Topics in multiscale modeling: numerical analysis and applications

A thesis submitted for the degree of
Doctor of Philosophy of Imperial College London
by

Urbain Philippe Vaes

Department of Mathematics
Imperial College London
London SW7 2AZ

March 2019

Declaration of originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

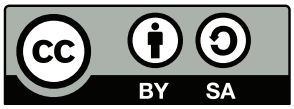
License

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative "Attribution-ShareAlike 4.0 Interational" (CC BY-SA).

Under this licence, you may copy and redistribute the material in any medium or format for both commercial and non-commercial purposes. You may also create and distribute modified versions of the work. This on the condition that: you credit the author and share any derivative works under the same licence.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.



Abstract

Most of the material presented in this thesis is based on [2, 11, 95, 177]. We explore several topics in multiscale modeling, with an emphasis on numerical analysis and applications. Throughout Chapters 2 to 4, our investigation is guided by asymptotic calculations and numerical experiments based on spectral methods.

In Chapter 2, we present a new method for the solution of multiscale stochastic differential equations at the diffusive time scale. In contrast to averaging-based methods, the numerical methodology that we present is based on a spectral method. We use an expansion in Hermite functions to approximate the solution of an appropriate Poisson equation, which is used in order to calculate the coefficients in the homogenized equation.

Extensions of this method are presented in Chapters 3 and 4 and employed for the investigation of the Desai–Zwanzig mean-field model with colored noise and the generalized Langevin dynamics in a periodic potential, respectively. In Chapter 3, we study in particular the effect of colored noise on bifurcations and phase transitions induced by variations of the temperature. In Chapter 4, we investigate the dependence of the effective diffusion coefficient associated with the generalized Langevin equation on the parameters of the equation.

In Chapter 5, which is independent from the rest of this thesis, we introduce a novel numerical method for phase-field models with wetting. More specifically, we consider the Cahn–Hilliard equation with a nonlinear wetting boundary condition, and we propose a class of linear, semi-implicit time-stepping schemes for its solution.

Acknowledgments

First and foremost, I would like thank my parents, Christophe and Nathalie, my brothers, Julien and Clément, and my grandparents, Jacqueline, Henri, Marie-Christine and Michel, for their support, their love, and for having encouraged me throughout life.

I would also like to offer my heartfelt thanks to my supervisors: to Grigorios Pavliotis, for his steady support, his kindness, his thoughtful scientific guidance, and for having believed in me in times when I doubted myself; and to Serafim Kalliadasis, for his equally steady support, his encouragements, his attention to detail, and for having welcomed me in his research group. Without either of you, I would not have been able to complete this work.

I also wish to thank the other people whose help or guidance I feel very fortunate to have benefited from during my PhD years. I thank in particular Benjamin Aymard, Assyr Abdulle, Gabriel Stoltz, Andrew Duncan, John Barrett and Marc Pradas, who all provided unconditional help with specific aspects of my research. I also thank Julien Roussel for useful discussions and for having kindly shared with me some unpublished research notes on hypocoercivity for the Langevin dynamics, Andy Thomas for his patient and continuous assistance in operating the departmental computing cluster, and Susana Gomes for the enjoyable collaboration and for her PhD-related advice.

I am also grateful to my viva examiners, Ben Goddard and Demetrios Papageorgiou, for their useful suggestions, their insightful comments and their interest in this work.

I shall take this opportunity to also thank Jean-François Remacle, my master's thesis supervisor, for his continuous support and his friendship, and Sébastien Combébis, for the inspirational role he played in my first year of undergrad by sharing with us his knowledge and love of the art of programming.

I would also like to express my gratitude to the friends that surrounded me during these years in London. Particular thanks to Khoi, for his generosity, for having introduced to the world of *GNU/Linux* and *Vim*, and for having shared with me his knowledge of and passion for mathematics; to Francesc, for his enthusiasm and his companionship; to Alexandre, Andrea, Istvan, Thomas and Nicolas, for the good years spent together in Earl's court and the little everyday gestures, such as the regular chocolate cakes Nicolas prepared for all of us; to Pierre, for the beer we had on my first day at college and for his constant friendship ever since; to Adam and Angelika, for being great office mates, and finally to Alex, Andrei, Cezary, Chloe, Ghalib, Kalle, Krystle, Ivan, Jens, Juliana, Luc, Luca, Marco, Marianna, Marina, Massimiliano, Nick, Pedro, Oana, Oriane, Ramandeep, Shalini, Silvia, Srikanth, Tom B, Tom L, Yidu, and the members of the Imperial social football club.

I also wish to thank my friends in Belgium, who always welcomed me as if no time had passed whenever I went back, as well as the friends who do not fit in any of the previous categories yet contributed to making this part of my life a happy one. Special thanks to Charles, for his lifelong friendship; to his partner Chaewon, for her kindness and for all the Korean meals she cooked for us; to Victor and Catherine, for their decade-long friendship and for always making me feel welcome and loved; to Hazimah, for her friendship since Delft and the numerous typos she spotted in an earlier version of this thesis; and finally to Allan, Arnaud, Aude-Line, Chloé, Colin, Coline, Corentin, Didier, Freddie, Grigoriy, Guillaume, Hugues, Laurent, Ludovic, Mathilde, Maxime, Nicolas V, Oksana, Pierre, Quentin, Tanguy, Victor X and Vincent.

Finally, I would like to extend my gratitude to the contributors to the free software projects that made the research presented in this thesis possible, or at least helped its realization. In particular, I thank Frédéric Hecht for *FreeFem++*, Christophe Geuzaine and Jean-François Remacle for *Gmsh*, Donald Knuth and Leslie Lamport for TEX and LATEX, Guido van Rossum for *Python*, Bram Moolenaar for *Vim*, Linus Thorvalds for *Linux* and *Git*, and Richard Stallman for the *GNU* project and his devotion to software freedom.

Contents

1 Introduction **25**

1.1 Brief introduction to the thesis . . . 26

1.2 Numerical solution of multiscale SDEs . . . 27

1.2.1 Averaging for multiscale SDEs . . . 27

1.2.2 Homogenization for multiscale SDEs . . . 30

1.2.3 The heterogeneous multiscale method . . . 33

1.3 The Desai–Zwanzig mean-field model with colored noise . . . 39

1.3.1 The white noise problem . . . 40

1.3.2 Colored noise problem . . . 44

1.4 Diffusion of a Brownian particle in a periodic potential . . . 45

1.4.1 Relationship between the GLE and Langevin dynamics . . . 46

1.4.2 Calculation of the effective diffusion coefficient . . . 47

1.4.3 Convergence to equilibrium . . . 52

1.5 The Cahn–Hilliard equation . . . 56

1.5.1 Formulation as a gradient flow . . . 57

1.5.2 Models for the free energies and the mobility . . . 59

1.5.3 The sharp interface limit . . . 62

1.5.4 Coupling with the Navier–Stokes equation . . . 63

1.6 Contributions of the thesis . . . 65

2 Hermite spectral method for multiscale SDEs **69**

2.1 Introduction . . . 70

2.2 Diffusion Approximation and Homogenization . . . 74

2.3 Numerical Method . . . 76

2.4 Main Results . . . 80

2.5 Proofs of the Main Results . . . 83

2.5.1 Convergence of the Spectral Method for the Poisson Equation . . . 83

2.5.2 Convergence of the Drift and Diffusion Coefficients . . . 85

2.5.3 Convergence of the Solution to the SDE . . . 86

2.6 Implementation and Numerical Experiments . . . 88

2.6.1 Implementation details . . . 88
2.6.2 Numerical experiments . . . 90
2.7 Conclusion and Further Work . . . 101
2.A Proof of the Lipschitz continuity of the effective coefficients . . . 102

3 Mean field limits for interacting diffusions with colored noise 105
3.1 Introduction . . . 106
3.2 The model . . . 110
3.2.1 Mean-field limit . . . 111
3.2.2 The white noise limit . . . 113
3.3 The numerical method . . . 115
3.3.1 Linear Fokker–Planck equation with white noise . . . 116
3.3.2 McKean–Vlasov equation with white noise . . . 119
3.3.3 Linear Fokker–Planck equation with colored noise . . . 122
3.3.4 McKean–Vlasov equation with colored noise . . . 125
3.3.5 Monte Carlo simulations . . . 127
3.4 Asymptotic analysis for the Galerkin formulation . . . 127
3.4.1 Suitable index sets . . . 128
3.4.2 Effective drift and diffusion coefficients . . . 130
3.4.3 Numerical verification of the rates of convergence . . . 131
3.5 Results: effect of colored noise on bifurcations . . . 133
3.5.1 Construction of the bifurcation diagrams for the mean-field equation . . . 133
3.5.2 Gaussian case . . . 134
3.5.3 Non-Gaussian noise . . . 135
3.5.4 Dependence of the critical temperature on ε . . . 136
3.6 Conclusions . . . 137
3.A Proof of Theorem 3.1 . . . 138
3.B Asymptotic analysis . . . 141
3.B.1 Ornstein–Uhlenbeck noise . . . 145
3.B.2 Harmonic noise . . . 146
3.B.3 General reversible noise . . . 148
3.C Additional numerical results . . . 148
3.C.1 Convergence of the method for the linear Fokker-Planck equation . . . 148
3.C.2 Convergence to the white noise limit . . . 150

4 Generalized Langevin equation in a periodic potential 153
4.1 Introduction . . . 154
4.2 Model and derivation of the effective diffusion . . . 156
4.3 Asymptotic analysis . . . 160

4.3.1 The overdamped limit . . . 163
4.3.2 The underdamped limit . . . 166
4.3.3 The short memory limit . . . 168
4.4 Numerical experiments . . . 172
4.4.1 Calculation of the velocity autocorrelation function . . . 173
4.4.2 Convergence of the method . . . 175
4.4.3 Effective diffusion . . . 177
4.5 Convergence of the GLE dynamics to equilibrium . . . 178
4.6 Conclusions . . . 182
4.7 Outlook: extension to two-dimensional periodic potentials . . . 183
4.A Effective diffusion coefficient for model (4.1a) . . . 185
4.B Attempt at proving discrete hypocoercivity in $H^1(\mu)$ . . . 186

5 A linear finite element method for the Cahn–Hilliard equation with wetting 189
5.1 Introduction . . . 190
5.2 Phase-field model for wetting phenomena . . . 192
5.2.1 Existence of a solution . . . 194
5.3 Numerical method . . . 195
5.3.1 OD2-W scheme . . . 195
5.4 Numerical results . . . 200
5.4.1 Equilibrium contact angle . . . 200
5.4.2 Convergence of the method . . . 201
5.4.3 Time-adaptation scheme . . . 203
5.4.4 Wetting in complex geometries and with heterogeneous substrates . . . 205
5.5 Extension to the Cahn–Hilliard–Navier–Stokes system . . . 212
5.6 Conclusions . . . 213
5.A Proof of Theorem 5.1 . . . 214

6 Software contributions 219
6.1 A Python library for the Hermite spectral method . . . 219
6.1.1 Library settings . . . 221
6.1.2 Calculation of the discrete Hermite transform . . . 222
6.1.3 Discretization of operators . . . 224
6.1.4 Examples . . . 225
6.1.5 Outlook . . . 228
6.2 Finite-element framework for Cahn–Hilliard equation . . . 229

7 Conclusions 233

Bibliography **237**

A Background material **255**

A.1 Stochastic differential equations 255

A.2 Weighted Sobolev Spaces 258

A.3 Hermite polynomials and Hermite functions 261

A.3.1 Hermite polynomials and Hermite functions in 1 dimension 262

A.3.2 Hermite polynomials and Hermite functions in n dimensions 267

A.4 Discrete transforms and interpolation error 269

List of Tables

3.1 Value of ζ 114

4.1 Numerical parameters used to generate the data presented in Fig. 4.10. 177

5.1 CPU times (hh:mm:ss) using an Intel i7-3770 processor for the simulations presented in Section 5.4.3 (two droplets on a substrate), with or without time-step adaptation. The method OD2-W was used for the simulations with an adaptive time step, and the method OD1-W was used for the simulations with a fixed time step. In both cases, an adaptive mesh was used, with the parameter $h_{\min}$ equal to $\varepsilon/10 = 0.001$. 205

5.2 CPU times (hh:mm:ss) using an Intel i7-3770 processor for the simulations presented in Section 5.4.4 (nucleation in a domain with complex boundaries), with or without time-step adaptation. The method OD2-W was used for the simulations with an adaptive time step, and the method OD1-W was used for the simulations with a fixed time step. In all cases, we used a fixed mesh with mesh size $h = 0.01$ (the size of the domain is roughly 5 by 5) and $P1$ elements, leading to 181587 unknowns. The parameter ε was set to 0.02. 212

6.1 Lines of code of *Hermipy*, obtained by using the `cloc` utility with the command `cloc $(git ls-files)` in the *Git* repository of the code. 220

6.2 Test coverage of the submodules comprised in *Hermipy*, obtained using the `coverage` utility, `coverage run -source=hermipy -m unittest discover -v -f tests`. 221

List of Figures

1.1 Numerical solutions obtained using the HMM (blue) and by direct numerical integration of the effective equation (1.29) (red), for different values of the precision p. 38

1.2 Error E_p (1.30) as a function of p. The slope of the blue line, obtained by linear fitting in the $p - \log_2(E_p)$ plane, is equal to -1.06, which is close to the theoretical value of -1 predicted in [65]. 38

1.3 Effective diffusion coefficient of the mean position of two particles interacting via the potential (1.37) (left), and of their difference (right), against the interaction strength. 42

1.4 Self-consistency map for two different values of the temperature, when $\theta = 1$ and V is the bistable potential. When the temperature decreases (β increases) enough, the steady state corresponding to $m = 0$ loses its stability and two new solutions of Eq. (1.40) emerge. 44

1.5 Bifurcation diagram of m against β when $\theta = 1$ (left), and stability of the $m = 0$ steady state in the β–θ plane (right). 44

1.6 Illustration of the effective diffusion by Monte–Carlo simulation of the GLE with Ornstein–Uhlenbeck noise; see Chapter 4 for more details on this model of the noise. We calculated the histograms and the approximation of the effective diffusion coefficient based on 10^5 trajectories, all started at $q = p = 0$. The exact effective diffusion coefficient in this case, calculated using an accurate spectral method, is equal to $D = 0.841$. The orange curve is the Green function associated with the one-dimensional heat equation with diffusivity D, i.e. $(4\,\pi\,D\,t)^{-1/2}\,\exp(-\frac{x^2}{4\,D\,t})$. It appears clearly from the local oscillations of the empirical density that the convergence of the rescaled position only holds in a weak sense, not pointwise or in $L^p(\mathbf{R})$. 48

1.7 Marginals of the steady state solution of the Langevin dynamics with forcing, Eq. (1.53a). The simulation were carried using *Hermipy*, the software library we developed for Fourier/Hermite spectral methods; see Chapter 6 and [216]. Here we used the potential $V(x) = (1/2)\,(1 - \cos(x))$ and the periodic cell $[-\pi, \pi]$. . . 50

1.8 Illustration of the dynamics Eq. (1.61), with in the background the level sets of the norms $|\mathbf{x}|^2$ (left) and $(\!(\mathbf{x}, \mathbf{x})\!)$ (right). 54

1.9 Bulk free energy (1.70) for different values of the absolute temperature (left) and Helmoltz double-well potential $\frac{1}{4}(\phi^2-1)^2$ (right). 60

1.10 Schematic of a droplet at equilibrium on an B-philic (therefore A-phobic) substrate. Here the equilibrium contact angle is equal to 70 degrees. 61

1.11 Solution of the Cahn–Hilliard equation at time $t = 1$ for different values of ε. Here we employed the initial condition $\phi_0(x) = -1 + 2\,H(r^2 - (x - x_1)^2 + y^2) + 2\,H(r^2 - (x - x_2)^2 + y^2)$, where $x_1 = -1/4$, $x_2 = 1/4$, $r = 1/4$ and H is the Heaviside function, as well as homogeneous Neumann boundary conditions for both the phase field and the chemical potential. 63

2.1 Error $E(d)$, see Eq. (2.42), for the fast-slow SDE (2.43). The blue dots were obtained by numerical experiments, and the straight line is the function $478.63 \times 10^{-0.45\,d}$ obtained by linear regression in the $d - \log_{10}(E(d))$ plane. A super-algebraic convergence is observed. 91

2.2 Relative error of the homogenized coefficients, $e(d, x)$, see Eq. (2.45), for the fast/slow SDE (2.44) at $x = (0.2, 0.2)$. In this case, as the points are more or less aligned, the convergence is also super-algebraic. The straight line, also obtained by linear regression in the $d - \log_{10}(E(d))$ plane, is the function $18.20 \times 10^{-0.31\,d}$. We notice that the convergence is slightly slower than for Eq. (2.43). 93

2.3 Error $E(d)$, see Eq. (2.42), for the fast/slow SDE (2.46). The straight line is the linear function $0.40 \times 10^{-0.19\,d}$, and we observe that the convergence is not as fast as in the unimodal cases. 93

2.4 Probability density $\mathrm{e}^{-V(\cdot)}/\mathcal{Z}$ associated to the potential (2.47). 94

2.5 Error $E(d)$, see Eq. (2.42), for the fast/slow system (2.48). The straight line is the linear function $1.17 \times 10^{-0.30\,d}$, so we see that the convergence is faster than for the bistable potential (Fig. 2.3). 95

2.6 Potential (2.49), used in equation (2.50). 96

2.7 Error $E(d)$, see Eq. (2.42), for the fast/slow system (2.50). 96

2.8 Evolution of the coefficient x_1 of the first term in the Fourier expansion (2.56) of the the solution to the SPDE (2.55), obtained numerically by the HMM (black) and the Hermite spectral method (red), for one sample of the driving Brownian motion. 100

2.9 Evolution of the coefficient x_2 of the second term in the Fourier expansion (2.56) of the the solution to the SPDE (2.55), obtained numerically by the HMM (black) and the Hermite spectral method (red), for one sample of the driving Brownian motion. 100

2.10 Error between the homogenized coefficients (see Eq. (2.59)) for the SPDE (2.55), as a function of the precision parameter p. The green line, obtained by linear fitting, has slope -1.01 in the $p - \log_2(E_p)$ plane, which is close to the theoretical value of -1, showing a perfect agreement with the theory. 101

3.1 Free energy (3.3) of the one-parameter family (3.6) of probability densities that solve Eq. (3.4a) for some value of m (in blue), and associated first moment $R(m)$ (in green), for fixed $\theta = 1$ and $\beta = 5$. Along the one-parameter family, $m = 0$ is a local maximum of the free energy, and it therefore corresponds to an unstable steady state of the McKean–Vlasov equation. 108

3.2 Truncated asymptotic expansion of the self-consistency map, $R_0 + \varepsilon^2 R_2$, as a function of m (red line) compared to $y = m$ (blue line) for the scalar Ornstein–Uhlenbeck noise, with $\beta = 10$, $\theta = 1$, $\varepsilon = 0.1$. 116

3.3 Study of the Galerkin approximation (3.30). 120

3.4 Snapshots of the solution to (3.29) using either the finite volume method from [37] or the Galerkin approximation (3.30) with the RK45 method and 25 Hermite functions. 121

3.5 Difference, in the $L^1(\mathbf{R})$ norm, between the numerical solutions ρ_c, ρ_i and ρ_e. . 121

3.6 Simulation data when $V(x) = x^2/2$. 124

3.7 Simulation data when $V(x) = x^4/4 - x^2/2$. 125

3.8 Probability density solution of Eq. (3.35) (obtained using the spectral method) at times $0, 0.2, 0.5, 1$. 126

3.9 Convergence of the Hermite spectral method for the nonlinear McKean–Vlasov equation with colored noise. The error decreases exponentially for low enough values of d, and then reaches a plateau when it becomes dominated by the error induced by the time discretization. 127

3.10 Comparison of the numerical solutions for the steady-state Fokker–Planck equation with the bistable noise model (model **B** in Section 3.2), using either a square index set or a triangle index set. While not obvious from the figures, in the former case it follows from the fact that $\varepsilon = 0.01 \ll 1$ that the columns of Hermite coefficients in the η direction are essentially colinear. 129

3.11 Comparison of the numerical solutions for the steady-state Fokker–Planck equation when the noise is confined by potential (3.11), with and without the corrective drift term. Without correction, the parasitic drift completely deteriorates the accuracy of the solution. 131

3.12 Normalized bias $\varepsilon \mu_d$ defined in (3.40), against degree of approximation, for $\sigma_\eta^2 = 0.1$. (The values shown need to be divided by ε to obtain the effective parasitic bias for a given value of ε.) 132

3.13 L^1 norm of the difference between the solutions found using the Hermite spectral method and the asymptotic expansion up to order ε^2(given in Eq. (3.69)), in both x and η, in the case of **OU** noise with $\varepsilon = 2^{-i}, i = 0, \ldots, 5$ (left), and the corresponding L^1 error of the asymptotic expansion (right). The scaling of the error (with respect to ε) is close to the expected value of 3. 132

3.14 Convergence as $\varepsilon \to 0$ in the case of harmonic noise. The observed rate of convergence of the x-marginal to the white noise limit is 3.93, which is close to the theoretical value of 4. Here $\rho^{x,q}$ denotes the marginal of the solution on the x, q plane. 133

3.15 Bifurcation diagram of m against β for Ornstein–Uhlenbeck noise, obtained via MC simulation, the spectral method, and the asymptotic expansion (3.20). . . . 135

3.16 Bifurcation diagram of m against β for harmonic noise (model **H**), obtained via MC simulation, the Hermite spectral method, and the asymptotic expansion (3.20). 135

3.17 Bifurcation diagram of m against β for the bistable noise (model **B**), using the spectral method and a truncated asymptotic expansion including the first nonzero correction. We see that, overall, the agreement between the two methods is excellent. 136

3.18 Left: bifurcation diagram of m against β for the nonsymmetric noise (model **NS**), using the spectral method and a truncated asymptotic expansion including the first nonzero correction. Right: $R_0 + \varepsilon R_1(m) - m$ against m for $\varepsilon = 0.1$ and $\beta = 2.6$. 136

3.19 Critical β against ε. 137

3.20 Convergence of the Galerkin discretization for the harmonic noise. The points correspond to the simulation data, and the lines are the best linear (in the *semilogy* scale) approximations. We observe that the accuracy achieved is better in the case of a quadratic potential. 149

3.21 Left and middle: steady-state solution to (3.32) when $V_{\text{eff}}(x) = x^4/4 - x^2/2$, $V_\eta(\eta) = \eta^4/4 - \eta^2/2$ and its x-marginal. Right: Convergence of the method when $d \to \infty$. 150

3.22 Steady-state solution of Eq. (3.32) and its marginals for the nonsymmetric noise (**NS**). These figures correspond to the parameters $\beta = 2$, $\varepsilon = 0.25$, $d = 100$, $\sigma_x^2 = 1/30$, $\sigma_y^2 = 1/10$. 151

3.23 Convergence of the steady-state solution of Eq. (3.32) when the driving noise is confined by the bistable potential $V_\eta(\eta) = \eta^4/4 - \eta^2/2$. Here $\rho^x(x)$ denotes the x marginal of $\rho(x, \eta)$. 152

4.1 Memory kernels associated with the models considered. As $\alpha \to \infty$, the memory kernel of the model GL1 is recovered. 159

4.2 *Left*: Some contour lines of the Hamiltonian and separatrix (in red). *Right*: Associated graph, over which the Hamiltonian performs a diffusion process in the underdamped limit. 162

4.3 Approximate solution to the Poisson equation (4.9), obtained by solving Eq. (4.39), when $\mathcal{L}$ is the generator of Langevin dynamics. 162

4.4 Error of the spectral method for the calculation of $I_\nu(E)$ (see Eq. (4.31)) when $\nu = 1$, estimated as the difference between the values computed with 50 and 100 Fourier modes. We observe that the error is largest close to $E = E_0 = 1$. For this value of ν, the error on $I_\nu(E)$ is uniformly bounded from above by 10^{-4}, and the difference between the approximations of the effective diffusion coefficient obtained with $N = 50$ and $N = 100$ is below 10^{-7}. 169

4.5 Comparison between the effective diffusion coefficients (multiplied by γ) for the Langevin and the generalized Langevin dynamics in the underdamped limit. . . 169

4.6 VAF and its spectrum for the Langevin and the GL1 dynamics, with the parameters $\gamma = 1$, $\beta = 2$, and $\nu \in \{1/4, 1/2, 3/4, 1\}$. In both figures, the curves corresponding to GLE with $\nu = 1/4$ are very close to those associated with the Langevin dynamics. 174

4.7 VAF and its spectrum for the Langevin, GL1 and GL2 dynamics, with the parameters $\gamma = 1$, $\beta = 2$, $\nu = 1/\sqrt{5}$, $\alpha = 1/\sqrt{5}$. 175

4.8 Spectrum of the VAF and comparison and with the results from [118, 119], which are displayed in the background of the figures, for the Langevin, GL1 and GL2 dynamics. Here $\omega_0 = \sqrt{2}/2$ is the natural frequency of oscillations of the Hamiltonian dynamics, and $C_{\hat{v}} = \sqrt{2}/(2\pi)^2$ is the scaling factor used in [118, 119] to obtain a dimensionless quantity. 175

4.9 Numerical error on the effective diffusion coefficient associated with Langevin dynamics, when an equal number of basis functions is used in each direction. Precise solutions, used for estimating the error, were calculated using $N = 250$, and for the scaling parameter $\sigma = \frac{1}{4}$ in the cases $\gamma = 0.001, 0.01$ and $\sigma = 1$ otherwise. Notice that the data points corresponding to $\gamma = 10, 100, 1000$ are not visible on the graph; this is because the corresponding errors are all below 10^{-10}. 176

4.10 Diffusion coefficient as a function of γ, for the parameters $\nu = 1$ (for the models GL1 and GL2) and $\alpha = 1$ (for GL2). We observe that, for values of γ in the range $[1, 10]$, the GL1 diffusion coefficient is slightly larger than $\overline{D}/\gamma$. 177

4.11 Effective diffusion coefficient against ν, whose square encodes the characteristic time of the autocorrelation function of the noise, for fixed values $\beta = \gamma = 1$. . . 178

4.12 Deviation of the effective diffusion coefficient from its limiting value as $\nu \to 0$, and comparison with the asymptotic expansion (4.38). For $\alpha = 1$, the straight line was fitted using a subset of the data points excluding the data corresponding to small ν (for which numerical errors dominate) and to large ν (for which the asymptotic expansion is no longer accurate). 178

4.13 Convergence rate of the dynamics for the three models when $\mathcal{X} = \mathbf{R}, V(q) = q^2/2$ (left) and $\mathcal{X} = \mathbf{T}, V(q) = (1/2)(1 - \cos(q))$ (right), obtained by numerical solution of the maximization problem (4.53). In both cases, $\nu = \alpha^{-1} = 1/3$. 183

4.14 *Left*: convergence to the effective diffusion coefficient of the one-dimensional case as $\delta \to 0$. *Right*: convergence to the overdamped limit as $\gamma \to \infty$. 185

5.1 Schematic of the profile geometry of a fluid-fluid interface intersecting a solid boundary and illustration of the stationary solution (5.7). 193

5.2 Equilibrium position of a droplet on substrates with different wetting properties. In all cases, θ denotes the angle appearing in boundary condition (5.16) and θ^* denotes the angle calculated numerically. Blue corresponds to $\phi = 1$ and green to $\phi = -1$. 200

5.3 Convergence of OD2-W with respect to the mesh size, without mesh adaptation. In this case, h corresponds to the uniform mesh size, given as input to the mesh generator of *gmsh*, and $h^* = 0.01$. 201

5.4 Convergence of the numerical method with respect to the time step, without mesh adaptation. In the case of OD1-W, the rate of convergence is close to the expected value of 1. In the other two cases, the rate of convergence is close to the expected value of 2. 202

5.5 Total numerical dissipation generated by the numerical schemes in the simulation used to produce Fig. 5.4. OD2-W is by far the scheme producing the least numerical dissipation, even for relatively large time steps. OD2mod-W, on the other hand, introduces significant numerical dissipation for large time steps, owing to the large value of β that was chosen for the simulation, but is much less dissipative than OD1-W for small time steps. 202

5.6 Phase field and chemical potential during the coalescence of two sessile droplets on a hydrophilic substrate for $\theta = \pi/4$. The snapshots correspond to iterations 500, 1000, 1500, 2000, 2500, and 3000, which correspond to times 2.26, 19.55, 34.13, 49.03, 83.20 and 273.64. Blue color represents phase $\phi = 1$ and green phase $\phi = -1$. 203

5.7 Phase field and chemical potential when the contact angle is set to $3\pi/4$. The snapshots correspond to iterations 500, 1000, 1500, which correspond to times 1.47, 17.32, and 115.71. 204

5.8 Simulation data for the numerical experiments presented in Section 5.4.3 (two droplets on a substrate), when using the adaptive time-stepping scheme 1. As expected, the total free energy decreases with time for all three values of the contact angle. In the case $\theta = \pi/4$, we note a peak in the mixing component of the free energy and a refinement of the time step at the coalescence time. See Table 5.1 for the associated computational costs; we notice that the simulations with an adaptive time step are about 3 times faster or more. 206

5.9 Evolution of the contact line (top) and the isosurface $\phi = 0$ (bottom) of the phase field, on a chemically heterogeneous substrate with a contact angle defined by (5.18). The balance of the wall component and the mixing component of the free energy determines the motion of the drop. The field represented in the top figures is the value of the contact angle at the triple line. In the bottom figures, the field represented on the plane representing the substrate is the imposed contact angle. Interestingly, the heterogeneities of the substrate cause the length of the contact line to increase. 207

5.10 Evolution of the isosurface $\phi = 0$ of the phase field when a constant flux is imposed at the bottom boundary; The pictures correspond to iterations 0, 200, 400, 800 and 1000. Note that, because of the neutral boundary condition imposed at the spheres, the isosurface tends to stay normal to them as long as they are not completely covered. 208

5.11 Evolution of $M(\phi)$ and $E_m(\phi)$ as a function of time in the case of a 3D flow through a porous medium (Figure 5.10). In this case, the mass increases linearly because we impose a constant mass inflow at the pore. The free energy increases as well, because the size of the interface grows, in agreement with both the mass and energy laws (5.4) and (5.6). 209

5.12 Evolution of the phase field and chemical potential for the nucleation in a domain with complex boundaries, when starting from a random distribution. As before, blue corresponds to $\phi = 1$ and green to $\phi = -1$. The contact angle imposed at the boundaries is $\theta = \frac{\pi}{4}$. 210

5.13 Comparison of the solutions at the final time, with (left) and without (right) time-step adaptation. At the initial time, the phase field is set to a random value at each grid point, drawn from a random normal distribution with variance 0.1. Although one could expect the solutions for $\theta = \pi/4$ and $\theta = 3\pi/4$ to differ only by a sign, this is not the case. There is also a significant difference between the solutions obtained with and without time-step adaptation. These differences can be explained by the sensitivity of the evolution to perturbations of the initial condition and to numerical errors in the early stage of the simulation. 210

5.14 Simulation data for the numerical experiments presented in Section 5.4.4 (nucleation in a geometry with complex boundaries). Overall, the time step increases steadily when the adaptive time-stepping scheme is used, which is consistent with the decreasing frequency of coalescence events. The time step is refined at times to ensure that the incremental decrease of free energy at each iteration is approximately constant. 211

Chapter 1

Introduction

1.1 Brief introduction to the thesis 26
1.2 Numerical solution of multiscale SDEs 27
1.2.1 Averaging for multiscale SDEs 27
1.2.2 Homogenization for multiscale SDEs 30
1.2.3 The heterogeneous multiscale method 33
1.3 The Desai–Zwanzig mean-field model with colored noise 39
1.3.1 The white noise problem 40
1.3.2 Colored noise problem 44
1.4 Diffusion of a Brownian particle in a periodic potential 45
1.4.1 Relationship between the GLE and Langevin dynamics 46
1.4.2 Calculation of the effective diffusion coefficient 47
1.4.3 Convergence to equilibrium 52
1.5 The Cahn–Hilliard equation 56
1.5.1 Formulation as a gradient flow 57
1.5.2 Models for the free energies and the mobility 59
1.5.3 The sharp interface limit 62
1.5.4 Coupling with the Navier–Stokes equation 63
1.6 Contributions of the thesis 65

This chapter is organized as follows. In Section 1.1 we present a brief introduction to the research problems addressed in this thesis. In Sections 1.2 to 1.5, we introduce in more details the chapters that constitute the bulk of this work, Chapters 2 to 5 respectively, and we present some background material on the corresponding topics. In Section 1.6, we present an outline of the rest of the thesis and we summarize the original contributions of each chapter.

1.1 Brief introduction to the thesis

In this thesis several topics related to multiscale modeling are investigated. Although each chapter addresses a different research question, the use of spectral methods and of multiscale techniques constitute unifying themes for Chapters 2 to 4. Chapter 5 is mostly independent from the rest of the thesis, and Chapter 6 focuses on software contributions.

In Chapter 2, we develop a numerical method for the solution of multiscale stochastic differential equations (SDEs) of the following type:

$$\begin{aligned} \mathrm{d}X_t^\varepsilon &= \frac{1}{\varepsilon}\mathbf{f}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t + \sqrt{2}\,\boldsymbol{\sigma}_x\,\mathrm{d}W_x(t), \\ \mathrm{d}Y_t^\varepsilon &= \frac{1}{\varepsilon^2}\mathbf{h}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t + \frac{\sqrt{2}}{\varepsilon}\boldsymbol{\sigma}_y\,\mathrm{d}W_y(t), \end{aligned}$$

where $X_t^\varepsilon \in \mathbf{R}^m$, $Y_t^\varepsilon \in \mathbf{R}^n$, ε is a small parameter encoding the scale separation between the processes X^ε and Y^ε, $\boldsymbol{\sigma}_x \in \mathbf{R}^{m\times d_1}$, $\boldsymbol{\sigma}_y \in \mathbf{R}^{n\times d_2}$ are constant matrices, and W_x, W_y are independent d_1 and d_2-dimensional Brownian motions, respectively. When $\varepsilon \ll 1$, traditional methods for the numerical solution of SDEs are not suitable for this problem, as a prohibitively small time step would be required to perform a stable and accurate integration in time. In contrast, the method we propose enjoys a computational cost independent of ε. It aims to capture the effective dynamics of the slow process X^ε in the limit $\varepsilon \to 0$, which under appropriate assumptions can be shown to be well-defined, and it is based on the numerical solution, using a Hermite spectral method, of an elliptic partial differential equation (PDE) involving the generator of the fast dynamics Y^ε.

In Chapter 3, we study the Desai–Zwanzig mean-field model with colored noise. More precisely, we consider the following system:

$$\mathrm{d}X_t^i = \left(-V'(X_t^i) - \theta\left(X_t^i - \frac{1}{N}\sum_{j=0}^{N} X_t^j\right)\right)\mathrm{d}t + \sqrt{2\beta^{-1}}\,\xi_t^i\,\mathrm{d}t, \quad 1 \le i \le N,$$

where N is the number of particles, $V(\cdot)$ is a confining potential, θ is an interaction strength, β is the inverse temperature of the system, and ξ_t^i are independent, identically distributed (i.i.d.) noise processes. When these processes are modeled by white noise and the confining potential has multiple minima, particle systems of this type are known to exhibit rich dynamics and, in particular, one or several phase transitions [213, 51, 94]. Our main aim is to study whether similar phenomena occur in the presence of colored noise, i.e. noise that is not delta-correlated in time. To this end, we develop a spectral method for the nonlinear and nonlocal Fokker–Planck equation that governs the evolution of the system in the mean-field limit.

In Chapter 4, we study the diffusion in a periodic potential $V(\cdot)$ of a Brownian particle governed by the *generalized Langevin equation* (GLE):

$$\ddot{q} = -V'(q) - \int_0^t \gamma(t-s)\,\dot{q}(s)\,\mathrm{d}s + F(t),$$

where q is the position of the particle, γ is a memory kernel, and F is a mean-zero stationary

Gaussian process. It is well-known that, in the limit $\varepsilon \to 0$, the rescaled particle position $q^\varepsilon := \varepsilon\, q(t/\varepsilon^2)$ converges weakly to a Brownian motion with a positive diffusion coefficient [103]. Our goal in this part is to study the dependence of this *effective* diffusion coefficient on the parameters of the problem, both theoretically and numerically via a mixed Fourier/Hermite spectral method extending the method developed in Chapter 2. We will also investigate the convergence to equilibrium of some Markovian approximations of the GLE, through hypocoercivity techniques applied to the associated backward Kolmogorov equation.

Finally, in Chapter 5, we present a novel numerical method for the Cahn–Hilliard equation, a fourth order parabolic PDE widely used to model phase separation and binary fluid flows. In order to model the interaction between a binary mixture and a substrate, we employ a nonlinear Neumann boundary condition for the phase field, which enables the specification of a contact angle at the boundary of the computational domain. Denoting the phase-field by ϕ and the chemical potential by μ, the system of equations that we consider is the following:

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= \nabla \cdot (b\, \nabla \mu), & \mu &= \frac{1}{\varepsilon} f_m(\phi) - \varepsilon\, \Delta \phi, & &\text{for } x \in \Omega,\ t \in (0, T], \\ \varepsilon \nabla \phi \cdot \mathbf{n} &= -f_w(\phi), & b\, \nabla \mu \cdot \mathbf{n} &= 0, & &\text{for } x \in \partial\Omega,\ t \in (0, T], \end{aligned}$$

where f_m and f_w are model-dependent functions determined by the internal and wall energies, respectively, b is a mobility parameter and ε is a small parameter encoding the thickness of the interface between the two phases. The symbols Ω, $\partial\Omega$ and $\mathbf{n}$ denote the computational domain, its boundary and the outward unit normal vector. The numerical method we present for this system of equations is linear and it satisfies discrete mass and energy laws consistent with those of the continuous equation.

1.2 Numerical solution of multiscale SDEs

As mentioned in Section 1.1, our aim in Chapter 2 will be the development of an efficient method for the solution of multiscale systems of SDEs of the type

$$\mathrm{d}X_t^\varepsilon = \frac{1}{\varepsilon} \mathbf{f}(X_t^\varepsilon, Y_t^\varepsilon)\, \mathrm{d}t + \sqrt{2}\, \boldsymbol{\sigma}_x\, \mathrm{d}W_x(t), \tag{1.2a}$$

$$\mathrm{d}Y_t^\varepsilon = \frac{1}{\varepsilon^2} \mathbf{h}(X_t^\varepsilon, Y_t^\varepsilon)\, \mathrm{d}t + \frac{\sqrt{2}}{\varepsilon} \boldsymbol{\sigma}_y\, \mathrm{d}W_y(t). \tag{1.2b}$$

In order to motivate the study of this equation, it is appropriate to first review the method of averaging for SDEs. We will then briefly review the method of homogenization for SDEs, which is relevant for studying the effective dynamics of the slow variables in Eq. (1.2a).

1.2.1 Averaging for multiscale SDEs

Our presentation of the method of averaging in this subsection, and of the method of homogenization in Section 1.2.2, is based on [175, 215]. We consider the following fast/slow system:

$$\mathrm{d}X_t^\varepsilon = \mathbf{f}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t, \qquad X_0^\varepsilon = x_0, \tag{1.3a}$$

$$\mathrm{d}Y_t^\varepsilon = \frac{1}{\varepsilon}\,\mathbf{h}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t + \sqrt{\frac{2}{\varepsilon}}\,\boldsymbol{\sigma}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}W(t), \qquad Y_0^\varepsilon = y_0, \tag{1.3b}$$

where we now assume that $X^\varepsilon(t) \in \mathcal{X}$, $Y^\varepsilon(t) \in \mathcal{Y}$, W is a standard Wiener process on $\mathcal{Y}$, and $\boldsymbol{\sigma}$ is a square matrix of appropriate dimension. The method of averaging provides an effective equation for the slow dynamics, X^ε, when the scale separation between X^ε and Y^ε is very large, that is when $\varepsilon \ll 1$. We begin by presenting the derivation of the effective equation formally by means of a perturbation expansion of the solution to the associated *backward Kolmogorov equation*, the parabolic PDE that governs the evolution of observables of the solution. For a smooth, compactly supported function ϕ on $\mathcal{X} \times \mathcal{Y}$, Proposition A.3 in the appendix stipulates that the evolution of $v^\varepsilon(x, y, t) := \mathbf{E}(\phi(X_t^\varepsilon, Y_t^\varepsilon) \,|\, X_0^\varepsilon = x,\, Y_0^\varepsilon = y)$, where the expectation is taken over the Brownian paths, satisfies the following parabolic PDE:

$$\frac{\partial v^\varepsilon}{\partial t} = \mathcal{L}\, v^\varepsilon \qquad \text{for } (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbf{R}_{>0}, \tag{1.4a}$$

$$v^\varepsilon(x, y, 0) = \phi(x, y) \qquad \text{for } (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{1.4b}$$

Here $\mathcal{L}$ is the generator of the diffusion process associated with Eq. (1.3):

$$\mathcal{L} = \frac{1}{\varepsilon}\mathcal{L}_0 + \mathcal{L}_1,$$

with

$$\mathcal{L}_0 = \mathbf{h}(x, y) \cdot \nabla_y + \mathbf{B}(x, y) : \nabla_y \nabla_y, \tag{1.5}$$

$$\mathcal{L}_1 = \mathbf{f}(x, y) \cdot \nabla_x, \tag{1.6}$$

where $\mathbf{B}(x, y) = \boldsymbol{\sigma}(x, y)\boldsymbol{\sigma}(x, y)^T$. To obtain an averaging result, it is necessary that the fast dynamics converge in some sense to an equilibrium when X_t^ε is considered a fixed parameter in Eq. (1.3b). It is thus natural to make an *ergodicity assumption*: denoting by $\mathcal{L}_0^*$ the formal $L^2(\mathcal{Y})$ adjoint of $\mathcal{L}_0$, we assume

$$\mathcal{N}(\mathcal{L}_0) = \operatorname{span}\{1\}, \tag{1.7a}$$

$$\mathcal{N}(\mathcal{L}_0^*) = \operatorname{span}\{\rho^\infty(\,\cdot\,; x)\}, \text{ with } \inf_{z \in \mathcal{Y}} \rho^\infty(z; x) > 0, \tag{1.7b}$$

where $\mathcal{N}(\cdot)$ denotes the null space and $\rho^\infty(\cdot; x)$ is a probability density on $\mathcal{Y}$. A detailed general discussion of ergodicity is beyond the scope of this work, but we note that Eqs. (1.7a) and (1.7b) are in fact equivalent when $\mathcal{L}$ satisfies the Fredholm alternative; see [175, Chapter 6] for more details and for the presentation of other equivalent conditions.

Let us apply a perturbation expansion of the solution to Eq. (1.4). If X^ε is to converge to the solution of an effective equation independent of Y^ε in the limit $\varepsilon \to 0$, then we expect the solution to (1.4) to converge to a solution independent of y in the same limit, provided

that the initial condition $\phi(x, y)$ in Eq. (1.4b) depends only on x. Writing $v^\varepsilon = v_0 + \varepsilon\, v_1 + \cdots$, substituting in (1.4), and gathering the terms multiplying an equal power of ε, we obtain:

$$\frac{1}{\varepsilon}\,\mathcal{L}_0\, v_0 + \left(\mathcal{L}_1\, v_0 + \mathcal{L}_0\, v_1 - \frac{\partial v_0}{\partial t}\right) + \mathcal{O}(\varepsilon^2) = 0 \qquad \text{for } (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbf{R}_{>0}, \tag{1.8}$$

$$v_0(x, y, 0) + \varepsilon\, v_1(x, y, 0) + \mathcal{O}(\varepsilon^2) = \phi(x) \qquad \text{for } (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{1.9}$$

For Eq. (1.8) to hold true for every value of ε, the following equations must be satisfied:

$$\mathcal{L}_0\, v_0 = 0, \tag{1.10a}$$

$$\mathcal{L}_1\, v_0 + \mathcal{L}_0\, v_1 - \frac{\partial v_0}{\partial t} = 0. \tag{1.10b}$$

Together with the ergodicity assumption (1.7a), Eq. (1.10a) imply that v_0 depends only on x and t. Based on Eq. (1.7b), we notice that

$$\int_{\mathcal{Y}} \mathcal{L}_0 f(y)\, \rho^\infty(y; x)\, \mathrm{d}y = \int_{\mathcal{Y}} f(y)\, \mathcal{L}_0^* \rho^\infty(y; x)\, \mathrm{d}y = 0,$$

for any regular enough function f, so a necessary condition for an equation of the type $-\mathcal{L}_0 f(y) = R(y)$ to admit a solution is that the right-hand side R satisfies a so-called *centering condition*:

$$\int_{\mathcal{Y}} R(y)\, \rho^\infty(y; x)\, \mathrm{d}y = 0.$$

When the operator $\mathcal{L}_0$ satisfies the Fredholm alternative, which will always be the case in this thesis, this condition is also sufficient to guarantee the existence of a weak solution; see, for example, [175, Theorem 7.9]. Equation (1.10b) therefore implies

$$\begin{aligned} &\int_{\mathcal{Y}} \left(\frac{\partial v_0}{\partial t} - \mathcal{L}_1 v_0\right) \rho^\infty(y; x)\, \mathrm{d}y = 0, \\ \Leftrightarrow\ &\frac{\partial v_0}{\partial t} = \int_{\mathcal{Y}} \left(\mathbf{f}(x, y) \cdot \nabla_x v_0(x, t)\right) \rho^\infty(y; x)\, \mathrm{d}y, \\ \Leftrightarrow\ &\frac{\partial v_0}{\partial t} = \left(\int_{\mathcal{Y}} \mathbf{f}(x, y)\, \rho^\infty(y; x)\, \mathrm{d}y\right) \cdot \nabla_x v_0. \end{aligned} \tag{1.11}$$

Equation (1.11) is the backward equation (see [175, Section 4.3]) associated with the ODE

$$X' = \mathbf{F}(X), \quad X(0) = x_0, \tag{1.12}$$

where the drift $\mathbf{F}(X)$ is defined as the average of $\mathbf{f}(x, y)$ with respect to the measure $\rho^\infty(y; x)\, \mathrm{d}y$:

$$\mathbf{F}(x) = \int_{\mathcal{Y}} \mathbf{f}(x, y)\, \rho^\infty(y; x)\, \mathrm{d}y. \tag{1.13}$$

Equation (1.12) is often referred to as the effective or averaged equation, and $\mathbf{F}(\cdot)$ as the effective drift coefficient. This concludes the formal derivation of the effective equation.

In order to give a flavor of some of the methods and obstacles often encountered in proofs related to multiscale analysis, we now briefly present how the convergence of X^ε to the solution

of the effective equation can be shown rigorously. The main idea is to consider the following *Poisson equation*:

$$\mathcal{L}_0 \mathbf{\Phi}(y;x) = \mathbf{f}(x,y) - \mathbf{F}(x), \qquad \int_{\mathcal{Y}} \mathbf{\Phi}(y;x)\, \rho^\infty(y;x)\, \mathrm{d}y = 0, \tag{1.14}$$

with appropriate (e.g. periodic, if $\mathcal{Y} = \mathbf{T}^n$) boundary conditions. Using this definition, we write

$$X^\varepsilon(t) - X(t) = \int_0^t (\mathbf{F}(X_s^\varepsilon) - \mathbf{F}(X_s))\, \mathrm{d}s + \int_0^t \mathcal{L}_0 \mathbf{\Phi}(X_s^\varepsilon, Y_s^\varepsilon)\, \mathrm{d}s. \tag{1.15}$$

Applying Itô's formula (Eq. (A.2) in the appendix) to $\mathbf{\Phi}(X_t^\varepsilon, Y_t^\varepsilon)$ leads to

$$\mathrm{d}\mathbf{\Phi}(X_t^\varepsilon, Y_t^\varepsilon) = \frac{1}{\varepsilon}\mathcal{L}_0 \mathbf{\Phi}(X_t^\varepsilon, Y_t^\varepsilon) + \mathcal{L}_1 \mathbf{\Phi}(X_t^\varepsilon, Y_t^\varepsilon) + \sqrt{\frac{2}{\varepsilon}}\, \nabla_y \mathbf{\Phi}(X_t^\varepsilon, Y_t^\varepsilon) \cdot \boldsymbol{\sigma}(X_t^\varepsilon, Y_t^\varepsilon)\, \mathrm{d}W_t.$$

Using this equation, Eq. (1.15) can be rewritten as

$$\begin{aligned} X^\varepsilon(t) - X(t) = \int_0^t (\mathbf{F}(X_s^\varepsilon) - \mathbf{F}(X_s))\, \mathrm{d}s + \varepsilon\, (\mathbf{\Phi}(X_t^\varepsilon, Y_t^\varepsilon) - \mathbf{\Phi}(x_0, y_0)) - \varepsilon \int_0^t \mathcal{L}_1 \mathbf{\Phi}(X_s^\varepsilon, Y_s^\varepsilon)\, \mathrm{d}t \\ + \sqrt{2\,\varepsilon} \int_0^t \nabla_y \mathbf{\Phi}(X_s^\varepsilon, Y_s^\varepsilon) \cdot \boldsymbol{\sigma}(X_s^\varepsilon, Y_s^\varepsilon)\, \mathrm{d}W_s. \end{aligned}$$

The rest of the proof is based on using estimates to bound the terms on the right-hand side, from where one can conclude by using the Grönwall inequality; see [175, Chapter 17]. Obtaining these estimates requires bounds and regularity results on the solution of the Poisson equation (1.14). In an unbounded domain, such bounds and results were obtained under rather general conditions in [170, 169]. The periodic case is considerably easier, especially when the leading-order generator $\mathcal{L}_0$ is elliptic, and it can be studied using standard tools from stochastic analysis and the theory of elliptic PDEs. In particular, considering the case $\mathcal{X} = \mathbf{T}^m$, $\mathcal{Y} = \mathbf{T}^n$ and assuming ellipticity

$$\langle \mathbf{B}(x,y)\, \xi, \xi \rangle > \bar{\beta}\, |\xi|^2 \quad \forall (x,y) \in \mathbf{T}^m \times \mathbf{T}^n, \tag{1.16}$$

it is possible to verify that the ergodicity assumptions (1.7a) and (1.7b) hold true, see Theorem A.5 in the appendix, and to show the following prototypical averaging result [175, Theorem 17.1], which characterizes the convergence of X^ε to X.

Theorem 1.1. *For any $p > 1$, the function X_t^ε solving (1.3) converges to the solution X_t of (1.12) in $L^p(\Omega, C([0,T]; \mathbf{T}^m))$: for any $T > 0$, there exists $C = C(T)$ such that*

$$\mathbf{E}\left(\sup_{0 \le t \le T} |X_t^\varepsilon - X_t|^p \right) \le C \varepsilon^{\frac{p}{2}}. \tag{1.17}$$

1.2.2 Homogenization for multiscale SDEs

When the effective drift $\mathbf{F}(X)$ defined in Eq. (1.13) vanishes for all values X, the method of averaging predicts only that $X_t^\varepsilon \approx x_0$ on a time scale $\mathcal{O}(1)$ when $\varepsilon \ll 1$. In this case, the drift

coefficient $\mathbf{f}(x,\cdot)$ is said to be *centered* with respect to the probability density $\rho^\infty(\,\cdot\,;x)$:

$$\int_{\mathcal{Y}} \mathbf{f}(x,y)\,\rho^\infty(y;x)\,\mathrm{d}y = 0.$$

To observe interesting dynamics, one must therefore look at the longer $\mathcal{O}(1/\varepsilon)$ time scale, which is often called the diffusive time scale because, as we will see, a diffusion term emerges in the corresponding effective equation for X^ε. Rescaling time in Eq. (1.3) with $t \to t/\varepsilon$ and using the scaling property of Brownian motion, that $W(c\,t) = \sqrt{c}\,W(t)$ in law, we obtain (keeping the same notations for simplicity)

$$\mathrm{d}X_t^\varepsilon = \frac{1}{\varepsilon}\,\mathbf{f}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t, \qquad X_0^\varepsilon = x_0, \tag{1.18a}$$

$$\mathrm{d}Y_t^\varepsilon = \frac{1}{\varepsilon^2}\,\mathbf{h}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t + \frac{\sqrt{2}}{\varepsilon}\,\boldsymbol{\sigma}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}W(t), \qquad Y_0^\varepsilon = y_0, \tag{1.18b}$$

which is of the same type as Eq. (1.2), except that here no noise term is present in the slow equations. The method of homogenization provides an effective equation for the slow dynamics X_t^ε in Eq. (1.18a). It can be applied to a class of SDEs that is in fact broader than Eq. (1.18), but this model equation will be sufficiently general for our purposes in this thesis. As in the previous subsection, the generator of multiscale system can be split into two parts, corresponding to the fast and slow dynamics respectively:

$$\mathcal{L} = \frac{1}{\varepsilon^2}\,\mathcal{L}_0 + \frac{1}{\varepsilon}\,\mathcal{L}_1,$$

where $\mathcal{L}_0$ and $\mathcal{L}_1$ are the same operators as in Eqs. (1.5) and (1.6). The effective equation can be derived formally in the same manner as in the previous section, under similar ergodicity assumptions; see [175, Chapter 11] and [215] for details. It reads

$$\mathrm{d}X_t = \mathbf{F}(X_t)\,\mathrm{d}t + \mathbf{A}(X_t)\,\mathrm{d}V_t, \qquad X(0) = x_0, \tag{1.19}$$

where V_t is a standard Brownian motion on $\mathbf{R}^m$. The effective drift coefficients, $\mathbf{F}$ and $\mathbf{A}$, are expressed in terms of the solution to a PDE involving the generator of the fast dynamics:

$$-\,\mathcal{L}_0\,\boldsymbol{\phi}(x,y) = \mathbf{f}(x,y), \qquad \int_{\mathcal{Y}} \boldsymbol{\phi}(x,y)\,\rho^\infty(y;x)\,\mathrm{d}y = 0. \tag{1.20}$$

This equation and the associated centering condition are sometimes referred to as the *cell problem* associated with the multiscale system. It can be seen as a PDE on $\mathcal{Y}$, where x plays the role of a fixed parameter. From the solution to Eq. (1.20), $\mathbf{F}(\,\cdot\,)$ and $\mathbf{A}(\,\cdot\,)$ can be expressed as follows:

$$\begin{aligned} \mathbf{F}(x) &= \int_{\mathcal{Y}} \mathbf{f}(x,y)\cdot\nabla_x\boldsymbol{\phi}(x,y)\,\rho^\infty(y;x)\,\mathrm{d}y, \\ \mathbf{A}(x)\,\mathbf{A}(x)^T &= \frac{1}{2}\left(\mathbf{A}_0(x) + \mathbf{A}_0(x)^T\right), \end{aligned} \tag{1.21}$$

where

$$\mathbf{A}_0(x) = 2\int_{\mathcal{Y}} \mathbf{f}(x,y)\otimes\boldsymbol{\phi}(x,y)\,\rho^\infty(y;x)\,\mathrm{d}y.$$

Remark 1.1. The matrix $\mathbf{A}(x)$ is well defined if and only if the right hand side of Eq. (1.21) is a positive semidefinite matrix. Taking into account that $\boldsymbol{\phi}(x,y)$ solves Eq. (1.20), we can write

$$\mathbf{A}_0(x) = -2\int_{\mathcal{Y}} \mathcal{L}_0\boldsymbol{\phi}(x,y) \otimes \boldsymbol{\phi}(x,y)\, \rho^\infty(y;x)\, \mathrm{d}y, \tag{1.22}$$

so, by linearity,

$$\boldsymbol{\xi}^T\mathbf{A}_0(x)\boldsymbol{\xi} = -2\int_{\mathcal{Y}} \mathcal{L}_0(\boldsymbol{\xi}^T\boldsymbol{\phi}(x,y))\ (\boldsymbol{\xi}^T\boldsymbol{\phi}(x,y))\ \rho^\infty(y;x)\, \mathrm{d}y,$$

for any vector $\boldsymbol{\xi} \in \mathbf{R}^m$, where m is the dimension of $\mathcal{X}$. It is thus sufficient that $\mathcal{L}_0$ is a negative operator on $L^2\left(\mathcal{Y}, \rho^\infty(\,\cdot\,;x)\right)$, which can be shown to always hold true in the periodic setting, see e.g. [175, Theorem 6.2]. This holds true also for general operators of the form $-A^\dagger A + B$, where the dagger † denotes adjunction in the weighted space $L^2\left(\mathcal{Y}, \rho^\infty(\,\cdot\,;x)\right)$ and B is an antisymmetric operator in that space; more on this later. ⊘

Remark 1.2. Let us also note that $\mathbf{A}(x)$ is not uniquely defined, and so the effective equation (1.19) is not unique either. This is consistent with the fact that the convergence of the solution of the original multiscale system to that of the effective equation holds in the sense of weak convergence of probability measures. ⊘

Proving the convergence of X_t^ε to X_t rigorously requires advanced techniques in all but the simplest cases. It was the subject of a series of studies by Pardoux and Veretennikov, who established the convergence in the unbounded setting in increasingly general situations: in [170] the authors address the case where the matrix $\mathbf{B}$ satisfies an ellipticity condition of the type (1.16) and the fast dynamics is independent of the slow processes, i.e. when $\mathbf{h}(X_t^\varepsilon, Y_t^\varepsilon) = \mathbf{h}(Y_t^\varepsilon)$ and $\boldsymbol{\sigma}(X_t^\varepsilon, Y_t^\varepsilon) = \boldsymbol{\sigma}(Y_t^\varepsilon)$; in [169] the assumption of independence of the fast dynamics on the slow processes is lifted; and in [171] the assumption of ellipticity is removed. A good reference proving a simple convergence result in the periodic setting is [175, Chapter 18].

Obtaining a rigorous convergence result relies crucially on bounds and regularity results on the solution of a Poisson equation, as in the previous subsection, and it also hinges on the martingale central limit theorem [175, Theorem 3.33]. Let us give a flavor of the proof in the simple setting where $\mathcal{X} = \mathbf{R}$, $\mathcal{Y} = \mathbf{T}^n$, $\mathbf{f}(X_t^\varepsilon, Y_t^\varepsilon) = f(Y_t^\varepsilon)$, $\boldsymbol{\sigma}(X_t^\varepsilon, Y_t^\varepsilon) = 1$ and $\mathbf{h}(X_t^\varepsilon, Y_t^\varepsilon) = -\nabla_y V(Y_t^\varepsilon)$ for some periodic potential V. In this case, the effective equation is purely diffusive with a diffusion coefficient independent of the slow dynamics, as will be the case in Chapter 4. The starting point of the proof is to use Itô's formula for $\phi(Y_t^\varepsilon)$, where ϕ denotes the solution of the cell problem (1.20),

$$\mathrm{d}\phi(Y_t^\varepsilon) = \frac{1}{\varepsilon^2}\underbrace{\mathcal{L}_0\phi(Y_t^\varepsilon)}_{=-f(Y_t^\varepsilon)} + \frac{1}{\varepsilon}\underbrace{\mathcal{L}_1\phi}_{=0}(Y_t^\varepsilon) + \frac{\sqrt{2}}{\varepsilon}\nabla_y\phi(Y_t^\varepsilon)\,\mathrm{d}W_t,$$

which leads to

$$X^\varepsilon(t) - x_0 = -\varepsilon\left(\phi(Y_t^\varepsilon) - \phi(y_0)\right) + \sqrt{2}\int_0^t \nabla_y\phi(Y_s^\varepsilon)\cdot \mathrm{d}W_s =: \varepsilon\,\theta(t) + M(t).$$

The idea then is to show that $(\varepsilon\,\theta, M) \Rightarrow (0, A\,B)$ in $C([0,T];\mathbf{R}^2)$ when $\varepsilon \to 0$, where '$\Rightarrow$' denotes the weak convergence of probability measures, B is a standard Brownian motion, and A is the (now scalar) diffusion coefficient given by Eq. (1.21). In the simple periodic setting under consideration in this paragraph, it is easy to show that ϕ is bounded, which immediately implies that $\theta \Rightarrow 0$. To study the martingale term, it is convenient to rewrite it in terms of $\varphi_t := Y^\varepsilon_{t/\varepsilon^2}$; by the scaling property of Brownian motion, it holds true in law that

$$M(t) = \sqrt{2}\,\varepsilon \int_0^{t/\varepsilon^2} \nabla_y \phi(\varphi_t) \cdot \mathrm{d}W_s = \sqrt{2}\,\varepsilon \sum_{j=1}^{n} \int_0^{t/\varepsilon^2} \frac{\partial \phi}{\partial y_j}(\varphi_t)\,\mathrm{d}W_s^j,$$

where W^j denotes the j-th component of the vectorial Brownian motion W. Using Itô's isometry, as well as the ergodicity of φ_t in order to replace the temporal average by a spatial one, we observe that, for $j = 1, 2, \ldots, n$,

$$\begin{aligned}\mathbf{E}\left[\left(\sqrt{2}\varepsilon \int_0^{t/\varepsilon^2} \frac{\partial \phi}{\partial y_j}(\varphi_t)\,\mathrm{d}W_t^i\right)^2\right] &= \mathbf{E}\left[2\varepsilon^2 \int_0^{t/\varepsilon^2} \left|\frac{\partial \phi}{\partial y_j}(\varphi_t)\right|^2 \mathrm{d}t\right] \\ &\to 2t \int_{\mathcal{Y}} \left|\frac{\partial \phi}{\partial y_j}(y)\right|^2 \rho^\infty(y)\,\mathrm{d}y \quad \text{as } \varepsilon \to 0.\end{aligned}$$

This suggests the convergence of $M(t)$ to a Brownian motion multiplied by the factor

$$\left(2 \int_{\mathcal{Y}} |\nabla \phi(y)|^2\, \rho^\infty(y)\,\mathrm{d}y\right)^{\frac{1}{2}}, \tag{1.23}$$

which is precisely the result obtained by direct application of the central limit theorem for martingales. That this coefficient (1.23) is indeed equal to A follows from (1.22) by integration by parts; taking into account that $\rho^\infty(y) = \frac{1}{\mathcal{Z}}\,\mathrm{e}^{-V(y)}$, with $\mathcal{Z}$ a normalization constant,

$$\begin{aligned}A^2 &= -2 \int_{\mathbf{T}^n} \big(-\nabla V(y) \cdot \nabla \phi(y) + \nabla\nabla : \phi(y)\big)\,\phi(y)\,\rho^\infty(y)\,\mathrm{d}y, \\ &= -2 \int_{\mathbf{T}^n} \mathrm{e}^{V(y)}\ \nabla \cdot \big(\mathrm{e}^{-V(y)}\ \nabla \phi(y)\big)\,\phi(y)\,\rho^\infty(y)\,\mathrm{d}y, \\ &= -\frac{2}{\mathcal{Z}} \int_{\mathbf{T}^n} \nabla \cdot \big(\mathrm{e}^{-V(y)}\ \nabla \phi(y)\big)\,\phi(y)\,\mathrm{d}y, \\ &= 2 \int_{\mathbf{T}^n} \nabla \phi(y) \cdot \nabla \phi(y)\,\rho^\infty(y)\,\mathrm{d}y.\end{aligned}$$

This concludes our discussion of the method of homogenization. We will now review the *heterogeneous multiscale method*, a commonly-used numerical method for the solution of multiscale stochastic equations of the type of Eq. (1.2), to which the spectral method we introduce in Chapter 2 is an alternative.

1.2.3 The heterogeneous multiscale method

Several methods exist for the numerical solutions of Eq. (1.2), some of them less general than others. In this subsection, we review only the *heterogeneous multiscale method* (HMM), which

will be used for comparison purposes later in this thesis. Some other methods, more specific to the context of statistical mechanics, will be reviewed in Section 1.4. In contrast with the Hermite spectral method we propose in Chapter 2, the HMM is based on Monte Carlo–type simulations. It was introduced and formalized in [218, 65, 3], and it relies on the link between the inverse of hypocoercive operators and the semigroup they generate.

Link between $\mathcal{L}^{-1}$ and $\mathrm{e}^{\mathcal{L}t}$

Consider a matrix $\mathbf{L} \in \mathbf{R}^{n\times n}$ with eigenvalues strictly in the left half-plane. Then

$$\int_0^\infty \mathrm{e}^{\mathbf{L}\,t}\,\mathbf{L}\,\mathrm{d}t = \int_0^\infty \mathbf{L}\,\mathrm{e}^{\mathbf{L}\,t}\,\mathrm{d}t = \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{e}^{\mathbf{L}\,t})\,\mathrm{d}t = -\mathbf{I},$$

where $\mathbf{I}$ is the identity matrix, because $\mathrm{e}^{\mathbf{L}\,t}$ converges to 0 as $t \to \infty$. This shows that

$$-\mathbf{L}^{-1} = \int_0^\infty \mathrm{e}^{\mathbf{L}t}\,\mathrm{d}t. \tag{1.24}$$

Remark 1.3. Equation (1.24) can also be derived by considering the time-dependent equation

$$\dot{\mathbf{x}}(t) = \mathbf{L}\mathbf{x}(t) + \mathbf{r}, \qquad \mathbf{x}(0) = \mathbf{x}_0, \tag{1.25}$$

where $\mathbf{r} \in \mathbf{R}^n$ is an arbitrary vector. Indeed, using the *variation-of-constants* formula, the solution reads

$$\begin{aligned}\mathbf{x}(t) &= \mathrm{e}^{\mathbf{L}t}\,\mathbf{x}_0 + \int_0^t \mathrm{e}^{\mathbf{L}(t-s)}\,\mathbf{r}\,\mathrm{d}s \\ &= \mathrm{e}^{\mathbf{L}t}\,\mathbf{x}_0 + \int_0^t \mathrm{e}^{\mathbf{L}s}\,\mathbf{r}\,\mathrm{d}s \quad \xrightarrow{t\to\infty} \quad \mathbf{x}_\infty := \int_0^\infty \mathrm{e}^{\mathbf{L}t}\,\mathbf{r}\,\mathrm{d}t,\end{aligned}$$

where we used the change of variable $s \mapsto t - s$ to pass from the first to the second line. Taking the limit as $t \to \infty$ in Eq. (1.25), we obtain

$$\mathbf{L}\mathbf{x}_\infty + \mathbf{r} = \lim_{t\to\infty} \dot{\mathbf{x}}(t) = \lim_{t\to\infty} \mathrm{e}^{\mathbf{L}t}(\mathbf{L}\mathbf{x}_0 + \mathbf{r}) = 0.$$

Since this equation is valid for all $\mathbf{r}$, we deduce Eq. (1.24). ⊘

We observe that the only assumption necessary to prove the identity (1.24) is that $\mathrm{e}^{\mathbf{L}t} \to 0$ as $t \to \infty$. It is therefore not surprising that this identity can be generalized to an infinite-dimensional operator $\mathcal{L}$ provided that $\mathrm{e}^{\mathcal{L}t} \to 0$ in the space of bounded operators on the Banach space on which $\mathrm{e}^{\mathcal{L}t}$ is defined. The following result holds:

Proposition 1.2. *Let E be a Banach space and assume that*

$$\|\,\mathrm{e}^{\mathcal{L}t}\,h\|_E \le C\,\mathrm{e}^{-\lambda t}\,\|h\|_E, \qquad \forall h \in E.$$

then

$$-\mathcal{L}^{-1} = \int_0^\infty \mathrm{e}^{t\mathcal{L}}\,\mathrm{d}t$$

on E and

$$\|\mathcal{L}^{-1}\|_{\mathcal{B}(E)} \leq \frac{C}{\lambda}, \quad \textit{where} \quad \|\mathcal{L}^{-1}\|_{\mathcal{B}(E)} = \sup_{h \in E \setminus \{0\}} \frac{\|\mathcal{L}^{-1} h\|_E}{\|h\|_E}.$$

Proof. This result is from [142, Proposition 2.1]. See also [175, Result 11.8]. □

Remark 1.4. When $\mathcal{L}$ is the generator of an SDE, the requirement that $\mathrm{e}^{\mathcal{L}t} \to 0$ is an ergodicity assumption. Indeed, denoting by Y_t the solution to the associated SDE, $\mathrm{e}^{\mathcal{L}t}\,\phi$ is the solution to the backward Kolmogorov equation with initial datum ϕ, which coincides with $\mathbf{E}(\phi(Y_t)|Y_0 = y)$; see Proposition A.3 in the appendix. ⊘

Using Proposition 1.2, we can obtain alternative expressions for the effective drift and diffusion coefficients that do not contain any explicit reference to the solution of the cell problem (1.20).

Alternative expressions of the effective coefficients

This section is loosely based on [175, Section 11.6], where the formal derivations presented in this section are rigorously justified. Considering the same setting as in Eq. (1.18), let us define, for $\xi \in \mathbf{R}^m$, the process $\varphi_\xi^t \in \mathbf{R}^n$ as the solution to:

$$\mathrm{d}\varphi_\xi^t(y) = \mathbf{h}\left(\xi, \varphi_\xi^t(y)\right) + \sqrt{2}\,\boldsymbol{\sigma}(\xi, \varphi_\xi^t(y))\,\mathrm{d}W_t, \qquad \varphi_\xi^0 = y.$$

Roughly speaking, $\varphi_\xi(y)$ corresponds to the rescaled fast dynamics with initial condition y when X_t^ε is taken as fixed parameter equal to ξ. With this notation, using Proposition 1.2 in the space of centered functions, the solution of the cell problem (1.20) can be expressed as

$$\boldsymbol{\phi}(x,y) = \int_0^\infty (e^{\mathcal{L}_0 t}\mathbf{f})(x,y)\,\mathrm{d}t = \int_0^\infty \mathbf{E}\left[\mathbf{f}(x, \varphi_x^t(y))\right]\,\mathrm{d}t.$$

Here the expectation is taken with respect to the Wiener measure. Using this representation formula together with ergodicity, the diffusion and drift coefficients can be expressed as

$$\begin{aligned}
\mathbf{A}_0(x) &= 2\int_{\mathcal{Y}} \mathbf{f}(x,y) \otimes \int_0^\infty \mathbf{E}\left[\mathbf{f}(x, \varphi_x^t(y))\right]\,\mathrm{d}t\,\rho^\infty(y;x)\,\mathrm{d}y, \\
&= 2\int_0^\infty \int_{\mathcal{Y}} \mathbf{E}\left[\mathbf{f}(x,y) \otimes \mathbf{f}(x, \varphi_x^t(y))\right]\,\rho^\infty(y;x)\,\mathrm{d}y\,\mathrm{d}t, \qquad (1.26a) \\
&= 2\int_0^\infty \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{f}(x, \varphi_x^s(y)) \otimes \mathbf{f}(x, \varphi_x^{t+s}(y))\,\mathrm{d}s\,\mathrm{d}t. \qquad (1.26b)
\end{aligned}$$

Equation (1.26a) is an example of a *Green–Kubo* formula: it relates an effective transport coefficient to the integral of an appropriate autocorrelation function. We will revisit the Green–Kubo formula in Section 1.4.2 in the context of the diffusion of a Brownian particle in a periodic potential. The same reasoning can be applied to derive an alternative expression of the effective drift coefficient provided that the fast dynamics is independent of the slow one, which we assume from now on. Denoting $\varphi_\xi^t(y)$ by just $\varphi^t(y)$,

$$\mathbf{F}(x) = \int_0^\infty \left(\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{f}(x, \varphi^s(y)) \cdot \nabla_x \mathbf{f}(x, \varphi^{t+s}(y))\,\mathrm{d}s\right)\mathrm{d}t. \qquad (1.27)$$

How the method works

As discussed above, classical methods for the integration of SDEs are inefficient for multiscale systems. To accurately capture the invariant measure of the fast dynamics, both explicit and implicit methods require a time step that scales as $\mathcal{O}(\varepsilon^2)$, rendering the computational cost of carrying out the numerical integration prohibitively expensive when $\varepsilon \ll 1$. The HMM is based on approximating the effective transport coefficients at discrete times by truncation and discretization of the integrals in Eqs. (1.26b) and (1.27). More precisely, the algorithm comprises three building blocks:

1. The *macro-solver* is the numerical method used to integrate the effective equation, the coefficients of which are approximated using the *estimator*.

