CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Warwick Research Archives Portal Repository

Manuscript version: Author's Accepted Manuscript
The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/135196

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

warwick.ac.uk/lib-publications

COARSE-GRAINING OF OVERDAMPED LANGEVIN DYNAMICS VIA THE MORI-ZWANZIG FORMALISM

THOMAS HUDSON AND XINGJIE HELEN LI

Abstract. The Mori–Zwanzig formalism is applied to derive an equation for the evolution of linear observables of the overdamped Langevin equation. To illustrate the resulting equation and its use in deriving approximate models, a particular benchmark example is studied both numerically and via a formal asymptotic expansion. The example considered demonstrates the importance of memory effects in determining the correct temporal behaviour of such systems.

1. Introduction

Molecular dynamics (MD) is a widely–used simulation technique which captures the atomistic details of material systems, allowing the prediction of their properties and behavior [19, 51]. However, despite the vast increases in computational capacity over recent decades, it is still not always possible to work with MD models at full resolution, particularly when studying large, complex systems over long time–scales. Fortunately, in many cases, the objectives of a simulation occur within a small region of interest. This observation has led to the development of *coarse–grained MD* (CGMD) models, in which excess degrees of freedom are incorporated implicitly [16–18, 20, 27, 29, 49, 55, 62].

Building reliable and efficient CGMD models attuned to the quantities of interest is a difficult problem. First, the simulator must find appropriate variables which capture the quantities of interest [24], often termed *reaction coordinates* or *resolved variables*. Once these are fixed, an appropriate proxy model for the reaction coordinates must be obtained, which implicitly incorporates the interaction between the reaction coordinates and unresolved degrees of freedom [16, 17, 34, 39, 51, 65]. If the objectives of a simulation are 'static' macroscopic equilibrium properties such as free energy or reaction rates, then a wide variety of choices of proxy dynamics which appropriately sample the relevant measures are available. However, if the objective is to capture a dynamical property of the physical system such as kinematic viscosity or a diffusion rate, then it is important to capture the correct effective dynamics of the reaction coordinates arising due to the relevant dynamics of the full system over moderate timescales [9, 19, 25, 47, 51, 64].

In recent years, a variety of studies of CGMD schemes have been undertaken, aiming to analyse the predictions of such schemes. In all cases, the ultimate goal is to obtain verifiable, statistically accurate predictions of the true dynamics for various applications. The wide variety of mathematical techniques used includes

- optimal prediction techniques [8, 10, 11, 21];
- information-theoretic tools [5, 15, 28, 48, 49];

Mathematics Institute, University of Warwick, Coventry, CV4 7AL
Fretwell 350C, 9201 Univ City Blvd., Charlotte, NC, 28023

2010 *Mathematics Subject Classification.* 41A60, 82C31, 60H10.

Key words and phrases. Mori–Zwanzig formalism, overdamped Langevin dynamics, memory effects.

The work of T. Hudson is supported by the Leverhulme Trust through Early Career Fellowship ECF-2016-526.

The work of X. Li is supported in part by the Simons Foundation Collaboration Grant with Award ID: 426935 and NSF DMS CAREER-1847770.

- statistical filtering and ensemble methods [1, 6, 45];
- identification of an appropriate parametrization [12, 23, 38, 50];
- series expansion [37, 60, 69];
- pathwise estimates [30, 35, 36, 40]; and
- conditional expectations [33].

Here, our focus is on the Mori–Zwanzig (MZ) approach to CGMD benchmark problem [3, 4, 32, 54, 57, 71, 72]. The MZ formalism provides an exact expression of the dynamics for a CGMD scheme, and is governed by three terms which separate out different contributions to the true dynamics, each of which has a different statistical physical meaning. This decomposition allows a study of the sources of error: the first term accounts for a conservative dynamics due to the effective interactions between the coarse grained variables; the second is a history-dependent term determined by a time integral of a memory kernel which represents the interactions between the resolved and unresolved variables; and the third term represents the random thermal fluctuations arising from unresolved variables. In different situations, each of these terms may have a more or less important role, but to correctly capture the dynamical properties and validate an effective model, it is critical to measure the relative size and behaviour of these terms accurately.

Our study concentrates particularly on the *memory term*, which may heuristically be thought of as measuring the extent to which the set of reaction coordinates is decoupled from the unresolved degrees of freedom. In recent years, there have been tremendous efforts to investigate memory terms from MZ projections for a variety of classes of dynamics, see for example [3, 7, 10, 13, 24, 41–43, 63, 66–68]. One common approach is to hope that a time-scale separation between the resolved and unresolved variables occurs, i.e. the fluctuations of the unresolved variables occur on a much faster timescale than those of the resolved variables, and therefore the two sets of variables are weakly correlated. In such cases, the memory kernel decay rapidly, approximating a delta function in time [9].

Our aim in this paper is to demonstrate that while such delta approximations of the memory kernel are appropriate in many situations, it is not generally to be expected that the memory kernel is independent of the value of the reaction coordinate, even in the simple situation where the chosen reaction coordinates are linear. To capture the correct dynamics, further careful analysis and sampling of the memory is therefore required.

As an illustration of this issue, we consider the dynamical behaviour of a gradient flow with stochastic forcing (often called the overdamped Langevin equation), demonstrating that at least in this case, a naïve approach to approximating the memory kernel yields a poor approximation of the dynamics. We hope that the benchmark problem we consider here will provide insight which will enable the study of CGMD derived from full Langevin dynamics based on reliable asymptotic analysis in future.

1.1. **Outline.** This paper is organized as follows. In Section 2, we review the Mori–Zwanzig formalism applied to general gradient flow systems, and in Theorem 2.1, derive an exact equation for the evolution of linear observables within an abstract framework. Our benchmark example is discussed in Section 3, and an asymptotic analysis is performed to obtain approximations of the various terms in the MZ equation. Finally, we study this particular example numerically in Section 4. Throughout the rest of this paper, we choose to refer exclusively to resolved degrees of freedom as *reaction coordinates*, and unresolved degrees of freedom as *orthogonal variables*.

2. Formulation of the problem

As our reference fine–scale dynamical system, we consider the following overdamped Langevin dynamics defined on $\mathbb{R}^N$:

$$d\mathbf{X}_t = -\nabla_{\mathbf{x}} V(\mathbf{X}_t)dt + \sqrt{2\beta^{-1}}d\mathbf{B}_t. \tag{2.1}$$

Here, $\mathbf{B}_t$ denotes a standard vector–valued Brownian motion, $V(\cdot)$ is a potential energy and β is the inverse temperature. Throughout this work, we assume that V is at least of class C^2, and satisfies the following conditions:

(1) $V(\mathbf{x}) \to +\infty$ as $|\mathbf{x}| \to +\infty$ and $e^{-\beta V(\mathbf{x})} \in L^1(\mathbb{R}^N)$.

(2) The gradient $\nabla V(\mathbf{x})$ is globally Lipschitz, i.e. there exists $\alpha > 0$ such that

$$\big|\nabla V(\mathbf{x}) - \nabla V(\mathbf{y})\big| \leq \alpha\,|\mathbf{x} - \mathbf{y}|.$$

Under the regularity assumption and condition (1), it is well–known (see for example Proposition 4.2 in [59], Theorem 2.1 in [61], or the general results of [52, 53]) that the dynamics defined by (2.1) are ergodic with respect to the Gibbs measure, $\boldsymbol{\mu}_G$, given by $d\boldsymbol{\mu}_G(\mathbf{x}) = \frac{1}{Z}e^{-\beta V(\mathbf{x})}d\mathbf{x}$ where Z is the partition function

$$Z := \int e^{-\beta V(\mathbf{x})}d\mathbf{x}.$$

Given a regular function $\mathbf{F} : \mathbb{R}^N \to \mathbb{R}^m$, which may be thought of as describing a family of *reaction coordinates* which are our variables of interest, we may apply Itô's formula to deduce that the value of $\mathbf{F}(\mathbf{X}_t)$ is governed by the Itô SDE

$$d\mathbf{F}(\mathbf{X}_t) = \Big(-\nabla V(\mathbf{X}_t)\cdot\nabla\mathbf{F}(\mathbf{X}_t) + \beta^{-1}\Delta\mathbf{F}(\mathbf{X}_t)\Big)dt + \sqrt{2\beta^{-1}}\nabla\mathbf{F}(\mathbf{X}_t)\cdot d\mathbf{B}_t. \tag{2.2}$$

To be compatible with the notion that we are interested only in the dynamics of $\mathbf{F}(\mathbf{X}_t)$, we will assume throughout that the initial conditions $\mathbf{X}_0$ for (2.1) are distributed according to the marginal of the Gibbs measure conditioned on the value of the reaction coordinates at time zero, $\mathbf{F}(\mathbf{X}_0)$.

In particular, if we consider a linear coarse–graining selector $\mathbf{F}(\mathbf{x}) := \Phi\mathbf{x}$, where $\Phi \in \mathbb{R}^{m\times N}$ is a constant matrix, (2.2) becomes

$$d\mathbf{F}(\mathbf{X}_t) = -\Phi\nabla V(\mathbf{X}_t)dt + \sqrt{2\beta^{-1}}\Phi\, d\mathbf{B}_t. \tag{2.3}$$

Such linear coordinates are commonly used in CGMD schemes, particularly for large molecules such as polymers [14, 18, 58]. If we are interested in the value of the reaction coordinates described by $\mathbf{F}$ alone, then (2.3) provides an equation for their evolution. In general however, since the first term on the right–hand side of the equation depends on the full process $\mathbf{X}_t$, this is not a closed equation for the value of $\mathbf{F}(\mathbf{X}_t)$.

In order to formulate a closed approximate equation for $\mathbf{F}(\mathbf{X}_t)$, we use the *Mori–Zwanzig formalism*, which uses projection operators to decompose the equations describing observables of a dynamical system into terms involving the value of the observables alone, and 'error' terms describing the contribution of variations of $\mathbf{X}_t$ which do not directly change the value of the observable.

In this case, the natural projection operator we choose to apply is the *Zwanzig projection*, which involves taking a conditional expectation with respect to the Gibbs distribution, i.e.

$$\mathcal{P}\mathbf{g} = \mathbb{E}_{\boldsymbol{\mu}_G}\big[\mathbf{g}\,\big|\,\mathbf{F}(\mathbf{x}) = \mathbf{h}\big] := \frac{\int_{\mathbf{F}^{-1}(\mathbf{h})}\mathbf{g}(\mathbf{x})\,e^{-\beta V(\mathbf{x})}d\mathbf{x}}{\int_{\mathbf{F}^{-1}(\mathbf{h})}e^{-\beta V(\mathbf{x})}d\mathbf{x}}; \tag{2.4}$$

note that in the above formula, we have cancelled the common factor $\frac{1}{Z}$ from the numerator and denominator.[1] It may be verified that $\mathcal{P}$ is an orthogonal projection on the space of square–integrable observables, i.e. $L^2(\mathbb{R}^N; e^{-\beta V(\mathbf{x})}d\mathbf{x})$, and we can therefore define its orthogonal counterpart,

$$\mathcal{Q} := \mathcal{I} - \mathcal{P}.$$

In particular, we note that the evolution of (2.3) can be divided into the stationary, mean–zero process induced by the Brownian motion, and the evolution of the mean value of $\mathbf{F}(\mathbf{X}_t)$. To consider the behaviour of the latter quantity given knowledge of $\mathbf{X}_0$, we define

$$\mathbf{h}_t(\mathbf{x}) = \mathbb{E}\big[\mathbf{F}(\mathbf{X}_t) \,\big|\, \mathbf{F}(\mathbf{X}_0) = \mathbf{x}\big]. \tag{2.5}$$

The evolution of this quantity is governed by the usual generator of the SDE (2.1),

$$\mathcal{L} := -\nabla V \cdot \nabla + \beta^{-1}\Delta. \tag{2.6}$$

Using this definition, the Feynmann–Kac formula governing the evolution of $\mathbf{h}$ states that the function $\mathbf{h}$ solves the PDE

$$\partial_t \mathbf{h}_t = \mathcal{L}\mathbf{h}_t \qquad \text{with} \qquad \mathbf{h}_0 = \mathbf{F}. \tag{2.7}$$

Using the definition of $\mathcal{P}$, we apply the Mori–Zwanzig formalism to provide a different expression of (2.7), stated in the following theorem.

