilon(\cdot, \cdot, \cdot)\|_{L^2(1/\rho_\infty^\varepsilon)}$$

$$+ \|\rho_0(\cdot, \cdot, \cdot) - \rho_\infty^\varepsilon(\cdot, \cdot, \cdot)\|_{L^2(1/\rho_\infty^\varepsilon)}$$

$$\leq 2\|\rho_0(\cdot, \cdot, \cdot) - \rho_\infty^\varepsilon(\cdot, \cdot, \cdot)\|_{L^2(1/\rho_\infty^\varepsilon)}.$$

Using Cauchy-Schwarz's inequality, we obtain

$$\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^{3d}} \left[\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}} \left(e^{(t-s)\mathcal{L}_0} \varphi \right) \right] \rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{z} \mathrm{d}s \right|$$

$$\leq 2 \| \rho_0(\cdot, \cdot, \cdot) - \rho_\infty^\varepsilon(\cdot, \cdot, \cdot) \|_{L^2(1/\rho_\infty^\varepsilon)} \cdot \int_0^t \left\| \nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}} \left(e^{s\mathcal{L}_0} \varphi \right) \right\|_{L^2(\rho_\infty^\varepsilon)} \mathrm{d}s.$$

Since W is of compact support, by (4.29), we have

$$\int_0^t \left\| \nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}} \left(e^{s\mathcal{L}_0} \varphi \right) \right\|_{L^2(\rho_\infty^\varepsilon)} \mathrm{d}s \leq C \int_0^t e^{-\lambda s} \mathrm{d}s \leq C'.$$

Here, C' is a constant independent of t and ε. Hence,

$$\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^{3d}} \left[\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}} \left(e^{(t-s)\mathcal{L}_0} \varphi \right) \right] \rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{z} \mathrm{d}s \right|$$

$$\leq 2C' \| \rho_0(\cdot, \cdot, \cdot) - \rho_\infty^\varepsilon(\cdot, \cdot, \cdot) \|_{L^2(1/\rho_\infty^\varepsilon)},$$

and the right-hand side tends to 0 as $\varepsilon \to 0$. Thus, (4.30) holds, and the convergence is uniform in t.

For (ii), by (4.29), for any $0 < T_1 < T_2$, we have

$$\left| \int_{T_1}^{T_2} \int_{\mathbb{R}^{3d}} \left[\nabla_{\mathbf{q}} W(\mathbf{q}) \cdot \nabla_{\mathbf{p}} \left(e^{(t-s)\mathcal{L}_0} \varphi \right) \right] \rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z}) \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{z} \mathrm{d}s \right|$$

$$\leq \| \nabla_{\mathbf{q}} W \|_{L^2(\rho_0)} \cdot \int_{T_1}^{T_2} e^{-\lambda s} \mathrm{d}s \leq C (e^{-T_1 \lambda} - e^{-T_2 \lambda}).$$

Here, C is a constant depending on V, W, and α. Thus, (4.31) holds, and the convergence in t is exponentially fast.

Next, for (iii), this is a consequence of convergence in L^1. For any given $\varepsilon > 0$, we have

$$\left| \frac{1}{\varepsilon} \left(\int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}) \rho_\infty^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}) \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{z} - \int_{\mathbb{R}^{3d}} \varphi(\mathbf{q}, \mathbf{p}, \mathbf{z}) \rho^\varepsilon(\mathbf{q}, \mathbf{p}, \mathbf{z}, t) \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{z} \right) \right|$$

$$\leq \frac{1}{\varepsilon} \| \varphi \|_{L^\infty} \| \rho^\varepsilon(\cdot, \cdot, \cdot, t) - \rho_\infty^\varepsilon(\cdot, \cdot, \cdot) \|_{L^1}$$

$$\leq C_1 e^{-r_\varepsilon t} / \varepsilon.$$

Here, $r_\varepsilon > 0$ and $C_1 > 0$ are constants depending on ε. Thus, we have convergence as $t \to \infty$.

Finally, for (iv), by (i), for any $\eta > 0$, there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$, we have

$$\left| R(t, \varepsilon; \varphi) - \int_0^t \int_{\mathbb{R}^{3d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}} (e^{s\mathcal{L}_0} \varphi))] \rho_0 \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{z} \mathrm{d}s \right| < \eta.$$

Passing the limit $t \to \infty$, we have

$$\left| \lim_{t \to \infty} R(t, \varepsilon; \varphi) - \int_0^\infty \int_{\mathbb{R}^{3d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}} (e^{s\mathcal{L}_0} \varphi))] \rho_0 \mathrm{d}\mathbf{q} \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{z} \mathrm{d}s \right| \leq \eta.$$

Thus, we conclude

$$\lim_{\varepsilon \to 0^+} \lim_{t \to \infty} R(t, \varepsilon; \varphi) = \int_0^\infty \int_{\mathbb{R}^{3d}} [\nabla_{\mathbf{q}} W \cdot (\nabla_{\mathbf{p}}(e^{s\mathcal{L}_0}\varphi))]\rho_0 d\mathbf{q}d\mathbf{p}d\mathbf{z}ds.$$

By (iii), this also implies (4.35). This concludes the proof. □

Appendix A: Omitted Proofs for Reversibility Results

This appendix provides some preliminary definitions and the omitted proofs for the reversibility of overdamped and underdamped Langevin dynamics.

Proof of Lemma 2.1 Suppose that (2.4) with the initial density function $\rho_0(\mathbf{q})$ describes a reversible process. According to Definition 2.3, ρ_0 must be a stationary measure. Consider the distribution of the time-reversed process $\mathbf{q}^*(t)$ with respect to a fixed time T (see Definition 2.3)

$$\mathbf{q}_t^* = \mathbf{q}_{T-t}. \tag{A.1}$$

Because $\rho_0(\mathbf{q})$ is stationary, the time-reversed process $\mathbf{q}_t^*$ has the same distribution as the original process. Thus, for any $\varphi_1, \varphi_2 \in C_b^\infty(\mathbb{R}^d)$, reversibility implies

$$\mathbb{E}[\varphi_1(\mathbf{q}_t)\varphi_2(\mathbf{q}_0) \mid \mathbf{q}_0 \sim \rho_0] = \mathbb{E}[\varphi_1(\mathbf{q}_t^*)\varphi_2(\mathbf{q}_0^*) \mid \mathbf{q}_0 \sim \rho_0]. \tag{A.2}$$

Using the Markov property and the invariance of ρ_0, for any $0 \leq t \leq T$, we have

$$\begin{aligned} \mathbb{E}[\varphi_1(\mathbf{q}_t^*)\varphi_2(\mathbf{q}_0^*) \mid \mathbf{q}_0 \sim \rho_0] &= \mathbb{E}[\varphi_1(\mathbf{q}_{T-t})\varphi_2(\mathbf{q}_T) \mid \mathbf{q}_0 \sim \rho_0] \\ &= \mathbb{E}[\varphi_1(\mathbf{q}_{T-t})\varphi_2(\mathbf{q}_T) \mid \mathbf{q}_{T-t} \sim \rho_0] \\ &= \mathbb{E}[\varphi_1(\mathbf{q}_0)\varphi_2(\mathbf{q}_t) \mid \mathbf{q}_0 \sim \rho_0]. \end{aligned} \tag{A.3}$$

Thus, we obtain the symmetry

$$\mathbb{E}[\varphi_1(\mathbf{q}_t)\varphi_2(\mathbf{q}_0) \mid \mathbf{q}_0 \sim \rho_0] = \mathbb{E}[\varphi_1(\mathbf{q}_0)\varphi_2(\mathbf{q}_t) \mid \mathbf{q}_0 \sim \rho_0]. \tag{A.4}$$

Conversely, if for arbitrary φ_1 and φ_2, equation (2.8) holds, we can apply smooth functions to approximate simple functions. By the monotone convergence theorem, this allows us to conclude reversibility. □

Proof of Theorem 2.1 We will prove the theorem using the following logical steps

$$(i) \implies (ii) \implies (iv) \implies (iii) \implies (iv) \implies (i).$$

Step 1. $(i) \implies (ii)$.
Denote the solution of (2.4) as $\mathbf{q}_t$. For arbitrary $\varphi_1, \varphi_2 \in C_b^\infty(\mathbb{R}^d)$, we have

$$\begin{aligned} \frac{\mathbb{E}[\varphi_1(\mathbf{q}_t)\varphi_2(\mathbf{q}_0)] - \mathbb{E}[\varphi_1(\mathbf{q}_0)\varphi_2(\mathbf{q}_0)]}{t} &= \frac{\mathbb{E}[(\varphi_1(\mathbf{q}_t) - \varphi_1(\mathbf{q}_0))\varphi_2(\mathbf{q}_0)]}{t} \\ &= \frac{\mathbb{E}[\mathbb{E}[(\varphi_1(\mathbf{q}_t) - \varphi_1(\mathbf{q}_0))\varphi_2(\mathbf{q}_0) \mid \mathbf{q}_0]]}{t} \end{aligned}$$

$$= \frac{\mathbb{E}[\varphi_2(\mathbf{q}_0)\mathbb{E}[(\varphi_1(\mathbf{q}_t) - \varphi_1(\mathbf{q}_0)) \mid \mathbf{q}_0]]}{t}$$

$$= \int_{\mathbb{R}^d} \varphi_2(\mathbf{q}) \frac{\mathbb{E}[(\varphi_1(\mathbf{q}_t) - \varphi_1(\mathbf{q}_0)) \mid \mathbf{q}_0 = \mathbf{q}]}{t} \rho_0(\mathbf{q}) \, \mathrm{d}\mathbf{q}.$$

(A.5)

Similarly, we have

$$\frac{\mathbb{E}[\varphi_2(\mathbf{q}_t)\varphi_1(\mathbf{q}_0)] - \mathbb{E}[\varphi_2(\mathbf{q}_0)\varphi_1(\mathbf{q}_0)]}{t} = \int_{\mathbb{R}^d} \varphi_1(\mathbf{q}) \frac{\mathbb{E}[(\varphi_2(\mathbf{q}_t) - \varphi_2(\mathbf{q}_0)) \mid \mathbf{q}_0 = \mathbf{q}]}{t} \rho_0(\mathbf{q}) \, \mathrm{d}\mathbf{q}.$$

(A.6)

By Definition 2.3, we know

$$\mathbb{E}[\varphi_1(\mathbf{q}_t)\varphi_2(\mathbf{q}_0)] = \mathbb{E}[\varphi_1(\mathbf{q}_0)\varphi_2(\mathbf{q}_t)].$$

Therefore, we have

$$\int_{\mathbb{R}^d} \varphi_2(\mathbf{q}) \frac{\mathbb{E}[(\varphi_1(\mathbf{q}_t) - \varphi_1(\mathbf{q}_0)) \mid \mathbf{q}_0 = \mathbf{q}]}{t} \rho_0(\mathbf{q}) \, \mathrm{d}\mathbf{q}$$

$$= \int_{\mathbb{R}^d} \varphi_1(\mathbf{q}) \frac{\mathbb{E}[(\varphi_2(\mathbf{q}_t) - \varphi_2(\mathbf{q}_0)) \mid \mathbf{q}_0 = \mathbf{q}]}{t} \rho_0(\mathbf{q}) \, \mathrm{d}\mathbf{q}.$$

(A.7)

Passing the limit as $t \to 0^+$, using the definition of the generator, we obtain

$$\int_{\mathbb{R}^d} (\mathcal{L}\varphi_1)(\mathbf{q})\varphi_2(\mathbf{q})\rho_0(\mathbf{q}) \, \mathrm{d}\mathbf{q} = \int_{\mathbb{R}^d} (\mathcal{L}\varphi_2)(\mathbf{q})\varphi_1(\mathbf{q})\rho_0(\mathbf{q}) \, \mathrm{d}\mathbf{q},$$

which is exactly (2.11).