2. The *micro-solver* is the numerical method employed to integrate one or several replicas of the fast dynamics over a short, $\mathcal{O}(\varepsilon^2)$ time interval at each macro time step, treating the slow variables as constant parameters. It may for example be based on the well-known Euler–Maruyama scheme:

$$Y_{n,m+1,k} = Y_{n,m,k} + \left(\frac{\delta t}{\varepsilon^2}\right) \mathbf{h}(Y_{n,m,k}) + \sqrt{\frac{2\,\delta t}{\varepsilon^2}}\, \boldsymbol{\sigma}(Y_{n,m,k})\, \xi_{n,m,k}. \tag{1.28}$$

   Here n and m are the indices of the macroscopic and microscopic steps, respectively, and k is the index of the replica. The random increments $\xi_{n,m,k}$ are vectors of random variables independently drawn from a standard normal distribution. Each replica of the discretized fast processes in Eq. (1.28) must be supplemented by an initial condition at each macroscopic time step. For the first step, a possible choice is to take $Y_{0,0,k} = y_0$ for all replicas, where y_0 is the initial condition in Eq. (1.18b). For the subsequent macroscopic time steps, it is efficient [65] to initialize the fast processes at the last value of the previous macroscopic step, i.e. to set $Y_{n+1,0,k} = Y_{n,n_T+M+M'-1,k}$, where $n_T + M + M' - 1$ is the number of microscopic time steps per macroscopic step. The meaning of the symbols n_T, M and M' is explained in the next item.

3. The *estimator* provides approximations of the effective transport coefficients based on the solution calculated by the *micro-solver*:

$$\hat{\mathbf{F}}(X_n) = \frac{(\delta_t/\varepsilon^2)}{M\,K} \sum_{k=1}^{K} \sum_{m=n_T}^{n_T+M-1} \sum_{m'=0}^{M'} \mathbf{f}(X_n, Y_{n,m,k}) \cdot \nabla_x \mathbf{f}(X_n, Y_{n,m+m',k}),$$

$$\hat{\mathbf{A}}_0(X_n) = \frac{2(\delta_t/\varepsilon^2)}{M\,K} \sum_{k=1}^{K} \sum_{m=n_T}^{n_T+M-1} \sum_{m'=0}^{M'} \mathbf{f}(X_n, Y_{n,m,k}) \otimes \mathbf{f}(X_n, Y_{n,m+m',k}).$$

   Here X_n is the approximation of the slow variables at step n, K is the number of replicas used to calculate the ensemble averages, M and M' correspond to the upper bounds of the integrals in Eqs. (1.26b) and (1.27) after truncation, and n_T is a parameter used to reduce the bias of the ergodic average introduced by the initial condition.

The weak convergence of the method was proved rigorously in [65] under some assumptions on the coefficients $\mathbf{f}$, $\mathbf{h}$ and $\boldsymbol{\sigma}$ of the multiscale SDE, Eq. (1.18). The HMM has many more parameters than traditional numerical methods for SDEs, and it is important to choose these parameters carefully in order to achieve a good accuracy. A detailed analysis of the efficiency of the algorithm was presented in [65], where the authors obtained the optimal scaling of the parameters, with respect to a *precision parameter* p, in order to guarantee that the error decreases as $\mathcal{O}(2^{-p})$. When using the Euler–Maruyama scheme in both the micro and macro solvers, for example, they showed that the parameters $(\delta t/\varepsilon^2, n_T, K, M, M')$ leading to the minimal computational cost (asymptotically as $p \to \infty$) scale with p as $(2^{-p}, 1, 1, 2^{3p}, 2^p p)$.

A numerical example

Before concluding this section, we present an example of a multiscale system for which the associated effective equation can be calculated explicitly, as well as some numerical results obtained in previous work [215]. We consider the fast/slow system

$$\begin{cases} \mathrm{d}x = \dfrac{1}{\varepsilon}(x-2)\, y\, z\, \mathrm{d}t, & x(0) = x_0, \\ \mathrm{d}y = -\dfrac{1}{\varepsilon^2}\, y\, \mathrm{d}t + \dfrac{1}{\varepsilon}\, x\, z\, \mathrm{d}t + \dfrac{1}{\varepsilon}\, \mathrm{d}W_1, & y(0) = y_0, \\ \mathrm{d}z = -\dfrac{2}{\varepsilon^2}\, z\, \mathrm{d}t + \dfrac{1}{\varepsilon}\, x\, y\, \mathrm{d}t + \dfrac{1}{\varepsilon}\, \mathrm{d}W_2, & z(0) = z_0, \end{cases}$$

where W_1 and W_2 are independent standard Wiener processes. This system is of a slightly more general type than the prototypical multiscale SDE (1.18) we considered in Section 1.2.2, but it can be studied using very similar methods. The effective equation for the slow process is

$$\mathrm{d}X = \left(\frac{X(X-2)}{4} + \frac{X-2}{24} \right) \mathrm{d}t + \frac{|X-2|}{2\sqrt{3}}\, \mathrm{d}W, \tag{1.29}$$

where W is a standard Brownian motion. For the numerical solution, we considered one realization of the Brownian motion and we approximated the effective dynamics by both the HMM and direct numerical integration of Eq. (1.29). The Euler–Maruyama scheme was used with the same (macroscopic) time step in both cases, and the parameters used for the *micro-solver* and *estimator* were $(\delta t/\varepsilon^2, n_T, M, N, N') = (2^{-p}, 16, 1, 10\times 2^{3p}, 2^p p)$. In Fig. 1.1, the two paths are compared for various values of p. We observe that the solutions become almost indistinguishable from $p = 5$.

In Fig. 1.2, a plot of the error in the effective transport coefficients as a function of the precision parameter p is presented. The formula used as a measure of the error was

$$E_p = \frac{\Delta t}{T_0} \left(\sum_{n \le T_0/\Delta t} |\mathbf{F}(\hat{X}_n) - \hat{\mathbf{F}}(\hat{X}_n)| + |\mathbf{A}(\hat{X}_n) - \hat{\mathbf{A}}(\hat{X}_n)| \right), \tag{1.30}$$

where $\hat{X}_n$ is the approximation of the slow dynamics obtained by the HMM.

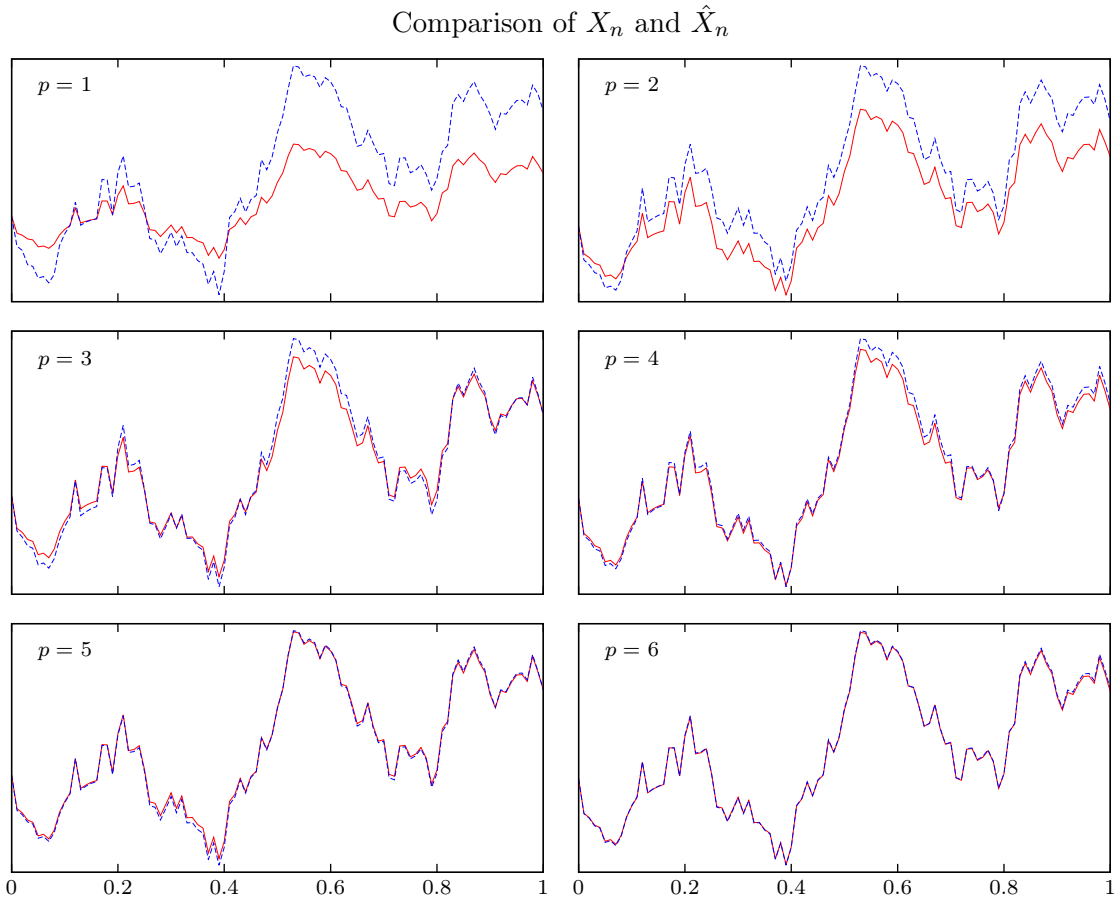


Figure 1.1: Numerical solutions obtained using the HMM (blue) and by direct numerical integration of the effective equation (1.29) (red), for different values of the precision p.

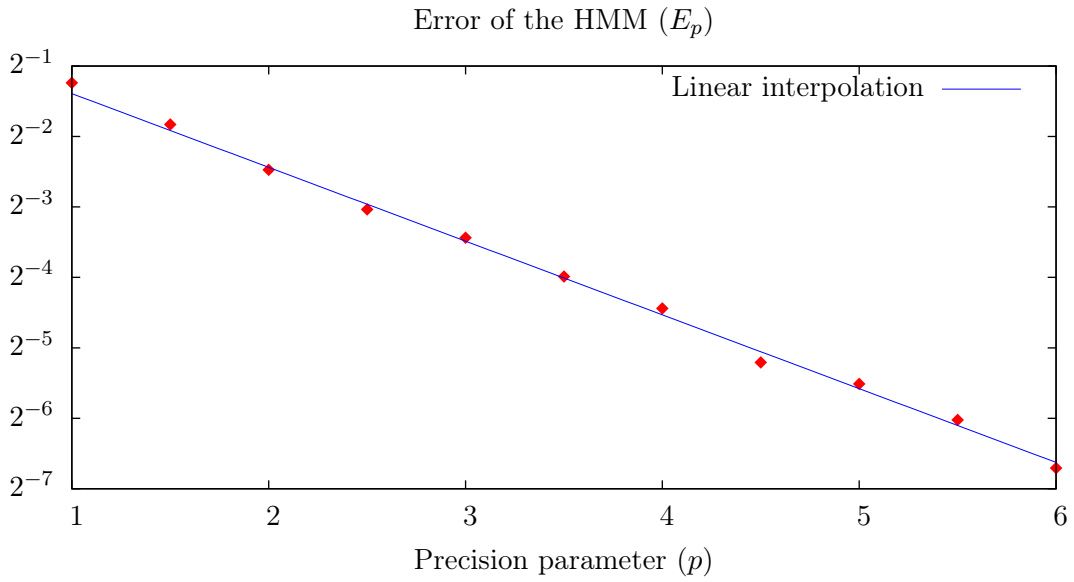


Figure 1.2: Error E_p (1.30) as a function of p. The slope of the blue line, obtained by linear fitting in the $p - \log_2(E_p)$ plane, is equal to -1.06, which is close to the theoretical value of -1 predicted in [65].

1.3 The Desai–Zwanzig mean-field model with colored noise

In Chapter 3, we will study systems of interacting particles subject to thermal noise. In contrast with most studies on the subject, this noise will be modeled by *colored noise*, i.e. noise with a nonzero correlation time. Specifically, we will consider that the noise is described by a stochastic differential equation (SDE) such as the *Ornstein–Uhlenbeck* or the *Langevin* equation. In this section, we introduce the subject matter and we present some background material on systems of interacting particles, their mean-field limit, and colored noise. In doing so, we motivate the appeal of using a spectral method, as opposed to generally more structure-preserving finite volume schemes, as a means of investigation of the nonlinear Fokker–Planck equation that arises in the mean-field limit.

Confining ourselves to the one-dimensional setting for simplicity, we consider systems of interacting particles of the following form, where the force on the particles derives from a potential and the interaction term depends only on the distance between the particles:

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -V'(X_t^i) - \frac{\theta}{N}\sum_{j=1}^{N} W'(X_t^i - X_t^j) + \sqrt{2\beta^{-1}}\,\xi_t^i, \quad i = 1, \dots, N. \tag{1.31}$$

Here $X_t^i \in \mathbf{R}$, $i = 1, \dots, N$ are the positions of the particles at time t, θ is an interaction strength, $V(\cdot)$ is an external potential, $W(\cdot)$ is an even interaction potential, and ξ^i are independent, identically distributed stationary noise processes. Rewriting Eq. (1.31) in vector form with the notations $\mathbf{X} := (X^1, \dots, X^N)$ and $\boldsymbol{\xi} := (\xi^1, \dots, \xi^N)$, one obtains

$$\frac{\mathrm{d}\mathbf{X}_t}{\mathrm{d}t} = -\nabla U(\mathbf{X}) + \sqrt{2\beta^{-1}}\,\boldsymbol{\xi}_t, \quad i = 1, \dots, N, \tag{1.32}$$

where $U(\mathbf{X})$ is defined by

$$U(\mathbf{X}_t) = \sum_{i=1}^{N} V(X_t^i) + \frac{\theta}{2N}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} W(X_t^i - X_t^j).$$

When the external potential V is of the form $V(x) = a\,\frac{x^4}{4} - b\,\frac{x^2}{2}$, for positive constants a and b, and the interaction potential W is the quadratic monomial $W(x) = \frac{x^2}{2}$ (also known as the *Curie–Weiss* interaction potential), Eq. (1.31) is the so-called *Desai–Zwanzig model*. In Chapter 3 we always take $a = b = 1$, in which case V is the usual bistable potential. Developing the quadratic form of W in Eq. (1.31), we notice that the particles interact only via their mean

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -V'(X_t^i) - \theta\left(X_t^i - \frac{1}{N}\sum_{j=1}^{N} X_t^j\right) + \sqrt{2\beta^{-1}}\,\xi_t^i, \quad i = 1, \dots, N. \tag{1.33}$$

If the processes ξ_t^i are white noise processes, the invariant measure of the system is well-known and the passage to the *mean-field limit*, $N \to \infty$, is well established [213, 158]. In this case, both the finite-dimensional Fokker–Planck equation associated with Eq. (1.33) and the nonlinear Fokker–Planck equation obtained in the mean-field limit can be viewed as gradient flows, which

is the subject of the next subsection.

1.3.1 The white noise problem

When ξ_i, $i = 1, \dots, N$, are white-noise processes, Eq. (1.32) can be viewed as an *overdamped Langevin equation* in a high dimension and the unique invariant distribution of the dynamics is the well-known Gibbs measure $\mu(\mathrm{d}\mathbf{X}) = \frac{1}{\mathcal{Z}} \mathrm{e}^{-\beta U(\mathbf{X})} \mathrm{d}\mathbf{X}$, where $\mathcal{Z}$ is a normalization constant. The associated Fokker–Planck equation reads

$$\begin{aligned} \frac{\partial \rho}{\partial t}(x,t) &= \nabla \cdot \big(\nabla U(x)\, \rho(x,t) + \beta^{-1} \nabla \rho(x,t) \big) \\ &= \nabla \cdot \Big(\rho(x,t) \nabla \Big(U(x) + \beta^{-1} \log\big(\rho(x,t)\big) \Big) \Big) \\ &= \nabla \cdot \left(\rho(x,t) \nabla \left(\frac{\delta E}{\delta \rho}(x,t) \right) \right), \end{aligned} \tag{1.34}$$

where $\frac{\delta E}{\delta \rho}$ denotes the functional derivative of the standard *free energy functional* associated with the particle system:

$$E(f) = \int_{\mathbf{R}^N} U(x)\, f(x) + \beta^{-1}\, f(x)\, \log\big(f(x)\big)\, \mathrm{d}x.$$

The first term in this expression is often referred to as the *internal energy*, whereas the negative of the integral $\int_{\mathbf{R}^N} f(x) \log\big(f(x)\big)\, \mathrm{d}x$ appearing in the second term is known as the *Gibbs–Boltzmann entropy*. For a detailed introduction to free energy Fokker–Planck equations and their structure, see [75, Chapter 4]. When the Fokker–Planck equation is written as (1.34), it appears clearly, by integration by parts and assuming sufficiently fast decay of $\frac{\delta E}{\delta \rho}(x,t)$ as $|x| \to \infty$, that $E(\rho)$ is a Lyapunov functional of the evolution:

$$\frac{\mathrm{d}}{\mathrm{d}t} E(\rho) = - \int_{\mathbf{R}^N} \left| \nabla \frac{\delta E}{\delta \rho}(x,t) \right|^2 \rho(x,t)\, \mathrm{d}x \leq 0. \tag{1.35}$$

This result is known as an H-theorem for the Fokker–Planck equation (1.34). In fact, it is possible to show that Eq. (1.34) is a gradient flow of $E(\rho)$ for the quadratic *Wasserstein metric* on probability measures, a fundamental result established by Jordan, Kinderlehrer, and Otto in [124]. When a result such as (1.35) holds, it is often desirable for numerical schemes to ensure a similar property at the discrete level, as estimates of this type are often instrumental in proving the associated stability and convergence results.

We now examine the strong interaction limit for the finite-dimensional particle system, which provides an intuitive justification for the existence of multiple steady states of the corresponding mean-field McKean–Vlasov equation.

Limit of strong interaction

In the limit of very strong interactions $\theta \gg 1$, the particles tend to cluster in a very small space; it holds approximately that $X_1 \approx \cdots \approx X_N$. In this regime, an approximate equation for the

mean position $\bar{X}_t := \frac{1}{N}\sum_{i=1}^N X_t^i$ can be obtained by summing Eq. (1.31) for $i = 1, \dots, N$:

$$\begin{aligned}\bar{X}_t' &= -\frac{1}{N}\sum_{i=1}^N V'(X_t^i) + \sqrt{2\beta^{-1}}\,\frac{1}{N}\sum_{i=1}^N \xi_t^i, \\ &\approx -V'(\bar{X}_t) + \sqrt{2\beta^{-1}}\,\frac{1}{N}\sum_{i=1}^N \xi_t^i = -V'(\bar{X}) + \sqrt{\frac{2\beta^{-1}}{N}}\,\chi_t,\end{aligned} \tag{1.36}$$

where the scaling property of Brownian motion was used and and a new white noise process χ_t was introduced. Equation (1.36) is the one-dimensional overdamped Langevin equation for the confining potential $V(\cdot)$, and it is well-known (see for example [174, Chapter 7]) that its solutions exhibit a metastable behavior when the diffusion coefficient, i.e. the factor multiplying the white noise, is low enough: they remain close to the minima of the potential for most of the evolution, and only occasionally jump from one well to the other. Denoting by ΔV the height of the potential barrier and by σ the diffusion coefficient of the SDE, the average time between jumps is known to scale as $\mathrm{e}^{2\Delta V/\sigma^2}$ when $\sigma \to 0$, i.e. as $\tau_N := \mathrm{e}^{N\beta\Delta V}$ in the case of Eq. (1.36). Since $\tau_N \to \infty$ as $N \to \infty$, it is not surprising that the mean-field equation should support two (or more, if V admits more than two local minima) stable steady states in the strong interaction regime.

Remark 1.5. That the mean particle position exhibits fewer jumps between the wells of the potential in the strong interaction regime can also be seen by studying its effective diffusion coefficient in a periodic potential. To this end, we consider that $V(x) = \frac{1}{2}(1 - \cos(x))$ and we approximate the interaction potential by a trigonometric function, in order to ease the numerical computation of the effective diffusion coefficient:

$$W(x) = 1 - \cos(x) = \frac{x^2}{2} + \mathcal{O}(x^4). \tag{1.37}$$

The effective diffusion coefficient of the mean particle position of the two-particle system ($N = 2$) governed by Eq. (1.31), with V and W as defined above, is illustrated in Fig. 1.3. We observe that it is significantly lower than that of a single particle for all the values of θ that we considered, but we caution that the periodic approximation (1.37) of the quadratic interaction potential leads to accurate results only in the strong interaction regime. Here a spectral method of the same type as the one presented in Chapter 4 was used for the numerical calculations. For more information on the link between the effective diffusion coefficient and the hopping rate between wells, see [178]. ⊘

Propagation of chaos and mean-field limit

Our presentation in this paragraph is based on [213] and [207]. The idea behind *propagation of chaos*, a terminology introduced by Kac, is the following. Let us assume that Eq. (1.34) is supplemented by a *chaotic initial condition*, $\rho(x, 0) = \prod_{i=1}^N \rho_0(x_i)$, i.e. that the initial particle positions are independently drawn from a given distribution ρ_0. In the presence of an interaction term, the particle positions will not remain independent for positive times, and indeed the Gibbs measure of the system cannot be expressed as tensor product of one-dimensional functions. It can

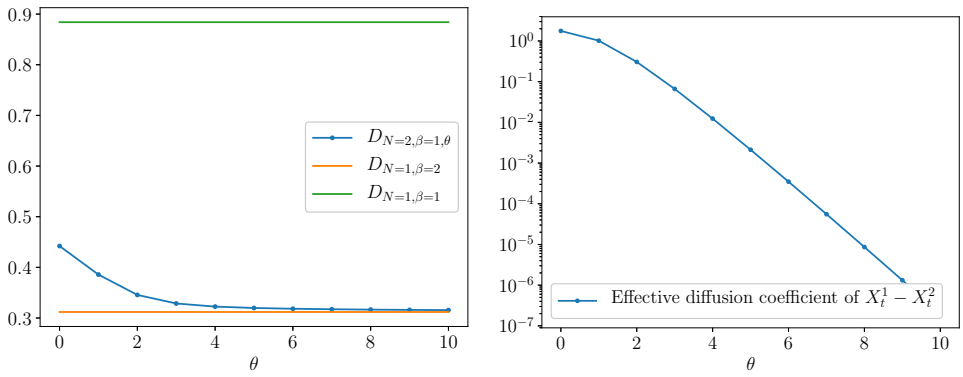


Figure 1.3: Effective diffusion coefficient of the mean position of two particles interacting via the potential (1.37) (left), and of their difference (right), against the interaction strength.

be shown, however, that the positions of a fixed subset of particles $X_t^1, \dots, X_t^k$ are asymptotically independent when $N \to \infty$; their distribution converges to $\prod_{i=1}^k u(x_i, t)$, where u is the solution to a one-dimensional nonlinear Fokker–Planck equation, as we shall see below.

Let us denote by $\mu_t^N := \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}$ the *empirical measure* of the particle system (1.31). In order to formally derive the mean-field equation, we employ Itô's formula, leading to

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{E}\left[\int_{\mathbf{R}} f\,\mathrm{d}\mu_t^N\right] = \mathbf{E}\left[-\int_{\mathbf{R}} f'\,V'\,\mathrm{d}\mu_t^N - \theta\int_{\mathbf{R}} f'\,\left(W'*\mu_t^N\right)\,\mathrm{d}\mu_t^N + \beta^{-1}\int_{\mathbf{R}} f''\,\mathrm{d}\mu_t^N\right],$$

for all smooth compactly supported function f, where $*$ denotes the convolution:

$$(W'*\mu_t^N)(x) := \int_{\mathbf{R}} W'(x-y)\,\mu_t^N(\mathrm{d}y) \quad \forall x \in \mathbf{R}.$$

If the assumption of propagation of chaos holds, we expect μ_t^N to converge, as $N \to \infty$ and in the sense of weak convergence of probability measures, to some deterministic measure μ_t satisfying the following equation in the distributional sense:

$$\frac{\partial \mu_t}{\partial t} = \frac{\partial}{\partial x}\left(V'\mu_t + \theta\,\left(W'*\mu_t\right)\,\mu_t + \beta^{-1}\frac{\partial \mu_t}{\partial x}\right), \qquad \mu_0 = \rho_0. \tag{1.38}$$

The equivalence between the convergence of μ_t^N to a deterministic measure μ_t and propagation of chaos as described above is proved in [207, Proposition 2.2]. Equation (1.38) is known as the *McKean–Vlasov* equation; it is the Fokker–Planck equation corresponding to the *McKean SDE*:

$$\begin{cases} \mathrm{d}X_t = -V'(X_t)\,\mathrm{d}t - \theta(W'*\mu_t)(X_t)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W_t, \qquad X_0 \sim \rho_0, \\ \mu_t = \mathrm{Law}(X_t). \end{cases}$$

The McKean SDE can also be obtained directly (formally) by using the law of large numbers in Eq. (1.31). Like the Fokker–Planck equation (1.34) for the finite-dimensional system, Eq. (1.38) can be shown to be a gradient flow for the Wasserstein metric of the free energy

$$E(f) = \int_{\mathbf{R}} V(x)\,f(x) + \frac{\theta}{2}(W*f)(x)\,f(x) + \beta^{-1}\,f(x)\,\log\big(f(x)\big)\,\mathrm{d}x,$$

where the second term corresponds to the interaction energy. The convergence to the mean-field limit was proved rigorously in [63] for non-convex potentials $V(\cdot)$, under relatively weak assumptions. There is extensive literature on the McKean–Vlasov equation, see for example [51, 213, 158]. Recently, a positivity preserving, energy decreasing finite-volume scheme was proposed in [37] for a large class of nonlinear nonlocal equations with a gradient flow structure, which was employed for the study of the McKean–Vlasov equation in [94].

Steady states and self-consistency equation

First notice that, using the property $(W * \mu_\infty)' = W' * \mu_\infty$ of the convolution, a probability density μ_∞ is a smooth steady state of Eq. (1.38) (denoting both the measure and its density by the same symbol, by a slight abuse of notation) if and only if

$$\mu_\infty(x) = \frac{\mathrm{e}^{-\beta(V(x)+\theta\, W*\mu_\infty(x))}}{\int_{\mathbf{R}} \mathrm{e}^{-\beta(V(x)+\theta\, W*\mu_\infty(x))}\,\mathrm{d}x}. \tag{1.39}$$

Solving this equation explicitly is in general not possible, but the situation simplifies when W is a polynomial because in this case $W * \mu_\infty$ depends on only a finite number of moments of μ_∞. For the Curie–Weiss interaction potential of interest in Chapter 3, for example,

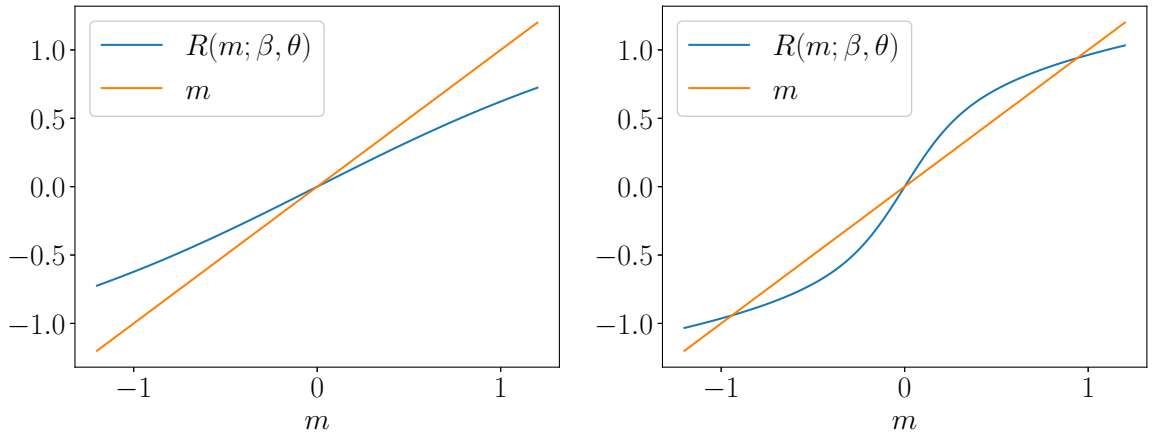
$$W * \mu_\infty = \frac{1}{2}\int_{\mathbf{R}} (x-y)^2\,\mu_\infty(\mathrm{d}y) = \frac{1}{2}(x^2 - 2\,m\,x + m_2),$$

where m and m_2 are the first and second moments of μ_∞, respectively. Multiplying Eq. (1.39) by x and integrating both sides, an equation for the first moment is obtained:

$$m = \frac{\int_{\mathbf{R}} x\ \mathrm{e}^{-\beta\left(V(x)+\frac{\theta}{2}(x^2-m)^2\right)}\,\mathrm{d}x}{\int_{\mathbf{R}} \mathrm{e}^{-\beta\left(V(x)+\frac{\theta}{2}(x^2-m)^2\right)}\,\mathrm{d}x} =: R(m;\beta,\theta). \tag{1.40}$$

Unlike Eq. (1.39), this equation can be easily solved numerically, for example by using a fixed point method. Each solution can then be substituted in Eq. (1.39) to obtain the corresponding invariant density μ_∞. Equation (1.40) is known in the literature as a *self-consistency* equation; see [75, Section 2.3] for more details. It always admits $m = 0$ as a solution when V is an even potential, but whether this solution is stable or not depends on the value of the interaction strength (θ) and the temperature (β^{-1}) [213, 51]. As suggested by our discussion above, there can be multiple steady states for sufficiently strong interactions or a sufficiently low temperature, and in fact it is possible to show that, in the limit $\beta \to \infty$ or $\theta \to 0$, the number of steady states is equal to the number of local minima of V [94]. The self-consistency map and its fixed points are illustrated in Fig. 1.4 for fixed $\theta = 1$ and two different values of β. Collating these points for a range of values of β, a *bifurcation diagram* of the first moment can be constructed, as depicted in Fig. 1.5a (left), revealing the existence of a *pitchfork bifurcation* corresponding to a so-called *second-order phase transition* of the particle system.

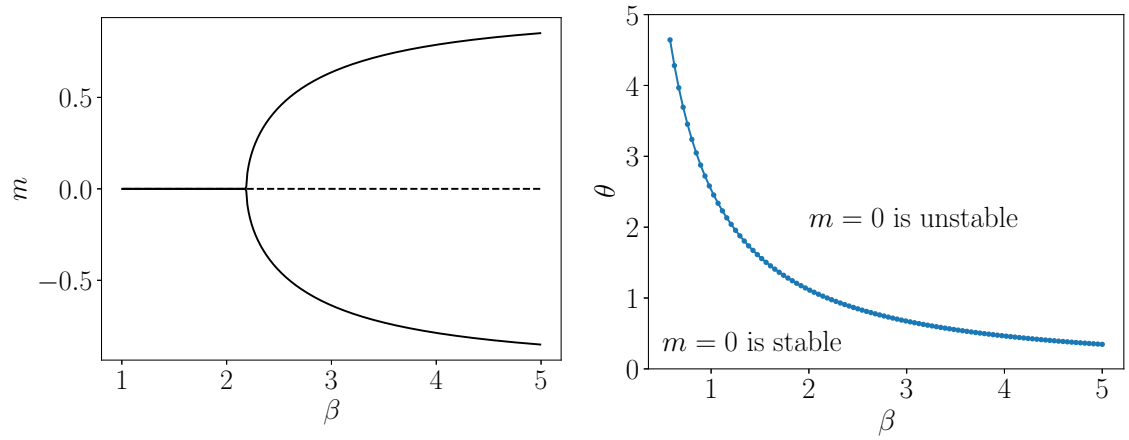
By solving the equation $\partial_m R(m;\beta,\theta) = 1$, the critical curve at which the mean-zero solution loses its stability can be obtained in the β – θ plane. This curve is depicted in Fig. 1.5b, where we observe that the critical temperature decreases as the interaction strength decreases.



(a) $\beta = 1$: $m = 0$ is stable.

(b) $\beta = 10$: $m = 0$ is unstable.

Figure 1.4: Self-consistency map for two different values of the temperature, when $\theta = 1$ and V is the bistable potential. When the temperature decreases (β increases) enough, the steady state corresponding to $m = 0$ loses its stability and two new solutions of Eq. (1.40) emerge.



(a) Bifurcation diagram.

(b) Stability of the even solution.

Figure 1.5: Bifurcation diagram of m against β when $\theta = 1$ (left), and stability of the $m = 0$ steady state in the β–θ plane (right).

Roughly speaking, our goal in Chapter 3 will be to examine the effect of colored noise on the steady states of the system and phase transitions. We will study not only the influence of the correlation time of the noise, but also that of the specific form of the invariant measure of the noise. We now present a brief introduction to colored noise and show the connection between our work in Chapter 3 and multiscale modeling.

1.3.2 Colored noise problem

As mentioned above, in Chapter 3 we will consider that the noise processes ξ_t^i in Eq. (1.33) are mean-zero independent stationary processes with nonzero correlation time. More specifically, we will consider, among other models, that the noise process (appropriately rescaled) can be

expressed as the solution to the overdamped Langevin equation:

$$\varepsilon \, \mathrm{d}\xi_t^i = -\frac{1}{\varepsilon^2} V_\eta'(\varepsilon \, \xi_t^i) \, \mathrm{d}t + \frac{\sqrt{2}}{\varepsilon} \, \mathrm{d}W_t^i, \qquad i = 1, \dots, N,$$

where W^i, $i = 1, \dots, N$ are independent Brownian motions, V_η is a confining potential, and ε is a parameter encoding the correlation time of the noise. The assumption that the noise is mean-zero requires that V_η satisfy the now familiar (from Section 1.2) *centering condition*:

$$\int_{\mathbf{R}} y \, \mathrm{e}^{-V_\eta(y)} \, \mathrm{d}y = 0.$$

Although it is possible to obtain, at least formally, a McKean–Vlasov equation of the type (1.38) in the mean-field limit of the particle system with colored noise, this equation is no longer a gradient flow. Numerical methods relying explicitly on a gradient flow structure, such as the finite volume method introduced in [37] and successfully employed for the simulation of the McKean–Vlasov equation with white noise [94], are thus not applicable in this case. Several options remain available: one could for example employ a standard finite volume method or a finite element method, both of which usually enjoy the probability conservation property. It might also be possible, using the ideas of Chang and Cooper [39], to design a finite difference method that is both positivity and probability preserving. A third option, when positivity and probability preservation can be forgone, is to use a spectral method, which often yields much more accurate (in the $L^2(\mathbf{R})$ norm, for example) results for the same computational cost. This will be our approach in Chapter 3.

The presence of colored noise also complicates the calculation of steady states of the mean-field McKean–Vlasov equation, because these cannot in general be obtained by solving a scalar self-consistency equation of the type (1.40). When the *correlation time* of the noise is small ($\varepsilon \ll 1$), however, a multiscale analysis based on [115] reveals that it is possible to approximate the steady states by solving a modified self-consistency equation. This is not surprising, since by homogenization (see Section 1.2.2) the noise processes satisfy

$$\int_0^{\cdot} \xi_s^i \, \mathrm{d}s \to W^i \quad \text{as} \quad \varepsilon \to 0,$$

in law and also pathwise, see [174, Chapter 5]. Using results from this multiscale analysis, we will be able to validate our numerical results obtained via a spectral method.

1.4 Diffusion of a Brownian particle in a periodic potential

In Chapter 4, we study the diffusion in a periodic potential of a Brownian particle whose motion is governed by the *generalized Langevin equation* (GLE):

$$\ddot{q}(t) = -V'(q) - \int_0^t \gamma(t-s) \, \dot{q}(s) \, \mathrm{d}s + F(t). \tag{1.41}$$

Here V is a periodic potential, q is the (non-periodized) position of the particle, $\gamma(\cdot)$ is the so-called *memory kernel*, and F is a mean-zero stationary Gaussian noise. The second term on the right-hand side of Eq. (1.41) is a dissipation term that depends on the history of the particle. The GLE is often used in molecular dynamics and nonequilibrium statistical mechanics to model a particle interacting with a heat bath at equilibrium. Both the dissipation and forcing terms originate from the interaction of the particle with the heat bath; they are related by the *fluctuation/dissipation theorem*, which stipulates that

$$\mathbf{E}(F(t)F(s)) = \beta^{-1}\,\gamma(t-s),$$

where β is the inverse temperature; see for example [174]. A formal derivation of the GLE, based on a simple model for the bath and its interaction with the particle, is presented in [104].

When the memory kernel $\gamma(t)$ in Eq. (1.41) is of the form $\left\langle \mathrm{e}^{-\mathbf{A}t}\,\boldsymbol{\lambda}, \boldsymbol{\lambda}\right\rangle$, for $\boldsymbol{\lambda} \in \mathbf{R}^n$ and $\mathbf{A} \in \mathbf{R}^{n\times n}$, it is possible to represent the solution as a Markovian process on an extended phase space, making the problem more amenable to analysis: introducing a vector $\mathbf{z}$ of n auxiliary variables, $(q, p, \mathbf{z})$ satisfy (in law):

$$\mathrm{d}q = p\,\mathrm{d}t, \tag{1.42a}$$

$$\mathrm{d}p = -V'(q)\,\mathrm{d}t + \langle \boldsymbol{\lambda}, \mathbf{z}\rangle\ \mathrm{d}t, \tag{1.42b}$$

$$\mathrm{d}\mathbf{z} = -p\,\boldsymbol{\lambda}\,\mathrm{d}t - \mathbf{A}\,\mathbf{z}\,\mathrm{d}t + \boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t, \qquad \mathbf{z}(0) \sim \mathcal{N}(0, \beta^{-1}\mathbf{I}), \tag{1.42c}$$

where $\mathbf{W}(t)$ is an n-dimensional standard Brownian motion and the matrix $\boldsymbol{\Sigma}$ is determined from $\mathbf{A}$ by the fluctuation/dissipation theorem: $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = \beta^{-1}(\mathbf{A} + \mathbf{A}^T)$. That such a representation holds is related to the result, due to Joseph L. Doob, that a mean-zero stationary Gaussian process with an exponential autocorrelation function and continuous paths is necessarily an Ornstein–Uhlenbeck process.

1.4.1 Relationship between the GLE and Langevin dynamics

In this subsection, we show formally how the GLE is related to the *Langevin dynamics* and to the *overdamped Langevin dynamics*. To this end, let us first consider the scaling $\gamma(t) = \frac{1}{\tau}\,\hat{\gamma}\left(\frac{t}{\tau}\right)$ for some autocorrelation function $\hat{\gamma}$ (that is, for a positive definite function $\hat{\gamma}$ such that $\hat{\gamma}(0) > 0$), and let us define

$$\gamma = \int_0^\infty \gamma(t)\,\mathrm{d}t = \int_0^\infty \hat{\gamma}(t)\,\mathrm{d}t.$$

To distinguish the memory kernel from its integral, the functional nature of the former will always be written explicitly, for example by using the notation $\gamma(\cdot)$. When the amount of memory of the system, here parametrized by τ, tends to zero, the memory kernel converges to a multiple of the Dirac delta distribution: $\gamma(\cdot) \to 2\,\gamma\,\delta(\cdot)$ in the sense of distributions over $\mathbf{R}$, where the factor 2 originates from the fact $\gamma(\cdot)$ is an even function. In that limit, we recover formally from Eq. (1.41) the *Langevin equation* with friction coefficient γ,

$$\ddot{q} = -V'(q) - \gamma\,\dot{q} + \sqrt{2\,\gamma\,\beta^{-1}}\dot{W}(t), \tag{1.43}$$

where W is a standard Wiener process, which is consistent with the fluctuation/dissipation theorem since $\mathbf{E}\big(\dot{W}(t)\dot{W}(s)\big) = \delta(t-s)$. Thus, the GLE is indeed a generalization of the Langevin equation. The interested reader may refer to [174, Result 8.4] and [168] for a more complete formal argument and a rigorous proof, respectively. As usual, the rigorous interpretation of Eq. (1.43) is in integral form.

Rescaling time appropriately, it is possible to obtain from Eq. (1.43) an even simpler model in the limit $\gamma \to \infty$, known as the *overdamped Langevin equation*:

$$\dot{q} = -V'(q) + \sqrt{2\,\beta^{-1}}\,\dot{W}_\infty(t), \tag{1.44}$$

with W_∞ a standard Brownian motion. To justify this limit formally, let us introduce the (dimensionless) momentum variable $p = \dot{q}$ and rescale q and p in time as $q_\gamma(t) := q(\gamma t)$ and $p_\gamma(t) := p(\gamma t)$. In terms of these rescaled variables, Equation (1.43) reads

$$\begin{aligned}\dot{q}_\gamma &= \gamma\, p_\gamma,\\ \dot{p}_\gamma &= -\gamma\, V'(q_\gamma) - \gamma^2\, p_\gamma + \gamma\,\sqrt{2\beta^{-1}}\,\dot{W}_\gamma(t),\end{aligned}$$

where $W_\gamma(t)$ is another Wiener process. Using the variation-of-constants formula, the second equation can be integrated and substituted into the first: denoting the initial momentum by p_0 and assuming that the initial position is 0 for simplicity,

$$\begin{aligned}q_\gamma(t) &= \int_0^t \left(\gamma\, p_0\, \mathrm{e}^{-\gamma^2 s} - \gamma^2 \int_0^s \mathrm{e}^{-\gamma^2(s-r)}\, V'(q_\gamma(r))\,\mathrm{d}r + \gamma^2\sqrt{2\,\beta^{-1}} \int_0^s \mathrm{e}^{-\gamma^2(s-r)}\,\mathrm{d}W_\gamma(r)\right)\mathrm{d}s\\ &\approx -\int_0^t V'(q_\gamma(s))\,\mathrm{d}s + \int_0^t \sqrt{2\,\beta^{-1}}\,\mathrm{d}W_\infty(s) \quad \text{when} \quad \gamma \gg 1.\end{aligned}$$

Here, we used the fact that the stochastic integral in the above equation is a Gaussian process with mean zero and an autocorrelation function that can be calculated using the *Itô isometry*: omitting some details,

$$\begin{aligned}&\mathbf{E}\left[\left(\gamma^2\int_0^s \mathrm{e}^{-\gamma^2(s-r)}\,\mathrm{d}W_\gamma(r)\right)\left(\gamma^2\int_0^t \mathrm{e}^{-\gamma^2(t-r)}\,\mathrm{d}W_\gamma(r)\right)\right]\\ &\quad= \gamma^4\,\mathrm{e}^{-\gamma^2|t-s|}\int_0^{\min(s,t)} \mathrm{e}^{-2\gamma^2(\min(s,t)-r)}\,\mathrm{d}r \to \delta(t-s) \quad \text{as} \quad \gamma\to\infty.\end{aligned}$$

1.4.2 Calculation of the effective diffusion coefficient

We saw that Eqs. (1.41), (1.43), and (1.44) form a hierarchy of models. For each of these models, it is possible to prove a *central limit theorem* similar to the homogenization result presented in Section 1.2: the rescaled process $\varepsilon\, q(t/\varepsilon^2)$ converges weakly, in the sense of probability measures on $C([0,T];\mathbf{R})$, to a pure diffusion:

$$\varepsilon\, q(t/\varepsilon^2) \Rightarrow \sqrt{2\,D}\,W(t), \quad \text{in } C([0,T];\mathbf{R}) \quad \text{when } \varepsilon \to 0. \tag{1.46}$$

The coefficient D, which appears as a simple multiplier of the Laplacian in the Fokker–Planck equation associated with the effective diffusion, is in general different for each model. Despite

being defined slightly differently from the effective diffusion coefficient in Section 1.2, it will also be referred to as the *effective diffusion coefficient*. Our main aim in Chapter 4 is to study the behavior of this coefficient for the generalized Langevin dynamics and how it relates to the diffusive behavior of the two other models.

By the properties of Brownian motion, it is simple to show that $\mathbf{E}(W(t)W(s)) = \min(s,t)$, so it follows from Eq. (1.46) that the effective diffusion coefficient is equal to

$$D = \lim_{t\to\infty} \frac{\mathbf{E}\,|q(t) - q(0)|^2}{2t}. \tag{1.47}$$

This immediately suggests approximating D by generating a high number of trajectories of the dynamics via a numerical method for SDEs and by employing an ensemble average to approximate the expectation in Eq. (1.47). This procedure is illustrated in Fig. 1.6 for the GLE. For Markovian

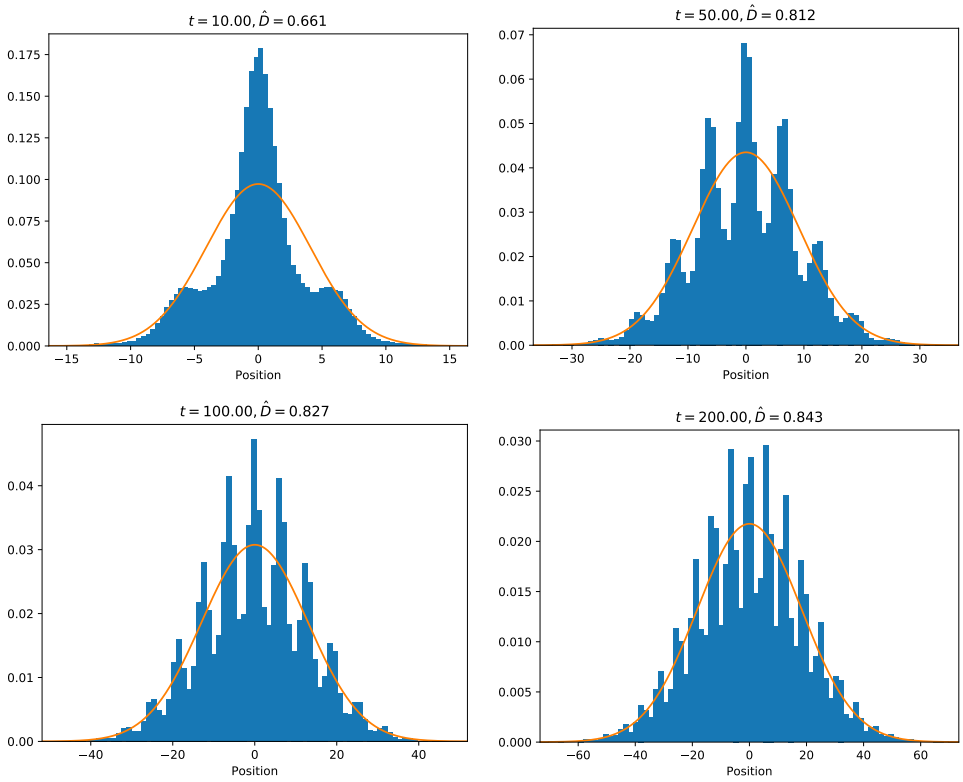


Figure 1.6: Illustration of the effective diffusion by Monte–Carlo simulation of the GLE with Ornstein–Uhlenbeck noise; see Chapter 4 for more details on this model of the noise. We calculated the histograms and the approximation of the effective diffusion coefficient based on 10^5 trajectories, all started at $q = p = 0$. The exact effective diffusion coefficient in this case, calculated using an accurate spectral method, is equal to $D = 0.841$. The orange curve is the Green function associated with the one-dimensional heat equation with diffusivity D, i.e. $(4\,\pi\,D\,t)^{-1/2}\,\exp(-\frac{x^2}{4\,D\,t})$. It appears clearly from the local oscillations of the empirical density that the convergence of the rescaled position only holds in a weak sense, not pointwise or in $L^p(\mathbf{R})$.

approximations of the GLE and for the Langevin equation, Eqs. (1.42) and (1.43) respectively, several alternative methods of calculating the effective diffusion have been developed, which we present concisely below. In this work, we mostly take a deterministic approach, based on the solution of an appropriate Poisson equation.

Poisson equation

To derive the effective diffusion coefficient in terms of the solution to an appropriate Poisson equation, we rewrite the rescaled dynamics as a fast/slow system of SDEs of the type introduced in Section 1.2 and studied in Chapter 2. Relabeling q as x, re-introducing q as the periodized position, i.e. the position within the periodic cell, and applying the scalings $x \mapsto x/\varepsilon$, $t \mapsto t/\varepsilon^2$, we obtain from Eq. (1.42):

$$\mathrm{d}x = \frac{1}{\varepsilon}\, p\, \mathrm{d}t, \tag{1.48a}$$

$$\mathrm{d}q = \frac{1}{\varepsilon^2}\, p\, \mathrm{d}t, \tag{1.48b}$$

$$\mathrm{d}p = \frac{1}{\varepsilon^2}\left(-V'(q) + \langle \boldsymbol{\lambda}, \mathbf{z}\rangle\right)\, \mathrm{d}t, \tag{1.48c}$$

$$\mathrm{d}\mathbf{z} = -\frac{1}{\varepsilon^2}\left(p\,\boldsymbol{\lambda} + \mathbf{A}\,\mathbf{z}\right)\, \mathrm{d}t + \frac{1}{\varepsilon}\,\boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}(t), \quad \mathbf{z}(0) \sim \mathcal{N}(0, \beta^{-1}\mathbf{I}). \tag{1.48d}$$

The fast dynamics is independent of the slow variable x, and it is ergodic with an invariant distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbf{T} \times \mathbf{R} \times \mathbf{R}^n$:

$$\mu(\mathrm{d}q\, \mathrm{d}p\, \mathrm{d}\mathbf{z}) = \frac{1}{\mathcal{Z}}\, \mathrm{e}^{-H(q,p) + \frac{|\mathbf{z}|^2}{2}}\, \mathrm{d}q\, \mathrm{d}p\, \mathrm{d}\mathbf{z}, \qquad H(q,p) = V(q) + \frac{p^2}{2},$$

with $\mathcal{Z}$ a normalization constant. Using the multiscale technique seen in Section 1.2, we conclude that the associated effective drift is equal zero, as expected, and we obtain an effective diffusion coefficient given by:

$$D = \int_{\mathbf{T}\times\mathbf{R}\times\mathbf{R}^n} \phi\, p\, \mathrm{d}\mu(\mathrm{d}q\, \mathrm{d}p\, \mathrm{d}q), \tag{1.49}$$

where ϕ is the solution of the following Poisson equation:

$$-\mathcal{L}\phi = p. \tag{1.50}$$

Here $\mathcal{L}$ is the generator of the fast dynamics in Eq. (1.48).

Green–Kubo formula

The Green–Kubo approach is related to the *heterogeneous multiscale method* reviewed in Section 1.2.3. For the problem under consideration, it can be presented in a self-contained manner, relying only on the expression Eq. (1.47) of the effective diffusion coefficient and on ergodicity. Assuming for simplicity that the dynamics is started at stationarity and writing q as the integral of p in Eq. (1.47), the effective diffusion coefficient can be rewritten as the integral of the autocorrelation function of the momentum, which we denote by C_p:

$$D = \lim_{t\to\infty} \frac{1}{2t}\, \mathbf{E}\left[\int_0^t p(s)\, \mathrm{d}s \int_0^t p(r)\, \mathrm{d}r\right] = \lim_{t\to\infty} \frac{1}{t}\, \mathbf{E}\left[\int_0^t \int_0^{t-s} p(r)\, p(r+s)\, \mathrm{d}r\, \mathrm{d}s\right] \tag{1.51a}$$

$$= \lim_{t\to\infty} \frac{1}{t} \int_0^t (t-s)\, C_p(s)\, \mathrm{d}s = \lim_{t\to\infty} \int_0^\infty \max\left(0, 1 - \frac{s}{t}\right) C_p(s)\, \mathrm{d}s = \int_0^\infty C_p(t)\, \mathrm{d}t, \tag{1.51b}$$

by dominated convergence and assuming that $C_p \in L^1(\mathbf{R})$. An advantage of this approach is that D can be calculated from only one trajectory; indeed by ergodicity we have

$$D = \int_0^\infty \lim_{T\to\infty} \left(\frac{1}{T} \int_0^T p(t)\, p(t+\delta)\, \mathrm{d}t \right) \mathrm{d}\delta = \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(\int_0^\infty p(t)\, p(t+\delta)\, \mathrm{d}\delta \right) \mathrm{d}t$$
$$\approx \frac{1}{T} \int_{T_0}^{T_0+T} \int_0^S p(t)\, p(t+\delta)\, \mathrm{d}\delta\, \mathrm{d}t, \tag{1.52}$$

for T and S large enough. The effective diffusion coefficient can then be approximated numerically by discretizing the integrals in Eq. (1.52). We remark that, in fact, this approach is exactly the one prescribed by the *heterogeneous multiscale method* when only one replica of the fast dynamics is employed.

Non-equilibrium steady-state technique

Another approach based on particle simulations is to simulate the dynamics when a constant force $\eta\,\mathbf{F}$, where $\eta \in \mathbf{R}$ and $\mathbf{F}$ is a unit vector in the direction of the force, is applied to the system. Considering the Langevin dynamics in one dimension for simplicity, the modified dynamics read

$$\mathrm{d}q = p\,\mathrm{d}t, \tag{1.53a}$$
$$\mathrm{d}p = \eta - V'(q)\,\mathrm{d}t - \gamma\, p + \sqrt{2\,\gamma\,\beta^{-1}}\,\mathrm{d}W(t). \tag{1.53b}$$

It is possible to show that, in the presence of this external forcing, the (periodized) dynamics is still ergodic for every η, with a unique invariant measure μ_η that is absolutely continuous with respect to the Lebesgue measure; see [142] and references therein. In contrast with the equilibrium case, however, the mean particle velocity at the steady state is nonzero, as illustrated in Fig. 1.7. The

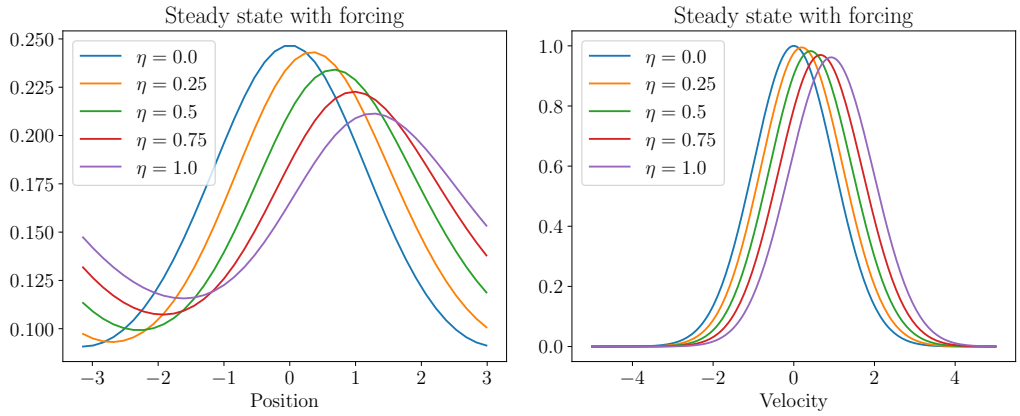


Figure 1.7: Marginals of the steady state solution of the Langevin dynamics with forcing, Eq. (1.53a). The simulation were carried using *Hermipy*, the software library we developed for Fourier/Hermite spectral methods; see Chapter 6 and [216]. Here we used the potential $V(x) = (1/2)\,(1 - \cos(x))$ and the periodic cell $[-\pi, \pi]$.

equilibrium probability density for nonzero η can be approximated by a power series expansion of the form $\mu_\eta = \mu_0 + \eta\mu_1 + \cdots$. Denoting by $\mathcal{L}^* + \eta\,\mathcal{L}_1^*$, with $\mathcal{L}_1^* = -\eta\,\partial_p$, the Fokker–Planck or forward Kolmogorov operator associated with (1.53), the densities corresponding to μ_i, which

will be denoted by the same symbols, satisfy:

$$-\mathcal{L}^*\mu_0 = 0, \qquad -\mathcal{L}^*\mu_1 = \mathcal{L}_1^*\mu_0, \qquad \dots$$

The first equation implies as expected that $\mu_0 = \mu$, where μ denotes the unique invariant density of the unperturbed dynamics. After substitution of the explicit expression of this density, the second equation reads

$$-\mathcal{L}^*\mu_1 = p\,\mu,$$

which implies, taking into account that the right-hand side satisfies the centering condition, that $\mu_1 = (-\mathcal{L}^*)^{-1}(p\,\mu)$. For small η, the steady-state flux can therefore be estimated as follows:

$$\begin{aligned}\int_{\mathbf{T}\times\mathbf{R}} p\,\mu_\eta\,\mathrm{d}q\,\mathrm{d}p &\approx \eta \int_{\mathbf{T}\times\mathbf{R}} p\,(-\mathcal{L}^*)^{-1}(p\mu)\,\mathrm{d}q\,\mathrm{d}p,\\ &= \eta \int_{\mathbf{T}\times\mathbf{R}} (-\mathcal{L}^{-1}p)\,p\,\mu\,\mathrm{d}q\,\mathrm{d}p,\end{aligned}$$

which, up to the factor η, is exactly the effective diffusion coefficient as defined in Eq. (1.49). It is possible to turn this formal argument into a rigorous one; see for example [191, 142]. Thus, this approach yields yet another characterization the effective diffusion coefficient:

$$D = \lim_{\eta\to 0}\frac{1}{\eta}\int_{\mathbf{T}\times\mathbf{R}} p\,\mu_\eta\,\mathrm{d}q\,\mathrm{d}p = \lim_{\eta\to 0}\frac{1}{\eta}\mathbf{E}_{\mu_\eta}\,p.$$

To carry out the calculation of the expectation in this equation, several options are possible. If the dimension of the problem is low enough, it is possible to calculate μ_η by a deterministic method, as was done to generate Fig. 1.7. In higher dimensions, Monte–Carlo approaches are more suitable and, by ergodicity of the dynamics Fig. 1.7, they can rely on both ensemble and time averages.

An advantage of this approach compared to the other probabilistic methods presented above, based on either Eq. (1.47) or Eq. (1.51), is that it can be combined with variance reduction techniques with control variates. Letting $\mathcal{L}_\eta$ denote the generator of the perturbed dynamics (1.53), we observe that

$$\mathbf{E}_{\mu_\eta} p = \mathbf{E}_{\mu_\eta}(p + \mathcal{L}_\eta\psi),$$

since $\mathbf{E}_{\mu_\eta}(\mathcal{L}_\eta\psi) = 0$ for any function ψ. In order approximate $\mathbf{E}_{\mu_\eta}p$ more efficiently, the function ψ may be chosen in such a way that the asymptotic variance of $p + \mathcal{L}_\eta\psi$ is less than that of p. This approach was introduced and rigorously analyzed in [189], and it was implemented in [191] to study the Langevin dynamics in the underdamped regime. The optimal choice of ψ, leading to a constant observable $p + \mathcal{L}_\eta\psi$, is the solution to the Poisson equation

$$-\mathcal{L}_\eta\psi = p - \mathbf{E}_{\mu_\eta}p. \tag{1.54}$$

In practice, unless the dimension of the problem is low, in which case the deterministic approach via the Poisson equation (1.50) is preferable, solving Eq. (1.54) precisely is not possible. Control

variates are therefore usually obtained from simplified models.

1.4.3 Convergence to equilibrium

In addition to the problem of effective diffusion in a periodic potential, Chapter 4 will also address the convergence to equilibrium of the generalized Langevin dynamics by means of PDE tools. In this section, we review some of the general techniques that can be employed for that purpose, taking as examples the Langevin and overdamped Langevin dynamics. For simplicity, we consider only the dynamics on the torus, i.e. we assume that $q \in \mathbf{T}$.

Instead of looking at the evolution of the probability distribution of the solution, we adopt the dual viewpoint of approaching the problem via the backward Kolmogorov equation. This approach makes the algebra simpler because the null space of the generator of the dynamics (overdamped Langevin, Langevin, or generalized Langevin), which we denote by $\mathcal{L}$, is spanned by constants. More precisely, we examine whether the solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\phi = \mathcal{L}\,\phi, \tag{1.55}$$

converges in some functional space to $\mathbf{E}_\mu\phi$, where $\mathbf{E}_\mu$ denotes the expectation with respect to the invariant measure μ associated with the dynamics. Since the kernel of $\mathcal{L}$ is spanned by constants, Eq. (1.55) implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\,(\phi - \mathbf{E}_\mu\phi) = \mathcal{L}\,(\phi - \mathbf{E}_\mu\phi),$$

so it is sufficient to consider the convergence to 0 of mean-zero functions with respect to μ. This suggests working in a space of mean-zero functions; in this subsection and in Chapter 4, we focus on the exponential decay of the semigroup generated by $\mathcal{L}$ on

$$L^2_0(\mu) = \left\{\phi \in L^2\,(\mu)\,;\, \int_{\mathcal{E}} \phi\,\mathrm{d}\mu = 0\right\},$$

where $\mathcal{E}$ denotes the state space of the dynamics, or on a Hilbertian subspace of this space. Specifically, we examine whether it is possible to construct a densely embedded Hilbert space $H \subset L^2_0(\mu)$ such that $\mathcal{L}$ is *hypocoercive* on that space.

Definition 1.3. The unbounded operator $\mathcal{L}$ is said to be λ-*hypocoercive* on H if there exists a finite constant C such that

$$\|\,\mathrm{e}^{-\mathcal{L}t}\,h\|_H \leq C\;\mathrm{e}^{-\lambda t}\;\|h\|_H, \qquad \forall h \in H. \tag{1.56}$$

It is said to be *hypocoercive* if it is λ-hypocoercive for some $\lambda > 0$.

If an estimate such as Eq. (1.56) holds, then a bound on the resolvent operator $\mathcal{L}^{-1}$ can be obtained by Proposition 1.2. Proving an exponential convergence result such as (1.56) is immediate when the (negative of the) generator is *coercive* on H.

Definition 1.4. The unbounded operator $\mathcal{L}$ is said to be λ-*coercive* on H if

$$\langle \mathcal{L}h, h\rangle_H \geq \lambda\|h\|_H^2 \quad \forall h \in D(\mathcal{L}). \tag{1.57}$$

It is said to be *coercive* if it is λ-coercive for some $\lambda > 0$.

Indeed, it is possible to show that λ-coercivity not only implies (by Grönwall's lemma) but is *equivalent* to the exponential decay (1.56) of the semigroup with the same constant λ and $C = 1$; see [220, Proposition 9].

We note that it is also possible, although this will not be necessary for our purposes, to study convergence in non-Hilbertian settings, notably in weighted L^∞ spaces; see for example [102].

Convergence of the overdamped Langevin dynamics

The generator of the overdamped Langevin dynamics provides us with an example of a coercive operator. Recalling that the associated invariant probability density is given by $\mu = \frac{1}{\mathcal{Z}_\beta}\, \mathrm{e}^{-\beta V}$, with a normalization constant $\mathcal{Z}_\beta$, we can show coercivity on $L^2_0(\mu)$ by first employing integration by parts to notice that

$$\begin{aligned} \langle \mathcal{L}h, h\rangle_{L^2_0(\mu)} &= \int_{\mathbf{T}} \left(-V'\, h' + \beta^{-1}\, h''\right)\, h\, \left(\frac{\mathrm{e}^{-\beta V}}{\mathcal{Z}_\beta}\right) \mathrm{d}q \\ &= \frac{1}{\beta\, \mathcal{Z}_\beta} \int_{\mathbf{T}} \frac{\mathrm{d}}{\mathrm{d}q}\left(\mathrm{e}^{-\beta V}\; h'\right) h\, \mathrm{d}q = -\frac{1}{\beta} \int_{\mathbf{T}} (h')^2\, \mu(\mathrm{d}q), \end{aligned} \tag{1.58}$$

for any function $h(q)$ that is regular enough. The coercivity of $-\mathcal{L}$ then follows from the Poincaré inequality (which holds for all smooth potentials on $\mathbf{T}$):

$$\int_{\mathbf{T}} h^2\, \mathrm{d}\mu \leq C_V \int_{\mathbf{T}} (h')^2\, \mathrm{d}\mu \qquad \forall h \in H^1(\mu) \cap L^2_0(\mu),$$

for a constant C_V depending on the potential.

In view of Eq. (1.58), we introduce the notation $\mathrm{d}_q^* = V' - \beta^{-1}\mathrm{d}_q$ to denote the formal adjoint of the derivative operator in $L^2_0(\mu)$, so that $\mathcal{L} = -\mathrm{d}_q^*\mathrm{d}_q$. The coercivity of $-\mathcal{L}$ on more regular weighted Sobolev spaces, of the type $H^i(\mu) \cap L^2_0(\mu)$, can be proved similarly provided that we make additional assumptions on V. Assuming, for example, that $\max_{q\in\mathbf{T}} |V''| \leq 1$ and using the notation $[A, B]$ to denote the commutator $(AB - BA)$ between two operators A and B,

$$\begin{aligned} \langle \mathcal{L}h, h\rangle_{H^1(\mu)\cap L^2_0(\mu)} &= \left\langle \mathrm{d}_q(\mathrm{d}_q^*\mathrm{d}_q)h, \mathrm{d}_q h\right\rangle_{L^2_0(\mu)} + \left\langle (\mathrm{d}_q^*\mathrm{d}_q)h, h\right\rangle_{L^2_0(\mu)}, \\ &= \left\langle [\mathrm{d}_q, \mathrm{d}_q^*]\mathrm{d}_q h, \mathrm{d}_q h\right\rangle_{L^2_0(\mu)} + \left\langle \mathrm{d}_q^2 h, \mathrm{d}_q^2 h\right\rangle_{L^2_0(\mu)} + \langle \mathrm{d}_q h, \mathrm{d}_q h\rangle_{L^2_0(\mu)}, \\ &= \left\langle \mathrm{d}_q^2 h, \mathrm{d}_q^2 h\right\rangle_{L^2_0(\mu)} + \left\langle (1 + V'')\, \mathrm{d}_q h, \mathrm{d}_q h\right\rangle_{L^2_0(\mu)} \geq C'_V\, \|h\|^2_{H^1(\mu)\cap L^2_0(\mu)}, \end{aligned}$$

where the Poincaré inequality was used for $\mathrm{d}_q^2 h$. By assigning more weight to the terms corresponding to derivatives of higher order in the Sobolev norm, the boundedness assumption on V'' can be lifted, but this is not necessary for our purposes.

Convergence of the Langevin dynamics

Proving exponential convergence to equilibrium for the Langevin dynamics is a more delicate task, because in this case the (negative of the) generator $\mathcal{L}$ is no longer coercive. To see why, we

decompose the generator into its symmetric and antisymmetric parts:

$$\mathcal{L} = \gamma \left(-p \frac{\partial}{\partial p} + \beta^{-1} \frac{\partial^2}{\partial p^2} \right) + \left(p \frac{\partial}{\partial q} - \frac{\partial V}{\partial q} \frac{\partial}{\partial p} \right) =: -A^* A + B, \tag{1.59}$$

where $A = \sqrt{\gamma}\, \partial_p$ and $A^* = \sqrt{\gamma}(p - \beta^{-1}\partial_p)$ denotes the adjoint operator in $L^2(\mu)$. We take for granted that $\mathcal{L}$ generates a strongly continuous semigroup on $L^2_0(\mu)$, and that the domain of $\mathcal{L}$ in $L^2_0(\mu)$ is dense in $L^2_0(\mu)$; see [147] for details. The decomposition (1.59) makes apparent the nonexpansivity of the semigroup generated by $\mathcal{L}$: for $h \in D(A^*A) \cap D(B)$, it holds

$$\left(\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \| \mathrm{e}^{\mathcal{L}t} h \|^2_{L^2_0(\mu)} \right)_{t=0} = \langle \mathcal{L}h, h \rangle_{L^2_0(\mu)} = - \langle Ah, Ah \rangle_{L^2_0(\mu)} \leq 0. \tag{1.60}$$

By density of the domain in $L^2_0(\mu)$, the semigroup property, and the continuity of the semigroup, this implies that $\| \mathrm{e}^{t\mathcal{L}} \|_{\mathcal{B}(L^2_0(\mu))} \leq 1$. In contrast with the coercive case, the inequality (1.60) is not strict, because $\|Ah\|_{L^2_0(\mu)} = 0$ for all functions depending only on q. It is still possible, nevertheless, to show the *hypocoercivity* of $-\mathcal{L}$, on both $H^1(\mu) \cap L^2_0(\mu)$ and $L^2_0(\mu)$, and these two approaches also enable the calculation of explicit bounds for the rate of convergence.

Example 1.1. Before summarizing the two approaches, it is instructive to look at one of the simplest possible hypocoercive systems:

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & -1 \\ 1 & -\gamma \end{pmatrix} \mathbf{x} =: \mathbf{L}\mathbf{x}, \quad \gamma > 0, \ x \in \mathbf{R}^2. \tag{1.61}$$

Writing $\mathbf{x} = (x, y)^T$, it is clear that $\mathrm{d}_t |\mathbf{x}|^2 = -2\,\gamma\, y^2$, which is negative everywhere except on the straight line $y = 0$. Let us define a modified, equivalent norm $(\!(\mathbf{x}, \mathbf{x})\!) = x^2 + 2\,\alpha\, x\, y + \beta y^2$, with $\alpha^2 < \beta$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t} (\!(\mathbf{x}, \mathbf{x})\!) = \mathbf{x}^T \begin{pmatrix} 2\,\alpha & -\alpha\,\gamma + \beta - 1 \\ -\alpha\,\gamma + \beta - 1 & -2\,\alpha - 2\,\beta\,\gamma \end{pmatrix} \mathbf{x}.$$

The matrix is clearly negative definite for suitably chosen parameters α and β, e.g. $(\alpha, \beta) = (-.2, 1)$, implying the convergence to zero of $(\!(\mathbf{x}, \mathbf{x})\!)$ and, by equivalence of $|\mathbf{x}|^2$. Graphically, the parameter α adds a tilt to the level-set of $(\!(\mathbf{x}, \mathbf{x})\!)$ at the line $y = 0$, ensuring that the tangent vector of the trajectories are never parallel to the level sets of the norm; see Fig. 1.8. ⊘

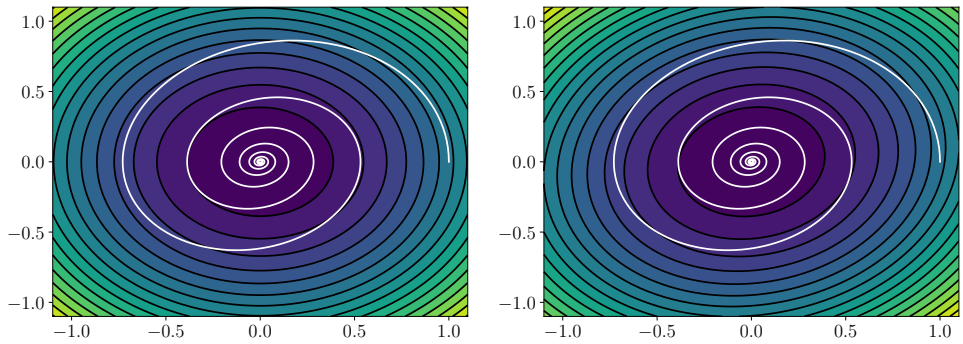


Figure 1.8: Illustration of the dynamics Eq. (1.61), with in the background the level sets of the norms $|\mathbf{x}|^2$ (left) and $(\!(\mathbf{x}, \mathbf{x})\!)$ (right).