Theorem 2.1. *Let $\mathbf{X}_t$ satisfy the SDE on $\mathbb{R}^N$*

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

and given a constant matrix of full rank $\Phi \in \mathbb{R}^{N\times m}$, the observable

$$\mathbf{h}_t(\mathbf{x}) = \mathbb{E}\big[\mathbf{F}(\mathbf{X}_t) \,\big|\, \mathbf{F}(\mathbf{X}_0) = \mathbf{x}\big]; \tag{2.8}$$

satisfies the following integro–differential equation:

$$\partial_t \mathbf{h}_t(\mathbf{x}) = -\Phi\Phi^T \nabla\mathcal{S}\big(\mathbf{h}_t(\mathbf{x})\big) + \int_0^t \mathcal{M}_s\big(\mathbf{h}_{t-s}(\mathbf{x})\big)\cdot\nabla\mathcal{S}\big(\mathbf{h}_{t-s}(\mathbf{x})\big) - \frac{1}{\beta}\mathsf{div}\,\mathcal{M}_s\big(\mathbf{h}_{t-s}(\mathbf{x})\big)ds + \mathcal{F}_t(\mathbf{x}), \tag{2.9}$$

where:

(1) $\mathcal{S} : \mathbb{R}^m \to \mathbb{R}$ *is the* effective potential*, defined to be*

$$\mathcal{S}(\mathbf{h}) := -\frac{1}{\beta}\log Z_\Phi(\mathbf{h}) \text{ with } Z_\Phi(\mathbf{h}) := \int_{\mathbf{F}^{-1}(\mathbf{h})} e^{-\beta V(\mathbf{x})}d\mathbf{x}, \tag{2.10}$$

(2) $\mathcal{M}_s : \mathbb{R}^m \to \mathbb{R}^{m\times m}$ *is the* memory kernel*, defined to be*

$$\mathcal{M}_s(\mathbf{h}) := \beta\,\mathbb{E}\big[e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}\otimes\mathcal{Q}\mathcal{L}\mathbf{F} \,\big|\, \mathbf{F}(\mathbf{x}) = \mathbf{h}\big] = \beta\,\mathbb{E}\big[\mathcal{F}_s\otimes\mathcal{F}_0 \,\big|\, \mathbf{F}(\mathbf{x}) = \mathbf{h}\big], \tag{2.11}$$

(3) *and $\mathcal{F}_t : \mathbb{R}^N \to \mathbb{R}$ is the* fluctuating force*, defined to be*

$$\mathcal{F}_t := e^{t\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}. \tag{2.12}$$

A proof of this result is given in Appendix A, and involves adapting standard variants of the Mori–Zwanzig formalism already present in the literature to this stochastic setting.

[1] Note that to be completely technically correct, the right hand side should be understood in the sense of Radon–Nikodym differentiation of measures.

Remark 2.1. The Mori–Zwanzig formalism [54, 56, 70, 71] uses projection operators to rewrite the equations governing observables of a dynamical system. Various formulations have been developed in recent years with a variety of applications in mind, and influence our own derivation, including: [41] treating crystalline solids via the harmonic approximation; [69] for the harmonic oscillators based on operator series expansions of the orthogonal dynamics propagator; [46] for the full Langevin dynamics model based on reduced-order modeling; [26] for a model based on dissipative particle dynamics; and [58] for a 'hybrid' coarse–graining map of a Hamiltonian model.

Recombining the evolution of the mean $\mathbf{h}_t$ given by (2.9) and adding back the Brownian motion, we find that (2.3) can be written

$$\begin{aligned} d\mathbf{F}(\mathbf{X}_t) = -\Phi\Phi^T \nabla\mathcal{S}\big(\mathbf{F}(\mathbf{X}_t)\big)dt + \int_0^t \mathcal{M}_s\big(\mathbf{F}(\mathbf{X}_{t-s})\big) \cdot \nabla\mathcal{S}\big(\mathbf{F}(\mathbf{X}_{t-s})\big) ds\, dt \\ - \int_0^t \frac{1}{\beta} \mathsf{div}\, \mathcal{M}_s\big(\mathbf{F}(\mathbf{X}_{t-s})\big) ds\, dt + d\mathcal{F}_t + \sqrt{2\beta^{-1}}\Phi d\mathbf{B}_t. \end{aligned} \tag{2.13}$$

It is important to note at this point that (2.13) is equivalent to considering the full evolution (2.3), in particular because $\mathcal{F}_t$ (which appears in (2.13) through the definition of $\mathcal{M}_s$) is unknown. Equation (2.13) therefore remains unclosed; however, the power of this formulation is that if $\mathcal{F}_t$ has statistics which are well–captured by some proxy process $\widetilde{\mathcal{F}}_t$, and $\mathcal{S}$ is either known or accurately approximated, then we can obtain a closed–form approximate dynamics

$$\begin{aligned} d\mathbf{h}_t = -\Phi\Phi^T\nabla\mathcal{S}(\mathbf{h}_t)dt + \left(\int_0^t \widetilde{\mathcal{M}}_s(\mathbf{h}_{t-s})\nabla\mathcal{S}(\mathbf{h}_{t-s}) - \frac{1}{\beta}\mathsf{div}\,\widetilde{\mathcal{M}}_s(\mathbf{h}_{t-s})ds \right) dt \\ + d\widetilde{\mathcal{F}}_t + \sqrt{2\beta^{-1}}\Phi d\mathbf{B}_t, \end{aligned}$$

where $\widetilde{\mathcal{M}}_s(\mathbf{h})$ is the autocovariance function of $\widetilde{\mathcal{F}}_t(\mathbf{h})$ (see for example §1.3.1 of [44]).

To explore this formulation and better understand the relationship between the terms involved, the remainder of the paper is devoted to an exploration of a particular illustrative example where an accurate approximation of the terms within (2.9) can be performed.

3. A benchmark problem

In this section, we will consider the overdamped Langevin equation (2.1) in the particular case where $\mathbf{x} \in \mathbb{R}^2$, and the potential energy is defined to be

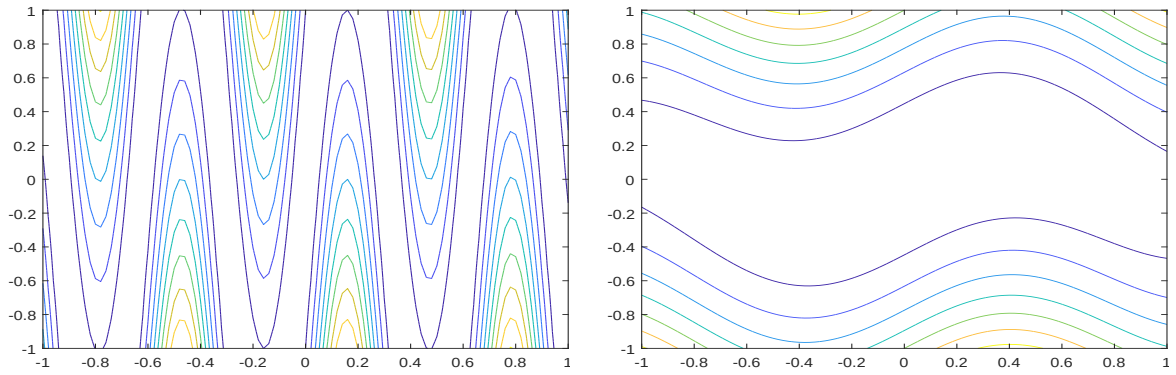
$$V(x, y) := \frac{\mu}{2}x^2 + \frac{\lambda}{2}\big(\tau\sin(\omega x) - y\big)^2 \tag{3.1}$$

with μ, λ, τ, $\omega \geq 0$ being parameters. Specifying even further, we will focus on the case where $\lambda \gg \mu$, so that there is a separation between the timescale of relaxation for the x and y variables. Here and throughout the paper, the symbols $\ll$, $\gg$ and $\sim$ are all intended in the formal asymptotic sense, as described in Section 3.4 of [2]. Typically, if we require that $f \ll g$, then $f \leq \frac{1}{10}g$ is usually sufficient to provide a good approximation.

As such, x is a 'slow' variable, and is a natural candidate for a reaction coordinate of the system; as in Section 2, we therefore consider

$$\mathbf{F}(\mathbf{x}) = \Phi\mathbf{x} \qquad \text{where} \qquad \Phi := \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

The second term in (3.1), i.e. $\frac{\lambda}{2}\big(\tau\sin(\omega x) - y\big)^2$, has been chosen to emulate a form of free energy barrier to the dynamics, since when $\tau \sim 1$, $\omega \gg 1$ and $\beta \gg 1$, we expect trajectories of the dynamics to remain close to the manifold $y = \sin(\omega x)$; see Figure 1 for representations of different



(a) $\mu = 2, \lambda = 20, \omega = 10, \tau = 2$ (b) $\mu = 2, \lambda = 20, \omega = 4, \tau = 0.2$

FIGURE 1. Contour plots of the potential energy $V(x, y)$ defined in (3.1): $\lambda/\mu \gg 1$ and $\omega\tau$ is chosen to be $\gg 1$ for figure (a) and < 1 for figure (b), respectively.

energy landscapes. As a consequence, we expect that as τ and ω increase with $\frac{\lambda}{\mu} \gg 1$, the dynamics should to take progressively longer to approach a fixed neighbourhood of the global equilibrium at 0 from a generic initial condition, since the dynamics must effectively 'travel further' along the meandering valley in the potential to get there.

3.1. **Derivation of approximate dynamics.** Under the assumptions described above, we compute $\mathcal{P}\mathcal{L}\mathbf{F}$ and $\mathcal{Q}\mathcal{L}\mathbf{F}$, and use these to derive formal approximations of the terms involved in (2.9).

(1) *Computation of* $\mathcal{P}\mathcal{L}\mathbf{F}$. The effective potential $\mathcal{S}(x)$ defined in (2.10) is equal to

$$\mathcal{S}(x) = -\frac{1}{\beta}\log\left(\int_{\mathbb{R}} e^{-\beta V(x,y)}dy\right) = \frac{\mu}{2}x^2 + \mathsf{const}, \tag{3.2}$$

as the orthgonal variable y follows normal distribution $y \sim N(\tau\sin(\omega x), \frac{1}{\beta\lambda})$, and hence

$$\mathcal{P}\mathcal{L}\mathbf{F} = \begin{pmatrix} -\mu x \\ 0 \end{pmatrix}. \tag{3.3}$$

(2) *Computation of* $\mathcal{Q}\mathcal{L}\mathbf{F}$. Clearly $\mathcal{Q}\mathcal{L}\mathbf{F} = \mathcal{L}\mathbf{F} - \mathcal{P}\mathcal{L}\mathbf{F}$, so using (3.3), we have

$$\mathcal{Q}\mathcal{L}\mathbf{F} = \mathcal{Q}\mathcal{L}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -\lambda\tau\omega\left(\tau\sin(\omega x) - y\right)\cos(\omega x) \\ -\lambda\left(y - \tau\sin(\omega x)\right) \end{pmatrix}.$$

(3) *Formal approximation of* $\mathcal{M}_s$. Recalling the definition of the memory function $\mathcal{M}_s(\mathbf{F})$ from (2.11), we must compute or otherwise approximate the expression $\mathcal{Q}\mathcal{L}e^{s\mathcal{Q}\mathcal{L}}\mathbf{F} \otimes \mathcal{Q}\mathcal{L}\mathbf{F}$. To do so, we define characteristic curves of the orthogonal dynamics, $\widetilde{\mathbf{x}}_s = (\widetilde{x}_s, \widetilde{y}_s) = e^{s\mathcal{Q}\mathcal{L}}\widetilde{\mathbf{x}}_0$, and find that

$$\begin{aligned} \mathcal{Q}\mathcal{L}e^{s\mathcal{Q}\mathcal{L}}\mathbf{F}(\widetilde{\mathbf{x}}_0) \otimes \mathcal{Q}\mathcal{L}\mathbf{F}(\widetilde{\mathbf{x}}_0) &= \Phi\dot{\widetilde{\mathbf{x}}}_s \otimes \Phi\dot{\widetilde{\mathbf{x}}}_0 \\ &= \lambda^2\tau^2\omega^2\left(\tau\sin(\omega\widetilde{x}_s) - \widetilde{y}_s\right)\cos(\omega\widetilde{x}_s)\cdot\left(\tau\sin(\omega\widetilde{x}_0) - \widetilde{y}_0\right)\cos(\omega\widetilde{x}_0). \end{aligned} \tag{3.4}$$