Step 2. $(ii) \implies (iv)$.
Rewrite (2.11) as

$$\int_{\mathbb{R}^d} \frac{1}{\rho_0(\mathbf{q})} \rho_0(\mathbf{q})\varphi_1(\mathbf{q})\mathcal{L}^*(\rho_0(\mathbf{q})\varphi_2(\mathbf{q})) \, \mathrm{d}\mathbf{q} = \int_{\mathbb{R}^d} \frac{1}{\rho_0(\mathbf{q})} \rho_0(\mathbf{q})\varphi_2(\mathbf{q})\mathcal{L}^*(\rho_0(\mathbf{q})\varphi_1(\mathbf{q})) \, \mathrm{d}\mathbf{q}.$$

This holds for arbitrary $\varphi_1, \varphi_2 \in C_0^\infty(\mathbb{R}^d)$. Hence, for arbitrary $\phi_1(\mathbf{q}), \phi_2(\mathbf{q}) \in C_0^\infty(\mathbb{R}^d)$, we get

$$\int_{\mathbb{R}^d} \frac{1}{\rho_0(\mathbf{q})} \phi_1(\mathbf{q})\mathcal{L}^*(\phi_2(\mathbf{q})) \, \mathrm{d}\mathbf{q} = \int_{\mathbb{R}^d} \frac{1}{\rho_0(\mathbf{q})} \phi_2(\mathbf{q})\mathcal{L}^*(\phi_1(\mathbf{q})) \, \mathrm{d}\mathbf{q}.$$

Let $1/\rho_0(\mathbf{q}) = e^{U(\mathbf{q})}$ for some $U(\mathbf{q}) : \mathbb{R}^d \to \mathbb{R}$. Substituting (2.5), we have

$$\int_{\mathbb{R}^d} e^{U(\mathbf{q})} \phi_1(\mathbf{q}) \left[-\nabla \cdot (\phi_2 \mathbf{b}) + \frac{1}{2}\nabla^2 : (\phi_2 \boldsymbol{\sigma}\boldsymbol{\sigma}^T) \right] \mathrm{d}\mathbf{q}$$

$$= \int_{\mathbb{R}^d} e^{U(\mathbf{q})} \phi_2(\mathbf{q}) \left[-\nabla \cdot (\phi_1 \mathbf{b}) + \frac{1}{2}\nabla^2 : (\phi_1 \boldsymbol{\sigma}\boldsymbol{\sigma}^T) \right] \mathrm{d}\mathbf{q}.$$

Integration by parts yields

$$
\int_{\mathbb{R}^d} e^{U(\mathbf{q})} \phi_2(\mathbf{q}) \nabla \phi_1(\mathbf{q}) \cdot \mathbf{b}(\mathbf{q}) + \frac{1}{2} \nabla^2 (e^{U(\mathbf{q})} \phi_1(\mathbf{q})) : (\phi_2 \boldsymbol{\sigma} \boldsymbol{\sigma}^T) \, d\mathbf{q}
$$
$$
= \int_{\mathbb{R}^d} e^{U(\mathbf{q})} \phi_1(\mathbf{q}) \nabla \phi_2(\mathbf{q}) \cdot \mathbf{b}(\mathbf{q}) + \frac{1}{2} \nabla^2 (e^{U(\mathbf{q})} \phi_2(\mathbf{q})) : (\phi_1 \boldsymbol{\sigma} \boldsymbol{\sigma}^T) \, d\mathbf{q}. \tag{A.8}
$$

Next, using the fact that $\boldsymbol{\sigma} \boldsymbol{\sigma}^T$ is constant, we obtain

$$
\nabla^2 (e^{U(\mathbf{q})} \phi_1(\mathbf{q})) : (\phi_2 \boldsymbol{\sigma} \boldsymbol{\sigma}^T) - \nabla^2 (e^{U(\mathbf{q})} \phi_2(\mathbf{q})) : (\phi_1 \boldsymbol{\sigma} \boldsymbol{\sigma}^T)
$$
$$
= 2[\boldsymbol{\sigma} \boldsymbol{\sigma}^T \nabla (e^{U(\mathbf{q})})] \cdot (\phi_2(\mathbf{q}) \nabla \phi_1(\mathbf{q}) - \phi_1(\mathbf{q}) \nabla \phi_2(\mathbf{q}))
$$
$$
+ e^{U(\mathbf{q})} \nabla \cdot (\boldsymbol{\sigma} \boldsymbol{\sigma}^T (\phi_2(\mathbf{q}) \nabla \phi_1(\mathbf{q}) - \phi_1(\mathbf{q}) \nabla \phi_2(\mathbf{q}))).
$$

Substituting this into (A.8) and integrating by parts again, we find

$$
\int_{\mathbb{R}^d} e^{U(\mathbf{q})} (\phi_2(\mathbf{q}) \nabla \phi_1(\mathbf{q}) - \phi_1(\mathbf{q}) \nabla \phi_2(\mathbf{q})) \cdot \left(\mathbf{b} + \frac{1}{2} \boldsymbol{\sigma} \boldsymbol{\sigma}^T \nabla U(\mathbf{q}) \right) \, d\mathbf{q} = 0.
$$

By Lemma A.1 and the smoothness of $\mathbf{b}$ and ρ_0, we obtain

$$
\mathbf{b} + \frac{1}{2} \boldsymbol{\sigma} \boldsymbol{\sigma}^T \nabla U(\mathbf{q}) = 0.
$$

Since $\boldsymbol{\sigma}$ is nonsingular, we have

$$
-\nabla U(\mathbf{q}) = 2(\boldsymbol{\sigma} \boldsymbol{\sigma}^T)^{-1} \mathbf{b}. \tag{A.9}
$$

Thus, $\rho_0(\mathbf{q}) = e^{-U(\mathbf{q})}$, which proves (iv).

Step 3. $(iv) \implies (iii)$.

If (iv) holds, we can derive that $\mathcal{L}$ generates a strongly continuous semigroup on $L^2(\rho_0 \, d\mathbf{q})$ by Hille-Yosida's theorem (see proof of Lemma A.2). Therefore, the diffusion process is well-defined, and $\mathbf{x}(t)$ admits a probability density $\rho(\mathbf{q}, t)$, which is the unique local solution of (2.5) with initial value $\rho(\mathbf{q}, 0) = \rho_0$ (see [3]).

Notice

$$
\mathcal{L}^* \rho_0 = \frac{1}{Z} \nabla \cdot \left(-e^{-U(\mathbf{q})} \mathbf{b}(\mathbf{q}) + \frac{1}{2} \boldsymbol{\sigma} \boldsymbol{\sigma}^T \nabla e^{-U(\mathbf{q})} \right)
$$
$$
= -\frac{1}{Z} \nabla \cdot \left(e^{-U(\mathbf{q})} \frac{1}{2} \boldsymbol{\sigma} \boldsymbol{\sigma}^T \nabla U(\mathbf{q}) - e^{-U(\mathbf{q})} \frac{1}{2} \boldsymbol{\sigma} \boldsymbol{\sigma}^T \nabla U(\mathbf{q}) \right)
$$
$$
= 0.
$$

Thus, the unique local solution of (2.5) with initial value $\rho_0(\mathbf{q})$ is exactly $\rho(\mathbf{q}, t) = \rho_0(\mathbf{q})$, which is actually a global solution. Hence

$$
\mathbf{j}(\rho(t)) = \mathbf{j}(\rho_0) = \mathbf{b} \rho_0 - \frac{1}{2} \boldsymbol{\sigma} \boldsymbol{\sigma}^T \nabla \rho_0 = \mathbf{0}. \tag{A.10}
$$

Therefore, $\mathbf{j}(\rho) = \mathbf{0}$, i.e., the probability flux is zero.

Step 4. $(iii) \implies (iv)$.
Since $\mathbf{j}(\rho(t)) = 0$, we know that ρ_0 satisfies

$$\mathbf{b}(\mathbf{q})\rho_0(\mathbf{q}) - \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla\rho_0(\mathbf{q}) = 0.$$

Since $\rho_0(\mathbf{q}) > 0$, we have

$$\mathbf{b}(\mathbf{q}) = \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla\log(\rho_0(\mathbf{q}))}{2}.$$

Thus, (iv) is proved by setting $U(\mathbf{q}) = -\log(\rho_0(\mathbf{q}))$.
Step 5. $(iv) \implies (i)$.
From the proof of Lemma A.2 and Hille-Yosida's theorem, we know that $\mathcal{L}$ generates a strongly continuous semigroup on $L^2(\rho_0\,\mathrm{d}\mathbf{q})$, denoted as $e^{t\mathcal{L}}$, for $t \geq 0$. Therefore, for any $\varphi_1 \in C_0^\infty(\mathbb{R}^d)$, $(e^{t\mathcal{L}}\varphi_1)(\mathbf{q}, t)$ is the solution of (2.6) with initial value $f(\mathbf{q}, 0) = \varphi_1(\mathbf{q})$.
Direct calculation yields that for arbitrary f and $g \in C_0^\infty(\mathbb{R}^d)$

$$\int_{\mathbb{R}^d} \rho_0(\mathbf{q})g(\mathbf{q})(\mathcal{L}f)(\mathbf{q})\,\mathrm{d}\mathbf{q}$$

$$= \frac{1}{Z}\int_{\mathbb{R}^d} e^{-U(\mathbf{q})}g(\mathbf{q})\left(\mathbf{b}\cdot\nabla f(\mathbf{q}) + \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T : \nabla^2 f(\mathbf{q})\right)\mathrm{d}\mathbf{q}$$

$$= \frac{1}{Z}\int_{\mathbb{R}^d} g(\mathbf{q})\left(\frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla e^{-U(\mathbf{q})}}{2}\cdot\nabla f(\mathbf{q}) + e^{-U(\mathbf{q})}\frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T : \nabla^2 f(\mathbf{q})\right)\mathrm{d}\mathbf{q}$$

$$= \frac{1}{2Z}\int_{\mathbb{R}^d} g(\mathbf{q})\nabla\cdot(e^{-U(\mathbf{q})}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla f(\mathbf{q}))\,\mathrm{d}\mathbf{q}$$

$$= -\frac{1}{2Z}\int_{\mathbb{R}^d} e^{-U(\mathbf{q})}(\nabla g(\mathbf{q}))^T\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla f(\mathbf{q})\,\mathrm{d}\mathbf{q}.$$

The last equality is symmetric with respect to f and g, thus $\mathcal{L}$ is symmetric in $C_0^\infty(\mathbb{R}^d) \subset L^2(\rho_0\,\mathrm{d}\mathbf{q})$. By [3, Corollary 7.3.2], the dual semigroup of $e^{t\mathcal{L}}$ in $L^2(\rho_0\,\mathrm{d}\mathbf{q})$ (denoted as $\widetilde{e^{t\mathcal{L}}}$) is generated by the dual operator of $\mathcal{L}$ in $L^2(\rho_0\,\mathrm{d}\mathbf{q})$ (denoted as $\widetilde{\mathcal{L}}$).

Since $\mathcal{L}$ is symmetric on $C_0^\infty(\mathbb{R}^d)$, we know $\widetilde{e^{t\mathcal{L}}}\varphi = e^{t\mathcal{L}}\varphi$ for any $\varphi \in C_0^\infty(\mathbb{R}^d)$. Thus, for arbitrary $\varphi_1, \varphi_2 \in C_0^\infty(\mathbb{R}^d)$, we have

$$\int_{\mathbb{R}^d} \rho_0(\mathbf{q})\varphi_2(\mathbf{q})(e^{t\mathcal{L}}\varphi_1)(\mathbf{q}, t)\,\mathrm{d}\mathbf{q} = \int_{\mathbb{R}^d} \rho_0(\mathbf{q})\varphi_1(\mathbf{q})(\widetilde{e^{t\mathcal{L}}}\varphi_2)(\mathbf{q}, t)\,\mathrm{d}\mathbf{q}$$

$$= \int_{\mathbb{R}^d} \rho_0(\mathbf{q})\varphi_1(\mathbf{q})(e^{t\mathcal{L}}\varphi_2)(\mathbf{q}, t)\,\mathrm{d}\mathbf{q}.$$

Thus, $\mathbb{E}[\varphi_1(\mathbf{q}_0)\varphi_2(\mathbf{q}_t) \mid \mathbf{q}(0) \sim \rho_0] = \mathbb{E}[\varphi_2(\mathbf{q}_0)\varphi_1(\mathbf{q}_t) \mid \mathbf{q}(0) \sim \rho_0]$, which proves reversibility. $\qquad\square$

Proof of Lemma 2.2 First, if the process with initial density function $\rho_0(\mathbf{q}, \mathbf{p})$ is a reversible process, then ρ_0 is a stationary measure. Consider the distribution of the time-reversed process $(\mathbf{q}^*(t), \mathbf{p}^*(t))$ with respect to a fixed time T, as defined in Definition 2.3. Since $\mathbf{q}$ is even and $\mathbf{p}$ is odd, we have

$$\mathbf{q}_t^* = \mathbf{q}_{T-t}, \quad \mathbf{p}_t^* = -\mathbf{p}_{T-t}. \tag{A.11}$$

Since ρ_0 is invariant, it follows that

$$(\mathbf{q}_t^*, \mathbf{p}_t^*) = (\mathbf{q}_{T-t}, -\mathbf{p}_{T-t}) \sim \rho_0(\mathbf{q}, -\mathbf{p}), \tag{A.12}$$

i.e., the distribution of the time-reversed process has density $\rho_0(\mathbf{q}, -\mathbf{p})$.