Remark 1.6. Equation (1.61) is not as far removed from the Langevin dynamics as one might think. In fact it coincides, up to constant coefficients, to the system of equations obtained when studying the convergence to equilibrium of $\mathrm{e}^{\mathcal{L}t}$ on the subspace of linear functions in q and p when V is the confining potential $q^2/2$, using the ansatz $x(t)\,q + y(t)\,p$. More generally, systems of this type are obtained after discretizing a hypocoercive operator on a finite-dimensional subspace. ⊘

$H^1(\mu)$ approach

In order to prove the hypocoercivity of $-\mathcal{L}$ on $H^1(\mu) \cap L^2_0(\mu)$, the approach is to introduce an auxiliary norm that is equivalent to the $H^1(\mu)$ norm but includes mixed derivatives. This idea was first used in [208] and then generalized in [220]. A possible choice of the auxiliary norm, for the Langevin dynamics, is following: for $h \in H^1(\mu) \cap L^2_0(\mu)$,

$$|\!|\!|(h,h)|\!|\!| = \langle h, h\rangle_\mu + a\left\langle \frac{\partial h}{\partial q}, \frac{\partial h}{\partial q}\right\rangle_\mu + b\left\langle \frac{\partial h}{\partial q}, \frac{\partial h}{\partial p}\right\rangle_\mu + c\left\langle \frac{\partial h}{\partial p}, \frac{\partial h}{\partial p}\right\rangle_\mu, \tag{1.62}$$

where $\langle\cdot,\cdot\rangle_\mu$ denotes the $L^2_0(\mu)$ inner product. This norm is equivalent to the $H^1(\mu)$ norm provided that $ac - b^2 > 0$, and it defines an inner product by the polarization identity, $4\,|\!|\!|(h_1,h_2)|\!|\!| = |\!|\!|(h_1+h_2,h_1+h_2)|\!|\!| - |\!|\!|(h_1-h_2,h_1-h_2)|\!|\!|$. As we will see in Chapter 4 in a broader context, it is possible to choose a, b and c such that the coercivity condition (1.57) holds for the norm (1.62), showing the exponential decay of the semigroup in $H^1(\mu)$ by equivalence of the norms. Using the techniques of *hypoelliptic regularization*, introduced in [109] and further developed in [220], it is then possible to also prove convergence when the initial datum lies in $L^2_0(\mu)$ but not necessarily in $H^1(\mu)$. This program was performed in [167] for the GLE in an abstract setting.

$L^2_0(\mu)$ approach

For certain kinetic equations, it is possible to show hypocoercivity in $L^2_0(\mu)$ directly, without recourse to hypoelliptic regularization. This approach was pioneered in [58] under abstract assumptions and later particularized to a variety of problems, including the Langevin dynamics in [190]. It is also based on the introduction of an alternative equivalent norm that assigns some weight to mixed terms. For the Langevin dynamics, this norm reads

$$|\!|\!|(h,h)|\!|\!| = \langle h, h\rangle_\mu - \varepsilon\,\langle Ah, h\rangle_\mu, \quad A = (1 + (B\Pi_p)^*(B\Pi_p))^{-1}\,(B\,\Pi_p)^*,$$

where B is as defined in Eq. (1.59) and

$$\forall \phi \in L^2(\mu), \qquad \Pi_p(\phi) = \int_{\mathbf{R}} \phi(q,p)\,\kappa(\mathrm{d}p),$$

with κ the p marginal of μ. It is possible to show that, for suitably chosen ε, the norm $|\!|\!|(\cdot,\cdot)|\!|\!|$ is equivalent to the $L^2\,(\mu)$ norm and it satisfies the coercivity condition (1.57).

1.5 The Cahn–Hilliard equation

In Chapter 5, we present a novel numerical method for the *Cahn–Hilliard equation* with wetting boundary condition. The Cahn–Hilliard equation was derived by John W. Cahn and John E. Hilliard in their seminal 1958 paper [36] on the free energy of isotropic systems with nonuniform composition or density. It was originally introduced to model phase transitions and spinodal decomposition [35], the mechanism by which a binary mixture can separate to form two coexisting phases, and has since been employed to study many other physical phenomena, ranging from the dynamics of multiphase fluids [127] to electrostatically induced pattern formation in thin polymer films [130, 212] and the electromigration of intergranular voids in microelectronic circuits [18]. More recently, it was also used in image processing for the inpainting of binary images [20].

In contrast with sharp interface models for binary mixtures, which are based on the assumption that the interfaces between the fluid components have a zero thickness, the Cahn–Hilliard equation belongs to the class of *diffuse interface* models: the composition of the mixture is described by a continuous scalar function, called the *phase field*, which varies rapidly but smoothly in the interfacial region. The thickness of the interfacial region is encoded by a small parameter in the equation, and the interfaces are conventionally identified with the level set of the phase field corresponding to equal concentrations of the two phases. In the context of interfacial dynamics, the Cahn–Hilliard equation enjoys several advantages over sharp interface approaches: it is physically more accurate [36], can be analyzed using classical PDE techniques [68], and it lends itself well to numerical simulation as the explicit tracking of topological changes of the solution is not required.

To fix ideas, let us consider a mixture of two immiscible fluids A and B, and let us define the phase field ϕ as the difference between the local mole fractions of components A and B:

$$\phi = \frac{m_B - m_A}{m_B + m_A}.$$

With this definition, the pure phases correspond to $\phi = 1$ and $\phi = -1$. Let us also consider a domain Ω and denote by $\partial\Omega$ and $\mathbf{n}$ its boundary and the outward unit normal vector, respectively. The Cahn–Hilliard equation is most easily derived from the Ginzburg–Landau free energy functional:

$$E(\phi) = \int_\Omega \frac{1}{2}\,\varepsilon\,|\nabla\phi|^2 + \frac{1}{\varepsilon}\,F_m(\phi)\,\mathrm{d}\Omega + \int_{\partial\Omega} F_w(\phi, x)\,\mathrm{d}\sigma, \tag{1.63}$$

where $F_m(\phi)$ is the free energy per unit volume of a mixture of uniform phase field ϕ, ε is a small positive parameter encoding the interface thickness, and $F_w(\phi)$ is a surface free energy per unit area. We remark that the contributions to $E(\phi)$ of a volume element $\mathrm{d}\Omega$ depends only on the norm of $\nabla\phi$, and not on its direction; this isotropy property was used as a postulate by Cahn and Hilliard in order to derive the form of the free energy functional.

To guarantee that an equilibrium configuration of the binary mixture corresponds to a (possibly local) minimum of the free energy functional (1.63), the variation of this functional induced by any smooth perturbation ψ around the equilibrium must vanish:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}E(\phi + \alpha\psi) = \int_\Omega \left(\frac{1}{\varepsilon}\,F_m'(\phi) - \varepsilon\,\Delta\phi\right)\,\psi\,\mathrm{d}\Omega + \int_{\partial\Omega} \left(\varepsilon\,\nabla\phi\cdot\mathbf{n} + F_w'(\phi)\right)\psi\,\mathrm{d}\sigma = 0. \tag{1.64}$$

This equation suggests the natural boundary condition

$$\varepsilon \nabla\phi \cdot \mathbf{n} = F_w' =: f_w(\phi), \tag{1.65}$$

which is the usual boundary condition used to model wetting phenomena [121, 226, 56]. It also leads to the following expression of the chemical potential, equal by definition to the functional derivative of the free energy functional:

$$\mu := \frac{\delta E}{\delta \phi} = \frac{1}{\varepsilon} F_m'(\phi) - \varepsilon \Delta\phi =: \frac{1}{\varepsilon} f_m(\phi) - \varepsilon \Delta\phi, \tag{1.66}$$

where we introduced $f_m = F_m'$.

The Cahn–Hilliard equation encodes that the phase field is subject to diffusion with a flux proportional to the gradient of the chemical potential, a fact that can be justified from physical arguments [36, 138]. Denoting the proportionality coefficient by $b(\phi)$, this leads to the system of equations announced in Section 1.1:

$$\frac{\partial \phi}{\partial t} = \nabla \cdot (b(\phi)\, \nabla\mu), \tag{1.67a}$$

$$\mu = \frac{1}{\varepsilon} f_m(\phi) - \varepsilon \Delta\phi, \qquad \text{for } x \in \Omega,\ t \in (0, T], \tag{1.67b}$$

$$\varepsilon \nabla\phi \cdot \mathbf{n} = -f_w(\phi), \tag{1.67c}$$

$$b(\phi)\, \nabla\mu \cdot \mathbf{n} = \dot{m}, \qquad \text{for } x \in \partial\Omega,\ t \in (0, T]. \tag{1.67d}$$

The second boundary condition is a flux boundary condition, with $\dot{m}$ the specified flux. From these equations, we can obtain conservation laws for the *mass* $\int_\Omega \phi \,\mathrm{d}\Omega$ and the free energy (1.63). Integrating Eq. (1.67a) over Ω and integrating by parts, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \phi \,\mathrm{d}\Omega = -\int_{\partial\Omega} b(\phi)\, (\nabla\mu \cdot \mathbf{n})\,\mathrm{d}\sigma = -\int_{\partial\Omega} \dot{m}\,\mathrm{d}\sigma, \tag{1.68}$$

showing that $\dot{m}$ indeed corresponds to an outward diffusive mass flux. Multiplying Eqs. (1.67a) and (1.67b) by $\partial_t\phi$ and μ, respectively, subtracting the two resulting equations, and using integration by parts, we obtain a conservation law for the energy:

$$\frac{\mathrm{d}}{\mathrm{d}t} E(\phi) = -\int_\Omega b(\phi)\ |\nabla\mu|^2 \ \mathrm{d}\Omega - \int_{\partial\Omega} \mu\, \dot{m}\,\mathrm{d}\sigma.$$

In particular, the free energy is always decreasing when the mass flux is zero. In fact, it turns out that the Cahn–Hilliard describes a gradient flow under appropriate assumptions.

1.5.1 Formulation as a gradient flow

Let us assume here that the mobility $b(\phi)$ is a constant, equal to 1 for simplicity, and that the boundary conditions are homogeneous, $f_w = \dot{m} = 0$. Then, when the mass of the solution is zero, the Cahn–Hilliard equation describes a gradient flow of $E(\phi)$ over an appropriate Hilbert space

$H_0^{-1}(\Omega)$, defined as the mean-zero subspace of the dual space of $H^1(\Omega)$, i.e.

$$H_0^{-1}(\Omega) := \{u \in H^1(\Omega)' : \langle u, 1\rangle_{H^1(\Omega)',H^1(\Omega)} = 0\},$$

where $H^1(\Omega)'$ is the dual of $H^1(\Omega)$ and $\langle\cdot,\cdot\rangle_{H^1(\Omega)',H^1(\Omega)}$ denotes the corresponding duality pairing. A detailed presentation of this viewpoint is beyond the scope of this work, so we will content ourselves with a mostly formal presentation of the subject, but the interested reader can find more information in [47]. Let $\mathcal{G} : H_0^{-1}(\Omega) \to H^1(\Omega)$ denote the linear map that associates with a function $v \in H_0^{-1}(\Omega)$ the unique mean-zero solution (in the distributional sense) $u \in H^1(\Omega)$ to the Laplace equation with homogeneous Neumann boundary condition:

$$\begin{cases} -\Delta u = v & \text{in } \Omega, \\ \nabla u \cdot \mathbf{n} = 0 & \text{in } \partial\Omega. \end{cases}$$

With this definition, we define the inner product of $u, v \in H_0^{-1}(\Omega)$ as

$$\langle u, v\rangle_{H_0^{-1}(\Omega)} = \langle \nabla\mathcal{G}u, \nabla\mathcal{G}v\rangle_{L^2(\Omega)}.$$

Let us denote by X the space of smooth, compactly supported and mean-zero functions on Ω, and assume that ϕ is smooth and satisfies the homogeneous boundary conditions, that is $\nabla\phi\cdot\mathbf{n} = 0 = \nabla\Delta\phi\cdot\mathbf{n}$. Taking into account that $\mathcal{G}\Delta u = u$ if u is smooth, mean-zero and satisfies $\nabla u\cdot\mathbf{n} = 0$, and that $\Delta\mathcal{G} = \mathrm{id}$ on H_0^{-1}, it holds $\forall\psi \in X$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}E(\phi+\alpha\psi) &= \int_\Omega \left(\frac{1}{\varepsilon}f_m(\phi) - \varepsilon\,\Delta\phi\right)\psi\,\mathrm{d}\Omega \\
&= -\int_\Omega \left(\frac{1}{\varepsilon}\mathcal{G}\Delta f_m(\phi) + \frac{1}{\varepsilon\,|\Omega|}\int_\Omega f_m(\phi)\,\mathrm{d}\Omega - \varepsilon\,\mathcal{G}\Delta^2\phi\right)\Delta\mathcal{G}\psi\,\mathrm{d}\Omega \\
&= -\int_\Omega \nabla\mathcal{G}\left(\frac{1}{\varepsilon}\Delta f_m(\phi) - \varepsilon\,\Delta^2\phi\right)\cdot\nabla\mathcal{G}\psi\,\mathrm{d}\Omega \\
&= -\left\langle \Delta\left(\frac{1}{\varepsilon}f_m(\phi) - \varepsilon\,\Delta\phi\right), \psi\right\rangle_{H_0^{-1}(\Omega)} =: -\left\langle \mathrm{Grad}_{H_0^{-1}}E(\phi), \psi\right\rangle_{H_0^{-1}(\Omega)},
\end{aligned}$$

which shows, at least formally, that the right-hand side of the Cahn–Hilliard equation is indeed the Gâteaux gradient of $E(\phi)$ in H_0^{-1}.

Remark 1.7. In view of the previous calculation, we remark that the equation

$$\frac{\partial\phi}{\partial t} = \frac{1}{\varepsilon}f_m(\phi) - \varepsilon\,\Delta\phi, \tag{1.69}$$

known as the Allen–Cahn equation [6], also describes a gradient flow, but this time over the space $L^2(\Omega)$. Consequently, the functional $E(\phi)$ decreases along solutions of this equation also:

$$\frac{\mathrm{d}}{\mathrm{d}t}E(\phi) = \langle \mathrm{Grad}_{L^2}E(\phi), \partial_t\phi\rangle_{L^2(\Omega)} = -\|\mathrm{Grad}_{L^2}E(\phi)\|^2_{L^2(\Omega)} \le 0.$$

Unlike the Cahn–Hilliard equation, however, Eq. (1.69) does not enjoy a mass conservation property of the type of Eq. (1.68), and it is therefore not usually employed for modeling multiphase

flows. ⊘

In addition to emphasizing the fact that the Cahn–Hilliard equation dissipates the free energy functional $E(\phi)$, reframing the equation as a gradient flow over a Hilbert space enables the use of powerful tools from functional analysis and PDEs; see [73] and [47]. The gradient structure can also be exploited in order to devise stable numerical schemes; see for example [71].

1.5.2 Models for the free energies and the mobility

Much of the literature on Cahn–Hilliard equation has revolved around the study of different models for F_m, F_w, and b. Several models are commonly used for each of these quantities, each with their set of advantages and disadvantages. Arguably the most physical model for the bulk free energy F_m, derived using a mean-field argument in e.g. [138], is of the form

$$\begin{cases} F_m(\phi) = \dfrac{\theta}{2}\left[(1+\phi)\,\log\left(\dfrac{1+\phi}{2}\right) + (1-\phi)\log\left(\dfrac{1-\phi}{2}\right)\right] + \Omega\,(1-\phi^2), & \text{if } |\phi| \le 1, \\ F_m(\phi) = +\infty, & \text{if } |\phi| > 1, \end{cases} \tag{1.70}$$

where θ is the absolute temperature and Ω is a physical constant independent of the temperature. The potential (1.70) is illustrated in Fig. 1.9 for different values of θ. Since $F_m(\phi) = +\infty$ when $|\phi| > 1$, the solution to the corresponding Cahn–Hilliard equation is confined between the physical bounds -1 and 1.

When the temperature is sufficiently high, the potential admits only one minimum, in which case the configuration of minimum free energy is a mixture with uniform concentration. In contrast, below a critical temperature, F_m admits two minima at an equal distance from the origin. In this situation, interpreting these minima as phases, it is clear the two terms constituting the overall volume free energy in Eq. (1.63), respectively proportional to the integrals of $\frac{1}{\varepsilon}F_m(\phi)$ and of $\varepsilon\,|\nabla\phi|^2$, have opposing effects on the evolution of the mixture: while the latter favors the mixing of the two phases, the former tends to induce their separation. The relative strength of these two effects is determined by the parameter ε. It has been shown in [69, Theorem 2.2] that if $\varepsilon > \varepsilon_0$, for some constant ε_0 that depends on the dimensions of the domain, then $\lim_{t\to\infty} \|\phi - M\|_\infty \to 0$, where M is the average mass density. When $\varepsilon \ll 1$, on the other hand, thin interfaces of thickness approximately proportional to ε appear, away from which the phase field takes values close to the minimizers of F_m.

In the limit $\theta \to 0$, the function $F_m(\phi)$ converges uniformly to $\Omega\,(1-\phi^2)$ on $[-1,1]$, and the two minimizers of F_m converge to -1 and 1, respectively. The bulk free energy function obtained when $\theta = 0$, known as the obstacle potential, was initially proposed in [24] and has since been studied extensively in a variety of contexts, e.g. in [17, 18].

In this work, we employ the Helmoltz double-well potential, $F_m(\phi) = \frac{1}{4}(\phi^2-1)^2$, for $\phi \in \mathbf{R}$, which is illustrated in Fig. 1.9 and can be viewed as a smooth approximation of (1.70). With this model, the Cahn–Hilliard equation is generally more amenable to analysis and numerical simulation, but the solution is not confined to the physical interval $[-1,1]$. In an infinite one-dimensional domain, the Cahn–Hilliard equation with the Helmoltz potential admits the well-known hyperbolic tangent equilibrium solution, $\phi_1(x) = \tanh(\frac{x}{\sqrt{2}\varepsilon})$. Using this solution, and

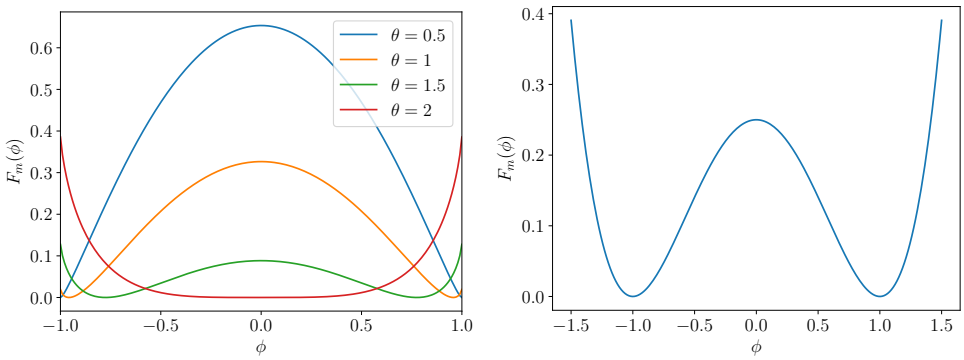


Figure 1.9: Bulk free energy (1.70) for different values of the absolute temperature (left) and Helmoltz double-well potential $\frac{1}{4}(\phi^2-1)^2$ (right).

noticing that $\phi_1' = \frac{1}{\sqrt{2}\varepsilon}(1-\phi_1^2)$, the contribution of the interface to the total free energy can be calculated explicitly:

$$\sigma = \int_{\mathbf{R}} \frac{1}{2}\,\varepsilon\,|\phi'|^2 + \frac{1}{\varepsilon}\,F_m(\phi)\,\mathrm{d}x = \frac{2\sqrt{2}}{3}. \tag{1.71}$$

In higher dimensions, by extrapolation, σ corresponds approximately to the contribution per unit length (in 2D) or area (in 3D) of the interface to the total free energy, because the variation of the phase field in the direction orthogonal to the interface has a profile similar to Eq. (1.71) when $\varepsilon \ll 1$; see for example [131].

Likewise, several models exist for the wall free energy $F_w(\phi)$ in Eq. (1.63), most of which were reviewed recently in [116]. The wall free energy enables modeling the wettability of the solid substrate surrounding the binary mixture; it allows specifying a *contact angle* between the interfaces and the substrate. Let us introduce $\sigma_A = F_w(-1)$ and $\sigma_B = F_w(1)$. By using an elementary energetic argument, it is possible to derive the dependence of the contact angle on the interfacial energies σ, σ_A, σ_B, where σ is as defined in Eq. (1.71). To this end we consider a multiphase system consisting of a two-dimensional circular droplet of fluid B deposited on a solid substrate, as illustrated in Fig. 1.10. If the interfaces are thin enough, i.e. $\varepsilon \ll 1$, then the total free energy of the system can be approximated based on Eq. (1.71),

$$E(\phi_\theta) = R\,(2\,\theta\,\sigma + \sin(\theta)\,(\sigma_A - \sigma_B))\,.$$

Minimizing the total free energy under the constraint that the area of the droplet remains constant directly provides us with *Young's equation* for the equilibrium contact angle:

$$\theta_0 = \operatorname*{arg\,min}_{\theta\in(0,\pi)} \frac{2\,\theta\,\sigma + \sin(\theta)\,(\sigma_A-\sigma_B)}{\sqrt{2\,\theta - \sin(2\,\theta)}} = \arccos\left(\frac{\sigma_B-\sigma_A}{\sigma}\right). \tag{1.72}$$

This shows us that, in order for the model of F_w to be a physical, σ_A and σ_B should correspond, up to appropriate dimensional constants, to the physical interfacial energies between the substrate and the pure phases A and B, respectively. We note that our derivation above is only formal, since we assumed *a priori* that the Cahn–Hilliard equation would lead the droplet to an equilibrium configuration where the fluid B occupies a circular segment adjacent to the

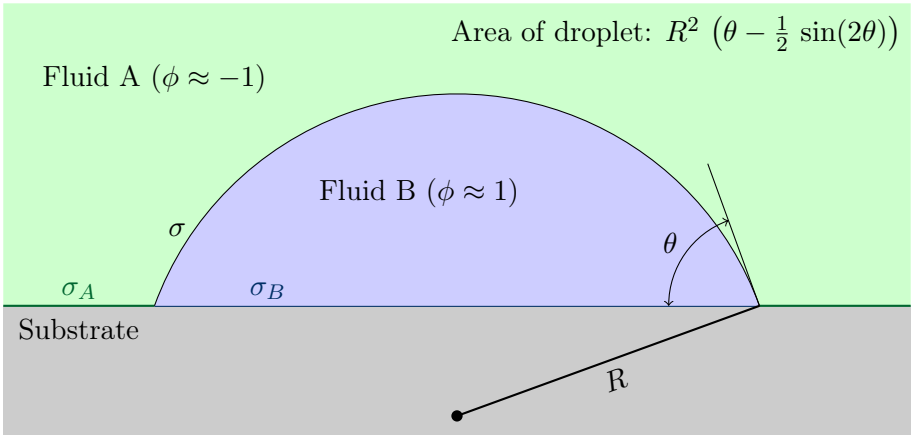


Figure 1.10: Schematic of a droplet at equilibrium on an B-philic (therefore A-phobic) substrate. Here the equilibrium contact angle is equal to 70 degrees.

boundary, irrespectively of the exact form of F_w. We also assumed that the phase field was approximately constant away from the fluid/fluid interface, which does not hold in general. Finally, we omitted to define precisely the contact angle in the case of a diffuse interface; clearly, if $f_w(0) = 0$, the 0 level set of the phase field is orthogonal to the boundary, by Eq. (1.67c), leading to an apparent contradiction with our derivation. Some of these issues are studied in [226], to which we point the interested reader for more details.

Three popular forms of the wall energy are the *linear*, *cubic*, and *sine* models. The *linear* model, which specifies a linear dependence between the phase field and the wall energy, is widely used in both numerical and theoretical works for its simplicity. Its main disadvantage is that it leads to the appearance of a parasitic *boundary layer* at the substrate. The *cubic* and *sine* forms, given by

$$F_w^{\text{cubic}}(\phi) = \frac{3}{4}\sigma\,\cos(\theta_0)\left(\frac{\phi^3}{3} - \phi\right) \quad \text{and} \quad F_w^{\text{sine}}(\phi) = -\frac{1}{2}\sigma\,\cos(\theta_0)\sin(\phi), \tag{1.73}$$

avoid this issue because $f_w(\pm 1) = F_w'(\pm 1) = 0$ in both cases. We can check that both of these forms satisfy Eq. (1.72). Of these two models, it can be argued that F_w^{cubic} is the most physical, at least when used in conjunction with the Helmoltz potential, because the associated Cahn–Hilliard equation posed in the half plane admits planar hyperbolic tangent solutions; see Chapter 5 for more details.

We note that other boundary conditions have also been employed in the literature to model the interaction between the binary mixture and the substrate. In particular, it has been shown that including a wall energy relaxation term in Eq. (1.67c) can improve the agreement between the model and experiments. It is also possible to impose the contact angle without relying on the surface energy formalism encoded in Eq. (1.63); see [194, 116] and references therein for details.

Before concluding this section, we briefly discuss two of the most popular models for the mobility coefficient b. The simplest possible approach is to assume that the mobility is independent of the phase field. As discussed above in Section 1.5.1, this choice simplifies the analysis of the Cahn–Hilliard equation and its numerical solution, but it allows a diffusive flux away from the interface, leading to sometimes unphysical or unwanted effects akin to Ostwald ripening. For

this reason, many authors employ the form $b(\phi) \propto 1 - \phi^2$, which has the advantage of being zero in the pure components. With this choice, the Cahn–Hilliard equation is degenerate, making its analysis considerably more difficult; in fact the problem of showing uniqueness in this case remains open after decades of research. Seminal publications on the subject are [68, 17].

1.5.3 The sharp interface limit

An introduction to the Cahn–Hilliard equation would not be complete without a discussion of the sharp interface limit, $\varepsilon \to 0$, which we present concisely in this section. Let us consider the Cahn–Hilliard equation, Eqs. (1.67a) and (1.67b), with constant mobility and homogeneous Neumann boundary conditions,

$$\begin{cases} \nabla\phi \cdot \mathbf{n} = 0 \\ \nabla\mu \cdot \mathbf{n} = 0 \end{cases} \quad \text{on} \quad \partial\Omega_T := \partial\Omega \times (0, T). \tag{1.74}$$

Let us also assume that the initial condition is such that there exists a closed hypersurface $\Gamma_0 \subset \Omega$ separating the domain into two distinct regions Ω_0^- and Ω_0^+, where Ω_0^- and Ω_0^+ denote respectively the region enclosed by Γ_0 and its complement, such that $\phi(x, t = 0) = -1$ for $x \in \Omega_0^-$ and $\phi_0(x, t = 0) = 1$ for $x \in \Omega_0^+$. Then when $\varepsilon \to 0$, the solution (ϕ, μ) converges, in an appropriate sense and under suitable assumptions, to a limit described by a sharp interface problem for (μ, Γ), known as the *Hele–Shaw problem*:

$$\Delta\mu(x, t) = 0 \qquad \text{for } x \in \Omega\backslash\Gamma(t) \text{ and } t \in [0, T], \tag{1.75a}$$

$$\nabla\mu \cdot \mathbf{n} = 0 \qquad \text{on } \partial\Omega_T := \partial\Omega \times (0, T], \tag{1.75b}$$

$$\mu = \frac{\sigma}{2}\kappa \qquad \text{on } \Gamma(t), \text{ for } t \in [0, T], \tag{1.75c}$$

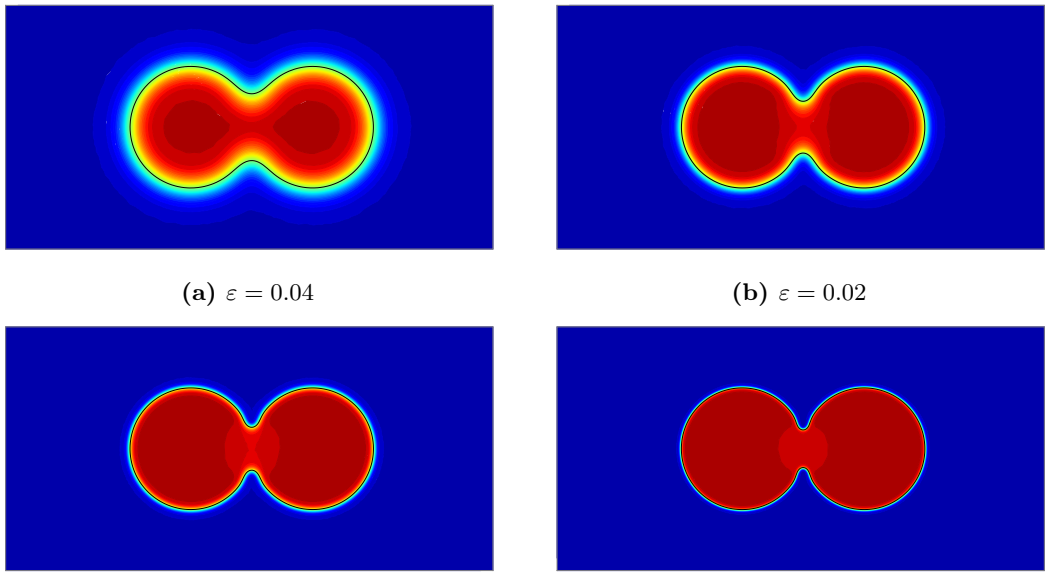
$$V = \frac{1}{2}[\nabla_{\mathbf{n}}\mu]_{\Gamma(t)} \qquad \text{on } \Gamma(t), \text{ for } t \in [0, T], \tag{1.75d}$$

$$\Gamma(0) = \Gamma_0. \tag{1.75e}$$

Here, κ denotes the mean curvature of Γ_t, with the sign convention that convex hypersurfaces have a positive curvature, σ is the interfacial energy, which depends only on F_m and was obtained explicitly in Eq. (1.71) for the Helmoltz potential, and $[\nabla_{\mathbf{n}}\mu]_{\Gamma(t)} = \nabla_{\mathbf{n}}\mu^+ - \nabla_{\mathbf{n}}\mu^-$, where μ^+ and μ^- are the restrictions of μ to Ω_t^+ and Ω_t^-, respectively. The vector $\mathbf{n}$ is the unit outward vector normal to $\partial\Omega$ or $\Gamma(t)$, V is the velocity of the hypersurface $\Gamma(t)$ in the direction of $\mathbf{n}$, and $[\nabla_{\mathbf{n}}\phi]_{\Gamma(t)} = \nabla_{\mathbf{n}}\phi^+ - \nabla_{\mathbf{n}}\phi^-$, where ϕ^- and ϕ^+ denote respectively the restrictions of ϕ to Ω_t^-, the area enclosed by $\Gamma(t)$, and Ω_t^+, its complement.

Intuitively, Eq. (1.75a) corresponds to Eq. (1.67a) when $\partial_t\phi \approx 0$, and the right-hand in Eq. (1.75d) ensures the conservation of mass when the fluxes are mismatched. The curvature term in Eq. (1.75c) shows that, in the sharp interface limit, the solution to the Cahn–Hilliard equation obeys a law similar to the well-known Young–Laplace relation.

Remark 1.8. The fact that the chemical potential approximately satisfies the Laplace equation in the pure phases when $\varepsilon \ll 1$ shows us that, in order to obtain a good overall accuracy, mesh-refinement strategies of numerical methods for the Cahn–Hilliard equation should increase



(a) $\varepsilon = 0.04$ **(b)** $\varepsilon = 0.02$

(c) $\varepsilon = 0.01$ **(d)** $\varepsilon = 0.005$

Figure 1.11: Solution of the Cahn–Hilliard equation at time $t = 1$ for different values of ε. Here we employed the initial condition $\phi_0(x) = -1 + 2\,H(r^2 - (x - x_1)^2 + y^2) + 2\,H(r^2 - (x - x_2)^2 + y^2)$, where $x_1 = -1/4$, $x_2 = 1/4$, $r = 1/4$ and H is the Heaviside function, as well as homogeneous Neumann boundary conditions for both the phase field and the chemical potential.

the resolution not only in the interfacial region but also in the pure phases. ⊘

The first formal derivation of the sharp interface limit was presented by Pego in [179], under the assumption that $\Gamma(t)$ does not intersect the boundaries of the domain. Approximately 5 years later, Alikakos, Bates, and Chen proved this result rigorously in [5], assuming the existence of a smooth solution to the *Hele–Shaw problem* (1.75). To the best of our knowledge, no rigorous sharp interface result exists in the case of non-homogeneous boundary conditions, although some works have investigated the question in the presence of a flow by means of asymptotics and numerical experiments, see e.g. [226, 116, 121].

The sharp interface limit is illustrated in Fig. 1.11, for the case of two droplets coalescing. The domain here is $[-1, 1] \times [-1/2, 1/2]$, and the simulations were started from two circular droplets of radius $1/4$, centered respectively at $(-1/4, 0)$ and $(1/4, 0)$. The mobility parameter employed was equal to 10^{-5}, and the simulations were run from $t = 0$ to $t = 1$, at which point the droplets are partially merged. We observe that, although further analysis of the solutions would be necessary to conclude to the existence of a limit when $\varepsilon \to 0$, the solutions corresponding to different values of ε are qualitatively very similar, aside from the interface thickness.

1.5.4 Coupling with the Navier–Stokes equation

Our work in Chapter 5, on the development of a numerical method for a Cahn–Hilliard equation with wetting boundary condition, is part of a larger effort in the research group of Serafim Kalliadasis to employ phase field models for the study of specific engineering problems in microfluidics. Of particular interest to the research group is the study and design of a *microseparator*, a device enabling the separation, by capillary effects, of the constitutive components of a binary

mixture. Both gas-liquid [192] and liquid-liquid [82] separators have a wide range of applications in engineering and technology. To date, however, most of the research works on the subject have been experimental or theoretical, and there is thus scope for in-depth investigation of the problem using numerical methods.

As is the case with most microfluidic devices, the flow through a typical microseparator is laminar, quasi-incompressible, and heavily influenced by capillary forces. A large part of the difficulty in simulating such multiphase flows originates from the topological changes that the interfaces can undergo during the course of one simulation. In addition to the phase field approach we present in this subsection, commonly-used approaches to the study of multiphase flows include the volume-of-fluid [97] and front-tracking [211, 89] methods, as well as the level-set [40, 166, 206] approach.

In order to account for the fluid flow in the Cahn–Hilliard equation, the standard approach is to include an advection term in the equation for the phase field, effectively replacing the time derivative by a *material derivative*:

$$\frac{\partial \phi}{\partial t} + \mathbf{v} \cdot \nabla \phi = \nabla \cdot (b(\phi)\, \nabla \mu)\,, \tag{1.76a}$$

$$\mu = \frac{1}{\varepsilon}\, f_m(\phi) - \varepsilon\, \Delta \phi. \tag{1.76b}$$

On the other hand, accounting for the phase field in the Navier–Stokes equation, in order to model the well-known *surface tension* and the associated *Laplace pressure*, is more delicate, and several competing models, reviewed in [130], are commonly used in the literature. Here we will present the approach proposed by Jacqmin in his celebrated 1999 paper [120], where the form of the capillary term is obtained by requiring that the variation of free energy, as defined Eq. (1.63), induced by the fluid flow be compensated exactly by a variation of kinetic energy induced by capillary effects; see also [101]. For simplicity of exposition, we will consider two fluids with equal densities and viscosities and we will assume that no exchange of mass or energy, convective or diffusive, occurs through the boundaries of the domain. In particular, we assume that a no-slip boundary condition holds at the boundaries. Let us define the total energy as

$$\mathcal{E}(\phi, \mathbf{v}) = E(\phi) + \lambda\, K(\mathbf{v}), \qquad \text{with} \qquad K(\mathbf{v}) = \int_\Omega \frac{\mathbf{v}^2}{2}\, \mathrm{d}\Omega, \tag{1.77}$$

where $\mathbf{v}$ denotes the velocity field and λ is a physical parameter. The incompressible Navier–Stokes equations, augmented by a placeholder term for the capillary force, read

$$\nabla \cdot \mathbf{v} = 0, \tag{1.78a}$$

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} = -\nabla p + \frac{1}{\mathrm{Re}} \Delta \mathbf{v} + \mathbf{F}_c, \tag{1.78b}$$

where Re is the Reynolds number of the flow, p is the pressure, and $\mathbf{F}_c$ is the capillary force. The

time derivative of the total energy is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\phi,\mathbf{v}) = -\int_\Omega b(\phi)\,|\nabla\mu|^2\,\mathrm{d}\Omega - \frac{\lambda}{\mathrm{Re}}\int_\Omega (\nabla\mathbf{v}:\nabla\mathbf{v})\,\mathrm{d}\Omega \\ -\int_\Omega \mu\,\nabla\phi\cdot\mathbf{v}\,\mathrm{d}\Omega + \lambda\int_\Omega \mathbf{F}_c\cdot\mathbf{v}\,\mathrm{d}\Omega.$$

Therefore we obtain, on requiring that the two terms in the second line balance out,

$$\mathbf{F}_c = \frac{1}{\lambda}\,\mu\,\nabla\phi. \tag{1.79}$$

Together, Eqs. (1.76a), (1.76b), (1.78a), and (1.78b) constitute the Cahn–Hillliard–Navier–Stokes equations. We observe that λ measures the relative importance of inertia compared to capillary effects; in the engineering literature, it is often referred to as the *Weber number*.

Remark 1.9. The capillary force could be replaced by $\frac{1}{\lambda}\phi\,\nabla\mu$, because the difference between the two expressions is exactly the gradient of $\frac{1}{\lambda}\,\phi\,\mu$, which by the continuity equation (1.78a) can be absorbed in the pressure term. Another possibility is to write the capillary force in divergence form, from which the associated stress tensor can be read off: remembering that $\nabla(\nabla u\otimes\nabla u) = \Delta u\,\nabla u + \frac{1}{2}\,\nabla\,|\nabla u|^2$ for any regular enough scalar function u, we have

$$\mu\nabla\phi = \left(\frac{1}{\varepsilon}f_m(\phi) - \varepsilon\,\Delta\phi\right)\nabla\phi = -\varepsilon\,\nabla\cdot(\nabla\phi\otimes\nabla\phi) + \nabla\left(\frac{1}{\varepsilon}F_m(\phi) + \frac{1}{2}\,\varepsilon\,|\nabla\phi|^2\right).$$

Here, the second term in the rightmost equation is the gradient of the free energy density, which can also be absorbed in the pressure term. When $\varepsilon \ll 1$, both the divergence formulation and Eq. (1.79) are appropriate for the calculation of the "true" pressure difference between points located away from interfaces, because the free energy density is approximately zero in those regions (assuming that F_m is zero in the pure phases). ⊘

Example 1.2. Assume that a two-dimensional circular bubble of fluid B, of radius R, is immersed at equilibrium in a domain containing fluid A, and that the evolution of the system is governed by the Cahn–Hilliard–Navier–Stokes equations with the formulation (1.79) of the capillary term. When $\varepsilon \ll 1$, $\mu(x) = \mu \approx \frac{\sigma}{2R}$ for all $x\in\Omega$ at equilibrium, by Eq. (1.75c). Since in addition $\mathbf{v}=0$ at equilibrium, Eq. (1.78b) implies $\nabla p - \frac{1}{\lambda}\,\mu\nabla\phi = 0$, and so the pressure difference between the interior and the exterior of the bubble is approximately equal to

$$p_{\mathrm{in}} - p_{\mathrm{out}} = \frac{\sigma}{\lambda\,R}.$$

Up to a constant factor, this is the familiar equation for the *Laplace pressure* in two dimensions. Since here σ depends only on the model employed for F_m, the only parameter to calibrate in order to approximate the physical surface tension associated with the fluid components is λ. ⊘

1.6 Contributions of the thesis

In Chapter 2, we present a novel methodology for the simulation of multiscale SDEs at the diffusive time scale, based on a Hermite spectral method for the Poisson equation (1.20). Our

analysis extends previous work in [83]. We prove the convergence of the method theoretically and present numerical experiments confirming our findings. We show in particular that, under some regularity and growth assumptions on the confining potential, the numerical error decreases *spectrally* with the degree of approximation, i.e. faster than any inverse polynomial. We also apply the method to multiscale stochastic partial differential equations (SPDEs), and we show how it fares in that case compared to the *heterogeneous multiscale method*, another popular method for multiscale SDEs [65, 1]. This chapter is based in most parts on [2]; it resulted from a collaboration with Assyr Abdulle and Grigorios Pavliotis.

In Chapter 3, we study systems of particles driven by colored noise in a bistable potential, interacting through their mean via a Curie–Weiss (quadratic) potential. Our main aim in this chapter is to study the influence of the correlation time of the noise on the phase transitions that the system can undergo following a change of temperature, and in particular to determine whether well-known results in the case of white noise [51, 196] also apply in the presence of colored noise. We begin by showing formally that, as with the case of white noise, a nonlinear and nonlocal Fokker–Planck equation, known in the literature as a *McKean–Vlasov equation*, can be obtained for the 1-particle probability density in the mean-field limit. We then study phase transitions of the system, both by direct solution of the McKean–Vlasov equation using a Hermite spectral method and by means of Monte–Carlo simulation of the particle system. Although related to the method presented in Chapter 2, the spectral method we use in this chapter exhibits original features. First, the Galerkin discretization is employed in conjunction with a semi-implicit linear scheme for the time-dependent McKean–Vlasov equation, and, second, it is based on the systematic use of *generalized Hermite functions*, which enable tailoring the finite-dimensional space of approximation to the problem under consideration. In order to validate our numerical results in the small correlation time regime, we also construct an approximate bifurcation diagram, of the first moment of the steady-state solution against the temperature, using techniques from [115] on singular perturbation theory for Markov processes. At the physical level, our findings reveal that the correlation time of the noise generally does not influence the topology of the bifurcation diagram, unless the noise has an asymmetric distribution. It does, however, influence the temperature at which new branches emerge in the bifurcation diagram. The material presented in this chapter is the product of a collaboration with Susana Gomes and Grigorios Pavliotis; it is based in most parts on [95].

In Chapter 4, we study the diffusion in a periodic potential of a Brownian particle whose motion is governed by the *generalized Langevin equation* (GLE) in one dimension. Our work in this chapter is based on previous work in [178, 167]. We begin by approaching the problem theoretically using classical techniques from multiscale analysis [175]. We study three limits of particular interest, namely the small correlation time limit, in which the solution of the GLE converges to that of the Langevin equation, with an effective friction determined by the autocorrelation function of the noise, as well as the overdamped and underdamped limits, corresponding to the effective friction coefficient going to $+\infty$ and 0, respectively. We show in particular that, while in the overdamped limit the leading order term in the asymptotic expansion of the effective diffusion coefficient is independent of the model of the noise, this does not hold in the underdamped limit; our analysis does suggest, however, that the scaling of the leading-order

term in this limit is model-independent. We also obtain quantitative lower bounds on the rate of convergence to equilibrium using the $H^1(\mu)$ framework developed in [220]. In the second part of this chapter, we confirm our findings numerically using a mixed Fourier/Hermite spectral method based on the work in [190], thereby corroborating and extending the results of the early studies [118, 119]. The research that lead to this chapter was conducted under the supervision of Grigorios Pavliotis.

In Chapter 5, we present a novel finite element method for the solution of the Cahn–Hilliard equation with nonlinear wetting boundary condition. Our method builds on the linear schemes proposed in [98] for the Cahn–Hilliard equation with homogeneous Neumann boundary condition. It is semi-implicit and linear, and it satisfies discrete mass and energy conservation laws similar to the ones satisfied by the exact solution. Instead of the cubic boundary condition traditionally used in the literature [132], we propose an alternative boundary condition that coincides with the cubic on the physical interval for the phase field but enables proving the existence of a solution to the continuous problem. Through careful numerical experiments, we verify the rates of convergence with respect to both the time step and the mesh size. We also demonstrate the flexibility and robustness of our method by applying it to test cases characterized by complex geometries or the presence of chemical heterogeneities. We conclude by showing that the method can be coupled to a numerical solver for the Navier–Stokes equation, enabling the study of phase separation within a fluid flow. This chapter is based in most parts on [11]; it is the result of joint work with Benjamin Aymard, Marc Pradas and Serafim Kalliadasis.

Finally in Chapter 6, we present the free software contributions [216, 217] that emerged from the research presented in this thesis. First we introduce *Hermipy*, a *Python* library for the numerical solution of PDEs in unbounded or periodic domains via Fourier/Hermite spectral methods. *Hermipy* aims to simplify the use of Fourier/Hermite spectral methods for researchers, by providing high-level functions and classes abstracting most of the technical implementation details of these methods, much like *FEniCS* [146] and *FreeFem++* [106] endeavored for the finite element method. It enables calculating the solution of PDEs efficiently and robustly with just a few lines of codes, and can read input functions or operators from either strings or *SymPy* [155] expressions. In order to be as user-friendly as possible, *Hermipy* relies heavily on the object-oriented paradigm available in *Python*; operators and functions, for example, as well as their discrete representations, namely stiffness matrices and Fourier/Hermite series, are all represented by *Python* objects. It also takes advantage of the fact that, at the discrete level, some operators and functions can be obtained as tensor products of lower-dimensional operators and functions, respectively. When applicable, this enables their fast discretization even in high dimensions, which would otherwise require a prohibitively high computational cost. Finally, *Hermipy* offers the possibility of working with *generalized* (Fourier or Hermite) basis functions, i.e. basis functions obtained from the standard Hermite polynomials or Fourier modes by spatial rescaling, multiplication by an appropriate (space-dependent) weight, or both. This enables, in particular, a similar treatment of Kolmogorov backward and forward equations, although the corresponding solutions live in different functional spaces.

Chapter 2

Hermite spectral method for multiscale SDEs

Abstract

This chapter presents a new method for the solution of multiscale stochastic differential equations at the diffusive time scale. In contrast to averaging-based methods, e.g., the heterogeneous multiscale method (HMM) or the equation-free method, which rely on Monte Carlo simulations, in this chapter we introduce a new numerical methodology that is based on a spectral method. In particular, we use an expansion in Hermite functions to approximate the solution of an appropriate Poisson equation, which is used in order to calculate the coefficients of the homogenized equation. Spectral convergence is proved under suitable assumptions. Numerical experiments corroborate the theory and illustrate the performance of the method. A comparison with the HMM and an application to singularly perturbed stochastic PDEs are also presented.

2.1 Introduction 70
2.2 Diffusion Approximation and Homogenization 74
2.3 Numerical Method 76
2.4 Main Results 80
2.5 Proofs of the Main Results 83
2.5.1 Convergence of the Spectral Method for the Poisson Equation 83
2.5.2 Convergence of the Drift and Diffusion Coefficients 85
2.5.3 Convergence of the Solution to the SDE 86
2.6 Implementation and Numerical Experiments 88
2.6.1 Implementation details 88
2.6.2 Numerical experiments 90
2.7 Conclusion and Further Work 101
2.A Proof of the Lipschitz continuity of the effective coefficients 102

2.1 Introduction

Multiscale stochastic systems arise frequently in applications. Examples include atmosphere/ocean science [151] and materials science [64]. For systems with a clear scale separation it is possible, in principle, to obtain a closed – averaged or homogenized – equation for the slow variables [175]. The calculation of the drift and diffusion coefficients that appear in this effective (coarse-grained) equation requires appropriate averaging over the fast scales. Several numerical methods for multiscale stochastic systems that are based on scale separation and on the existence of a coarse-grained equation for the slow variables have been proposed in the literature. Examples include the heterogeneous multiscale method (HMM) [218, 65, 3] and the equation-free approach [128]. These techniques are based on evolving the coarse-grained dynamics, while calculating the drift and diffusion coefficients "on-the-fly" using short simulation bursts of the fast dynamics.

A prototype fast/slow system of stochastic differential equations (SDEs) for which the aforementioned techniques can be applied is [1]

$$\mathrm{d}X_t^\varepsilon = \frac{1}{\varepsilon}\mathbf{f}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t + \sqrt{2}\,\boldsymbol{\sigma}_x\,\mathrm{d}W_x(t), \tag{2.1a}$$

$$\mathrm{d}Y_t^\varepsilon = \frac{1}{\varepsilon^2}\mathbf{h}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}t + \frac{\sqrt{2}}{\varepsilon}\boldsymbol{\sigma}_y\,\mathrm{d}W_y(t). \tag{2.1b}$$

where $X_t^\varepsilon \in \mathbf{R}^m$, $Y_t^\varepsilon \in \mathbf{R}^n$, $\varepsilon \ll 1$ is the parameter measuring scale separation, $\boldsymbol{\sigma}_x \in \mathbf{R}^{m\times d_1}$, $\boldsymbol{\sigma}_y \in \mathbf{R}^{n\times d_2}$ are constant matrices, and W_x, W_y are independent d_1 and d_2-dimensional Brownian motions, respectively.[2] For fast-slow systems of this form, a direct numerical approximation of the full dynamics would be prohibitively expensive, because resolving the fine scales would require a time step δt that scales as $\mathcal{O}(\varepsilon^2)$. Under appropriate assumptions on the coefficients and on the ergodic properties of the fast process Y_t^ε, it is well-known that the slow process converges, in the limit as ε tends to 0, to a homogenized equation that is independent of the fast process and of ε [175, Ch. 11]:

$$\mathrm{d}X_t = \mathbf{F}(X_t)\,\mathrm{d}t + \mathbf{A}(X_t)\,\mathrm{d}W_t. \tag{2.2}$$

The drift and diffusion coefficients in (2.2) can be calculated by solving a Poisson equation involving the generator of the fast process,[3]

$$-\mathcal{L}_y\boldsymbol{\phi} = \mathbf{f}, \tag{2.3}$$

where $\mathcal{L}_y = \mathbf{h}(x,y)\cdot\nabla_y + \boldsymbol{\sigma}_y\boldsymbol{\sigma}_y^T : \nabla_y\nabla_y$, together with appropriate boundary conditions, and

[1] In this chapter we will consider the fast/slow dynamics at the diffusive time scale, or, using the terminology of [175], the homogenization problem.

[2] It is straightforward to consider problems where the Brownian motions driving the fast and slow processes are correlated. This scenario might be relevant in applications to mathematical finance. See e.g. [49].

[3] We are assuming that the centering condition is satisfied, see Eq. (H_f) below.

calculating averages with respect to the invariant measure $\mu_x(\mathrm{d}y)$ of Y_t^ε:

$$\mathbf{F}(x) = \int_{\mathbf{R}^n} \nabla_x \boldsymbol{\phi}(x,y)\, \mathbf{f}(x,y)\, \mu_x(\mathrm{d}y), \tag{2.4a}$$

$$\mathbf{A}(x)\mathbf{A}(x)^T = \int_{\mathbf{R}^n} \left[\mathbf{f}(x,y) \otimes \boldsymbol{\phi}(x,y) + \boldsymbol{\phi}(x,y) \otimes \mathbf{f}(x,y)\right]\, \mu_x(\mathrm{d}y). \tag{2.4b}$$

Once the drift and diffusion coefficients have been calculated, then it becomes computationally advantageous to solve the homogenized equations, in particular since we are usually interested in the evolution of observables of the slow process alone. The main computational task, thus, is to calculate the drift and diffusion coefficients that appear in the homogenized equation (2.2). When the state space of the fast process is high dimensional, the numerical solution of the Poisson equation and calculation of the integrals in (2.4) using deterministic methods become prohibitively expensive and Monte Carlo-based approaches have to be employed. In recent years different methodologies have been proposed for the numerical solution of the fast-slow system (2.1) that are based on the strategy outlined above, for example the Heterogeneous Multiscale Method (HMM) [218, 65, 3] and the equation-free approach [128]. In particular, the PDE-based formulas (2.4) are replaced by Green–Kubo type formulas [65, Sec. 1] that involve time averages and numerically calculated autocorrelation functions. The equivalence between the homogenization and the Green–Kubo formalism has been shown for a quite general class of fast/slow systems of SDEs [173]. See also [140, 142]. Although they offer several advantages, these methods, being based on time and ensemble averages, are computationally very expensive for the calculation of accurate solutions. Based on the analysis of [65], one deduces that the computational cost needed to obtain an error of order 2^{-p} scales as $\mathcal{O}(2^{p(2+1/l)})$, where l is the weak order of accuracy of the micro-solver used.

When the dimension of the state space of the fast process is relatively low, numerical approaches that are based on the accurate and efficient numerical solution of the Poisson equation (2.3) using "deterministic" techniques become preferable. Such an approach was taken in [28] for the study of the diffusion approximation of a kinetic model for swarming [38]. In dimensionless variables, the equation for the distribution function $f^\varepsilon(x,v,t)$ reads

$$\frac{\partial f^\varepsilon}{\partial t} + \frac{1}{\sqrt{\varepsilon}}(v \cdot \nabla_r f^\varepsilon - \nabla_r \Psi \cdot \nabla_v f^\varepsilon) = \frac{1}{\varepsilon} Q(f^\varepsilon),$$

where Ψ is a potential that is defined self-consistently through the solution of a Poisson equation, $Q(\cdot)$ denotes a linearized "collision" operator, with the appropriate number and type of collision invariants. It was shown in [28] that in the limit as ε tends to 0, the spatial density $\rho(x,t) = \int f(x,v,t)\,\mathrm{d}v$ of swarming particles converges to the solution of an aggregation-diffusion equation of the form

$$\frac{\partial \rho}{\partial t} - \nabla \cdot (\mathcal{D}\nabla\rho + \mathcal{K}(\nabla U \star \rho)\rho) = 0, \tag{2.5}$$

where $\star$ denotes the convolution product, U is the interaction potential, and the drift and diffusion tensors $\mathcal{K}$ and $\mathcal{D}$, respectively, can be calculated using an approach identical to (2.3) and (2.4):

we first have to solve the Poisson equations[4]

$$-\mathcal{H}u_\chi = v\sqrt{M} \quad \text{and} \quad -\mathcal{H}u_\kappa = \frac{1}{\theta}\nabla_v W\sqrt{M}, \tag{2.6}$$

where $W(\cdot)$ is a potential in velocity, $M(v) = Z^{-1}\,\mathrm{e}^{-W(v)/\theta}$ is the Maxwellian distribution at temperature θ, with Z being the normalization constant, $\mathcal{H} = -\theta\Delta_v + \Phi(v)$ and

$$\Phi(v) = -\frac{1}{2}\Delta_v W(v) + \frac{1}{4\theta}\left|\nabla_v W(v)\right|^2. \tag{2.7}$$

Then the effective coefficients can be calculated by the integrals

$$\mathcal{D} = \int_{\mathbf{R}^d} \mathcal{H}(u_\chi)\otimes u_\chi \,\mathrm{d}v \quad \text{and} \quad \mathcal{K} = \int_{\mathbf{R}^d} \mathcal{H}(u_\chi)\otimes u_\kappa \,\mathrm{d}v. \tag{2.8}$$

We note that the operator $\mathcal{H}$ that appears in (2.6) is a Schrödinger operator whose spectral properties are very well understood [184, 113]. In particular, under appropriate growth assumptions on the potential Φ given in (2.7), the operator $\mathcal{H}$ is essentially selfadjoint, has discrete spectrum and its eigenfunctions form an orthonormal basis in $L^2\left(\mathbf{R}^d\right)$. The computational methodology that was introduced and analyzed in [28] for calculating the homogenized coefficients in (2.5) is based on the numerical calculation of the eigenvalues and eigenfunctions of an appropriate Schrödinger operator using a high-order finite element method. The authors showed rigorously that for sufficiently smooth interaction potentials the proposed numerical scheme performs extremely well; in particular, the numerical calculation of the first few eigenvalues and eigenfunctions are sufficient for the very accurate calculation of the drift and diffusion coefficients.

Methods that are based on the caclculation of the eigenvalues and eigenfunctions of the transfer (Koopman) operator have been introduced in [79, 80]. In addition diffusion maps have also been applied to multiscale stochastic systems [45, 44]. Techniques that are based on the transfer operator and diffusion maps are related to our approach since, unlike the HMM and the equation-free approach, they do not require the solution of the fast dynamics. On the other hand it should be noted that these techniques are, in principle, more powerful than the methodology developed in this chapter, since they enable us to identify the coarse-grained variables.

In this chapter we develop further the methodology introduced in [28] and we apply it to the numerical solution of fast/slow systems of SDEs, including singularly perturbed stochastic partial differential equations (SPDEs) in bounded domains. Thus, we complement the work presented in [1], in which a hybrid HMM/spectral method for the numerical solution of singularly perturbed SPDEs with quadratic nonlinearities [23] at the diffusive time scale was developed.[5] The main difference between the methodology presented in [28] and the approach we take in this chapter is that, rather than obtaining the orthonormal basis by solving the eigenvalue problem for an appropriate Schrödinger operator, we fix the orthonormal basis (Hermite functions) and expand

[4] We first perform a unitary transformation that maps the generator of a diffusion process of the form $\mathcal{L}_y$ that appears in (2.3) to an appropriate Schrödinger-type operator; see [174, Sec. 4.9] for details.

[5] When the centering condition (see Equation (H_f)) is not satisfied, one needs to study the problem at a shorter time scale (called the advective time scale) before considering the diffusive limit. This problem is easier to study since it does not require the solution of a Poisson equation. The rigorous analysis of the HMM method for singularly perturbed SPDEs at the advective time scale was presented in [32].

the solution of the Poisson equation (2.3) (after the unitary transformation that maps it to an equation for a Schrödinger operator) in this basis. We show rigorously and by means of numerical experiments that our proposed methodology achieves spectral convergence for a wide class of fast processes in (2.1). Consequently, our method outperforms Monte Carlo-based methodologies such as the HMM and the equation-free method, at least for problems with low-dimensional fast processes. We discuss how our method can be modified so that it becomes efficient when the fast process has a high-dimensional state space in the conclusions section, Section 2.7.

In this chapter we will consider fast/slow systems of SDEs for which the fast process is reversible, i.e. it has a gradient structure [174, Sec. 4.8][6]

$$\mathrm{d}X_t^\varepsilon = \frac{1}{\varepsilon}\mathbf{f}(X_t^\varepsilon, Y_t^\varepsilon)\mathrm{d}t + \boldsymbol{\alpha}(X_t^\varepsilon, Y_t^\varepsilon)\,\mathrm{d}W_x(t), \qquad X_0^\varepsilon = x_0, \tag{2.9a}$$

$$\mathrm{d}Y_t^\varepsilon = -\frac{1}{\varepsilon^2}\nabla V(Y_t^\varepsilon)\mathrm{d}t + \frac{\sqrt{2}}{\varepsilon}\mathrm{d}W_y(t), \qquad Y_0^\varepsilon = y_0, \tag{2.9b}$$

where $X_t^\varepsilon(t) \in \mathbf{R}^m$, $Y_t^\varepsilon(t) \in \mathbf{R}^n$, $\boldsymbol{\alpha}(\cdot,\cdot) \in \mathbf{R}^{m\times p}$, W_x and W_y are standard p and n-dimensional Brownian motions, and $V(\cdot)$ is a smooth confining potential.

Remark 2.1. Our algorithm applies to the more general situation when the fast dynamics depends on the slow one in the following way

$$\mathrm{d}Y_t^\varepsilon = -\frac{1}{\varepsilon^2}\nabla V(Y_t^\varepsilon)\mathrm{d}t + \frac{1}{\varepsilon}\mathbf{h}(X_t^\varepsilon(t), Y_t^\varepsilon(t))\,\mathrm{d}t + \frac{\sqrt{2}}{\varepsilon}dW_y(t), \quad Y_0^\varepsilon = y_0.$$

The analysis presented in this chapter also generalizes with minor changes for such a fast process and the effective drift (2.4a) becomes

$$\mathbf{F}(x) = \int_{\mathbf{R}^n} \nabla_x \boldsymbol{\phi}(x,y)\,\mathbf{f}(x,y)\,\mu_x(\mathrm{d}y) + \int_{\mathbf{R}^n} \nabla_y \boldsymbol{\phi}(x,y)\,\mathbf{h}(x,y)\,\mu_x(\mathrm{d}y).$$

Hence the additional term appearing above involves the solution of a Poisson problem already computed. ⊘

SDEs of this form appear in several applications, e.g. in molecular dynamics [60, 141]. Furthermore, several interesting semilinear singularly perturbed SPDEs can be written in this form, see Section 2.6. It is well-known [174, Sec. 4.9] that the generator of a reversible SDE is unitarily equivalent to an appropriate Schrödiner operator. Consequently, the calculation of the drift and diffusion coefficients in the homogenized equation corresponding to (2.9) reduces to a problem that is very similar to (2.6) and (2.8). Our approach is to first solve this Poisson equation for the Schrödinger operator via a spectral method using Hermite functions and then use this solution in order to calculate the integrals in (2.4). For smooth potentials that increase sufficiently fast at infinity our method has spectral accuracy, i.e. the error decreases faster than any negative power of the number of floating point operations performed. This, in turn, via a comparison for SDEs argument, implies that we can approximate very accurately the evolution of observables of the slow variable X_t^ε in (2.9) by solving an approximate homogenized equation in which the drift and diffusion coefficients are calculated using our spectral method. For relatively

[6] We could, in principle, also consider reversible SDEs with a diffusion tensor that is not a multiple of the identity.

low dimensional fast-processes, this leads to a much more accurate and computationally efficient numerical method than any Monte Carlo-based methodology. We remark that our proposed numerical methodology becomes (analytically) exact when the fast process is, to leading order, an Ornstein–Uhlenbeck process, since in this case, for a suitable choice of the mean and the covariance matrix, the Hermite functions are the eigenfunctions of the corresponding Schrödinger operator.

The rest of the chapter is organized as follows. In Section 2.2, we summarize the results from homogenization theory for the fast/slow system (2.9) that we will need in this work. In Section 2.3 we present our numerical method in an algorithmic manner. In Section 2.4, we summarize the main theoretical results of this chapter; in particular we show that our method, under appropriate assumptions on the coefficients of the fast/slow system, is spectrally accurate. The proofs of our main results are given in Section 2.5. In Section 2.6 we present details on the implementation of our numerical method, discuss the computational efficiency and present several numerical examples, including an example of the numerical solution of a singularly perturbed SPDE; for this example, we also present a brief qualitative comparison of our method with the HMM method. Section 2.7 is reserved for conclusions and discussion of further work. Finally in the appendices we present some results related to approximation theory in weighted Sobolev spaces that are needed in the proof of the main convergence theorem.

2.2 Diffusion Approximation and Homogenization

In this section, we summarize some of our working hypotheses and the results from the theory of homogenization used to derive the effective SDE for the system (2.9). Throughout this chapter, the notation $|\cdot|$ denotes the Euclidian norm when applied to vectors or the 2-norm when applied to matrices. In addition, we denote the components of a vector $\mathbf{v} \in \mathbf{R}^d$ by $v_1, v_2 \cdots, v_d$. We start by assuming that $V(\cdot)$ is a smooth confining potential, [174, Definition 4.2]:

$$V \in C^\infty(\mathbf{R}^n), \qquad \lim_{|y|\to\infty} V(y) = \infty \quad \text{and} \quad \mathrm{e}^{-V(\cdot)} \in L^1\left(\mathbf{R}^n\right). \tag{H_V}$$

These hypotheses guarantee that the fast process has a well defined solution for all positive times, with a unique invariant measure whose density is given by $\frac{1}{\mathcal{Z}}\,\mathrm{e}^{-V(\cdot)}$, where $\mathcal{Z}$ is the normalization constant. Without loss of generality, we may assume that $\mathcal{Z} = 1$. To these assumptions, we add

$$\lim_{|y|\to\infty} \nabla V \cdot \left(\frac{y}{|y|}\right) = \infty \quad \text{and} \quad \lim_{|y|\to\infty} W(y) := \lim_{|y|\to\infty} \left(\frac{1}{4}|\nabla V(y)|^2 - \frac{1}{2}\Delta V(y)\right) = \infty, \tag{H_W}$$

which guarantee that the law of Y_t^ε converges to its invariant distribution e^{-V} exponentially fast (e.g. in relative entropy), see [152]. We assume furthermore that the drift coefficient in the slow

equation of system (2.9) satisfies

$$\begin{aligned}&\mathbf{f}(x,y)\in(C^\infty(\mathbf{R}^m\times\mathbf{R}^n))^m,\\&\int_{\mathbf{R}^n}\mathbf{f}(x,y)\,\mathrm{e}^{-V(y)}\,\mathrm{d}y=\mathbf{0},\ \text{and}\\&|\mathbf{f}(x,y)|\le K(1+|x|)(1+|y|^\beta)\quad\forall x\in\mathbf{R}^m\text{ and }\forall y\in\mathbf{R}^n,\end{aligned}\tag{H_f}$$

with β a positive integer and K a positive constant. Under Assumptions (H_V) and (H_W), we can use Proposition A.12 and Lax–Milgram theorem to show that there exists for all $x\in\mathbf{R}^m$ a unique $\boldsymbol{\phi}(x,\cdot)\in H^1(\mathbf{R}^n,\mathrm{e}^{-V})^m$, where $H^1(\mathbf{R}^n,\mathrm{e}^{-V})$ is the weighted Sobolev space defined in Definition A.7, such that

$$-\mathcal{L}_y\phi_i(x,y):=-\left(\Delta_y-\nabla_yV\cdot\nabla_y\right)\phi_i(x,y)=f_i(x,y)\quad\text{for } i=1,\dots,m\tag{2.10}$$

and

$$\int_{\mathbf{R}^n}\boldsymbol{\phi}(x,y)\,\mathrm{e}^{-V(y)}\,\mathrm{d}y=\mathbf{0},\quad\forall x\in\mathbf{R}^m.$$

By the theory of elliptic regularity, the uniform ellipticity of $\mathcal{L}$ implies that $\boldsymbol{\phi}$ is smooth in y, and by [170, Theorem 1], $\boldsymbol{\phi}$ grows at most polynomially in y.

We also make the following assumption on the Lipschitz continuity with respect to x of the coefficients,

$$\left|\mathbf{f}(x,y)-\mathbf{f}(x',y)\right|+\left|\boldsymbol{\alpha}(x,y)-\boldsymbol{\alpha}(x',y)\right|\le C(y)\left|x-x'\right|,\quad\forall x,x'\in\mathbf{R}^m,\tag{H_L}$$

with $C(y)$ a function bounded from above by a polynomial, and the following assumptions on the growth of the coefficients:

$$\begin{aligned}&|\nabla_x\mathbf{f}(x,y)|+\left|\nabla_x^2\mathbf{f}(x,y)\right|\le K(1+|y|^{m_1}),\\&|\boldsymbol{\alpha}(x,y)|\le K(1+|x|^{1/2})(1+|y|^{m_2}),\end{aligned}\tag{H_G}$$

for positive integers m_1,m_2 and a positive constant K. It follows from this that $\boldsymbol{\phi}(\cdot,y)$ belongs to $(C(2)\mathbf{R}^m)^m$ for all values of y. This can be shown by using the Feynman–Kac representation of the solution of (2.10) that was studied in [170]:

$$\phi_i(x,y)=\int_0^\infty\mathbf{E}_yf_i(x,z_t^y)\,\mathrm{d}t,\qquad i=1,\dots,m,\tag{2.11}$$

where z_t^y is the solution of

$$\mathrm{d}z_t^y=-\nabla_yV(z_t^y)\,\mathrm{d}t+\sqrt{2}\,\mathrm{d}W_t\quad\text{with}\quad z_0^y=y.$$

Using the Feynman–Kac formula (2.11), one can show [170, p. 1073] that there exist $L,q>0$ such that:

$$\begin{aligned}&|\boldsymbol{\phi}(x,y)|+|\nabla_y\boldsymbol{\phi}(x,y)|\le L(1+|x|)(1+|y|^q),\\&|\nabla_x\boldsymbol{\phi}(x,y)|+|\nabla_y\nabla_x\boldsymbol{\phi}(x,y)|+|\nabla_x\nabla_x\boldsymbol{\phi}(x,y)|+|\nabla_y\nabla_x\nabla_x\boldsymbol{\phi}(x,y)|\le L(1+|y|^q).\end{aligned}\tag{2.12}$$

Using the previous assumptions it is possible to prove the following homogenization/diffusion approximation result [170, Theorem 3].

Theorem 2.1. *Let Eqs. (H_V) to (H_f), (H_L) and (H_G) be satisfied. Then for any $T > 0$, the family of processes $\{X_t^\varepsilon, 0 \leq t \leq T\}$ solving (2.9) is weakly relatively compact in $(C()[0,T])^m$. Any accumulation point X_t is a solution of the martingale problem associated to the operator:*

$$\mathcal{G} = \frac{1}{2}\mathbf{D}(x) : \nabla_x \nabla_x + \mathbf{F}(x) \cdot \nabla_x$$

where

$$\mathbf{F}(x) = \int_{\mathbf{R}^n} \nabla_x \boldsymbol{\phi}(x,y)\, \mathbf{f}(x,y)\, \mathrm{e}^{-V(y)}\, \mathrm{d}y, \tag{2.13}$$

and

$$\mathbf{D}(x) = \int_{\mathbf{R}^n} \left(\boldsymbol{\alpha}(x,y)\boldsymbol{\alpha}(x,y)^T + \mathbf{f}(x,y) \otimes \boldsymbol{\phi}(x,y) + \boldsymbol{\phi}(x,y) \otimes \mathbf{f}(x,y)\right) \mathrm{e}^{-V(y)}\, \mathrm{d}y, \tag{2.14}$$

and $\boldsymbol{\phi}(x,y)$ is the centered solution of the Poisson equation (2.10). If, moreover, the martingale problem associated to $\mathcal{G}$ is well-posed, then $X_t^\varepsilon \Rightarrow X_t$ (convergence in law), where X_t is the unique diffusion process (in law) with generator $\mathcal{G}$.

The matrix $\mathbf{D}(x)$ is clearly symmetric, and one can show that it is also positive semi-definite, see [175, Section 11.5]. In view of Theorem 2.1, writing $\mathbf{D}(x) = \mathbf{A}(x)\mathbf{A}(x)^T$ we obtain the functions $\mathbf{F}(x)$, $\mathbf{A}(x)$ that appear in the homogenized SDE (2.2). Though the choice of $\mathbf{A}(x)$ is not unique, all choices lead to the same associated Fokker–Planck or backward Kolmogorov equations, hence the probability measures in path space are the same.

2.3 Numerical Method

In this section, we describe our method for the approximation of the effective dynamics, the analysis of which is postponed to Section 2.5. We start by introducing the necessary notation. We will denote by $L^2(\mathbf{R}^n)$ the space of square integrable functions on $\mathbf{R}^n$. The notation $L^2(\mathbf{R}^n, \rho)$, for a probability density ρ, will be used to denote the space of functions f such that $\sqrt{\rho} f \in L^2(\mathbf{R}^n)$. Weighted Sobolev spaces associated to a probability density are defined in Definition A.7. whereas scales of Sobolev spaces, associated to an operator, are defined in Definition A.11.

In addition to these function spaces, we will denote by $\mathbf{P}(d)(\mathbf{R}^n)$ the space of polynomials in n variables of degree less than or equal to d, and by $\mathbf{H}_\alpha(y; \mu, \Sigma)$ the Hermite polynomials on $\mathbf{R}^n$ defined in Appendix A.3:

$$\mathbf{H}_\alpha(y;\mu,\Sigma) = \mathbf{H}_\alpha(S^{-1}(y-\mu); 0, I), \text{ with } \mathbf{H}_\alpha(z;0,I) = \prod_{k=1}^n H_{\alpha_k}(z_k) \tag{2.15}$$

and $\alpha \in \mathbf{N}^n$. Here $H_{\alpha_k}(\cdot)$ denotes the one-dimensional Hermite polynomial of degree α_k, $\Sigma \in \mathbf{R}^{n\times n}$ is a symmetric positive definite matrix, D and Q are diagonal and orthogonal matrices such that $\Sigma = QDQ^T$, $S = QD^{1/2}$ and $\mu \in \mathbf{R}^n$. We recall from Appendix A.3 that these polynomials form a complete orthonormal basis of $L^2(\mathbf{R}^n, g_{\mu,\Sigma})$, where $g_{\mu,\Sigma}$ denotes the Gaussian

density on $\mathbf{R}^n$ with mean μ and covariance matrix Σ. Finally, we will use the notation $\mathbf{h}_\alpha(y;\mu,\Sigma)$ to denote the Hermite functions corresponding to the Hermite polynomials (2.15), see Eq. (A.28).