We now change variable with the intention of linearizing, setting $\widetilde{u}_s := \widetilde{x}_s - \widetilde{x}_0$ and $\widetilde{v}_s := \widetilde{y}_s - \tau\sin(\omega\widetilde{x}_0)$. Expressed in these new variables, the action of the orthogonal dynamics is equivalent to solving the ODE system

$$\begin{pmatrix} \dot{\widetilde{u}}_s \\ \dot{\widetilde{v}}_s \end{pmatrix} = \begin{pmatrix} -\lambda\tau\omega\left(\tau\sin(\omega(\widetilde{x}_0 + \widetilde{u}_s)) - \tau\sin(\omega\widetilde{x}_0) - \widetilde{v}_s\right)\cos\left(\omega(\widetilde{x}_0 + \widetilde{u}_s)\right) \\ -\lambda\left(\tau\sin(\omega\widetilde{x}_0) + \widetilde{v}_s - \tau\sin\left(\omega(\widetilde{x}_0 + \widetilde{u}_s)\right)\right) \end{pmatrix}.$$

Since we have assumed that $\lambda \gg \mu$, given knowledge of $\widetilde{x}_0$ alone, we expect that initial conditions for the orthogonal dynamics to be concentrated near $(\widetilde{x}_0, \widetilde{y}_0) = (\widetilde{x}_0, \tau\sin(\omega\widetilde{x}_0))$, and so linearising on this basis, we obtain

$$\begin{pmatrix} \dot{\widetilde{u}}_s \\ \dot{\widetilde{v}}_s \end{pmatrix} = \lambda \begin{pmatrix} -\tau^2\omega^2\cos^2(\omega\widetilde{x}_0) & \tau\omega\cos(\omega\widetilde{x}_0) \\ \tau\omega\cos(\omega\widetilde{x}_0) & -1 \end{pmatrix} \begin{pmatrix} \widetilde{u}_s \\ \widetilde{v}_s \end{pmatrix} + \mathcal{O}(\widetilde{u}_s^2, \widetilde{v}_s^2, \widetilde{u}_s\widetilde{v}_s). \tag{3.5}$$

Noting in particular that $\widetilde{u}_0 = 0$ since $\widetilde{u}_s = \widetilde{x}_s - \widetilde{x}_0$, and neglecting higher–order terms in (3.5), the solution is approximately

$$\begin{aligned} \begin{pmatrix} \widetilde{u}_s \\ \widetilde{v}_s \end{pmatrix} \approx \frac{1}{1+\tau^2\omega^2\cos^2(\omega\widetilde{x}_0)} \begin{pmatrix} \widetilde{v}_0\tau\omega\cos(\omega\widetilde{x}_0) \\ \widetilde{v}_0\tau^2\omega^2\cos^2(\omega\widetilde{x}_0) \end{pmatrix} \\ + \frac{e^{-\lambda\left(1+\tau^2\omega^2\cos^2(\omega\widetilde{x}_0)\right)s}}{1+\tau^2\omega^2\cos^2(\omega\widetilde{x}_0)} \begin{pmatrix} -\widetilde{v}_0\tau\omega\cos(\omega\widetilde{x}_0) \\ \widetilde{v}_0 \end{pmatrix}, \end{aligned}$$

and therefore

$$\dot{\widetilde{x}}_s = \dot{\widetilde{u}}_s \approx \lambda\tau\omega\widetilde{v}_0\cos(\omega\widetilde{x}_0)e^{-\lambda\left(1+\tau^2\omega^2\cos^2(\omega\widetilde{x}_0)\right)s}. \tag{3.6}$$

When conditioning on knowledge of $\widetilde{x}_0$, it follows that $\widetilde{y}_0 \sim \mathcal{N}(\tau\sin(\omega\widetilde{x}_0), \frac{1}{\lambda\beta})$, and therefore $\widetilde{v}_0 \sim \mathcal{N}(0, \frac{1}{\lambda\beta})$: the memory kernel is therefore approximately

$$\begin{aligned} \mathcal{M}_s(\mathbf{h}) &= \beta\,\mathbb{E}[\mathcal{Q}\mathcal{L}e^{s\mathcal{Q}\mathcal{L}}\mathbf{F}\otimes\mathcal{Q}\mathcal{L}\mathbf{F}\,|\,\mathbf{F}(\mathbf{x})=\mathbf{h}] \\ &\approx \beta\int_{\mathbb{R}} \dot{\widetilde{u}}_s(\mathbf{h},\widetilde{y}_0)\otimes\dot{\widetilde{u}}_0(\mathbf{h},\widetilde{y}_0)e^{-\beta\lambda(\widetilde{y}_0-\tau\sin(\omega\mathbf{h}))^2}d\widetilde{y}_0 \\ &= \lambda\tau^2\omega^2\cos^2(\omega\mathbf{h})e^{-\lambda\left(1+\tau^2\omega^2\cos^2(\omega\mathbf{h})\right)s}, \end{aligned} \tag{3.7}$$

where in the above formula, $\dot{\widetilde{u}}_0 = \lim_{s\to 0+}\dot{\widetilde{u}}_s$. Recalling the form of (2.9), we note that we must also approximate the divergence of $\mathcal{M}_s$; using the expression derived in (3.7), in this case we obtain

$$\mathsf{div}\mathcal{M}_s(\mathbf{h}) \approx -\lambda\tau^2\omega^3\sin(2\omega\mathbf{h})\left(1-\lambda\tau^2\omega^2\cos^2(\omega\mathbf{h})s\right)e^{-\lambda\left(1+\tau^2\omega^2\cos^2(\omega\mathbf{h})\right)s}. \tag{3.8}$$

(4) *Formal approximation of memory integral.* Our next step is to approximate the first integral term involving the memory kernel in (2.9). Noting the form of (3.7), we see that this is an integral of exponential type, and therefore we apply the method of steepest descent (also known as Laplace's method) to derive a formal approximation of the memory integral (see Chapter 6 of [2]).

Viewing λ as a large parameter, using (3.7) and (3.2) we note that

$$\int_0^t \mathcal{M}_s(\mathbf{h}_{t-s})\cdot\nabla\mathcal{S}(\mathbf{h}_{t-s})ds \approx \int_0^t \lambda\tau^2\omega^2\cos^2(\omega\mathbf{h}_{t-s})\,\mu\mathbf{h}_{t-s}e^{-\lambda\left(1+\tau^2\omega^2\cos^2(\omega\mathbf{h}_{t-s})\right)s}ds.$$

Setting

$$p(s) := \tau^2\omega^2\cos^2(\omega\mathbf{h}_{t-s})\,\mu\mathbf{h}_{t-s} \quad\text{and}\quad q(s) := -\left(1+\tau^2\omega^2\cos^2(\omega\mathbf{h}_{t-s})\right)s,$$

this integral takes the general form

$$I(\lambda) = \int_0^t \lambda p(s)e^{\lambda q(s)}ds.$$

Noting that $q(s)$ is maximal on the domain of integration at $s=0$, since

$$-\left(1+\tau^2\omega^2\cos^2(\omega\mathbf{h}_{t-s})\right)s < 0 \qquad \text{for all } s>0,$$

the main contribution to the integral therefore comes from the interval $s \in [0, \varepsilon)$, and arguing as in Section 6.4 of [2] and noting $q(0) = 0$, $q'(0) = -(1 + \tau^2\omega^2 \cos^2(\omega \mathbf{h}_t))$ and $|p(0)| \sim \mu$, we have

$$\int_0^t \lambda p(s) e^{\lambda q(s)} ds \approx \int_0^\infty \lambda p(0) e^{\lambda q'(0) s} + \mathcal{O}\left(\frac{\mu}{\lambda}\right) = \frac{p(0)}{q'(0)} + \mathcal{O}\left(\frac{\mu}{\lambda}\right).$$

This approximation therefore yields

$$\begin{aligned}
&\int_0^t \mathcal{M}_s(\mathbf{h}_{t-s}) \cdot \nabla \mathcal{S}(\mathbf{h}_{t-s}) ds \\
&\quad \approx \int_0^\infty \lambda \tau^2 \omega^2 \cos^2(\omega \mathbf{h}_t)\, \mu \mathbf{h}_t e^{-\lambda\left(1 + \tau^2\omega^2 \cos^2(\omega \mathbf{h}_t)\right) s} ds + \mathcal{O}\left(\frac{\mu}{\lambda}\right) \\
&\quad \approx \frac{\tau^2 \omega^2 \cos^2(\omega \mathbf{h}_t)}{1 + \tau^2 \omega^2 \cos^2(\omega \mathbf{h}_t)} \mu \mathbf{h}_t + \mathcal{O}\left(\frac{\mu}{\lambda}\right),
\end{aligned}$$

and via a similar procedure applied to the divergence term, we have

$$\begin{aligned}
&\int_0^t -\frac{1}{\beta} \mathsf{div}\left(\mathcal{M}_s(\mathbf{h}_{t-s})\right) ds \\
&\qquad \approx \frac{1}{\beta} \frac{\tau^2 \omega^3 \sin(2\omega \mathbf{h}_t)}{\left(1 + \tau^2 \omega^2 \cos^2(\omega \mathbf{h}_t)\right)^2} + \mathcal{O}\left(\frac{1}{\beta\lambda}\right).
\end{aligned}$$

Combining these expressions, we obtain a final approximation of the memory contributions in (2.9) as

$$\begin{aligned}
&\int_0^t \mathcal{M}_s(\mathbf{h}_{t-s}) \cdot \nabla \mathcal{S}(\mathbf{h}_{t-s}) - \frac{1}{\beta} \mathsf{div}\, \mathcal{M}_s(\mathbf{h}_{t-s}) ds \\
&\qquad \approx \frac{\tau^2 \omega^2 \cos^2(\omega \mathbf{h}_t)}{1 + \tau^2 \omega^2 \cos^2(\omega \mathbf{h}_t)} \mu \mathbf{h}_t + \frac{1}{\beta} \frac{\tau^2 \omega^3 \sin(2\omega \mathbf{h}_t)}{\left(1 + \tau^2 \omega^2 \cos^2(\omega \mathbf{h}_t)\right)^2} + \mathcal{O}\left(\frac{\mu}{\lambda}, \frac{1}{\beta\lambda}\right).
\end{aligned} \tag{3.9}$$

Notably, if $\tau \sim 1$, the second term is negligible compared with the first both when $\omega \ll 1$, and when $\omega \gg 1$. We will therefore discard the second term in these cases.