Let $\widetilde{\rho}_0(\mathbf{q}, \mathbf{p}) = \rho_0(\mathbf{q}, -\mathbf{p})$. For any $\varphi_1, \varphi_2, \psi_1, \psi_2 \in C_b^\infty(\mathbb{R}^d)$, the reversibility condition implies

$$\mathbb{E}[\varphi_1(\mathbf{q}_0)\varphi_2(\mathbf{p}_0)\psi_1(\mathbf{q}_t)\psi_2(\mathbf{p}_t) \mid (\mathbf{q}_0, \mathbf{p}_0) \sim \rho_0]$$
$$= \mathbb{E}[\varphi_1(\mathbf{q}_0^*)\varphi_2(\mathbf{p}_0^*)\psi_1(\mathbf{q}_t^*)\psi_2(\mathbf{p}_t^*) \mid (\mathbf{q}_0, \mathbf{p}_0) \sim \rho_0].$$

Let $\widetilde{\varphi}_2(\mathbf{p}) = \varphi_2(-\mathbf{p})$ and $\widetilde{\psi}_2(\mathbf{p}) = \psi_2(-\mathbf{p})$. Using the Markov property and the invariance of ρ_0, for any $t \in [0, T]$, we have

$$\mathbb{E}[\varphi_1(\mathbf{q}_0^*)\varphi_2(\mathbf{p}_0^*)\psi_1(\mathbf{q}_t^*)\psi_2(\mathbf{p}_t^*) \mid (\mathbf{q}_0, \mathbf{p}_0) \sim \rho_0]$$
$$= \mathbb{E}[\varphi_1(\mathbf{q}_T)\varphi_2(-\mathbf{p}_T)\psi_1(\mathbf{q}_{T-t})\psi_2(-\mathbf{p}_{T-t}) \mid (\mathbf{q}_0, \mathbf{p}_0) \sim \rho_0]$$
$$= \mathbb{E}[\varphi_1(\mathbf{q}_T)\widetilde{\varphi}_2(\mathbf{p}_T)\psi_1(\mathbf{q}_{T-t})\widetilde{\psi}_2(\mathbf{p}_{T-t}) \mid (\mathbf{q}_0, \mathbf{p}_0) \sim \rho_0]$$
$$= \mathbb{E}[\varphi_1(\mathbf{q}_T)\widetilde{\varphi}_2(\mathbf{p}_T)\psi_1(\mathbf{q}_{T-t})\widetilde{\psi}_2(\mathbf{p}_{T-t}) \mid (\mathbf{q}_{T-t}, \mathbf{p}_{T-t}) \sim \rho_0]$$
$$= \mathbb{E}[\varphi_1(\mathbf{q}_t)\widetilde{\varphi}_2(\mathbf{p}_t)\psi_1(\mathbf{q}_0)\widetilde{\psi}_2(\mathbf{p}_0) \mid (\mathbf{q}_0, \mathbf{p}_0) \sim \rho_0].$$

Therefore, equation (2.16) holds.

Conversely, if equation (2.16) holds for arbitrary φ_1 and φ_2, then by using smooth functions to approximate simple functions, one can conclude reversibility through the monotone convergence theorem. Hence, we employ (2.16) to prove reversibility instead of checking the definition directly. $\square$

Lemma A.1 *For any $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_d) \in C_0^\infty(\mathbb{R}^d; \mathbb{R}^d)$ with $\phi_i \geq 0$, $i = 1, 2, \ldots, d$, there exist $f, g \in C_0^\infty(\mathbb{R}^d)$ such that*

$$f\nabla g - g\nabla f = \boldsymbol{\phi}.$$

Proof For each i, let $f_i = \sqrt{\phi_i}$ and $g_i = x_i f_i$. Then

$$f_i\nabla g_i - g_i\nabla f_i = f_i^2\nabla\left(\frac{g_i}{f_i}\right) = \phi_i \boldsymbol{e}_i,$$

where $\boldsymbol{e}_i = (0, 0, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^d$ has 1 in the i-th position and 0 elsewhere. Thus, $\boldsymbol{\phi} = \sum_{i=1}^d f_i\nabla g_i - g_i\nabla f_i$. $\square$

Proof of Theorem 2.2 **Step 1.** We prove $(i) \implies (ii)$.

Because the process is reversible (and therefore stationary), $(\mathbf{q}_0^*, \mathbf{p}_0^*)$ and $(\mathbf{q}_0, \mathbf{p}_0)$ have the same distribution. Thus $(\mathbf{q}_0^*, \mathbf{p}_0^*) = (\mathbf{q}_T, -\mathbf{p}_T)$ has the same distribution as $(\mathbf{q}(0), \mathbf{p}(0))$, meaning that $\rho_0(\mathbf{q}, \mathbf{p}) = \rho_0(\mathbf{q}, -\mathbf{p})$.

From the definition of reversibility, we know that (i) is equivalent to

$$\mathbb{E}[\varphi_1(\mathbf{q}_0)\varphi_2(\mathbf{p}_0)\psi_1(\mathbf{q}_t)\psi_2(\mathbf{p}_t) \mid (\mathbf{q}_0, \mathbf{p}_0) \sim \rho_0]$$
$$= \mathbb{E}[\varphi_1(\mathbf{q}_t)\varphi_2(-\mathbf{p}_t)\psi_1(\mathbf{q}_0)\psi_2(-\mathbf{p}_0) \mid (\mathbf{q}_0, \mathbf{p}_0) \sim \rho_0] \tag{A.13}$$

for any $\varphi_1(\mathbf{q}), \varphi_2(\mathbf{p}), \psi_1(\mathbf{q}), \psi_2(\mathbf{p}) \in C_0^\infty(\mathbb{R}^d)$.

By similar calculations to those in (A.5), we know

$$
\iint \varphi_1(\mathbf{q})\varphi_2(\mathbf{p}) \frac{\mathbb{E}\left[\psi_1(\mathbf{q}_t)\psi_2(\mathbf{p}_t) - \psi_1(\mathbf{q}_0)\psi_2(\mathbf{p}_0) \mid \mathbf{q}_0 = \mathbf{q}, \mathbf{p}_0 = \mathbf{p}\right]}{t} \rho_0(\mathbf{q}, \mathbf{p}) \, d\mathbf{q} \, d\mathbf{p}
$$
$$
= \iint \psi_1(\mathbf{q})\psi_2(-\mathbf{p}) \frac{\mathbb{E}\left[\varphi_1(\mathbf{q}_t)\varphi_2(-\mathbf{p}_t) - \varphi_1(\mathbf{q}_0)\varphi_2(-\mathbf{p}_0) \mid \mathbf{q}_0 = \mathbf{q}, \mathbf{p}_0 = \mathbf{p}\right]}{t} \rho_0(\mathbf{q}, \mathbf{p}) \, d\mathbf{q} \, d\mathbf{p}.
$$

Taking the limit as $t \to 0^+$, by the definition of the generator for the process $\mathbf{q}_t$, $\mathbf{p}_t$, we conclude that (i) implies

$$
\iint \varphi_1(\mathbf{q})\varphi_2(\mathbf{p})\mathcal{L}\left[\psi_1(\mathbf{q})\psi_2(\mathbf{p})\right] \rho_0(\mathbf{q}, \mathbf{p}) \, d\mathbf{q} \, d\mathbf{p}
$$
$$
= \iint \mathcal{L}\left[\varphi_1(\mathbf{q})\widetilde{\varphi}_2(\mathbf{p})\right] \psi_1(\mathbf{q})\widetilde{\psi}_2(\mathbf{p})\rho_0(\mathbf{q}, \mathbf{p}) \, d\mathbf{q} \, d\mathbf{p},
$$

for any $\varphi_1(\mathbf{q}), \varphi_2(\mathbf{p}), \psi_1(\mathbf{q}), \psi_2(\mathbf{p}) \in C_0^\infty(\mathbb{R}^d)$, where $\tilde{\varphi}_2(\mathbf{p}) := \varphi_2(-\mathbf{p})$ and $\tilde{\psi}_2(\mathbf{p}) := \psi_2(-\mathbf{p})$.

As a consequence, taking $\psi_1 \equiv \psi_2 \equiv 1$ in this equation, we have

$$
\iint \mathcal{L}\left[\varphi_1(\mathbf{q})\widetilde{\varphi}_2(\mathbf{p})\right] \rho_0(\mathbf{q}, \mathbf{p}) \, d\mathbf{q} \, d\mathbf{p} = 0, \quad \forall \varphi_1, \varphi_2 \in C_0^\infty(\mathbb{R}^d),
$$

which implies that $\rho_0(\mathbf{q}, \mathbf{p})$, satisfying $\mathcal{L}^*\rho_0 = 0$, is an invariant measure. Therefore, we conclude that $(i) \implies (ii)$.

Step 2. We prove (ii) is equivalent to (iii).

Let $H(\mathbf{q}, \mathbf{p}) = -\frac{1}{\beta} \ln \rho_0$. Then (ii) is equivalent to

$$
\iint \varphi_1(\mathbf{q})\varphi_2(\mathbf{p})\mathcal{L}\left[\psi_1(\mathbf{q})\psi_2(\mathbf{p})\right] e^{-\beta H(\mathbf{q}, \mathbf{p})} \, d\mathbf{q} \, d\mathbf{p}
$$
$$
= \iint \mathcal{L}\left[\varphi_1(\mathbf{q})\widetilde{\varphi}_2(\mathbf{p})\right] \psi_1(\mathbf{q})\widetilde{\psi}_2(\mathbf{p})e^{-\beta H(\mathbf{q}, \mathbf{p})} \, d\mathbf{q} \, d\mathbf{p},
$$

for any $\varphi_1(\mathbf{q}), \varphi_2(\mathbf{p}), \psi_1(\mathbf{q}), \psi_2(\mathbf{p}) \in C_0^\infty(\mathbb{R}^d)$. Substituting the operator $\mathcal{L}$ as defined in (2.15) and using $\widetilde{\varphi}_2(\mathbf{p}) = \varphi_2(-\mathbf{p})$ and $\widetilde{\psi}_2(\mathbf{p}) = \psi_2(-\mathbf{p})$, we obtain

$$
\iint \varphi_1(\mathbf{q})\varphi_2(\mathbf{p})\Bigg[\psi_2(\mathbf{p})\mathbf{p} \cdot \nabla_\mathbf{q}\psi_1(\mathbf{q}) + \psi_1(\mathbf{q})\mathbf{b} \cdot \nabla_\mathbf{p}\psi_2(\mathbf{p})
$$
$$
+ \psi_1(\mathbf{q})\left(-\frac{1}{2}(\sigma\sigma^T\mathbf{p}) \cdot \nabla_\mathbf{p}\psi_2(\mathbf{p}) + \frac{1}{2\beta}(\sigma\sigma^T) : \nabla_\mathbf{p}^2\psi_2(\mathbf{p})\right) \Bigg] e^{-\beta H(\mathbf{q}, \mathbf{p})} \, d\mathbf{q} \, d\mathbf{p}
$$
$$
= \iint \psi_1(\mathbf{q})\psi_2(\mathbf{p})\Bigg[-\varphi_2(\mathbf{p})\mathbf{p} \cdot \nabla_\mathbf{q}\varphi_1(\mathbf{q}) - \varphi_1(\mathbf{q})\mathbf{b} \cdot \nabla_\mathbf{p}\varphi_2(\mathbf{p})
$$
$$
+ \varphi_1(\mathbf{q})\left(-\frac{1}{2}(\sigma\sigma^T\mathbf{p}) \cdot \nabla_\mathbf{p}\varphi_2(\mathbf{p}) + \frac{1}{2\beta}(\sigma\sigma^T) : \nabla_\mathbf{p}^2\varphi_2(\mathbf{p})\right) \Bigg] e^{-\beta H(\mathbf{q}, \mathbf{p})} \, d\mathbf{q} \, d\mathbf{p}.
$$

We now separate the Hamiltonian flow terms on the LHS from the Fokker-Planck terms on the RHS. Then, we get

$$I_1 := \iint \left[\varphi_2(\mathbf{p})\psi_2(\mathbf{p})\mathbf{p} \cdot \nabla_{\mathbf{q}}\left(\psi_1(\mathbf{q})\varphi_1(\mathbf{q})\right) + \varphi_1(\mathbf{q})\psi_1(\mathbf{q})\mathbf{b} \cdot \nabla_{\mathbf{p}}\left(\varphi_2(\mathbf{p})\psi_2(\mathbf{p})\right) \right]$$
$$\times\, e^{-\beta H(\mathbf{q},\mathbf{p})}\, d\mathbf{q}\, d\mathbf{p}$$
$$= I_2 := \iint \left[\psi_1(\mathbf{q})\varphi_1(\mathbf{q})\left(\frac{1}{2}\varphi_2(\mathbf{p})(\sigma\sigma^T\mathbf{p}) \cdot \nabla_{\mathbf{p}}\psi_2(\mathbf{p}) - \frac{1}{2\beta}\varphi_2(\mathbf{p})(\sigma\sigma^T) : \nabla_{\mathbf{p}}^2\psi_2(\mathbf{p}) \right.\right.$$
$$\left.\left. -\frac{1}{2}\psi_2(\mathbf{p})(\sigma\sigma^T\mathbf{p}) \cdot \nabla_{\mathbf{p}}\varphi_2(\mathbf{p}) + \frac{1}{2\beta}\psi_2(\mathbf{p})(\sigma\sigma^T) : \nabla_{\mathbf{p}}^2\varphi_2(\mathbf{p}) \right)\right] e^{-\beta H(\mathbf{q},\mathbf{p})}\, d\mathbf{q}\, d\mathbf{p}.$$