We recall from Section 2.2 that obtaining the drift and diffusion coefficients $\mathbf{F}(X)$ and $\mathbf{A}(X)$, respectively, of the homogenized equation

$$\mathrm{d}X = \mathbf{F}(X)\,\mathrm{d}t + \mathbf{A}(X)\,\mathrm{d}W_t, \tag{2.16}$$

requires the solution of the Poisson equations (2.10). To emphasize the fact that x appears as a parameter in (2.10), we will use the notations $\boldsymbol{\phi}^x(\cdot) := \boldsymbol{\phi}(x,\cdot)$ and $\mathbf{f}^x(\cdot) := \mathbf{f}(x,\cdot)$. The weak formulation of the Poisson equation (2.10) is to find $\boldsymbol{\phi}^x \in H^1(\mathbf{R}^n, \mathrm{e}^{-V})^m$ such that for $i = 1,\ldots,m$,

$$a_V(\phi_i^x, v) := \int_{\mathbf{R}^n} \nabla\phi_i^x \cdot \nabla v \; \mathrm{e}^{-V}\,\mathrm{d}y = \int_{\mathbf{R}^n} f_i^x \, v \; \mathrm{e}^{-V}\,\mathrm{d}y \qquad \forall v \in H^1(\mathbf{R}^n, \mathrm{e}^{-V}), \tag{2.17}$$

with the centering condition

$$\mathcal{M}(\boldsymbol{\phi}^x) := \int_{\mathbf{R}^n} \boldsymbol{\phi}^x \; \mathrm{e}^{-V} \; \mathrm{d}y = \mathbf{0}. \tag{2.18}$$

We recall that in order to be well-posed the condition $\mathcal{M}(\mathbf{f}^x) = 0$ must be satisfied.

We begin by performing the standard unitary transformation that maps the generator of a reversible Markov process to a Schrödinger operator: $\mathrm{e}^{-V/2} : L^2\left(\mathbf{R}^n, \mathrm{e}^{-V}\right) \to L^2\left(\mathbf{R}^n\right)$. Introducing

$$\mathcal{H} := \mathrm{e}^{-V/2}\,\mathcal{L}_y\left(\mathrm{e}^{V/2}\,\cdot\right) = \Delta - \left(\frac{1}{4}|\nabla V|^2 - \frac{1}{2}\Delta V\right) = \Delta - W(y), \tag{2.19}$$

and $\boldsymbol{\psi}^x = \mathrm{e}^{-V/2}\,\boldsymbol{\phi}^x$, the Poisson equation (2.10) can be rewritten in terms of the operator (2.19) as:

$$-\mathcal{H}\boldsymbol{\psi}^x = \mathrm{e}^{-V/2}\,\mathbf{f}^x. \tag{2.20}$$

Note that, because the superscript x already indicates that x plays the role of a parameter, we have have not appended the subscript y to $\mathcal{H}$. The weak formulation of this mapped problem reads: find $\boldsymbol{\psi}^x \in H^1(\mathbf{R}^n, \mathcal{H})$ such that, for $i = 1,\ldots,m$,

$$a(\psi_i^x, v) := \int_{\mathbf{R}^n} \nabla\psi_i^x \cdot \nabla v + W(y)\,\psi_i^x\,v\,\mathrm{d}y = \int_{\mathbf{R}^n} f_i^x \, v \; \mathrm{e}^{-V/2}\,\mathrm{d}y \qquad \forall v \in H^1(\mathbf{R}^n, \mathcal{H}), \tag{2.21}$$

where $H^1(\mathbf{R}^n, \mathcal{H}) = \left\{u \in H^1(\mathbf{R}^n) : \int_{\mathbf{R}^n} |W|\,u^2\,\mathrm{d}y < \infty\right\}$, and such that the following centering condition is satisfied:

$$\hat{\mathcal{M}}(\boldsymbol{\psi}^x) := \int_{\mathbf{R}^n} \boldsymbol{\psi}^x\,\mathrm{e}^{-V/2}\,\mathrm{d}y = 0. \tag{2.22}$$

The formulas for the effective drift and diffusion coefficients can be written as

$$\mathbf{F}(x) = \int_{\mathbf{R}^n} \nabla_x \boldsymbol{\psi}^x \left(\mathbf{f}^x \, \mathrm{e}^{-V/2}\right) \mathrm{d}y, \tag{2.23a}$$

$$\mathbf{D}(x) = \int_{\mathbf{R}^n} \boldsymbol{\alpha}\boldsymbol{\alpha}^T(x,y)\, \mu^x(\mathrm{d}y) + \mathbf{A}_0(x) + \mathbf{A}_0(x)^T, \tag{2.23b}$$

where

$$\mathbf{A}_0(x) = \int_{\mathbf{R}^n} \boldsymbol{\psi}^x \otimes \left(\mathbf{f}^x \, \mathrm{e}^{-V/2}\right) \mathrm{d}y.$$

The advantage of using the unitary transformation is that the solution of this new problem and its derivative lie in $L^2(\mathbf{R}^n)$, rather than in a weighted space.

To approximate numerically the coefficients of the effective SDE, we choose a finite-dimensional subspace $\hat{S}_d$ of $H^1(\mathbf{R}^n, \mathcal{H})$, specified below, and consider the finite-dimensional approximation problem: find $\boldsymbol{\psi}_d^x \in (\hat{S}_d)^m$ such that, for $i = 1, \ldots, m$,

$$a(\psi_{d,i}^x, v_d) = \int_{\mathbf{R}^n} f_i^x \, v_d \, \mathrm{e}^{-V/2} \, \mathrm{d}y \qquad \forall v_d \in \hat{S}_d. \tag{2.24}$$

Existence and uniqueness, up to a function in the kernel of $\mathcal{H}$, if that space is contained in $\hat{S}_d$, of the solution to the finite-dimensional problem are inherited from the infinite-dimensional problem (2.21).

For a given basis $\{e_\alpha\}_{|\alpha|\le d}$ of $\hat{S}_d$, the finite-dimensional approximation of $\boldsymbol{\psi}^x$ can be expanded as $\boldsymbol{\psi}_d^x = \sum_{|\alpha|\le d} \boldsymbol{\psi}_\alpha^x \, e_\alpha$, and from the variational formulation (2.24) we obtain the following linear systems:

$$\sum_{|\beta|\le d} a(e_\alpha, e_\beta)\, \boldsymbol{\psi}_\beta^x = \mathbf{f}_\alpha^x \quad \text{with} \quad \mathbf{f}_\alpha^x = \int_{\mathbf{R}^n} \mathbf{f}^x \, e_\alpha \, \mathrm{e}^{-V/2} \, \mathrm{d}y. \tag{2.25}$$

We will use the notation $S_{\alpha\beta} = a(e_\alpha, e_\beta)$ for the stiffness matrix. In view of formula (2.23) we see that we also need an approximation of the gradient of the solution, which we denote by $(\nabla_x \boldsymbol{\psi}^x)_d$ and decompose in the basis as $(\nabla_x \boldsymbol{\psi}^x)_d = \sum_{|\alpha|\le d} (\nabla_x \boldsymbol{\psi}^x)_\alpha \, e_\alpha$. This approximation can be obtained by solving (2.25) with the right-hand side $(\nabla_x \mathbf{f}^x)_\alpha = \int_{\mathbf{R}^n} (\nabla_x \mathbf{f}^x)\, e_\alpha \, \mathrm{e}^{-V/2} \, \mathrm{d}y$. In practice, we solved (2.25) by using the function *solve* of the C++ linear algebra library Armadillo. As $d \to \infty$, the lowest eigenvalue of $S_{\alpha\beta}$ tends to zero. This could in principle induce numerical instabilities, because the system becomes ill-conditioned, but didn't cause any problem in any of the numerical experiments we present in Section 2.6.

Then, by substituting the approximations of $\boldsymbol{\psi}^x$ and $\nabla_x \boldsymbol{\psi}^x$ in (2.23), we calculate the approximate drift and diffusion as follows:

$$\mathbf{F}_d(x) = \int_{\mathbf{R}^n} (\nabla_x \boldsymbol{\psi}^x)_d \left(\mathbf{f}^x \, \mathrm{e}^{-V/2}\right) \mathrm{d}y = \sum_{|\alpha|\le d} (\nabla_x \boldsymbol{\psi}^x)_\alpha \, \mathbf{f}_\alpha^x, \tag{2.26a}$$

$$\mathbf{A}_{0,d}(x) = \int_{\mathbf{R}^n} \boldsymbol{\psi}_d^x \otimes \left(\mathbf{f}^x \, \mathrm{e}^{-V/2}\right) \mathrm{d}y = \sum_{|\alpha|\le d} \boldsymbol{\psi}_\alpha^x \otimes \mathbf{f}_\alpha^x, \tag{2.26b}$$

$$\mathbf{D}_d(x) = \int_{\mathbf{R}^n} \boldsymbol{\alpha}\boldsymbol{\alpha}^T(x,y) \, \mathrm{e}^{-V} \, \mathrm{d}y + \mathbf{A}_{0,d}(x) + \mathbf{A}_{0,d}(x)^T, \quad \mathbf{A}_d(x)\mathbf{A}_d(x)^T = \mathbf{D}_d(x). \tag{2.26c}$$

For $\mathbf{A}_d(x)$ to be well defined, the symmetric matrix $\mathbf{D}_d$ must be positive semi-definite. Using

the notations $\psi_\xi^x = \boldsymbol{\psi}_d^x \cdot \boldsymbol{\xi}$ and $f_\xi^x = (\mathbf{f}^x \, \mathrm{e}^{-V/2}) \cdot \boldsymbol{\xi}$, we can show this by noticing that, for any $\boldsymbol{\xi} \in \mathbf{R}^m$, $\boldsymbol{\xi}^T (\mathbf{A}_{0,d}(x) + \mathbf{A}_{0,d}(x)^T)\boldsymbol{\xi} = 2 \int_{\mathbf{R}^n} \psi_\xi^x f_\xi^x \, \mathrm{d}y = 2\, a(\psi_\xi^x, \psi_\xi^x)$, by Eq. (2.24).

Although $\boldsymbol{\psi}_d^x$ and $(\nabla_x \boldsymbol{\psi}^x)_d$ are not uniquely defined when $\hat{S}_d$ contains the kernel of $\mathcal{H}$, the coefficients defined in Eqs. (2.26a) and (2.26b) are unique, because by the centering condition f is orthogonal in $L^2(\mathbf{R}^n)$ to the kernel of $\mathcal{H}$. The matrix $\mathbf{A}_d(x)$ is not, however, uniquely determined by Eq. (2.26c). Using these coefficients, we obtain the approximate homogenized SDE

$$\mathrm{d}X_d \;=\; \mathbf{F}_d(X_d)\,\mathrm{d}t + \mathbf{A}_d(X_d)\,\mathrm{d}W_t. \tag{2.27}$$

This equation can now be easily solved using a standard numerical method, e.g. Euler–Maruyama.

Our numerical methodology is based on the expansion of the solution to (2.20) in Hermite functions:

$$\hat{S}_d = \operatorname{span}\{\mathbf{h}_\alpha(y; \mu, \Sigma)\}_{|\alpha| \le d}. \tag{2.28}$$

A good choice of the mean and covariance, μ and Σ, respectively, is important for the efficiency of the algorithm. In our implementation we choose

$$\mu = \int_{\mathbf{R}^n} y \; \mathrm{e}^{-V(y)} \, \mathrm{d}y \quad \text{and} \quad \Sigma = \lambda^2 \int_{\mathbf{R}^n} (y-\mu)(y-\mu)^T \; \mathrm{e}^{-V(y)} \, \mathrm{d}y, \tag{2.29}$$

where $\lambda > 0$ is a free parameter independent of the first two moments of e^{-V}. This choice for the mean and covariance guarantees that our method is invariant under the rescaling $\tilde{Y}_t^\varepsilon = \sigma(Y_t^\varepsilon - m)$. An example illustrating why this is desirable is when the mass of the probability density e^{-V} is concentrated far away from the origin. Using Hermite functions centered at 0 would provide a very poor approximation in this case, but choosing Hermite functions around the center of mass of e^{-V} leads to a much better approximation. Note that this is not the only choice that guarantees invariance under rescaling, but it is the most natural one.

Remark 2.2. When the potential V is quadratic, say $V(y) = \frac{1}{2}(y-m)^T S(y-m)$, the eigenfunctions of the operator $\mathcal{H}$ (defined in (2.19)) are precisely the Hermite functions $\mathbf{h}_\alpha(y; m, S)$. Hence choosing these as a basis, i.e. $e_\alpha = \mathbf{h}_\alpha(y; m, S)$, leads to a diagonal matrix A in the linear systems (2.25), because $a(e_\alpha, e_\beta) = \lambda_\alpha \delta_{\alpha\beta}$, with λ_α defined in Eq. (A.26). This choice corresponds to $\lambda = 1$ in (2.29). The optimal choice for the parameters μ and Σ for a general density e^{-V} and function $\mathbf{f}$ has been partially studied. In particular, it was shown in [96] that $O(p^2)$ Hermite polynomials are necessary to resolve p wavelengths of a sine function, when keeping the scaling parameter fixed. This result carries over to the case of normalized Hermite functions, where the associated covariance matrix would play the role of the scaling parameter. In [209], it was shown that much better results could be obtained by choosing the scaling parameter as a function of the degree of approximation. In particular, the author shows that by choosing this parameter inversely proportional to the number of Hermite functions, only $O(p)$ functions are needed in order to resolve p wavelengths in one spatial dimension. More recently, the question of the optimal choice of the scaling parameter has also been studied in the framework of spectral methods for the Fokker–Planck equation, see [74]. ⊘

Summary of the Method In short, the method can be summarized as follows.

For a given initial condition $X^\varepsilon(0) = X_0$, $n = 0, 1, 2, \ldots$, a given stochastic integrator $X_d^{n+1} = \Psi(X_d^n, \mathbf{F}_d, \mathbf{A}_d, \Delta t, \xi_n)$, and a chosen time step Δt, set $X_0^n = X_0$ and

1. Compute the solution $\boldsymbol{\psi}_d^{X_d^n}$ and $(\nabla_x \boldsymbol{\psi}^{X_d^n})_d$ of (2.25);
2. Evaluate $\mathbf{F}_d(X_d^n), \mathbf{A}_d(X_d^n)$ from (2.26);
3. Compute a time step $X_d^{n+1} = \Psi(X_d^n, \mathbf{F}_d, \mathbf{A}_d, \Delta t, \xi_n)$, and go back to 1.

2.4 Main Results

In this section we present the main results on the analysis of our numerical method, the proof of which will be presented in Section 2.5. We first need to introduce some new notations. We will denote by $\langle \cdot, \cdot \rangle_{\mathrm{e}^{-V}}$ the inner product of $L^2\left(\mathbf{R}^n, \mathrm{e}^{-V}\right)$, defined by $\langle u, v \rangle_{\mathrm{e}^{-V}} = \int_{\mathbf{R}^n} u\, v\, \mathrm{e}^{-V}\, \mathrm{d}y$, and by $\| \cdot \|_{\mathrm{e}^{-V}}$ the associated norm. We will also use the notation $\| \cdot \|_{k,\mathrm{e}^{-V}}$ for the norm of $H^k(\mathbf{R}^n, \mathrm{e}^{-V})$, and $\| \cdot \|_{k,\mathcal{O}}$, where $\mathcal{O}$ is a negative selfadjoint operator, for the norm of $H^k(\mathbf{R}^n, \mathcal{O})$, see Appendix A.2. We will denote by $\pi(\cdot)$ the projection onto mean-zero functions of $L^2\left(\mathbf{R}^n, \mathrm{e}^{-V}\right)$, defined by

$$\pi(v) = v - \langle v, 1 \rangle_{\mathrm{e}^{-V}}, \quad v \in L^2\left(\mathbf{R}^n, \mathrm{e}^{-V}\right).$$

We will work mostly with the Schrödinger formulation (2.20) of the Poisson equation. In that context, we will employ the $L^2(\mathbf{R}^n)$ projection operator on $\{\hat{v} \in L^2(\mathbf{R}^n) : \hat{M}(\hat{v}) = 0\}$, see Eq. (2.22), which we denote by $\hat{\pi}(\cdot)$:

$$\hat{\pi}(\hat{v}) = \hat{v} - \left\langle \hat{v}, \mathrm{e}^{-V/2} \right\rangle_0 \mathrm{e}^{-V/2}, \quad \hat{v} \in L^2(\mathbf{R}^n),$$

where $\langle \cdot, \cdot \rangle_0$ denotes the $L^2(\mathbf{R}^n)$ inner product.

Finally, we will say that a function $g \in L^2(\mathbf{R}^n) \cap C^\infty(\mathbf{R}^n)$ decreases faster than any exponential function in the $L^2(\mathbf{R}^n)$ sense if

$$\int_{\mathbf{R}^n} g(x)^2\, \mathrm{e}^{\mu|y|}\, \mathrm{d}y < \infty \quad \forall \mu \in \mathbf{R}, \tag{2.30}$$

and denote by $E(\mathbf{R}^n)$ the space of all such functions.

In addition to the hypotheses presented in Section 2.2, we will use the following assumptions.

Assumption 2.1. *The potential $W(y)$, introduced in (H_W), is bounded from above by a polynomial of degree $4k$, for some $k \in \mathbf{N}$. Furthermore, for every multi-index α, there exist constants $c_\alpha > 0$ and $\mu_\alpha \in \mathbf{R}$ such that*

$$\left|\partial_y^\alpha V\right| \leq c_\alpha\, \mathrm{e}^{\mu_\alpha |y|},$$

where $V(\cdot)$ is the potential that appears in (2.9b).

Assumption 2.2. *The drift vector* $\mathbf{f}(x,y)$ *in* (2.9a) *is such that* $\mathrm{e}^{-V(\cdot)/2}\,\partial_y^\alpha \mathbf{f}(x,\cdot) \in (E(\mathbf{R}^n))^m$ *and* $\mathrm{e}^{-V(\cdot)/2}\,\partial_y^\alpha \nabla_x \mathbf{f}(x,\cdot) \in (E(\mathbf{R}^n))^{m\times m}$ *for all* $\alpha \in \mathbf{N}^n$ *and* $x \in \mathbf{R}^m$.

For the proof of our main theorem we will need to have control on higher order derivatives of the solution to the Poisson equation (2.10). To obtain such bounds we need to strengthen our assumptions on $\mathbf{f}(x,y)$ in (2.9a). In particular, in addition to (H_G), we assume the following:

Assumption 2.3. *For all* $\alpha \in \mathbf{N}^n$, *there exist constants* $C_\alpha > 0$ *and* $\ell_\alpha \in \mathbf{N}$ *such that*

$$\left|\partial_y^\alpha \mathbf{f}\right| + \left|\partial_y^\alpha \nabla_x \mathbf{f}\right| \leq C_\alpha\,(1+|y|^{\ell_\alpha}).$$

In addition, the diffusion coefficient on the right-hand side of (2.9a) *satisfies*

$$|\boldsymbol{\alpha}(x,y)| \leq K(1+|y|^{m_2}),$$

for constants K *and* m_2 *independent of* x.

From the Pardoux–Veretennikov bounds (2.12), a bootstrapping argument, Assumptions 2.1 and 2.3 and the integrability of monomials with respect to Gaussian weights we obtain the bounds

$$\|\boldsymbol{\phi}^x\|_{s,\mathcal{L}_{\mu,\Sigma}} \vee \|\nabla_x \boldsymbol{\phi}^x\|_{s,\mathcal{L}_{\mu,\Sigma}} \vee \|\mathbf{f}^x\|_{\mathrm{e}^{-V}} \leq C(s), \tag{2.31}$$

for $s \in \mathbf{N}$ and a constant $C(s)$ independent of x, and where $a \vee b$ denotes the maximum between a and b.

Remark 2.3. In Assumption 2.3 we assumed that the derivatives of the drift vector in (2.9a) with respect to y are bounded uniformly in x. This is a very strong assumption and it can be replaced by a linear growth bound as in (H_G). Under such an assumption the proof of Theorem 2.4 has to be modified using a localization argument that is based on the introduction of appropriate stopping times. Although tedious, this is a standard argument, see e.g. [112], and we will not present it in this chapter. Details can be found in [215]. ⊘

Theorem 2.2 (Spectral convergence of the Hermite–Galerkin method)**.** *Under Assumptions 2.1 to 2.3, for any* $s \in \mathbf{N}$ *there exists* $C(s)$, *independent of* x, *such that the approximate solutions* $\boldsymbol{\psi}_d^x$ *and* $(\nabla_x \boldsymbol{\psi}^x)_d$ *satisfy the following error estimate:*

$$\|\hat{\boldsymbol{\pi}}(\boldsymbol{\psi}_d^x) - \boldsymbol{\psi}^x\|_0 \vee \|\hat{\pi}((\nabla_x \boldsymbol{\psi}^x)_d) - \nabla_x \boldsymbol{\psi}^x\|_0 \leq C(s)\,d^{-s},$$

where $\|\cdot\|_0$ *is the usual* $L^2(\mathbf{R}^n)$ *norm.*

Using this result, we can prove spectral convergence for the calculation of the drift and diffusion coefficients.

Theorem 2.3 (Convergence of the drift and diffusion coefficients F_d and A_d)**.** *Suppose that Assumptions 2.1, 2.2 and 2.3 hold. Then the error on the approximate drift and diffusion coefficients decreases faster than any negative power of* d, *uniformly in* x, *i.e. for all* $s \in \mathbf{N}$ *there exists* $D(s)$ *such that*

$$\sup_{x\in\mathbf{R}^m} \left(|\mathbf{F}_d(x) - \mathbf{F}(x)| \vee \left|\mathbf{A}_d(x)\,\mathbf{A}_d(x)^T - \mathbf{A}(x)\,\mathbf{A}(x)^T\right|\right) \leq D(s)\,d^{-s}.$$

Using the spectral convergence of the approximate calculation of the drift and diffusion coefficients, we can now control the distance between the solution of the homogenized SDE and its approximation (2.27).

As we have already mentioned, homogenization/diffusion approximation theorems are generally of the weak convergence type. Furthermore, the effective diffusion coefficient of the simplified equation is not uniquely defined – see Equation (2.14) and the fact that $\mathbf{D}(x) = \mathbf{A}(x)\mathbf{A}(x)^T$. Consequently, it is not possible to prove the strong convergence of the solution to the approximate SDE (2.27) to the solution to the homogenized SDE (2.16) without constraining the choice of the factorization. To establish the next result, we will therefore assume that $\mathbf{A}(x)$ and $\mathbf{A}_d(x)$ are the unique symmetric positive semi-definite square roots of $\mathbf{D}(x)$ and $\mathbf{D}_d(x)$.

Denoting by $X(t)$ the exact solution of the homogenized equation and by $X_d(t)$ the approximate solution, we use the following norm to measure the error:

$$|||X(t) - X_d(t)||| := \left(\mathbf{E} \left[\sup_{0 \le t \le T} |X(t) - X_d(t)|^2 \right] \right)^{1/2}.$$

Theorem 2.4. *Let Assumptions 2.1 to 2.3 hold. Then the error between the approximate and exact solutions of the simplified equation satisfies*

$$|||X(t) - X_d(t)||| \le \sqrt{4\,(T+4)\,\beta(s)\,T\,d^{-s}}\,\exp\left(2\,(T+4)\,C_L\,T\right),$$

for any $s \in \mathbf{N}$ and $T > 0$, and where $\beta(s)$ is a constant depending only on s.

Now we consider the fully discrete scheme. We need to consider an appropriate discretization of the approximate homogenized equation (2.27). For simplicity we present the convergence results for the case when we discretize the homogenized SDE using the Euler–Maruyama method:

$$X_d^{n+1} = X_d^n + \Delta t\,\mathbf{F}_d(X_d^n) + \mathbf{A}_d(X_d^n)\,\Delta W_n, \tag{2.32}$$

but we emphasize that any higher order integrator, e.g. the Milstein scheme, could be used [133, 156]. The following is a classical result on the convergence of X_d^n for which we refer to [133, 156, 112] for a proof.

Theorem 2.5 (Convergence of the SDE solver)**.** *Assume that X_0 is a random variable such that $\mathbf{E}|X_0|^2 < \infty$ and that Assumptions 2.1 to 2.3 hold. Then*

$$\left(\mathbf{E} \left[\sup_{n\Delta t \in [0,T]} |X_d^n - X_d(t_n)|^2 \right] \right)^{\frac{1}{2}} \le C(T)\sqrt{\Delta t}.$$

for any choice of T, where X_d^n denotes the solution of (2.32).

Combined, Theorem 2.4 and Theorem 2.5 imply the weak convergence of the solution of (2.32) to the solution of the homogenized equation (2.16).

2.5 Proofs of the Main Results

2.5.1 Convergence of the Spectral Method for the Poisson Equation

In this section we present the proof of Theorem 2.2, establishing the convergence of the spectral method for the solution of the Poisson equation (2.17). Since the variable x only appears as a parameter in the Poisson equation, we will consider in this section that it takes an arbitrary value and will omit it from the notation. Additionally, to disencumber ourselves of vectorial notations, we will consider an arbitrary direction of $\mathbf{R}^n$, defined through a unit vector $\mathbf{e}$, and denote by f the projection $\mathbf{f} \cdot \mathbf{e}$.

We recall from [170, 169] that there exists a unique smooth mean-zero function of $\phi \in H^1(\mathbf{R}^n, \mathrm{e}^{-V})$ satisfying the variational formulation

$$a_V(\phi, v) := \langle \nabla\phi, \nabla v \rangle_{\mathrm{e}^{-V}} = \langle f, v \rangle_{\mathrm{e}^{-V}} \quad \forall v \in H^1(\mathbf{R}^n, \mathrm{e}^{-V}). \tag{2.33}$$

Let S_d be the finite-dimensional subset of $H^1(\mathbf{R}^n, \mathrm{e}^{-V})$ defined by $S_d = \mathrm{e}^{V/2}\, \hat{S}_d$, where $\hat{S}_d$ is the approximation space defined in eq. (2.28), and consider the following problem: find $\phi_d \in S_d$ satisfying:

$$a_V(\phi_d, v_d) = \langle f, v_d \rangle_{\mathrm{e}^{-V}} \quad \forall v_d \in S_d. \tag{2.34}$$

Note that, by definition of f, $\phi = \boldsymbol{\phi} \cdot \mathbf{e}$ and $\phi_d = \boldsymbol{\phi}_d \cdot \mathbf{e}$. The convergence of ϕ_d to ϕ can be obtained using techniques from the theory of finite elements, in particular Céa's lemma and an approximation argument. We will use the notation that was introduced at the beginning Section 2.4.

Lemma 2.6 (Céa's lemma)**.** *Let ϕ be the solution of (2.33) satisfying $\mathcal{M}(\phi) = 0$ and ϕ_d be a solution of (2.34). Then,*

$$\|\phi - \pi(\phi_d)\|_{1,\mathrm{e}^{-V}} \leq C \inf_{v_d \in S_d} \|\phi - v_d\|_{1,\mathrm{e}^{-V}}.$$

Proof. The proof is classical. It uses Poincaré inequality for the measure $\mathrm{e}^{-V}\,\mathrm{d}x$, recalled in Appendix A.2, Proposition A.12. From this inequality, we obtain the coercivity estimate $c\, a(v, v) \geq \|\pi(v)\|^2_{1,\mathrm{e}^{-V}}$ for all $v \in H^1(\mathbf{R}^n, \mathrm{e}^{-V})$. Combining this with Galerkin orthogonality, $a_V(\phi - \phi_d, v_d) = 0$ for all $v_d \in S_d$ and the continuity estimate $a_V(v_1, v_2) \leq \|v_1\|_{1,\mathrm{e}^{-V}} \|v_2\|_{1,\mathrm{e}^{-V}}$ for all $v_1, v_2 \in H^1(\mathbf{R}^n, \mathrm{e}^{-V})$ gives the result. More precisely,

$$\begin{aligned}
\|\pi(\phi - \phi_d)\|_{1,\mathrm{e}^{-V}}{}^2 &\leq c\, a_V(\phi - \phi_d, \phi - \phi_d) \\
&= c\, a_V(\pi(\phi - \phi_d), \phi - \phi_d) \\
&= c\, a_V(\pi(\phi - \phi_d), \phi - v) \\
&\leq c\, \|\pi(\phi - \phi_d)\|_{1,\mathrm{e}^{-V}} \|\phi - v\|_{1,\mathrm{e}^{-V}},
\end{aligned}$$

which leads to the conclusion. □

Since we will be working mostly with the Schrödinger formulation of Poisson equation, we need an analogue of Lemma 2.6 for the transformed PDE. We recall from Appendix A.2 that the

space $H^1(\mathbf{R}^n, \mathcal{H})$ is equipped with the norm

$$\|\psi\|_{1,\mathcal{H}}^2 = \|\psi\|_0^2 + \int_{\mathbf{R}^n} |\nabla\psi|^2 \, \mathrm{d}y + \int_{\mathbf{R}^n} W\psi^2 \, \mathrm{d}y.$$

Lemma 2.7. *Let $\boldsymbol{\psi}$ be the unique solution of (2.21) satisfying $\hat{\mathcal{M}}(\psi) = 0$ and $\boldsymbol{\psi}_d$ be a solution of (2.24). Then the projections $\psi = \boldsymbol{\psi} \cdot \mathbf{e}$ and $\psi_d = \boldsymbol{\psi}_d \cdot \mathbf{e}$ satisfy*

$$\|\psi - \hat{\pi}(\psi_d)\|_{1,\mathcal{H}} \leq C \inf_{v_d \in \hat{S}_d} \|\psi - v_d\|_{1,\mathcal{H}}. \tag{2.35}$$

Proof. The result follows directly by using the fact that $\mathrm{e}^{-V/2}$ is also a unitary transformation from $H^1(\mathbf{R}^n, \mathrm{e}^{-V})$ to $H^1(\mathbf{R}^n, \mathcal{H})$. □

Next, we focus on establishing a result that will allow us to control the right-hand side of (2.35). In [83, Lemma 2.3] the authors show that any smooth square integrable function such that $(-\Delta + W)v = g$ lies in the space $E(\mathbf{R}^n)$ introduced in (2.30), provided that $g \in E(\mathbf{R}^n)$ and that Assumption (H_W) holds. Differentiating the equation with respect to y_i, we obtain:

$$(-\Delta + W)\, \partial_{y_i} v = \partial_{y_i} g - \partial_{y_i} W \, v,$$

so it is clear by Assumption 2.1 that $\partial_\alpha \psi \in E(\mathbf{R}^n)$ for all values of $\alpha \in \mathbf{N}^n$. This implies that ψ belongs to the Schwartz space $S(\mathbf{R}^n)$. We now generalize slightly [83, Lemma 3.1]. This result will enable to control the norm $\|\cdot\|_{1,\mathcal{H}}$ on the right-hand side of (2.35) by a norm $\|\cdot\|_{k,\mathcal{H}_{\mu,\Sigma}}$, where $\mathcal{H}_{\mu,\Sigma}$ is an operator defined in Appendix A.2. From this appendix, we recall that the operator $\mathcal{H}_{\mu,\Sigma}$, with $\mu \in \mathbf{R}^n$ and Σ a symmetric positive definite matrix, is defined by $\mathcal{H}_{\mu,\Sigma} = -\Delta + W_{\mu,\Sigma}(y)$, where $W_{\mu,\Sigma}$ denotes the quadratic function $(y-\mu)^T\Sigma^{-2}(y-\mu)/4 - \operatorname{tr}\Sigma^{-1}/2$.

Lemma 2.8. *For every $k \in \mathbf{N}$ and $v \in S(\mathbf{R}^n)$,*

$$\int_{\mathbf{R}^n} |y|^{4k} \, v^2(y) \, \mathrm{d}y \leq C(k, \mu, \Sigma) \|v\|_{2k,\mathcal{H}_{\mu,\Sigma}}^2,$$

where $C(k, \mu, \Sigma)$ is a constant independent of v.

Proof. We set $Q_{\mu,\Sigma} = (y-\mu)^T\Sigma^{-2}(y-\mu)/4$. Following the methodology used to prove lemma 3.1 in [83], we establish that:

$$\|Q_{\mu,\Sigma}(y)^{k+1} v\|_0^2 \leq \|Q_{\mu,\Sigma}(y)^k \left(\mathcal{H}_{\mu,\Sigma} + \operatorname{tr}\Sigma^{-1}/2\right) v\|_0^2 + C_1(k,\Sigma)\|Q_{\mu,\Sigma}(y)^k v\|_0^2,$$

for all $k \in \mathbf{N}$, and where $C_1(k,\Sigma) = (4k+2)(k\,\rho(\Sigma^{-2}) + \operatorname{tr}\Sigma^{-2}/4)$. Reasoning by recursion and applying the triangle inequality, this immediately implies

$$\begin{aligned}\|Q_{\mu,\Sigma}(y)^k v\|_0^2 &\leq \sum_{i=0}^{k} c_i(k,\Sigma) \, \| \left(\mathcal{H}_{\mu,\Sigma} + \operatorname{tr}\Sigma^{-1}/2\right)^i v\|_0^2 \\ &\leq C_2(k,\Sigma)\|v\|_{2k,\mathcal{H}_{\mu,\Sigma}}^2,\end{aligned}$$

To conclude, note that

$$|y|^{4k} \leq C_3 + C_4\, Q_{\mu,\Sigma}(y)^{2k},$$

for suitably chosen C_3 and C_4 depending on Σ and μ. □

A finer version of the previous inequality could be obtained by following the argumentation in [83, Theorem 3.2], but this will not be necessary for our purposes. Lemma 2.8 can be used to show the following result.

Lemma 2.9. *If $W(y)$ is bounded above by a polynomial of degree $4k$, there exists a constant C depending on k, μ, Σ, and W such that any $v \in S(\mathbf{R}^n)$ satisfies*

$$\|v\|_{1,\mathcal{H}} \leq C\, \|v\|_{2k,\mathcal{H}_{\mu,\Sigma}}.$$

Proof. This follows from the considerations of Appendix A.2. First we note that

$$\|v\|^2_{1,\mathcal{H}} = \|v\|^2_{1,\mathcal{H}_{\mu,\Sigma}} + \int_{\mathbf{R}^n} (W - W_{\mu,\Sigma}) v^2 \,\mathrm{d}y.$$

To bound the second term, we use Assumption 2.1 on W, together with Lemma 2.8:

$$\int_{\mathbf{R}^n} (W - W_{\mu,\Sigma}) v^2 \,\mathrm{d}y \leq \int_{\mathbf{R}^n} (C_1 + C_2\ |y|^{4k}) v^2 \,\mathrm{d}y \leq C_3 \|v\|^2_{2k,\mathcal{H}_{\mu,\Sigma}},$$

with C_1, C_2, C_3 depending on k, μ, Σ. □

Upon combining the results presented so far in this section, we can complete the proof of Theorem 2.2.

Proof of Theorem 2.2. In this proof, the constant C is independent of d and changes from line to line. By Lemmas 2.7 and 2.8, and the fact that the exact solution ψ and its derivatives are smooth and decrease faster than exponentials, we have:

$$\|\psi - \hat{\pi}(\psi_d)\|_{1,\mathcal{H}} \leq C \inf_{v_d \in \hat{S}_d} \|\psi - v_d\|_{1,\mathcal{H}} \leq C \inf_{v_d \in \hat{S}_d} \|\psi - v_d\|_{2k,\mathcal{H}_{\mu,\Sigma}}.$$

Using Corollary A.22 on approximation by Hermite functions, we have for any $s > 2k$

$$\begin{aligned} \|\psi - \hat{\pi}(\psi_d)\|_{1,\mathcal{H}} &\leq C(d+1)^{-\frac{s-2k}{2}} \|\psi\|_{s,\mathcal{H}_{\mu,\Sigma}}, \\ &\leq C(d+1)^{-\frac{s-2k}{2}}, \end{aligned}$$

where we used the first estimate of (2.31) and the fact that $\|\psi\|_{s,\mathcal{H}_{\mu,\Sigma}} = \|\phi\|_{s,\mathcal{L}_{\mu,\Sigma}}$. The same reasoning can be applied to $\nabla_x \psi$. Since s was arbitrary, this proves the statement. □

2.5.2 Convergence of the Drift and Diffusion Coefficients

In this section we prove the convergence of the drift and diffusion coefficients obtained from the approximate solution of the Poisson equation.

Proof of Theorem 2.3. From the expressions of $\mathbf{F}$ and $\mathbf{F}_d$ we have:

$$\begin{aligned}\mathbf{F}(x) - \mathbf{F}_d(x) &= \int_{\mathbf{R}^n} (\nabla_x \boldsymbol{\psi}^x - (\nabla_x \boldsymbol{\psi}^x)_d) \cdot (\mathbf{f}^x \, \mathrm{e}^{-V/2}) \, \mathrm{d}y \\ &= \int_{\mathbf{R}^n} (\nabla_x \boldsymbol{\psi}^x - \hat{\pi}((\nabla_x \boldsymbol{\psi}^x)_d)) \cdot (\mathbf{f}^x \, \mathrm{e}^{-V/2}) \, \mathrm{d}y,\end{aligned}$$

where we used the fact that $\mathbf{f}$ is centered, thus orthogonal to $(\nabla_x \boldsymbol{\psi}^x)_d - \hat{\pi}((\nabla_x \boldsymbol{\psi}^x)_d)$. Using Theorem 2.2 and Cauchy–Schwarz inequality we deduce that there exists for any value of $s \in \mathbf{N}$ a constant $C(s)$ such that

$$\begin{aligned}|\mathbf{F}_d(x) - \mathbf{F}(x)| &\leq \|\nabla_x \boldsymbol{\psi}^x - \hat{\pi}((\nabla_x \boldsymbol{\psi}^x)_d)\|_0 \, \|\mathbf{f}^x \, \mathrm{e}^{-V/2} \, \|_0 \\ &\leq C(s) \, d^{-s} \|\mathbf{f}^x\|_{\mathrm{e}^{-V}}.\end{aligned}$$

The error on the diffusion term can be bounded similarly:

$$\begin{aligned}|\mathbf{A}_{0,d}(x) - \mathbf{A}_0(x)| &= \int_{\mathbf{R}^n} (\boldsymbol{\psi}_d^x - \boldsymbol{\psi}^x) \otimes (\mathbf{f}^x \, \mathrm{e}^{-V/2}) \, \mathrm{d}y \\ &\leq C(s) \, d^{-s} \, \|\mathbf{f}^x\|_{\mathrm{e}^{-V}}.\end{aligned}$$

The proof can then be concluded using the last bound from (2.31). □

2.5.3 Convergence of the Solution to the SDE

The proof of Theorem 2.4 relies on two results. First, it relies on the existence of a constant C_L, proved in Section 2.A, such that the effective drift and diffusion coefficients, obtained using Eqs. (2.13) and (2.14) and by calculating $\mathbf{A}(x)$ as the unique symmetric positive semi-definite square root of $\mathbf{D}(x)$, satisfy

$$|\mathbf{F}(a) - \mathbf{F}(b)|^2 \vee |\mathbf{A}(a) - \mathbf{A}(b)|^2 \leq C_L |a - b|^2, \tag{2.36}$$

for all $a, b \in \mathbf{R}^m$. Next, by Theorem 2.3 and Lemma 2.10, there exists for every $s \in \mathbf{N}$ a constant $\beta(s)$ independent of d and x such that

$$|\mathbf{F}_d(x) - \mathbf{F}(x)|^2 \vee |\mathbf{A}_d(x) - \mathbf{A}(x)|^2 \leq \beta(s) \, d^{-s}, \tag{2.37}$$

for any $x \in \mathbf{R}^m$. Upon combining (2.36) and (2.37), Theorem 2.4 can be proved.

Proof of Theorem 2.4. The error $e_d(t) = X(t) - X_d(t)$ satisfies

$$e_d(t) = \int_0^t \mathbf{F}(X(\tau)) - \mathbf{F}_d(X_d(\tau)) \, d\tau + \int_0^t \mathbf{A}(X(\tau)) - \mathbf{A}_d(X_d(\tau)) \, \mathrm{d}W_\tau.$$

Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and Cauchy–Schwarz, we have

$$\begin{aligned}\mathbf{E}\left[\sup_{0\leq t\leq T}|e_d(t)|^2\right] \leq\ & 2\,T\,\mathbf{E}\left[\int_0^T |\mathbf{F}(X(\tau)) - \mathbf{F}_d(X_d(\tau))|^2\,\mathrm{d}\tau\right] \\ & + 2\,\mathbf{E}\left[\sup_{0\leq t\leq T}\left|\int_0^t \mathbf{A}(X(\tau)) - \mathbf{A}_d(X_d(\tau))\,\mathrm{d}W_\tau\right|^2\right].\end{aligned} \tag{2.38}$$

The first term on the right-hand side can be bounded by using the triangle inequality with the decomposition $\mathbf{F}(X(\tau)) - \mathbf{F}_d(X_d(\tau)) = (\mathbf{F}(X(\tau)) - \mathbf{F}(X_d(\tau))) + (\mathbf{F}(X_d(\tau)) - \mathbf{F}_d(X_d(\tau)))$, the Lipschitz continuity of $\mathbf{F}(\cdot)$ and the convergence of $\mathbf{F}_d$ to $\mathbf{F}$:

$$\begin{aligned}&\mathbf{E}\left[\int_0^T |\mathbf{F}(X(\tau)) - \mathbf{F}_d(X_d(\tau))|^2\,\mathrm{d}\tau\right] \\ &\qquad \leq \mathbf{E}\left[2\,\beta(s)\,T\,d^{-s} + 2\,C_L\int_0^T |X(\tau) - X_d(\tau)|^2\,d\tau\right] \\ &\qquad \leq 2\,\beta(s)\,T\,d^{-s} + 2\,C_L\int_0^T \mathbf{E}\left[\sup_{0\leq t\leq \tau}|e_d(t)|^2\right]d\tau\end{aligned} \tag{2.39}$$

The second term can be bounded in a similar manner by using Burkholder–Davis–Gundy inequality, see for example [126, Theorem 3.28], and Itô isometry :

$$\begin{aligned}&\mathbf{E}\left[\sup_{0\leq t\leq T}\left|\int_0^t \mathbf{A}(X(\tau)) - \mathbf{A}_d(X_d(\tau))\,\mathrm{d}W_\tau\right|^2\right] \\ &\qquad \leq 4\,\mathbf{E}\left|\int_0^T \mathbf{A}\left(X(\tau)\right) - \mathbf{A}_d(X_d(\tau))\,\mathrm{d}W_\tau\right|^2 \\ &\qquad \leq 8\,\beta(s)\,T\,d^{-s} + 8\,C_L\int_0^T \mathbf{E}\left[\sup_{0\leq t\leq \tau}|e_d(t)|^2\right]d\tau.\end{aligned} \tag{2.40}$$

Using (2.39) and (2.40) in (2.38), we obtain:

$$\mathbf{E}\left[\sup_{0\leq t\leq T}|e_d(t)|^2\right] \leq 4\,(T+4)\left(\beta(s)\,T\,d^{-s} + C_L\int_0^T \mathbf{E}\left[\sup_{0\leq t\leq \tau}|e_d(t)|^2\right]\mathrm{d}\tau\right).$$

By Gronwall's inequality, this implies:

$$\mathbf{E}\left[\sup_{0\leq t\leq T}|e_d(t)|^2\right] \leq 4\,(T+4)\,\beta(s)\,T\,d^{-s}\exp\left(4\,(T+4)\,C_L\,T\right), \tag{2.41}$$

which finishes the proof. □

Remark 2.4. Note that, as mentioned in Section 2.4, the convergence of the solution can still be proved if we assume that the Lipschitz continuity and convergence of the coefficients holds only locally, provided there exists $p > 2$ and a constant K independent of d such that the solutions of the equations

$$\mathrm{d}X = \mathbf{F}(X)\,\mathrm{d}t + \mathbf{A}(X)\,\mathrm{d}W_t, \qquad X(0) = X_0,$$

and

$$\mathrm{d}X_d = \mathbf{F}_d(X_d)\,\mathrm{d}t + \mathbf{A}_d(X_d)\mathrm{d}W_t, \qquad X_d(0) \;=\; X_0,$$

satisfy the moment bounds

$$\mathbf{E}\left[\sup_{0\le t\le T}|X(t)|^p\right] \vee \mathbf{E}\left[\sup_{0\le t\le T}|X_d(t)|^p\right] \le K.$$

With these alternative assumptions, we can show that

$$\begin{aligned}\mathbf{E}\left[\sup_{0\,\le\, t\,\le\, T}|X(t)\;-\;X_d(t)|^2\right] \;\le\; 4\,(T\;+\;4)\,D_R(s)\,T\,d^{-s}\,\exp\left(4\,(T\;+\;4)\,C_R\,T\right)\\ +\,2\,K\left(\frac{2^p\,\delta}{p}\;+\;\frac{p-2}{R^p\,p\,\delta^{\frac{2}{p-2}}}\right)\end{aligned}$$

for any $\delta > 0$ and $R > |X_0|$, and where C_R and D_R are the local constants for the assumptions. The proof of this estimate is very similar to the one of the strong convergence of Euler–Maruyama scheme in [112, Theorem 2.2], and will thus not be presented here. From this estimate, we deduce that the solution of the approximate homogenized equation converges to the exact solution when $d \to \infty$. ⊘

2.6 Implementation and Numerical Experiments

In this section, we discuss the implementation of the algorithm and present some numerical experiments to validate the method and illustrate our theoretical findings.

2.6.1 Implementation details

Below we discuss the quadrature rules used and the approach taken for the calculation of the matrix and right-hand side of the linear system of equations (2.25).

The algorithm requires the calculation of Gaussian integrals of the type:

$$I = \int_{\mathbf{R}^n} f(y)\,g_{\mu,\Sigma}(y)\,\mathrm{d}y.$$

Several approaches, either Monte Carlo-based or deterministic, can be used for the calculation of such integrals. Probabilistic methods offer an advantage when the dimension n of the state space of the fast process is large, but since the HMM is more efficient than our approach in that case, in practice we don't use them. Instead, we use a multi-dimensional quadrature rule obtained by tensorization of one-dimensional Gauss–Hermite quadrature rules.

For the calculation of the stiffness matrix $\mathbf{S}$, we can take advantage of the fact that this matrix is diagonal when the potential is equal to $V_{\mu,\Sigma} := \frac{1}{2}(y-\mu)\Sigma^{-1}(y-\mu) + \log(\sqrt{(2\pi)^n\,|\Sigma|})$.[7] Using the notation $\mathcal{H}_{\mu,\Sigma}$ to denote the same operator as in Lemma 2.8, and the shorthand notations

[7] The constant $\log(\sqrt{(2\pi)^n\,|\Sigma|})$ in $V(\mu,\Sigma)$ is chosen so that $\int_{\mathbf{R}^n}\mathrm{e}^{-V}\,\mathrm{d}y = 1$.

$\mathbf{H}_\alpha$ and $\mathbf{h}_\alpha$, for $\alpha \in \mathbf{N}^n$, in place of $\mathbf{H}_\alpha(y;\mu,\Sigma)$ and $\mathbf{h}_\alpha(y;\mu,\Sigma)$, respectively, we have:

$$S_{\alpha\beta} = -\int_{\mathbf{R}^n} (\mathcal{H} - \mathcal{H}_{\mu,\Sigma})\, \mathbf{h}_\alpha\, \mathbf{h}_\beta \,\mathrm{d}y - \int_{\mathbf{R}^n} \mathcal{H}_{\mu,\Sigma}\, \mathbf{h}_\alpha\, \mathbf{h}_\beta \,\mathrm{d}y =: S^\delta_{\alpha\beta} + Q_{\alpha\beta},$$

where $\mathbf{Q}$ is a diagonal matrix with entries $Q_{\alpha\beta}$ that can be computed explicitly, and

$$S^\delta_{\alpha\beta} = \int_{\mathbf{R}^n} (W - W_{\mu,\Sigma})\, \mathbf{h}_\alpha \mathbf{h}_\beta \,\mathrm{d}y = \int_{\mathbf{R}^n} (W - W_{\mu,\Sigma})\, g_{\mu,\Sigma}\, \mathbf{H}_\alpha \mathbf{H}_\beta \,\mathrm{d}y,$$

where $W_{\mu,\Sigma}$ is the potential obtained from $V_{\mu,\Sigma}$ according to Eq. (H_W). To simplify the calculation of these coefficients, we expand the Hermite polynomials in terms of monomials:

$$\mathbf{H}_\alpha(y;\mu,\Sigma) = \sum_{|\beta|\le d} c_{\alpha\beta}\, y^\beta.$$

With this notation, we write:

$$S^\delta_{\alpha\beta} = \sum_{|\rho|\le d} \sum_{|\sigma|\le d} c_{\alpha\rho}\, c_{\beta\sigma} \int_{\mathbf{R}^n} (W - W_{\mu,\Sigma})\, g_{\mu,\Sigma}\, y^{\rho+\sigma} \,\mathrm{d}y =: \sum_{|\rho|\le d} \sum_{|\sigma|\le d} c_{\alpha\rho}\, c_{\beta\sigma} I_{\rho+\sigma},$$

The integrals I_α are computed by numerical quadrature. Denoting by w_i and q_i the weights and nodes of the Gauss–Hermite quadrature, respectively, I_α is approximated as

$$I_\alpha \approx \sum_{i=1}^{N_q} w_i \left(W(q_i) - W_{\mu,\Sigma}(q_i)\right)\, g_{\mu,\Sigma}(q_i)\, q_i^\alpha, \qquad |\alpha| \le 2d,$$

where N_q denotes the number of points in the quadrature. Only the last factor of the previous expression depends on α, so the numerical calculation of these integrals can be performed by evaluating for each grid point the value of $w_i \left(W(q_i) - W_{\mu,\Sigma}(q_i)\right)\, g_{\mu,\Sigma}(q_i)$ and the values of q_i^α for $|\alpha| \le 2d$. For the numerical experiments presented in this chapter, we used a quadrature rule with 100 points in each dimension, so $N_q = 100^n$. This quadrature rule was used also to approximate the exact effective coefficients from Eq. (2.23).

A similar method can be applied for the calculation of the right-hand side. Assuming for simplicity that the slow process is scalar (i.e., $m = 1$), and denoting by $f(x,y)$ the corresponding scalar drift coefficient, the entries of the right-hand side are expressed as:

$$b_\alpha = \int_{\mathbf{R}^n} \mathrm{e}^{-V(y)/2}\, f(x,y)\, \mathbf{h}_\alpha(y) \,\mathrm{d}y.$$

By expanding the Hermite functions in terms of Hermite polynomials multiplying $g_{\mu,\Sigma}^{1/2}$, the previous equation can be rewritten as

$$b_\alpha = \sum_{|\beta|\le d} c_{\alpha\beta} \int_{\mathbf{R}^n} \left(\frac{\mathrm{e}^{-V(y)}}{g_{\mu,\Sigma}(y)}\right)^{\frac{1}{2}} f(x,y)\, y^\beta\, g_{\mu,\Sigma}(y) \,\mathrm{d}y,$$

which is a Gaussian integral that can also be calculated using a multi-dimensional Gauss–Hermite quadrature.

2.6.2 Numerical experiments

Now we present the results of some numerical experiments.

The Euler–Maruyama scheme is used to approximate both $X(t)$ and $X_d(t)$ with a time step of 0.01 for $T = 1$, and $N_r = 100$ replicas of the driving Brownian motion are used for the numerical computation of expectations. The ith replica of the discretized approximations of $X(t)$ and $X_d(t)$ are noted $X^{n,i}$ and $X_d^{n,i}$ respectively. In most of the numerical experiments below, the error is measured by:

$$E(d) = \left(\frac{1}{N_r} \sum_{i=1}^{N_r} \max_{0 \leq n \, \Delta t \leq 1} |X^{n,i} - X_d^{n,i}|^2 \right)^{\frac{1}{2}}, \tag{2.42}$$

which is an approximation of the norm $|||\cdot|||$ used in Theorem 2.4.

In the numerical experiments presented in this chapter, we chose the scaling parameter λ in Eq. (2.29) by trial-and-error, guided by the considerations presented in Appendix A.3.1; see, in particular, the discussion that follows Eq. (A.22). Confining our discussion here to the one-dimensional case for simplicity, and denoting by $h_d(x; \mu, \sigma^2)$ the one-dimensional Hermite function associated with the Hermite polynomial of degree d, we recall from this appendix that the final inflection points of $h_d(x; \mu, \sigma^2)$ occur at $x_d^{\pm} = \mu \pm \sqrt{4d+2}\,\sigma$ [139] and that, beyond these points, $h_d(x; \mu, \sigma^2)$ decreases to 0 rapidly. It is therefore crucial to ensure that σ (and therefore also λ) is sufficiently large. Increasing λ, however, leads to a poorer resolution of the higher frequencies, so a trade off needs to be made.

In practice, a suitable value of λ is found iteratively: starting with a sufficiently large value of λ, a rough approximation of the solution to Eq. (2.20) is obtained with our spectral method. A lower value of λ, but still large enough to guarantee that most of the mass (in the L^2 norm, for example) of this solution is concentrated in $[x_d^-, x_d^+]$, is then calculated, enabling a better resolution of the high frequencies. This procedure is then repeated until a satisfactory numerical solution is obtained. By numerical exploration, we observed super-algebraic convergence with respect to d regardless of the choice of λ. For a given value of the degree d, however, variations of λ can have a considerable effect on the accuracy of the method. This is not surprising, considering that a division of σ by a factor f corresponds to a multiplication, by the same factor, of the number of wavelengths that can be resolved; see the discussion in Remark 2.2. A natural extension of the work presented in this chapter is to develop a systematic methodology for identifying the optimal scaling parameter.

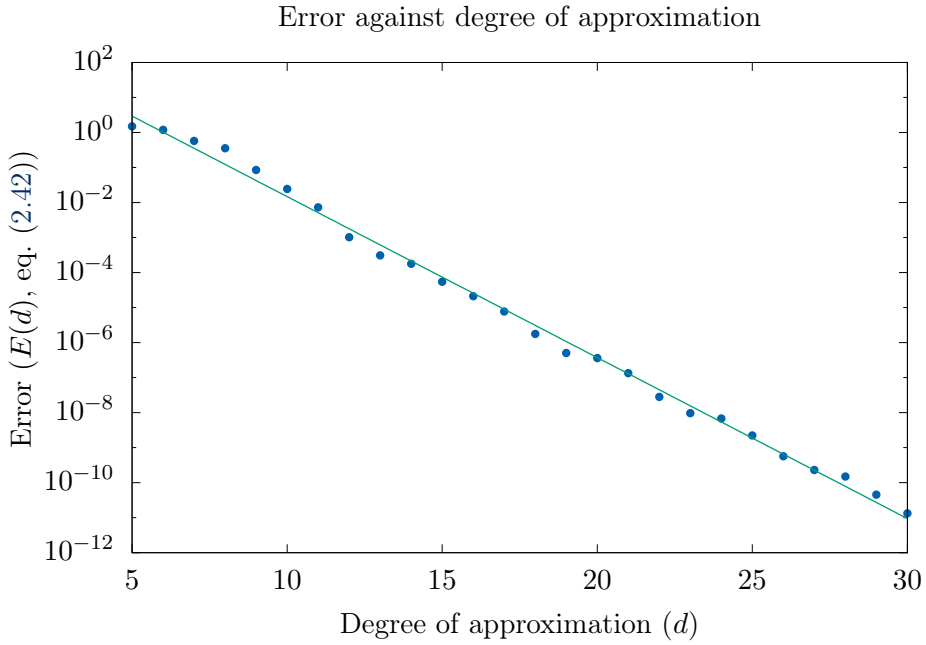


Figure 2.1: Error $E(d)$, see Eq. (2.42), for the fast-slow SDE (2.43). The blue dots were obtained by numerical experiments, and the straight line is the function $478.63 \times 10^{-0.45\,d}$ obtained by linear regression in the $d - \log_{10}(E(d))$ plane. A super-algebraic convergence is observed.

Test of the method for single well potentials

For the two problems in this section, the scaling parameter is chosen as $\lambda = 0.5$ for all degrees of approximation. We start by considering the following problem.

$$\begin{cases} \mathrm{d}x_0(t) = -\dfrac{1}{\varepsilon}\,\mathcal{L}_y\left[\cos\left(x_0(t) + y_0(t) + y_1(t)\right)\right]\,\mathrm{d}t, \\ \mathrm{d}x_1(t) = -\dfrac{1}{\varepsilon}\,\mathcal{L}_y\left[\sin\left(x_1(t)\right)\sin\left(y_0(t) + y_1(t)\right)\right]\,\mathrm{d}t, \\ \mathrm{d}y_0(t) = -\dfrac{1}{\varepsilon^2}\,\partial_{y_0}V(y)\,\mathrm{d}t + \dfrac{1}{\varepsilon}\left[\cos\left(x_0(t)\right)\cos\left(y_0(t)\right)\cos\left(y_1(t)\right)\right]\,\mathrm{d}t + \dfrac{\sqrt{2}}{\varepsilon}\mathrm{d}W_0(t), \\ \mathrm{d}y_1(t) = -\dfrac{1}{\varepsilon^2}\,\partial_{y_1}V(y)\,\mathrm{d}t + \dfrac{1}{\varepsilon}\left[\cos\left(x_0(t)\right)\cos\left(y_0(t) + y_1(t)\right)\right]\,\mathrm{d}t + \dfrac{\sqrt{2}}{\varepsilon}\mathrm{d}W_1(t), \end{cases} \tag{2.43}$$

with

$$V(y) = y_0^2 + y_1^2 + 0.5\left(y_0^2 + y_1^2\right)^2,$$

and where $\mathcal{L}_y = -\nabla V \cdot \nabla + \Delta$. We have written the right-hand side of the equations for the slow processes $x_0(t)$ and $x_1(t)$ in this form to ensure that the centering condition is satisfied. The convergence of the approximate solution of the effective equation for this problem is illustrated in Fig. 2.1. Here the probability density associated with the potential is unimodal and it decreases to 0 rapidly as $|y| \to \infty$, so Hermite functions are well suited for the approximation of the solution, which is reflected in the very good convergence observed.

In the next example, the state space of the fast process has dimension 3:

$$\begin{cases} \mathrm{d}x_0(t) = -\dfrac{1}{\varepsilon}\,\mathcal{L}_y\left[\cos\left(x_0(t)+y_0(t)+y_1(t)\right)\right]\,\mathrm{d}t, \\ \mathrm{d}x_1(t) = -\dfrac{1}{\varepsilon}\,\mathcal{L}_y\left[\sin\left(x_1(t)\right)\sin\left(y_0(t)+y_1(t)+2y_2(t)\right)\right]\,\mathrm{d}t, \\ \mathrm{d}y_0(t) = -\dfrac{1}{\varepsilon^2}\,\partial_{y_0}V(y)\,\mathrm{d}t + \dfrac{1}{\varepsilon}\left[\cos\left(x_0(t)\right)\cos\left(y_1(t)\right)\cos\left(y_0(t)+y_2(t)\right)\right]\,dt + \dfrac{\sqrt{2}}{\varepsilon}\,\mathrm{d}W_0(t), \\ \mathrm{d}y_1(t) = -\dfrac{1}{\varepsilon^2}\,\partial_{y_1}V(y)\,\mathrm{d}t + \dfrac{1}{\varepsilon}\left[\cos\left(x_0(t)\right)\cos\left(y_0(t)+y_1(t)\right)\right]\,dt + \dfrac{\sqrt{2}}{\varepsilon}\,\mathrm{d}W_1(t), \\ \mathrm{d}y_2(t) = -\dfrac{1}{\varepsilon^2}\,\partial_{y_2}V(y)\,\mathrm{d}t + \dfrac{\sqrt{2}}{\varepsilon}\,\mathrm{d}W_2(t), \end{cases} \tag{2.44}$$

with

$$V(y) = y_0^4 + 2y_1^4 + 3y_2^4.$$

Since, given the higher dimensionality of the problem, computing the effective coefficients is much more expensive computationally than in the previous case, we measure the error for a given value of the slow variables, by

$$e(d,x) = \frac{|\mathbf{F}(x) - \mathbf{F}_d(x)|}{|\mathbf{F}(x)|} + \frac{|\mathbf{A}(x) - \mathbf{A}_d(x)|}{|\mathbf{A}(x)|}. \tag{2.45}$$

The value we chose for the comparison is $x = (0.2, 0.2)$, for which the denominators in the previous equation are nonzero. The relative error on the homogenized coefficients is illustrated in Fig. 2.2. In this case, the method also performs very well, although it is slightly less accurate than in the previous example.

Test of the method for potentials with multiple wells

Now we consider multiple-well potentials that lead to multimodal distributions. The first potential that we analyze is the standard bistable potential,

$$V(y) = \frac{y^4}{4} - \frac{y^2}{2}.$$

We consider the fast/slow SDE system:

$$\begin{cases} \mathrm{d}x_t = -\dfrac{1}{\varepsilon}\mathcal{L}_y\left(x_t\,\sin(y_t)\right)\,\mathrm{d}t, \\ \mathrm{d}y_t = -\dfrac{1}{\varepsilon^2}\partial_y V(y_t)\,\mathrm{d}t + \dfrac{\sqrt{2}}{\varepsilon}\,\mathrm{d}W_t. \end{cases} \tag{2.46}$$

We choose the parameter λ in Eq. (2.29) to be $\lambda = 0.5$. Since the potential is associated with a bimodal distribution, we expect that, in order to reach a given accuracy, more Hermite functions will be required than in the unimodal cases considered previously. The convergence of the method is illustrated in Fig. 2.3. Although the method is slightly less accurate than in the previous cases, a super-algebraic convergence can still be observed. We emphasize that the computational cost for this one-dimensional problem is very low: the numerical solution can be calculated in a

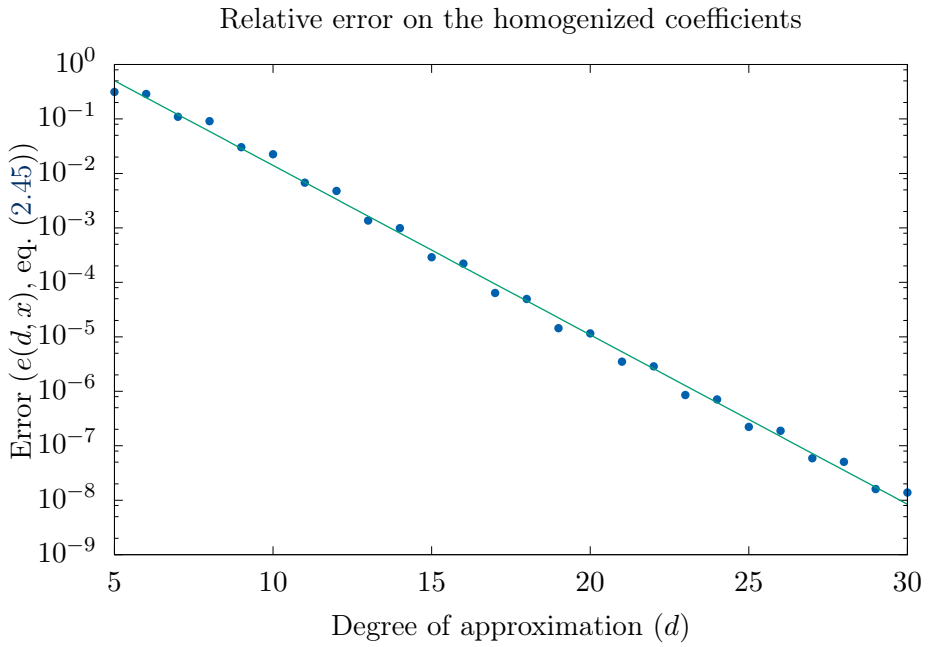


Figure 2.2: Relative error of the homogenized coefficients, $e(d,x)$, see Eq. (2.45), for the fast/slow SDE (2.44) at $x = (0.2, 0.2)$. In this case, as the points are more or less aligned, the convergence is also super-algebraic. The straight line, also obtained by linear regression in the $d - \log_{10}(E(d))$ plane, is the function $18.20 \times 10^{-0.31\,d}$. We notice that the convergence is slightly slower than for Eq. (2.43).

matter of seconds on a personal computer.

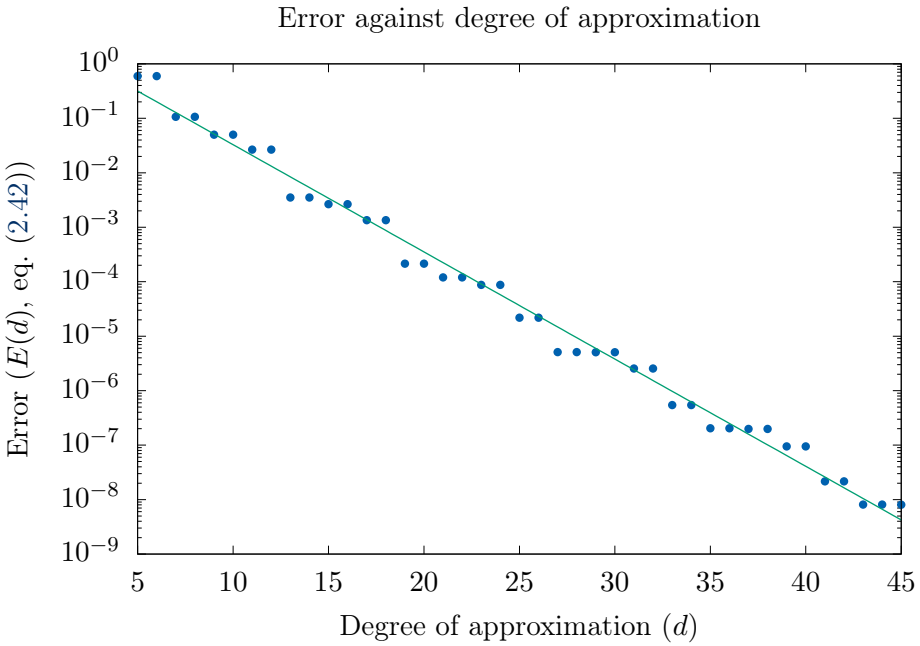


Figure 2.3: Error $E(d)$, see Eq. (2.42), for the fast/slow SDE (2.46). The straight line is the linear function $0.40 \times 10^{-0.19\,d}$, and we observe that the convergence is not as fast as in the unimodal cases.

Next we consider the tilted bistable potential

$$V(y) = \frac{y^4}{4} - \frac{y^2}{2} + 10\, y, \tag{2.47}$$

which corresponds to the case $\gamma = 1$, $\delta = 10$ in the examples considered in [28], and the fast/slow SDE

$$\begin{cases} \mathrm{d}x_t = -\dfrac{1}{\varepsilon}\mathcal{L}_y \left(x_t \, \sin(y_t) + \cos(y_t) + y_t^3 \right) \, \mathrm{d}t, \\ \mathrm{d}y_t = -\dfrac{1}{\varepsilon^2}\partial_y V(x_t, y_t)\, \mathrm{d}t + \dfrac{\sqrt{2}}{\varepsilon}\, \mathrm{d}W_t. \end{cases} \tag{2.48}$$

The convergence of the solution in this case is presented in Fig. 2.5, for the scaling parameter $\lambda = 1.5$. Due to the presence of a strong linear term, the potential is unimodal, see Fig. 2.4. As a result, the convergence of the spectral method is good, though it does not appear to be super-algebraic.

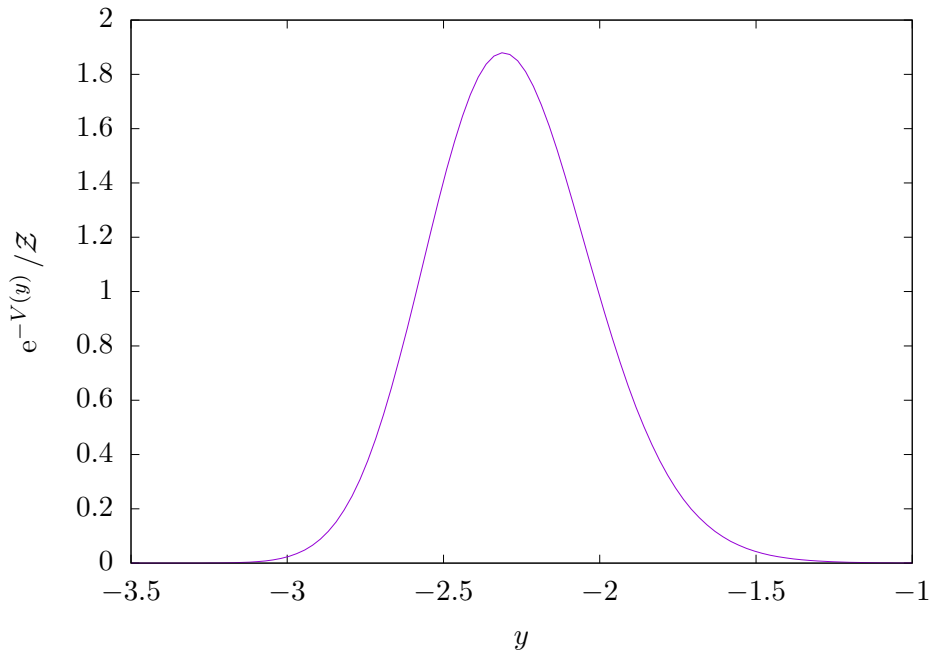


Figure 2.4: Probability density $\mathrm{e}^{-V(\cdot)}/\mathcal{Z}$ associated to the potential (2.47).

Finally, we consider a three-well potential in $\mathbf{R}^2$,

$$V(y) = \left((y_0-1)^2 + y_1^2\right)\left(\left(y_0+\tfrac{1}{2}\right)^2 + \left(y_1 - \tfrac{\sqrt{3}}{2}\right)^2\right)\left(\left(y_0+\tfrac{1}{2}\right)^2 + \left(y_1 + \tfrac{\sqrt{3}}{2}\right)^2\right), \tag{2.49}$$

and the following fast/slow SDE:

$$\begin{cases} \mathrm{d}x_0(t) = -\dfrac{1}{\varepsilon}\, \mathcal{L}_y \left[\cos\left(x_0(t) + y_0(t) + y_1(t)\right)\right] \, \mathrm{d}t, \\ \mathrm{d}x_1(t) = -\dfrac{1}{\varepsilon}\, \mathcal{L}_y \left[\sin\left(x_1(t)\right) \sin\left(y_0(t) + y_1(t)\right)\right] \, \mathrm{d}t, \\ \mathrm{d}y_0(t) = -\dfrac{1}{\varepsilon^2}\, \partial_{y_0} V(y)\, \mathrm{d}t + \dfrac{1}{\varepsilon}\left[\cos\left(x_0(t)\right) \cos\left(y_0(t)\right) \cos\left(y_1(t)\right)\right] \, \mathrm{d}t + \dfrac{\sqrt{2}}{\varepsilon}\, \mathrm{d}W_0(t), \\ \mathrm{d}y_1(t) = -\dfrac{1}{\varepsilon^2}\, \partial_{y_1} V(y)\, \mathrm{d}t + \dfrac{1}{\varepsilon}\left[\cos\left(x_0(t)\right) \cos\left(y_0(t) + y_1(t)\right)\right] \, \mathrm{d}t + \dfrac{\sqrt{2}}{\varepsilon}\, \mathrm{d}W_1(t). \end{cases} \tag{2.50}$$

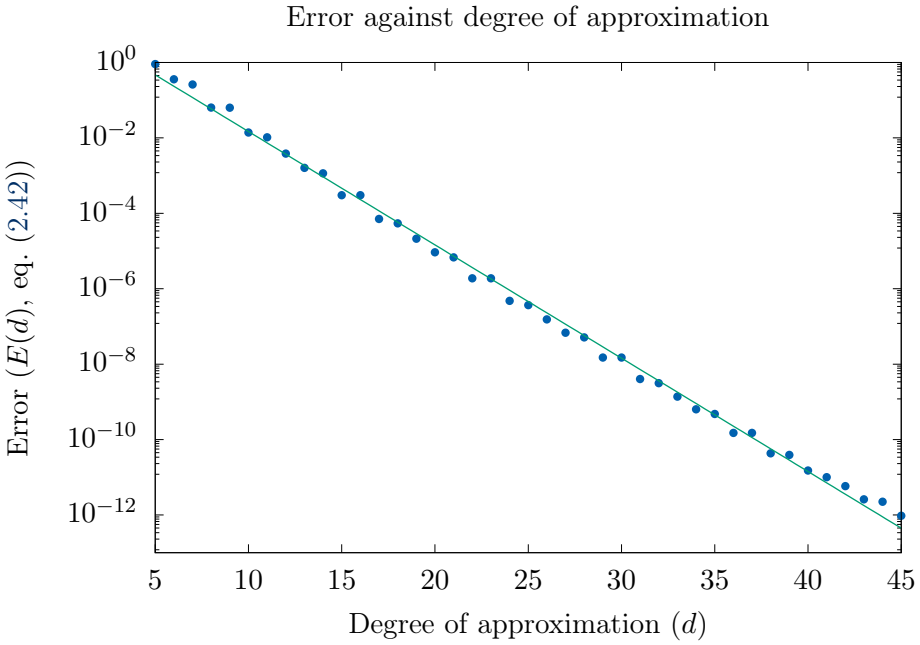


Figure 2.5: Error $E(d)$, see Eq. (2.42), for the fast/slow system (2.48). The straight line is the linear function $1.17 \times 10^{-0.30\,d}$, so we see that the convergence is faster than for the bistable potential (Fig. 2.3).