(5) *Formal approximation of fluctuating force.* Above, we have shown that the memory kernel can be approximated as

$$\mathcal{M}_s(\mathbf{h}) \approx \frac{\tau^2 \omega^2 \cos^2(\omega \mathbf{h})}{1 + \tau^2 \omega^2 \cos^2(\omega \mathbf{h})} \delta_0(s).$$

Indeed, defining

$$\gamma(\mathbf{h}) = 1 + \tau^2 \omega^2 \cos^2(\omega \mathbf{h}),$$

with the delta approximation of the memory derived above, and neglecting the term arising from the divergence of $\mathcal{M}_s$, we may combine the first three terms in (2.13) to obtain

$$-\nabla \mathcal{S}(\mathbf{h}_t) + \int_0^t \mathcal{M}_s(\mathbf{h}_{t-s}) \cdot \nabla \mathcal{S}(\mathbf{h}_{t-s}) - \frac{1}{\beta} \mathsf{div}\, \mathcal{M}_s\big(\mathbf{h}_{t-s}(\mathbf{x})\big) ds \approx -\gamma^{-1}(\mathbf{h}_t) \nabla \mathcal{S}(\mathbf{h}_t). \tag{3.10}$$

We next note that since the dynamics of $\mathbf{X}_t$ are ergodic with respect to the Gibbs measure $\boldsymbol{\mu}_G$ under assumption (1) given in Section 2, it follows that $\mathbf{F}(\mathbf{X}_t)$ is ergodic with respect to the pushforward $\mathbf{F}_* \boldsymbol{\mu}_G$, which in this case has a density proportional to $e^{-\beta \mathcal{S}(\mathbf{h})}$. If we wish to maintain this property in our choice of approximate dynamics, it is natural to approximate the combination of the fluctuating force and the Brownian drift by

$$d\mathcal{F}_t + \sqrt{2\beta^{-1}} \Phi d\mathbf{B}_t \approx \sqrt{2\beta^{-1} \gamma^{-1}(\mathbf{h}_t)} d\mathbf{W}_t, \tag{3.11}$$

where $\mathbf{W}_t$ is a 1–dimensional Brownian motion. This choice ensures that the infinitesimal generator of the process is

$$\Omega f = -\gamma^{-1}\nabla\mathcal{S}\cdot\nabla f + \beta^{-1}\gamma^{-1}\Delta f,$$

with formal adjoint

$$\Omega^* g = \gamma^{-1}\,\mathrm{div}\Big(g\nabla\mathcal{S} + \beta^{-1}\nabla g\Big),$$

and hence the unique invariant measure under this dynamics remains proportional to $e^{-\beta\mathcal{S}(\mathbf{h})}$ as in the case of the true dynamics.

Alternatively, a physical justification for this choice arises from the Fluctuation–Dissipation Theorem, which in the case of overdamped Langevin dynamics (see for example Section 3 of [31]) requires that if

$$d\mathbf{Q}_t = -\gamma^{-1}\nabla U(\mathbf{Q}_t)\,dt + \sigma\,d\mathbf{W}_t,$$

where $-\nabla U$ are forces derived from a potential energy U, then at thermal equilibrium it must hold that $\sigma^2 = 2\beta^{-1}\gamma^{-1}$. The choice made in (3.11) indeed therefore satisfies this relation.

Combining the approximations above, in the case where $\omega \gg 1$, we obtain the closed–form approximate equation

$$d\mathbf{h}_t = -\frac{\mu\mathbf{h}_t}{1+\tau^2\omega^2\cos^2(\omega\mathbf{h}_t)}dt + \sqrt{\frac{2\beta^{-1}}{1+\tau^2\omega^2\cos^2(\omega\mathbf{h}_t)}}d\mathbf{W}_t \tag{3.12}$$

Notably, the drift term in this equation is independent of β.

We remark that the derivation given above relies on formal asymptotic methods and arguments, and we will assess the approximation numerically in Section 4, leaving a rigorous treatment of our approach in a more general setting for future study.

3.2. **Other choices of approximate dynamics.** The derivation of the effective dynamics (3.9) was informed by the Mori–Zwanzig formalism, but other choices could be made, and may be more appropriate in other circumstances.

(1) *Discarding memory and fluctuating force.* In [33], the authors consider another choice of effective dynamics, which in our setting, amounts to considering

$$d\boldsymbol{\xi}_t = -\Phi\Phi^T\nabla\mathcal{S}(\boldsymbol{\xi}_t)dt + \sqrt{2\beta^{-1}\Phi\Phi^T}d\mathbf{B}_t. \tag{3.13}$$

This is equivalent to (2.13) where the memory and fluctuating force terms have been neglected entirely. For the evolution of the mean $\mathbf{h}_t$, this choice of dynamics yields

$$\partial_t\mathbf{h}_t = -\Phi\Phi^T\nabla\mathcal{S}(\mathbf{h}_t)dt \tag{3.14}$$

as the effective equation for the observable we consider. The authors have proved error bounds on the time marginals of the resulting probability distribution when compared the true dynamics captured by (2.3); applying [33, Proposition 3.1] to our case gives the bound

$$H(\psi_t|\phi_t) \le \frac{\beta\tau^2\omega^2}{4}\left[H\left(\psi_0|\mu\right) - H\left(\psi_t|\mu\right)\right], \tag{3.15}$$

where:

(a) $H(\mu|\nu)$ is the relative entropy of a measure μ with respect to ν, i.e.

$$H(\mu|\nu) := \int \log\left(\frac{d\mu}{d\nu}\right)d\mu;$$

(b) μ is the Gibbs measure;

(c) ψ_t is the distribution of the 'true' dynamics at time t; and
(d) ϕ_t is the distribution of solutions to (3.13) at time t.
Clearly, the constant in (3.15) is large when $\tau\omega \gg 1$; this reflects the fact that neglecting the memory in this case is not sufficient to accurately capture the dynamical properties of the system, and a more sophisticated approach is needed.

(2) *A more naïve memory approximation.* To highlight the need to conduct dynamical sampling to approximate $\mathcal{M}_s$ correctly, we remark that the approximation of the memory terms obtained in (3.9) is notably *not* the same as simply choosing to approximate

$$\begin{aligned}\mathcal{M}_s(\mathbf{h}) \approx \widetilde{\mathcal{M}}_s(\mathbf{h}) &:= \beta\,\mathbb{E}[\mathcal{Q}\mathcal{L}\mathbf{x}\otimes\mathcal{Q}\mathcal{L}\mathbf{x}\,|\,\mathbf{F}(\mathbf{x})=\mathbf{h}]\,\delta_0(s),\\ &= \left(\sqrt{\frac{\lambda\beta}{2\pi}}\int \beta\,\lambda^2\tau^2\omega^2(\tau\sin(\omega\mathbf{h})-y)^2\cos^2(\omega\mathbf{h})e^{-\frac{1}{2}\beta\lambda(\tau\sin(\omega\mathbf{h})-y)^2}dy\right)\delta_0(s)\\ &= \lambda\tau^2\omega^2\cos^2(\omega\mathbf{h})\delta_0(s).\end{aligned}$$

Using $\widetilde{\mathcal{M}}_s$ would result in the approximate dynamics

$$\partial_t\mathbf{h}_t = \big(\lambda\tau^2\omega^2\cos^2(\omega\mathbf{h}_t)-1\big)\mu\mathbf{h}_t + \frac{\lambda}{\beta}\tau^2\omega^3\sin(2\omega\mathbf{h}_t). \tag{3.16}$$

Since we have chosen $\lambda \gg 1$, we see that this approximation will yield qualitatively different dynamics to both the true dynamics for $\mathbf{F}$, (2.3), and the approximate dynamics given by (3.9); we investigate this numerically in Section 4.

The remarks above suggests that in general, careful dynamical sampling of the memory kernel is required to accurately capture the interaction between chosen reaction coordinates and the neglected degrees of freedom.

4. Numerical simulations

In this section, we conduct a numerical study of the various choices of approximate effective dynamics for the observable in the benchmark example considered in Section 3. We first study the temporal and spatial behaviour of the memory kernel and the approximation derived in (3.7), and propose a possible measure for the quality of the chosen reaction coordinate based on the covariance of the orthogonal dynamics. We then compare different approximation strategies for the full dynamics, both with and without thermostat.

4.1. **Investigation of the memory kernel.**

(1) *Time decay and spatial oscillation of the memory kernel.* The derivation of the effective dynamics for $\mathbf{h}_t$ made in Section 3.1 relies crucially upon a series of formal approximations to both the orthogonal dynamics and the memory integral in (2.9); we therefore first numerically test the validity of these assumptions by computing the memory kernel empirically.

To do so, trajectories of the orthogonal dynamics were statistically sampled. The results of these simulations are shown in Figure 2, and are compared to the explicit approximate form derived in (3.7). The empirical memory exhibits rapid exponential decay in time for all values of x_0 considered, in agreement with (3.7).

(2) *Assessing the choice of reaction coordinate.* For practical applications, it is difficult to compute and then integrate over long trajectories of the memory term, so a well–chosen coarse–graining selector should lead to both a small correlation between the reaction coordinate and the orthogonal dynamics and rapid decay of the memory kernel [24].

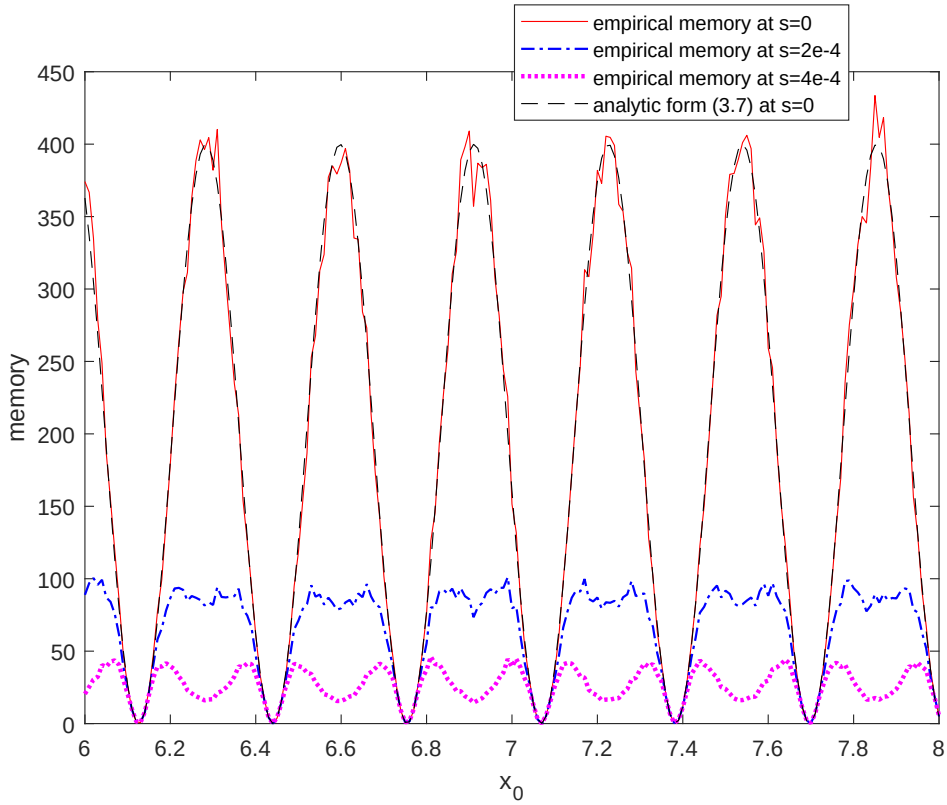
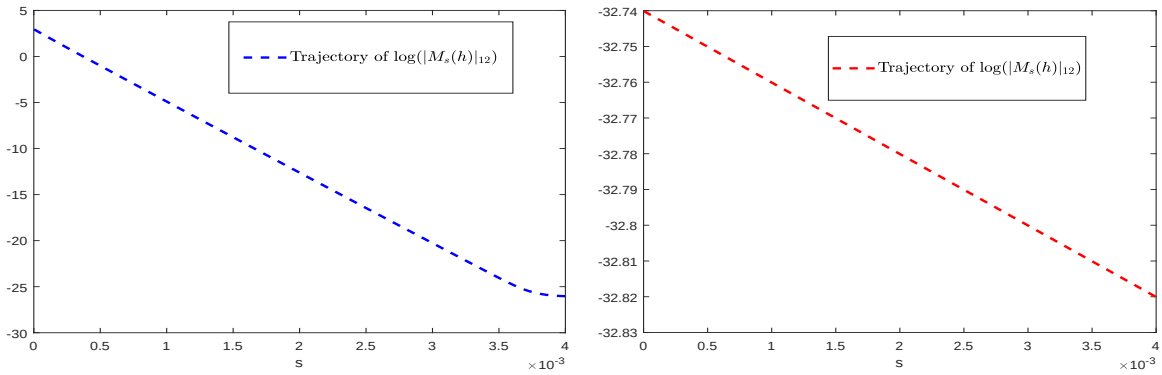


FIGURE 2. Empirical computation of the memory kernel $\mathcal{M}_s(\mathbf{h})$ for various values of s and $\mathbf{h} = \mathbf{F}(\mathbf{X}_0)$, compared with the approximate $\mathbf{h}$–dependent form of the memory (3.7) evaluated at $s = 0$. The empirical memory clearly shows rapid decay with increasing s, and oscillates strongly in space. Parameters for V in this case are $\omega = 10$, $\tau = 2$, $\lambda = 20$ and $\mu = 2$.