Using integration by parts, I_1 simplifies as

$$I_1 = \iint \varphi_1(\mathbf{q})\psi_1(\mathbf{q})\varphi_2(\mathbf{p})\psi_2(\mathbf{p})\left(-\mathbf{p} \cdot \nabla_{\mathbf{q}}e^{-\beta H} - \mathbf{b} \cdot \nabla_{\mathbf{p}}e^{-\beta H}\right)d\mathbf{q}d\mathbf{p}. \qquad (A.14)$$

Similarly, I_2 simplifies as

$$I_2 = \frac{1}{2}\iint \varphi_1(\mathbf{q})\psi_1(\mathbf{q})\Big[e^{-\beta H}(\sigma\sigma^T\mathbf{p}) \cdot \left(\varphi_2(\mathbf{p})\nabla_{\mathbf{p}}\psi_2(\mathbf{p}) - \psi_2(\mathbf{p})\nabla_{\mathbf{p}}\varphi_2(\mathbf{p})\right)$$
$$+ \frac{1}{\beta}\nabla_{\mathbf{p}}\left(\varphi_2(\mathbf{p})e^{-\beta H}\right)(\sigma\sigma^T)\nabla_{\mathbf{p}}\psi_2(\mathbf{p})$$
$$- \frac{1}{\beta}\nabla_{\mathbf{p}}\left(\psi_2(\mathbf{p})e^{-\beta H}\right)(\sigma\sigma^T)\nabla_{\mathbf{p}}\varphi_2(\mathbf{p})\Big]d\mathbf{q}d\mathbf{p} \qquad (A.15)$$
$$= \frac{1}{2}\iint \varphi_1(\mathbf{q})\psi_1(\mathbf{q})\left[\left(e^{-\beta H}(\sigma\sigma^T\mathbf{p}) + \frac{1}{\beta}\sigma\sigma^T\nabla_{\mathbf{p}}e^{-\beta H}\right)\right.$$
$$\left. \cdot\left(\varphi_2(\mathbf{p})\nabla_{\mathbf{p}}\psi_2(\mathbf{p}) - \psi_2(\mathbf{p})\nabla_{\mathbf{p}}\varphi_2(\mathbf{p})\right)\right]d\mathbf{q}d\mathbf{p}.$$

Now, let $\varphi_2 = \psi_2$, then $I_2 = 0$, and from (A.14), we obtain

$$I_1 = \iint \varphi_1(\mathbf{q})\psi_1(\mathbf{q})\varphi_2^2(\mathbf{p})\left(-\mathbf{p} \cdot \nabla_{\mathbf{q}}e^{-\beta H} - \mathbf{b} \cdot \nabla_{\mathbf{p}}e^{-\beta H}\right)d\mathbf{q}d\mathbf{p} = 0$$

for any $\varphi_1(\mathbf{q})$, $\psi_1(\mathbf{q})$, and $\varphi_2(\mathbf{p})$. Thus, we deduce

$$-\mathbf{p} \cdot \nabla_{\mathbf{q}}e^{-\beta H} - \mathbf{b} \cdot \nabla_{\mathbf{p}}e^{-\beta H} = 0. \qquad (A.16)$$

This implies that $I_1 = I_2 = 0$ for arbitrary test functions $\varphi_1(\mathbf{p})$, $\varphi_2(\mathbf{p})$, $\psi_1(\mathbf{q})$, and $\psi_2(\mathbf{q})$. By the auxiliary lemma, equation (A.15) yields

$$e^{-\beta H}(\sigma\sigma^T\mathbf{p}) + \frac{1}{\beta}\sigma\sigma^T\nabla_{\mathbf{p}}e^{-\beta H} = 0.$$

Since σ is nonsingular, we conclude

$$e^{-\beta H}\mathbf{p} + \frac{1}{\beta}\nabla_{\mathbf{p}}e^{-\beta H} = 0 \qquad (A.17)$$

holds for any $\mathbf{p}$.

Next, we use (A.16) and (A.17) to derive $H(\mathbf{q}, \mathbf{p}) = \frac{|\mathbf{p}|^2}{2} + U(\mathbf{q})$ for some potential $U(\mathbf{q})$. Equation (A.17) implies

$$\nabla_{\mathbf{p}} e^{-\beta H + \beta \frac{|\mathbf{p}|^2}{2}} = 0,$$

and thus there exists a potential $U(\mathbf{q})$ such that

$$e^{-\beta H} = e^{-\beta \left(\frac{|\mathbf{p}|^2}{2} + U(\mathbf{q}) \right)}. \tag{A.18}$$

Substituting (A.18) into (A.16), we derive

$$-\mathbf{p} \cdot \nabla_{\mathbf{q}} e^{-\beta H} - \mathbf{b} \cdot \nabla_{\mathbf{p}} e^{-\beta H} = 0, \tag{A.19}$$

which simplifies to

$$\mathbf{p} \cdot \left(\mathbf{b} + \nabla_{\mathbf{q}} U(\mathbf{q}) \right) = 0, \quad \forall \mathbf{p}, \mathbf{q}. \tag{A.20}$$

Therefore, we conclude

$$\mathbf{b} = -\nabla_{\mathbf{q}} U(\mathbf{q}),$$

and $U(\mathbf{q})$ is unique up to a constant. Thus, the invariant measure is given by

$$\rho_0(\mathbf{q}, \mathbf{p}) = \frac{1}{Z} e^{-\beta \left(\frac{|\mathbf{p}|^2}{2} + U(\mathbf{q}) \right)}, \quad Z = \iint e^{-\beta \left(\frac{|\mathbf{p}|^2}{2} + U(\mathbf{q}) \right)} d\mathbf{q} d\mathbf{p}, \tag{A.21}$$

which is the unique invariant measure, and hence (ii) is equivalent to (iii).

Step 3. We prove that (iii) implies (i).

First, from (iii) and for $\mathcal{L}$ defined in (2.15) with $\mathbf{b} = -\nabla_{\mathbf{q}} U(\mathbf{q})$, it is easy to verify (2.17). Now, we use (2.17) and (iii) to derive (i).

From Lemma A.2, $-\mathcal{L}$ is a maximal monotone operator in $L^2(\rho_0 \, d\mathbf{q} d\mathbf{p})$, where

$$\rho_0(\mathbf{q}, \mathbf{p}) = \frac{1}{Z} e^{-\beta \left(\frac{|\mathbf{p}|^2}{2} + U(\mathbf{q}) \right)}.$$

Recall the strongly continuous semigroup generated by $-\mathcal{L}$, denoted by $S(t) = e^{t\mathcal{L}}$.

From [3, p381, Cor 7.3.2], we know that the dual semigroup $\widetilde{S}(t) = \widetilde{(e^{t\mathcal{L}})}$ is generated by the dual operator $-\widetilde{\mathcal{L}}$, which is also a maximal monotone operator. Here, the dual operators are understood in $L^2(\rho_0 \, d\mathbf{q} d\mathbf{p})$. Therefore, we have

$$\iint \varphi_1(\mathbf{q}) \varphi_2(\mathbf{p}) e^{t\mathcal{L}} \left(\psi_1(\mathbf{q}) \psi_2(\mathbf{p}) \right) \rho_0(\mathbf{q}, \mathbf{p}) \, d\mathbf{q} \, d\mathbf{p}$$

$$= \iint \widetilde{(e^{t\mathcal{L}})} \left(\varphi_1(\mathbf{q}) \varphi_2(\mathbf{p}) \right) \psi_1(\mathbf{q}) \psi_2(\mathbf{p}) \rho_0(\mathbf{q}, \mathbf{p}) \, d\mathbf{q} \, d\mathbf{p}$$

$$= \iint \left(e^{t\widetilde{\mathcal{L}}} \right) \left(\varphi_1(\mathbf{q}) \varphi_2(\mathbf{p}) \right) \psi_1(\mathbf{q}) \psi_2(\mathbf{p}) \rho_0(\mathbf{q}, \mathbf{p}) \, d\mathbf{q} \, d\mathbf{p},$$

where $\left(e^{t\widetilde{\mathcal{L}}}\right)(\varphi_1(\mathbf{q})\varphi_2(\mathbf{p}))$ is the solution to

$$\frac{\partial}{\partial t}f = \widetilde{\mathcal{L}}f := -\left(\mathbf{p}\cdot\nabla_\mathbf{q}f + \mathbf{b}\cdot\nabla_\mathbf{p}f\right) + \left(-\frac{1}{2}\left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p}\right)\cdot\nabla_\mathbf{p}f + \frac{1}{2\beta}\left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T\right):\nabla_\mathbf{p}^2 f\right).$$
(A.22)

By changing variables from $\mathbf{p}$ to $-\mathbf{p}$, it is easy to verify

$$\left(e^{t\widetilde{\mathcal{L}}}\right)(\varphi_1(\mathbf{q})\varphi_2(\mathbf{p})) = \left(e^{t\mathcal{L}}\right)(\varphi_1(\mathbf{q})\varphi_2(-\mathbf{p}))\Big|_{(\mathbf{q},-\mathbf{p})} = \left(e^{t\mathcal{L}}\right)(\varphi_1(\mathbf{q})\tilde{\varphi}_2(\mathbf{p}))\Big|_{(\mathbf{q},-\mathbf{p})}. \quad \text{(A.23)}$$

Thus, we have

$$\iint \varphi_1(\mathbf{q})\varphi_2(\mathbf{p})e^{t\mathcal{L}}\left(\psi_1(\mathbf{q})\psi_2(\mathbf{p})\right)\rho_0(\mathbf{q},\mathbf{p})\,d\mathbf{q}\,d\mathbf{p}$$

$$= \iint \left(e^{t\mathcal{L}}\right)(\varphi_1(\mathbf{q})\tilde{\varphi}_2(\mathbf{p}))\Big|_{(\mathbf{q},-\mathbf{p})}\psi_1(\mathbf{q})\psi_2(\mathbf{p})\rho_0(\mathbf{q},\mathbf{p})\,d\mathbf{q}\,d\mathbf{p}$$

$$= \iint (e^{t\mathcal{L}})(\varphi_1(\mathbf{q})\tilde{\varphi}_2(\mathbf{p}))\,\psi_1(\mathbf{q})\psi_2(-\mathbf{p})\rho_0(\mathbf{q},-\mathbf{p})\,\mathrm{d}\mathbf{q}\,\mathrm{d}\mathbf{p}$$

$$= \iint \left(e^{t\mathcal{L}}\right)(\varphi_1(\mathbf{q})\tilde{\varphi}_2(\mathbf{p}))\,\psi_1(\mathbf{q})\tilde{\psi}_2(\mathbf{p})\rho_0(\mathbf{q},\mathbf{p})\,d\mathbf{q}\,d\mathbf{p},$$

and we conclude (i).

Step 4. We prove that (iii) and (iv) are equivalent.

It is obvious that (iii) implies (iv) since $H(\mathbf{q},\mathbf{p}) = H(\mathbf{q},-\mathbf{p})$, and the Gibbs measure solves (2.5). Thus, we just need to prove that (iv) implies (iii). Let $\widetilde{\rho}_0(\mathbf{q},\mathbf{p}) = \rho_0(\mathbf{q},-\mathbf{p})$. Since $\rho(\mathbf{q},\mathbf{p}) = \rho(\mathbf{q},-\mathbf{p})$, they both solve (2.5). Thus

$$-\mathbf{p}\cdot\nabla_\mathbf{q}\rho_0 - \mathbf{b}\cdot\nabla_\mathbf{p}\rho_0 + \frac{1}{2}\nabla_\mathbf{p}\cdot\left(\rho_0\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p} + \frac{1}{\beta}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla_\mathbf{p}\rho_0\right) = 0,$$
$$\text{(A.24)}$$
$$-\mathbf{p}\cdot\nabla_\mathbf{q}\widetilde{\rho}_0 - \mathbf{b}\cdot\nabla_\mathbf{p}\widetilde{\rho}_0 + \frac{1}{2}\nabla_\mathbf{p}\cdot\left(\widetilde{\rho}_0\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p} + \frac{1}{\beta}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla_\mathbf{p}\widetilde{\rho}_0\right) = 0.$$

Substituting $\widetilde{\rho}_0(\mathbf{q},\mathbf{p}) = \rho_0(\mathbf{q},-\mathbf{p})$ into (A.24), we get

$$-\mathbf{p}\cdot\nabla_\mathbf{q}\rho_0 - \mathbf{b}\cdot\nabla_\mathbf{p}\rho_0 + \frac{1}{2}\nabla_\mathbf{p}\cdot\left(\rho_0\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p} + \frac{1}{\beta}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla_\mathbf{p}\rho_0\right) = 0,$$
$$\text{(A.25)}$$
$$\mathbf{p}\cdot\nabla_\mathbf{q}\rho_0 + \mathbf{b}\cdot\nabla_\mathbf{p}\rho_0 + \frac{1}{2}\nabla_\mathbf{p}\cdot\left(\rho_0\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p} + \frac{1}{\beta}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla_\mathbf{p}\rho_0\right) = 0,$$

which implies

$$\mathbf{p}\cdot\nabla_\mathbf{q}\rho_0 + \mathbf{b}\cdot\nabla_\mathbf{p}\rho_0 = 0, \quad \nabla_\mathbf{p}\cdot\left(\rho_0\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p} + \frac{1}{\beta}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla_\mathbf{p}\rho_0\right) = 0. \quad \text{(A.26)}$$