For this fast/slow SDE, we choose $\lambda = 0.35$. A contour plot of the potential is shown in Fig. 2.6, and the convergence graph is presented in Fig. 2.7. In this case the error is very large for degrees of approximation lower than 10, beyond which the convergence is clear and super-algebraic. The accuracy reached with a degree of approximation equal to 30 is of the order of 1×10^{-4}, which is good in comparison with the accuracy that can be achieved using Monte Carlo-based methods.

We emphasize that, for Eq. (2.50), calculating the effective drift and diffusion coefficients with our spectral method takes only a few seconds, even for $d = 30$. In comparison, it takes more than a minute to consistently reach an accuracy of 0.001 with the HMM. The computational bottleneck of our method is the solution of the linear system of equations (2.25), the size of which scales as $\mathcal{O}(d^n)$, where n is the dimension of the state space of the fast process. Consequently, assuming that a naive direct method is employed for the solution of this linear system, the overall computational cost of our spectral method scales as $\mathcal{O}(d^{3n})$. The computational cost of the HMM, on the other hand, is discussed in detail in [65] and briefly in Eq. (2.58) below; for the parameters we employed, it scales as $\mathcal{O}(2^{3p})$ in order to guarantee that the error is bounded from above by 2^{-p}, where p is a precision parameter.

Discretization of a multiscale stochastic PDE

As mentioned in the introduction, our numerical method is particularly well-suited for the solution of singularly perturbed stochastic PDEs (SPDEs), and constitutes a very good complement to the method proposed in [1]. Let us recall how the method introduced in [1] works for a singularly perturbed SPDE of the following form

$$\frac{\partial u}{\partial t} = \frac{1}{\varepsilon^2}\mathcal{A}u + \frac{1}{\varepsilon}F(u) + \frac{1}{\varepsilon}Q\dot{W}, \tag{2.51}$$

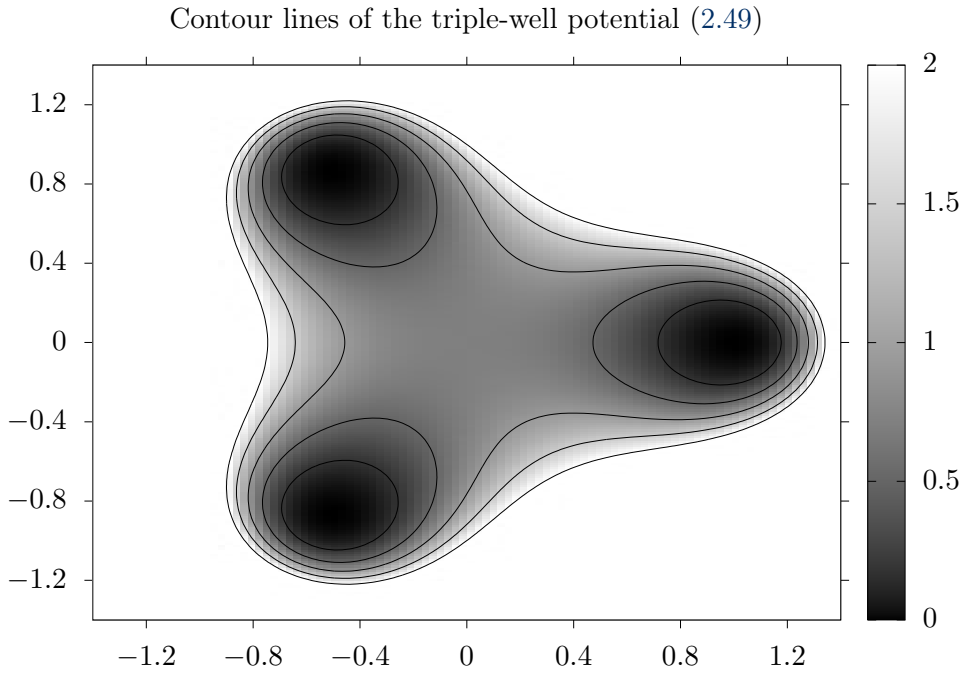


Figure 2.6: Potential (2.49), used in equation (2.50).

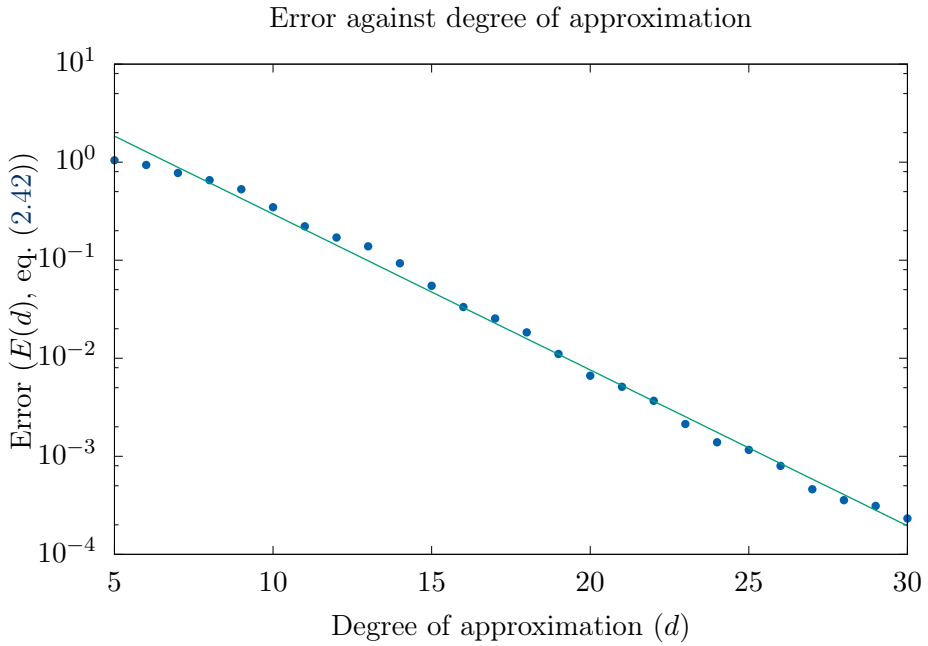


Figure 2.7: Error $E(d)$, see Eq. (2.42), for the fast/slow system (2.50).

posed in a bounded domain of $\mathbf{R}^m$ with suitable boundary conditions. We consider the set-up from [23]. In particular, the operator $\mathcal{A}$ in Eq. (2.51) is a differential operator, assumed to be negative and selfadjoint in a Hilbert space $\mathcal{H}$, and with compact resolvent. It is furthermore assumed that $\mathcal{A}$ has a finite dimensional kernel, denoted by $\mathcal{M}$. The term W denotes a cylindrical Wiener process on $\mathcal{H}$ and Q denotes the covariance operator of the noise, which is assumed to be positive, selfadjoint, and bounded. It is assumed that Q and $\mathcal{A}$ commute, and that the noise acts only on the orthogonal complement of $\mathcal{M}$, denoted by $\mathcal{M}^{\perp}$, see [23, Assumption 2.4]. The

function $F(\cdot)$ is a polynomial function representing a nonlinearity that has to be such that the above scaling makes sense.[8]

Since $\mathcal{A}$ is selfadjoint with compact resolvent, there exists an orthonormal basis of $\mathcal{H}$ consisting of eigenfunctions of $\mathcal{A}$. We denote by $\{\lambda_k, e_k\}$ the eigenvalues and corresponding eigenfunctions of $\mathcal{A}$. We arrange the eigenpairs by increasing absolute value of the eigenvalues, so the m first eigenfunctions are in the kernel of the differential operator, $\mathcal{M} = \text{span}\{e_1, \ldots, e_m\}$. Formally, the cylindrical Brownian motion can be expanded in the basis as $W(t) = \sum_{i=1}^{\infty} e_i\, w_i(t)$, where $\{w_i\}_{i=1}^{\infty}$ are independent Brownian motions. The assumption that the covariance operator Q commutes with the differential operator $\mathcal{A}$ means that this operator satisfies $Q\, e_i = q_i\, e_i$, while the assumption that the noise only acts on $\mathcal{M}^{\perp}$ implies that $q_i = 0$ for $i = 1, 2, \ldots, m$.

We now summarize how the dynamics of the slow modes in (2.51) can be approximated by solving a multiscale system of SDEs using the methodology developed in [1].

First, we write the solution of (2.51) as

$$u = x + y, \quad \text{with } x = \sum_{k=1}^{m} x_k\, e_k \quad \text{and } y = \sum_{k=m+1}^{\infty} y_k\, e_k.$$

Note that $x = \mathcal{P}u$, and $y = (I - \mathcal{P})u$, where $\mathcal{P}$ is the projection operator from $\mathcal{H}$ onto $\mathcal{M}$. By assumption, the noise term can be expanded in the same way, as $\sum_{k=1}^{\infty} q_k\, e_k\, \dot{w}_k(t)$. Substitution of these expansions in the SPDE gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{k=1}^{m} x_k\, e_k \; + \sum_{k=m+1}^{\infty} y_k\, e_k\right) = -\frac{1}{\varepsilon^2}\sum_{k=m+1}^{\infty} \lambda_k\, y_k\, e_k + \frac{1}{\varepsilon}F(u) + \frac{1}{\varepsilon}\sum_{k=m+1}^{\infty} q_k\, e_k\, \dot{w}_k(t).$$

The equations that govern the evolution of the coefficients x_k and y_k can be obtained by taking the inner product (of $\mathcal{H}$) of both sides of the above equation by each of the eigenfunctions of the operator, and using orthonormality :

$$\begin{cases} \dot{x}_i = \dfrac{1}{\varepsilon}\langle F(u), e_i\rangle & i = 1, \ldots, m; \\ \dot{y}_i = -\dfrac{1}{\varepsilon^2}\lambda_i\, y_i + \dfrac{1}{\varepsilon}\langle F(u), e_i\rangle + \dfrac{1}{\varepsilon} q_i\, \dot{w}_i & i = m+1, m+2, \ldots \end{cases} \tag{2.52}$$

Equation (2.52) can be written in the form

$$\begin{cases} \dot{x} = \dfrac{1}{\varepsilon} a(x, y), \\ \dot{y} = \dfrac{1}{\varepsilon^2}\mathcal{A}\, y + \dfrac{1}{\varepsilon} b(x, y) + \dfrac{1}{\varepsilon} Q\, \dot{W}, \end{cases} \tag{2.53}$$

where $a(x, y)$ and $b(x, y)$ are the projections of $F(u)$ on $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively:

$$a(x, y) = \sum_{i=1}^{m} a^i(x, y)\, e_i \quad \text{with} \quad a^i(x, y) = \langle F(x + y), e_i\rangle,$$

[8] i.e., the centering condition is satisfied.

and

$$b(x,y) = \sum_{i=m+1}^{\infty} b^i(x,y)\, e_i \quad \text{with} \quad b^i(x,y) = \langle F(x+y), e_i \rangle.$$

The scale separation now appears clearly. We now truncate the fast process in Eq. (2.53) as $y \approx \sum_{i=m+1}^{m+n} y_i\, e_i$ to derive the following finite dimensional system:

$$\begin{cases} \dot{x}_i \;=\; \dfrac{1}{\varepsilon} a^i(x,y) & i = 1, \ldots, m; \\ \dot{y}_i \;=\; -\dfrac{1}{\varepsilon^2}\lambda_i y_i \;+\; \dfrac{1}{\varepsilon}\, b^i(x,y) \;+\; \dfrac{1}{\varepsilon} q_i\, \dot{w}_i & i = m+1, \ldots m+n, \end{cases} \tag{2.54}$$

In [143], the authors investigate the use of the heterogeneous multiscale method (HMM) for solving the problem (2.54), and show that a good approximation can be obtained using this method. However, when the nonlinearity is a polynomial function of u, the function a in the system above, which also appears on the right-hand side of the Poisson equation, is polynomial in x and y. In addition, the generator of this system of stochastic differential equations is of Ornstein–Uhlenbeck type to leading order, and so its eigenfunctions are Hermite polynomials. This means that the right-hand side can be expanded exactly in Hermite polynomials, and so the exact effective coefficients can be computed. Note that although equivalent, applying the unitary transformation is not necessary in this case, as we can work directly with Hermite polynomials in the appropriate weighted L^2 space.

We consider the SPDE (2.51), with $\mathcal{A} = \frac{\partial^2}{\partial x^2} + 1$ and $F(u) = u^2 \frac{\partial u^2}{\partial x}$, posed on $[-\pi, \pi]$ with periodic boundary conditions:

$$\frac{\partial u}{\partial t} \;=\; \frac{1}{\varepsilon^2}\left(\frac{\partial^2}{\partial x^2} \;+\; 1\right) u \;+\; \frac{1}{\varepsilon}\, u^2\, \frac{\partial u^2}{\partial x} \;+\; \frac{1}{\varepsilon}\, Q\dot{W}. \tag{2.55}$$

In this case the null space of $\mathcal{A}$ is two-dimensional. The eigenfunctions of $\mathcal{A}$ on $[-\pi, \pi]$ with periodic B.C. are

$$e_i \;=\; \begin{cases} \dfrac{1}{\sqrt{\pi}} \sin\left(\dfrac{i+1}{2}\, x\right) & \text{if } i \text{ is odd}, \\ \dfrac{1}{\sqrt{\pi}} \cos\left(\dfrac{i}{2}\, x\right) & \text{if } i \text{ is even}, \end{cases}$$

and the corresponding eigenvalues are $\lambda_i = 1 - \frac{(i+1)^2}{4}$ if i is odd and $\lambda_i = 1 - \frac{i^2}{4}$ if i is even. For the numerical experiments, we chose $q_i = 1/i$ if $3 \le i \le 5$, $q_i = 0$ otherwise, and we used 7 modes for the truncation of the Fourier series. We consider a noise process of the form:

$$Q\dot{W} = \sum_{i=3}^{\infty} q_i\, \dot{w}_i.$$

Following the methodology outlined above, we approximate the solution by a truncated Fourier series:

$$u \;=\; x_1\, e_1 \;+\; x_2\, e_2 \;+\; \sum_{i\,=\,3}^{n+2} y_i\, e_i. \tag{2.56}$$

Substituting in the nonlinearity and taking the inner product with each of the eigenfunctions,

a system of equation of the type (2.54) is obtained. The operator $\mathcal{A}$ and the nonlinearity were chosen so that the centering condition is satisfied. The homogenized equation for the slow variables (x_1, x_2) reads

$$\mathrm{d}X_t = \mathbf{F}(X_t)\,\mathrm{d}t + \mathbf{A}(X_t)\,\mathrm{d}W_t, \tag{2.57}$$

where $\mathbf{F}(\cdot)$ and $\mathbf{A}(\cdot)$ are given by equations (2.4a) and (2.4b), respectively, and W is a standard Wiener process in $\mathbf{R}^2$. The Euler–Maruyama solver was used for both the macro and micro solvers, and the parameters of the HMM were chosen as

$$(\delta t/\varepsilon^2, N_T, M, N, N') = (2^{-p}, 16, 1, 10{\times}2^{3p}, 2^p p). \tag{2.58}$$

Here δt is the time step of the micro-solver, N_T is the number of steps that are omitted in the time-averaging process to reduce transient effects, M is the number of samples used for ensemble averages, and N, N' are the number of time steps employed for the calculation of time averages and the discretization of integrals originating from Feynman–Kac representation formula (2.11), respectively. See [65, 218] for a more detailed description of the method and a detailed explanation of the parameters in (2.58). In Figs. 2.8 and 2.9, we compare the solution obtained using the HMM method (denoted by $\hat{X}^n$, where n is the index of the time step) with the one obtained using our approach (denoted by X^n), using the same macro-solver and the same replica of the driving Brownian motion for both. The initial condition was taken to be $x_1(0) = x_2(0) = 1.2$. Notice that when the value of the parameter p increases, the solution obtained using the HMM converges to the exact solution obtained using the Hermite spectral method.

We now investigate the dependence on the precision parameter p of the error between the homogenized coefficients. The same error measure as in [65] is used to compare the two methods:

$$E_p = \frac{\Delta t}{T}\left(\sum_{n \le T/\Delta t} |\mathbf{F}^p_{HMM}(X^n) - \mathbf{F}_{Sp}(X^n)| + |\mathbf{A}^p_{HMM}(X^n) - \mathbf{A}_{Sp}(X^n)|\right). \tag{2.59}$$

Here $\mathbf{F}^p_{HMM}$ and $\mathbf{A}^p_{HMM}$ are the drift and diffusion coefficients obtained using the HMM with the precision parameter equal to p, while $\mathbf{F}_{Sp}$ and $\mathbf{A}_{Sp}$ are the coefficients given by the Hermite spectral method developed in this chapter. Given the choice of parameters (2.58), the theory developed in [65] predicts that the error should decrease as $\mathcal{O}(2^{-p})$. This error is presented in Fig. 2.10 as a function of the precision parameter p, showing a good agreement with the theory developed in [1, 65].

For the SPDE described above, solving the Poisson equation associated with (2.54) using Hermite polynomials does recover exactly the corresponding effective parameters, and the only source of error is the macroscopic discretization scheme. This is in sharp contrast with the HMM-based method developed in [3], for which the micro-averaging process to recover the effective coefficients represents a non-negligible computational cost.

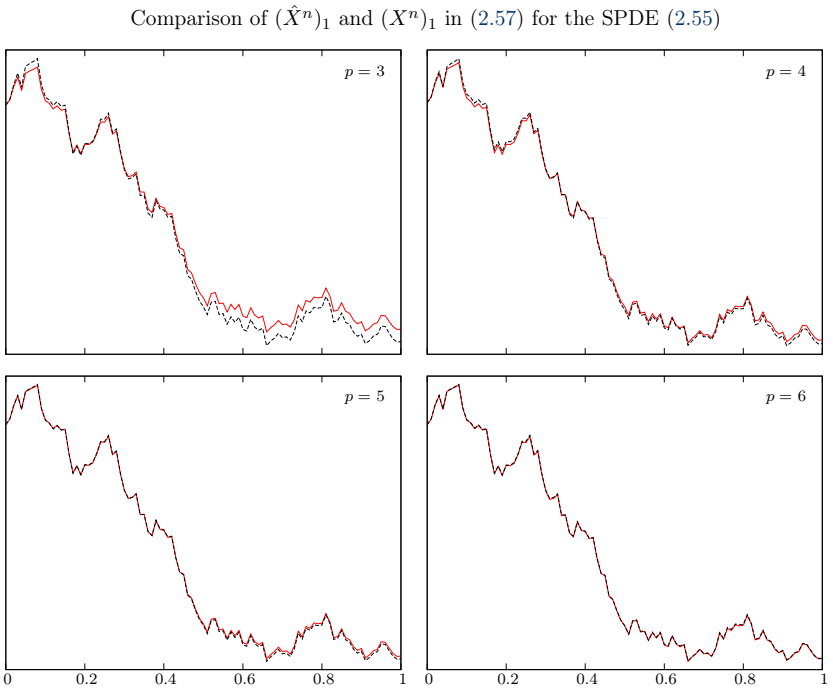


Figure 2.8: Evolution of the coefficient x_1 of the first term in the Fourier expansion (2.56) of the the solution to the SPDE (2.55), obtained numerically by the HMM (black) and the Hermite spectral method (red), for one sample of the driving Brownian motion.

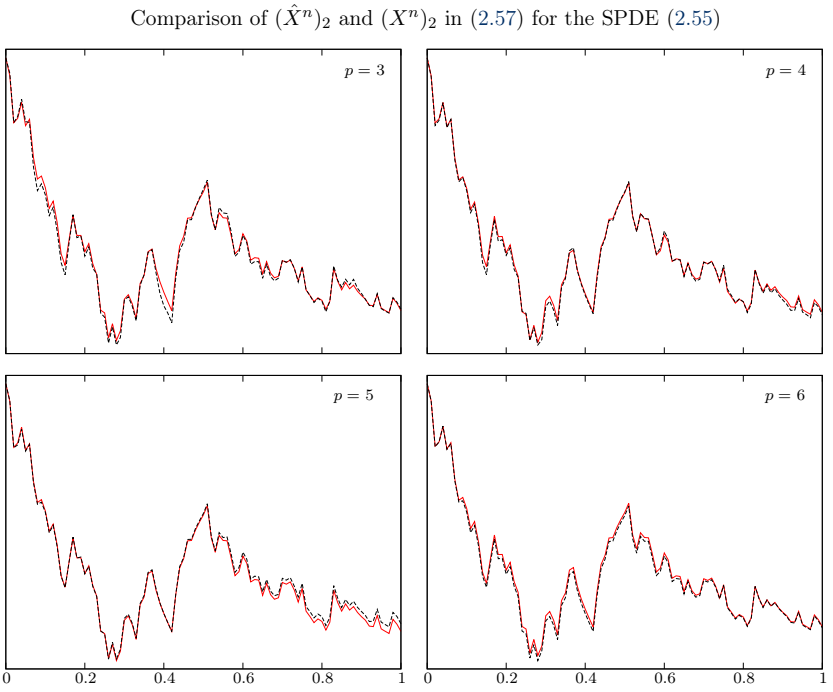


Figure 2.9: Evolution of the coefficient x_2 of the second term in the Fourier expansion (2.56) of the the solution to the SPDE (2.55), obtained numerically by the HMM (black) and the Hermite spectral method (red), for one sample of the driving Brownian motion.

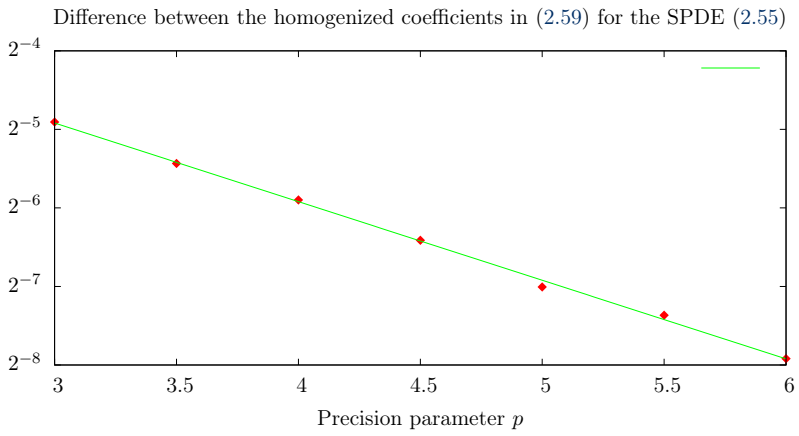


Figure 2.10: Error between the homogenized coefficients (see Eq. (2.59)) for the SPDE (2.55), as a function of the precision parameter p. The green line, obtained by linear fitting, has slope -1.01 in the $p - \log_2(E_p)$ plane, which is close to the theoretical value of -1, showing a perfect agreement with the theory.

2.7 Conclusion and Further Work

In this chapter, we proposed a new approach for the numerical approximation of the slow dynamics of fast/slow SDEs for which a homogenized equation exists. Starting from the appropriate Poisson equation, the same unitary transformation as in [28] was utilized to obtain formulas for the drift and diffusion coefficients in terms of the solution to a Schrödinger equation. This equation is solved at each discrete time by means of a spectral method using Hermite functions, from which approximations of the homogenized drift and diffusion coefficients were calculated. A stochastic integrator was then used to evolve the slow variables by one time step, and the procedure is repeated.

Building on the work of [83], spectral convergence of the homogenized coefficients was rigorously established, from which weak convergence of the discrete approximation in time to the exact homogenized solution was derived. In the final section, the accuracy and efficiency of the proposed methodology were examined through numerical experiments.

The method presented, although not as general as the HMM, has proven more precise and more efficient for a broad class of problems. It performs particularly well for singularly perturbed SPDEs, and constitutes in this case a good complement to the HMM-based method presented in [1]. It also works comparatively very well when the fast dynamics is of relatively low dimension – typically less than or equal to 3 – and especially so when the potential is localized, since fewer Hermite functions are required to accurately resolve the Poisson equations in this situation. Our method also has several advantages compared to the approach taken in [28]: it does not require truncation of the domain, does not require the calculation of the eigenvalues and eigenfunctions of the Schrödinger operator, and has better asymptotic convergence properties.

The limitations of the method are two-fold; its generality is limited by the requirement of the gradient structure for fast dynamics, and its efficiency is limited by the curse of dimensionality, which causes the computation time to become prohibitive when the dimension of the state space

of the fast process increases.

The extent to which some of these constraints can be lifted constitutes an interesting topic for future work. We believe that it is possible to generalize our method to a broader class of problems while retaining its efficiency and accuracy. When the fast process is not reversible, then tools from the spectral theory for non-selfadjoint problems could be used [108]. In the case of the underdamped Langevin dynamics, expansion of the solution in Hermite polynomials in both position and momentum can be performed [137].

On the other hand, high-dimensional integrals could be computed more efficiently. For example, an alternative to the tensorized quadrature approach taken in this work is to use a sparse grid method; such a method can in principle offer the same degree of polynomial exactness with a significantly lower number of nodes, see e.g. [87, 125]. Furthermore, efficient preconditioning techniques can be used in order to reduce the computational cost of the numerical solution. Preconditioning can be naturally implemented for nonreversible fast processes that have a well-defined structure, such as the underdamped Langevin dynamics, [178], or the Markovian approximation of the generalized Langevin equation, [168].

2.A Proof of Eq. (2.36)

In this appendix, we show the Lipschitz continuity of the effective coefficients. The proof relies on the following technical lemma.

Lemma 2.10. *For any $m \in \mathbf{N}$ and any symmetric positive semi-definite matrices $M_1, M_2 \in \mathbf{R}^{m\times m}$, there exists $c(m) > 0$ such that*

$$\sup_{|x|=1} \left| \sqrt{x^T M_2^2 x} - \sqrt{x^T M_1^2 x} \right| \geq c(m) \left| M_2 - M_1 \right|,$$

where $|\cdot|$ is the matrix 2-norm.

Proof. To alleviate the notations, let $M = M_1$ and $\Delta = M_2 - M_1$. The inequality is trivially satisfied when $\Delta = 0$; we assume now that $\Delta \neq 0$. Without loss of generality, we assume that $|\Delta| = 1$, that Δ has an eigenvalue equal to $+1$, and that M is diagonal with its eigenvalues occuring in descending order on the diagonal. For $x \in \mathbf{R}^m$ arbitrary,

$$\begin{aligned}\sqrt{x^T(M+\Delta)^2 x} &= \sqrt{x^T M^2 x + x^T(M\Delta + \Delta M)x + x^T\Delta^2 x} \\ &= \sqrt{\left(\sqrt{x^T M^2 x} + \sqrt{x^T \Delta^2 x}\right)^2 - 2\left(\sqrt{x^T M^2 x}\sqrt{x^T\Delta^2 x} - x^T M\Delta x\right)}.\end{aligned}$$

By Cauchy–Schwarz inequality, the expression in the second brackets is non-negative, so assuming that x is such that the expression in first brackets is positive,

$$\begin{aligned}\left|\sqrt{x^T(M+\Delta)^2 x} - \sqrt{x^T M^2 x}\right| &\geq \frac{\left|2\, x^T M\Delta x + x^T\Delta^2 x\right|}{2(\sqrt{x^T M^2 x} + \sqrt{x^T\Delta^2 x})} \\ &\geq \frac{2\, x^T M\Delta x + x^T\Delta^2 x}{2(\sqrt{x^T M^2 x} + 1)} =: R_{M,\Delta}(x).\end{aligned}$$

To conclude the proof, we show that for any m, M, Δ as above, there exists $x \in \mathbf{R}^m$ such that $R_{M,\Delta}(x)$ is bounded from below by a constant depending only on m. Assume that there exists $\{M_i\}_{i=1}^{\infty}$ and $\{\Delta_i\}_{i=1}^{\infty}$ such that $\sup_{|x|=1} R_{M_i,\Delta_i}(x) \to 0$, and let x_i denote an eigenvector corresponding to the eigenvalue 1 of Δ_i. Passing to a subsequence if necessary, we can assume that $x_i \to x_\infty$. By assumption,

$$R_{M_i,\Delta_i}(x_i) = \frac{2\, x_i^T M_i x_i + 1}{2\sqrt{x_i^T M_i^2 x_i} + 2} \to 0 \Rightarrow \frac{x_i^T M_i x_i}{\sqrt{x_i^T M_i^2 x_i}} \to 0 \Rightarrow x_i^T \left(\frac{M_i}{|M_i|}\right) x_i \to 0,$$

so in particular the first component of x_∞ is equal to 0. Let now $\{y_i\}_{i=1}^{\infty}$ be the sequence obtained from $\{x_i\}_{i=1}^{\infty}$ by replacing the first component by 0. By assumption,

$$R_{M_i,\Delta_i}(y_i) = \frac{(M_i\, y_i)^T (\Delta_i y_i)}{|M_i y_i| + 1} + \frac{|\Delta_i y_i|^2}{2\,|M_i y_i| + 2} \to 0.$$

Since $(\Delta_i y_i - y_i) \to 0$, both terms must tend to 0. This implies that $y_i^T M_i y_i / |M_i y_i| \to 0$, so the second component of x_∞ is 0 too. Continuing in this manner, we obtain $x_\infty = 0$, which is a contradiction because $x_i \to x_\infty$ and $|x_i| = 1$. □

Proof of Eq. (2.36). For the drift coefficient, we use Eq. (2.12) and Assumption 2.3:

$$\begin{aligned}
|\mathbf{F}(a) - \mathbf{F}(b)| \le & \int_{\mathbf{R}^n} |\nabla_x \boldsymbol{\phi}(a,y) - \nabla_x \boldsymbol{\phi}(b,y)|\; |\mathbf{f}(a,y)|\; \mathrm{e}^{-V} \\
& + |\nabla_x \boldsymbol{\phi}(b,y)|\; |\mathbf{f}(a,y) - \mathbf{f}(b,y)|\; \mathrm{e}^{-V}\; \mathrm{d}y \\
\le & |a-b| \int_{\mathbf{R}^n} \sup_{x\in(a,b)} |\nabla_x \nabla_x \boldsymbol{\phi}(x,y)\, \mathbf{f}(a,y)|\; \mathrm{e}^{-V} \\
& + \sup_{x\in(a,b)} |\nabla_x \boldsymbol{\phi}(b,y)\, \nabla_x \mathbf{f}(x,y)|\; \mathrm{e}^{-V}\; \mathrm{d}y \\
\le & |a-b| \int_{\mathbf{R}^n} 2\, C_0\, L\, (1+|y|^{\ell_0})(1+|y|^q)\; \mathrm{e}^{-V}\; \mathrm{d}y.
\end{aligned}$$

The proof of the Lipschitz continuity of $A(\cdot)$ relies on Lemma 2.10. By virtue of this result, it is sufficient to show that, for all $\boldsymbol{\xi} \in \mathbf{R}^m$ with $\boldsymbol{\xi}^T\boldsymbol{\xi} = 1$, $(\boldsymbol{\xi}^T \mathbf{D}(x) \boldsymbol{\xi})^{1/2}$ is Lipschitz continuous with a Lipschitz constant independent of $\boldsymbol{\xi}$. Let $\boldsymbol{\xi} \in \mathbf{R}^m$ be given, and let us introduce the notations $f_\xi = \mathbf{f}^T\, \boldsymbol{\xi}$, $\phi_\xi = \boldsymbol{\phi}^T\, \boldsymbol{\xi}$, and $\boldsymbol{\alpha}_\xi = \boldsymbol{\alpha}^T \boldsymbol{\xi}$. Clearly, ϕ_ξ is the centered solution of $-\mathcal{L}\phi_\xi = f_\xi$, so

$$\boldsymbol{\xi}^T \mathbf{D}(x) \boldsymbol{\xi} = \int_{\mathbf{R}^n} \left(|\boldsymbol{\alpha}_\xi|^2 \; - \; \mathcal{L}\phi_\xi\, \phi_\xi \right) \mathrm{e}^{-V}\; \mathrm{d}y =: D_1(x) + D_2(x).$$

From the triangle inequality and (H_L),

$$\begin{aligned}
\left| \sqrt{D_1(x+\Delta x)} - \sqrt{D_1(x)} \right| &\le \left(\int_{\mathbf{R}^n} |\mathbf{a}_\xi(x+\Delta x, y) - \mathbf{a}_\xi(x,y)|^2\, \mathrm{e}^{-V(y)}\; \mathrm{d}y \right)^{\frac{1}{2}} \\
&\le |\Delta x| \left(\int_{\mathbf{R}^n} C(y)^2\, \mathrm{e}^{-V(y)}\; \mathrm{d}y \right)^{\frac{1}{2}}.
\end{aligned}$$

Likewise, using Eqs. (2.12) and (H_G),

$$\begin{aligned}\left|\sqrt{D_2(x+\Delta x)}-\sqrt{D_2(x)}\right| &\leq \left(\int_{\mathbf{R}^n} |\phi_\xi(x+\Delta x,y)-\phi_\xi(x,y)|\ |f_\xi(x+\Delta x,y)-f_\xi(x,y)|\ \mathrm{e}^{-V(y)}\,\mathrm{d}y\right)^{\frac{1}{2}} \\ &\leq |\Delta x|\left(\int_{\mathbf{R}^n} KL(1+|y|^q)(1+|y|^{m_1})\ \mathrm{e}^{-V(y)}\ \mathrm{d}y\right)^{\frac{1}{2}}.\end{aligned}$$

Thus $\sqrt{\boldsymbol{\xi}^T\mathbf{D}(x)\boldsymbol{\xi}}$ is Lipschitz continuous with a constant independent of $\boldsymbol{\xi}$. □

Chapter 3

Mean field limits for interacting diffusions with colored noise

Abstract

We consider systems of weakly interacting particles driven by colored noise in a bistable potential, and we study the effect of the correlation time of the noise on the bifurcation diagram for the equilibrium states. We accomplish this by solving the corresponding McKean–Vlasov equation using a Hermite spectral method, and we verify our findings using Monte Carlo simulations of the particle system. We consider both Gaussian and non-Gaussian noise processes, and for each model of the noise we also study the behavior of the system in the small correlation time regime using singular perturbation theory for Markov processes. The spectral method that we develop in this paper can be used for solving both linear and nonlinear and nonlocal (mean-field) Fokker–Planck equations, without requiring that they have a gradient structure.

3.1 Introduction . . . 106
3.2 The model . . . 110
3.2.1 Mean-field limit . . . 111
3.2.2 The white noise limit . . . 113
3.3 The numerical method . . . 115
3.3.1 Linear Fokker–Planck equation with white noise . . . 116
3.3.2 McKean–Vlasov equation with white noise . . . 119
3.3.3 Linear Fokker–Planck equation with colored noise . . . 122
3.3.4 McKean–Vlasov equation with colored noise . . . 125
3.3.5 Monte Carlo simulations . . . 127
3.4 Asymptotic analysis for the Galerkin formulation . . . 127
3.4.1 Suitable index sets . . . 128
3.4.2 Effective drift and diffusion coefficients . . . 130
3.4.3 Numerical verification of the rates of convergence . . . 131
3.5 Results: effect of colored noise on bifurcations . . . 133

3.5.1 Construction of the bifurcation diagrams for the mean-field equation . . 133
3.5.2 Gaussian case 134
3.5.3 Non-Gaussian noise 135
3.5.4 Dependence of the critical temperature on ε 136
3.6 Conclusions 137
3.A Proof of Theorem 3.1 138
3.B Asymptotic analysis 141
3.B.1 Ornstein–Uhlenbeck noise 145
3.B.2 Harmonic noise 146
3.B.3 General reversible noise 148
3.C Additional numerical results 148
3.C.1 Convergence of the method for the linear Fokker-Planck equation 148
3.C.2 Convergence to the white noise limit 150

3.1 Introduction

Systems of interacting particles appear in a wide variety of applications, ranging from plasma physics and galactic dynamics [21] to mathematical biology [72, 149], the social sciences [84, 161], active media [13], dynamical density functional theory (DDFT) [91, 90] and machine learning [148, 188, 204]. They can also be used in models for cooperative behavior [51], opinion formation [84], and risk management [85], and also in algorithms for global optimization [180].

In most of the existing works on the topic, the particles are assumed to be subject to thermal additive noise that is modeled as a white noise process, i.e. a mean-zero Gaussian stationary process that is delta-correlated in time. There is extensive literature studying the behavior of these systems; we mention for example works on the rigorous passage to the mean-field limit [165], the long-time behavior of solutions (see [51, 196] for a case of a ferromagnetic (quartic) potential, and [93] for more general potentials), multiscale analysis [94], and phase transitions [213].

In a more realistic scenario, the system has memory and the hypothesis of Markovianity does not hold [115, 118, 119]. This memory can be modeled by using *colored* noise, i.e. noise with a nonzero correlation time (or, more precisely, a nonsingular autocorrelation function), which is the approach we take in this paper. For simplicity, we will assume that the noise is additive and that it can be represented by a finite-dimensional Markov process, as in the recent study [62] on mean-field limits for non-Markovian interacting particles.

In this paper we will study the dynamics of a system of interacting particles of the Desai–Zwanzig type, interacting via a quadratic Curie–Weiss potential. The system of interacting particles is modeled by a system of stochastic differential equations (SDEs):

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = -\left(V'(X_t^i) + \theta\left(X_t^i - \frac{1}{N}\sum_{j=1}^{N} X_t^j\right)\right) + \sqrt{2\beta^{-1}}\,\xi_t^i, \qquad i = 1, \dots, N, \tag{3.1}$$

where N is the number of particles, $V(\cdot)$ is a confining potential, θ is the interaction strength, β is the inverse temperature of the system, and ξ_t^i are independent, identically distributed (i.i.d.) noise processes.

Before discussing the Desai–Zwanzig model with colored noise, we present a brief overview of known results [51, 196] for the white noise problem. When ξ_t^i are white noise processes, we can pass to the mean-field limit $N \to \infty$ in Eq. (3.1) and obtain a nonlinear and nonlocal Fokker–Planck equation, known in the literature as a McKean–Vlasov equation, for the one-particle distribution function $\rho(x,t)$:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left(V'(x)\,\rho + \theta\,\left(x - \int_{\mathbf{R}} x\,\rho(x,t)\,\mathrm{d}x \right)\,\rho + \beta^{-1}\,\frac{\partial \rho}{\partial x} \right). \tag{3.2}$$

The McKean–Vlasov equation (3.2) is a gradient flow with respect to the quadratic Wasserstein metric for the free energy functional

$$\mathcal{F}[\rho] = \beta^{-1} \int_{\mathbf{R}} \rho(x)\,\ln \rho(x)\,\mathrm{d}x + \int_{\mathbf{R}} V(x)\,\rho(x)\,\mathrm{d}x + \frac{\theta}{2}\int_{\mathbf{R}}\int_{\mathbf{R}} F(x-y)\,\rho(x)\,\rho(y)\,\mathrm{d}x\,\mathrm{d}y, \tag{3.3}$$

where $F(x) := x^2/2$ is the interaction potential. The long-time behavior of solutions depends on the number of local minima of the confining potential V [213]. It follows directly from Eq. (3.2) that any steady-state solution $\rho_\infty(x)$ solves, together with its first moment, the following system of equations:

$$\frac{\partial}{\partial x}\left(V'(x)\,\rho_\infty(x) + \theta\,(x-m)\,\rho_\infty(x) + \beta^{-1}\,\frac{\partial \rho_\infty}{\partial x}(x) \right) = 0, \tag{3.4a}$$

$$m = \int_{\mathbf{R}} x\,\rho_\infty(x)\,\mathrm{d}x. \tag{3.4b}$$

Since Eq. (3.4a) is, for m fixed, the stationary Fokker–Planck equation associated with the overdamped Langevin dynamics in the confining potential

$$V_{\mathrm{eff}}(x; m, \theta) = V(x) + \frac{\theta}{2}(x-m)^2, \tag{3.5}$$

solutions can be expressed explicitly as

$$\rho_\infty(x; m, \beta, \theta) := \frac{1}{\mathcal{Z}(m,\beta,\theta)}\,\mathrm{e}^{-\beta V_{\mathrm{eff}}(x;m,\theta)}, \tag{3.6}$$

where $\mathcal{Z}(m,\beta,\theta)$ is the normalization constant (partition function); see [51, 93, 94] for more details. By substitution in Eq. (3.4b), a scalar fixed-point problem is obtained for m, the *self-consistency equation*:

$$m = \int_{\mathbf{R}} x\,\rho_\infty(x; m, \beta, \theta)\,\mathrm{d}x =: R(m,\beta,\theta). \tag{3.7}$$

The stability of solutions to Eq. (3.4) depends on whether they correspond to a local minimum (stable) or to a local maximum/saddle point (unstable) of the free energy functional. The free energy along the one-parameter family (3.6), with parameter m, can be calculated explicitly as

follows [94],

$$\mathcal{F}[\rho_\infty(x;m,\beta,\theta)] = -\beta^{-1}\,\ln \mathcal{Z}(m,\beta,\theta) - \frac{\theta}{2}\big(R(m,\beta,\theta)-m\big)^2,$$

from which we conclude that

$$\frac{\partial}{\partial m}\mathcal{F}[\rho_\infty(x;m,\beta,\theta)] = -\beta\theta^2\big(R(m,\beta,\theta)-m\big)\int_{\mathbf{R}}\big(x-R(m,\beta,\theta)\big)^2\,\rho_\infty(x;m,\beta,\theta)\,\mathrm{d}x.$$

Though incomplete, this informal argument suggests that the stability of a steady-state solution can also be inferred from the slope of $R(m,\beta,\theta)-m$ at the corresponding value of m: if this slope is positive, the equilibrium is unstable, and conversely. The self-consistency map and the free energy of $\rho_\infty(x;m,\beta,\theta)$, for a range of values of m, are illustrated in Fig. 3.1 for the bistable potential $V(x)=\frac{x^4}{4}-\frac{x^2}{2}$. It is well-known that, when $V(\cdot)$ is an even potential, Eq. (3.2)

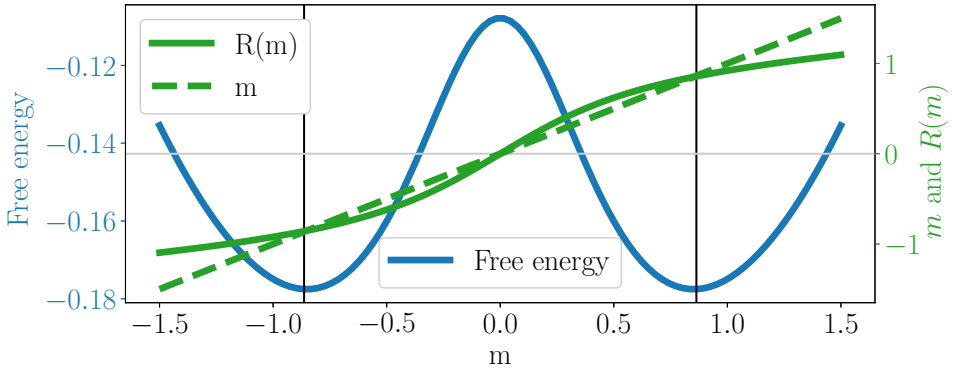


Figure 3.1: Free energy (3.3) of the one-parameter family (3.6) of probability densities that solve Eq. (3.4a) for some value of m (in blue), and associated first moment $R(m)$ (in green), for fixed $\theta=1$ and $\beta=5$. Along the one-parameter family, $m=0$ is a local maximum of the free energy, and it therefore corresponds to an unstable steady state of the McKean–Vlasov equation.

possesses a unique, mean-zero steady-state solution for sufficiently large temperatures (i.e., small β). As the temperature decreases, this solution loses its stability and two new solutions of the self-consistency equation emerge, corresponding to a pitchfork bifurcation; see [51, 94] for details.

As mentioned above, in this paper we focus on the case where the noise processes ξ_t^i in Eq. (3.1) have a nonzero correlation time, and in particular we assume that each noise process can be represented using a (possibly multi-dimensional) SDE, in which case Eq. (3.1) leads to a Markovian system of SDEs in an extended phase space. The colored noise will be modeled by either an Ornstein–Uhlenbeck process, harmonic noise [174, Example 8.2], or a non-Gaussian reversible diffusion process.

Though more realistic, the use of colored noise presents us with some difficulties. First, the introduction of an extra SDE for the noise breaks the gradient structure of the problem; while we can still pass formally to the limit $N\to\infty$ in Eq. (3.1) and obtain a McKean–Vlasov equation for the associated one-particle distribution function, it is no longer possible to write a free energy functional, such as Eq. (3.3), that is dissipated by this equation. Second, the McKean–Vlasov equation is now posed in an extended phase space, which increases the computational cost of its numerical solution via PDE methods. And third, it is no longer possible to obtain an explicit

expression for the one-parameter family of (possible) stationary solutions to the mean-field equation, as was possible in Eq. (3.6), which renders the calculation of steady states considerably more difficult.

When the correlation time of the noise is small, this latter difficulty can be somewhat circumvented by constructing an approximate one-parameter family of solutions through appropriate asymptotic expansions in terms of the correlation time, from which steady-state solutions of the McKean–Vlasov dynamics can be extracted by solving a self-consistency equation similar to (3.7), see Eq. (3.22). See Section 3.2.2 and Section 3.B in the appendix. Outside of the small correlation time regime, however, finding the steady-states of the McKean–Vlasov equation requires a numerical method for PDEs in all but the simplest cases.

In this work, we propose a novel Hermite spectral method for the time-dependent and steady-state equations, applicable to the cases of both white and colored noise. Discretized in a basis of Hermite functions, the McKean–Vlasov equation becomes a system of ordinary differential equations with a quadratic nonlinearity originating from the interaction term. In contrast with other discretization methods for PDEs, the use of (possibly rescaled) Hermite functions for the problem under consideration leads to an efficient numerical method, first because Hermite functions have very good approximation properties in L^2, but also because all the differential operators appearing in the McKean Vlasov equation lead to sparse matrices in Hermite space, with a small bandwidth related to the polynomial degree of V (provided that a suitable ordering of the multi-indices is employed). To solve the finite-dimensional system of equations obtained after discretization of the time-dependent equation, we employ either the Runge–Kutta 45 method (RK45) or a linear, semi-implicit time-stepping scheme.

To assess the performance of our numerical method, we compare its efficiency in the white noise case with that of the finite volume scheme developed in [37], the applicability of which depends on the existence a gradient structure of Eq. (3.2). We also verify that our results agree with known analytical solutions in simple settings, and with explicit asymptotic expansions in the small correlation time regime. We then use our spectral method, together with asymptotic expansions and Monte Carlo (MC) simulations of the particle system, to construct the bifurcation diagram of the first moment of the steady-state solutions as a function of the inverse temperature.

For the reader's convenience, we summarize here the main results of this paper:

1. The systematic study of the effect of colored noise, both Gaussian and non-Gaussian, on the long-time behavior of the McKean–Vlasov mean-field equation, including the effect of colored noise on the structure and properties of phase transitions.

2. The development and analysis of a spectral numerical method for the solution of linear or nonlinear, local or nonlocal Fokker–Planck-type equations. In particular, our method does not depend on an underlying gradient structure for the PDE.

The rest of the paper is organized as follows. In Section 3.2, we present the models for the colored noise and we derive formally the mean-field McKean–Vlasov equation associated with the interacting particle system. In Section 3.3, we present the numerical methods used to (a) solve the time-dependent and steady-state Fokker–Planck (or McKean–Vlasov) equations and

(b) solve the finite-dimensional system of interacting diffusions (3.1). In Section 3.4, we study the performance of our numerical method in the small correlation time regime, and we verify numerically the convergence rates to the white noise solution in the limit where the correlation time tends to 0. In Section 3.5, we describe our methodology for constructing the bifurcation diagrams and we present the associated results. Section 3.6 is reserved for conclusions and perspectives for future work.

3.2 The model

We consider the following system of weakly interacting diffusions,

$$\mathrm{d}X_t^i = -\left(V'(X_t^i) + \theta\left(X_t^i - \frac{1}{N}\sum_{j=1}^{N} X_t^j\right)\right)\mathrm{d}t + \sqrt{2\beta^{-1}}\,\eta_t^i\,\mathrm{d}t, \quad 1 \le i \le N, \tag{3.8}$$

where the noise processes η_t^i are independent, mean-zero, second-order stationary processes with almost surely continuous paths and an autocorrelation function $K(t)$. In the rest of this paper we will assume that the interaction strength θ is fixed and equal to 1, and we will use the inverse temperature β^{-1} as the bifurcation parameter. We will consider two classes of models for the noise: Gaussian stationary noise processes with an exponential correlation function, and non-Gaussian noise processes that we construct by using the overdamped Langevin dynamics in a non-quadratic potential.

Gaussian noise Stationary Gaussian processes in $\mathbf{R}^n$ with continuous paths and an exponential autocorrelation function are solutions to an SDE of Ornstein–Uhlenbeck type:

$$\mathrm{d}\mathbf{Y}_t^i = \mathbf{A}\,\mathbf{Y}_t^i\,\mathrm{d}t + \sqrt{2}\,\mathbf{D}\,\mathrm{d}\mathbf{W}_t^i, \qquad i = 1, \dots, N, \tag{3.9}$$

where $\mathbf{A}, \mathbf{D}$ are $n \times n$ matrices satisfying Kalman's rank condition [147, Chapter 9], and $\mathbf{W}_t^i$, $1 \le i \le n$, are independent white noise processes in $\mathbf{R}^n$. We assume here that the noise is obtained by projection as $\eta_t^i = \langle \mathbf{Y}_t^i, \mathbf{y}_\eta \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidian inner product, for some vector $\mathbf{y}_\eta \in \mathbf{R}^n$. Throughout this paper we will consider two particular examples, namely the scalar OU process and the harmonic noise [174, Chapter 8].

(OU) Scalar Ornstein–Uhlenbeck process:

$$\mathrm{d}\eta_t^i = -\eta_t^i\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i.$$

The associated autocorrelation function is

$$K_{OU}(t) = e^{-|t|}.$$

(H) Harmonic noise:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & -\gamma \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{\gamma} \end{pmatrix}, \quad \mathbf{y}_\eta = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

In this case the noise is the solution to the Langevin equation, with the first and second components of $\mathbf{Y}$ corresponding to the position and velocity, respectively. Throughout this paper we will assume $\gamma = 1$ for simplicity. The associated autocorrelation function of η^i is given by

$$K_H(t) = e^{-\frac{|t|}{2}} \left(\cos\left(\frac{\sqrt{3}}{2}t\right) + \frac{\sqrt{3}}{3} \sin\left(\frac{\sqrt{3}}{2}t\right) \right).$$

Non-Gaussian noise In this case, instead of Eq. (3.9) we consider

$$\mathrm{d}\eta_t^i = -V_\eta'(\eta_t^i)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i,$$

where now V_η is a smooth non-quadratic confining potential satisfying the mean-zero condition:

$$\int_{\mathbf{R}} \eta\, \mathrm{e}^{-V_\eta(\eta)}\, \mathrm{d}\eta = 0. \tag{3.10}$$

We consider the following choices for V_η:

(B) The bistable potential $V_\eta(\eta) = \eta^4/4 - \eta^2/2$.

(NS) The shifted tilted bistable potential

$$V_\eta(\eta) = \frac{(\eta-\alpha)^4}{4} - \frac{(\eta-\alpha)^2}{2} + (\eta - \alpha), \tag{3.11}$$

with the constant $\alpha \approx 0.885$ such that Eq. (3.10) is satisfied.

Remark 3.1. For the two Gaussian noise processes we consider, it would have been equivalent (by a change of variables) to include the inverse temperature β in the noise equation Eq. (3.9) rather than in Eq. (3.8). This is not the case for non-Gaussian noise processes, for which including the temperature in the noise equation leads to an effective diffusion coefficient, in the limit as the correlation time tends to 0, with a nonlinear dependence on β. ⊘

3.2.1 Mean-field limit

In this subsection, we pass to the limit $N \to \infty$ in Eq. (3.8) and we derive the corresponding McKean–Vlasov equation. These formal calculations presented below can be justified rigorously using the results in [63, 158]. We consider the system of SDEs (3.8) and we denote by $P_N(x_1, \dots, x_N, \mathbf{y}_1, \dots, \mathbf{y}_N, t)$ the corresponding N-particle distribution function, the solution of the N-particle Fokker–Planck equation. Here $\mathbf{y}_i$ are the noise variables of the i-th particle, and we denote its components $\mathbf{y}_i = \eta_i$ for scalar noise and $\mathbf{y}_i = (\eta_i, \lambda_i)$ for harmonic noise. We first note that the stochastic system (3.1) is exchangeable [51], i.e. the law of $\{x_i : i = 1, \dots, N\}$ is

equal to that of $\{x_{\pi_i} : i = 1, \dots, N\}$ for every permutation π of $\{1, \dots, N\}$. We assume chaotic initial data,

$$P_N(x_1, \dots, x_N, \mathbf{y}_1, \dots, \mathbf{y}_N, 0) = \prod_{\ell=1}^{N} \rho_0(x_\ell, \mathbf{y}_\ell), \tag{3.12}$$

for some probability density function ρ_0. The density P_N satisfies the N-particle Fokker–Planck equation

$$\frac{\partial P_N}{\partial t} = \sum_{i=1}^{N} \frac{\partial}{\partial x_i} \left(V'(x_i) P_N + \theta \left(x_i - \frac{1}{N} \sum_{j=1}^{N} x_j \right) P_N - \sqrt{2\beta^{-1}} \left\langle \mathbf{y}_i, \mathbf{y}_\eta \right\rangle P_N \right) + \sum_{i=1}^{N} \mathcal{L}^*_{\mathbf{y}_i} P_N, \tag{3.13}$$

where the operator $\mathcal{L}^*_{\mathbf{y}_i}$ depends on the noise process:

$$\mathcal{L}^*_{\mathbf{y}_i} \rho = \begin{cases} \partial_{\eta_i} \left(\eta_i \, \rho + \partial_{\eta_i} \rho \right), & \text{for scalar OU noise,} \\ \partial_{\lambda_i} \left(\lambda_i \, \rho + \partial_{\lambda_i} \rho \right) + \left(\eta_i \, \partial_{\lambda_i} \rho - \lambda_i \, \partial_{\eta_i} \rho \right), & \text{for harmonic noise,} \\ \partial_{\eta_i} \left(V'_{\eta_i} \, \rho + \partial_{\eta_i} \rho \right), & \text{for non-Gaussian noise.} \end{cases}$$

We make the standard mean-field ansatz, see [15, 153]:

$$P_N(x_1, \dots, x_N, \mathbf{y}_1, \dots, \mathbf{y}_N, t) \approx \prod_{\ell=1}^{N} \rho(x_\ell, \mathbf{y}_\ell, t), \tag{3.14}$$

where $\rho(\cdot, \cdot, t)$ is a normalized probability density for all values of t. We note that the right-hand side in Eq. (3.14) is invariant under permutations of the coordinate pairs $(x_i, \mathbf{y}_i)$, which is consistent with the exchangeability of the stochastic system. Substituting Eq. (3.14) into (3.13), and integrating out the $2(N-1)$ variables $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N$ and $\mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, \dots, \mathbf{y}_N$, we obtain the following PDE for $\rho(x_i, \mathbf{y}_i, t)$:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x_i} \left(V' \rho + \theta \left(\frac{N-1}{N} \right) \left(x_i - \int_{\mathbf{R}} \int_{\mathbf{R}^n} x \, \rho(x, \mathbf{y}, t) \, \mathrm{d}\mathbf{y} \, \mathrm{d}x \right) \rho - \sqrt{2\beta^{-1}} \left\langle \mathbf{y}_i, \mathbf{y}_\eta \right\rangle \rho \right) + \mathcal{L}^*_{\mathbf{y}_i} \rho, \tag{3.15}$$

We see that, apart from the nonlinear, nonlocal interaction term, all the terms in Eq. (3.15) are similar to those in Eq. (3.13). Taking formally the limit as $N \to \infty$, relabeling $(x_i, \mathbf{y}_i)$ as $(x, \mathbf{y})$, and ignoring the $\mathcal{O}(1/N)$ terms, we obtain the following McKean–Vlasov equation

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x} \left(V' \rho + \theta \left(x - m(t) \right) \rho - \sqrt{2\beta^{-1}} \left\langle \mathbf{y}_i, \mathbf{y}_\eta \right\rangle \rho \right) + \mathcal{L}^*_{\mathbf{y}} \rho, \tag{3.16a}$$

with the dynamic constraint

$$m(t) = \int_{\mathbf{R}} \int_{\mathbf{R}^n} x \, \rho(x, \mathbf{y}, t) \, \mathrm{d}\mathbf{y} \, \mathrm{d}x, \tag{3.16b}$$

and with the initial condition $\rho(x, \mathbf{y}, 0) = \rho_0(x, \mathbf{y})$, which follows from Eq. (3.12).

Remark 3.2. It is possible to derive the McKean–Vlasov equation (3.16) under a weaker assumption than the product form (3.14). From the N-particle density P_N, we can define the reduced one-

body $\rho_i(x_i, \mathbf{y}_i, t)$ and two-body $\rho_{i,j}(x_i, \mathbf{y}_i, x_j, \mathbf{y}_j, t)$ probability distributions as

$$\rho_i(x_i, \mathbf{y}_i, t) = \int_{(\mathbf{R}\times\mathbf{R}^n)^{N-1}} P_N(x_1, \dots, x_N, \mathbf{y}_1, \dots, \mathbf{y}_N, t) \prod_{k\neq i} (\mathrm{d}\mathbf{y}_k \, \mathrm{d}x_k),$$
$$\rho_{i,j}(x_i, \mathbf{y}_i, x_j, \mathbf{y}_j, t) = \int_{(\mathbf{R}\times\mathbf{R}^n)^{N-2}} P_N(x_1, \dots, x_N, \mathbf{y}_1, \dots, \mathbf{y}_N, t) \prod_{\substack{k\neq i \\ k\neq j}} (\mathrm{d}\mathbf{y}_k \, \mathrm{d}x_k), \quad i \neq j.$$

Since the stochastic system (3.1) is exchangeable and we assumed chaotic initial data, the reduced one-body and two-body probability distributions are independent of the particle indices: $\rho_i = \rho$ and $\rho_{i,j} = \rho_2$ when $i \neq j$, for some probability distributions ρ and ρ_2. Integrating the N-particle Fokker–Planck equation (3.13) in all directions but the i-th, we obtain

$$\begin{aligned}\frac{\partial \rho}{\partial t} =& \frac{\partial}{\partial x_i}\left(V'\rho - \sqrt{2\beta^{-1}}\,\langle \mathbf{y}_i, \mathbf{y}_\eta\rangle\,\rho\right) + \mathcal{L}^*_{\mathbf{y}_i}\rho \\ &+ \frac{\partial}{\partial x_i}\left(\frac{\theta}{N}\sum_{j=1}^{N}\int_{\mathbf{R}}\int_{\mathbf{R}^n}(x_i - x_j)\,\rho_2(x_i, \mathbf{y}_i, x_j, \mathbf{y}_j, t)\,\mathrm{d}\mathbf{y}_j\,\mathrm{d}x_j\right).\end{aligned}$$

To further simplify this equation, we need the product form for only the two-body distribution: employing the ansatz $\rho_2(x_i, \mathbf{y}_i, x_j, \mathbf{y}_j, t) = \rho(x_i, \mathbf{y}_i, t)\,\rho(x_j, \mathbf{y}_j, t)$, we recover Eq. (3.15). The reason why this alternative approach is possible is that only two-body interactions are present in the particle system; see e.g. [91] for more details. ⊘

The main goal of this paper is the study of the effect of colored noise on the structure of the bifurcation diagram for the McKean–Vlasov equation with colored noise, Eqs. (3.16a) and (3.16b). In other words, we want to gain insight into the number of solutions to the following stationary PDE and associated constraint (self-consistency equation):

$$\frac{\partial}{\partial x}\left(V'(x)\,\rho + \theta\,(x - m)\,\rho - \sqrt{2\beta^{-1}}\,\langle \mathbf{y}_i, \mathbf{y}_\eta\rangle\,\rho\right) + \mathcal{L}^*_{\mathbf{y}}\rho = 0, \tag{3.17a}$$

$$m = \int_{\mathbf{R}}\int_{\mathbf{R}^n} x\,\rho(x, \mathbf{y})\,\mathrm{d}\mathbf{y}\,\mathrm{d}x. \tag{3.17b}$$

Although there still exists, for fixed β, a one-parameter family of solutions to (3.17a) (with parameter m), which we will denote by $\{\rho_\infty(x, \mathbf{y}; m, \beta)\}_{m\in\mathbf{R}}$, no closed form is available for these solutions, because the detailed balance condition no longer holds in the presence of colored noise, i.e. the probability flux at equilibrium does not vanish; see [174, Section 4.6].

3.2.2 The white noise limit

To study the limit of small correlation time, it will be convenient to rescale the noise as

$$\eta^i_t \to \zeta\,\eta^i_{t/\varepsilon^2}/\varepsilon,$$

where ε is a time scaling parameter, and ζ is a model-dependent parameter ensuring that the autocorrelation function of the rescaled noise, given by $\zeta^2 K(t/\varepsilon^2)/\varepsilon^2$, satisfies

$$\int_0^\infty \zeta^2 K(t/\varepsilon^2)/\varepsilon^2 \,\mathrm{d}t = \int_0^\infty \zeta^2 K(t)\,\mathrm{d}t = \frac{1}{2}.$$

Then the autocorrelation of the noise converges to a Dirac delta when $\varepsilon \to 0$, and it can be shown that, in this limit, the solution of Eq. (3.8) converges to that of

$$\mathrm{d}X_t^i = \left(-V'(X_t^i) - \theta\left(X_t^i - \frac{1}{N}\sum_{j=0}^{N} X_t^j\right)\right)\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W_t^i, \quad i = 1, \dots N,$$

where W^i, $i = 1, \dots N$, are independent Wiener processes; see [22] and [175, Chapter 11]. While not strictly necessary, including the parameter ζ is convenient to obtain simpler formulas. The value of ζ for each of the noise models considered in this paper is presented in Table 3.1. For the models **B** and **NS**, ζ was calculated numerically and rounded to three significant figures in this table.

Table 3.1: Value of ζ

Model	**OU**	**H**	**B**	**NS**
ζ	$1/\sqrt{2}$	$1/\sqrt{2}$	0.624	0.944

In view of the convergence of the solution of the finite-dimensional particle system when $\varepsilon \to 0$, we expect that also the x-marginals of the steady-state solutions to the McKean–Vlasov equation with colored noise, obtained by solving Eqs. (3.17a) and (3.17b), should converge to their white-noise counterparts as $\varepsilon \to 0$. It turns out that this is the case and, using asymptotic techniques from [115], it is possible to approximate the solutions $\rho_\infty(x, \mathbf{y}; m, \beta)$ to Eq. (3.17a) by a power series expansion in ε; using a superscript to emphasize the dependence on ε,

$$\rho_\infty^\varepsilon(x, \mathbf{y}; m, \beta) = p_0(x, \mathbf{y}; m, \beta) + \varepsilon\, p_1(x, \mathbf{y}; m, \beta) + \varepsilon^2\, p_2(x, \mathbf{y}; m, \beta) + \cdots, \tag{3.18}$$

From Eq. (3.18), we obtain a power series expansion for the x-marginal by integrating out the noise variable:

$$\begin{aligned}\rho_\infty^\varepsilon(x; m, \beta) &= \int_{\mathbf{R}^n} \rho_\infty^\varepsilon(x, \mathbf{y}; m, \beta)\,\mathrm{d}\mathbf{y} \\ &=: \rho_\infty(x; m, \beta) + \varepsilon\, p_1(x; m, \beta) + \varepsilon^2\, p_2(x; m, \beta) + \cdots,\end{aligned} \tag{3.19}$$

The methodology to obtain expressions for the terms works by substituting Eq. (3.18) in Eq. (3.17a) and grouping the terms in powers of ε in the resulting equation. This leads to a sequence of equations that can be studied using standard techniques. Details of the analysis leading to an explicit expression of the first nonzero correction in (3.19) can be found in [115, Section 8] for the particular case of the **OU** noise, and in Section 3.B for the other noise models we consider.

The order of the first nonzero correction in this expansion depends on the model: it is equal to 1 for model **NS**, to 2 for models **OU** and **B**, and to 4 for model **H**. In all cases, the first

nontrivial term in the series expansion (3.19) can be calculated explicitly (possibly up to constant coefficients that have to be calculated numerically). For completeness, we present the expression of the first nonzero correction for the scalar Ornstein–Uhlenbeck and harmonic noise models. For scalar Ornstein–Uhlenbeck noise, omitting the dependence of V_{eff} (the effective potential defined in Eq. (3.5)) on m for notational convenience, we have

$$\rho_\infty^\varepsilon(x;m,\beta) = \rho_\infty(x;m,\beta)\left[1+\varepsilon^2\left(C_{OU} - \frac{\beta}{2}\left(V_{\text{eff}}'(x)\right)^2 + V_{\text{eff}}''(x)\right)\right] + \mathcal{O}(\varepsilon^4), \tag{3.20}$$

and for the case of harmonic noise,

$$\begin{aligned}\rho_\infty^\varepsilon(x;m,\beta) = \rho_\infty(x)\Bigg[1+\varepsilon^4\Bigg(&C_H + \frac{1}{2}\left(V_{\text{eff}}''(x)\right)^2 - \beta\left(V_{\text{eff}}'(x)\right)^2 V_{\text{eff}}''(x)\\ &+2\,V_{\text{eff}}'(x)V_{\text{eff}}'''(x) - \frac{1}{\beta}V_{\text{eff}}^{(iv)}(x) + \beta\int_0^x \left(V_{\text{eff}}''(\xi)\right)^2 V_{\text{eff}}'(\xi)\,\mathrm{d}\xi\Bigg)\Bigg] + \mathcal{O}(\varepsilon^6).\end{aligned} \tag{3.21}$$

Here $C_{OU} = C_{OU}(m,\beta)$ and $C_H = C_H(m,\beta)$ are constants such that the corrections integrate to 0.

Taking into account only the first nontrivial correction, the order of which we denote by δ, the steady-state solutions to the McKean–Vlasov equation with colored noise can be approximated by solving the following approximate self-consistency equation:

$$\begin{aligned} m &= R_0(m,\beta) + \varepsilon^\delta R_\delta(m,\beta)\\ &:= \int_{\mathbf{R}} x\,\rho_\infty(x;m,\beta)\,\mathrm{d}x + \varepsilon^\delta\int_{\mathbf{R}} x\,p_\delta(x;m,\beta)\,\mathrm{d}x\\ &\approx R(m,\beta) := \int_{\mathbf{R}} x\,\rho_\infty^\varepsilon(x;m,\beta)\,\mathrm{d}x.\end{aligned} \tag{3.22}$$

We show in Fig. 3.2 that the equation $R_0(m) + \varepsilon^2 R_2(m) = m$ admits three solutions for $\beta = 10$ in the case of **OU** noise, similarly to the case of white noise. This figure was generated using the asymptotic expansion (3.20) with the parameters $\theta = 1$, $\varepsilon = 0.1$.

3.3 The numerical method

In this section, we describe the spectral numerical method that we will use in order to solve the time-dependent McKean–Vlasov equation, Eqs. (3.16a) and (3.16b), as well as the steady-state equation, Eqs. (3.17a) and (3.17b). Before looking at colored noise, we consider the case of white noise, for which our method can be tested against the results in [94], which were obtained using the finite volume scheme developed in [37].

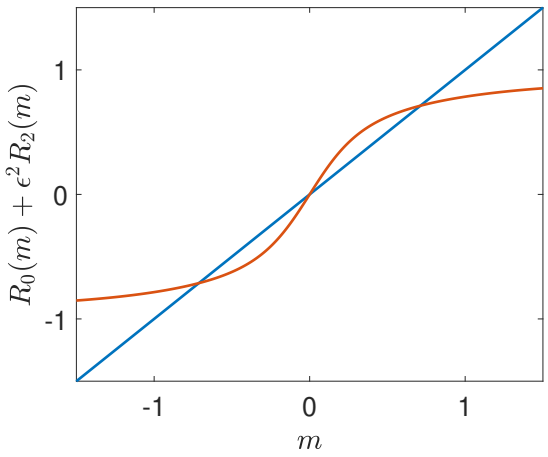


Figure 3.2: Truncated asymptotic expansion of the self-consistency map, $R_0 + \varepsilon^2 R_2$, as a function of m (red line) compared to $y = m$ (blue line) for the scalar Ornstein–Uhlenbeck noise, with $\beta = 10$, $\theta = 1$, $\varepsilon = 0.1$.

3.3.1 Linear Fokker–Planck equation with white noise

We start by presenting the methodology used in the absence of an interaction term, in which case Eq. (3.16a) reduces to a linear Fokker–Planck equation:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left(V'\,\rho + \beta^{-1}\,\frac{\partial \rho}{\partial x}\right) =: \mathcal{L}_x^*\rho, \quad \rho(x, t=0) = \rho_0(x). \tag{3.23}$$

We assume that $V(\cdot)$ is a smooth confining potential and, consequently, the unique invariant distribution is given by $\rho_s = \frac{1}{\mathcal{Z}}\,\mathrm{e}^{-\beta V}$, where $\mathcal{Z}$ is the normalization constant [174, Proposition 4.2]. The Fokker–Planck operator in Eq. (3.23) is unitarily equivalent to a Schrödinger operator; see [2] and [174, Section 4.9]. Defining $u = \rho/\sqrt{\rho_s}$, the function u satisfies

$$\frac{\partial u}{\partial t} = \sqrt{\rho_s^{-1}}\,\mathcal{L}_x^*\left(\sqrt{\rho_s}\,u\right) = \beta^{-1}\,\frac{\partial^2 u}{\partial x^2} + \left(\frac{1}{2}\,V''(x) - \frac{\beta}{4}\,|V'(x)|^2\right) u =: \mathcal{H}_x u, \tag{3.24}$$

with the initial condition $u(x, t=0) = \rho_0/\sqrt{\rho_s} =: u_0$. Several works made use of Hermite spectral methods to study equations of this type, e.g. [2, 74, 83]. The Schrödinger operator on the right-hand side of Eq. (3.24) is selfadjoint in $L^2(\mathbf{R})$ and it has nonpositive eigenvalues. Under appropriate growth assumptions on the potential $V(x)$ as $x \to \infty$, it can be shown that its eigenfunctions decrease more rapidly than any exponential function in the $L^2(\mathbf{R})$ sense, in that they satisfy $\mathrm{e}^{\mu|x|}\,\varphi(x) \in L^2(\mathbf{R})$ for all $\mu \in \mathbf{R}$; see [83] and also [4] for a detailed study. Under appropriate decay assumptions at infinity on the initial condition, we expect the solution to Eq. (3.24) to also decrease rapidly as $|x| \to \infty$.