(a) $|\cos(\omega\mathbf{h})| = 1$ (b) $\cos(\omega\mathbf{h}) = 0$

FIGURE 3. Empirical computation of $\log|\mathbf{M}_s(\mathbf{h})_{12}|$ to assess the coupling between reaction coordinate and orthogonal variable for different initial values of the reaction coordinate $\mathbf{h} = \mathbf{F}(\mathbf{X}_0)$. 2000 compatible samples of the orthgonal variable as prescribed in (4.5) have been used. In both cases parameters of V are $\lambda = 20$, $\mu = 2$, $\tau = 2$ and $\omega = 10$, and we observe exponential decay, but with significantly different initial magnitude.

With this in mind, we set $\Sigma = \sqrt{\Phi\Phi^T}$, and consider the covariance matrix

$$\begin{aligned}\mathbf{M}_s(\mathbf{h}) &:= \beta\mathbb{E}\big[e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{X}\otimes\mathcal{Q}\mathcal{L}\mathbf{X}\,\big|\,\Phi\mathbf{X}=\mathbf{h}\big]\\ &= \beta\mathbb{E}\left[\begin{pmatrix} e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\Sigma^{-1}\Phi\mathbf{X}\otimes\mathcal{Q}\mathcal{L}\Sigma^{-1}\Phi\mathbf{X} & e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\Sigma^{-1}\Phi\mathbf{X}\otimes\mathcal{Q}\mathcal{L}\Psi\mathbf{X}\\ e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\Psi\mathbf{X}\otimes\mathcal{Q}\mathcal{L}\Sigma^{-1}\Phi\mathbf{X} & e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\Psi\mathbf{X}\otimes\mathcal{Q}\mathcal{L}\Psi\mathbf{X}\end{pmatrix}\,\middle|\,\Phi\mathbf{X}=\mathbf{h}\right] \qquad (4.1)\\ &= \begin{pmatrix}\mathbf{M}_s(\mathbf{h})_{11} & \mathbf{M}_s(\mathbf{h})_{12}\\ \mathbf{M}_s(\mathbf{h})_{21} & \mathbf{M}_s(\mathbf{h})_{22}\end{pmatrix}.\end{aligned}$$

We see that $\mathbf{M}_s(\mathbf{h})_{11} = \Sigma^{-1}\mathcal{M}_s(\mathbf{h})\Sigma^{-1}$, recalling the definition of $\mathcal{M}_s$ from (2.11). We also note that if dynamical sampling of the orthogonal dynamics is used to approximate $\mathcal{M}_s$ in practice, all of the information needed to compute $\mathbf{M}_s$ is available.

Intuitively, the off–diagonal blocks of $\mathbf{M}_s$ describe the correlation between the action of the fluctuating force on the reaction coordinates and orthogonal variables, and in particular, $\mathbf{M}_s(\mathbf{h})_{12}$ describes the influence of the orthogonal variables $\Psi\mathbf{X}$ at the current time on the dynamics of the reaction coordinates $\Phi\mathbf{X}$ at later times. We can therefore test the strength of the 'coupling' between the reaction coordinates and the other degrees of freedom by considering the magnitude of $\mathbf{M}_s(\mathbf{h})_{12}$.

In Figure 3, we perform such a comparison for our benchmark example, providing a log scale plot of $\mathbf{M}_s(\mathbf{h})_{12}$ for two different initial conditions. In both cases, we observe exponential decay of the corresponding entry, although the initial value is significantly different. Despite the considerable spatial oscillation of the memory kernel, the exponential decay in time indicates rapid decorrelation between the reaction coordinate and orthogonal variable, and hence suggests that the formula derived in (3.7) provides a good approximation uniformly in space.

4.2. **Comparison of different effective dynamics.**

(1) *Simulations without thermostat.* To compare the different approximation strategies proposed in Section 3, we first simulate the evolution of the mean value of the observable $\mathbf{F}(\mathbf{X}_t)$ for various choices of dynamics without thermostat. For convenience, we recall the relevant governing equations are:

$$\text{Effective system:} \qquad \partial_t\mathbf{h}_t = -\mu\mathbf{h}_t \qquad (4.2)$$

$$\text{Approach 1:} \qquad \partial_t\mathbf{h}_t = -\frac{\mu\mathbf{h}_t}{1+\tau^2\omega^2\cos^2(\omega\mathbf{h}_t)} \qquad (4.3)$$

$$\text{Approach 2:} \qquad \partial_t\mathbf{h}_t = \big(\lambda\tau^2\omega^2\cos^2(\omega\mathbf{h}_t)-1\big)\mu\mathbf{h}_t + \frac{\lambda}{\beta}\tau^2\omega^3\sin(2\omega\mathbf{h}_t), \qquad (4.4)$$

where the Effective system is driven by the effective potential only, Approach 1 is based on the approximation of MZ derived in Section 3.1, and Approach 2 is based on the naïve approximation of the memory proposed in (3.16). In each case, we fix an initial condition $\mathbf{h}_0 = x_0$, and compare with the full dynamics also without thermostat, i.e. we consider $\mathbf{h}_t := \mathbb{E}[\mathbf{F}(\mathbf{X}_t)|\mathbf{F}(\mathbf{X}_0) = x_0]$ by sampling

$$d\mathbf{F}(\mathbf{X}_t) = -\Phi\nabla V(\mathbf{X}_t)dt, \quad \text{where} \quad \mathbf{F}(\mathbf{X}_0) = x_0 \quad \text{and} \quad \Psi\mathbf{X}_0 \sim \mathcal{N}\left(\tau\sin(\omega x_0), \tfrac{1}{\beta\lambda}\right), \qquad (4.5)$$

where as before $\Psi\mathbf{X}_0$ denotes the orthogonal variables. This corresponds to the assumption that initial conditions are distributed according to the marginal of the Gibbs distribution conditioned on the value of $\mathbf{F}(\mathbf{X}_0)$, as assumed in Section 2. Here, we choose $\beta = 10$.

Figure 4 shows the results of these simulations, and indicates that the average trajectories of (4.3) and (4.5) closely correspond. On the other hand, the effective system (4.2) in which the memory contribution is neglected relaxes much faster than the full dynamics, and the naïve choice of memory made to arrive at (4.4) yields qualitatively incorrect behaviour.

(2) *Simulations with thermostat.* As a second comparison of our approximation strategies, we now include the thermostat once more, and consider sample averages of the different dynamics, which are respectively governed by

$$\text{Effective system:} \qquad d\mathbf{h}_t = -\mu \mathbf{h}_t + \sqrt{2\beta^{-1}\Phi\Phi^T}\, d\mathbf{W}_t \tag{4.6}$$

$$\text{Approach 1:} \qquad d\mathbf{h}_t = -\frac{\mu \mathbf{h}_t}{1+\tau^2\omega^2\cos^2(\omega \mathbf{h}_t)}dt + \sqrt{\frac{2\beta^{-1}\Phi\Phi^T}{1+\tau^2\omega^2\cos^2(\omega \mathbf{h}_t)}}\, d\mathbf{W}_t. \tag{4.7}$$

These choices are again compared with the full dynamics of $\mathbf{h}_t$ including the thermostat, which is driven by

$$d\mathbf{F}(\mathbf{X}_t) = -\Phi\nabla V(\mathbf{X}_t)dt + \sqrt{2\beta^{-1}}\Phi d\mathbf{B}_t. \tag{4.8}$$

Simulations are performed at different choices of inverse temperature, and averages are taken over 500 identical realisations of the Brownian motion, aiming to minimise statistical error. In each case, initial conditions are randomly chosen such that

$$\cos(\omega\Phi\mathbf{X}_0) = 0, \qquad \Psi\mathbf{X}_0 \sim \mathcal{N}\big(\tau\sin(\omega\Phi\mathbf{X}_0), \tfrac{1}{\beta\lambda}\big).$$

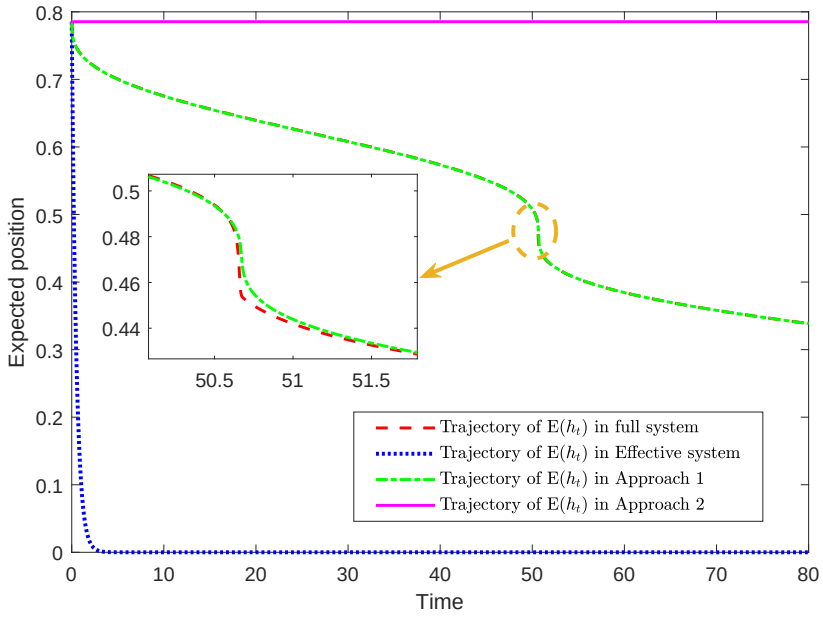
The results of these simulations are shown in Figure 5.

We note the surprising level of agreement between the results obtained using Approach 1 (4.7) and those using the full dynamics (4.8) for both choices of β shown here. Moreover, we also observe improving accuracy as $\beta \to \infty$. This seems to be consistent with the observation that the factor γ which appears in (3.10) is independent of β, and that some of the error terms in the formal expansion derived in (3.9) are $\mathcal{O}(\frac{1}{\lambda\beta})$. In comparison, the results obtained using the Effective system (4.6) again poorly reflect the dynamical properties of the system in all cases.

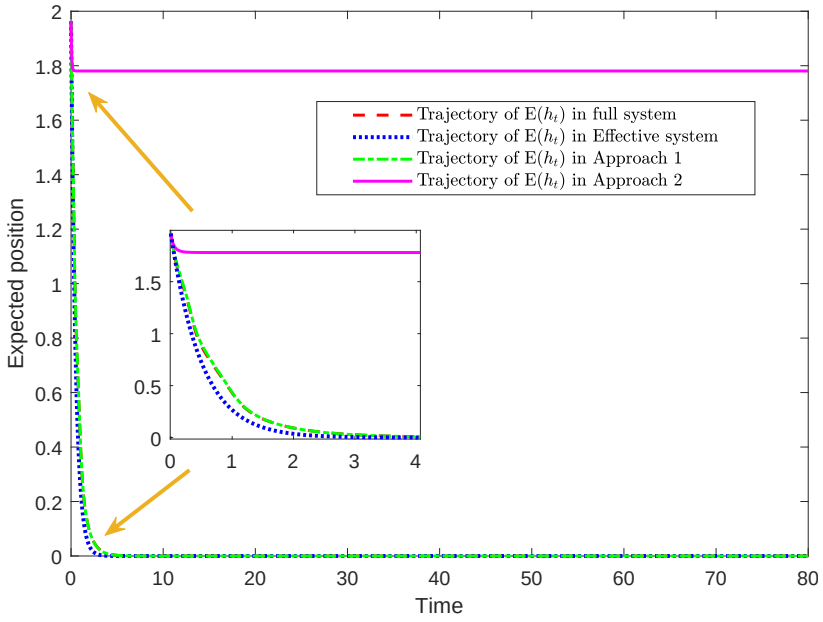
5. CONCLUSION

In this paper, we have employed the Mori–Zwanzig framework to rigorously derive an effective equation for linear reaction coordinates, describing features of an underlying overdamped Langevin dynamics. Such models are appropriate for a variety of applications where we wish to capture only limited aspects of a complex model, such as MD systems in the high friction limit. The equation we derived enables us to understand the sources of error and thereby inform a choice of effective dynamics which better captures dynamical features of the evolution which are not well–represented by the dynamics of the effective potential alone. We hope that this approach can serve to aid practitioners in understanding the sources of error in a coarse–grained model, particularly in the presence of entropic barriers.