Since $\rho(\mathbf{q},\mathbf{p}) > 0$ and $\rho_0(\mathbf{q},\mathbf{p}) \in L^1(\mathbb{R}^{2d})$, by Fubini's theorem, for almost every $\mathbf{q} \in \mathbb{R}^d$, $m(\mathbf{q}) := \int_{\mathbb{R}^d}\rho_0(\mathbf{q},\mathbf{p})\,d\mathbf{p} \in (0,\infty)$ exists, and $m(\mathbf{q}) \in L^1(\mathbb{R}^d)$. Now, for these $\mathbf{q}$'s, from (A.26), $\rho_0(\mathbf{q},\mathbf{p})$ solves the following Fokker-Planck equation (in the $\mathbf{p}$-variable)

$$\nabla_\mathbf{p}\cdot\left(\rho_0\boldsymbol{\sigma}\boldsymbol{\sigma}^T\mathbf{p} + \frac{1}{\beta}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\nabla_\mathbf{p}\rho_0\right) = 0,$$

and $\rho_0(\mathbf{q}, \mathbf{p})$ is a positive L^1-solution (in the $\mathbf{p}$-variable). Thus, by Proposition 3.2, we know that $e^{-\beta|\mathbf{p}|^2/2}$ is the unique non-zero solution (up to a constant). Hence, there exists $c(\mathbf{q}) > 0$ such that

$$\rho_0 = c(\mathbf{q})e^{-\beta|\mathbf{p}|^2/2}. \tag{A.27}$$

Substituting the above equality into (A.26), we get

$$e^{-\beta|\mathbf{p}|^2/2}\mathbf{p} \cdot \left(\nabla_{\mathbf{q}}c(\mathbf{q}) - c\beta\mathbf{b}\right) = 0.$$

This holds for arbitrary $\mathbf{p}$, so $\mathbf{b} = -\nabla_{\mathbf{q}}\log(c)/\beta$, and (iii) is proved.

   Step 5. We prove that (v) and (iii) are equivalent.

   It is clear that (iii) implies (v), so we only need to prove that (v) implies (iii). First, we have

$$-U_2(\mathbf{p})\,\mathbf{p} \cdot \nabla_{\mathbf{q}}U_1(\mathbf{q}) - U_1(\mathbf{q})\,\mathbf{b} \cdot \nabla_{\mathbf{p}}U_2(\mathbf{p})$$
$$+ \frac{U_1(\mathbf{q})}{2}\nabla_{\mathbf{p}} \cdot \left(U_2(\mathbf{p})\sigma\sigma^T\mathbf{p} + \frac{1}{\beta}\sigma\sigma^T\nabla_{\mathbf{p}}U_2(\mathbf{p})\right) = 0.$$

Dividing by $U_1(\mathbf{q})U_2(\mathbf{p})$, we get

$$-\mathbf{p} \cdot \nabla_{\mathbf{q}}\log U_1(\mathbf{q}) - \mathbf{b} \cdot \nabla_{\mathbf{p}}\log U_2(\mathbf{p}) + \frac{1}{2U_2(\mathbf{p})}\nabla_{\mathbf{p}} \cdot \left(U_2(\mathbf{p})\sigma\sigma^T\mathbf{p} + \frac{1}{\beta}\sigma\sigma^T\nabla_{\mathbf{p}}U_2(\mathbf{p})\right) = 0. \tag{A.28}$$

Notice that the last term only depends on $\mathbf{p}$. Now, evaluating (A.28) at $(\mathbf{q}_1, \mathbf{p})$ and $(\mathbf{q}_2, \mathbf{p})$ and taking the difference yields

$$\mathbf{p} \cdot \nabla_{\mathbf{q}}[\log U_1(\mathbf{q}_1) - \log U_1(\mathbf{q}_2)] + [\mathbf{b}(\mathbf{q}_1) - \mathbf{b}(\mathbf{q}_2)] \cdot \nabla_{\mathbf{p}}\log U_2(\mathbf{p}) = 0.$$

This holds for any $\mathbf{p}, \mathbf{q}_1, \mathbf{q}_2 \in \mathbb{R}^d$. Taking the derivative with respect to $\mathbf{p}$, we get

$$\nabla_{\mathbf{q}}[\log U_1(\mathbf{q}_1) - \log U_1(\mathbf{q}_2)] + \nabla_{\mathbf{p}}^2\log U_2(\mathbf{p})[\mathbf{b}(\mathbf{q}_1) - \mathbf{b}(\mathbf{q}_2)] = 0. \tag{A.29}$$

Notice that the first term only depends on $\mathbf{q}_1$ and $\mathbf{q}_2$, and so does the second term. Thus, for any $\mathbf{p}_1$ and $\mathbf{p}_2$, we have

$$[\nabla_{\mathbf{p}}^2\log U_2(\mathbf{p}_1) - \nabla_{\mathbf{p}}^2\log U_2(\mathbf{p}_2)][\mathbf{b}(\mathbf{q}_1) - \mathbf{b}(\mathbf{q}_2)] = \mathbf{0}. \tag{A.30}$$

Equation (A.30) implies that, for any $\mathbf{p}_1$ and $\mathbf{p}_2$, we must have

$$\nabla_{\mathbf{p}}^2\log U_2(\mathbf{p}_1) = \nabla_{\mathbf{p}}^2\log U_2(\mathbf{p}_2).$$

Thus, $\nabla_{\mathbf{p}}^2\log U_2(\mathbf{p})$ is a constant, meaning there exist $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathbf{a} \in \mathbb{R}^d$, and $c \in \mathbb{R}$, all constant, such that

$$\log U_2(\mathbf{p}) = \mathbf{p}^T\mathbf{A}\mathbf{p} + \mathbf{a} \cdot \mathbf{p} + c. \tag{A.31}$$

We assume that $\mathbf{A}$ is symmetric. Substituting this back into (A.29), we get

$$\nabla_{\mathbf{q}}[\log U_1(\mathbf{q}_1) - \log U_1(\mathbf{q}_2)] + 2\mathbf{A}[\mathbf{b}(\mathbf{q}_1) - \mathbf{b}(\mathbf{q}_2)] = 0,$$

which holds for arbitrary $\mathbf{q}_1$ and $\mathbf{q}_2$. Thus, there exists a constant $\mathbf{c}'$ such that, for any $\mathbf{q} \in \mathbb{R}^d$, we have

$$\nabla_{\mathbf{q}} \log U_1(\mathbf{q}) + 2\mathbf{A}\mathbf{b} = \mathbf{c}'. \tag{A.32}$$

Substituting (A.32) and (A.31) back into (A.28), we get

$$- \mathbf{p} \cdot \mathbf{c}' - \mathbf{b} \cdot \mathbf{a}$$
$$+ \frac{1}{2} \left[(2\mathbf{A}\mathbf{p} + \mathbf{a})^T \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{p} + \mathrm{tr}(\boldsymbol{\sigma} \boldsymbol{\sigma}^T) + \frac{1}{\beta}(2\mathbf{A}\mathbf{p} + \mathbf{a})^T \boldsymbol{\sigma} \boldsymbol{\sigma}(2\mathbf{A}\mathbf{p} + \mathbf{a}) + \frac{2}{\beta} \mathrm{tr}(\boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{A}) \right] = 0.$$

Here, $\mathbf{b}$ is the only term that depends on $\mathbf{q}$. Thus, from the previous argument, we know that $\mathbf{a} = 0$. The above equation holds for all $\mathbf{p} \in \mathbb{R}^d$, so the coefficients are all zero. Hence

$$\mathbf{c}' = \mathbf{0}, \quad \mathbf{A}^T \boldsymbol{\sigma} \boldsymbol{\sigma}^T + \frac{2}{\beta} \mathbf{A}^T \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{A} = \mathbf{0}, \quad \mathrm{tr}(\boldsymbol{\sigma} \boldsymbol{\sigma}^T (2\mathbf{A}/\beta + \mathbf{I}_d)) = 0.$$

Since $\mathbf{A} = \mathbf{A}^T$, the second equality implies that $\mathbf{A}^T \boldsymbol{\sigma} \boldsymbol{\sigma}^T = -2\mathbf{A}^T \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{A}/\beta$ is symmetric, so $\mathbf{A}$ and $\boldsymbol{\sigma} \boldsymbol{\sigma}^T$ are commutative. Thus

$$\boldsymbol{\sigma} \boldsymbol{\sigma}^T (\mathbf{A} + 2\mathbf{A}^2/\beta) = 0. \tag{A.33}$$

Since $\boldsymbol{\sigma} \boldsymbol{\sigma}^T$ is invertible, it follows that $\mathbf{A} + 2\mathbf{A}^2/\beta = \mathbf{0}$. The eigenvalues of $\mathbf{A}$ are either 0 or $-\beta/2$. Therefore, the eigenvalues of $(2\mathbf{A}/\beta + \mathbf{I}_d)$ are either 0 or 1. Suppose the eigenvectors of $(2\mathbf{A}/\beta + \mathbf{I}_d)$ are $\mathbf{u}_i$, $i = 1, 2, \ldots, d$, satisfying $\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij}$, with corresponding eigenvalues λ_i, $i = 1, 2, \ldots, d$. Then $\lambda_i = 0$ or 1, and

$$0 = \mathrm{tr}(\boldsymbol{\sigma} \boldsymbol{\sigma}^T (2\mathbf{A}/\beta + \mathbf{I}_d)) = \sum_{i=1}^{d} \mathbf{u}_i^T \boldsymbol{\sigma} \boldsymbol{\sigma}^T (2\mathbf{A}/\beta + \mathbf{I}_d) \mathbf{u}_i = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i^T \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{u}_i \geq 0.$$

Thus, $\lambda_i = 0$ for all $i = 1, 2, \ldots, d$, since $\boldsymbol{\sigma}$ is nonsingular. Hence, the only eigenvalue of $\mathbf{A}$ is $-\beta/2$, and therefore $\mathbf{A} = -\beta \mathbf{I}_d/2$. Since $\mathbf{c}' = \mathbf{0}$ from (A.32), we have $\mathbf{b} = \frac{1}{\beta} \nabla_{\mathbf{q}} \log U_1(\mathbf{q})$, which is a gradient. Thus, (iii) is proved. $\qquad \square$

Lemma A.2 *Let $\mathbf{b} = -\nabla_{\mathbf{q}} U(\mathbf{q})$ in (2.15). Consider the operator*

$$\mathcal{L}f = \left(-\mathbf{p} \cdot \nabla_{\mathbf{q}} f - \mathbf{b} \cdot \nabla_{\mathbf{p}} f \right) + \left(-\frac{1}{2}(\boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{p}) \cdot \nabla_{\mathbf{p}} f + \frac{1}{2\beta}(\boldsymbol{\sigma} \boldsymbol{\sigma}^T) : \nabla_{\mathbf{p}}^2 f \right) =: Tf + L_v f. \tag{A.34}$$

Then $-\mathcal{L}$ is a maximal monotone operator in $L^2(\rho_0 \, \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p})$ with the weight $\rho_0(\mathbf{q}, \mathbf{p})$ defined in (2.19).

Proof First, observe that

$$\langle -Tf, f \rangle_{L^2(\rho_0 \, \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p})} = 0, \quad \langle -L_v f, f \rangle_{L^2(\rho_0 \, \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p})} \geq 0, \tag{A.35}$$

which shows that both $-T$ and $-L_v$ are monotone in $L^2(\rho_0 \, \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p})$.

Second, we prove that $\text{Ran}(I - L_v) = L^2(\rho_0 \, d\mathbf{q}d\mathbf{p})$, i.e., given any $h \in L^2(\rho_0 \, d\mathbf{q}d\mathbf{p})$, there exists a solution $f \in D(-L_v) \subset L^2(\rho_0 \, d\mathbf{q}d\mathbf{p})$ such that $(I - L_v)f = h$. Indeed, define a weighted Hilbert space $H^1(e^{-\frac{1}{2}|\mathbf{p}|^2} d\mathbf{p})$ with the norm

$$\|f\|_1 := \int \rho_0 f^2 \, d\mathbf{p} + \int \rho_0 |\nabla f|^2 \, d\mathbf{p},$$

and a bilinear form on $H^1(e^{-\frac{1}{2}|\mathbf{p}|^2} d\mathbf{p}) \times H^1(e^{-\frac{1}{2}|\mathbf{p}|^2} d\mathbf{p})$:

$$a(u, v) := \langle (I - L_v)u, v \rangle_{L^2(\rho_0)}.$$

It is clear that $a(u, v)$ is a coercive and bounded bilinear form on $H^1(e^{-\frac{1}{2}|\mathbf{p}|^2} d\mathbf{p}) \times H^1(e^{-\frac{1}{2}|\mathbf{p}|^2} d\mathbf{p})$. Therefore, for any $\mathbf{q}$, we conclude the existence of $(I - L_v)f(\mathbf{q}, \cdot) = h(\mathbf{q}, \cdot)$ by the Lax-Milgram theorem.