We denote by $\mathbf{P}(d)$ the space of polynomials of degree less than or equal to d, and by $\langle \cdot, \cdot \rangle$ the usual $L^2(\mathbf{R})$ inner product. For a quadratic potential $V_q = \frac{1}{2}\left(\frac{x}{\sigma}\right)^2$, with σ a scaling parameter,

the Galerkin method we employ consists in finding $u_d(t) \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(d)$ such that

$$\left\langle \frac{\partial u_d}{\partial t}, w_d \right\rangle = \langle \mathcal{H}_x u_d, w_d \rangle_{\hat{d}} \qquad \forall w_d \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(d), \quad \forall t > 0, \tag{3.25a}$$

$$\langle u_d(0), w_d \rangle = \langle u_0, w_d \rangle_{\hat{d}} \qquad \forall w_d \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(d). \tag{3.25b}$$

Here the subscript $\hat{d} \geq d$ on the right-hand side of Eqs. (3.25a) and (3.25b) indicates that the inner product is performed using a numerical quadrature with $\hat{d}+1$ points. With appropriately rescaled Gauss–Hermite points, inner products calculated using the quadrature are *exact* for functions in $\mathrm{e}^{-V_q/2}\,\mathbf{P}(\hat{d})$,

$$\langle v_d, w_d \rangle_{\hat{d}} = \langle v_d, w_d \rangle \quad \forall v_d, w_d \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(\hat{d}),$$

which is why we did not append the subscript $\hat{d}$ to the inner products in the left-hand side of Eqs. (3.25a) and (3.25b). When V is a polynomial, it is possible to show using the recursion relations Eqs. (A.13) and (A.15) in Appendix A.3 that the inner product $\langle \mathcal{H}_x u_d, w_d \rangle_{\hat{d}}$ on the right-hand side of Eq. (3.25a) is *exactly* $\langle \mathcal{H}_x u_d, w_d \rangle$ when $\hat{d} \geq d + \deg(|V'|^2)$. This is the approach we take in all the numerical experiments presented in this paper, and we will therefore omit the subscript $\hat{d}$ in Eq. (3.25a) from now on.

The natural basis of $\mathbf{P}(d)$ (from which a basis of $e^{-V_q/2}\,\mathbf{P}(d)$ follows) to obtain a finite-dimensional system of differential equations from the variational formulation (3.25a) is composed of rescaled Hermite polynomials $H_i^\sigma(x) := H_i(x/\sigma)$, $0 \leq i \leq d$, where $H_i(x)$ are the Hermite polynomials orthonormal for the Gaussian weight $\mathcal{N}(0,1)$; the corresponding basis functions of $\mathrm{e}^{-V_q/2}\,\mathbf{P}(d)$ are then rescaled Hermite functions. For completeness, we summarize in Appendix A.3 the fundamental results on Hermite polynomials, Hermite functions and the related approximation results that are used in this paper. In general, in addition to the rescaling parametrized by σ, a translation could be applied in order to generate more suitable basis functions (e.g. when most of the mass of $\mathrm{e}^{-\beta V}$ is itself localized away from $x = 0$), but for simplicity we confine ourselves in this work to the case where V_q is symmetric around $x = 0$.

Remark 3.3. Although Eq. (3.24) and the associated variational formulation (3.25a) are convenient for analysis purposes, for numerical purposes it is useful to perform a second unitary transformation; defining $v = \mathrm{e}^{V_q/2}\,u =: u/\sqrt{\rho_q}$, the function v satisfies

$$\frac{\partial v}{\partial t} = \sqrt{\rho_s^{-1}\,\rho_q^{-1}}\,\mathcal{L}_x^*\left(\sqrt{\rho_s\,\rho_q}\,v\right) = \frac{1}{\beta}\left(-V_q'\,\frac{\partial}{\partial x} + \frac{\partial^2}{\partial x^2} + \frac{\beta}{2}\,V'' - \frac{1}{2}\,V_q'' - \frac{1}{4}\,|\beta V'|^2 + \frac{1}{4}\,|V_q'|^2\right) v, \tag{3.26}$$

with the initial condition $v(x, t=0) = \rho_0/\sqrt{\rho_s\,\rho_q}$, and to approximate the solution to this equation in $\mathbf{P}(d)$. While clearly equivalent, this approach enables us to work directly with Hermite polynomials, for which a range of free and open-source software tools are available, e.g. in the *NumPy* package for scientific computing in *Python* [123]. Building upon the tools provided by *NumPy*, we have developed a thin *Python* library, available online [216], that offers the possibility of fully automating spectral methods based on Hermite polynomials. We note that, when $V_q = \beta V$, Eq. (3.26) is merely the backward Kolmogorov equation corresponding to Eq. (3.23). ⊘

It is possible to prove the convergence of the method presented above when $d \to \infty$ given appropriate additional assumptions on the confining potential $V(\cdot)$. For simplicity we will make the following assumption, which is satisfied for the bistable potential that we consider in this work, but we note that less restrictive conditions would be sufficient.

Assumption 3.1. *The confining potential $V(\cdot)$ is a polynomial of (even) degree greater than or equal to 2. Consequently, it satisfies*

$$C_1(1+|x|^2) \leq C_2 + W := C_2 + \left(\frac{\beta}{4}\left|V'\right|^2 - \frac{1}{2}V''\right) \leq C_3(1+|x|^{2k}),$$

for constants $C_1, C_2, C_3 > 0$ and a natural number $k \geq 1$.

Theorem 3.1. *Assume that Assumption 3.1 holds and that the initial condition u_0 is smooth and satisfies*

$$\langle(-\mathcal{H}_x^m + 1)u_0, u_0\rangle < \infty \qquad \text{for a natural number } m \geq 2k,$$

where k is as in Assumption 3.1. Then for any final time T and when $d \geq m-1$

$$\sup_{t\in[0,T]} \|u(t) - u_d(t)\|^2 \leq C\ \mathrm{e}^T\ \frac{(d-m+1)!}{(d-2k+1)!}\,\langle(-\mathcal{H}_x^m + 1)u_0, u_0\rangle\,,$$

for a constant C not depending on d, u_0, or T, and where $\|\cdot\|$ denotes the $L^2(\mathbf{R})$ norm.

Proof. See Section 3.A. □

Remark 3.4. When the initial condition is smooth and decreases exponentially, together with all its derivatives, as $x \to \infty$, Theorem 3.1 implies that the error decreases faster than any inverse polynomial. In most practical examples, we observed numerically that the convergence is in fact exponential. ⊘

Remark 3.5. An alternative manner of solving Eq. (3.23) numerically is to approximate the eigenvalues and eigenfunctions of the operator $\mathcal{H}_x$ in terms of Hermite functions, after which integration in time becomes trivial. This approach requires finding (or approximating) the $d+1$ solutions $(\varphi_d^i, \lambda_d^i)$, $\varphi_d^i \in \mathrm{e}^{-V_q/2}\ \mathbf{P}(d)$, $0 \leq i \leq d$, of the eigenvalue problem

$$\left\langle \mathcal{H}_x\,\varphi_d^i, w_d\right\rangle = \lambda_d^i\left\langle\varphi_d^i, w_d\right\rangle \quad \forall w_d \in \mathrm{e}^{-V_q/2}\ \mathbf{P}(d). \tag{3.27}$$

Once these have been calculated, an approximation of the solution to Eq. (3.24) is obtained simply as

$$u_d(t) = \sum_{i=0}^{d} c_{0i}\,\varphi_d^i\ \mathrm{e}^{\lambda_d^i t}, \quad c_{0i} = \left\langle u_0, \varphi_d^i\right\rangle_{\hat{d}}. \tag{3.28}$$

It is readily seen that u_d, defined by this equation, is also the unique solution of (3.25a). This is the approach taken in [28] for the calculation of drift and diffusion coefficients in the diffusion approximation of kinetic equations. While equivalent, this methodology is more computationally expensive because it requires the full solution of the eigenvalue problem (3.27). ⊘

3.3.2 McKean–Vlasov equation with white noise

In the presence of an interaction term, the Fokker–Planck equation becomes nonlinear:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left(V'\,\rho + \theta(x - m(t))\,\rho + \beta^{-1}\frac{\partial \rho}{\partial x}\right) =: (\mathcal{L}_x^m)^*\rho, \quad m(t) = \int_{\mathbf{R}} x\,\rho\,\mathrm{d}x. \tag{3.29}$$

For this equation the weighted $L^2\left(\mathbf{R}, \mathrm{e}^V\right)$ energy estimate of the linear case (3.50), based on observing that $\langle \partial_t\rho, \rho\rangle_{\mathrm{e}^V} \leq 0$, does not hold, and there is therefore no longer a natural space for the Galerkin approximation. We will thus use Hermite functions to approximate the solution to Eq. (3.29) directly, i.e. we will look for an approximate solution in the space $\mathrm{e}^{-V_q/2}\,\mathbf{P}(d)$. The variational formulation corresponding to the Galerkin approximation is then to find $\rho \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(d)$ such that

$$\left\langle \frac{\partial \rho_d}{\partial t}, w_d \right\rangle = \langle (\mathcal{L}_x^{m_d})^*\rho_d, w_d\rangle \qquad \forall w_d \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(d), \tag{3.30a}$$

$$m_d = \frac{\langle x, \rho_d\rangle_{\hat{d}}}{\langle 1, \rho_d\rangle_{\hat{d}}} \approx \frac{\int_{\mathbf{R}} x\,\rho_d\,\mathrm{d}x}{\int_{\mathbf{R}} \rho_d\,\mathrm{d}x}, \tag{3.30b}$$

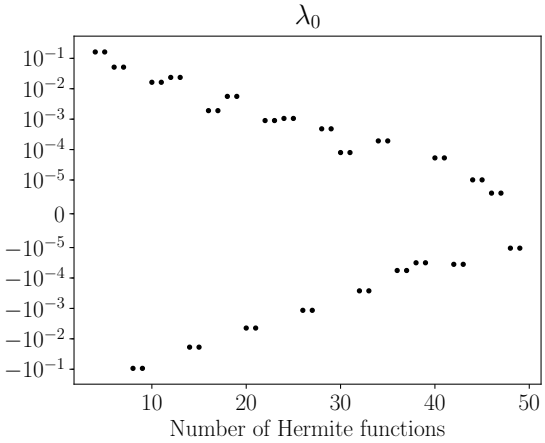
$$\langle \rho_d(0), w_d\rangle = \langle \rho_0, w_d\rangle_{\hat{d}} \qquad \forall w_d \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(d). \tag{3.30c}$$

Dividing by $\langle 1, \rho_d\rangle_{\hat{d}}$ in Eq. (3.30b) is useful to account for changes in the total mass of ρ_d, which can compromise the accuracy of the method when d is low, but doing so becomes unnecessary for large enough d. In contrast with the operator $\mathcal{H}_x$ in Eq. (3.25a), the operator $(\mathcal{L}_x^{m_d})^*$ is not selfadjoint in $L^2(\mathbf{R})$, and therefore the associated stiffness matrix is not symmetric. In addition, the quadratic form $\langle (\mathcal{L}_x^m)^*\cdot, \cdot\rangle$ is not necessarily negative for the usual $L^2(\mathbf{R})$ inner product, and indeed we observe numerically that the eigenvalue with smallest real part of the discrete operator is often negative, although small when d is large enough. This is illustrated in Fig. 3.3a for the same parameters as in the subsequent convergence study. We also notice that the sign of the eigenvalue with smallest real part of $-\hat{\Pi}_d\,(\mathcal{L}_x^{m=0})^*\,\hat{\Pi}_d$, where $\hat{\Pi}_d$ denotes the $L^2(\mathbf{R})$ projection operator on $\mathrm{e}^{-V_q/2}\,\mathbf{P}(d)$, exhibits an oscillatory behavior, which is consistent with the fact that $(\mathcal{L}_x^m)^*$ is not symmetric in the space $L^2(\mathbf{R})$.

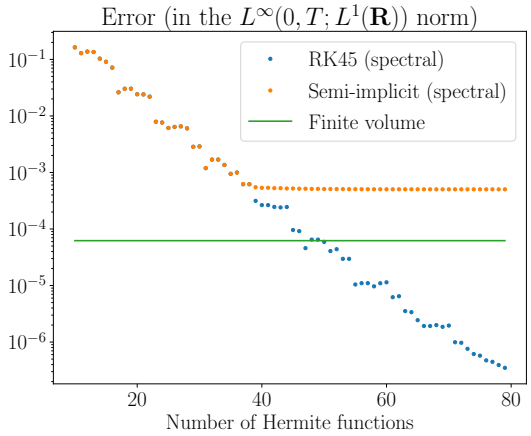
For the integration in time, we used either the RK45 method (using the `solve_ivp` method from *SciPy integrate* module), or a linear semi-implicit method obtained by treating m_d explicitly and the other terms implicitly at each time step. The former is most useful when an accurate solution is required, while the latter enables the use of larger time steps and is therefore more convenient when only the steady-state solution is sought, as will be the case for the construction of bifurcation diagrams. Denoting the time step by Δt and the Galerkin approximation of $\rho_d(n\,\Delta t)$ by ρ_d^n, the semi-implicit method is based on obtaining ρ_d^{n+1} by solving:

$$\left\langle \rho_d^{n+1} - \rho_d^n, w_d\right\rangle = \Delta t\,\left\langle (\mathcal{L}_x^{m_d^n})^*\rho_d^{n+1}, w_d\right\rangle \qquad \forall w_d \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(d), \tag{3.31a}$$

$$m_d^{n+1} = \frac{\left\langle x, \rho_d^{n+1}\right\rangle_{\hat{d}}}{\left\langle 1, \rho_d^{n+1}\right\rangle_{\hat{d}}}. \tag{3.31b}$$



(a) Eigenvalue with smallest real part of $-\hat{\Pi}_d \left(\mathcal{L}_x^m\right)^* \hat{\Pi}_d$ for $m = 0$. A mixed scale, linear in the interval $[-10^{-5}, 10^{-5}]$ and logarithmic elsewhere, is used for the y axis.



(b) Error of the Hermite Galerkin discretization, with either RK45 or the semi-implicit method (3.31), against degree of approximation (d), and comparison with the error of the finite volume scheme from [37].

Figure 3.3: Study of the Galerkin approximation (3.30).

Convergence study The analysis of the Hermite spectral method for general types of McKean–Vlasov equations will not be presented here. For the purposes of this work, it will be sufficient to present a detailed numerical study of the convergence of the method. To study empirically the validity of the Galerkin method (3.30) and of the associated time-stepping scheme (3.31), we compare our method with the positivity preserving, entropy decreasing finite volume method proposed in [37] for nonlinear, nonlocal gradient PDEs[1]. The parameters used here are $\beta = 3$, $\theta = 1$, and the initial condition was the Gaussian $\mathcal{N}(10^{-1}, 1)$. The same time points were used for the finite volume method and semi-implicit Galerkin method (with a mean time step of approximately 0.002), and for RK45 the absolute and relative tolerances were both set to 10^{-11}. For the finite volume method, 600 equidistant mesh points were used between $x = -6$ and $x = 6$.

Figure 3.3b presents the $L^\infty\left(0, T; L^1\left(\mathbf{R}\right)\right)$ norm of the errors associated with the solutions obtained, for values of d, the degree of Hermite polynomials used, ranging from 10 to 80. A very accurate solution, obtained by using our spectral method with $d = 120$, was employed for the calculation of the errors. We observe that, as d increases initially, the solutions obtained using the semi-implicit (3.31) and the RK45 methods are indistinguishable and converge exponentially fast. From $d \approx 40$, the accuracy of the semi-implicit method no longer improves, indicating that the error introduced by the time-stepping scheme dominates from that point on. From $d \approx 50$, the Galerkin/RK45 approximation becomes more precise than the finite volume method. We therefore conclude that an accuracy as good as that obtained using the finite volume scheme can be reached with roughly *ten times fewer unknowns* using the spectral discretization (3.30). Our spectral method also enjoys a low computational cost: it ran in only about a minute with a Intel i7-3770 processor, even for a value of d as high as 80, whereas the finite volume simulation took over an hour.

Figure 3.4 presents snapshots of the solutions at different times. We observe that, although the number of Hermite functions employed in the expansion is relatively low (=25), the solutions

[1] We reiterate the fact that one of the main advantages of our numerical method is that it *does not* require that the PDE has a gradient flow structure.

are in extremely good agreement. To estimate the contribution of the initial error to the error

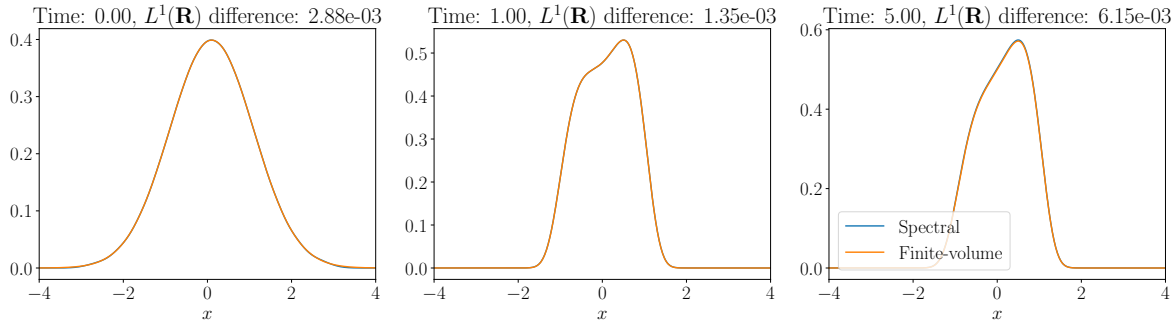


Figure 3.4: Snapshots of the solution to (3.29) using either the finite volume method from [37] or the Galerkin approximation (3.30) with the RK45 method and 25 Hermite functions.

at later times, we compared three solutions. Let us denote by $\rho_{d_1}^{d_2}$, for $d_2 \leq d_1$, the numerical solution obtained via the Galerkin method (3.30) with a number of Hermite modes equal to d_1 but with the initial condition truncated at $d = d_2$, i.e. with the Hermite coefficients corresponding to $d_2 < d \leq d_1$ of the numerical initial condition equal to zero. We calculated the following numerical solutions: $\rho_c := \rho_{25}^{25}$, the coarse solution depicted in Fig. 3.4; $\rho_i := \rho_{120}^{25}$, a solution calculated with a precise discretization of the equation but a coarse approximation of the initial condition; and $\rho_e := \rho_{120}^{120}$, a precise solution considered exact for our purposes. The pairwise $L^1(\mathbf{R})$ differences between these solutions are illustrated in Fig. 3.5. We observe the truncation of the Hermite series associated with the initial condition accounts for a significant part of the error on ρ_c only at short times; at longer times, such as $t = 5$, the solution ρ_i is much closer to the exact solution than is ρ_c, although ρ_i and ρ_c coincide at $t = 0$.



Figure 3.5: Difference, in the $L^1(\mathbf{R})$ norm, between the numerical solutions ρ_c, ρ_i and ρ_e.

In the simulations presented in this section, the scaling factor was set to $\sigma^2 = \frac{1}{10}$. As discussed in Appendix A.3, choosing this factor appropriately can significantly improve the accuracy of the method. In particular, given that the solution to Eq. (3.29) decreases rapidly as $|x| \to \infty$, σ should decrease with d, with the optimal scaling being $\sigma \propto \sqrt{d}$, as demonstrated in [209]. For convergence studies, however, it is convenient to use a fixed σ, first because this is assumed by most convergence results (such as Theorem 3.1) and, second, because this simplifies the

calculation of the matrices involved in the Galerkin formulation (only the last row and the last column have to be calculated upon incrementing d).

Remark 3.6 (Computational considerations). Discretizing the operators appearing in the Galerkin approximations (3.25a) and (3.30) requires the calculation of multiple matrices corresponding to operators of the type $\Pi_d \, (f \, \partial_x) \, \Pi_d$, where f is a polynomial and Π_d is the $L^2 \left(\mathbf{R}, \mathrm{e}^{-V_q}\right)$ projection operator onto $\mathbf{P}(d)$. These calculations can be carried out by noticing that

$$\Pi_d \left(f \, \frac{\mathrm{d}^m}{\mathrm{d}x^m} \right) \Pi_d = \left(\Pi_d \, f \, \Pi_d \right) \left(\Pi_d \frac{\mathrm{d}^m}{\mathrm{d}x^m} \Pi_d \right).$$

The matrix representation of the first operator on the right-hand side, in a basis of Hermite polynomials, can be obtained from the Hermite transform of f. The matrix representation of the second operator, on the other hand, is a matrix with zero entries everywhere except on the m-th superdiagonal, in view of the recursion relation (A.15). ⊘

3.3.3 Linear Fokker–Planck equation with colored noise

In this section, we turn our attention to the case of Gaussian or non-Gaussian colored noise given in terms of overdamped Langevin dynamics. The case of harmonic noise can be treated in a similar fashion, and for conciseness we do not present the associated Galerkin formulation explicitly here. We start by considering the linear, without the interaction term, Fokker–Planck equation with colored noise:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x} \left(\frac{\partial V}{\partial x} \, \rho - \frac{\zeta}{\varepsilon} \sqrt{2 \, \beta^{-1}} \eta \, \rho \right) + \frac{1}{\varepsilon^2} \frac{\partial}{\partial \eta} \left(V_\eta' \, \rho + \frac{\partial \rho}{\partial \eta} \right) =: \mathcal{L}_\varepsilon^* \, \rho. \tag{3.32}$$

We recall that ε^2 controls the correlation time of the colored noise and ζ is a parameter such that the white noise limit is recovered (with inverse temperature β) when $\varepsilon \to 0$. We include ε in Eq. (3.32) because, although we do not consider the white noise limit in this section, large values of ε are in general more difficult to tackle numerically, and it will be therefore convenient to use smaller correlation times in the numerical experiments below. The problem is now two-dimensional and the operator on the right-hand side of Eq. (3.32) is no longer elliptic. In contrast with the white noise case, there does not exist an explicit formula for the steady-state solution for Eq. (3.32).

The procedure for obtaining a Galerkin formulation is the same as in Section 3.3.1, except that we now use tensorized Hermite polynomials/functions. To retain some generality, we will consider that the Galerkin approximation space is of the form $S_d = \mathrm{e}^{-U(x,\eta)/2} \, \mathrm{e}^{-V_q(x,\eta)/2} \, \mathbf{P}(\mathcal{I}_d)$ for some function $U : \mathbf{R}^2 \mapsto \mathbf{R}$, a nondegenerate quadratic potential V_q to be determined, and where $\mathbf{P}(\mathcal{I}_d) := \operatorname{span} \{ x^{\alpha_x} \, \eta^{\alpha_\eta} : (\alpha_x, \alpha_\eta) \in \mathcal{I}_d \}$ for some index set $\mathcal{I}_d \subset \mathbf{N}^2$ that grows with $d \in \mathbf{N}$. Compared to the one-dimensional case, there are now two scaling parameters, $V_q := x^2/2\sigma_x^2 + \eta^2/2\sigma_\eta^2$. The Galerkin approximation we propose consists in finding $\rho_d \in S_d$ such that

$$\left\langle \frac{\partial \rho_d}{\partial t}, w_d \right\rangle_{\mathrm{e}^U} = \langle \mathcal{L}_\varepsilon^* \, \rho_d, w_d \rangle_{\mathrm{e}^U} \qquad \forall w_d \in S_d, \quad \forall t > 0, \tag{3.33}$$

with appropriate initial conditions. The choice of the weight e^{U} in the inner products of Eq. (3.33) is motivated by the fact that differential operators admit sparse representations in the Hermite-type basis naturally associated with S_d, and we note that $\mathrm{e}^{-U(x,\eta)/2}\,\mathrm{e}^{-V_q(x,\eta)/2}\,\mathbf{P}(\mathbf{N}^2)$, where $\mathbf{P}(\mathbf{N}^2)$ is the space of polynomials in two dimensions, is dense in $L^2\left(\mathbf{R}^2, \mathrm{e}^{U}\right)$. In practice, we obtain ρ_d as $\mathrm{e}^{-U(x,\eta)/2}\,\mathrm{e}^{-V_q(x,\eta)/2}\,v_d$, where v_d is obtained by solving

$$\left\langle \frac{\partial v_d}{\partial t}, w_d \right\rangle_{\mathrm{e}^{-V_q}} = \langle \mathcal{H}_\varepsilon\, v_d, w_d \rangle_{\mathrm{e}^{-V_q}} \qquad \forall w_d \in \mathbf{P}(\mathcal{I}_d), \quad \forall t > 0, \tag{3.34}$$

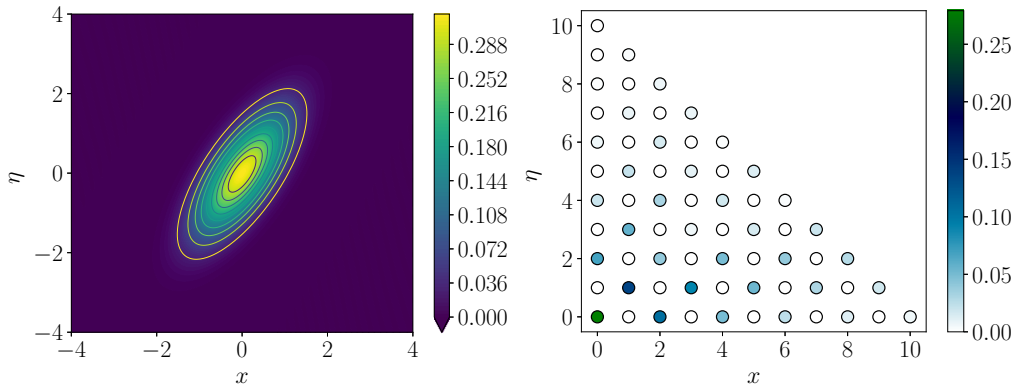
where $\mathcal{H}_\varepsilon := (\mathrm{e}^{U/2}\,\mathrm{e}^{V_q/2}\,v_d)\mathcal{L}_\varepsilon^*\,(\mathrm{e}^{-U/2}\,\mathrm{e}^{-V_q/2})$, and the basis functions used for Eq. (3.34) are Hermite polynomials orthonormal with respect to the Gaussian weight e^{-V_q}. Regarding the index set, several choices are possible, with the simplest ones being the triangle $\{\alpha \in \mathbf{N}^2 : |\alpha|_1 \leq d\}$ and the square $\{\alpha \in \mathbf{N}^2 : |\alpha|_\infty \leq d\}$, see Figs. 3.6a and 3.7a below. We demonstrate in Section 3.4 that, in order to study the limit $\varepsilon \to 0$, a rectangle-shaped index set is usually the only suitable choice. When studying the behavior as d increases, however, we observed spectral convergence irrespectively of the index set utilized.

Clearly, it is necessary that $\rho \in L^2\left(\mathbf{R}^2, \mathrm{e}^{U}\right)$ for the Galerkin discretization (3.33) to produce good results. Since the $1/\varepsilon^2$ part of the operator on the right-hand side of Eq. (3.32), $\mathcal{L}_0^* \cdot = \mathrm{d}_\eta(V_\eta'(\eta)\cdot + \mathrm{d}_\eta \cdot)$, is selfadjoint in $L^2\left(\mathbf{R}, \mathrm{e}^{V_\eta}\right)$, it is natural to choose $\mathrm{e}^{-U(x,\eta)/2} = \mathrm{e}^{-U_x(x)/2 - V_\eta(\eta)/2}$ for some one-dimensional potential U_x, which guarantees that the matrix representation of $\mathcal{L}_0^*$ is symmetric and negative semi-definite, but this is not a requirement.

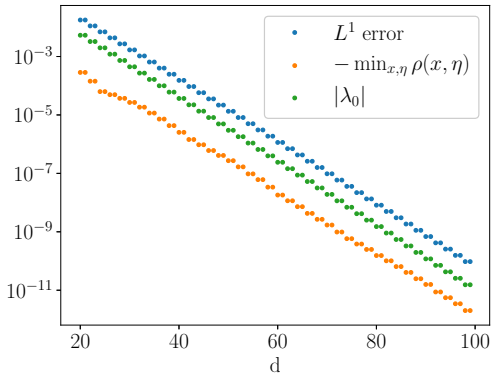
Before moving to the nonlinear case, we examine the performance of the Galerkin approximation (3.33) through numerical experiments. Here we consider only the cases where $V(\cdot)$ is a quadratic or a bistable potential and where the noise is described by an OU process, but results of additional numerical experiments, corresponding to harmonic noise and non-Gaussian noise, are presented in Section 3.C.

We start with the case $V(x) = x^2/2$, for which the exact solution to the Fokker–Planck equation (3.32) can be calculated explicitly by substitution of a Gaussian ansatz, see [174, Section 3.7]. We study the convergence of the steady-state solution, obtained by calculating the eigenfunction associated with the eigenvalue of lowest magnitude of $\hat{\Pi}_d\,\mathcal{L}_\varepsilon^*\,\hat{\Pi}_d$, where $\hat{\Pi}_d$ is the $L^2\left(\mathbf{R}^2, \mathrm{e}^{U}\right)$ projection operator on S_d, directly using the method `eigs` from the *SciPy* toolbox. The parameters used for this simulation are the following: $\beta = \varepsilon = 1$, $\sigma_x^2 = \frac{1}{10}$, $\sigma_\eta^2 = 1$, $\mathrm{e}^{-U(x,\eta)/2} = \mathrm{e}^{-V_\eta(\eta)/2} = \mathrm{e}^{-\eta^2/4}$. With these parameters, the steady-state solution to Eq. (3.32) is equal to $\rho_\infty(x,\eta) = \mathrm{e}^{-2x^2+2x\eta-\eta^2}/\pi$, and clearly $\rho_\infty \in L^2\left(\mathbf{R}^2, \mathrm{e}^{U}\right)$. Figure 3.6a presents the steady-state solution, obtained using the spectral method with Hermite polynomials up to degree 100 ($d = 100$) and a triangular index set, and Fig. 3.6b presents the convergence of the method. Since the solution satisfies $\rho_\infty(x,\eta) = \rho_\infty(-x,-\eta)$, the Hermite coefficients corresponding to even values of $i + j$ are zero, where i and j are the indices in the x and η directions, respectively.

Now we consider that V is the bistable potential $x^4/4 - x^2/2$, which was solved numerically in [105] using generalized Hermite functions and a variation of the matrix continued fraction technique. For this case an explicit analytical solution is not available. The parameters we use are: $\beta = 1$, $\varepsilon = \frac{1}{2}$, $\sigma_x^2 = \frac{1}{20}$, $\sigma_\eta^2 = 1$. Through numerical exploration, we noticed that a good convergence could be obtained by using the multiplier function $\mathrm{e}^{-U(x,\eta)/2} = \mathrm{e}^{-\beta V(x)/2 - \eta^2/4}$,



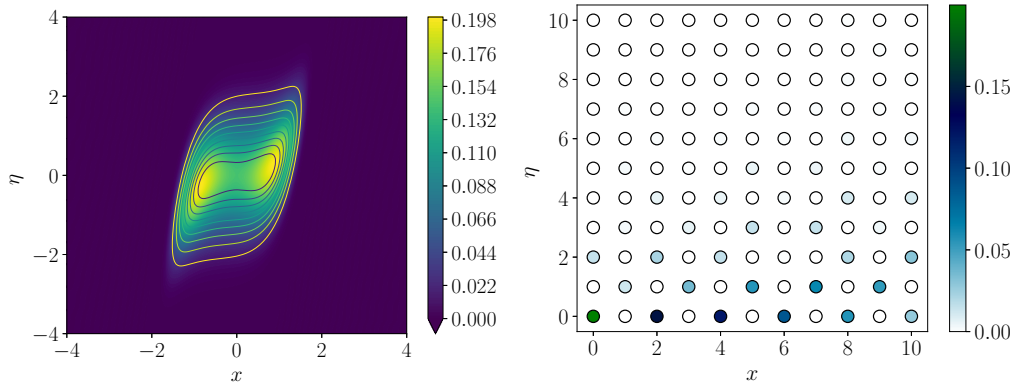
(a) Steady-state solution of the Fokker–Planck equation (3.33) with the associated field lines of the probability flux (left) and absolute value of the coefficients of degree less than equal to 10 in the Hermite expansion (right). See Appendix A.1, and in particular Eq. (A.6), for the methodology used for the calculation of the field lines.



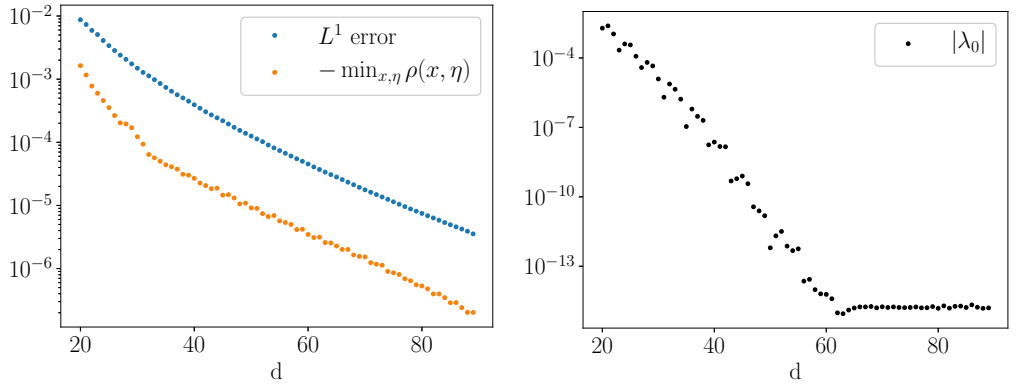
(b) Convergence of the method, using three different metrics for the error: the L^1 norm of the error between the numerical and exact solutions, the negative of the minimum of the numerical solution, and the absolute value of the eigenvalue with smallest real part.

Figure 3.6: Simulation data when $V(x) = x^2/2$.

rather than just $\mathrm{e}^{-\eta^2/4}$ in the previous paragraph. We note that this would have been the natural choice if the noise in the x direction had been white noise. The solution obtained using a square-shaped index set and $d = 100$, as well as the corresponding Hermite coefficients up to degree 10, is illustrated in Fig. 3.7a. We observe that the Hermite coefficients corresponding to the degree 0 in the η direction (i.e. to the basis function $\mathrm{e}^{-\eta^2/2}$) are significantly larger than the other coefficients, which is consistent with the fact that, as $\varepsilon \to 0$, the steady-state solution approaches $\mathrm{e}^{-\beta V(x)}\ \mathrm{e}^{-\eta^2/2}$ (up to a constant factor). The associated convergence curves are presented in Fig. 3.7b.



(a) Steady-state solution to Eq. (3.33) and associated field lines of the probability flux (left), and absolute value of the coefficients of degree less than or equal to 10 in the Hermite expansion (right).



(b) Convergence of the method using the same measures of the error as in Fig. 3.6a, except that the L^1 error is calculated by comparison with the numerical solution obtained when $d = 100$.

Figure 3.7: Simulation data when $V(x) = x^4/4 - x^2/2$.

3.3.4 McKean–Vlasov equation with colored noise

We consider now the nonlinear McKean–Vlasov initial value problem with OU noise: recalling that $\zeta = 1/\sqrt{2}$ in this case,

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left(\frac{\partial V}{\partial x}\,\rho + \theta\,(x - m(t))\,\rho - \frac{1}{\varepsilon}\sqrt{\beta^{-1}}\,\eta\,\rho\right) + \frac{1}{\varepsilon^2}\,\frac{\partial}{\partial \eta}\left(\eta\,\rho + \frac{\partial \rho}{\partial \eta}\right), \tag{3.35a}$$

$$m(t) = \int_{\mathbf{R}}\int_{\mathbf{R}} x\,\rho(x,\eta,t)\,\mathrm{d}\eta\,\mathrm{d}x, \tag{3.35b}$$

$$\rho(x,\eta,t=0) = \rho_0(x,\eta), \tag{3.35c}$$

for some initial distribution $\rho_0(x,\eta)$ such that the noise is not necessarily started at stationarity. The method that we use in this case, which applies *mutatis mutandis* to the other noise models, is the same as in Eq. (3.33), with the addition of the interaction term, and we use the same time-stepping schemes as in Section 3.3.1. When the potential $V(\cdot)$ is quadratic and the initial condition is Gaussian, it is well-known that the McKean–Vlasov equation has an explicit solution and that this solution is Gaussian. We assume that $V(x) = x^2/2$ and we rewrite Eq. (3.35) in

the formalism of [62], as

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \left(B\,\mathbf{x}\,\rho + \int_{\mathbf{R}^2} K(\mathbf{x}-\mathbf{x}')\,\rho(\mathbf{x}',t)\,\mathrm{d}\mathbf{x}'\,\rho - D\,\nabla\rho \right),$$

where $\mathbf{x} = (x,\eta)^T$ and

$$B = \begin{pmatrix} -1 & \varepsilon^{-1}\beta^{-1/2} \\ 0 & -\varepsilon^{-2} \end{pmatrix}, \quad K = \begin{pmatrix} -\theta & 0 \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 \\ 0 & \varepsilon^{-2} \end{pmatrix}.$$

Adapting [62, Proposition 2.3] to our case, we deduce that the solution is of the type

$$\rho(\mathbf{x},t) = \frac{1}{(2\pi)\,|\Sigma(t)|}\,\exp\left(-\frac{1}{2}\,(\mathbf{x}-\mu(t))^T\Sigma^{-1}(t)(\mathbf{x}-\mu(t))\right),$$

where $\mu(t)$ and $\Sigma(t)$ are given by

$$\mu(t) = \mathrm{e}^{Bt}\,\mu(0), \quad \Sigma(t) = \mathrm{e}^{t(B+K)}\,\Sigma(0)\,\mathrm{e}^{t(B+K)^T} + 2\int_0^t \mathrm{e}^{s(B+K)}\,D\,\mathrm{e}^{s(B+K)^T}\,\mathrm{d}s. \tag{3.36}$$

This solution can be obtained by introducing $g = -\ln\rho$, rewriting Eq. (3.35) as an equation for g, and using a quadratic ansatz for g. The eigenvalue decomposition of $B+K$ is

$$(B+K)\begin{pmatrix} 1 & -\varepsilon \\ 0 & \sqrt{\beta}\,(1-\varepsilon^2(1+\theta)) \end{pmatrix} = \begin{pmatrix} 1 & -\varepsilon \\ 0 & \sqrt{\beta}\,(1-\varepsilon^2(1+\theta)) \end{pmatrix}\begin{pmatrix} -1-\theta & 0 \\ 0 & -\varepsilon^{-2} \end{pmatrix},$$

which enables the explicit calculation of the integral in the expression of $\Sigma(t)$. From Eq. (3.36) and the structure of B and K, we notice that, as $t\to\infty$, $\mu\to 0$ and

$$\Sigma(t) \to \Sigma_\infty = 2\int_0^\infty \mathrm{e}^{s(B+K)}\,D\,\mathrm{e}^{s(B+K)^T}\,\mathrm{d}s,$$

which coincides with the solution of the steady state linear Fokker–Planck equation corresponding to the McKean–Vlasov equation when m is a parameter equal to 0. For this test case, we use the following parameters: $\beta = \theta = 1$, $\varepsilon = 1/2$, $\sigma_x^2 = \sigma_\eta^2 = 1/5$, $\mathrm{e}^{-U(x,\eta)/2} = \mathrm{e}^{-V(x)/2}\,\mathrm{e}^{-V_\eta(\eta)/2}$. The initial condition is taken to be the Gaussian density $\mathcal{N}\left((1,1)^T, I_{2\times 2}\right)$. The evolution of the probability density is illustrated in Fig. 3.8, and the convergence of the method, in the $L^\infty\left(0,T;L^1\left(\mathbf{R}^2\right)\right)$ norm, is illustrated in Fig. 3.9.

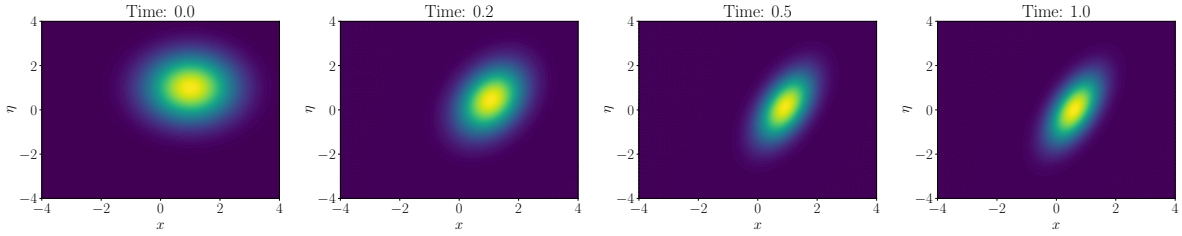


Figure 3.8: Probability density solution of Eq. (3.35) (obtained using the spectral method) at times $0, 0.2, 0.5, 1$.

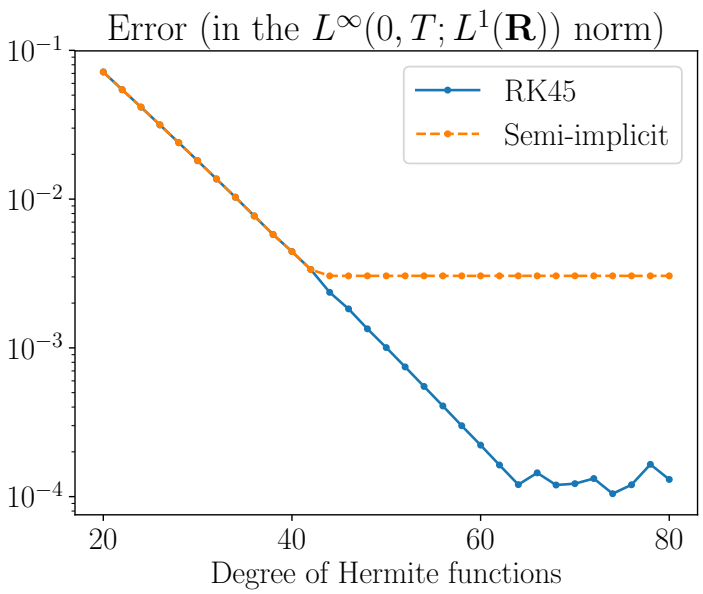


Figure 3.9: Convergence of the Hermite spectral method for the nonlinear McKean–Vlasov equation with colored noise. The error decreases exponentially for low enough values of d, and then reaches a plateau when it becomes dominated by the error induced by the time discretization.

3.3.5 Monte Carlo simulations

We will compare the bifurcation diagrams obtained using the spectral method described above to those obtained by direct MC simulations of the system of interacting particles (3.8). We use the Euler–Maruyama method:

$$X_{k+1}^i = X_k^i - V'(X_k^i)\,\Delta t - \theta\left(X_k^i - \frac{1}{N}\sum_{j=1}^N X_k^j\right)\Delta t + \frac{\zeta}{\varepsilon}\sqrt{2\beta^{-1}}\,\eta_k^i\,\Delta t,$$

where η_k^i is the appropriate projection of the stochastic process $\mathbf{Y}_t$. In the case of Gaussian noise, this is discretized as follows

$$\mathbf{Y}_{k+1}^i = \mathbf{Y}_k^i + \frac{1}{\varepsilon^2} A\,\mathbf{Y}_k^i\,\Delta t + \frac{1}{\varepsilon}\sqrt{2\,\Delta t}\,D\xi,$$

where $\xi \sim N(0,1)$, and X_k, $\mathbf{Y}_k$ and η_k are the approximations to $X(k\Delta t)$, $\mathbf{Y}(k\Delta t)$ and $\eta(k\Delta t)$, respectively. The time step used was always $\mathcal{O}(\varepsilon^2)$, to ensure the accurate solution of the equation. Since the noise in all the equations we consider is additive, this scheme has strong order of convergence one, see [111, 112], and we find that we capture the correct behavior as long as the time step is sufficiently small.

3.4 Asymptotic analysis for the Galerkin formulation

In Section 3.5, we will construct bifurcation diagrams of m as a function of β for different values of ε, and we will verify that the bifurcation diagram of the white noise case is recovered when $\varepsilon \to 0$. Since the spectral method presented in Section 3.3 will be used to that purpose, it is

useful to study the behavior of the solution to the Galerkin formulation (3.33) in the limit $\varepsilon \to 0$, which is the purpose of this section. We will then confirm numerically the rates of convergence to the white noise limit presented in Section 3.2.2, i.e. $\mathcal{O}(\varepsilon^2)$ for Ornstein–Uhlenbeck noise and $\mathcal{O}(\varepsilon^4)$ for harmonic noise.

For simplicity, we confine ourselves for the analysis to the case where the noise process is one-dimensional and the weight function $\mathrm{e}^{-U(x,\eta)/2}$ can be decomposed as $\mathrm{e}^{-U(x,\eta)/2} = \mathrm{e}^{-U_x(x)/2}\,\mathrm{e}^{-V_\eta(\eta)/2}$. As before, $\hat{\Pi}_d$ denotes the $L^2\left(\mathbf{R}^2, \mathrm{e}^{U}\right)$ projection operator on the space of Hermite functions (with appropriate scalings). Decomposing the operator $(\hat{\Pi}_d\,\mathcal{L}_\varepsilon^*\,\hat{\Pi}_d)$ in Eq. (3.33) in powers of ε, we obtain the equation

$$\begin{aligned}\frac{\partial \rho_d}{\partial t} = (\hat{\Pi}_d\,\mathcal{L}_\varepsilon^*\,\hat{\Pi}_d)\,\rho_d = \hat{\Pi}_d\left(\frac{1}{\varepsilon^2}\mathcal{L}_0^* + \frac{1}{\varepsilon}\mathcal{L}_1^* + \mathcal{L}_2^*\right)\hat{\Pi}_d\,\rho_d,\\ =: \left(\frac{1}{\varepsilon^2}\hat{\mathcal{L}}_0 + \frac{1}{\varepsilon}\hat{\mathcal{L}}_1 + \hat{\mathcal{L}}_2\right)\rho_d.\end{aligned}$$

As a consequence of the choice of e^{U}, the largest (sign included) eigenvalue of $\hat{\mathcal{L}}_0$ is a nonpositive, non-decreasing function of d. Since we cannot expect the leading order term of the discrete generator to have an eigenvalue exactly equal to 0, we look for a solution of the form $\rho_d = \mathrm{e}^{-|\lambda_{0,d}|\,t/\varepsilon^2}(\varrho_0 + \varepsilon\,\varrho_1 + \varepsilon^2\,\varrho_2 + \cdots)$, where $\lambda_{0,d}$ is the largest eigenvalue of $\hat{\mathcal{L}}_0$. Denoting by id the identity operator and gathering equal powers of ε, we obtain the equations:

$$0 = \left(\hat{\mathcal{L}}_0 + |\lambda_{0,d}|\ \mathrm{id}\right)\ \varrho_0, \tag{3.38a}$$

$$0 = \left(\hat{\mathcal{L}}_0 + |\lambda_{0,d}|\ \mathrm{id}\right)\ \varrho_1 + \hat{\mathcal{L}}_1\,\varrho_0, \tag{3.38b}$$

$$\frac{\partial \varrho_i}{\partial t} = \left(\hat{\mathcal{L}}_0 + |\lambda_{0,d}|\ \mathrm{id}\right)\ \varrho_{i+2} + \hat{\mathcal{L}}_1\,\varrho_{i+1} + \hat{\mathcal{L}}_2\,\varrho_i, \qquad i = 0, 1, \ldots \tag{3.38c}$$

3.4.1 Suitable index sets

Let H_i^x and H_j^η denote the (possibly rescaled) Hermite functions in the x and η directions, respectively. Let also $\mathcal{I}_{i,\eta}$ be a slice of the index set $\mathcal{I}_{i,\eta} := \{j : (i,j) \in \mathcal{I}_d\}$ and $\Pi_x\mathcal{I}_d$ be the projected index $\{i : (\exists j \in \mathbf{N})[(i,j) \in \mathcal{I}_d]\}$. Expanding ϱ_0 in terms of the basis functions used for the Galerkin discretization in the first equation, we obtain

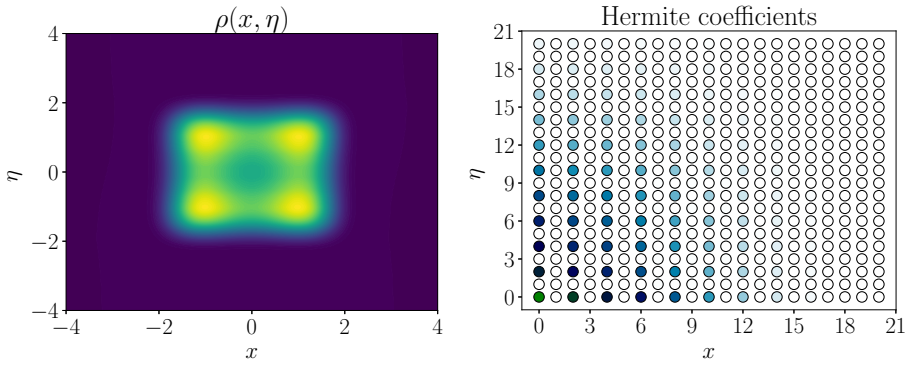
$$\begin{aligned}0 &= (\hat{\mathcal{L}}_0 + |\lambda_{0,d}|\,\mathrm{id}) \sum_{(i,j)\in\mathcal{I}_d} c_{ij}\left(\mathrm{e}^{-U_x/2}\,H_i^x \otimes e^{-V_\eta/2}\,H_j^\eta\right)\\ &= \sum_{i\in\Pi_x\mathcal{I}_d} \mathrm{e}^{-U_x/2}\,H_i^x \otimes \left(\sum_{j\in\mathcal{I}_{i,\eta}} c_{ij} \sum_{k\in\mathcal{I}_{i,\eta}} (L_{jk} + |\lambda_{0,d}|\ \delta_{jk})\ \mathrm{e}^{-V_\eta/2}\ H_k^\eta\right),\end{aligned}$$

where $L_{jk} := \left\langle \mathcal{L}_0^*(\mathrm{e}^{-V_\eta/2}\ H_j^\eta), \mathrm{e}^{-V_\eta/2}\ H_k^\eta \right\rangle_{\mathrm{e}^{V_\eta}}$. This leads to:

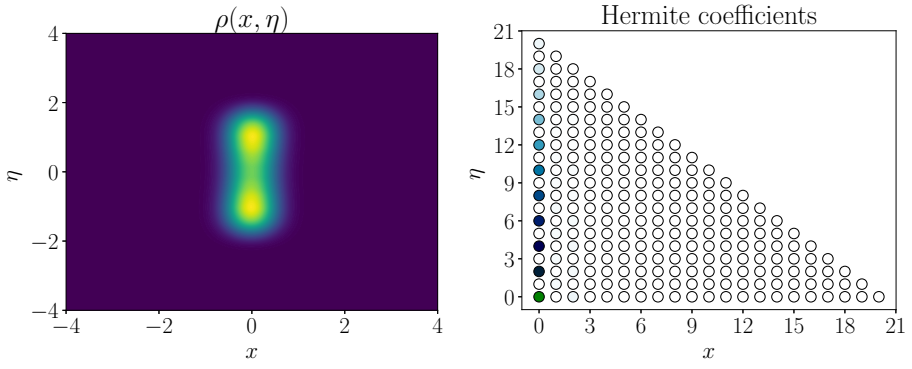
$$\sum_{j\in\mathcal{I}_{i,\eta}} c_{ij}\,(L_{jk} + |\lambda_{0,d}|\ \delta_{jk}) = 0, \qquad \forall i \in \Pi_x\mathcal{I}_d, \tag{3.39}$$

implying that the vector or $(c_{ij})_{j\in\mathcal{I}_{i,\eta}}$ is in the kernel of the matrix $(L_{jk} + |\lambda_{0,d}|)_{j,k\in\mathcal{I}_{i,\eta}}$. Therefore, if for some i this kernel is empty, the corresponding Hermite coefficients must be 0. This is of particular relevance when the eigenfunction in the kernel of $\mathcal{L}_0$ cannot be exactly represented in terms of the approximating basis functions, as is the case with the noise processes **B** and **NS** considered in Section 3.2. For the noise model **B**, in particular, Eq. (3.39) implies that, when using a triangular index set, $c_{ij} = 0$ for all $i > 0$ or $i > 1$, depending on whether the maximal degree is even or odd. From this we conclude that, in order to capture the correct solution as $\varepsilon \to 0$, it is necessary to choose a rectangularly shaped index set, which is consistent with the fact that, in the limit $\varepsilon \to 0$, the solution can be expressed as a tensor product $\rho_\infty(x, \eta) = \rho^x(x)\,\rho^\eta(\eta)$.

To illustrate the point made in the previous paragraph, we present side by side in Fig. 3.10 the results of numerical experiments performed using either a triangular index set or a square index set, for the parameters $\varepsilon = 0.01$, $d = 20$, $\beta = 15$, $\theta = 0$, $\sigma_x^2 = \sigma_\eta^2 = 1/15$, $\mathrm{e}^{-U_x(x)} = 1$. While the probability density obtained using a square index set is close to the exact solution and clearly exhibits four local maxima, the solution obtained using a triangle index set is concentrated around $x = 0$, and all the associated Hermite coefficients c_{ij} with $i > 0$ are very close to zero.



(a) Square index set.



(b) Triangle index set.

Figure 3.10: Comparison of the numerical solutions for the steady-state Fokker–Planck equation with the bistable noise model (model **B** in Section 3.2), using either a square index set or a triangle index set. While not obvious from the figures, in the former case it follows from the fact that $\varepsilon = 0.01 \ll 1$ that the columns of Hermite coefficients in the η direction are essentially colinear.

3.4.2 Effective drift and diffusion coefficients

We assume from now on that the index-set has a rectangular shape, $\mathcal{I}_d = \{0, 1, \dots, d_x\} \times \{0, 1, \dots, d_\eta\}$. At the continuous, infinite dimensional level, the absence of an effective drift term for the models of the noise **B** and **NS** as $\varepsilon \to 0$ is ensured by the centering condition (3.10). Any deviation from zero would lead to an effective drift term, scaling as $1/\varepsilon$ and proportional to

$$\sqrt{2\,\beta^{-1}}\,\zeta \int_{\mathbf{R}} \eta \, \exp\left(-\beta V_\eta(\eta)\right) \, \mathrm{d}\eta.$$

At the finite-dimensional, numerical level, a parasitic effective drift can arise even when V_η satisfies (3.10), as we demonstrate below. This can occur when V_η is not an even function, and it is especially critical when the number of basis functions used to approximate the solution in the η direction is relatively low, leading to a nonzero first moment of the approximate equilibrium probability density of the noise process. In these cases, it is useful to introduce an artificial drift term μ_d/ε in Galerkin formulation, for some constant μ_d to be determined. To formulate Result 3.2 below, let $\varphi_{0,d}$ denote the (assumed unique) normalized one-dimensional eigenfunction associated with $\lambda_{0,d}$,

$$\varphi_{0,d} = \underset{\varphi_d \in S^\eta_{d_\eta},\, \|\varphi_d\|_{\mathrm{e}^{V_\eta}}=1}{\arg\max} \langle \mathcal{L}_0^* \varphi_d, \varphi_d \rangle_{\mathrm{e}^{V_\eta}}\,, \qquad \text{where } S^\eta_{d_\eta} = \mathrm{e}^{-V_\eta/2}\, \mathrm{e}^{-\eta^2/4\sigma_\eta^2}\, \mathbf{P}(d_\eta).$$

Result 3.2. *Let μ_d be defined by*

$$\mu_d = -\sqrt{2\,\beta^{-1}}\,\zeta \int_{\mathbf{R}} \eta\, \varphi_{0,d}^2\, \mathrm{e}^{V(\eta)}\, \mathrm{d}\eta. \tag{3.40}$$

Then, when $\varepsilon \ll 1$, the solution ρ_d of

$$\frac{\partial \rho_d}{\partial t} = \hat{\Pi}_d \left(\mathcal{L}_\varepsilon^* + |\lambda_{0,d}|\, \mathrm{id} - \left(\frac{\mu_d}{\varepsilon}\right) \frac{\partial}{\partial x} \right) \hat{\Pi}_d\, \rho_d, \qquad \rho_d(x, \eta, t=0) = \rho_{d,0}(x)\, \varphi_{0,d}(\eta)$$

with $\mathcal{L}_\varepsilon^*$ as in Eq. (3.32), can be approximated by $\rho_d^x(x,t)\,\varphi_{0,d}(\eta)$, where ρ_d^x satisfies

$$\frac{\partial \rho_d^x}{\partial t} = \hat{\Pi}_d \left[\frac{\partial}{\partial x} \left(V'\, \rho_d^x + A_d\, \frac{\partial \rho_d^x}{\partial x} \right) \right], \qquad \rho_d^x(x, t=0) = \rho_{d,0}(x). \tag{3.41}$$

Here the effective diffusion A_d is equal to

$$A_d := \int_{\mathbf{R}} \left(-\hat{\Pi}_d\, \mathcal{L}_0\, \hat{\Pi}_d + |\lambda_{0,d}|\, \mathrm{id}\right)^{-1} (b_\eta\, \varphi_{0,d})\, (b_\eta\, \varphi_{0,d})\, \mathrm{e}^{V_\eta(\eta)}\, \mathrm{d}\eta, \tag{3.42}$$

with $b_\eta := \left(\mu_d + \sqrt{2\,\beta^{-1}}\,\zeta\,\eta\right)$ and where $\mathcal{L}_0$ is the formal L^2 adjoint of $\mathcal{L}_0^*$.

Proof. The argument below is formal, but it can be turned into a rigorous proof using standard methods in multiscale analysis; see e.g. [174]. Expanding the solution in powers of ε and gathering terms multiplying equal powers of ε, a system of equations similar to Eqs. (3.38a) to (3.38c) can be obtained, differing only by the presence of the corrective drift term next to $\hat{\mathcal{L}}_1$. The solvability condition for the first equation implies that $\varrho_0 = \rho_d^x(x,t)\,\varphi_{0,d}(\eta)$. For the second equation, we

see from the definition of μ_d and using the symmetry of $\hat{\mathcal{L}}_0$ in $L^2\left(\mathbf{R}^2, \mathrm{e}^{V_\eta(\eta)+U_x(x)}\right)$, that the Fredholm solvability condition is automatically satisfied, which enables solving for ϱ_1:

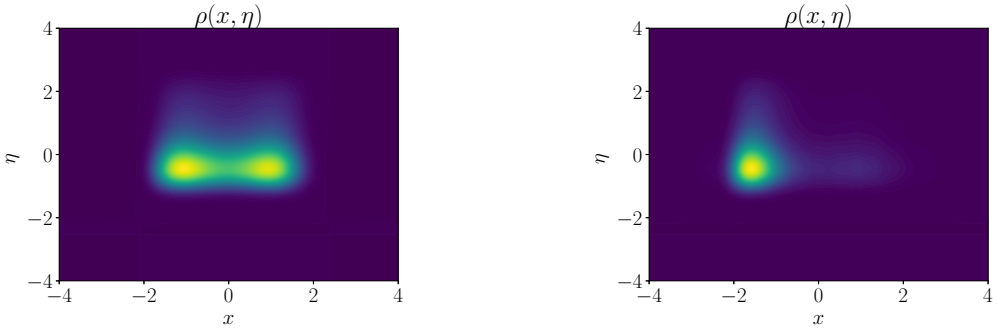
$$\varrho_1 = \left(-\hat{\Pi}_d\,\mathcal{L}_0^*\,\hat{\Pi}_d + |\lambda_{0,d}|\ \mathrm{id}\right)^{-1}\left(-(\mu_d + \sqrt{2\,\beta^{-1}}\,\zeta\,\eta)\,\varphi_{0,d}\right)\,\frac{\partial \rho_d^x}{\partial x} + \Phi_1(x,t)\,\varphi_{0,d}.$$

Writing out the solvability condition for the third equation, we obtain the effective equation for ρ_d^x:

$$\frac{\partial \rho_d^x}{\partial t} = \hat{\mathcal{L}}_2\,\rho_d^x(x,t) + \int_{\mathbf{R}}\left(\hat{\mathcal{L}}_1 - \mu_d\,(\hat{\Pi}_d\,\partial_x\hat{\Pi}_d)\right)\,\varrho_1\,\varphi_{0,d}\,\mathrm{e}^{V_\eta}\,\mathrm{d}\eta,$$

which after expansion of the terms is Eq. (3.41). □

In the examples we considered in Section 3.3, both $\varphi_{0,d}$ and V_η were even functions, so the corrective drift term defined by Eq. (3.40) was zero. This is not the case for the noise model **NS**, where the noise is confined by the uneven potential (3.11). To illustrate the importance of including the corrective drift term (3.40), we present simulation results with and without it in the case of the noise model **NS**. We consider the following parameters for the equation: $\varepsilon = 2^{-5}$, $\theta = 0$, $\beta = 1$, $V(x) = x^4/4 - x^2/2$, and for the numerics we choose $d = 20$, $\sigma_x^2 = \sigma_\eta^2 = 0.1$, $\mathrm{e}^{-U_x(x)} = 1$. The solutions obtained are presented in Figs. 3.11a and 3.11b.



(a) With the corrective drift.

(b) Without the corrective drift.

Figure 3.11: Comparison of the numerical solutions for the steady-state Fokker–Planck equation when the noise is confined by potential (3.11), with and without the corrective drift term. Without correction, the parasitic drift completely deteriorates the accuracy of the solution.

We note however that, for fixed ε, $\mu_d \to 0$ as $d \to \infty$, by convergence of $\varphi_{0,d}$ to $C\ \mathrm{e}^{-V_\eta}$. This is illustrated in Fig. 3.12, where the same parameters as those used in Fig. 3.11 were used. While the value of the bias decreases exponentially with the degree of approximation, we have found through numerical experiments that, when ε is of the order of 0.01, failure to account for the parasitic drift is very detrimental to the accuracy of the solution for d as high as 30.

3.4.3 Numerical verification of the rates of convergence

In this section, we verify numerically the rate of convergence of the stationary solution to the Galerkin formulation Eq. (3.33) in the limit $\varepsilon \to 0$. We recall that this Galerkin formulation is associated to the linear Fokker–Planck equation (3.33). In addition to verifying the rates of convergence, examining the limit $\varepsilon \to 0$ numerically will enable us to gain insight into the accuracy of the asymptotic expansions for moderate values of ε.

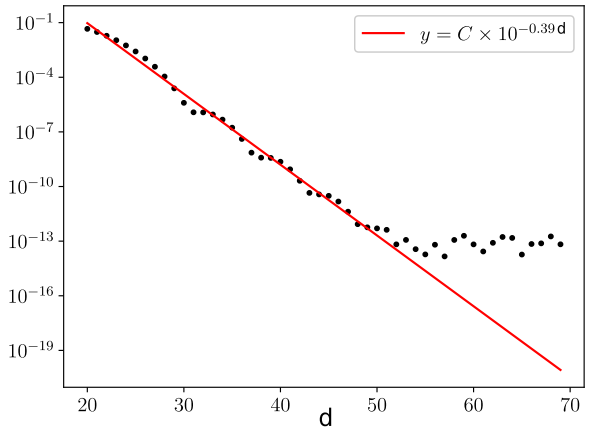


Figure 3.12: Normalized bias $\varepsilon\,\mu_d$ defined in (3.40), against degree of approximation, for $\sigma_\eta^2 = 0.1$. (The values shown need to be divided by ε to obtain the effective parasitic bias for a given value of ε.)

One-dimensional Ornstein–Uhlenbeck noise For this test we use the same parameters as in the convergence study for the bistable potential in Section 3.3.3. We verify the accuracy of the asymptotic expansion up to order ε^2 of the full solution in the x–η plane, given in Eq. (3.69), by comparing it to numerical results obtained using the spectral method introduced in Section 3.3. The convergence is presented in Fig. 3.13. We notice that, even for the smallest value of ε considered (2^{-6}), the norm of the difference between the asymptotic and spectral solutions appears to be roughly constant for $d \geq 20$. This is because, beyond this point, the spectral method is more accurate than the asymptotic expansion.

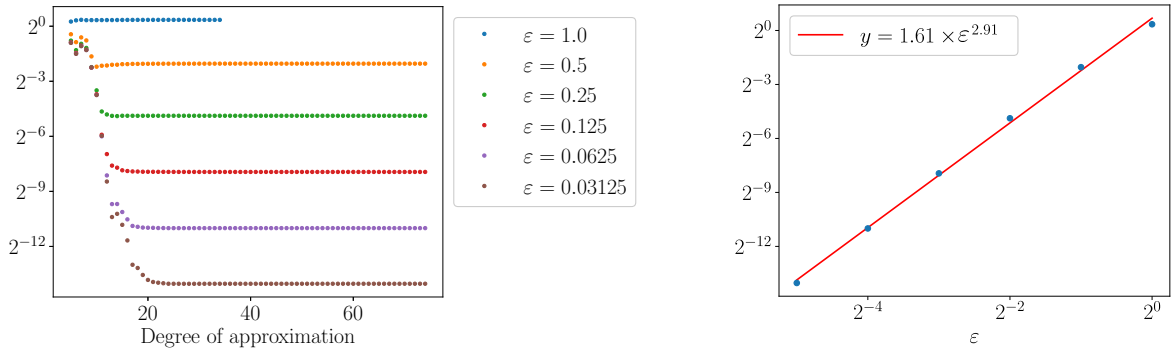


Figure 3.13: L^1 norm of the difference between the solutions found using the Hermite spectral method and the asymptotic expansion up to order ε^2(given in Eq. (3.69)), in both x and η, in the case of **OU** noise with $\varepsilon = 2^{-i}, i = 0, \ldots, 5$ (left), and the corresponding L^1 error of the asymptotic expansion (right). The scaling of the error (with respect to ε) is close to the expected value of 3.

Harmonic noise The second case we consider is that of harmonic noise, for which the order of the first nontrival correction in the expansion of the solution is $\mathcal{O}(\varepsilon^4)$, see Eq. (3.21). We confirmed this numerically for $V(x) = x^4/4 - x^2/2$ and $\beta = 5$ using 50 basis functions in each direction, with scaling factors $\sigma_x^2 = 1/30, \sigma_p^2 = \sigma_q^2 = 1$. The results are illustrated in Fig. 3.14.

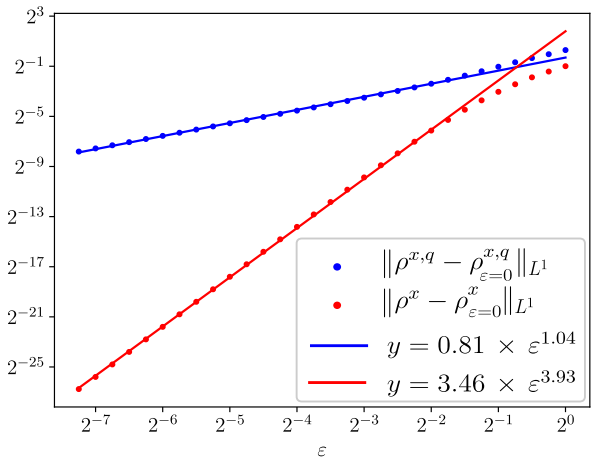


Figure 3.14: Convergence as $\varepsilon \to 0$ in the case of harmonic noise. The observed rate of convergence of the x-marginal to the white noise limit is 3.93, which is close to the theoretical value of 4. Here $\rho^{x,q}$ denotes the marginal of the solution on the x, q plane.

3.5 Results: effect of colored noise on bifurcations

In this section we present the bifurcation diagrams corresponding to the four models of the noise introduced in Section 3.2. We begin with the case of Gaussian noise, and later move to the case of non-Gaussian noise.

3.5.1 Construction of the bifurcation diagrams for the mean-field equation

We constructed the bifurcation diagrams using three different approaches:

Monte Carlo simulations We solved the system of interacting particles (3.8) with a large number of particles, and we approximated the first moment by ergodic average over an interval $(T, T + \Delta T)$, where T is sufficiently large to guarantee that the system can be considered to have reached its stationary state and ΔT is sufficiently large to ensure that the ergodic averages are accurate. By applying this procedure for a range of inverse temperatures, $\beta = 0.1, 0.15, 0.2, \ldots, 10$, we obtained the desired bifurcation diagram.

Perturbation expansions This approach, which we already outlined to in Section 3.2, relies on the fact that the self-consistency map can be approximated as $R(m, \beta) \approx R_0(m, \beta) + \varepsilon^\delta R_\delta(m, \beta)$, with good accuracy when $\varepsilon \ll 1$. Here we used the same notations as in Section 3.2, and in particular δ denotes the order of the first nontrivial correction in Eq. (3.19). Using arclength continuation[2] for the resulting approximate self-consistency equation, $m = R_0(m, \beta) + \varepsilon^\delta R_\delta(m, \beta)$, we can plot the first moment m as a function of β for a fixed value of ε. We note that, in view of the typical shape of the self-consistency map, depicted in a particular case in Fig. 3.2, a rudimentary root finding algorithm can be employed to initiate the arclength continuation at some initial inverse temperature β_0.

[2]We do this using the Moore–Penrose quasi-arclength continuation algorithm. Rigorous mathematical construction of the arclength continuation methodology can be found, e.g., in [134] and [7]. Some useful practical aspects of implementing arclength continuation are also given in [55]. See also [93].

The spectral method Finally, we employed the Galerkin method presented in Section 3.3.3. We considered two different methodologies: on the one hand, by calculating numerically an approximation $\rho_{d,\infty}(x, \eta; \beta, m)$ of the steady-state solution of the linear Fokker–Planck equation (3.17a) with fixed m and β, the self-consistency map can be approximated as $R(m, \beta) \approx \int_{\mathbf{R}} \int_{\mathbf{R}^n} x\, \rho_{d,\infty}(x, \eta; \beta, m)\, \mathrm{d}x\, \mathrm{d}\mathbf{y}$, after which a bifurcation diagram can be constructed by using the same method as in the previous paragraph. Each evaluation of the self-consistency map requires the computation of the eigenvector with eigenvalue of smallest magnitude of the discretized operator, which can be performed efficiently for sufficiently small systems using the *SciPy* toolbox. On the other hand, the time-dependent (nonlinear) McKean–Vlasov equation can be integrated directly using our spectral method. Since only the final solution is of interest to us, the semi-implicit time-stepping scheme (3.31) can be used with a large time step, which enables a quick and accurate approximation of the steady-state solutions. While both methodologies work well in the two-dimensional case, in three dimensions (harmonic noise) solving the McKean–Vlasov equation directly proved more efficient, so this is the approach we employed for all the tests presented in this section.

3.5.2 Gaussian case

The one-dimensional Ornstein–Uhlenbeck noise provides an ideal testbed for the three methods we use to construct bifurcation diagrams. Figure 3.15 below plots the bifurcation diagram of the first moment m as a function of β for $\varepsilon = 0.1, 0.2, \ldots, 0.5$. Three different initial conditions ($\mathbf{X}_0 \sim N(0, 0.1)$, $\mathbf{X}_0 \sim N(0.1, 0.1)$, and $\mathbf{X}_0 \sim N(-0.1, 0.1)$) were used for the MC simulations. Although we observe that the results of MC simulations tend to be less precise around the bifurcation point, the agreement between the three methods overall is excellent for $\varepsilon = 0.1, 0.2$. For the other values of ε, while the results of MC simulations and of our spectral method continue to agree, those obtained from the asymptotic expansion are significantly less accurate, which is consistent with the observations presented in Fig. 3.13.

With an Intel i7-3770 processor, the construction of each of the bifurcation diagrams in Fig. 3.13 took about 30 hours with the spectral method, and about 20 hours with the MC simulations.