We validated our analytic results by considering a benchmark example of overdamped Langevin system in a case where relaxation is impeded by a winding free energy barrier. This necessitated the careful asymptotic treatment of interactions between reaction coordinates and orthogonal variables in order to correctly capture the dynamical behaviour of the system. In particular, although a time-scale separation occurs within the system we considered, we nevertheless showed that careful asymptotic analysis or dynamical sampling is required in general to ensure accuracy. The approximate model we constructed based upon the equations we derived exhibited a drastic improvement

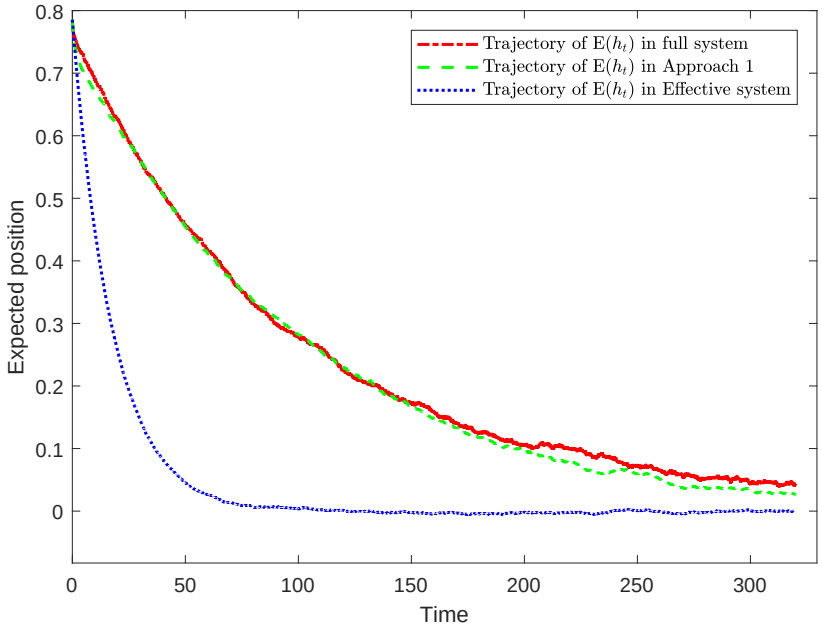


(a) $\tau = 2$, $\omega = 10$

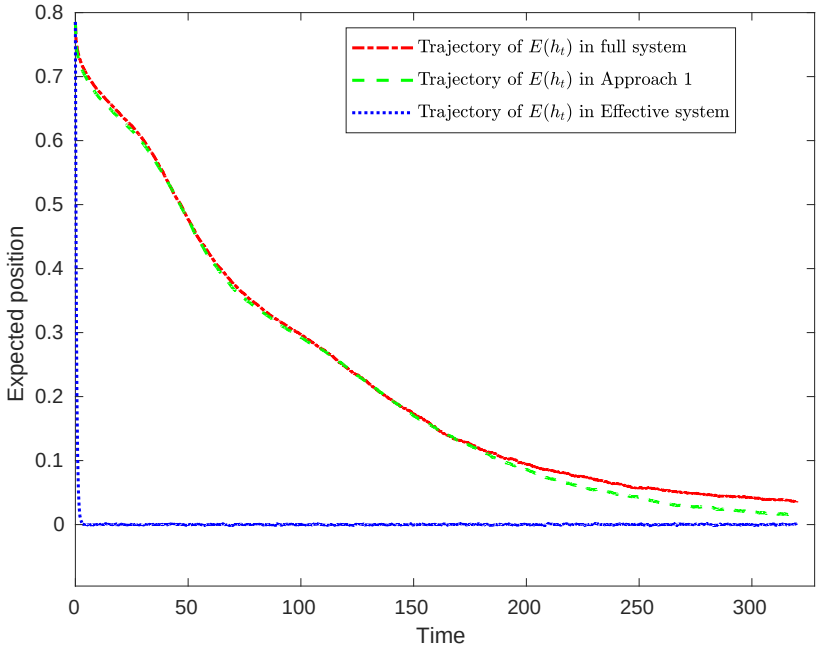


(b) $\tau = 0.2$, $\omega = 4$

FIGURE 4. Mean trajectories of $\mathbf{h}_t$ computed over 1000 realisations of the full dynamics (4.5), the Effective system (4.2), approximation Approach 1 (4.3) and approximation Approach 2 (4.4). Parameters of V are $\lambda = 20$, $\mu = 2$, and the time step was $\Delta t = 10^{-5}$ and $T = 80$.



(a) $\beta = 10$



(b) $\beta = 100$

FIGURE 5. Mean trajectories of $\mathbf{h}_t$ averaged over 500 realisations for two different values of β, respectively computed via the full overdamped Langevin dynamics (4.8), via the Effective system (4.6) and via the approximation Approach 1 (4.7). Identical realisations of Brownian motion are used in each case. Parameters of V are $\lambda = 20$ and $\mu = 2$, $\tau = 2$, $\omega = 10$, and the time step was $\Delta t = 10^{-5}$, total time was $T = 320$.

in predicting the dynamical behaviour of the reaction coordinates over the common approach of using the effective potential alone to describe the dynamics.

Our work prompts several questions, which we hope to address in future:

(1) *Practical sampling algorithms and error analysis.* It would be of practical interest to devise an algorithm to generate effective dynamics based on the asymptotic approximation we considered here, and conduct a rigorous error analysis in this case.
(2) *Extensions to nonlinear coarse–grained variables and full Langevin dynamics.* In this work, we have considered the overdamped Langevin setting with linear reaction coordinates. It would be of significant interest to extend this analysis to nonlinear variables and full Langevin dynamics using our reliable asymptotic analysis approach in future.

Appendix A. Proof of Theorem 2.1

In this section, we provide a proof of our main mathematical result, Theorem 2.1. The proof follows a similar strategy to other derivations using the Mori–Zwanzig formalism given in the literature, notably [11, 22, 26].

A.1. **Construction of 'orthogonal' variables.** Our first step to construct variables which capture directions in phase space which are 'orthogonal' to those captured by the reaction coordinates $\mathbf{F}$, i.e. a foliation of the phase space.

Recall that $\Phi \in \mathbb{R}^{m\times N}$ is a matrix of full rank by assumption, and therefore it follows that the symmetric strictly positive definite square root matrix $\Sigma := \sqrt{\Phi\Phi^T} \in \mathbb{R}^{m\times m}$ exists. Recalling the construction of the singular value decomposition, we find that there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{N\times N}$ such that

$$\Sigma^{-1}\Phi = UDV^T, \quad \text{where} \quad D = \begin{pmatrix} \mathrm{I}_m & 0 \end{pmatrix} \in \mathbb{R}^{m\times N},$$

where $\mathrm{I}_m \in \mathbb{R}^{m\times m}$ is the identity matrix, and 0 denotes a submatrix of zeros. Defining

$$E := \begin{pmatrix} 0 & \mathrm{I}_{N-m} \end{pmatrix} \in \mathbb{R}^{(N-m)\times N},$$

where again $\mathrm{I}_{N-m} \in \mathbb{R}^{(N-m)\times(N-m)}$ is an identity matrix and 0 denotes a matrix of zeros, we set

$$\Psi := EV^T \in \mathbb{R}^{(N-m)\times N}, \qquad \text{and} \qquad \Phi^* := \Phi^T\Sigma^{-2} \in \mathbb{R}^{N\times m}$$

and it follows that

$$\Phi^*\Phi + \Psi^T\Psi = \mathrm{I}_{N\times N}. \tag{A.1}$$

From the construction above, we see that $\Phi^*\Phi \in \mathbb{R}^{N\times N}$ is an orthogonal projection acting on the phase space $\mathbb{R}^N$, and the matrix Ψ 'selects' exactly the orthogonal variables, so that if $\Phi\mathbf{x} = \mathbf{h}$ and $\Psi\mathbf{x} = \widetilde{\mathbf{x}}$, we have

$$\mathbf{x} = \Phi^*\Phi\mathbf{x} + \Psi^T\Psi\mathbf{x} = \Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}.$$

Given Φ, we may use the construction of Ψ in order to define the partition function $Z_\Phi : \mathbb{R}^m \to \mathbb{R}$,

$$Z_\Phi(\mathbf{h}) := \int e^{-\beta V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})} d\widetilde{\mathbf{x}}.$$

A.2. **Dyson–Duhamel principle.** Next, we define

$$\mathbf{h}_t(\mathbf{x}) = \mathbb{E}\big[\mathbf{F}(\mathbf{X}_t) \,\big|\, \mathbf{X}_0 = \mathbf{x}\big].$$

Applying the Feynman–Kac formula, we recall that $\mathbf{h}_t$ solve the PDE

$$\partial_t \mathbf{h}_t = \mathcal{L}\mathbf{h}_t = -\nabla V \cdot \nabla \mathbf{h}_t + \beta^{-1}\Delta \mathbf{h}_t, \quad \text{with} \quad \mathbf{h}_0(\mathbf{x}) = \mathbf{F}(\mathbf{x}).$$

In semigroup notation, we will write $\mathbf{h}_t = e^{t\mathcal{L}}\mathbf{F}$, and so the Feynman–Kac formula becomes

$$\partial_t \mathbf{h}_t = e^{t\mathcal{L}}\mathcal{L}\mathbf{F}. \tag{A.2}$$

Given mutually orthogonal projection operators $\mathcal{P}$ and $\mathcal{Q}$, applying the Dyson–Duhamel principle entails that we have the identity

$$e^{t\mathcal{L}} = \int_0^t e^{(t-s)\mathcal{L}}\mathcal{P}\mathcal{L}e^{s\mathcal{Q}\mathcal{L}}ds + e^{t\mathcal{Q}\mathcal{L}}, \tag{A.3}$$

which can be verified by differentiation with respect to t. Writing $\mathcal{L}\mathbf{F} = \mathcal{P}\mathcal{L}\mathbf{F} + \mathcal{Q}\mathcal{L}\mathbf{F}$ in (A.2) and applying (A.3), we find that

$$\partial_t \mathbf{h}_t = e^{t\mathcal{L}}\mathcal{P}\mathcal{L}\mathbf{F} + \int_0^t e^{(t-s)\mathcal{L}}\mathcal{P}\mathcal{L}e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}\, ds + e^{t\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}. \tag{A.4}$$

Our main focus will now be on the first two terms in (A.4), since the latter term is $\mathcal{F}_t$ as defined in (2.12). To rewrite the former term, we apply the definition of projection operator $\mathcal{P}$, the definitions of Φ, Ψ and Z_Φ and the chain rule, giving

$$\begin{aligned}\mathcal{P}\mathcal{L}\mathbf{F}(\mathbf{x}) &= \Sigma^2\,\mathbb{E}\big[-\Sigma^{-2}\Phi\nabla V \,\big|\, \mathbf{F}(\mathbf{x})\big] \\ &= \frac{\Sigma^2}{Z_\Phi\big(\mathbf{F}(\mathbf{x})\big)}\int -(\Phi^*)^T\nabla V(\Phi^*\,\mathbf{F}(\mathbf{x}) + \Psi^T\widetilde{\mathbf{x}})e^{-\beta V(\Phi^*\,\mathbf{F}(\mathbf{x})+\Psi^T\widetilde{\mathbf{x}})}d\widetilde{\mathbf{x}} = -\Sigma^2\nabla\mathcal{S}\big(\mathbf{F}(\mathbf{x})\big),\end{aligned}$$

where we recall the definition of the effective potential $\mathcal{S}$ given in (2.10).