Third, we show that the sum of $-T$ and $-L_v$ is a maximal monotone operator by using [1, Cor 2.6]. From the above, we have already concluded that $-L_v$ is a maximal monotone operator in $L^2(\rho_0 \, d\mathbf{q}d\mathbf{p})$ and $-T$ is monotone. It remains to verify that $-T$ is demicontinuous (see [16]), i.e., if $f_n \in D(-T)$ converges strongly to $f \in D(-T)$ in $L^2(\rho_0 \, d\mathbf{q}d\mathbf{p})$, then $-Tf_n$ converges weak* to $-Tf$ in $(L^2(\rho_0 \, d\mathbf{q}d\mathbf{p}))^*$. It is clear that $-T$ is a linear operator in $L^2(\rho_0 \, d\mathbf{q}d\mathbf{p})$ and thus demicontinuous. $\qquad\square$

Remark A.1 Define the Hamiltonian $\mathcal{H}(\rho, \xi) := \langle e^{-\xi} \mathcal{L} e^{\xi}, \rho \rangle$. Then the symmetry condition (ii) for $\mathcal{L}$ is equivalent to

$$\mathcal{H}(\rho, \xi) = \mathcal{H}\left(\rho, \log \frac{\rho}{\rho_0} - \xi\right), \quad \forall \xi, \rho \in C_c^\infty(\mathbb{R}^d). \tag{A.36}$$

Indeed, if $\mathcal{L}$ is symmetric with respect to ρ_0, then

$$\mathcal{H}\left(\rho, \log \frac{\rho}{\rho_0} - \xi\right) = \left\langle \frac{\rho_0}{\rho} e^\xi \mathcal{L}\left(\frac{\rho}{\rho_0} e^{-\xi}\right), \rho \right\rangle$$

$$= \left\langle e^\xi, \mathcal{L}\left(\frac{\rho}{\rho_0} e^{-\xi}\right) \right\rangle_{\rho_0} = \langle \mathcal{L} e^\xi, \rho e^{-\xi} \rangle = \mathcal{H}(\rho, \xi).$$

If (A.36) holds, let

$$\xi = \log(\phi_2), \quad \rho = \phi_1 \phi_2 \rho_0,$$

then:

$$\mathcal{H}\left(\rho, \log \frac{\rho}{\rho_0} - \xi\right) = \langle \phi_2, \mathcal{L}\phi_1 \rangle_{\rho_0}, \quad \mathcal{H}(\rho, \xi) = \langle \phi_1, \mathcal{L}\phi_2 \rangle_{\rho_0}.$$

Therefore, (A.36) yields the symmetry of $\mathcal{L}$ with respect to $\langle \cdot, \cdot \rangle_{\rho_0}$.

Proof of Lemma 3.3 We first prove part (i). To start, we show that $-\mathcal{L}_\varepsilon^*$ is maximal accretive. Define

$$H := \left\{ u \mid \frac{u}{\rho_\infty^\varepsilon} \in H^1(\rho_\infty^\varepsilon) \right\},$$

and equip H with the following inner product

$$\langle u, v \rangle_H = \left\langle \frac{u}{\rho_\infty^\varepsilon}, \frac{v}{\rho_\infty^\varepsilon} \right\rangle_{H^1(\rho_\infty^\varepsilon)}.$$

Clearly, $H \subset L^2(1/\rho_\infty^\varepsilon)$, and H is a Hilbert space. Let the domain of $\mathcal{L}_\varepsilon^*$ be

$$D(\mathcal{L}_\varepsilon^*) := \{ u \mid u \in H, \ \mathcal{L}_\varepsilon^* u \in L^2(1/\rho_\infty^\varepsilon) \} \subset L^2(1/\rho_\infty^\varepsilon).$$

Define $\mathbf{b} = -\nabla V + \varepsilon \mathbf{M}$. Then, for any $u \in D(\mathcal{L}_\varepsilon^*)$, we have

$$
\begin{aligned}
\langle u, \mathcal{L}_\varepsilon^* u \rangle_{L^2(1/\rho_\infty^\varepsilon)} &= \int_{\mathbb{R}^d} \frac{u}{\rho_\infty^\varepsilon} \mathcal{L}_\varepsilon^* u \, d\mathbf{q} \\
&= \int_{\mathbb{R}^d} \frac{u}{\rho_\infty^\varepsilon} \left[\nabla \cdot \left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T \left(-\mathbf{b}\rho_\infty^\varepsilon \frac{u}{\rho_\infty^\varepsilon} + \nabla \left(\rho_\infty^\varepsilon \cdot \frac{u}{\rho_\infty^\varepsilon} \right) \right) \right) \right] d\mathbf{q} \\
&= \int_{\mathbb{R}^d} \frac{u}{\rho_\infty^\varepsilon} \left[\nabla \cdot \left(\frac{u}{\rho_\infty^\varepsilon} \boldsymbol{\sigma}\boldsymbol{\sigma}^T \left(-\mathbf{b}\rho_\infty^\varepsilon + \nabla \rho_\infty^\varepsilon \right) \right) \right] d\mathbf{q} \\
&\quad + \int_{\mathbb{R}^d} \frac{u}{\rho_\infty^\varepsilon} \left[\nabla \cdot \left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T \rho_\infty^\varepsilon \nabla \left(\frac{u}{\rho_\infty^\varepsilon} \right) \right) \right] d\mathbf{q} \\
&= -\int_{\mathbb{R}^d} \rho_\infty^\varepsilon \left| \boldsymbol{\sigma}^T \nabla \left(\frac{u}{\rho_\infty^\varepsilon} \right) \right|^2 d\mathbf{q} - \frac{1}{2} \int_{\mathbb{R}^d} \left| \frac{u}{\rho_\infty^\varepsilon} \right|^2 \mathcal{L}_\varepsilon^* \rho_\infty^\varepsilon \, d\mathbf{q} \\
&= -\int_{\mathbb{R}^d} \rho_\infty^\varepsilon \left| \boldsymbol{\sigma}^T \nabla \left(\frac{u}{\rho_\infty^\varepsilon} \right) \right|^2 d\mathbf{q} \leq 0.
\end{aligned}
\tag{A.37}
$$

Thus, $-\mathcal{L}_\varepsilon^*$ is accretive.

Next, we prove that $R(I - \mathcal{L}_\varepsilon^*) = L^2(1/\rho_\infty^\varepsilon)$. This can be derived by Lax-Milgram's theorem. For arbitrary $f \in L^2(1/\rho_\infty^\varepsilon)$, define the linear functional

$$f(u) \colon L^2(1/\rho_\infty^\varepsilon) \to \mathbb{R}, \quad u \mapsto f(u) := \langle u, f \rangle_{L^2(1/\rho_\infty^\varepsilon)},$$

which is continuous on H. Now, consider the bilinear form:

$$a(u, v) \colon H \times H \to \mathbb{R}, \quad (u, v) \mapsto a(u, v) := \langle (I - \mathcal{L}_\varepsilon^*)u, v \rangle_{L^2(1/\rho_\infty^\varepsilon)}.$$

By direct calculation and Cauchy-Schwarz inequality, we obtain

$$
\begin{aligned}
|a(u, v)| &= \left| \int_{\mathbb{R}^d} uv \frac{1}{\rho_\infty^\varepsilon} d\mathbf{q} - \int_{\mathbb{R}^d} \frac{1}{\rho_\infty^\varepsilon} u \mathcal{L}_\varepsilon^* v \, d\mathbf{q} \right| \\
&\leq \|u\|_{L^2(1/\rho_\infty^\varepsilon)} \|v\|_{L^2(1/\rho_\infty^\varepsilon)} \\
&\quad + \sqrt{\int_{\mathbb{R}^d} \rho_\infty^\varepsilon \left| \boldsymbol{\sigma}^T \nabla \left(\frac{u}{\rho_\infty^\varepsilon} \right) \right|^2 d\mathbf{q}} \cdot \sqrt{\int_{\mathbb{R}^d} \rho_\infty^\varepsilon \left| \boldsymbol{\sigma}^T \nabla \left(\frac{v}{\rho_\infty^\varepsilon} \right) \right|^2 d\mathbf{q}}.
\end{aligned}
$$

Therefore, we have

$$|a(u, v)| \leq c \|u\|_H \|v\|_H,$$

where $c > 0$ is a constant depending on $\boldsymbol{\sigma}$. Meanwhile,

$$a(u, u) = \|u\|^2_{L^2(1/\rho^\varepsilon_\infty)} + \int_{\mathbb{R}^d} \rho^\varepsilon_\infty \left| \boldsymbol{\sigma}^T \nabla \left(\frac{u}{\rho^\varepsilon_\infty} \right) \right|^2 d\mathbf{q} \geq c' \|u\|^2_H,$$

where $c' > 0$ is a constant depending on $\boldsymbol{\sigma}$. Therefore, a is bounded and coercive on H. By Lax-Milgram's theorem, there exists $u \in H$ such that $a(u, v) = f(v)$ holds for all $v \in H$. Hence, $(I - \mathcal{L}^*_\varepsilon)u = f$, and thus $R(I - \mathcal{L}^*_\varepsilon) = L^2(1/\rho^\varepsilon_\infty)$.

Finally, by the Hille-Yosida theorem, we conclude that $\mathcal{L}^*_\varepsilon$ generates a strongly continuous semigroup of contractions in $L^2(1/\rho^\varepsilon_\infty)$.

For part (ii), by [3], equation (2.5) admits a unique solution $\rho^\varepsilon(\mathbf{q}, t) \in C^1([0, T], L^2(1/\rho^\varepsilon_\infty))$ for any $T > 0$. $\qquad\square$

Proof of Theorem 2.3 Notice that the structure of the generator in (2.24) is quite similar to the generator in (2.15), with the exception that only $\mathbf{p}$ is an odd variable.

Step 1. $(i) \Rightarrow (ii)$ is exactly the same as Step 1 in the proof of Theorem 2.2.

Step 2. We prove $(ii) \Rightarrow (iii)$. From the definition of the generator $\mathcal{L}$ in (2.24), (2.26) implies

$$\iiint \Big[-\varphi_2(\mathbf{p})\psi_2(\mathbf{p})\varphi_3(\mathbf{z})\psi_3(\mathbf{z}) \, \mathbf{p} \cdot \nabla_\mathbf{q} (\psi_1(\mathbf{q})\varphi_1(\mathbf{q}))$$

$$- \varphi_1(\mathbf{q})\psi_1(\mathbf{q})\varphi_3(\mathbf{z})\psi_3(\mathbf{z})(\mathbf{b} + \mathbf{z}) \cdot \nabla_\mathbf{p} (\varphi_2(\mathbf{p})\psi_2(\mathbf{p}))$$

$$- \varphi_1(\mathbf{q})\psi_1(\mathbf{q})\varphi_2(\mathbf{p})\psi_2(\mathbf{p}) \, \mathbf{p} \cdot \nabla_\mathbf{z} (\varphi_3(\mathbf{z})\psi_3(\mathbf{z})) \Big] e^{-\beta H(\mathbf{q},\mathbf{p},\mathbf{z})} \, d\mathbf{q} \, d\mathbf{p} \, d\mathbf{z}$$

$$= \iiint \Big[\psi_1(\mathbf{q})\varphi_1(\mathbf{q})\varphi_2(\mathbf{p})\psi_2(\mathbf{p})\alpha\Big(\psi_3(\mathbf{z}) \, \mathbf{z} \cdot \nabla_\mathbf{z} \varphi_3(\mathbf{z})$$

$$- \varphi_3(\mathbf{z}) \, \mathbf{z} \cdot \nabla_\mathbf{z} \psi_3(\mathbf{z}) + \frac{1}{\beta} (\varphi_3(\mathbf{z}) \Delta_\mathbf{z} \psi_3(\mathbf{z}) - \psi_3(\mathbf{z}) \Delta_\mathbf{z} \varphi_3(\mathbf{z})) \Big) \Big] e^{-\beta H(\mathbf{q},\mathbf{p},\mathbf{z})} \, d\mathbf{q} \, d\mathbf{p} \, d\mathbf{z}.$$

$$\tag{A.38}$$

Using integration by parts, this simplifies as

$$\iiint \varphi_1 \psi_1 \varphi_2 \psi_2 \varphi_3 \psi_3 \left(\mathbf{p} \cdot \nabla_\mathbf{q} e^{-\beta H} + (\mathbf{b} + \mathbf{z}) \cdot \nabla_\mathbf{p} e^{-\beta H} - \mathbf{p} \cdot \nabla_\mathbf{z} e^{-\beta H} \right) d\mathbf{q} \, d\mathbf{p} \, d\mathbf{z}$$

$$= \iiint \varphi_1 \psi_1 \varphi_2 \psi_2 \left(\frac{1}{\beta} \nabla_\mathbf{z} e^{-\beta H} + \mathbf{z} e^{-\beta H} \right) (\psi_3 \nabla_\mathbf{z} \varphi_3 - \varphi_3 \nabla_\mathbf{z} \psi_3) \, d\mathbf{q} \, d\mathbf{p} \, d\mathbf{z}. \tag{A.39}$$