The case of harmonic noise, corresponding to a three-dimensional McKean–Vlasov equation, is more challenging to tackle using our spectral method. When using 40 basis functions in each direction, the CPU time required to construct the full bifurcation diagram was of the order of a week. As a consequence of the lower number of basis functions used in this case, we observe a small discrepancy between the results of the spectral method and those of MC simulations for large β in the case $\varepsilon = 0.4$. Nevertheless, as can be seen in Fig. 3.16, for small ε the overall agreement between the three methods is excellent. We note in particular that, as suggested by the asymptotic expansions, the use of harmonic noise produces results much closer to the white noise limit than scalar OU noise.

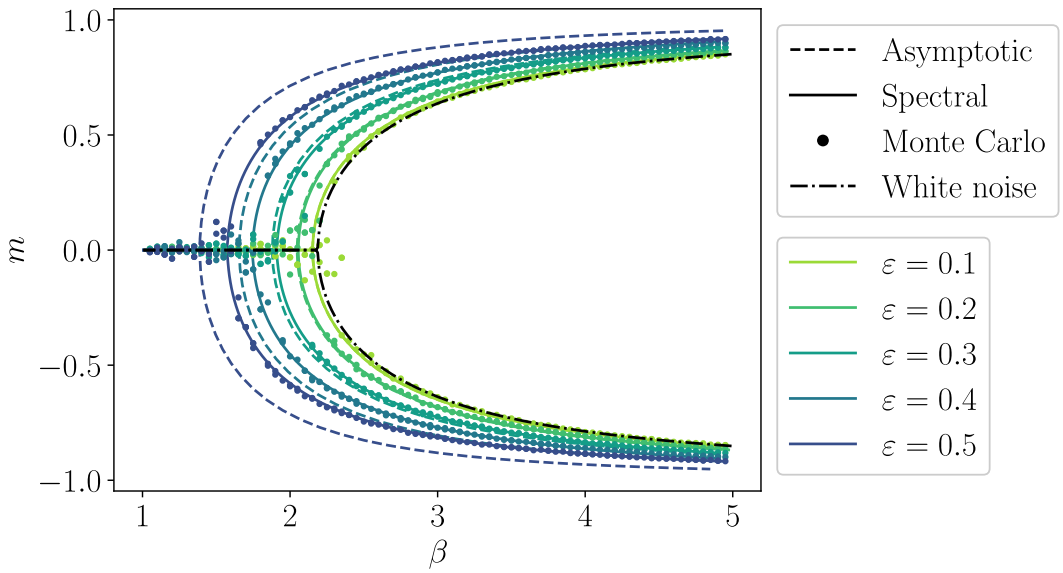


Figure 3.15: Bifurcation diagram of m against β for Ornstein–Uhlenbeck noise, obtained via MC simulation, the spectral method, and the asymptotic expansion (3.20).

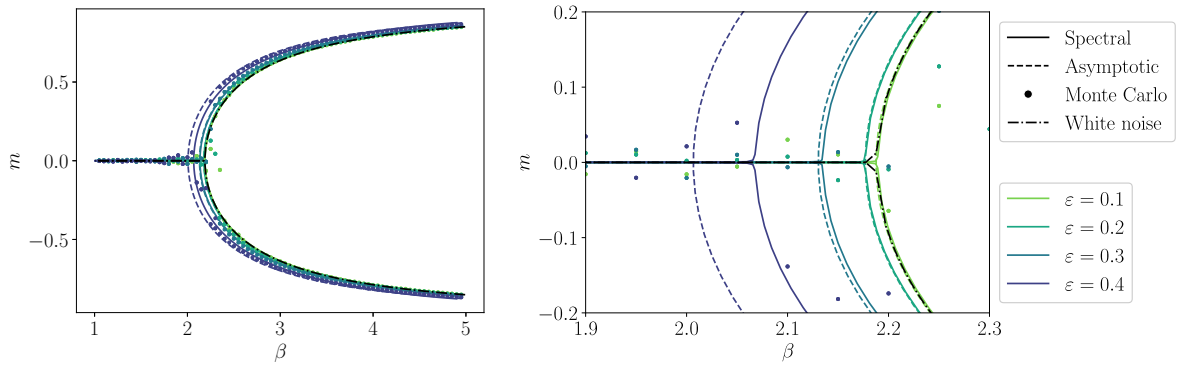


Figure 3.16: Bifurcation diagram of m against β for harmonic noise (model **H**), obtained via MC simulation, the Hermite spectral method, and the asymptotic expansion (3.20).

3.5.3 Non-Gaussian noise

For the non-Gaussian noise processes we consider, the x^4 asymptotic growth of the confining potentials in both directions causes the solution to the McKean–Vlasov equation to be stiffer than in the cases of OU and harmonic noise, especially for large values of ε. Consequently, we were not able to consider as wide a range of ε as in the previous subsection using the spectral method. Since, on the other hand, MC simulations become overly computationally expensive for small ε, the comparisons in this section comprise only results obtained using our spectral method and asymptotic expansions. Results of simulations for the bistable noise (model **B**) are presented in Fig. 3.17, in which a very good agreement can be observed.

For nonsymmetric noise (model **NS**), the two branches in the bifurcation diagram are separate, as illustrated in Fig. 3.18. Here too, the agreement between the spectral method and the asymptotic expansion is excellent. In contrast with the other models considered, the first nonzero term in the asymptotic expansion is of order ε, which is reflected by the manifestly higher sensitivity to the correlation time of the noise. In the right panel of Fig. 3.18, we present

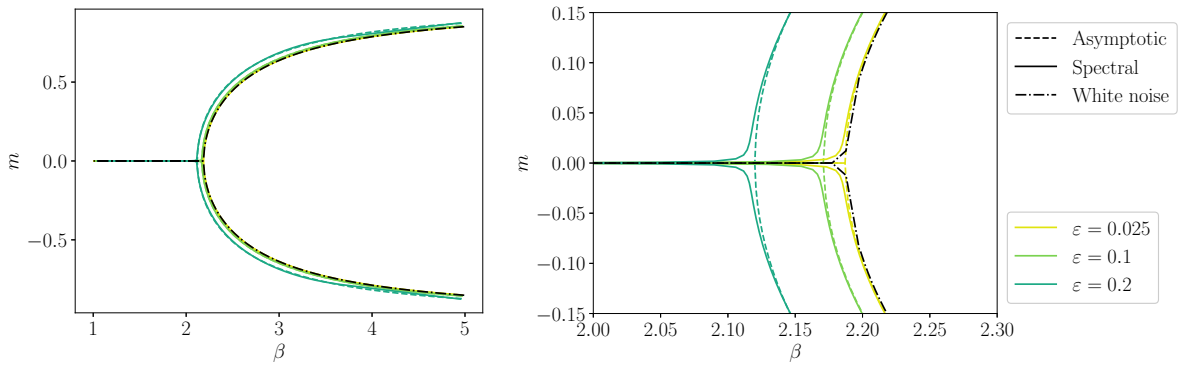


Figure 3.17: Bifurcation diagram of m against β for the bistable noise (model **B**), using the spectral method and a truncated asymptotic expansion including the first nonzero correction. We see that, overall, the agreement between the two methods is excellent.

the graph of $R_0(m;\beta) + \varepsilon R_1(m;\beta)$ for a value of β close to the point at which new branches (one stable and one unstable) emerge.

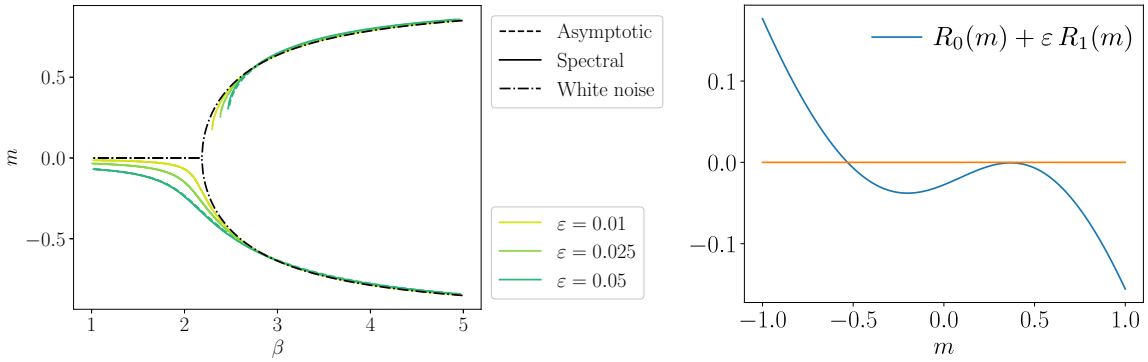


Figure 3.18: Left: bifurcation diagram of m against β for the nonsymmetric noise (model **NS**), using the spectral method and a truncated asymptotic expansion including the first nonzero correction. Right: $R_0 + \varepsilon R_1(m) - m$ against m for $\varepsilon = 0.1$ and $\beta = 2.6$.

3.5.4 Dependence of the critical temperature on ε

For the noise models **OU**, **H** and **B**, the effect of colored noise on the dynamics is a shift of the critical temperature: the pitchfork bifurcation occurs for *smaller* values of β (i.e., larger temperatures) as the correlation time increases. In order to further investigate the effect of the correlation time on the long time behavior of the system of interacting particles, we will compute the critical temperature as a function of ε based on the asymptotic expansions and compare with the results of spectral and MC simulations, see Fig. 3.19. Rather than finding the critical inverse temperature β_C for a range of values of ε (and for a fixed θ), it is convenient to fix β_C and find the corresponding ε, satisfying

$$\frac{\mathrm{d}}{\mathrm{d}m}\left(\int_{\mathbf{R}} x\, p_0(x;\beta_C,m)\,\mathrm{d}x\right)_{m=0} + \varepsilon^{\delta}\frac{\mathrm{d}}{\mathrm{d}m}\left(\int_{\mathbf{R}} p_\delta(x;\beta_C,m)\,\mathrm{d}x\right)_{m=0} = 1, \tag{3.43}$$

which is merely a polynomial equation in ε, the coefficient of which can be calculated by numerical differentiation. With this procedure, the dependence of the critical β upon ε can be calculated

on a fine mesh. In the case of **OU** noise, for example, both coefficients in the left-hand side of Eq. (3.43) are positive, implying that the equation has a solution (in fact, two, but one of them negative) only if β_C is lower than the inverse critical temperature in the white noise case.

Of the three methods employed in Fig. 3.19, the approach based on the asymptotic expansions has the lowest computational cost: calculating all the solid curves took only about a couple of minutes on a personal computer with an Intel i7-3770 processor. The data points associated with the spectral method and the MC simulations were obtained from the bifurcation diagrams presented above; refer to Section 3.5.2 for the corresponding computational costs.

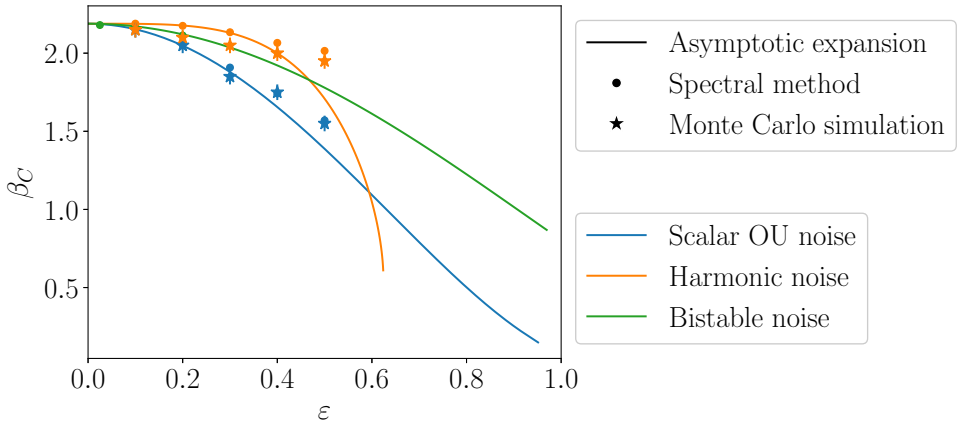


Figure 3.19: Critical β against ε.

3.6 Conclusions

In this paper, we introduced a robust spectral method for the numerical solution of linear and nonlinear, local and nonlocal Fokker–Planck-type PDEs that does not require that the PDE is a gradient flow. We then used our method to construct the bifurcation diagram for the stationary solutions of the mean-field limit of a system of weakly interacting particles driven by colored noise.

To verify our results, we also constructed the bifurcation diagrams by using two other independent approaches, namely by MC simulation of the N-particle system and by using explicit asymptotic expansions with respect to correlation time of the noise. In the small correlation time regime, we observed a very good agreement between all three methods. For larger values of the correlation time, the asymptotic expansions become inaccurate, but the results obtained via the spectral method and MC simulations continue to be in good agreement.

It appeared from our study that, unless the potential in which the noise process is confined is asymmetric, the correlation structure of the noise does not influence the topology of the bifurcation diagram: the mean-zero steady-state solution, which is stable for sufficiently large temperatures, becomes unstable as the temperature decreases below a critical value, at which point two new stable branches emerge, in the same manner as reported in [51, 196]. The correlation structure does, however, influence the temperature at which bifurcation occurs, and in general this temperature increases as the correlation time of the noise increases. In the presence

of an asymmetry in the confining potential of the noise, on the other hand, the two stable branches in the bifurcation diagram are separate, indicating that the system always reaches the same equilibrium upon decreasing the temperature. This behavior is similar to what has been observed previously in the white noise case when a tilt is introduced in the confining potential $V(\cdot)$, see [93, 94].

Several problems remain open for future work. On the theoretical front, we believe that the analysis we presented in Sections 3.A and 3.3.1 for the linear Fokker–Planck equation can be extended to both the linear Fokker–Planck equation with colored noise and the nonlinear McKean–Vlasov equation. Another direction for future research could be the rigorous study of bifurcations and, more specifically, of fluctuations and critical slowing down near the bifurcation point. On the modeling front, it would be interesting to consider more general evolution equations for the interacting particles, such as the generalized Langevin equation, and also to study systems of interacting particles subject to colored noise that is multiplicative.

3.A Proof of Theorem 3.1

Using the same notation as in Appendix A.3, we let Π_d be the $L^2\left(\mathbf{R}, e^{-V_q}\right)$ projection operator on $\mathbf{P}(d)$ and $\hat{\Pi}_d := e^{-V_q/2}\,\Pi_d\, e^{V_q/2}$. The solution u_d of Eq. (3.25a) satisfies $\partial_t u_d = \hat{\Pi}_d\,\mathcal{H}_x\,\hat{\Pi}_d\,u_d =: \mathcal{H}_d\,u_d$. Clearly, the operator $\mathcal{H}_d$ is selfadjoint on $\mathrm{e}^{-V_q/2}\,\mathbf{P}(d)$ with the $L^2(\mathbf{R})$ inner product, and it is also negative, by negativity of $\mathcal{H}_x$:

$$\langle \mathcal{H}_d w_d, w_d\rangle = \langle \mathcal{H}_x w_d, w_d\rangle \le 0 \quad \forall w_d \in \mathrm{e}^{-V_q/2}\,\mathbf{P}(d). \tag{3.44}$$

To prove the convergence of u_d when $d\to\infty$, we will rely on the following lemma.

Lemma 3.3. *Let $\hat{\partial}_x := \partial_x + x/2$, and assume that $\hat{\partial}_x^n u \in L^2(\mathbf{R})$ for $n = 0,\dots,m$. Then for all natural numbers m_1, m_2 such that $m_1+m_2\le m$, it holds that $x^{m_1}\,u^{(m_2)} \in L^2(\mathbf{R})$ and*

$$K_1(m)\max_{m_1+m_2\le m}\|x^{m_1}\,u^{(m_2)}\| \le \max_{0\le i\le m}\|\hat{\partial}_x^i u\| \le K_2(m)\max_{m_1+m_2\le m}\|x^{m_1}\,u^{(m_2)}\|, \tag{3.45}$$

where $K_1(m), K_2(m)$ are positive constants depending only on m and $\|\cdot\|$ is the usual $L^2(\mathbf{R})$ norm.

Proof. We denote by $H^m(\mathbf{R}, \mathrm{e}^{-x^2/2})$ the Sobolev space weighted by $\mathrm{e}^{-x^2/2}$,

$$H^m(\mathbf{R}, \mathrm{e}^{-x^2/2}) = \{v : v^{(i)} \in L^2\left(\mathbf{R}, \mathrm{e}^{-x^2/2}\right) \text{ for } i = 0,\dots,m\},$$

and by $\|\cdot\|_{m,\mathrm{e}^{-x^2/2}}$ the associated norm: $\|v\|^2_{m,\mathrm{e}^{-x^2/2}} = \sum_{i=0}^m \|v^{(i)}\|^2_{\mathrm{e}^{-x^2/2}}$. For the first inequality, we know from [195, Lemma B.6] that

$$\|xv\|_{\mathrm{e}^{-x^2/2}} \le 4\,\|v\|_{1,\mathrm{e}^{-x^2/2}} \qquad \forall v \in H^1(\mathbf{R}, \mathrm{e}^{-x^2/2}).$$

Applying this inequality repeatedly, we obtain

$$\|x^{m_1} v\|_{\mathrm{e}^{-x^2/2}} \leq C(m) \|v\|_{m,\mathrm{e}^{-x^2/2}}, \qquad m_1 = 0, \dots, m, \qquad \forall v \in H^m(\mathbf{R}, \mathrm{e}^{-x^2/2}), \tag{3.46}$$

for a constant $C(m)$ depending only on m. By definition, $\hat{\partial}_x u = \mathrm{e}^{-x^2/4}\, \partial_x(\mathrm{e}^{x^2/4}\, u)$, so the assumption implies that $\mathrm{e}^{x^2/4}\, u \in H^m(\mathbf{R}, \mathrm{e}^{-x^2/2})$, from which we obtain using Eq. (3.46) that, for $0 \leq m_1 \leq m$,

$$\|x^{m_1} u\| = \|x^{m_1} u\ \mathrm{e}^{x^2/4}\|_{\mathrm{e}^{-x^2/2}} \leq C(m) \|u\ \mathrm{e}^{x^2/4}\|_{m,\mathrm{e}^{-x^2/2}} = C(m) \sqrt{\sum_{i=0}^{m} \|\hat{\partial}_x^m u\|^2}.$$

This proves the first inequality of Eq. (3.45) in the case $m_2 = 0$. We assume now that the statement is proved up to $m_2 - 1$, and we show that it is valid for m_2. Using the triangle inequality we obtain

$$\|x^{m_1}\, u^{(m_2)}\| \leq \|x^{m_1}\, (u^{(m_2)} - \hat{\partial}_x^{m_2} u)\| + \|x^{m_1}\, \hat{\partial}_x^{m_2} u\|.$$

The derivatives in the first term are of order strictly lower than m_2, and therefore this term can be bounded by the induction assumption. The second term is bounded by applying the base case to $\hat{\partial}_x^{m_2} u$: introducing $v := \hat{\partial}_x^{m_2} u$, we notice that $\hat{\partial}_x^{m-m_2} v = \hat{\partial}_x^m u \in L^2(\mathbf{R})$ by assumption, so we can apply the first inequality in Eq. (3.45), without any derivative of v in the left-hand side, to deduce

$$\|x^{m-m_2} v\| \leq \max_{0 \leq i \leq m-m_2} \|\hat{\partial}_x^i v\| \leq \max_{0 \leq i \leq m} \|\hat{\partial}_x^i u\|.$$

The second inequality in (3.45) then holds trivially by expanding $\hat{\partial}_x$ and applying a triangle inequality. □

With Assumption 3.1, we can show that the two norms in Lemma 3.3 can be bounded from above by the norm $\sqrt{\langle (-\mathcal{H}_x + 1)^m u, u\rangle}$ for appropriate m.

Lemma 3.4 (Bound by alternative norm)**.** *If Assumption 3.1 holds, then*

$$\sum_{i=0}^{m} \|\hat{\partial}_x^m u\|^2 \leq C\ \langle (-\mathcal{H}_x + 1)^m u, u\rangle$$

for any smooth u for which the right-hand side is well-defined. Here C is a positive constant that depends on β, m, and on the particular expression of the potential W defined in Assumption 3.1.

Proof. Below C_1 and C_2 denote the same constants as in Assumption 3.1. First we notice that, for any constant $K > 1$,

$$\langle (-\mathcal{H}_x + K)^m u, u\rangle \leq K^m \langle (-\mathcal{H}_x + 1)^m u, u\rangle, \tag{3.47}$$

because $\mathcal{H}_x$ is a negative operator. Since W is a polynomial, its derivatives grow asymptotically

more slowly that W itself, and so it is possible for any $\varepsilon > 0$ to find $K \geq C_2$ large enough that:

$$\left|W^{(i)}(x)\right| \leq \varepsilon \left(W(x) + K\right) \qquad \forall x \in \mathbf{R}, \qquad \forall i \in \mathbf{N}. \tag{3.48}$$

We decompose $-\mathcal{H}_x + K$ as $(-\beta^{-1}\partial_x^2) + (W(x) + K)$. The two operators in this sum are positive because $K \geq C_2$ and by assumption $W(x) + C_2 \geq C_1(1 + |x|^2)$. Expanding the inner product in the left-hand side of Eq. (3.47) and using integration by parts,

$$\langle (-\mathcal{H}_x + K)^m u, u\rangle = \left(\sum_{i=0}^{m} \beta^{-i} \binom{m}{i} \int_{\mathbf{R}} \left(W(x) + K\right)^i \left(u^{(m-i)}(x)\right)^2 \mathrm{d}x\right) + \cdots, \tag{3.49}$$

where the remainder terms originate from the fact that the operators ∂_x and $(W(x) + K)$ do not commute. By using Eq. (3.48), these terms can be bounded for sufficiently large K by half the leading term in Eq. (3.49). To conclude, we further expand this leading term:

$$\begin{aligned}
\langle (-\mathcal{H}_x + K)^m u, u\rangle &\geq \frac{1}{2} \sum_{i=0}^{m} \binom{m}{i} \beta^{-i} \int_{\mathbf{R}} \left(W(x) + K\right)^i \left(u^{(m-i)}(x)\right)^2 \mathrm{d}x \\
&\geq \frac{1}{2} \sum_{i=0}^{m} \binom{m}{i} C_1^i \, \beta^{-i} \int_{\mathbf{R}} \left(1 + x^2\right)^i \left(u^{(m-i)}(x)\right)^2 \mathrm{d}x \\
&\geq \frac{1}{2} \sum_{i=0}^{m} \sum_{j=0}^{i} \binom{m}{i} \binom{i}{j} C_1^i \, \beta^{-i} \int_{\mathbf{R}} x^{2j} \left(u^{(m-i)}(x)\right)^2 \mathrm{d}x \\
&\geq C(m, \beta, C_1) \sum_{m_1 + m_2 \leq m} \|x^{m_1} \, u^{(m_2)}\|^2,
\end{aligned}$$

from which Lemma 3.3 allows us to conclude. □

Proof of Theorem 3.1. We assume for simplicity that $\sigma = 1$, and we begin by splitting the error as $u_d - u = (u_d - \hat{\Pi}_d u) + (\hat{\Pi}_d u - u) =: e_d + \delta_d$. The first term is related to the so-called consistency error, and the second to the approximation error. We obtain from Eqs. (3.24) and (3.25a)

$$\partial_t e_d = \hat{\Pi}_d \mathcal{H}_x \hat{\Pi}_d e_d + (\hat{\Pi}_d \mathcal{H}_x \hat{\Pi}_d - \hat{\Pi}_d \mathcal{H}_x) u.$$

Taking the inner product with e_d and using (3.44), this implies

$$\begin{aligned}
\langle \partial_t e_d, e_d\rangle &\leq \left\langle \mathcal{H}_x (\hat{\Pi}_d \, u - u), e_d \right\rangle \\
&\leq \frac{1}{2} \langle e_d, e_d\rangle + \frac{1}{2} \left\langle \mathcal{H}_x^2 (u - \hat{\Pi}_d \, u), (u - \hat{\Pi}_d \, u) \right\rangle.
\end{aligned}$$

We see from this equation that e_d can be controlled if one can bound the second inner product on the right-hand side. For this we use arguments similar to the ones employed in [2, 83]. Since

$\mathcal{H}_x$ is negative and selfadjoint, we notice

$$\begin{aligned}\left\langle (-\mathcal{H}_x)^i u(t), u(t)\right\rangle &= \left\langle (-\mathcal{H}_x)^i u_0, u_0\right\rangle + \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\left\langle (-\mathcal{H}_x)^i u(s), u(s)\right\rangle \mathrm{d}s \\ &= \left\langle (-\mathcal{H}_x)^i u_0, u_0\right\rangle - 2\int_0^t \left\langle (-\mathcal{H}_x)^{i+1} u(s), u(s)\right\rangle \mathrm{d}s \\ &\leq \left\langle (-\mathcal{H}_x)^i u_0, u_0\right\rangle, \end{aligned} \tag{3.50}$$

for $i = 1, 2, \ldots$, which implies that the inner products $\left\langle (-\mathcal{H}_x)^i u, u\right\rangle$ remain bounded for all positive times. We can now apply Theorem A.17 to obtain, using Lemmas 3.3 and 3.4 and Assumption 3.1,

$$\begin{aligned}\left\langle \mathcal{H}_x^2(u - \hat{\Pi}_d\, u), (u - \hat{\Pi}_d\, u)\right\rangle &\leq C \sum_{i=0}^{2k} \|\hat{\partial}_x^i (u - \hat{\Pi}_d\, u)\|^2 \\ &\leq C\, \frac{(d-m+1)!}{(d-2k+1)!}\, \|\hat{\partial}_x^m u\|^2 \\ &\leq C\, \frac{(d-m+1)!}{(d-2k+1)!}\, \left\langle (-\mathcal{H}_x^m + 1)u, u\right\rangle \\ &\leq C\, \frac{(d-m+1)!}{(d-2k+1)!}\, \left\langle (-\mathcal{H}_x^m + 1)u_0, u_0\right\rangle. \end{aligned}$$

We note that when V is quadratic, $k = 1$ is a valid choice in Assumption 3.1, and the bound above can be obtained by simply expanding u in terms of the eigenfunctions of $\mathcal{H}_x$, which in that case are just rescaled Hermite functions. Using Grönwall's inequality, we finally obtain

$$\begin{aligned}\|e_d(t)\|^2 &\leq \mathrm{e}^t\, \|e_d(0)\|^2 + \int_0^t \mathrm{e}^{(t-s)}\, \left\langle \mathcal{H}_x^2(u - \hat{\Pi}_d\, u), (u - \hat{\Pi}_d\, u)\right\rangle \mathrm{d}s, \\ &\leq \mathrm{e}^t \left(\|e_d(0)\|^2 + C\, \frac{(d-m+1)!}{(d-2k+1)!} \right). \end{aligned} \tag{3.51}$$

The first term, proportional to $\|e_d(0)\|^2$, depends only on the interpolation error of the initial condition, which is nonzero when using a Gauss–Hermite quadrature. It was proved that this error term also decreases spectrally, see e.g. [195, Theorems 7.17, 7.18], and in our case faster than the second error term. For the approximation error δ_d, similar inequalities to the ones used above can be used to obtain a bound of the type (3.51), which leads to the conclusion. □

3.B Asymptotic analysis

In this appendix, we present the methodology, presented in detail in [115] for the case of Ornstein–Uhlenbeck noise, that we used to derive asymptotic expansions of the type (3.20) and (3.21). For each model of the noise noise, we give an explicit expression of the first nonzero term in the expansion of the $x-$marginal. Although the asymptotic results in Section 3.B.1 were known, the results obtained in Sections 3.B.2 and 3.B.3 are novel. We also demonstrate why only the even degrees in the asymptotic expansion are nonzero when the noise is a general one-dimensional overdamped Langevin process in an even potential. The starting point is to decompose the

Fokker–Planck operator in Eq. (3.32) in powers of ε, which enables rewriting the stationary equation as

$$\left(\frac{1}{\varepsilon^2}\,\mathcal{L}_0^* + \frac{1}{\varepsilon}\,\mathcal{L}_1^* + \mathcal{L}_2^*\right)\rho = 0, \tag{3.52}$$

where

$$\mathcal{L}_1^*\cdot = -(\mathbf{y}_\eta^T\,\mathbf{y})\frac{\partial}{\partial x}\cdot, \tag{3.53a}$$

$$\mathcal{L}_2^*\cdot = \frac{\partial}{\partial x}\left(V'_{\text{eff}}(x)\,\cdot\right), \tag{3.53b}$$

and $\mathcal{L}_0^*$ is the Fokker–Planck operator corresponding to the model of the noise. For simplicity of notations, the constants have been absorbed in $\mathbf{y}_\eta$. We recall that $\mathcal{L}_0^*$ and $\mathcal{L}_2^*$ depend only on $\mathbf{y}$ and x, respectively.

We first show the uniqueness of the invariant measure by employing standard tools from [154]; see also [176, Appendix A]. For simplicity, we confine ourselves to the case of scalar noise governed by the overdamped Langevin equation in a potential V_η. It is convenient to consider the SDE associated with the Fokker–Planck equation:

$$\begin{cases} \mathrm{d}x = -\left(V'_{\text{eff}}(x) - \dfrac{c}{\varepsilon}\,\eta\right)\mathrm{d}t, \\ \mathrm{d}\eta = -\dfrac{1}{\varepsilon^2}V'_\eta(\eta)\mathrm{d}t + \dfrac{\sqrt{2}}{\varepsilon}\,\mathrm{d}W_t, \end{cases} \tag{3.54}$$

where $c = \zeta\sqrt{2\beta^{-1}}$. The associated transition kernel will be denoted by $P_t(\mathbf{x}, A)$, where here $\mathbf{x} = (x, \eta)^T$.

Proposition 3.5. *Assume that V_{eff} and V_η are polynomial confining potentials of degree greater than or equal to 2. Then the process $\{x(t), \eta(t)\}$ possesses a unique invariant measure.*

Proof. In [154] the authors show that, when simultaneously satisfied, the following conditions are sufficient to guarantee the existence of a unique invariant measure for Eq. (3.54).

- For all $\mathbf{x} \in \mathbf{R}^2$, $t > 0$ and nonempty open set $\mathcal{O}$, it holds that $P_t(\mathbf{x}, \mathcal{O}) > 0$;
- For all $t > 0$, the transition kernel $P_t(\mathbf{x}, A)$ possesses a jointly continuous density $p_t(\mathbf{x}, \mathbf{y})$ with respect to the Lebesgue measure;
- There exists a Lyapunov function $G : \mathbf{R}^2 \mapsto [1, \infty)$ such that $\lim_{|\mathbf{x}|\to\infty} = \infty$, as well as real numbers $a, d > 0$ such that

$$\mathcal{L}G(x, \eta) \leq -a\,G(x, \eta) + d, \qquad \forall (x, \eta) \in \mathbf{R}^2, \tag{3.55}$$

  where $\mathcal{L}$ is the generator of Eq. (3.54):

$$\mathcal{L} = \left(-V'_{\text{eff}}(x) + \frac{c}{\varepsilon}\,\eta\right)\frac{\partial}{\partial x} + \frac{1}{\varepsilon^2}\left(-V'_\eta(\eta)\,\frac{\partial}{\partial \eta} + \frac{\partial^2}{\partial \eta^2}\right).$$

Showing that the first and second conditions are satisfied can be achieved by following the same reasoning as in the proofs of [154, Theorem 3.2] and [154, Lemma 3.4], respectively, and the details

will not be presented here. For the third condition, it is easily checked that $G(x,\eta) = \frac{1}{2}x^2 + \frac{s}{2}\eta^2 + 1$ is a suitable Lyapunov function for sufficiently enough s: employing Young's inequality with parameter α, we find

$$\begin{aligned}\mathcal{L}G(x,\eta) &= -x\,V'_{\text{eff}}(x) - \frac{s}{\varepsilon^2}\eta\,V'_\eta(\eta) + \frac{c}{\varepsilon}x\eta + s,\\ &\leq -x\,V'_{\text{eff}}(x) - \frac{s}{\varepsilon^2}\eta\,V'_\eta(\eta) + \frac{c\,\alpha}{2\,\varepsilon}x^2 + \frac{c}{2\,\varepsilon\,\alpha}\eta^2 + s.\end{aligned}$$

Since V_{eff} and V_η are assumed to be polynomial confining potentials of degree at least 2, there exist positive constants C_1 and C_2 such that:

$$x\,V'_{\text{eff}}(x) \geq C_1(1+x^2) - C_2, \qquad \frac{s}{\varepsilon^2}\eta\,V'_{\text{eff}}(\eta) \geq C_1\,s\,\eta^2 - C_2,$$

so

$$\mathcal{L}G(x,\eta) \leq \left(\frac{c\,\alpha}{\varepsilon} - 2C_1\right)\frac{x^2}{2} + \left(\frac{c}{\varepsilon\alpha s} - 2C_1\right)\frac{s\,\eta^2}{2} - C_1 + s + 2C_2.$$

Taking $\alpha = \frac{\varepsilon C_1}{c}$ and $s = \frac{c}{\varepsilon\alpha C_1}$, we obtain

$$\mathcal{L}G(x,\eta) \leq -C_1 G(x,\eta) + s + 2C_2,$$

which is Eq. (3.55) with $a = C_1$ and $d = s + 2C_2$. Employing tools from [154], it is possible to also show that the convergence to the invariant measure is exponential, but this will not be necessary for our purposes. □

We denote the unique steady-state solution of Eq. (3.52) by $\rho^\varepsilon_\infty : \mathbf{R}\times\mathbf{R}^n \mapsto \mathbf{R}$, with $n = 1$ or 2 depending on the model of the noise, and we expand this function in powers of ε, omitting for simplicity the dependence on n and β:

$$\rho^\varepsilon_\infty(x,\mathbf{y}) = p_0(x,\mathbf{y}) + \varepsilon\,p_1(x,\mathbf{y}) + \varepsilon^2\,p_2(x,\mathbf{y}) + \dots. \tag{3.56}$$

To ensure that ρ^ε_∞ is a probability density for all $\varepsilon > 0$, we must impose that

$$\int_{\mathbf{R}\times\mathbf{R}^n} p_i\,\mathrm{d}x\,\mathrm{d}\mathbf{y} = \begin{cases} 1 \text{ if } i = 0,\\ 0 \text{ if } i > 0.\end{cases} \tag{3.57}$$

We are interested in the marginal distribution $\rho^\varepsilon_\infty(x) := \int \rho^\varepsilon_\infty(x,\mathbf{y})\,\mathrm{d}\mathbf{y}$. Substituting the expansion (3.56) into (3.52), we obtain

$$\mathcal{L}_0^*\,p_0 = 0, \tag{3.58a}$$

$$\mathcal{L}_0^*\,p_1 + \mathcal{L}_1^*\,p_0 = 0, \tag{3.58b}$$

$$\mathcal{L}_0^*\,p_{i+2} + \mathcal{L}_1^*\,p_{i+1} + \mathcal{L}_2^*\,p_i = 0 \qquad \text{for } i \geq 0. \tag{3.58c}$$

From Eq. (3.58a) and the ergodicity of the noise process, we deduce that the leading order term

$p_0(x, \mathbf{y})$ can be expressed as

$$p_0(x, \mathbf{y}) = U_0(x)\, p_s(\mathbf{y}), \tag{3.59}$$

where $p_s(\mathbf{y})$ is the unique invariant probability measure of the noise process and $U_0(x)$ is a function to be determined. For all of the noise models we consider in this chapter (see Section 3.2), $p_s(\mathbf{y})$ is of the form $\mathrm{e}^{-V_{\mathbf{y}}(\mathbf{y})}$, for a potential $V_{\mathbf{y}}$ that is known explicitly. This potential is quadratic for the models **OU** and **H**, and it coincides with V_η for the models **B** and **NS**, so in all cases $p_s(\mathbf{y}) > 0$ for all $\mathbf{y} \in \mathbf{R}^n$. To simplify the calculations, it is useful at this point to introduce new unknown functions $u_i := p_i/p_s$. Rewriting system (3.58) in terms of these, we obtain, after multiplication of each equation by p_s^{-1},

$$\mathcal{L}_0\, u_0 = 0, \tag{3.60a}$$

$$\mathcal{L}_0\, u_1 + \mathcal{L}_1^*\, u_0 = 0, \tag{3.60b}$$

$$\mathcal{L}_0\, u_{i+2} + \mathcal{L}_1^*\, u_{i+1} + \mathcal{L}_2^*\, u_i = 0 \qquad \text{for } i \geq 0, \tag{3.60c}$$

where $\mathcal{L}_0 := p_s^{-1}\mathcal{L}_0^* p_s$. If the dynamics of the noise is reversible, the operator $\mathcal{L}_0$ coincides with the formal $L^2(\mathbf{R} \times \mathbf{R}^n)$ adjoint of $\mathcal{L}_0^*$, i.e. it is the backward Kolmogorov operator associated with the noise process. Regardless of whether the noise process is reversible or not, the operator $\mathcal{L}_0$, when viewed as an operator on $L^2(\mathbf{R}^n, p_s)$, has a one-dimensional kernel spanned by constant functions in $\mathbf{y}$. The solvability conditions associated with Eqs. (3.60b) and (3.60c), provided by the Fredholm alternative, will enable us to obtain an equation for $U_0(x)$. It is useful here to introduce the integration operator w.r.t. p_s and the projection operator on the space of p_s-mean zero functions,

$$\mathcal{I}_Y u = \int_{\mathbf{R}^n} u(\,\cdot\,, \mathbf{y})\, p_s(\mathbf{y})\, \mathrm{d}\mathbf{y}, \qquad \Pi_0 u = u - \mathcal{I}_Y u, \tag{3.61}$$

as well as the solution operator of the Poisson equation $-\mathcal{L}_0 u = f$,

$$\mathcal{A} : f \mapsto -\Pi_0\, \mathcal{L}_0^{-1}\, \Pi_0\, f. \tag{3.62}$$

In view of Eqs. (3.53b) and (3.59), the solvability condition for (3.60b) is automatically satisfied so we can solve for u_1, whereby we obtain $u_1(x, y) = -\mathcal{A}(\mathbf{y}_\eta^T \mathbf{y})\, U_0'(x) + U_1(x)$. Solving Eq. (3.60c) formally and rewriting the resulting equations as a first order recursion, we obtain:

$$\begin{pmatrix} u_{i+2} \\ u_{i+1} \end{pmatrix} = \begin{pmatrix} \mathcal{A}\,\mathcal{L}_1^* & \mathcal{A}\,\mathcal{L}_2^* \\ \mathrm{id} & 0 \end{pmatrix} \begin{pmatrix} u_{i+1} \\ u_i \end{pmatrix} + \begin{pmatrix} U_{i+2} \\ 0 \end{pmatrix}, \qquad i = -1, 0, 1, \dots \tag{3.63}$$

where U_i are functions depending only on x and with the initial conditions $u_{-1} := 0$ and $u_0 = U_0$. The solution can be expressed as:

$$\begin{pmatrix} u_{i+1} \\ u_i \end{pmatrix} = \sum_{j=-1}^{i} \begin{pmatrix} \mathcal{A}\,\mathcal{L}_1^* & \mathcal{A}\,\mathcal{L}_2^* \\ \mathrm{id} & 0 \end{pmatrix}^{i-j} \begin{pmatrix} U_{j+1} \\ 0 \end{pmatrix}, \qquad i = -1, 0, 1, \dots \tag{3.64}$$

In particular, calculating explicitly the coefficients of the terms multiplying U_{i+1}, U_i, U_{i-1}, and

taking into account that $\mathcal{A}\mathcal{L}_2^* U_i = 0$ by definition of $\mathcal{A}$ and $\mathcal{L}_2^*$, one obtains

$$u_{i+1} = U_{i+1} - \mathcal{A}(\mathbf{y}_\eta^T \mathbf{y})\, U_i' + \mathcal{A}(\mathbf{y}_\eta^T \mathbf{y}\, \mathcal{A}(\mathbf{y}_\eta^T \mathbf{y}))U_{i-1}'' + \cdots \tag{3.65}$$

To find the functions U_j, we use the solvability conditions corresponding to (3.60c): applying $\mathcal{I}_Y$ to both sides of that equation, we obtain for $i = 0, 1, \ldots$

$$\begin{aligned} 0 = \mathcal{I}_Y \mathcal{L}_1^* u_{i+1} + \mathcal{L}_2^* U_i =& \frac{\mathrm{d}}{\mathrm{d}x}\left(V_{\mathrm{eff}}' U_i + \mathcal{I}_Y(\mathbf{y}_\eta^T \mathbf{y}\, \mathcal{A}(\mathbf{y}_\eta^T \mathbf{y}))U_i'\right) \\ &+ \mathcal{I}_Y\, \mathcal{L}_1^* \sum_{j=-1}^{i-2} \begin{pmatrix} \mathrm{id} \\ 0 \end{pmatrix}^T \begin{pmatrix} \mathcal{A}\,\mathcal{L}_1^* & \mathcal{A}\,\mathcal{L}_2^* \\ \mathrm{id} & 0 \end{pmatrix}^{i-j} \begin{pmatrix} U_{j+1} \\ 0 \end{pmatrix}. \end{aligned} \tag{3.66}$$

In the case $i = 0$, we obtain a stationary Fokker–Planck equation for U_0,

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(V_{\mathrm{eff}}' U_0 + \mathcal{I}_Y(\mathbf{y}_\eta^T \mathbf{y}\, \mathcal{A}(\mathbf{y}_\eta^T \mathbf{y}))U_0'\right) = 0. \tag{3.67}$$

From this equation, we can read off the effective diffusion coefficient in the limit where $\varepsilon \to 0$, $D := \sqrt{2\,\mathcal{I}_Y(\mathbf{y}_\eta^T \mathbf{y}\, \mathcal{A}(\mathbf{y}_\eta^T \mathbf{y}))}$, with the expression in the square root being non-negative because $\mathcal{A}$ is a positive operator. A detailed study of the limiting equation (3.67) is presented in [175, Result 11.1]. To recover the white noise limit when $\varepsilon \to 0$ with temperature β^{-1} and coefficient $\sqrt{2}$ in front of the noise, the coefficient $\mathbf{y}_\eta$ must satisfy:

$$\sqrt{2\,\mathcal{I}_Y(\mathbf{y}_\eta^T \mathbf{y}\, \mathcal{A}(\mathbf{y}_\eta^T \mathbf{y}))} = \sqrt{2\beta^{-1}} \;\to\; \|\mathbf{y}_\eta\|^2 = \frac{1}{\beta\, \mathcal{I}_Y(\hat{\mathbf{y}}_\eta^T \mathbf{y}\, \mathcal{A}(\hat{\mathbf{y}}_\eta^T \mathbf{y}))} \text{ with } \hat{\mathbf{y}}_\eta := \frac{\mathbf{y}_\eta}{\|\mathbf{y}_\eta\|} \tag{3.68}$$

Assuming that $\mathbf{y}_\eta$ is such that the condition above is satisfied, we can solve Eq. (3.67) together with the normalization condition (3.57) to obtain

$$U_0 = \frac{1}{\mathcal{Z}} \exp(-\beta V_{\mathrm{eff}}(x)), \qquad \text{with } \mathcal{Z} := \int_{\mathbf{R}} \exp(-\beta V_{\mathrm{eff}}(x))\, \mathrm{d}x.$$

From here on, we study the asymptotic expansion on a case by case basis.

3.B.1 Ornstein–Uhlenbeck noise

Details of the computations in this case can be found in [115]. According to (3.68), we find that the correct choice of $\mathbf{y}_\eta$ to recover the white noise limit is $\mathbf{y}_\eta = \sqrt{\beta^{-1}}$. Up to order ε^2, we obtain the following expansion, writing $\mathbf{y}$ as η to emphasize that the noise is 1-dimensional,

$$\rho_\infty^\varepsilon(x, \eta) = \frac{1}{\mathcal{Z}} \exp\left(-\frac{\eta^2}{2} - \beta V_{\mathrm{eff}}(x)\right)\left(1 + \varepsilon\left(\sqrt{\beta}\,\eta\, V_{\mathrm{eff}}'(x)\right)\right. \tag{3.69}$$

$$\left. + \varepsilon^2\left(C - \beta\left(1 - \frac{\eta^2}{2}\right) V_{\mathrm{eff}}'(x)^2 + \frac{1}{2}(3 - \eta^2) V_{\mathrm{eff}}''(x)\right) + \mathcal{O}(\varepsilon^3)\right) \tag{3.70}$$

This is obtained by developing (3.65) up to $u_4(x, \eta)$ and by considering the associated solvability conditions (3.66). The first order correction is averaged out by integration w.r.t. η, so the first

term with a nonzero $x-$marginal is an $\mathcal{O}(\varepsilon^2)$ correction:

$$\rho_\infty^\varepsilon(x) = \frac{\mathrm{e}^{-\beta V_{\mathrm{eff}}(x)}}{\mathcal{Z}} \left(1 + \varepsilon^2 \left(C - \frac{\beta}{2}\left(V'_{\mathrm{eff}}(x)\right)^2 + V''_{\mathrm{eff}}(x)\right) + \mathcal{O}(\varepsilon^4)\right), \tag{3.71}$$

The constant C is obtained from the normalization condition (3.57):

$$C = \int_{\mathbf{R}} \frac{\mathrm{e}^{-\beta V_{\mathrm{eff}}(x)}}{\mathcal{Z}} \left(\frac{\beta}{2}\left(V'_{\mathrm{eff}}(x)\right)^2 - V''_{\mathrm{eff}}(x)\right) \mathrm{d}x.$$

We note that, in the more general case where the coefficient multiplying the noise term depends on x, we would have obtained an effective SDE with the noise interpreted in the Stratonovich sense:

$$\mathrm{d}x_t = -V'_{\mathrm{eff}}(x_t)\,\mathrm{d}t + \sqrt{2\,\beta^{-1}}\, D(x_t) \circ \mathrm{d}W_t.$$

3.B.2 Harmonic noise

Here we use the notation $\mathbf{y} = (q, p)^T$, and we assume that the noise process is governed by

$$\mathrm{d}q_t = p_t\,\mathrm{d}t, \tag{3.72a}$$
$$\mathrm{d}p_t = -q_t\,\mathrm{d}t - p_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{3.72b}$$

which is the Langevin equation in the confining potential $q^2/2$, with friction coefficient 1 and inverse temperature $\beta^{-1} = 1$. The correct scaling of $\mathbf{y}_\eta$ to recover the white noise limit is in this case $\mathbf{y}_\eta = (1/\sqrt{\beta}, 0)^T$. To verify this, we can observe that the spectral density of $\frac{1}{\varepsilon^2}\, q\left(\frac{t}{\varepsilon^2}\right)$ converges pointwise to that of white noise (with coefficient $\sqrt{2}$) when $\varepsilon \to 0$:

$$S_q(\omega) = \frac{(1/\pi)}{(1-\varepsilon^4\omega^2)^2 + \varepsilon^4\omega^2} \to (1/\pi), \tag{3.73}$$

implying $\mathbf{E}(q_t q_s) \to 2\,\delta(t)$. The spectral properties of the generator of this noise process have been studied extensively, and explicit expressions for the eigenfunctions in terms of Hermite polynomials are available, see e.g. [174], enabling the calculation of the terms in Eq. (3.65), as was possible for the scalar OU noise.

In this case, it is also possible to apply a simple parity argument to show that the terms multiplying odd powers of ε in the asymptotic expansion are zero. Indeed, the operator $\mathcal{L}_0$ (which coincides with the generator of (3.72a) up to the sign of the antisymmetric part), and therefore also the corresponding solution operator $\mathcal{A}$ defined in (3.62), modify the parity of their argument in a well-defined manner. Decomposing $f(p, q)$ into four parts of different parities:

$$\begin{aligned}
4\,f(p,q) =& f(p,q) + f(-p,q) + f(p,-q) + f(-p,-q) && \text{(even-even)}\\
&+ f(p,q) + f(-p,q) - f(p,-q) - f(-p,-q) && \text{(even-odd)}\\
&+ f(p,q) - f(-p,q) + f(p,-q) - f(-p,-q) && \text{(odd-even)}\\
&+ f(p,q) - f(-p,q) - f(p,-q) + f(-p,-q) && \text{(odd-odd)}\\
=:& f_{ee}(p,q) + f_{eo}(p,q) + f_{oe}(p,q) + f_{oo}(p,q),
\end{aligned}$$

we will say that f has the same parity in p and q if only f_{ee} and f_{oo} above are nonzero, and that it has a different parity in p and q if only f_{eo} and f_{oe} are nonzero. Using the usual decomposition of the $\mathcal{L}_0$ into symmetric and antisymmetric parts

$$\mathcal{L}_0 = \left(-p\frac{\partial}{\partial p} + \frac{\partial^2}{\partial p^2}\right) + \left(p\,\frac{\partial}{\partial q} - q\,\frac{\partial}{\partial p}\right), \tag{3.74}$$

we see that the symmetric part changes neither the parity in p nor the parity in q, while the antisymmetric part changes both the parity in p and the parity in q. It follows from this that, if $f(p,q)$ has the same (resp. a different) parity in p and q, then $\mathcal{L}_0 f$ also has the same (resp. a different) parity in p and q. In contrast, the operator $\mathcal{L}_1^*$ maps functions with the same parity in p and q to functions with a different parity in p and q, and conversely. This explains, looking as in the previous section at powers of the matrix in Eq. (3.63), why every term multiplying an odd power of ε is zero in the asymptotic expansion (of the x-marginal). To justify why also the term of order ε^2 cancels out, it is thus sufficient to show that $U_2 = 0$, which we do by calculating explicitly the coefficients in the equation for U_2 in Eq. (3.66). Rearranging the terms, and taking into account that $\mathcal{A}\mathcal{L}_2^* U_0 = 0$ by definition of $\mathcal{A}$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\left(V'_{\mathrm{eff}}(x)\,U_2 + \beta^{-1}U_2'\right) &= -\,\mathcal{I}_Y\mathcal{L}_1^*\Big((\mathcal{A}\mathcal{L}_1^*)^3 + (\mathcal{A}\mathcal{L}_2^*)(\mathcal{A}\mathcal{L}_1^*)\Big)\,U_0 \\
&= -\,\frac{1}{\beta^2}\,U_0^{(iv)} \times \mathcal{I}_Y\left(-\frac{p^3q}{6} + \frac{p^2q^2}{2} - \frac{pq^3}{2} + \frac{q^4}{6} - q^2\right) \\
&\quad -\,\frac{1}{\beta}\,\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}}{\mathrm{d}x}\left(V'_{\mathrm{eff}}(x)\,U_0'\right)\right) \times \mathcal{I}_Y(-pq) = -0 - 0,
\end{aligned}$$

implying $U_2 = 0$. Looking at the solvability condition for the equation containing $\mathcal{L}_0 u_6$, we obtain an equation for the correction U_4, which becomes, after carrying out the integrals,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x}\left(V'_{\mathrm{eff}}\,U_4 + \frac{1}{\beta}U_4'\right) = \frac{1}{\beta^3}\,U_0^{(6)} + \frac{1}{\beta^2}\frac{\mathrm{d}^4}{\mathrm{d}x^4}\left(V'_{\mathrm{eff}}\,U_0'\right) \\
&+ \frac{1}{\beta^2}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(V'_{\mathrm{eff}}\,U_0'''\right) + \frac{1}{\beta}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(V'_{\mathrm{eff}}\frac{\mathrm{d}}{\mathrm{d}x}\left(V'_{\mathrm{eff}}\,U_0'\right)\right).
\end{aligned}$$

Integrating once and rearranging the terms, we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{\beta V_{\mathrm{eff}}}\,U_4\right) = C_1\,\beta\,e^{\beta V_{\mathrm{eff}}} + \left(-V'_{\mathrm{eff}}\,\frac{\mathrm{d}}{\mathrm{d}x} + \frac{1}{\beta}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)^2\left(e^{\beta V_{\mathrm{eff}}}U_0'\right).$$

Integrating a second time, we arrive at an expression for U_4, where we set $C_1 = 0$ in view of the normalization condition (3.57):

$$\frac{U_4}{U_0} = \left(C_2 + \frac{1}{2}\left(V''_{\mathrm{eff}}\right)^2 - \beta\left(V'_{\mathrm{eff}}\right)^2 V''_{\mathrm{eff}} + 2\,V'_{\mathrm{eff}}V'''_{\mathrm{eff}} - \frac{1}{\beta}\,V_{\mathrm{eff}}^{(iv)} + \beta\int_0^x \left(V''_{\mathrm{eff}}\right)^2\,V'_{\mathrm{eff}}\,\mathrm{d}x\right),$$

with C_2 a constant such that U_4 integrates to 0. We note that, as expected, U_4 is an even function if the potential V_{eff} is even.

3.B.3 General reversible noise

To conclude this appendix, we examine the case where the noise process is one-dimensional and reversible but not necessarily Gaussian:

$$\mathrm{d}\eta_t = -V_\eta'(\eta_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t. \tag{3.77}$$

Writing out Eq. (3.66) for $i = 1$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(V_{\mathrm{eff}}'(x)\,U_1 + \beta^{-1}U_1'\right) + \mathcal{I}_Y\left(\mathcal{L}_1^*(\mathcal{A}\mathcal{L}_1^*)^2 U_0\right) = 0. \tag{3.78}$$

When V_η is an even function of η, $\mathcal{A}$ is an operator that maintains the parity of its argument, while $\mathcal{L}_1^*$ inverts it. It follows from this that the second term in (3.78) is identically zero. In general, even powers of the matrix in Eq. (3.63) are such that the diagonal entries are operators that maintain parity while the off-diagonal entries are operators that invert parity, and conversely for odd powers, so in fact $U_i = 0$ for all odd values of i. We can calculate U_2 for a general symmetric potential V_η by developing Eq. (3.66) and solving the linear systems using our spectral method, see Section 3.3:

$$\begin{aligned}\begin{pmatrix} U_2 \\ U_0 \end{pmatrix} =& \frac{\mathcal{I}_Y(\eta\mathcal{A}(\eta\mathcal{A}(\eta\mathcal{A}\eta)))}{(\mathcal{I}_Y(\eta\,\mathcal{A}\eta))^2}\left(V_{\mathrm{eff}}'' - \frac{3\,\beta}{2}\left|V_{\mathrm{eff}}'\right|^2 + \beta^2\int_0^x \left(V_{\mathrm{eff}}'(y)\right)^3\,\mathrm{d}y\right) \\ &+ \frac{\mathcal{I}_Y(\eta\mathcal{A}(\mathcal{A}(\eta)))}{\mathcal{I}_Y(\eta\,\mathcal{A}\eta)}\left(\frac{\beta}{2}\left|V_{\mathrm{eff}}'\right|^2 - \beta^2\int_0^x \left(V_{\mathrm{eff}}'(y)\right)^3\,\mathrm{d}y\right) + C,\end{aligned} \tag{3.79}$$

with a constant C such that U_2 integrates to 0. We can verify that, for scalar OU noise, the coefficients are both equal to 1 and we recover (3.71). In the case of bistable noise (model **B**), the coefficients multiplying the first and second brackets are equal to 0.175 and 1.259, respectively. When V_η is not an even function, we can obtain a expression for U_1 by separating the operator on the right-hand side of Eq. (3.78) into two parts, acting respectively on functions of x or η:

$$\begin{pmatrix} U_1 \\ U_0 \end{pmatrix} = \frac{\mathcal{I}_Y(\eta\mathcal{A}(\eta\mathcal{A}\eta))}{(\mathcal{I}_Y(\eta\,\mathcal{A}\eta))^{3/2}}\left(C - \beta^{1/2}\,V_{\mathrm{eff}}' + \beta^{3/2}\int_0^x |V_{\mathrm{eff}}'(y)|^2\,\mathrm{d}y\right), \tag{3.80}$$

with C a constant such that U_1 integrates to 0. For the model of the noise **NS**, we calculate numerically that the coefficient on the right-hand side is equal to 1.556.

3.C Additional numerical results

In this appendix we present the results of supplementary numerical experiments, with the aim of further validating the Galerkin method presented in Section 3.3.3.

3.C.1 Convergence of the method for the linear Fokker-Planck equation

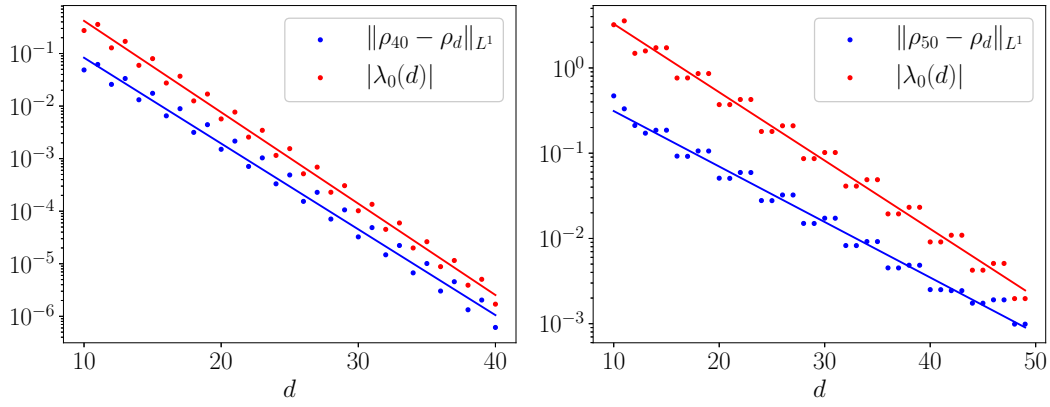
In this subsection, we study the performance of our spectral method for the calculation of the steady-state solution to the linear Fokker-Planck equation (3.32). For simplicity, we will consider that the interaction strength θ, which enters in the effective potential V_{eff}, is equal to zero.

Harmonic noise Let us begin by studying the convergence of the method for harmonic noise, i.e. model (**H**) in Section 3.2. In this case, the Fokker–Planck equation of interest, written in terms of the variables $\mathbf{y} =: (p, q)^T$, is the following:

$$\frac{\partial}{\partial x}\left(V'\,\rho - \frac{1}{\varepsilon}\sqrt{\beta^{-1}}\,q\,\rho\right) + \frac{1}{\varepsilon^2}\left(-p\,\frac{\partial}{\partial q} + q\frac{\partial}{\partial p} + \frac{\partial}{\partial p}\left(\rho\,p + \frac{\partial \rho}{\partial p}\right)\right) = 0. \tag{3.81}$$

Given the large size of the linear system obtained after Galerkin discretization, we calculated the steady-state solution by integrating the time-dependent equation, using the backward Euler method, until the difference between successive iterations was below an appropriate threshold.

The convergence of the method as the degree of approximation d increases is depicted in Fig. 3.20, for both the quadratic potential $V(x) = x^2/2$ and the bistable potential $V(x) = x^4/4 - x^2/2$. For the quadratic potential, we used a cubic index set and the parameters $\varepsilon = 1/5$, $\beta = 1$, $\sigma_x^2 = \sigma_p^2 = \sigma_q^2 = 0.2$, $\mathrm{e}^{-U(x,p,q)/2} = 1$. The error was estimated by comparison with the exact solution, which in this case is Gaussian and can be calculated explicitly. For the bistable potential, we used the index set $\mathcal{I}_d = \{\alpha \in \mathbf{N}^3 : \max(\alpha_x, 2\alpha_p, 2\alpha_q) \leq d\}$ and the parameters $\varepsilon = 1/5$, $\beta = 5$, $\sigma_x^2 = 0.05$, $\sigma_p^2 = \sigma_q^2 = 0.2$, $\mathrm{e}^{-U(x,p,q)/2} = \mathrm{e}^{-\beta V/2 - p^2/4 - q^2/4}$. The $L^1\left(\mathbf{R}^3\right)$ error, in this case, was obtained by comparison with the most precise solution available ($d = 50$). We observe from the figure that the convergence follows an exponential trend in both cases. The CPU time required to generate each of the convergence graphs in Fig. 3.20, with an Intel i7-3770 computer, was approximately one hour.



(a) $V = x^2/2$. **(b)** $V = x^4/4 - x^2/2$.

Figure 3.20: Convergence of the Galerkin discretization for the harmonic noise. The points correspond to the simulation data, and the lines are the best linear (in the *semilogy* scale) approximations. We observe that the accuracy achieved is better in the case of a quadratic potential.

Bistable noise We now consider that both the position process x and the noise process η are confined by the bistable potential, i.e. $V(\,\cdot\,) = V_\eta(\,\cdot\,) = \cdot^4/4 - \cdot^2/2$. In contrast with the examples considered before, the $\eta-$marginal of the steady-state solution cannot be exactly represented in terms of suitably rescaled Hermite functions in this case. Guided by preliminary numerical experiments exhibiting the presence of steeper directional derivatives of the solution in the x direction than in the η direction, we employed a higher degree of approximation and a smaller scaling factor in the x direction. More specifically, we used the parameters $\beta = 1$,

$\varepsilon = 0.25$, $\sigma_x^2 = 1/100$, $\sigma_\eta^2 = 1/10$, and the index set $\mathcal{I}_d = \{\alpha \in \mathbf{N}^2 : \max(\alpha_x, 2\,\alpha_\eta) \leq d\}$, with a maximum value of $d = 80$. We also set $\mathrm{e}^{-U(x,\eta)/2} = \mathrm{e}^{-\beta V(x)/2}\ \mathrm{e}^{-V_\eta(\eta)/2}$, as for the case of the bistable potential with OU noise.

In Fig. 3.21, we present the steady-state solution of equation (3.32), its x marginal and the convergence of the method. It appears clearly that, in comparison with the case of white noise, the solution decreases more steeply to 0 as $|x| \to \infty$. This is consistent with the fact that also the equilibrium density of the noise process, proportional to $\mathrm{e}^{-(\eta^4/4-\eta^2/2)}$, decreases to 0 faster than a Gaussian. From the rightmost panel in Fig. 3.21, we observe that a good accuracy, of approximately 10^{-5}, is reached from $d \approx 150$. In this case, calculating the data points of the convergence graph took approximately three hours of CPU time.

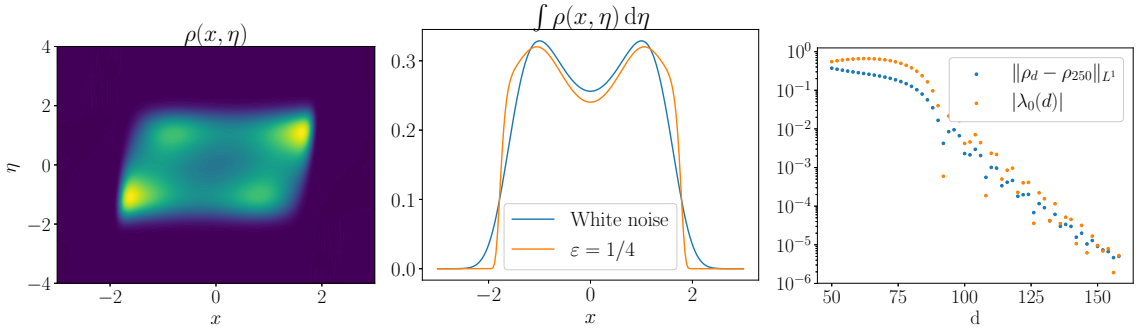
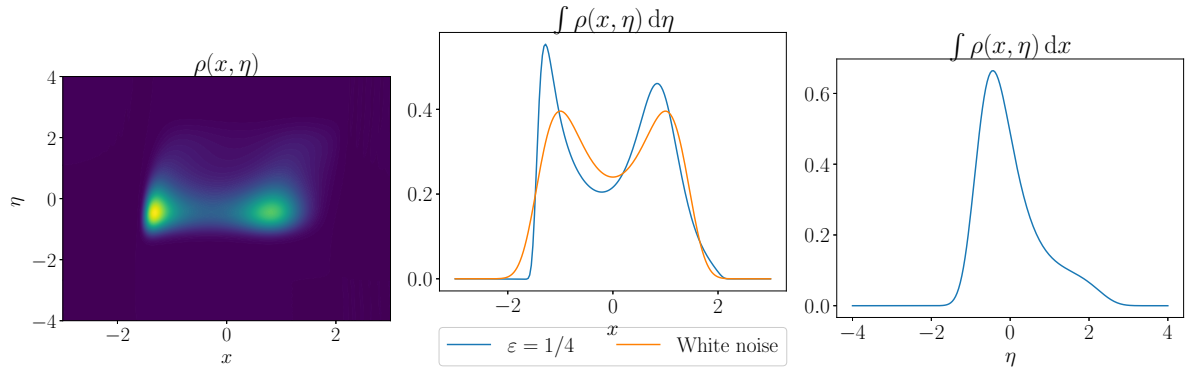


Figure 3.21: Left and middle: steady-state solution to (3.32) when $V_{\text{eff}}(x) = x^4/4 - x^2/2$, $V_\eta(\eta) = \eta^4/4 - \eta^2/2$ and its x-marginal. Right: Convergence of the method when $d \to \infty$.

Non-symmetric reversible noise An interesting feature of the noise process (**NS**), confined by the shifted tilted bistable potential, is that it breaks the symmetry of the problem, even in cases where the confining potential $V(x)$ is symmetric. The asymmetry can be observed at the level of both the steady-state solution to (3.32) and the bifurcation diagram. In Figs. 3.22a and 3.22b, we present the steady-state solution for $\varepsilon = 0.25$, a solution that looks qualitatively very different from that in the white noise limit. We observe in particular that the interval supporting most of the density is asymmetric, and that the rate at which the solution decreases to 0 is much higher at the left edge of that interval than at the right one, which is consistent with the asymptotic expansion of the x-marginal given in Eq. (3.80).

3.C.2 Convergence to the white noise limit

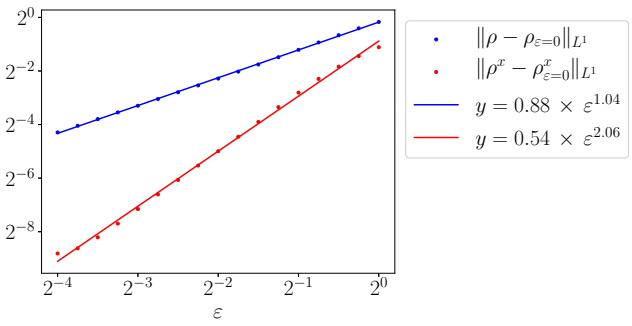
We showed formally in Section 3.B.3 that the rate of convergence to the white noise limit is as $\mathcal{O}(\varepsilon^2)$ when the noise process is given by the solution to the scalar overdamped Langevin equation in an even potential. To conclude this appendix, we verify this result numerically for $V(x) = x^2/2$ and bistable noise, $V_\eta(\eta) = \eta^4/4 - \eta^2/2$. The parameters we use in this case are $\beta = 1$, $\sigma_x^2 = \sigma_\eta^2 = 1/20$, $d_x = d_\eta = 80$, $f_x = 1$. From the numerical results presented in Fig. 3.23a, we estimate that the convergence rate of the x marginal of the steady state to the white noise limit as $\varepsilon \to 0$ is equal to 2.06, which is close to the theoretical value of 2. We note in Fig. 3.23b the presence of two maxima of the marginal density for sufficiently large ε. At the level of the



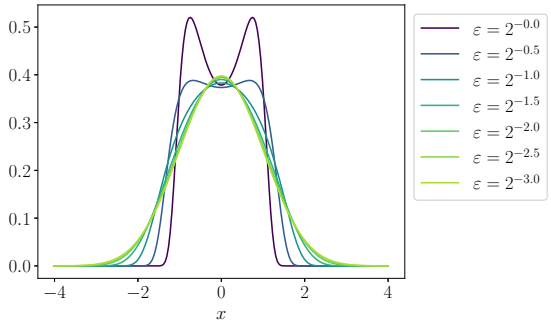
(a) Density. **(b)** Marginal in x. **(c)** Marginal in η.

Figure 3.22: Steady-state solution of Eq. (3.32) and its marginals for the nonsymmetric noise (**NS**). These figures correspond to the parameters $\beta = 2$, $\varepsilon = 0.25$, $d = 100$, $\sigma_x^2 = 1/30$, $\sigma_y^2 = 1/10$.

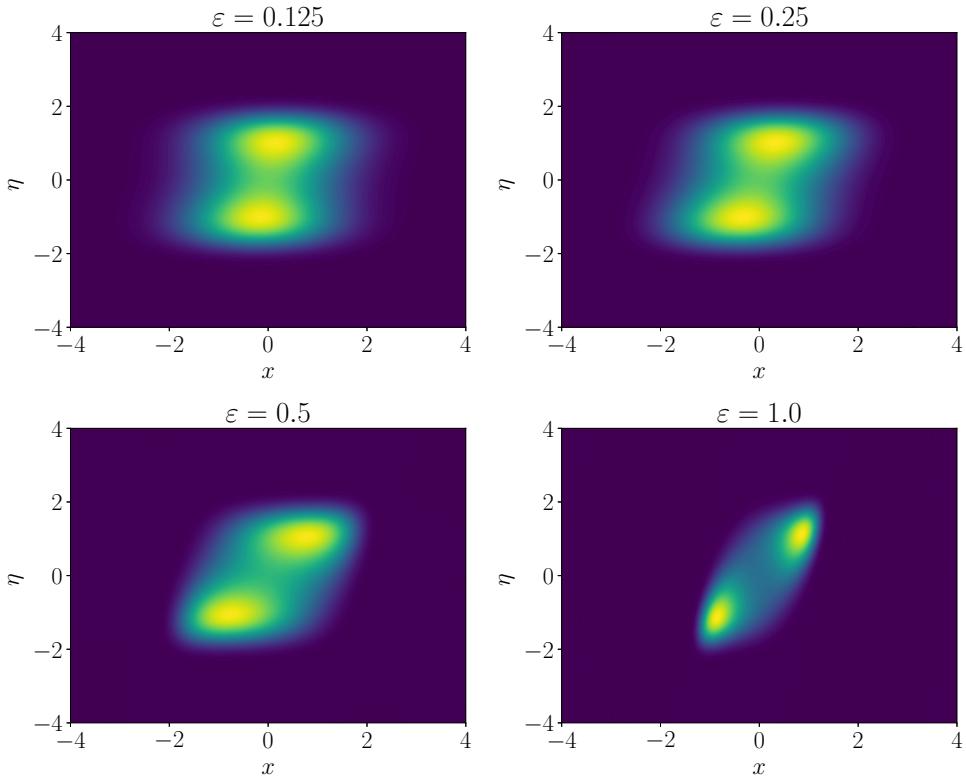
two-dimensional density $\rho(x, \eta)$, we see in Fig. 3.23c that an increase in ε results in a clearer tilt and a more localized solution.



(a) Convergence as $\varepsilon \to 0$.



(b) $\rho^x(x)$.



(c) $\rho(x, \eta)$.

Figure 3.23: Convergence of the steady-state solution of Eq. (3.32) when the driving noise is confined by the bistable potential $V_\eta(\eta) = \eta^4/4 - \eta^2/2$. Here $\rho^x(x)$ denotes the x marginal of $\rho(x, \eta)$.

Chapter 4

Generalized Langevin equation in a periodic potential

Abstract

In this chapter, we study the diffusion in a periodic potential of solutions to the *generalized Langevin equation* (GLE). We obtain asymptotic results in the small correlation time regime, as well as in the overdamped and underdamped limit. We also extend a recently developed numerical method [190] to calculate the effective diffusion for a wide range of effective frictions, confirming our asymptotic results. Finally, we study the convergence of solutions of the GLE to equilibrium using techniques from the theory of hypocoercivity.