A.3. **Orthogonal dynamics.** Next, we consider the action of $e^{s\mathcal{Q}\mathcal{L}}$, which will subsequently allow us to treat the integral term in (A.4). We begin by noting that

$$\mathcal{Q}\mathcal{L}\mathbf{F}(\mathbf{x}) = \mathcal{L}\mathbf{F}(\mathbf{x}) - \mathcal{P}\mathcal{L}\mathbf{F}(\mathbf{x}) = -\Phi\nabla V(\mathbf{x}) + \Sigma^2\nabla\mathcal{S}\big(\mathbf{F}(\mathbf{x})\big),$$

and define $\mathbf{g}_t : \mathbb{R}^N \to \mathbb{R}^m$ to be the solution to

$$\partial_t \mathbf{g}_t(\mathbf{x}) = \mathcal{Q}\mathcal{L}\mathbf{F}(\mathbf{x}) \cdot \nabla \mathbf{g}_t(\mathbf{x}), \qquad \mathbf{g}_0(\mathbf{x}) = \mathcal{Q}\mathcal{L}\mathbf{F}(\mathbf{x});$$

using semigroup notation, we write this as $\mathbf{g}_s = e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}$.

Under assumptions (1) and (2) made in Section 2, it can be verified that $\Sigma^2\nabla\mathcal{S}\circ\mathbf{F}$ is globally Lipschitz. Using this fact, it can therefore be shown using the method of characteristics that $\mathbf{g}_s$ exists and is a C^1 diffeomorphism on $\mathbb{R}^N$; for similar results in the Hamiltonian setting, see [22].

Moreover, defining $\mathbf{a}_s := \mathbf{g}_s(\mathbf{x}) \in \mathbb{R}^m$ and $\mathbf{b}_s := \mathbf{g}_s(\mathbf{y}) \in \mathbb{R}^m$, then we use the fact that $\nabla\mathcal{S}$ and ∇V are Lipschitz along with Young's inequality to deduce that

$$\begin{aligned}&\frac{d}{ds}\frac{1}{2}\,|\mathbf{a}_s - \mathbf{b}_s|^2 \\ &\quad= -\Big(\Phi\nabla V(\Phi^*\mathbf{a}_s + \Psi^T\Psi\mathbf{x}) - \Phi\nabla V(\Phi^*\mathbf{b}_s + \Psi^T\Psi\mathbf{y}) - \Sigma^2\nabla\mathcal{S}(\mathbf{a}_s) + \Sigma^2\nabla\mathcal{S}(\mathbf{b}_s)\Big)\cdot(\mathbf{a}_s - \mathbf{b}_s), \\ &\quad\lesssim \Big(\big|\Phi^*(\mathbf{a}_s - \mathbf{b}_s) + \Psi^T\Psi(\mathbf{x}-\mathbf{y})\big| + |\mathbf{a}_s - \mathbf{b}_s|\Big)|\mathbf{a}_s - \mathbf{b}_s|, \\ &\quad\lesssim |\mathbf{a}_s - \mathbf{b}_s|^2 + |\mathbf{x}-\mathbf{y}|^2.\end{aligned}$$

Applying Gronwall's inequality in the usual way, it follows that there exists $\alpha > 0$ such that

$$|\mathbf{g}_s(\mathbf{x}) - \mathbf{g}_s(\mathbf{y})| \leq |\mathbf{x} - \mathbf{y}|\sqrt{s}e^{\alpha s}, \qquad \text{and thus} \qquad |\nabla \mathbf{g}_s(\mathbf{x})| \leq \sqrt{s}e^{\alpha s}. \tag{A.5}$$

A.4. **Memory integral.** Now that we have established properties of $\mathbf{g}_s$, we return to the integral term in (A.4). For now, we fix $\mathbf{x} \in \mathbb{R}^N$, and set $\mathbf{h} := \mathbf{F}(\mathbf{x})$.

Consider $\mathcal{P}\mathcal{L}\mathbf{g}_s$; using the partition of the identity constructed in (A.1) we may write

$$\begin{aligned}
&\mathcal{P}\mathcal{L}\mathbf{g}_s(\mathbf{x}) \\
&= \frac{1}{Z_\Phi(\mathbf{h})} \int \Big(- \nabla \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) \cdot \nabla V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) + \beta^{-1}\Delta \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})\Big) e^{-\beta V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})} d\widetilde{\mathbf{x}} \\
&= \frac{1}{Z_\Phi(\mathbf{h})} \int \Big(- \nabla \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})\Phi^*\Phi \nabla V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) + \beta^{-1}\Delta \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})\Phi^*\Phi \\
&\qquad\qquad - \nabla \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})\Psi^T\Psi \nabla V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) + \beta^{-1}\Delta \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})\Psi^T\Psi\Big) \\
&\qquad\qquad\qquad\qquad \times e^{-\beta V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})} d\widetilde{\mathbf{x}}.
\end{aligned}$$

We collect terms involving matrix products with Φ and Ψ separately, and using the chain rule, we find that

$$\begin{aligned}
\mathcal{P}\mathcal{L}\mathbf{g}_s(\mathbf{x}) = &\underbrace{\frac{1}{Z_\Phi(\mathbf{h})} \int \mathsf{div}_{\widetilde{\mathbf{x}}} \left(\frac{1}{\beta} \nabla_{\widetilde{\mathbf{x}}} \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) e^{-\beta V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})} \right) d\widetilde{\mathbf{x}}}_{=:T_1} \\
&+ \underbrace{\frac{1}{Z_\Phi(\mathbf{h})} \int \mathsf{div}_{\mathbf{h}} \left(\frac{1}{\beta} \nabla_{\mathbf{h}} \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) \Sigma^2 e^{-\beta V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})} \right) d\widetilde{\mathbf{x}}}_{=:T_2},
\end{aligned}$$

where subscripts denote the variable with respect which derivatives are taken. In particular, in the formula above, $\mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\mathbf{x})$ is treated as a composition of functions. We now consider each of the terms T_1 and T_2 separately.

To treat T_1, we apply the divergence theorem. Truncating the domain of integration to $B_R(0) \subset \mathbb{R}^{N-m}$, a ball of radius R centred at 0, and considering the limit as $R \to \infty$, we have

$$\begin{aligned}
T_1 &= \frac{1}{Z_\Phi(\mathbf{h})} \lim_{R\to\infty} \int_{B_R(0)} \mathsf{div}_{\widetilde{\mathbf{x}}} \left(\frac{1}{\beta} \nabla_{\widetilde{\mathbf{x}}} \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) e^{-\beta V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})} \right) d\widetilde{\mathbf{x}} \\
&= \frac{1}{Z_\Phi(\mathbf{h})} \lim_{R\to\infty} \int_{\partial B_R(0)} \left(\frac{1}{\beta} \nabla_{\widetilde{\mathbf{x}}} \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) e^{-\beta V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})} \right) \cdot \vec{\nu}\, d\widetilde{S}.
\end{aligned} \tag{A.6}$$

Applying (A.5) and the growth assumptions on V to pass to the limit, we see that $T_1 = 0$.

For T_2, we note that we may commute differentiation and integration, and so multiplying and dividing by $Z_\Phi(\mathbf{h})$, we obtain, we obtain

$$\begin{aligned}
T_2 &= \frac{1}{Z_\Phi(\mathbf{h})} \mathsf{div}_{\mathbf{h}} \left(\int \frac{1}{\beta} \nabla_{\mathbf{h}} \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) \Sigma^2 e^{-\beta V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})} d\widetilde{\mathbf{x}} \right) \\
&= \frac{1}{Z_\Phi(\mathbf{h})} \mathsf{div}_{\mathbf{h}} \left(\frac{Z_\Phi(\mathbf{h})}{\beta} \frac{1}{Z_\Phi(\mathbf{h})} \int \nabla_{\mathbf{h}} \mathbf{g}_s(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}}) \Sigma^2 e^{-\beta V(\Phi^*\mathbf{h} + \Psi^T\widetilde{\mathbf{x}})} d\widetilde{\mathbf{x}} \right).
\end{aligned} \tag{A.7}$$

A.5. **Memory kernel.** To complete our analysis, we must show the identity

$$\mathcal{M}_s(\mathbf{h}) = -\frac{1}{Z_\Phi(\mathbf{h})}\int \nabla_{\mathbf{h}}\mathbf{g}_s(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})\Sigma^2 e^{-\beta V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})}d\widetilde{\mathbf{x}}, \tag{A.8}$$

where we recall that $\mathcal{M}_s$ was defined in (2.11).

Since we may again commute differentiation and integration, we use the product rule to write

$$\begin{aligned}
&-\frac{1}{Z_\Phi(\mathbf{h})}\int \nabla_{\mathbf{h}}\mathbf{g}_s(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})\Sigma^2 e^{-\beta V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})}d\widetilde{\mathbf{x}}\\
&\quad= -\frac{1}{Z_\Phi(\mathbf{h})}\int \nabla_{\mathbf{h}}\left(\mathbf{g}_s(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})e^{-\beta V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})}\right)\Sigma^2 d\widetilde{\mathbf{x}}\\
&\qquad -\frac{1}{Z_\Phi(\mathbf{h})}\int \mathbf{g}_s(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})\otimes\beta\,(\Phi^*)^T\nabla V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})e^{-\beta V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})}\,\Sigma^2 d\widetilde{\mathbf{x}}\\
&\quad= -\frac{1}{Z_\Phi(\mathbf{h})}\nabla_{\mathbf{h}}\left(\int \mathbf{g}_s(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})e^{-\beta V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})}d\widetilde{\mathbf{x}}\right)\\
&\qquad -\frac{1}{Z_\Phi(\mathbf{h})}\int \left(\mathbf{g}_s(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})\otimes\beta\,\Phi\nabla V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})e^{-\beta V(\Phi^*\mathbf{h}+\Psi^T\widetilde{\mathbf{x}})}\right)d\widetilde{\mathbf{x}},\\
&\quad= \underbrace{-\frac{1}{Z_\Phi(\mathbf{h})}\nabla_{\mathbf{h}}\Big(Z_\Phi(\mathbf{h})\mathbb{E}[\mathbf{g}_s|\mathbf{F}(\mathbf{x})=\mathbf{h}]\Big)}_{=:T_{11}} + \underbrace{\beta\,\mathbb{E}[\mathbf{g}_s\otimes\mathcal{L}\mathbf{F}|\mathbf{F}(\mathbf{x})=\mathbf{h}]}_{=:T_{12}}.
\end{aligned}$$

Next, we recall that $\mathbf{g}_s = e^{s\mathcal{Q}\mathcal{L}}\mathbf{F}$, and $\mathcal{P}\mathcal{Q}=0$, so

$$\mathbb{E}\big[\mathbf{g}_s\big|\mathbf{F}(\mathbf{x})\big] = \mathbb{E}\big[e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}\big|\mathbf{F}(\mathbf{x})\big] = \left(\mathcal{P}\mathcal{Q}\mathcal{L}e^{s\mathcal{Q}\mathcal{L}}\mathbf{F}\right)(\mathbf{x}) = 0, \tag{A.9}$$

and therefore $T_{11}=0$. To treat T_{12}, we note that since $\mathcal{P}+\mathcal{Q}=\mathcal{I}$, we may split T_{12} into

$$\begin{aligned}
T_{12} &= \beta\,\mathbb{E}[e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}\otimes\mathcal{Q}\mathcal{L}\mathbf{F}|\mathbf{F}(\mathbf{x})=\mathbf{h}] + \beta\,\mathbb{E}[e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}\otimes\mathcal{P}\mathcal{L}\mathbf{F}|\mathbf{F}(\mathbf{x})=\mathbf{h}]\\
&= \mathcal{M}_s(\mathbf{h}) + \beta\,\mathcal{P}e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\mathbf{F}\otimes\mathcal{P}\mathcal{L}\mathbf{F}.
\end{aligned}$$

Once again, the latter term vanishes thanks to (A.9), and so we have proved identity (A.8).