Taking $\varphi_3(\mathbf{z}) = \psi_3(\mathbf{z})$, the RHS of (A.39) becomes zero, yielding

$$\mathbf{p} \cdot \nabla_\mathbf{q} e^{-\beta H} + (\mathbf{b} + \mathbf{z}) \cdot \nabla_\mathbf{p} e^{-\beta H} - \mathbf{p} \cdot \nabla_\mathbf{z} e^{-\beta H} = 0. \tag{A.40}$$

Thus, by the auxiliary Lemma A.1, we obtain

$$\frac{1}{\beta} \nabla_\mathbf{z} e^{-\beta H} + \mathbf{z} e^{-\beta H} = 0. \tag{A.41}$$

This implies $e^{-\beta H + \frac{\beta}{2} |\mathbf{z}|^2}$ is independent of $\mathbf{z}$. Rewriting (A.40) as

$$\mathbf{p} \cdot \nabla_\mathbf{q} e^{-\beta H + \frac{\beta}{2} |\mathbf{z}|^2} + \mathbf{b} \cdot \nabla_\mathbf{p} e^{-\beta H + \frac{\beta}{2} |\mathbf{z}|^2} + \mathbf{z} \cdot \left(\nabla_\mathbf{p} e^{-\beta H + \frac{\beta}{2} |\mathbf{z}|^2} + \beta \mathbf{p} e^{-\beta H + \frac{\beta}{2} |\mathbf{z}|^2} \right) = 0, \tag{A.42}$$

we conclude

$$\nabla_{\mathbf{p}} e^{-\beta H + \frac{\beta}{2}|\mathbf{z}|^2} + \beta \mathbf{p} e^{-\beta H + \frac{\beta}{2}|\mathbf{z}|^2} = 0. \tag{A.43}$$

Therefore, $e^{-\beta H + \frac{\beta}{2}|\mathbf{p}|^2 + \frac{\beta}{2}|\mathbf{z}|^2}$ depends only on $\mathbf{q}$, and we conclude there exists a potential $U(\mathbf{q})$ such that $H = U(\mathbf{q}) + \frac{1}{2}|\mathbf{p}|^2 + \frac{1}{2}|\mathbf{z}|^2$. From:

$$\mathbf{p} \cdot \nabla_{\mathbf{q}} e^{-\beta H + \frac{\beta}{2}|\mathbf{z}|^2} + \mathbf{b} \cdot \nabla_{\mathbf{p}} e^{-\beta H + \frac{\beta}{2}|\mathbf{z}|^2} = 0,$$

we further conclude $b(\mathbf{q}) = -\nabla_{\mathbf{q}} U(\mathbf{q})$.

Step 3. $(iii) \Rightarrow (i)$ is exactly the same as Step 3 in the proof of Theorem 2.2.

Step 4. The statement (iii) implies both statements (iv) and (v), which are obvious.

To show that (iv) implies (iii), similar to (A.27), we use the evenness in $\mathbf{p}$ to obtain that $\rho_0(\mathbf{q}, \mathbf{p}, \mathbf{z})$ solves

$$\nabla_{\mathbf{z}} \cdot \left(\mathbf{z} \rho_0 + \frac{1}{\beta} \nabla_{\mathbf{z}} \rho_0 \right) = 0.$$

By Proposition 3.2, we conclude there exists $c(\mathbf{q}, \mathbf{p}) > 0$ such that

$$\rho_0 = c(\mathbf{q}, \mathbf{p}) e^{-\frac{\beta}{2}|\mathbf{z}|^2}.$$

Then, by similar arguments to (A.42) and (A.43), we can conclude (iii).

To show that (v) implies (iii), we plug $\rho_0 = U_1(\mathbf{q}) U_2(\mathbf{p}) U_3(\mathbf{z})$ into $\mathcal{L}^* \rho_0 = 0$ and obtain:

$$-\mathbf{p} \cdot \nabla_{\mathbf{q}} \log U_1(\mathbf{q}) - \mathbf{b} \cdot \nabla_{\mathbf{p}} \log U_2(\mathbf{p}) + \mathbf{z} \cdot \nabla_{\mathbf{p}} \log U_2(\mathbf{p}) + \mathbf{p} \cdot \nabla_{\mathbf{z}} \log U_3(\mathbf{z})$$

$$+ \frac{\alpha}{U_3(\mathbf{z})} \nabla_{\mathbf{z}} \cdot \left(U_3(\mathbf{z}) \mathbf{z} + \frac{1}{\beta} \nabla_{\mathbf{z}} U_2(\mathbf{z}) \right) = 0. \tag{A.44}$$

By similar arguments as in (A.31), we have:

$$\log U_2(\mathbf{p}) = \mathbf{p}^T A \mathbf{p} + \mathbf{a} \cdot \mathbf{p} + c, \quad \log U_3(\mathbf{z}) = \mathbf{z}^T B \mathbf{z} + \mathbf{d} \cdot \mathbf{z} + e,$$

for some constant matrices $A, B \in \mathbb{R}^{d \times d}$, and vectors $\mathbf{a}, \mathbf{d} \in \mathbb{R}^d$, and constants $c, e \in \mathbb{R}$. By similar arguments as in (A.33), we conclude that $A = -\frac{\beta}{2} I$, $B = -\frac{\beta}{2} I$, $\mathbf{a} = 0$, and $\mathbf{d} = 0$. Finally, we obtain

$$\nabla_{\mathbf{q}} \log U_1(\mathbf{q}) = \beta \mathbf{b} = -\nabla_{\mathbf{q}} U(\mathbf{q}),$$

for $U(\mathbf{q}) = -\log U_1(\mathbf{q})$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Appendix B: Omitted Proofs for Exponential Convergence and Smoothness of Fokker-Planck Equations

In this section, we present the omitted proofs for the exponential convergence and smoothness of the solution to the Fokker-Planck equation corresponding to overdamped Langevin dynamics, as well as the hypocoercivity and hypoellipticity results for underdamped Langevin dynamics.

B.1 Exponential Convergence and Smoothness for the Overdamped Case

We first provide the proof for the exponential convergence and smoothness in the over-damped case. The proof of Proposition 3.1 invokes the following version of Harris's theorem, as stated in [10, Theorem 3.6].

Theorem B.1 (Harris's theorem, Theorem 3.6 in [10]) *Suppose that $\mathcal{P}$ is the Markov operator of a discrete Markov semigroup $\mathcal{P}^n$ for $n = 1, 2, \ldots$ on $\mathbb{R}^d$. Let $p(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, denote the transition probability of $\mathcal{P}$. If the following two conditions hold*

(i) *(Lyapunov Function) There exists a function $U : \mathbb{R}^d \to [0, \infty)$ and constants $K > 0$ and $\gamma \in (0, 1)$ such that*

$$\mathcal{P}U(\mathbf{x}) \leq \gamma U(\mathbf{x}) + K \tag{B.1}$$

for all $\mathbf{x} \in \mathbb{R}^d$;

(ii) *(Minorisation) For every $R > 0$, there exists a constant $\alpha > 0$ such that*

$$\|p(\mathbf{x}, \cdot) - p(\mathbf{y}, \cdot)\|_{L^1(\mathbb{R}^d)} \leq 2(1 - \alpha), \tag{B.2}$$

for all $\mathbf{x}, \mathbf{y}$ such that $U(\mathbf{x}) + U(\mathbf{y}) \leq R$.

Define the following weighted supremum norm:

$$\|\varphi\|_U = \sup_{\mathbf{x} \in \mathbb{R}^d} \frac{|\varphi(\mathbf{x})|}{1 + U(\mathbf{x})}. \tag{B.3}$$

Then $\mathcal{P}$ admits a unique invariant measure μ^*. Furthermore, there exist constants $C > 0$ and $\rho \in (0, 1)$ such that

$$\left\| \mathcal{P}^n \varphi - \int_{\mathbb{R}^d} \varphi(\mathbf{x}) \, \mathrm{d}\mu^* \right\|_U \leq C\rho^n \left\| \varphi - \int_{\mathbb{R}^d} \varphi(\mathbf{x}) \, \mathrm{d}\mu^* \right\|_U. \tag{B.4}$$

In the proof of Proposition 3.1, we take $\mathcal{P} = e^{t_0 \mathcal{L}_\varepsilon}$ for some appropriately chosen t_0, and the Lyapunov function $U(\mathbf{x}) = e^{V(\mathbf{x})/3}$, where $V(\mathbf{x})$ is the potential function.

Proof of Proposition 3.1 We directly employ Theorem B in [15]. Consider $U(\mathbf{q}) = e^{V(\mathbf{q})/3}$, which is in $L^1(\rho_0)$. Moreover, we have

$$\mathcal{L}_\varepsilon U(\mathbf{q}) = \frac{1}{3} \left(\mathrm{tr}(\boldsymbol{\sigma} \boldsymbol{\sigma}^T \nabla^2 V(\mathbf{q})) - \frac{2}{3} |\boldsymbol{\sigma}^T \nabla V(\mathbf{q})|^2 + \varepsilon \boldsymbol{\sigma} \boldsymbol{\sigma}^T \mathbf{M}(\mathbf{q}) \cdot \nabla V(\mathbf{q}) \right) e^{V(\mathbf{q})/3}.$$

For $\mathbf{q}$ that is not in the support of $\mathbf{M}$ and sufficiently large, by Assumptions (I) and (II), we have

$$\mathcal{L}_\varepsilon U(\mathbf{q}) \leq -\frac{1}{18} |\boldsymbol{\sigma}^T \nabla V(\mathbf{q})|^2 e^{V(\mathbf{q})/3} \leq -C_1 e^{V(\mathbf{q})/3}.$$

Thus, for all $\mathbf{q} \in \mathbb{R}^d$, we have

$$\mathcal{L}_\varepsilon U(\mathbf{q}) \leq C_2 - C_1 U(\mathbf{q}) \tag{B.5}$$

for some constants $C_1 > 0$ and $C_2 > 0$. Therefore, $U(\mathbf{q})$ is a strongly unbounded Lyapunov function. By Theorem B in [15], we conclude exponential convergence.

Next, we prove that C_ε and r_ε can be selected uniformly. Indeed, the proof of Theorem B in [15] relies on Harris's theorem [10], which is based on the following two conditions:

1. Lyapunov condition: For some $t_0 > 0$, there exist constants $\gamma \in (0, 1)$, $K > 0$, and a function $U : \mathbb{R}^d \to \mathbb{R}^+$ such that

$$e^{t_0 \mathcal{L}_\varepsilon} U(\mathbf{q}) \leq \gamma U(\mathbf{q}) + K \tag{B.6}$$

holds for all $\mathbf{q} \in \mathbb{R}^d$. According to [15, Lemma 3.4], (B.6) holds by choosing $U(\mathbf{q}) = e^{V(\mathbf{q})/3}$, i.e., the unbounded strong Lyapunov function. Moreover, γ and K depend on $\mathbf{M}$, V, $\boldsymbol{\sigma}$, and constants C_1 and C_2 in (B.5), which are all uniform in ε.

2. Minorisation condition: It is required to verify the following local minorisation condition: for all $R > 0$, there exists $\alpha > 0$ such that

$$\|p^\varepsilon(\mathbf{q}_1, \cdot, t_0) - p^\varepsilon(\mathbf{q}_2, \cdot, t_0)\|_{L^1(\mathbb{R}^d)} \leq 2(1 - \alpha) \tag{B.7}$$

holds for all $\mathbf{q}_1$ and $\mathbf{q}_2$ in the set $\{(\mathbf{q}_1, \mathbf{q}_2) : U(\mathbf{q}_1) + U(\mathbf{q}_2) \leq R\}$. Here, $p^\varepsilon(\mathbf{q}_1, \mathbf{q}, t_0)$ is the transition probability. By [15, Lemma 3.3], we know that (B.7) holds for $\varepsilon = 0$ with some $\alpha = \alpha_0$. Then, by [2, Corollary 9.8.26], there exists a constant $C > 0$ (depending on $\mathbf{M}$, V, and $\boldsymbol{\sigma}$) such that for all $\mathbf{q}_1$ with $U(\mathbf{q}_1) \leq R$, we have

$$\|p^\varepsilon(\mathbf{q}_1, \cdot, t_0) - p^0(\mathbf{q}_1, \cdot, t_0)\|_{L^1(\mathbb{R}^d)} \leq C\varepsilon.$$

Thus, there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$,

$$\|p^\varepsilon(\mathbf{q}_1, \cdot, t_0) - p^\varepsilon(\mathbf{q}_2, \cdot, t_0)\|_{L^1(\mathbb{R}^d)} \leq 2(1 - \alpha_0/2). \tag{B.8}$$

Therefore, for each $R > 0$, one can find $\alpha > 0$ such that (B.7) holds uniformly for sufficiently small ε.