A.6. **Conclusion of the proof.** Applying identity (A.8) to (A.7) and using the product rule and the definition of the effective potential given in (2.10), we find that

$$T_2 = \frac{1}{Z_\Phi(\mathbf{h})}\mathsf{div}\left(-\frac{Z_\Phi(\mathbf{h})}{\beta}\mathcal{M}_s(\mathbf{h})\right) = \mathcal{M}_s(\mathbf{h})\nabla\mathcal{S}(\mathbf{h}) - \frac{1}{\beta}\mathsf{div}\mathcal{M}_s(\mathbf{h}).$$

Combining our analysis of each of the terms, we have therefore shown that

$$\partial_t\mathbf{h}_t = -\Sigma^2\nabla\mathcal{S}(\mathbf{h}_t) + \int_0^t \mathcal{M}_s(\mathbf{h}_{t-s})\nabla\mathcal{S}(\mathbf{h}_{t-s}) - \frac{1}{\beta}\mathsf{div}\mathcal{M}_s(\mathbf{h}_{t-s})\,ds + \mathcal{F}_t$$

Hence, we prove the theorem.

Acknowledgment

We would like to thank Dr Xiantao Li for helpful suggestions and encouragement during this project. Our thanks also go to the two anonymous referees, whose suggestions helped us to significantly improve and clarify many aspects of this paper.

References

[1] R. V. Abramov and A. J. Majda. Quantifying uncertainty for non-gaussian ensembles in complex systems. *SIAM Journal on Scientific Computing*, 26:411–447, 2004.

[2] C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Springer, 1999.

[3] M. Berkowitz, J. Morgan, and J. A. McCammon. Generalized Langevin dynamics simulations with arbitrary time-dependent memory kernels. *The Journal of Chemical Physics*, 78:3256, 1983.

[4] B. J. Berne and R. Pecora. Dynamic light scattering: With applications to chemistry, biology, and physics. Dover: New York, 2000.

[5] M. Branicki and A. J. Majda. Quantifying uncertainty for predictions with model error in non-Gaussian systems with intermittency. *Nonlinearity*, 25(9), 2012.

[6] M. Branicki and A. J. Majda. An information-theoretic framework for improving imperfect dynamical predictions via multi-model ensemble forecasts. *Journal of Nonlinear Science*, 25:489–538, 2015.

[7] M. Chen, X. Li, and C. Liu. Computation of the memory functions in the generalized Langevin models for collective dynamics of macromolecules. *The Journal of Chemical Physics*, 141:064112, 2014.

[8] A. Chorin and P. Stinis. Problem reduction, renormalization, and memory. *Communications in Applied Mathematics and Computational Science*, 1:1–27, 2006.

[9] A. J. Chorin and O. H. Hald. In *Stochastic Tools in Mathematics and Science*. Springer, 2013.

[10] A. J. Chorin, O. H. Hald, and R. Kupferman. Optimal prediction and the morizwanzig representation of irreversible processes. *Proceedings of the National Academy of Sciences of the United States of America*, 97:6253–6257, 2000.

[11] A. J. Chorin, O. H. Hald, and R. Kupferman. Optimal prediction with memory. *Physica D*, 166:239–257, 2002.

[12] A. J. Chorin and F. Lu. Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics, 2015.

[13] E. Darve, J. Solomon, and A. Kia. Computing generalized Langevin equations and generalized FokkerPlanck equations. *Proceedings of the National Academy of Sciences of the United States of America*, 106:10884–10889, 2009.

[14] N. di Pasquale, D. Marchisio, and P. Carbone. Mixing atoms and coarse-grained beads in modelling polymer melts. *The Journal of Chemical Physics*, 137(16):164111, 2012.

[15] M. Dobson. Information theoretic fitting for coarse-grained molecular dynamics. preprint. 2016 SIAM Conference on Mathematical Aspects of Materials Science.

[16] W. E. In *Principles of Multiscale Modeling*. Cambridge University Press, 2011.

[17] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. Heterogeneous multiscale methods: A review. *Communications in Computational Physics*, 2:367–450, 2007.

[18] P. Espanol and P. Warren. Statistical mechanics of dissipative particle dynamics. *Europhysics Letters*, 30:191, 1995.

[19] D. J. Evans and G. Morriss. Statistical mechanics of nonequilibrium liquids. Cambridge University Press, 2008.

[20] G. W. Ford, M. Kac, and P. Mazur. Statistical mechanics of assemblies of coupled oscillators. *Journal of Mathematical Physics*, 6:504–515, 1965.

[21] M. Frank and B. Seibold. Optimal prediction for radiative transfer: A new perspective on moment closure. *Kinetic and Related Models*, 4:717–733, 2011.

[22] D. Givon, R. Kupferman, and O. H. Hald. Existence proof for orthogonal dynamics and the Mori-Zwanzig formalism. *Israel Journal of Mathematics*, 145(1):221–241, Dec 2005.

[23] I. Grooms and A. J. Majda. Efficient stochastic superparameterization for geophysical turbulence. *Proceedings of the National Academy of Sciences of the United States of America*, 110:4464–4469, 2013.

[24] N. Guttenberg, J. F. Dama, M. G. Saunders, G. A. Voth, J. Weare, and A. R. Dinner. Minimizing memory as an objective for coarse-graining. *The Journal of Chemical Physics*, 138:094111, 2013.

[25] C. Hartmann. Model reduction in classical molecular dynamics, 2007. PhD Thesis. Fachbereich Mathematik und Informatik Freie Universitat Berlin.

[26] C. Hijn, P. Espaol, E. Vanden-Eijndenc, and R. Delgado-Buscalioni. MoriZwanzig formalism as a practical computational tool. *Faraday Discussions*, 144:301–322, 2010.

[27] S. Izvekov and G. A. Voth. Modeling real dynamics in the coarse-grained representation of condensed phase systems. *Journal of Chemical Physics*, 125:151101151104, 2006.

[28] M. A. Katsoulakis and P. Plechac. Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. *The Journal of Chemical Physics*, 139:74–115, 2013.

[29] T. Kinjo and S. aki Hyodo. Equation of motion for coarse-grained simulation based on microscopic description. *Physical Review E*, 75:051109, 2007.
[30] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28:985–1010, 2018.
[31] R. Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255–284, jan 1966.
[32] R. Kupferman. Fractional kinetics in kac–zwanzig heat bath models. *Journal of Statistical Physics*, 114(112), 2004.
[33] F. Legoll and T. Lelievre. Effective dynamics using conditional expectations. *Nonlinearity*, 23:2131–2163, 2010.
[34] F. Legoll and T. Lelievre. Some remarks on free energy and coarse-graining. In *Numerical Analysis and Multiscale Computations, B. Engquist, O. Runborg, R. Tsai eds., Springer Lecture Notes in Computational Science and Engineering*, volume 82, pages 279–329. Springer, 2012.
[35] F. Legoll, T. Lelievre, and S. Olla. Pathwise estimates for an effective dynamics. *Stochastic Processses and their Applications*, 127:2841–2863, 2017.
[36] F. Legoll, T. Lelievre, and U. Sharma. Effective dynamics for non-reversible stochastic differential equations: a quantitative study. manuscript.
[37] H. Lei, N. Baker, and X. Li. Data-driven parameterization of the generalized Langevin equation. *Proceedings of the National Academy of Sciences of the United States of America*, 113:14183–14188, 2016.
[38] H. Lei, B. Caswell, and G. E. Karniadakis. Direct construction of mesoscopic models from microscopic simulations. *Physical Review E*, 81:026704, 2010.
[39] T. Lelièvre, M. Rousset, and G. Stoltz. In *Free energy computations: A mathematical perspective*. Imperial College Pressr, 2010.
[40] T. Lelièvre and W. Zhang. Pathwise estimates for effective dynamics: the case of nonlinear vectorial reaction coordinates. 2018. hal-01794919.
[41] X. Li. A coarse-grained molecular dynamics model for crystalline solids. *International Journal for Numerical Methods in Engineering*, 83:986–997, 2010.
[42] Z. Li, X. Bian, X. Li, and G. E. Karniadakis. Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism. *The Journal of Chemical Physics*, 143:243128, 2015.
[43] Z. Li, H. S. Lee, E. Darve, and G. E. Karniadakis. Computing the non-Markovian coarse-grained interactions derived from the MoriZwanzig formalism in molecular systems: application to polymer melts. *The Journal of Chemical Physics*, 146, 2017.
[44] G. Lindgren. Stationary stochastic processes: Theory and applications. Chapman and Hall/CRC, 2012.
[45] F. Lu, X. Tu, and A. J. Chorin. Accounting for model error from unresolved scales in ensemble kalman filters by stochastic parameterization, 2017.
[46] L. Ma, X. Li, and C. Liu. Coarse-graining Langevin dynamics using reduced-order techniques, 2018. arxiv:1802.10133.
[47] A. Majda and X. Wang. Nonlinear dynamics and statistical theories for basic geophysical flows. Cambridge University Press, 2006.
[48] A. J. Majda. Introduction to turbulent dynamical systems in complex systems. In *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*. Springer, 2016.
[49] A. J. Majda, R. V. Abramov, and M. J. Grote. Information theory and stochastics for multiscale nonlinear systems. American Mathematical Society, 2005.
[50] A. J. Majda, M. Branicki, and Y. Frenkel. Improving complex models through stochastic parameterization and information theory, 2011. ECMWF-WCRP, Thorpex Workshop on Model Uncertainty.
[51] G. F. Mazenko. In *Nonequilibrium Statistical Mechanics*. Wiley-Vch, 2006.
[52] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993.
[53] S. P. Meyn and R. L. Tweedie. Stability of markovian processes iii: Foster-lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993.
[54] H. Mori. Transport, collective motion, and Brownian motion. *Progress of Theoretical Physics*, 33(3):423–455, 1965.
[55] T. Munakata. Generalized Langevin-equation approach to impurity diffusion in solids: perturbation theory. *Physical Review B*, 33:8017, 1985.
[56] S. Nordholm and R. Zwanzig. A systematic derivation of exact generalized Brownian motion theory. *Journal of Statistical Physics*, 13(4):340–370, 1975.

[57] M. Ottobre and G. A. Pavliotis. Asymptotic analysis for the generalized langevin equation. *Nonlinearity*, 24(5), 2011.
[58] N. D. Pasquale, T. Hudson, and M. Icardi. Systematic derivation of hybrid coarse-grained models, 2018. arxiv:1804.08157.
[59] G. A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.
[60] F. Pinski, G. Simpson, A. M. Stuart, and H. Weber. Algorithms for Kullback-Leibler approximation of probability measures in infinite dimensions. *SIAM Journal on Scientific Computing*, 37(6):2733–2757, 2014.
[61] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996.
[62] I. Snook. The langevin and generalised Langevin approach to the dynamics of atomic, polymeric and colloidal systems. Elsevier, Amsterdam, 2017.
[63] P. Stinis. Renormalized MoriZwanzig-reduced models for systems without scale separation. *Proceedings of the Royal Society A*, 471:20140446, 2015.
[64] T. D. Swinburne. Stochastic dynamics of crystal defects, 2015. PhD Thesis. Imperial College London, Department of Physics.
[65] M. E. Velinova. In *Coarse-grained Molecular Dynamics*. Delve Pub, 2017.
[66] D. Venturi, H. Cho, and G. E. Karniadakis. Mori-Zwanzig approach to uncertainty quantification. In *Handbook of Uncertainty Quantification*, pages 1–36. Springer, 2016.
[67] Y. Yoshimoto, I. Kinefuchi, T. Mima, A. Fukushima, T. Tokumasu, and S. Takagi. Bottom-up construction of interaction models of non-Markovian dissipative particle dynamics. *Physical Review E*, 88:043305, 2013.
[68] Y. Zhu, J. Dominy, and D. Venturi. On the estimation of the Mori-Zwanzig memory integral. *Journal of Mathematical Physics*, 59:103501, 2018.
[69] Y. Zhu and D. Venturi. Faber approximation to the Mori-Zwanzig equation. *Journal of Computational Physics*, 372:694–718, 2018.
[70] R. Zwanzig. Memory effects in irreversible thermodynamics. *Physical Review*, 124(4):983–922, 1961.
[71] R. Zwanzig. Nonlinear generalized Langevin equations. *Journal of Statistical Physics*, 9:215–220, 1973.
[72] R. Zwanzig. Nonequilibrium statistical mechanics. In *Nonequilibrium statistical mechanics*. Oxford University Press, 2001.