Finally, according to [10, Remark 3.10], the contraction constants C_ε and r_ε depend on $\mathbf{M}$, V, $\boldsymbol{\sigma}$, and the following constants: by selecting R sufficiently large such that

$$\gamma_0 := \gamma + \frac{2K}{R} < 1, \quad \beta := \frac{\alpha}{2K}, \quad \alpha' := \max\left\{1 - \frac{\alpha}{2}, \frac{2 + R\beta\gamma_0}{2 + R\beta}\right\}.$$

Here, γ and K are from (B.6), and α is from (B.7), all of which are uniform for $\varepsilon \in (0, \varepsilon_0)$. Thus, C_ε and r_ε are also uniform in $\varepsilon \in (0, \varepsilon_0)$. □

Proof of Lemma 3.4 Consider the coordinate transformation $\mathbf{q}' = \boldsymbol{\sigma}^{-1}\mathbf{q}$. Then we have $\nabla_{\mathbf{q}'} f = \boldsymbol{\sigma}^T \nabla_{\mathbf{q}} f$. Thus, one can reformulate $\mathcal{L}_\varepsilon^* - \frac{\partial}{\partial t}$ as

$$\mathcal{L}_\varepsilon^* - \frac{\partial}{\partial t} = \sum_{j=1}^d \left(\frac{\partial}{\partial \mathbf{q}_j'}\right)^2 + \left(\nabla_{\mathbf{q}_j'} V(\boldsymbol{\sigma}\mathbf{q}') - \varepsilon\boldsymbol{\sigma}^T \mathbf{M}(\boldsymbol{\sigma}\mathbf{q}')\right) \cdot \nabla_{\mathbf{q}_j'} - \frac{\partial}{\partial t}$$
$$+ \left(\Delta_{\mathbf{q}'} V(\boldsymbol{\sigma}\mathbf{q}') - \varepsilon\nabla_{\mathbf{q}_j'} \cdot \boldsymbol{\sigma}^T \mathbf{M}(\boldsymbol{\sigma}\mathbf{q}')\right).$$

Let

$$\mathbf{X}_j = \mathbf{e}_j \in \mathbb{R}^{d+1}, \quad j = 1, 2, \ldots, d, \quad \mathbf{X}_0 = -\mathbf{e}_{d+1} + \left[\nabla_{\mathbf{q}_j'} V(\boldsymbol{\sigma}\mathbf{q}') - \varepsilon\boldsymbol{\sigma}^T \mathbf{M}(\boldsymbol{\sigma}\mathbf{q}'), 0\right],$$

$$c = \Delta^2_{\mathbf{q}'} V(\boldsymbol{\sigma}\mathbf{q}') - \varepsilon \nabla_{\mathbf{q}'_j} \cdot \boldsymbol{\sigma}^T \mathbf{M}(\boldsymbol{\sigma}\mathbf{q}').$$

Then, we can express $\mathcal{L}^*_\varepsilon - \frac{\partial}{\partial t}$ as

$$\mathcal{L}^*_\varepsilon - \frac{\partial}{\partial t} = \sum_{i=1}^d (\mathbf{X}_j)^2 + \mathbf{X}_0 + c.$$

Note that $\mathbf{X}_i$ for $i = 1, 2, \ldots, d+1$ are linearly independent at any point $(\mathbf{q}', t) \in \mathbb{R}^d \times (0, +\infty)$. By Theorem 3.1, we know that $\mathcal{L}^*_\varepsilon - \frac{\partial}{\partial t}$ is hypoelliptic, and $\rho^\varepsilon(\mathbf{q}, t)$ is smooth in both $\mathbf{q}$ and t. □

Now we provide the well-posedness of the Fokker-Planck equation for the overdamped case. The existence and uniqueness are obtained via the semigroup method, while the smoothness is established using Hörmander's hypoellipticity theorem 3.1.

Proof of Lemma 4.1 We define a weighted Hilbert space $L^2(\rho^\varepsilon_\infty)$ with the weight $\rho^\varepsilon_\infty(\mathbf{q}, \mathbf{p}) > 0$ and the norm

$$\|f\|_{L^2(\rho^\varepsilon_\infty)} := \iint \rho^\varepsilon_\infty f^2 \, d\mathbf{q} \, d\mathbf{p}.$$

Additionally, we define a weighted Hilbert space $H^1(\rho^\varepsilon_\infty)$ with the norm

$$\|f\|_{H^1(\rho^\varepsilon_\infty)} := \iint \rho^\varepsilon_\infty f^2 \, d\mathbf{q} \, d\mathbf{p} + \iint \rho^\varepsilon_\infty |\nabla f|^2 \, d\mathbf{q} \, d\mathbf{p}.$$

For (i), we have

$$\frac{\rho_0(\mathbf{q}, \mathbf{p})}{\rho^\varepsilon_\infty(\mathbf{q}, \mathbf{p})} = \frac{e^{-\varepsilon W(\mathbf{q})} Z_\varepsilon}{Z_0}.$$

Since $W \in C^\infty_c(\mathbb{R}^d; \mathbb{R})$, for $\mathbf{q}$ outside the support of W, $\frac{\rho_0(\mathbf{q},\mathbf{p})}{\rho^\varepsilon_\infty(\mathbf{q},\mathbf{p})}$ is a constant. Therefore, it is in $L^2(\rho^\varepsilon_\infty)$, and any order derivative of $\frac{\rho_0}{\rho^\varepsilon_\infty}$ is also in $L^2(\rho^\varepsilon_\infty)$. Thus, (i) holds.

For (ii), by Lemma A.2, we know that if $\mathbf{b} = -\nabla_{\mathbf{q}} U(\mathbf{q})$, then $-\mathcal{L}_{\varepsilon,1}$ is a maximal monotone operator in $L^2(\rho^\varepsilon_\infty \, d\mathbf{q} \, d\mathbf{p})$, with the weight defined in (2.19). By the Hille-Yosida theorem, $\mathcal{L}_{\varepsilon,1}$ generates[1] a strongly continuous semigroup on $L^2(\rho^\varepsilon_\infty)$, denoted as $S(t) = e^{t\mathcal{L}_{\varepsilon,1}}$.

For (iii), from the contraction semigroup in (ii), we know that the Cauchy problem

$$\frac{\partial p^\varepsilon}{\partial t} = \mathcal{L}_{\varepsilon,1} p, \quad p(\mathbf{q}, \mathbf{p}, 0) = \frac{\rho_0}{\rho^\varepsilon_\infty},$$

is well-posed, and $p^\varepsilon(\mathbf{q}, \mathbf{p}, t) \in C^1([0, T]; L^2(\rho^\varepsilon_\infty))$. Thus, (4.2) admits a unique solution $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t) \in C^1([0, T]; L^2(\rho^\varepsilon_\infty))$. □

Proof of Lemma 4.2 We apply Theorem 3.1 again. For simplicity, we assume $\boldsymbol{\sigma} = \mathbf{I}$. If this is not the case, a coordinate transformation similar to that in Lemma 3.4 can be used.

---

[1]Sometimes $-\mathcal{L}_{\varepsilon,1}$ is called the infinitesimal generator instead of $\mathcal{L}$.

The operator $\mathcal{L}_\varepsilon^* - \dfrac{\partial}{\partial t}$ can be expressed as

$$\mathcal{L}_\varepsilon^* - \frac{\partial}{\partial t} = \sum_{i=1}^d (\mathbf{X}_i)^2 + \mathbf{X}_0 + c,$$

where $c = d$ and

$$\mathbf{X}_i = \frac{\partial}{\partial p_i}, \quad i = 1, 2, \ldots, d,$$

$$\mathbf{X}_0 = -\mathbf{p} \cdot \nabla_{\mathbf{q}} + \left(\nabla_{\mathbf{q}} (V - \varepsilon W) + \mathbf{p} \right) \cdot \nabla_{\mathbf{p}} - \frac{\partial}{\partial t}.$$

Notice that

$$[\mathbf{X}_i, \mathbf{X}_0] = -\frac{\partial}{\partial q_i} + \frac{\partial}{\partial p_i}, \quad i = 1, 2, \ldots, d.$$

Thus, at any point $(\mathbf{q}, \mathbf{p}, t) \in \mathbb{R}^{2d} \times (0, \infty)$, we have

$$\mathrm{span}\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d, [\mathbf{X}_1, \mathbf{X}_0], \ldots, [\mathbf{X}_d, \mathbf{X}_0], \mathbf{X}_0\} = \mathbb{R}^{2d+1}.$$

By Theorem 3.1, we conclude that $\rho^\varepsilon(\mathbf{q}, \mathbf{p}, t)$ is smooth. $\qquad \square$

Declarations

Competing Interests The authors declare no conflict of interest.

References

1. Barbu, V.: Nonlinear Differential Equations of Monotone Types in Banach Spaces. Springer Monographs in Mathematics. Springer, Berlin (2010)
2. Bogachev, V.I., Krylov, N.V., Röckner, M., Shaposhnikov, S.V.: Fokker-Planck-Kolmogorov Equations. Mathematical Surveys and Monographs, vol. 207. Am. Math. Soc., Providence (2015)
3. Bühler, T., Salamon, D.A.: Functional Analysis, vol. 191. Am. Math. Soc., Providence (2018)
4. Campisi, M., Hänggi, P., Talkner, P.: Colloquium: quantum fluctuation relations: foundations and applications. Rev. Mod. Phys. **83**(3), 771 (2011)
5. Dabelow, L., Bo, S., Eichhorn, R.: Irreversibility in active matter systems: fluctuation theorem and mutual information. Phys. Rev. X **9**(2), 021009 (2019)
6. Dolbeault, J., Mouhot, C., Schmeiser, C.: Hypocoercivity for linear kinetic equations conserving mass. Trans. Am. Math. Soc. **367**(6), 3807–3828 (2015)
7. Einstein, A.: Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Ann. Phys. **322**(8), 549–560 (1905)
8. Gallavotti, G., Cohen, E.G.D.: Dynamical ensembles in stationary states. J. Stat. Phys. **80**, 931–970 (1995)
9. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order, vol. 224. Springer, Berlin (2015)
10. Hairer, M.: Convergence of Markov Processes. Lecture Notes (2010)
11. Hairer, M., Majda, A.J.: A simple framework to justify linear response theory. Nonlinearity **23**(4), 909 (2010)

12. Hairer, M., Pavliotis, G.A.: From ballistic to diffusive behavior in periodic potentials. J. Stat. Phys. **131**, 175–202 (2008)
13. Hiroshi, K.: On backward stochastic differential equations. Stochastics **6**(3–4), 293–313 (1982)
14. Hörmander, L.: Hypoelliptic second order differential equations (1968)
15. Ji, M., Shen, Z., Yi, Y.: Convergence to equilibrium in Fokker–Planck equations. J. Dyn. Differ. Equ. **31**, 1591–1615 (2019)
16. Kato, T.: Demicontinuity, hemicontinuity and monotonicity. Bull. Am. Math. Soc. **70**(4), 548–551 (1964)
17. Kou, S.C., Xie, X.S.: Generalized Langevin equation with fractional Gaussian noise: subdiffusion within a single protein molecule. Phys. Rev. Lett. **93**(18), 180603 (2004)
18. Kubo, R.: The fluctuation-dissipation theorem. Rep. Prog. Phys. **29**(1), 255 (1966)
19. Lei, L., Liu, J.-G., Lu, J.: Fractional stochastic differential equations satisfying fluctuation-dissipation theorem. J. Stat. Phys. **169**(2), 316–339 (2017)
20. Marconi, U.M.B., Puglisi, A., Rondoni, L., Vulpiani, A.: Fluctuation–dissipation: response theory in statistical physics. Phys. Rep. **461**(4–6), 111–195 (2008)
21. Mori, H.: A continued-fraction representation of the time-correlation functions. Prog. Theor. Phys. **34**(3), 399–416 (1965)
22. Onsager, L.: Reciprocal relations in irreversible processes. I. Phys. Rev. **37**(4), 405 (1931)
23. Onsager, L.: Reciprocal relations in irreversible processes. II. Phys. Rev. **38**(12), 2265 (1931)
24. Pavliotis, G.A.: Asymptotic analysis of the Green–Kubo formula. IMA J. Appl. Math. **75**(6), 951–967 (2010)
25. Pavliotis, G.A.: Stochastic processes and applications p. 155
26. Raz, K.: Fractional kinetics in Kac–Zwanzig heat bath models. J. Stat. Phys. **114**, 291–326 (2004)
27. Ruelle, D.: A review of linear response theory for general differentiable dynamical systems. Nonlinearity **22**(4), 855 (2009)
28. Seifert, U.: Stochastic thermodynamics, fluctuation theorems and molecular machines. Rep. Prog. Phys. **75**(12), 126001 (2012)
29. Villani, C.: Hypocoercivity. Am. Math. Soc., Providence (2009)