4.1 Introduction . . . 154
4.2 Model and derivation of the effective diffusion . . . 156
4.3 Asymptotic analysis . . . 160
4.3.1 The overdamped limit . . . 163
4.3.2 The underdamped limit . . . 166
4.3.3 The short memory limit . . . 168
4.4 Numerical experiments . . . 172
4.4.1 Calculation of the velocity autocorrelation function . . . 173
4.4.2 Convergence of the method . . . 175
4.4.3 Effective diffusion . . . 177
4.5 Convergence of the GLE dynamics to equilibrium . . . 178
4.6 Conclusions . . . 182
4.7 Outlook: extension to two-dimensional periodic potentials . . . 183
4.A Effective diffusion coefficient for model (4.1a) . . . 185
4.B Attempt at proving discrete hypocoercivity in $H^1(\mu)$ . . . 186

4.1 Introduction

In this chapter, we consider diffusion approximations for the generalized Langevin equation (GLE) in a periodic potential V. We consider three different models, written here in 1D for simplicity:

$$\dot{q} = -V'(q) + \sqrt{2\,\beta^{-1}}\,\dot{W}, \tag{4.1a}$$

$$\ddot{q} = -V'(q) - \gamma\,\dot{q} + \sqrt{2\,\gamma\,\beta^{-1}}\,\dot{W}, \tag{4.1b}$$

$$\ddot{q} = -V'(q) - \int_0^t \gamma(t-s)\,\dot{q}(s)\,\mathrm{d}s + F(t), \quad \text{where} \quad \langle F(t)F(s)\rangle = \beta^{-1}\,\gamma(t-s). \tag{4.1c}$$

The GLE (4.1c) is often used in molecular dynamics and nonequilibrium statistical mechanics to represent the interaction of a particle with a heat bath at equilibrium, see for example [174]. Rather than the general GLE (4.1c), we consider a hierarchy of finite-dimensional Markovian models for the noise term, and we study in particular the cases of scalar Ornstein–Ulhenbeck (OU) noise and harmonic noise.

The behavior of a Brownian particle in a periodic potential has applications to many areas of science, including electronics [220], biology [185], surface diffusion [92] and Josephson tunneling [16]. While this problem has been studied extensively for the Langevin equation (4.1b), the study of non-Markovian Brownian motion in a periodic potential has received less attention. Early results were obtained in [119] by means of numerical experiments using the matrix-continued fraction method (see, e.g., [187, Section 9.1.2]), and verified in [118] by analog simulation, but few authors have investigated the problem since then. Recently, new theoretical results based on recent developments in multiscale analysis [175, 178] were obtained [168], establishing rigorously the passage from Eq. (4.1c) to Eq. (4.1b) and proving a homogenization result for the GLE. In light of these new results, and with the increased computational power available today, there is now scope for a more in-depth study of the problem than was possible in [119, 118].

For the Langevin (4.1b) and overdamped Langevin (4.1a) equations, as well as for all finite-dimensional approximations of the GLE (4.1c), a functional central limit theorem (CLT) holds: applying the diffusive rescaling, the dynamics of the position q converges weakly (i.e. in the sense of probability measures) to a Brownian motion:

$$\varepsilon\,q(t/\varepsilon^2) \to \sqrt{2\,D}\,W(t),$$

where the effective diffusion coefficient depends on the model and its parameters. This is shown in, for example, [178] for the overdamped Langevin and Langevin dynamics, and in [168, Theorem 2.5] for finite-dimensional approximations of the GLE. In one dimension, the effective diffusion coefficient in the low temperature regime is related to the hopping motion performed by particles between the local minimal of the potential: denoting by $\langle \ell^2 \rangle$ the mean square jump length and by κ the hopping rate, the effective diffusion coefficient can be expressed as $D = \langle \ell^2 \rangle\,\kappa$. The study of the effective diffusion coefficient, via an appropriate Poisson equation, is the main focus of this work.

For finite-dimensional Markovian models of the noise in Eq. (4.1c), it was also shown rigorously

in [168, Theorem 2.6] that, with appropriate scalings, the solution of the GLE converges weakly to that of the Langevin equation (4.1b) when the amount of memory of the system goes to zero, i.e. when the autocorrelation function of the noise converges to a Dirac delta. We will show, both theoretically and through numerical experiments, that this is also true at the level of the effective diffusion coefficient (in other words, the diffusive and short memory limits commute), and we will derive the convergence rate with respect to an appropriate parameter entering as a scaling coefficient in the memory kernel.

For each of the models of the noise that we consider, we also study the influence on the effective diffusion of the friction coefficient that appears in the short memory limiting Langevin equation, a coefficient that we will also refer to as the *friction coefficient* by a slight abuse of terminology. This problem has been studied for the Langevin equation (4.1b), and it is by now well-known [178] that the following asymptotic expansion of the effective diffusion coefficient associated with (4.1b) holds in the limit $\gamma \to \infty$,

$$D_\gamma = \frac{1}{\gamma} \frac{4\pi^2}{\beta \mathcal{Z}_q \hat{\mathcal{Z}}_q} - \frac{1}{\gamma^3} \frac{4\pi^2 \beta \mathcal{Z}_1}{\mathcal{Z}_q \hat{\mathcal{Z}}_q^2} + \mathcal{O}\left(\frac{1}{\gamma^5}\right), \tag{4.2}$$

where the term multiplying γ^{-1} in the first term corresponds to the effective diffusion coefficient for the overdamped dynamics (4.1a), and with

$$\mathcal{Z}_q = \int_{-\pi}^{\pi} \mathrm{e}^{-\beta V(q)} \, \mathrm{d}q, \quad \hat{\mathcal{Z}}_q = \int_{-\pi}^{\pi} \mathrm{e}^{\beta V(q)} \, \mathrm{d}q, \quad \mathcal{Z}_1 = \int_{-\pi}^{\pi} \left|V'(q)\right|^2 \mathrm{e}^{\beta V(q)} \, \mathrm{d}q.$$

The underdamped limit $\gamma \to 0$ of the Langevin equation is also well-understood, and it leads to a leading-order term scaling as γ^{-1}. We will carry out, for the case of the GLE, an analysis similar to that presented in [178] for the Langevin equation and we will show, in particular, that the leading order term in the expression of the effective diffusion in the limit where the friction increases to infinity is the same as in Eq. (4.2). In contrast with the Langevin equation, however, the effective diffusion coefficient for finite γ is no longer bounded from above by $\overline{D}/\gamma$, where $\overline{D}/\gamma$ is the leading order term in Eq. (4.2).

To validate the results obtained by asymptotic analysis, we will solve numerically the Poisson equation involved in the calculation of the effective diffusion coefficient, by adapting the Fourier/Hermite spectral method introduced and studied in [190]. Although no rigorous convergence result exists for this method for general hypocoercive linear PDEs, our numerical experiments suggest that spectral convergence occurs as the number of Fourier and Hermite basis functions increases. In the early studies [118, 119], very few 4D simulations could be achieved given the hardware limitations at the time, and only about 15 basis functions could be used for 3D simulations; in contrast, with today's hardware and wide range of mathematical software libraries, we were able to calculate the effective diffusion accurately in both 3D (with 100 basis functions) and 4D (with 40 basis functions), over the ranges $\gamma \in [1/100, 1000]$ and $\gamma \in [1/10, 1000]$ respectively.

In addition to studying the effective diffusion in a periodic potential, we investigate the rate of convergence of the GLE to equilibrium, in the particular cases of scalar OU and harmonic noises. While it is possible to explicitly calculate the spectrum of the generator for a quadratic confining

potential, such calculations cannot in general be carried out explicitly for more complicated confining potentials or in the periodic setting. In these cases it is possible to prove convergence to equilibrium using hypocoercivity techniques from [220], as was shown theoretically in [168]. In this work, we complement the findings of [168] by proposing a numerical methodology for obtaining quantitative estimates for the convergence rate, and we use this methodology to study the influence of the friction coefficient on the convergence to equilibrium.

The rest of the chapter is organized as follows. In Section 4.2, we present the finite-dimensional Markovian models of the GLE that we employ and we recall the derivation of the effective diffusion coefficient. In Section 4.3, we study the underdamped and overdamped limit, and we also carry out a multiscale analysis with respect to the amount of memory of the system. In Section 4.4, we adapt the numerical method introduced in [190] to the GLE and present numerical results corroborating the results obtained in Section 4.3. In Section 4.5, we obtain numerically quantitative lower bounds for the rate of convergence to equilibrium, for a range of values of the friction coefficient. Section 4.6 is reserved for conclusions and perspectives for future work.

4.2 Model and derivation of the effective diffusion

When the memory kernel is of the form

$$\gamma(t) = \left\langle \mathrm{e}^{-\mathbf{A}|t|}\,\boldsymbol{\lambda}, \boldsymbol{\lambda} \right\rangle, \tag{4.3}$$

for a (possibly nonsymmetric) matrix $\mathbf{A} \in \mathbf{R}^{n\times n}$ with eigenvalues strictly in the left half-plane and a vector $\boldsymbol{\lambda} \in \mathbf{R}^n$, it is well-known [174, Proposition 8.1] that the GLE is quasi-Markovian, in that it is equivalent to the stochastic differential equation (SDE)

$$\mathrm{d}q = p\,\mathrm{d}t, \tag{4.4a}$$

$$\mathrm{d}p = -V'(q)\,\mathrm{d}t + \langle \boldsymbol{\lambda}, \mathbf{z}\rangle\,\mathrm{d}t, \tag{4.4b}$$

$$\mathrm{d}\mathbf{z} = -p\,\boldsymbol{\lambda}\,\mathrm{d}t - \mathbf{A}\,\mathbf{z}\,\mathrm{d}t + \boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t, \quad \mathbf{z}(0) \sim \mathcal{N}(0, \beta^{-1}\mathbf{I}), \tag{4.4c}$$

where $\Sigma \in \mathbf{R}^{n\times n}$ is related to $\mathbf{A}$ by the fluctuation/dissipation theorem:

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = \beta^{-1}\,(\mathbf{A} + \mathbf{A}^T).$$

The Fourier and Laplace transforms of $\gamma(t)$ in Eq. (4.3) are rational functions, as can be seen by passing to the normal Jordan form of $\mathbf{A}$. Conversely, certain rational functions are the Fourier or Laplace transforms of functions of the form (4.3) (but not all since, by Bochner's theorem, the Fourier transform of the correlation function $\gamma(t)/\gamma(0)$ is necessarily a probability measure). In some particular cases, standard procedures exist to calculate $\mathbf{A}$ and $\boldsymbol{\lambda}$ from the Laplace or Fourier transform of $\gamma(t)$. For instance, if the Fourier transform of $\gamma(\cdot)$ reads

$$\hat{\gamma}(\omega) = \frac{1}{|p(\omega)|^2},$$

then [186, Proposition 2.3] states that $\gamma(t)$ is the autocorrelation function of the stochastic process $x(t)$ solving

$$\left(p\left(i\frac{\mathrm{d}}{\mathrm{d}t}\right)\,x(t)\right)\,\mathrm{d}t = \mathrm{d}W_t,$$

an autocorrelation function that can be shown to be of the form (4.3). When the Laplace transform can be expressed as a finite continued fraction, on the other hand, the parameters $\mathbf{A}$ and $\boldsymbol{\lambda}$ can be obtained from the following standard result.

Proposition 4.1. *If*

$$\mathbf{A} = \begin{pmatrix} \theta_1 & \varepsilon_1 & & & & \\ -\varepsilon_1 & \theta_2 & \varepsilon_2 & & & \\ & -\varepsilon_1 & \theta_3 & \varepsilon_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\varepsilon_{n-2} & \theta_{n-1} & \varepsilon_{n-1} \\ & & & & -\varepsilon_{n-1} & \theta_n, \end{pmatrix} \quad \textit{and} \quad \boldsymbol{\lambda} = \begin{pmatrix} \lambda \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

then the Laplace transform of $\gamma(t) = \left\langle \mathrm{e}^{-\mathbf{A}t}\,\boldsymbol{\lambda}, \boldsymbol{\lambda}\right\rangle$ *is given by the (finite) continued fraction*

$$\tilde{\gamma}(s) = \cfrac{\lambda^2}{\theta_1 + s + \cfrac{\varepsilon_1^2}{\theta_2+s+\cfrac{\varepsilon_2^2}{\theta_3+s+\cfrac{\varepsilon_3^2}{\ddots}}}}. \tag{4.5}$$

Proof. By definition of the Laplace transform,

$$\tilde{\gamma}(s) = \int_0^\infty \left\langle \mathrm{e}^{-\mathbf{A}t}\,\boldsymbol{\lambda}, \boldsymbol{\lambda}\right\rangle \mathrm{e}^{-st}\,\mathrm{d}s = \int_0^\infty \left\langle \mathrm{e}^{-(\mathbf{A}+s\mathbf{I})t}\,\boldsymbol{\lambda}, \boldsymbol{\lambda}\right\rangle \mathrm{d}s = \left\langle (\mathbf{A}+s\mathbf{I})^{-1}\boldsymbol{\lambda}, \boldsymbol{\lambda}\right\rangle,$$

where $\mathbf{I} \in \mathbf{R}^{n\times n}$ is the identity matrix. Introducing the vector $\mathbf{u} := (\mathbf{A}+s\mathbf{I})^{-1}\boldsymbol{\lambda}$ with components u_i and the ratios $r_i := u_i/u_{i+1}$ for $i = 1, \ldots, n-1$, we can rewrite the equation $(\mathbf{A}+s\mathbf{I})\mathbf{u} = \boldsymbol{\lambda}$ in terms of r_i:

$$\begin{aligned} &(\theta_1 + s) + \varepsilon_1\, r_1 = u_0^{-1}, \\ &-\varepsilon_{i-1}\, r_{i-1}^{-1} + (\theta_i + s) + \varepsilon_i\, r_i = 0, \qquad i = 1, \ldots, n-1, \\ &-\varepsilon_{n-1}\, r_{n-1}^{-1} + (\theta_n + s) = 0. \end{aligned}$$

Therefore

$$u_0 = \frac{1}{\theta_1 + s + \varepsilon_1\, r_1}, \qquad r_i = \frac{\varepsilon_i}{\theta_{i+1} + s + \varepsilon_{i+1}\, r_{i+1}}, \qquad r_{n-1} = \frac{\varepsilon_{n-1}}{\theta_n + s},$$

which can be combined to obtain Eq. (4.5). □

When the Laplace transform of the autocorrelation function of the noise is an infinite continued fraction, it is not possible to find finite-dimensional parameters $\mathbf{A}$ and $\boldsymbol{\lambda}$ such that Eq. (4.3) is satisfied but, using Proposition 4.1, a hierarchy of finite-dimensional Markovian approximations

can be constructed by retaining an increasing number of terms in the fraction, in order to approach the full, possibly non-Markovian dynamics; see for instance [159, 193, 135].

In this study we limit ourselves to one and two-dimensional approximations of the noise, because the computational cost of running deterministic numerical experiments becomes prohibitive for larger values of the dimension n. Higher-dimensional models, as well as the convergence of particular approximations of the noise when $n \to \infty$, will be studied from a theoretical viewpoint and using particle-based simulations in future work. We consider the two following models for the process $\mathbf{z}$:

GL1 The noise is modeled by a scalar OU process ($n = 1$), so $\boldsymbol{\lambda}$, $\mathbf{A}$, $\boldsymbol{\Sigma}$ and $\mathbf{z}$ are scalar quantities. We consider the parametrization $\boldsymbol{\lambda} = \sqrt{\gamma}/\nu$, $\mathbf{A} = 1/\nu^2$, for two positive parameters ν and γ, so Eq. (4.4c) becomes

$$\mathrm{d}z = (-\sqrt{\gamma}\, p/\nu - z/\nu^2)\,\mathrm{d}t + \sqrt{2\,\beta^{-1}/\nu^2}\,\mathrm{d}W_t, \quad z(0) \sim \mathcal{N}(0, \beta^{-1}).$$

The associated memory kernel is $\gamma(t) = \frac{\gamma}{\nu^2}\, \mathrm{e}^{-\frac{|t|}{\nu^2}}$.

GL2 The noise is modeled by a generalized version of harmonic noise:

$$\boldsymbol{\lambda} = \frac{1}{\nu}\begin{pmatrix}\sqrt{\gamma}\\ 0\end{pmatrix}, \quad \mathbf{A} = \frac{1}{\nu^2}\begin{pmatrix}0 & -\alpha\\ \alpha & \alpha^2\end{pmatrix} \to \quad \boldsymbol{\Sigma} = \sqrt{\frac{2\,\beta^{-1}\,\alpha^2}{\nu^2}}\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}.$$

The associated memory kernel is given by $\frac{\gamma}{\nu^2}\, \mathrm{e}^{-\frac{2|t|}{\nu^2}}$ when $\alpha = 2$ and otherwise by

$$\gamma(t) = \frac{\gamma}{\nu^2}\, e^{-\frac{\alpha^2\,|t|}{2\,\nu^2}} \left(\frac{\alpha}{\sqrt{|4-\alpha^2|}}\, s_\alpha\left(\frac{\sqrt{|4-\alpha^2|}}{2}\,\frac{\alpha\,|t|}{\nu^2}\right) + c_\alpha\left(\frac{\sqrt{|4-\alpha^2|}}{2}\,\frac{\alpha\,|t|}{\nu^2}\right)\right), \tag{4.6}$$

where (s_α, c_α) are the functions $(\sin, \cos)$ when $\alpha < 2$ and $(\sinh, \cosh)$ when $\alpha > 2$. In particular, since $\alpha^2 - \alpha\sqrt{\alpha^2 - 4} \to 2$ as $\alpha \to \infty$, we recover $\gamma(t) = \frac{\gamma}{\nu^2}\, \mathrm{e}^{-\frac{|t|}{\nu}}$ as $\alpha \to \infty$ (the overdamped limit of the noise), which is the autocorrelation function of the noise in the model GL1.

In both models, the parameters γ and ν (or, rather, ν^2) are scaling parameters, and, in model GL2, α encodes the shape of the memory kernel. The kernel is illustrated in Fig. 4.1 for the two models and different values of α.

Throughout this chapter, we will work with the one-dimensional periodic cosine potential $V(q) = \frac{1}{2}(1 - \cos q)$ and the periodic cell $[-\pi, \pi]$. The derivation of the effective diffusion coefficient for systems of the type (4.4) is well understood, see for example [175]. Relabeling q as x, re-introducing q as the periodized position $q := x - 2\,\pi\,\lfloor (x+\pi)/(2\pi)\rfloor \in [-\pi, \pi]$, which represents the position within the periodic cell, and applying the diffusive rescaling $x \mapsto x/\varepsilon$, $t \mapsto t/\varepsilon^2$, we can recast Eq. (4.4) as a fast/slow system of SDEs of the type studied in [175,

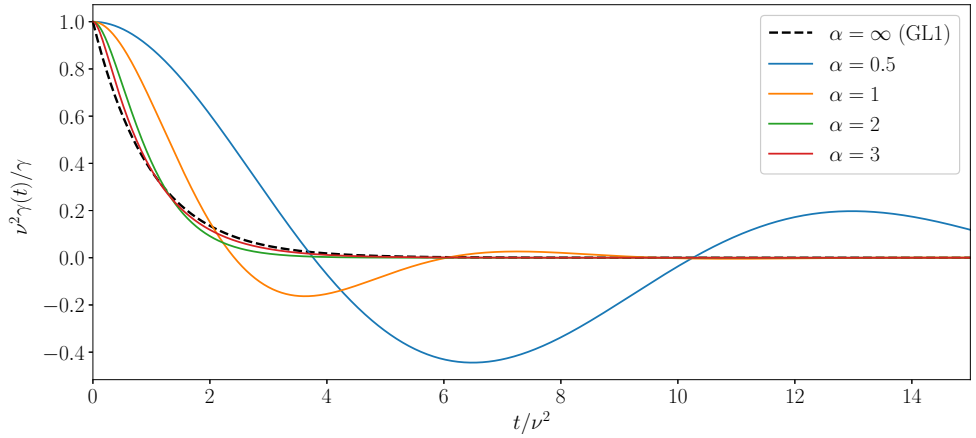


Figure 4.1: Memory kernels associated with the models considered. As $\alpha \to \infty$, the memory kernel of the model GL1 is recovered.

Chapter 11]:

$$
\begin{aligned}
\mathrm{d}x &= \frac{1}{\varepsilon}\, p\, \mathrm{d}t, \\
\mathrm{d}q &= \frac{1}{\varepsilon^2}\, p\, \mathrm{d}t, \\
\mathrm{d}p &= \frac{1}{\varepsilon^2}\left(-V'(q) + \langle \boldsymbol{\lambda}, \mathbf{z}\rangle\right) \mathrm{d}t, \\
\mathrm{d}\mathbf{z} &= -\frac{1}{\varepsilon^2}\left(p\,\boldsymbol{\lambda} + \mathbf{A}\,\mathbf{z}\right) \mathrm{d}t + \frac{1}{\varepsilon}\,\boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_t, \quad \mathbf{z}(0) \sim \mathcal{N}(0, \beta^{-1}\mathbf{I}).
\end{aligned}
$$

By expanding the solution to the backward Kolmogorov (BK) equation associated with this system as $v = v_0 + \varepsilon\, v_1 + \varepsilon^2\, v_2 + \cdots$, grouping terms multiplying equal powers of ε and using the Fredholm alternative repeatedly to obtain solvability conditions, it is possible to show formally that, in the limit $\varepsilon \to 0$, $x(t)$ satisfies the coarse-grained equation

$$
\mathrm{d}x = \sqrt{2D_n}\, \mathrm{d}W_t,
$$

where W is a standard Brownian motion in $\mathbf{R}$ and the diffusion coefficient D_n is obtained in terms of the solution to a Poisson equation:

$$
D_n = \int_{\mathbf{T}\times\mathbf{R}\times\mathbf{R}^n} \phi\, p\, \mu(\mathrm{d}q\, \mathrm{d}p\, \mathrm{d}\mathbf{z}), \tag{4.8}
$$

$$
-\mathcal{L}\,\phi = p. \tag{4.9}
$$

Here $\mathbf{T}$ is the $[-\pi, \pi]$ torus, $\mathbf{z}$ is the vector of extra processes involved in constructing the Markovian approximation of the GLE, see Eq. (4.4c), $\mathcal{L}$ is the generator of the Markov semigroup associated with the periodized dynamics, and μ is the invariant measure of the system:

$$
\mu(\mathrm{d}q\, \mathrm{d}p\, \mathrm{d}\mathbf{z}) = \frac{1}{\mathcal{Z}} \exp\left(-\beta\left(H(q,p) + |\mathbf{z}|^2\right)\right),
$$

with $\mathcal{Z}$ a normalization constant and $H(q,p) = V(q) + p^2/2$. Equation (4.9) is equipped with periodic boundary conditions in $[-\pi, \pi]$, and the condition that it is square integrable with

respect to the Gibbs measure, i.e. $\phi \in L^2(\mathbf{T} \times \mathbf{R} \times \mathbf{R}^n, \mu)$. The subscript n in Eq. (4.8) was included to emphasize the dependence on the finite-dimensional model of the noise. A natural question is the dependence of the diffusion coefficient on the parameters of the model. We note that Eq. (4.9) is also valid for the Langevin dynamics, provided that $\mathcal{L}$ is interpreted as the associated generator.

The generator of the dynamics (4.4) is

$$\begin{aligned}\mathcal{L} &= \left(p\,\frac{\partial}{\partial q} - \frac{\partial V}{\partial q}\,\frac{\partial}{\partial p} + \langle \boldsymbol{\lambda}, \mathbf{z} \rangle\,\frac{\partial}{\partial p} - p\,\boldsymbol{\lambda}\cdot\nabla_{\mathbf{z}} - \mathbf{A}_a\,\mathbf{z}\cdot\nabla_{\mathbf{z}} \right) \\ &\quad + \left(-\mathbf{A}_s\,\mathbf{z}\cdot\nabla_{\mathbf{z}} + \beta^{-1}\,\mathbf{A} : (\nabla_{\mathbf{z}}\nabla_{\mathbf{z}}) \right) \\ &=: \mathcal{L}_{\text{ham}} + \mathcal{L}_{\text{FD}},\end{aligned}$$

where $\mathcal{L}_{\text{ham}}$ corresponds to the Hamiltonian part of the dynamics and $\mathcal{L}_{FD}$ is the symmetric part of the generator, which is related to the fluctuation and dissipation terms in Eq. (4.4c). Here $\mathbf{A}_s = (\mathbf{A} + \mathbf{A}^T)/2$ and $\mathbf{A}_a = \mathbf{A} - \mathbf{A}_s$ are the symmetric and antisymmetric parts of $\mathbf{A}$, respectively, $\nabla_{\mathbf{z}}\nabla_{\mathbf{z}}$ is the Hessian operator and ":" is the Frobenius inner product. By using the decomposition of $\mathcal{L}$ into symmetric and antisymmetric parts, and noticing that $\mathcal{L}_{FD} = \beta^{-1}\,\mathrm{e}^{\beta|\mathbf{z}|^2}\,\nabla_{\mathbf{z}}\cdot(\mathrm{e}^{-\beta|\mathbf{z}|^2}\,\mathbf{A}_s\,\nabla_{\mathbf{z}})$ we obtain a simpler expression for D_n:

$$D_n = \beta^{-1} \int_{\mathbf{T}\times\mathbf{R}\times\mathbf{R}^n} \mathbf{A}_s : (\nabla_{\mathbf{z}}\varphi_n \otimes \nabla_{\mathbf{z}}\varphi_n)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}\mathbf{z})\,\mathrm{d}t \tag{4.10}$$

For the overdamped Langevin equation (4.1a), the above approach does not apply but it is possible to show, using homogenization theory for parabolic PDEs, see [175, Chapter 13], that the rescaled position also performs a diffusive motion, with a diffusion coefficient equal to

$$\overline{D} = \int_{\mathbf{T}} \beta^{-1}\,(1+\chi')^2\,\nu(\mathrm{d}q), \tag{4.11}$$

where $\nu(\mathrm{d}q) = \frac{1}{\mathcal{Z}_q}\,\mathrm{e}^{-\beta V(q)}\,\mathrm{d}q$, with $\mathcal{Z}_q$ the normalization constant, and χ is a solution to the following Poisson equation:

$$-\left(-V'\frac{\mathrm{d}}{\mathrm{d}q} + \beta^{-1}\frac{\mathrm{d}^2}{\mathrm{d}q^2} \right)\chi = -V'. \tag{4.12}$$

The detailed proof in the general, multi-dimensional case is available in [172], and a simplified proof for the one-dimensional case is reproduced in Section 4.A for the reader's convenience. Using the same notation as in the introduction, Eq. (4.12) leads to $\chi' = \frac{2\pi}{\hat{\mathcal{Z}}_q}\,\mathrm{e}^{\beta V} - 1$, which by substitution in Eq. (4.11) leads to the leading order term in Eq. (4.2) (up to the factor $\frac{1}{\gamma}$).

4.3 Asymptotic analysis

In this section, we study the overdamped and underdamped limits, as well the short memory limit, for the GLE. We start by reviewing a few fundamental results on the overdamped ($\gamma \to \infty$) and underdamped ($\gamma \to 0$) limits for the Langevin equation; see for example [174, Chapter 6] for

a more detailed overview.

In the overdamped limit, it is well-known that, if q denotes the solution to the Langevin equation, then $q_\gamma(t) = q(\gamma t)$ converges to the solution to (4.1a). In addition the diffusive and overdamped limits commute, implying that $\gamma\, D_\gamma \to \overline{D}$ as $\gamma \to \infty$, where D_γ denotes the effective diffusion coefficient of the Langevin dynamics with friction coefficient γ and, as before, $\overline{D}$ is the effective diffusion coefficient from Eq. (4.11).

In the underdamped regime, on the other hand, the Hamiltonian of the rescaled process $q_\gamma(t) = q(t/\gamma)$ becomes the slow variable and it converges weakly to a diffusion process on a graph, whose generator can be calculated explicitly in simple cases; see for example [103]. The phase portrait of the Hamiltonian dynamics and the corresponding graph are illustrated in Fig. 4.2 for the periodic potential under consideration. It was proved in [178] that, for any periodic potential with only one local minimum within the periodic cell, the leading order term (of order γ^{-1}) in the asymptotic expansion of the solution to the Poisson equation (4.9) in the limit $\gamma \to 0$, which we denote by $\phi_0(V(q) + p^2/2) =: \phi_0(E)$, solves the following differential equation:

$$\left(\beta^{-1}\, T(E) - S(E)\right)\, \phi_0'(E) + \beta^{-1}\, S(E)\, \phi_0''(E) = \begin{cases} -\,2\pi, & \text{for } p > 0, E > E_0, \qquad (4.13a) \\ \;\;2\pi, & \text{for } p < 0, E > E_0, \qquad (4.13b) \\ \;\;0, & \text{for } E_{\min} < E < E_0, \qquad (4.13c) \end{cases}$$

where

$$S(E) := \int_{\mathbf{T}} p(q,E)\, \mathrm{d}q, \qquad T(E) := \int_{\mathbf{T}} \frac{1}{p(q,E)}\, \mathrm{d}q. \tag{4.14}$$

Here $p(q,E) = \sqrt{2(E - V(q))}$, E_0 is the critical energy, beyond which deterministic orbits are no longer closed, and $E_{\min}$ is the minimal value that the Hamiltonian can take. To close the problem, Eq. (4.13) must be supplemented by the condition that the solution is square integrable with respect to the invariant measure of the dynamics and by an appropriate continuity condition at the critical energy E_0. It can be shown, see [174, Section 6.6], that this so-called *gluing* condition reads $2(\phi_0^c)' = (\phi_0^a)' + (\phi_0^b)'$, where ϕ_0^a, ϕ_0^b and ϕ_0^c denote the solutions of (4.13a), (4.13b) and (4.13c), respectively.

Starting from Eq. (4.13), a straightforward calculation leads to an expression of the effective diffusion in the underdamped regime:

$$D^* = 8\pi^2 Z_\beta^{-1} \beta^{-1} \int_{E_0}^{\infty} \frac{\mathrm{e}^{-\beta z}}{S(z)}\, \mathrm{d}z, \qquad \text{where} \qquad Z_\beta = \sqrt{\frac{2\pi}{\beta}} \int_{-\pi}^{\pi} \mathrm{e}^{-\beta V(q)}\, \mathrm{d}q. \tag{4.15}$$

For the cosine potential considered in this chapter, $E_{\min} = 0$ and $E_0 = 1$, and explicit formulas for Z_β and $S(z)$, obtained in [178], can be presented. We include them here for completeness:

$$Z_\beta = 2\pi \sqrt{\frac{\pi}{\beta}} \left(\frac{1}{\pi} \int_0^{\pi} \mathrm{e}^{\beta \cos(q)/2 - \beta/2}\, \mathrm{d}q \right) = \frac{(2\pi)^{3/2}}{(\beta)^{1/2}}\, \mathrm{e}^{-\beta/2}\, I_0(\beta/2),$$

$$S(z) = 4\sqrt{2z} \int_0^{\pi/2} \sqrt{1 - \frac{1}{z}\sin(q)^2}\, \mathrm{d}q = 2^{5/2} \sqrt{z}\, E(1/z),$$

where I_0 is the modified Bessel function of the first kind of order 0 and E is the complete elliptic integral of the second kind. After substituting these expressions in Eq. (4.15), the effective diffusion can be calculated by numerical quadrature.

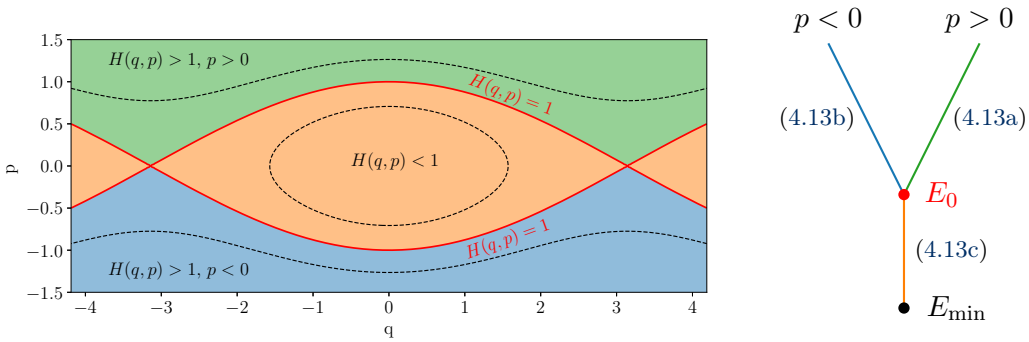


Figure 4.2: *Left*: Some contour lines of the Hamiltonian and separatrix (in red). *Right*: Associated graph, over which the Hamiltonian performs a diffusion process in the underdamped limit.

In addition to the overdamped and underdamped limits, it can be shown, see for example [103], that the effective diffusion coefficient, multiplied by γ, can be bounded from below and from above by its limiting values in the underdamped and overdamped limits, respectively:

$$D^* \leq \gamma D_\gamma \leq \overline{D} \quad \forall \gamma \in (0, \infty). \tag{4.16}$$

Figure 4.3 illustrates the solution to the Poisson Eq. (4.9) for the Langevin dynamics with the values of $\gamma = 1/10, 1, 10$; see Section 4.4 for details on the numerical method used to obtain these results. In the case $\gamma = 1/10$, we observe that the contours of the solution correspond loosely to the contours of the Hamiltonian, $H(q,p) = V(q) + p^2/2$, and that the solution takes values significantly different from 0 only when $H(q,p) \geq 1$, with positive values when $p > 0$ and negative values when $p < 0$. This is because, in the deterministic limit, orbits corresponding to $H(q,p) < 1$ are closed, as they correspond to particles with not enough energy to cross the energy barriers. As γ increases, the contour lines of the solution align with the vertical axis, which is consistent with the fact that $\phi = \phi(q)$ in the overdamped limit.

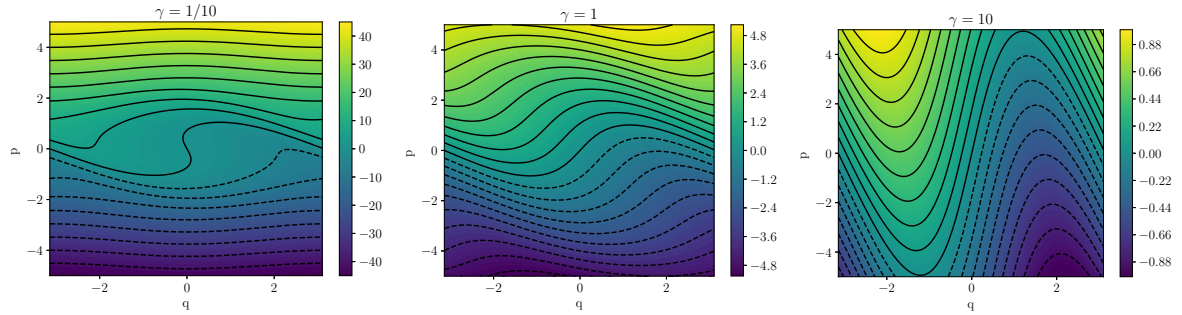


Figure 4.3: Approximate solution to the Poisson equation (4.9), obtained by solving Eq. (4.39), when $\mathcal{L}$ is the generator of Langevin dynamics.

4.3.1 The overdamped limit

The convergence of $q(\gamma t)$ to the solution of (4.1a) as $\gamma \to \infty$ also holds for the generalized Langevin dynamics; see for example [193, Theorem 8.1.2], which proves the convergence in probability. The idea of the proof in [193] is to rewrite the rescaled process Eq. (4.4b) as

$$\mathrm{d}\begin{pmatrix} p \\ \mathbf{z} \end{pmatrix} - \begin{pmatrix} 0 & \gamma^{3/2}\,\boldsymbol{\lambda}^T \\ -\gamma^{3/2}\,\boldsymbol{\lambda} & -\gamma\,\mathbf{A} \end{pmatrix} \begin{pmatrix} p \\ \mathbf{z} \end{pmatrix} \mathrm{d}t = \begin{pmatrix} -\gamma\,V'(q) \\ 0 \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \mathbf{0} \\ \gamma^{1/2}\,\boldsymbol{\Sigma} \end{pmatrix} \mathrm{d}\mathbf{W}_t,$$

and, denoting the matrix in the left-hand side as $\mathbf{M}$, to use the variation-of-constants formula to express the solution as

$$\begin{pmatrix} p \\ \mathbf{z} \end{pmatrix} = \mathrm{e}^{\mathbf{M}\,t} \begin{pmatrix} p_0 \\ \mathbf{z}_0 \end{pmatrix} + \int_0^t \mathrm{e}^{\mathbf{M}(t-s)} \begin{pmatrix} -\gamma\,V'(q)\,\mathrm{d}s \\ \gamma^{1/2}\,\boldsymbol{\Sigma}\,\mathrm{d}\mathbf{W}_s \end{pmatrix}.$$

Using this expression in the equation for q, the limiting equation when $\gamma \to \infty$ can be obtained. In this chapter we will obtain asymptotics for the effective diffusion directly, by expanding the solution to the Poisson equation (4.9) appropriately. In the following result $D_{\nu,\gamma}$ denotes the effective diffusion for the GLE with parameters ν and γ.

Result 4.2. *In the limit $\gamma \to \infty$, $\gamma\,D_{\nu,\gamma} \to \overline{D}$.*

Proof. For simplicity we assume that $\nu = \beta = 1$ and we consider only the model GL1, but the approach we take can be applied to the other Markovian approximations of the GLE. Rescaling time and introducing the small parameter $\delta = 1/\sqrt{\gamma}$, the generator of the generalized Langevin dynamics can be written as:

$$\begin{aligned} \mathcal{L} = \frac{1}{\delta}\left(z\,\frac{\partial}{\partial p} - p\,\frac{\partial}{\partial z}\right) \\ + \left(p\,\frac{\partial}{\partial q} - \frac{\partial V}{\partial q}\,\frac{\partial}{\partial p} - z\frac{\partial}{\partial z} + \frac{\partial^2}{\partial z^2}\right) =: \frac{1}{\delta}\,\mathcal{L}_0 + \mathcal{L}_1, \end{aligned}$$

Expanding the solution to Eq. (4.9) as $\phi_0 + \delta\,\phi_1 + \delta^2\,\phi_2 + \cdots$ and gathering terms that multiply the same power of δ, we obtain equations for the terms in the series:

$$\mathcal{O}(\delta^{-1}) \qquad \mathcal{L}_0\phi_0 = 0, \tag{4.17a}$$

$$\mathcal{O}(1) \qquad \mathcal{L}_0\phi_1 + \mathcal{L}_1\phi_0 = -p, \tag{4.17b}$$

$$\mathcal{O}(\delta^i) \qquad \mathcal{L}_0\phi_{i+1} + \mathcal{L}_1\phi_i = 0, \qquad i = 1, 2, \ldots \tag{4.17c}$$

The kernel of $\mathcal{L}_0$ consists of functions depending only on q and $p^2 + z^2$, so we deduce from the first equation that $\phi_0 = \phi_0(q, p^2 + z^2)$. To obtain the solvability condition for the second equation, let us introduce the $L^2\left(\mathbf{T} \times \mathbf{R}^2, \mu\right)$ (formal) adjoint operators $\partial_p^* := p - \partial_p$ and $\partial_z^* := z - \partial_z$, and let us observe that $\mathcal{L}_0 = \partial_z^*\partial_p - \partial_p^*\partial_z$ is formally antisymmetric on $L^2\left(\mathbf{T} \times \mathbf{R}^2, \mu\right)$:

$$\int_{\mathbf{T}\times\mathbf{R}^2} \mathcal{L}_0 u_1\,u_2\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) = -\int_{\mathbf{T}\times\mathbf{R}^2} u_1\,\mathcal{L}_0 u_2\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z),$$

for all sufficiently regular functions u_1 and u_2 on $\mathbf{T}\times\mathbf{R}^2$. Multiplying Eq. (4.17b) by $u_0(q,p^2+z^2)$, and integrating with respect to μ, we obtain

$$\begin{aligned}
0 &= \int_{\mathbf{T}\times\mathbf{R}^2} (\mathcal{L}_0\phi_1 + \mathcal{L}_1\phi_0 + p)\ \phi_0(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) \\
&= \int_{\mathbf{T}\times\mathbf{R}^2} -\phi_1\,\mathcal{L}_0\big(\phi_0(q,p^2+z^2)\big) + (\mathcal{L}_1\phi_0 + p)\ \phi_0(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) \\
&= 0 + \int_{\mathbf{T}\times\mathbf{R}^2} (\mathcal{L}_1\phi_0 + p)\ \phi_0(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z),
\end{aligned} \tag{4.18}$$

because, by the product rule:

$$\begin{aligned}
\mathcal{L}_0\big(\phi_0(q,p^2+z^2)\big) &= \left(z\,\frac{\partial}{\partial p} - p\,\frac{\partial}{\partial z}\right)\ \phi_0(q,p^2+z^2) \\
&= 2\,z\,p\,\frac{\partial\phi_0}{\partial R}(q,p^2+z^2) - 2\,p\,z\,\frac{\partial\phi_0}{\partial R}(q,p^2+z^2) = 0,
\end{aligned}$$

where ∂_R denotes the derivative with respect to the second argument of ϕ_0. It is also clear that

$$\int_{\mathbf{T}\times\mathbf{R}^2} p\,\phi_0(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) = 0,$$

because the integrand is an odd function of p. Going back to Eq. (4.18), we therefore obtain, observing that $\mathcal{L}_1 = \partial_p^*\partial_q - \partial_q^*\partial_p - \partial_z^*\partial_z$, where $\partial_q^* := V'(q) - \partial_q$:

$$\begin{aligned}
0 &= \int_{\mathbf{T}\times\mathbf{R}^2} \left(\partial_p^*\partial_q - \partial_q^*\partial_p - \partial_z^*\partial_z\right)\phi_0(q,p^2+z^2)\,\phi_0(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) \\
&= -\int_{\mathbf{T}\times\mathbf{R}^2} \left|\frac{\partial}{\partial z}\big(\phi_0(q,p^2+z^2)\big)\right|^2 \mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) \\
&= -\int_{\mathbf{T}\times\mathbf{R}^2} \left|2z\,\frac{\partial\phi_0}{\partial R}(q,p^2+z^2)\right|^2 \mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z),
\end{aligned}$$

where we used the antisymmetry of $\partial_p^*\partial_q - \partial_q^*\partial_p$ and the symmetry of $\partial_z^*\partial_z$ in $L^2\left(\mathbf{T}\times\mathbf{R}^2,\mu\right)$. In order for Eq. (4.17b) to be solvable, it is therefore necessary that $\phi_0 = \phi_0(q)$. Solving Eq. (4.17b), we obtain a general expression for ϕ_1,

$$\phi_1(q,p,z) = z(\phi_0'(q)+1) + \psi_1(q,p^2+z^2),$$

where a new function $\psi_1(\cdot,\cdot)$ was introduced because, at this point, ϕ_1 is defined only up to a function in the kernel of $\mathcal{L}_0$. To obtain the solvability condition of Eq. (4.17c) with $i=1$, we multiply this equation by $\psi_1(q,p^2+z^2)$ and we integrate with respect to μ, which, after using

the antisymmetry of $\mathcal{L}_0$ again, leads to the following:

$$\begin{aligned}
0 &= \int_{\mathbf{T}\times\mathbf{R}^2} \mathcal{L}_1\big(z(\phi_0'(q)+1)+\psi_1(q,p^2+z^2)\big)\,\psi_1(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z)\\
&= \int_{\mathbf{T}\times\mathbf{R}^2} \big(p\,z\,\phi_0''(q) - z\,(\phi_0'(q)+1) + \mathcal{L}_0\big(\psi_1(q,p^2+z^2)\big)\big)\;\psi_1(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z)\\
&= -\int_{\mathbf{T}\times\mathbf{R}^2} \left|2z\frac{\partial\psi_1}{\partial R}(q,p^2+z^2)\right|^2\;\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z),
\end{aligned}$$

so in fact $\psi_1=\psi_1(q)$. Here, for the last equation, we used the fact that the terms depending on ϕ_0, being the integrals of odd functions of z, are zero. Solving for ϕ_2,

$$\phi_2(q,p,z) = \frac{z^2}{2}\,\phi_0''(q) + p\,(\phi_0'(q)+1) + z\,\psi_1'(q) + \psi_2(q,p^2+z^2).$$

The solvability condition for the next order will give us an equation for the leading-order term of the asymptotic expansion, ϕ_0. Multiplying Eq. (4.17c) (with $i=2$) by $f(q,p^2+z^2)$, where f is any smooth function on $\mathbf{T}\times\mathbf{R}$ that grows sufficiently slowly in its second argument, integrating with respect to μ, using the antisymmetry of $\mathcal{L}_0$, and employing symmetry arguments of the same type as above, we obtain

$$\begin{aligned}
0 &= \int_{\mathbf{T}\times\mathbf{R}^2} \mathcal{L}_1\left(\frac{z^2}{2}\,\phi_0''(q) + p\,(\phi_0'(q)+1) + z\,\psi_1'(q) + \psi_2(q,p^2+z^2)\right)\,f(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z)\\
&= \int_{\mathbf{T}\times\mathbf{R}^2} \big((1+p^2-z^2)\phi_0''(q) - V'(q)\,(\phi_0'(q)+1)\big)\big)\;f(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z)\\
&\quad - \int_{\mathbf{T}\times\mathbf{R}^2} \mathcal{L}_1\big(\psi_2(q,p^2+z^2)\big)\,f(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z). \qquad (4.19)
\end{aligned}$$

Since $f(q,p^2+z^2)$ is invariant under rotations in the p–z plane, it is clear that

$$\int_{\mathbf{T}\times\mathbf{R}^2} (p^2-z^2)\,\phi_0''(q)\,f(q,p^2+z^2)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) = 0.$$

Taking $f(q,p^2+z^2)=f(q)$, we notice that the second integral in Eq. (4.19) cancels out:

$$\begin{aligned}
&\int_{\mathbf{T}\times\mathbf{R}^2} \mathcal{L}_1\big(\psi_2(q,p^2+z^2)\big)\,f(q)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z)\\
&\qquad = \int_{\mathbf{T}\times\mathbf{R}^2} \psi_2(q,p^2+z^2)\;\big(-\partial_p^*\partial_q + \partial_q^*\partial_p - \partial_z^*\partial_z\big)\,f(q)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z)\\
&\qquad = \int_{\mathbf{T}\times\mathbf{R}^2} \psi_2(q,p^2+z^2)\;\big(-p\,f'(q)\big)\;\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) = 0.
\end{aligned}$$

Therefore, Eq. (4.19) reduces to

$$\int_{\mathbf{T}\times\mathbf{R}^2} \big(\phi_0''(q) - V'(q)\,\phi_0'(q) - V'(q)\big)\;f(q)\,\mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) = 0. \qquad (4.20)$$

For this equation to be satisfied for any smooth periodic function f, it is necessary that

$$\mathcal{L}_q\phi_0 := \left(-V'\frac{\mathrm{d}}{\mathrm{d}q} + \frac{\mathrm{d}^2}{\mathrm{d}q^2}\right)\phi_0 = V',$$

which is exactly (4.12) (with $\beta = 1$), so the conclusion then follows by substitution in the expression (4.10) of the effective diffusion coefficient. We observe, going back to Eq. (4.19) and testing with $f(q, p^2 + z^2) = \psi_2(q, p^2 + z^2)$, that in fact $\psi_2(q, p^2 + z^2) = \psi_2(q)$. □

Remark 4.1. This result is also valid for the model GL2, in which case solving Eqs. (4.17a) to (4.17c) leads to

$$\phi_0(q) = \phi_0(q), \qquad \phi_1(q) = (z_1 + z_2)(\phi_0'(q) + 1) + \psi_1(q),$$
$$\phi_2(q) = \left(\frac{z_1^2}{2} + z_1\, z_2\right)\phi_0''(q) + p\left(\phi_0'(q) + 1\right) + \frac{z_2^2}{2}\,\phi_0''(q) + (z_1 + z_2)\psi_1'(q) + \psi_2(q),$$

after which the solvability condition for Eq. (4.17c) with $i = 2$ leads to (4.12) as above. ⊘

4.3.2 The underdamped limit

For simplicity, we assume again that $\beta = 1$ and we focus only on the model GL1. It is reasonable to assume that the leading order term in the solution to Poisson equation (4.9) is of order γ^{-1}, as this is the case for the Langevin dynamics, so we use the expansion:

$$\phi = \frac{1}{\gamma}\,\phi_0 + \frac{1}{\sqrt{\gamma}}\,\phi_1 + \phi_2 + \cdots. \tag{4.21}$$

Decomposing the generator as

$$\mathcal{L}_0 + \sqrt{\gamma}\mathcal{L}_1 := \left(p\,\frac{\partial}{\partial q} - \frac{\partial V}{\partial q}\,\frac{\partial}{\partial p} + \frac{1}{\nu^2}\left(-z\frac{\partial}{\partial z} + \frac{\partial^2}{\partial z^2}\right)\right) + \frac{\sqrt{\gamma}}{\nu}\left(z\,\frac{\partial}{\partial p} - p\,\frac{\partial}{\partial z}\right),$$

and substituting (4.21) in the Poisson equation, we obtain

$$\mathcal{O}(\gamma^{-1}) \qquad \mathcal{L}_0\phi_0 = 0, \tag{4.22}$$
$$\mathcal{O}(\gamma^{-1/2}) \qquad \mathcal{L}_0\phi_1 + \mathcal{L}_1\phi_0 = 0, \tag{4.23}$$
$$\mathcal{O}(\gamma^{(i-1)/2}) \qquad \mathcal{L}_0\phi_{i+1} + \mathcal{L}_1\phi_i = -p, \quad i = 1, 2, \ldots \tag{4.24}$$

Multiplying the first equation by ϕ_0, integrating with respect to μ, and taking into account that the contribution of the antisymmetric part of $\mathcal{L}_0$ vanishes, we obtain

$$\int_{\mathbf{T}\times\mathbf{R}\times\mathbf{R}} \left(\frac{\partial\phi_0}{\partial z}\right)^2 \mu(\mathrm{d}q\,\mathrm{d}p\,\mathrm{d}z) = 0,$$

implying $\phi_0 = \phi_0(q, p)$. Substituting again in Eq. (4.22), we obtain $\phi_0 = \phi_0(V(q) + p^2/2)$ by the associated solvability condition. We will denote the Hamiltonian associated with the Langevin dynamics, seen as a variable independent of q, by $E = H(q, p) = V(q) + p^2/2$. Using the ansatz

$\phi_1 = z\,\psi_1 + \omega_1(V(q) + p^2/2)$, we obtain an equation for ψ_1 from Eq. (4.23):

$$\left(p\,\frac{\partial}{\partial q} - \frac{\partial V}{\partial q}\,\frac{\partial}{\partial p} - \frac{1}{\nu^2}\right)\,\psi_1 + \frac{1}{\nu}\,\frac{\partial}{\partial p}\left(\phi_0\left(V(q) + \frac{p^2}{2}\right)\right) = 0. \tag{4.25}$$

In this subsection we present calculations only in the region $p \geq 0$, $E > E_0$. The same reasoning can be applied for the regions $E < E_0$ and $p \leq 0$, $E > E_0$, similarly to the derivation presented in detail in [174, Section 6.6] for the Langevin dynamics. In the coordinates (q, E), Eq. (4.25) simplifies to

$$\left(p(q,E)\frac{\partial}{\partial q} - \frac{1}{\nu^2}\right)\,\psi_1(q,E) + \frac{1}{\nu}\,p(q,E)\,\phi_0'(E) = 0, \tag{4.26}$$

with $p(q,E) := \sqrt{2(E - V(q))}$ and with periodic boundary conditions in q. For fixed E, this equation can be viewed as as a linear differential equation for $\psi_1(\cdot, E)$. In order to express the solution in a concise manner, let us define

$$F(q,E) = \int_{-\pi}^{q} \frac{1}{p(x,E)}\,\mathrm{d}x, \quad G(q,E) = \int_{-\pi}^{q} \exp\left(-\frac{F(x,E)}{\nu^2}\right)\,\mathrm{d}x.$$

For the cosine potential $F(q,E) = \sqrt{2E^{-1}}\,e(q/2|E^{-1})$, where $e(\cdot|m)$ is the elliptic integral of the first kind with parameter m. Dividing Eq. (4.26) by $p(q,E) > 0$, we obtain

$$0 = \left(\frac{\partial}{\partial q} - \frac{1}{\nu^2}\,\frac{\partial F}{\partial q}\right)\,\psi_1 + \frac{1}{\nu}\,\phi_0'(E) = \frac{\partial}{\partial q}\left(\mathrm{e}^{-F/\nu^2}\,\psi_1\right) + \frac{1}{\nu}\,\mathrm{e}^{-F/\nu^2}\,\phi_0'(E), \tag{4.27}$$

so the solution is

$$\psi_1(q,E) = \mathrm{e}^{F(q,E)/\nu^2}\,\left(C(E) - \frac{1}{\nu}\,\phi_0'(E)\,G(q,E)\right) =: \phi_0'(E)\,s_\nu(q,E), \tag{4.28}$$

where $C(E)$ can be calculated from the periodic boundary condition, and where the subscript ν was used to emphasize the dependence on the memory of the system. Moving on, integrating Eq. (4.24) in z with respect the Gaussian weight $g(z) = (2\pi)^{-1/2}\,\mathrm{e}^{-z^2/2}$, we obtain the equation:

$$\left(p\,\frac{\partial}{\partial q} - \frac{\partial V}{\partial q}\,\frac{\partial}{\partial p}\right)\,\int_{\mathbf{R}} \phi_2(p,q,z)\,g(z)\,\mathrm{d}z + \frac{1}{\nu}\left(p\,\frac{\partial \psi_1}{\partial E} - p\,\psi_1\right) + p = 0.$$

It can be shown [174, Chapter 6], using the (formal) antisymmetry of the operator $p\,\partial_q - V'(q)\,\partial_p$ in $L^2\left(\mathbf{R}^2\right)$, that the solvability condition for this equation is that, for any smooth, rapidly decaying function f on $\mathbf{R}$,

$$\int_{\mathbf{T}\times\mathbf{R}} \left(\frac{1}{\nu}\left(p\,\frac{\partial \psi_1}{\partial E} - p\,\psi_1\right) + p\right)\,f(H(q,p))\,\mathrm{d}q\,\mathrm{d}p = 0.$$

Expressing this integral in the (q,E) coordinates, and taking into account that the Jacobian of the transformation is $(p(q,E))^{-1}$, we are left with

$$\int_{\mathbf{T}\times\mathbf{R}^+} \left(\frac{1}{\nu}\left(\frac{\partial \psi_1}{\partial E} - \psi_1\right) + 1\right)\,f(E)\,\mathrm{d}q\,\mathrm{d}E = 0. \tag{4.29}$$

Substituting Eq. (4.28), integrating in q, and using the fact that (4.29) has to hold for all smooth rapidly decaying f to obtain a pointwise (in E) estimate, we obtain a differential equation for ϕ_0 similar to that obtained for the Langevin dynamics, Eq. (4.13):

$$\frac{1}{\nu}\left(S_\nu'(E)\,\phi_0'(E) + S_\nu(E)\,\phi_0''(E) - S_\nu(E)\,\phi_0'(E)\right) + 2\pi = 0, \quad S_\nu(E)\int_{-\pi}^{\pi} s_\nu(q,E)\,\mathrm{d}q, \tag{4.30}$$

with an appropriate continuity condition at the critical energy. From Eq. (4.30) and Eq. (4.8), we obtain the following expression for leading order term in the expansion of the effective diffusion:

$$D_\nu^* = 8\pi^2 Z_\beta^{-1}\beta^{-1}\int_{E_0}^{\infty}\frac{\mathrm{e}^{-\beta E}}{S_\nu(E)}\,\mathrm{d}E =: \int_{E_0}^{\infty} I_\nu(E)\,\mathrm{d}E. \tag{4.31}$$

Before closing the theoretical discussion of the underdamped limit, we note that, when $\nu \ll 1$, Eq. (4.26) implies that $\psi_1(q,E) \approx \nu\, p(q,E)\,\phi_0'(E)$ to leading order. In that case we see that Eqs. (4.30) and (4.31) are exactly the same as Eqs. (4.13) and (4.15). We therefore conclude, denoting again by $D_{\nu,\gamma}$ and D_γ the effective diffusion coefficients associated with the GLE and the Langevin equation, respectively, and by D^* the leading order term in the series expansion of D_γ in the underdamped regime, that

$$\lim_{\nu\to 0}\lim_{\gamma\to 0}\gamma\, D_{\nu,\gamma} = \lim_{\gamma\to 0}\gamma\, D_\gamma = D^*.$$

When $\nu > 0$, it is possible to solve Eq. (4.26) numerically in order to obtain an approximation of $s_\nu(q,E)$, and from there to calculate S_ν and D_ν^* by numerical integration. A Fourier spectral method is well suited for this purpose, as it leads to accurate results at a low computational cost. In our numerical simulation, we truncated the integral in Eq. (4.31) at $E = 25$ and we carried out the numerical integration by using the *SciPy* function `scipy.integrate.quad` with a relative tolerance of 1.5×10^{-8}. The number of Fourier modes employed to approximate $S_\nu(E)$ for fixed E was equal to 100. The error of the spectral method for the calculation of $I_\nu(E)$ (see Eq. (4.31)), calculated by comparison with the values obtained with only 50 Fourier modes, is depicted in Fig. 4.4 for values of E close to E_0, which is where the error is largest.

Results from the numerical simulation are illustrated in Fig. 4.5. We observe that, although the effective diffusion in the limit does vary with ν, the relative variation is very small ($< 5\%$ over the interval $\nu \in [0,1]$). We also notice that $D_\nu^* \to D^*$ as $\nu \to 0$, as expected.

4.3.3 The short memory limit

In this section, we will study the behavior of the solution to the Poisson equation (4.9) and of the associated diffusion coefficient (4.8) in the short memory limit. For models GL1 and GL2 the memory of the system is already encoded by the parameter ν. For general models of the type (4.4), we apply the rescaling $\lambda \mapsto \lambda/\nu$, $\mathbf{A} \mapsto \mathbf{A}/\nu^2$. For simplicity, $\mathbf{A}$, $\boldsymbol{\lambda}$ will always denote the parameters after rescaling (independent of ν) in this subsection. We remind the reader that, in the short memory limit, (q,p) converges strongly to the solution of the Langevin

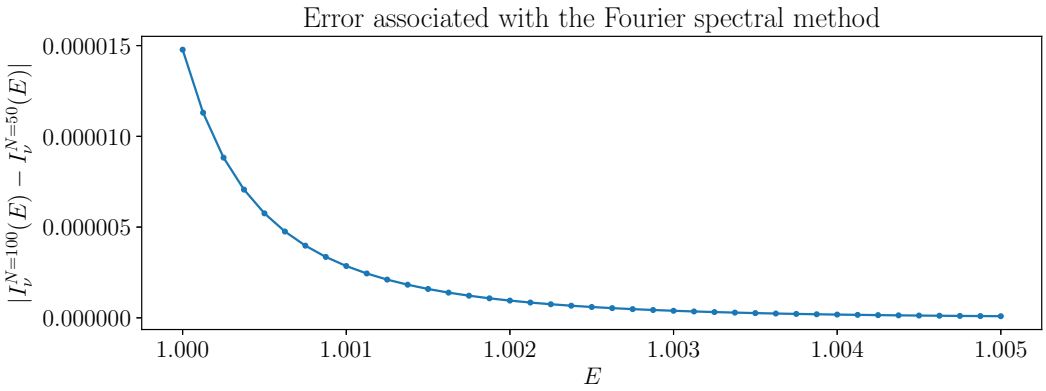


Figure 4.4: Error of the spectral method for the calculation of $I_\nu(E)$ (see Eq. (4.31)) when $\nu = 1$, estimated as the difference between the values computed with 50 and 100 Fourier modes. We observe that the error is largest close to $E = E_0 = 1$. For this value of ν, the error on $I_\nu(E)$ is uniformly bounded from above by 10^{-4}, and the difference between the approximations of the effective diffusion coefficient obtained with $N = 50$ and $N = 100$ is below 10^{-7}.

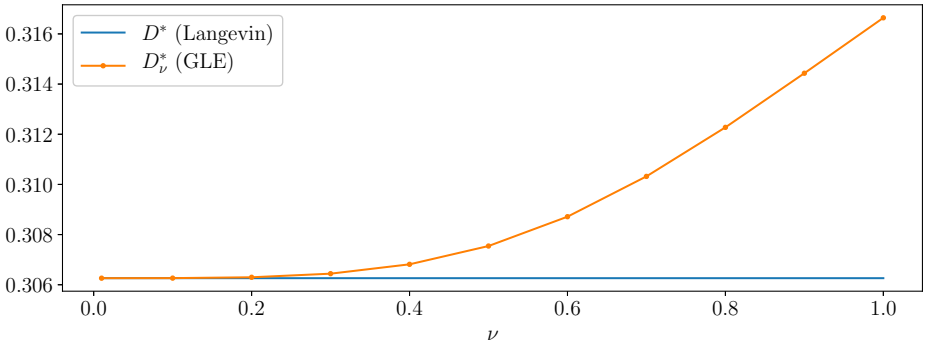


Figure 4.5: Comparison between the effective diffusion coefficients (multiplied by γ) for the Langevin and the generalized Langevin dynamics in the underdamped limit.

equation (4.1b) with friction coefficient

$$\gamma = \left\langle \boldsymbol{\lambda}, \mathbf{A}^{-1}\boldsymbol{\lambda} \right\rangle = \int_0^\infty \gamma(t)\,\mathrm{d}t, \tag{4.32}$$

see for example [167, Theorem 2.6] or [174, Result 8.4]. Expanding the solution to (4.9) as $\phi_0 + \nu\phi_1 + \nu^2\phi_2 + \cdots$ and gathering terms multiplying the same power of ν, we obtain a series of equations

$$\mathcal{L}_0\phi_0 = 0, \tag{4.33a}$$
$$\mathcal{L}_0\phi_1 + \mathcal{L}_1\phi_0 = 0, \tag{4.33b}$$
$$\mathcal{L}_0\phi_2 + \mathcal{L}_1\phi_1 + \mathcal{L}_2\phi_0 = -p, \tag{4.33c}$$
$$\mathcal{L}_0\phi_{i+2} + \mathcal{L}_1\phi_{i+1} + \mathcal{L}_2\phi_i = 0, \qquad i = 1, 2, \ldots \tag{4.33d}$$

where the operators $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2$ are defined by

$$\begin{aligned}\mathcal{L}_0 &= -\mathbf{A}\,\mathbf{z}\cdot\nabla_{\mathbf{z}} + \beta^{-1}\,\mathbf{A} : (\nabla_{\mathbf{z}}\nabla_{\mathbf{z}})\\ \mathcal{L}_1 &= \langle \boldsymbol{\lambda}, \mathbf{z}\rangle\ \frac{\partial}{\partial p} - p\,\boldsymbol{\lambda}\cdot\nabla_{\mathbf{z}}\\ \mathcal{L}_2 &= p\,\frac{\partial}{\partial q} - \frac{\partial V}{\partial q}\,\frac{\partial}{\partial p}.\end{aligned}$$

Since any constant added to ϕ_i would vanish in the calculation of the diffusion coefficient, we add for convenience the requirement that

$$\int_{\mathbf{T}\times\mathbf{R}\times\mathbf{R}^n} \phi_i \,\mathrm{d}\mu = 0, \quad i = 0, 1, \ldots \tag{4.35}$$

From (4.33a), we deduce that ϕ_0 must lie in the kernel of $\mathcal{L}_0$. When viewed as an unbounded operator on $L^p\left(\mathbf{R}^n, (2\pi)^{-n/2}\,\mathrm{e}^{-|\mathbf{z}|^2/2}\right)$ with $p \in (1, +\infty)$, the operator $\mathcal{L}_0$ is an Ornstein–Uhlenbeck operator, so its spectrum consists of all the linear combinations, with coefficients in the natural numbers (including 0), of the eigenvalues of $-\mathbf{A}$. In addition, the eigenfunctions associated with an eigenvalue λ are polynomials with degree at most $\lambda/r(\mathbf{A})$, where $r(\cdot)$ denotes the spectral radius; see [147, Section 9.3]. Consequently, the kernel of $\mathcal{L}_0$ consists of only constant (in $\mathbf{z}$) functions, and the solvability condition of Eq. (4.33a) gives $\phi_0 = \phi_0(q, p)$.

For (4.33b), we can check that the solvability condition provided by Fredholm alternative is automatically satisfied,

$$\int_{\mathbf{R}^n} \mathcal{L}_1\phi_0\ \mathrm{e}^{-\beta|\mathbf{z}|^2/2}\,\mathrm{d}z = 0.$$

The solution can be expressed as

$$\phi_1 = -\mathcal{L}_0^{-1}\,\langle \boldsymbol{\lambda}, \mathbf{z}\rangle\ \frac{\partial\phi_0}{\partial p} + \psi_1(q,p) = \left\langle \mathbf{A}^{-T}\boldsymbol{\lambda}, \mathbf{z}\right\rangle\ \frac{\partial\phi_0}{\partial p} + \psi_1(q,p).$$

For (4.33c), the solvability condition requires that

$$\begin{aligned}-p &= \left(\frac{\beta}{2\pi}\right)^{n/2} \int_{\mathbf{R}^n} (\mathcal{L}_1\phi_1 + \mathcal{L}_2\phi_0)\ \mathrm{e}^{-\beta|\mathbf{z}|^2/2}\,\mathrm{d}z,\\ &= \mathcal{L}_2\phi_0 + \left(-\left\langle \boldsymbol{\lambda}, \mathbf{A}^{-1}\boldsymbol{\lambda}\right\rangle\ p\,\frac{\partial\phi_0}{\partial p} + \beta^{-1}\ \mathrm{tr}(\mathbf{A}^{-T}\boldsymbol{\lambda}\boldsymbol{\lambda}^T)\,\frac{\partial^2\phi_0}{\partial p}\right),\\ &= \left(p\,\frac{\partial}{\partial q} - \frac{\partial V}{\partial q}\,\frac{\partial}{\partial p} - \gamma\,p\,\frac{\partial}{\partial p} + \gamma\,\beta^{-1}\,\frac{\partial^2}{\partial p}\right)\phi_0 = \mathcal{L}^L\phi_0,\end{aligned}$$

where we used the expression of the friction coefficient in the limiting equation (4.32), and where the superscript in $\mathcal{L}^L$ indicates that the operator obtained is the generator of the Langevin dynamics. With this condition satisfied, we can obtain the following expression for ϕ_2 by solving

Eq. (4.33c):

$$\begin{aligned}\phi_2 &= -\mathcal{L}_0^{-1}\left(\langle\boldsymbol{\lambda},\mathbf{z}\rangle\,\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{z}\right\rangle-\gamma\right)\frac{\partial^2\phi_0}{\partial p^2}-\mathcal{L}_0^{-1}\,\langle\boldsymbol{\lambda},\mathbf{z}\rangle\,\frac{\partial\psi_1}{\partial p}+\psi_2(q,p),\\ &= \frac{1}{2}\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{z}\right\rangle^2\frac{\partial^2\phi_0}{\partial p^2}+\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{z}\right\rangle\,\frac{\partial\psi_1}{\partial p}+\psi_2(q,p),\end{aligned}$$

where a function depending only on q,p was added to the function ψ_2 from the first to the second line. It can be verified that the centering condition for the next order gives

$$\mathcal{L}^L\psi_1=0,$$

implying that $\psi_1=0$ in view of Eq. (4.35). Solving Eq. (4.33d) for $i=1$, we obtain an expression for ϕ_3 in terms of ϕ_0 and ψ_2:

$$\begin{aligned}\phi_3 =& -\mathcal{L}_0^{-1}\left(\frac{1}{2}\,\langle\boldsymbol{\lambda},\mathbf{z}\rangle\,\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{z}\right\rangle^2\frac{\partial^3\phi_0}{\partial p^3}-p\,\left\langle\mathbf{A}^{-1}\boldsymbol{\lambda},\boldsymbol{\lambda}\right\rangle\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{z}\right\rangle\frac{\partial^2\phi_0}{\partial p^2}\right)\\ &-\mathcal{L}_0^{-1}\,\langle\boldsymbol{\lambda},\mathbf{z}\rangle\,\frac{\partial\psi_2}{\partial p}+\psi_3(q,p),\\ =&\left(\frac{1}{6}\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{z}\right\rangle^3+\gamma\,\beta^{-1}\,\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{A}^{-1}\mathbf{z}\right\rangle\right)\frac{\partial^3\phi_0}{\partial p^3}\\ &-\left(\gamma\,p\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{A}^{-1}\,\mathbf{z}\right\rangle\right)\frac{\partial^2\phi_0}{\partial p^2}+\left\langle\mathbf{A}^{-T}\boldsymbol{\lambda},\mathbf{z}\right\rangle\,\frac{\partial\psi_2}{\partial p}+\psi_3(q,p).\end{aligned}$$

One can obtain an equation for ψ_2 by looking at the solvability condition for (4.33d) with $i=2$. From here on, we limit ourselves to the model GL2, from which expressions for the model GL1 can be obtained by passing to the limit $\alpha\to\infty$. The equation for ψ_2 is:

$$\begin{aligned}&-\mathcal{L}^L\psi_2-\frac{\gamma}{\beta}\,\mathcal{L}^L\left(\frac{\partial^2\phi_0}{\partial p^2}\right)\\ &\qquad=\frac{\gamma}{2\beta^2}(1-\alpha^{-2})\left(\beta\mathcal{L}^L\left(\frac{\partial^2\phi_0}{\partial p^2}\right)-2\,\beta^2\,p\,\mathcal{L}^L\left(\frac{\partial\phi_0}{\partial p}\right)+2\beta\frac{\partial^2\phi_0}{\partial q\partial p}-2\beta\gamma\frac{\partial^2\phi_0}{\partial p^2}\right).\end{aligned}\tag{4.36}$$

We observe that, when $\alpha=1$, the equation simplifies and admits the solution

$$\psi_2=-(\gamma/\beta)\,\frac{\partial^2\phi_0}{\partial p^2}+C,\tag{4.37}$$

for a constant C such that Eq. (4.35) is satisfied. When $\alpha\neq1$, it is not possible to find an explicit expression for ψ_2, but it is still possible to solve equation Eq. (4.36) numerically, which we will do in Section 4.4 to ascertain the validity of our numerical results. Using Eq. (4.10), we

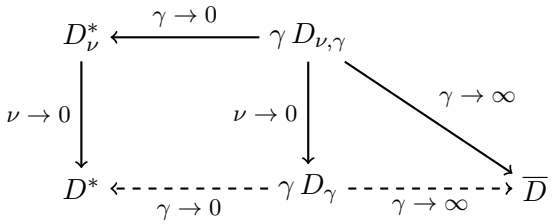
obtain an approximation of the diffusion coefficient:

$$\begin{aligned} D_n &= \beta^{-1} \int_{\mathbf{T}\times\mathbf{R}\times\mathbf{R}^n} \mathbf{A}_s : (\nabla_{\mathbf{z}}\phi_1 \otimes \nabla_{\mathbf{z}}\phi_1)\, \mu(\mathrm{d}q\, \mathrm{d}p\, \mathrm{d}\mathbf{z}) \\ &\quad + \beta^{-1}\, \nu^2 \int_{\mathbf{T}\times\mathbf{R}\times\mathbf{R}^n} \mathbf{A}_s : (\nabla_{\mathbf{z}}\phi_2 \otimes \nabla_{\mathbf{z}}\phi_2 + 2\, \nabla_{\mathbf{z}}\phi_1 \otimes \nabla_{\mathbf{z}}\phi_3)\, \mu(\mathrm{d}q\, \mathrm{d}p\, \mathrm{d}\mathbf{z}) \\ &= \beta^{-1}\, \gamma \int_{\mathbf{T}\times\mathbf{R}} \left| \frac{\partial \phi_0}{\partial p} \right|^2 \mu(\mathrm{d}q\, \mathrm{d}p) + \beta^{-1}\, \gamma\, \nu^2 \times \\ &\quad \int_{\mathbf{T}\times\mathbf{R}} 2\, \frac{\partial \phi_0}{\partial p} \frac{\partial \psi_2}{\partial p} + \frac{\gamma}{\beta} \left(1 + \frac{1}{\alpha^2}\right) \frac{\partial \phi_0}{\partial p} \frac{\partial^3 \phi_0}{\partial p^3} + \frac{\gamma}{\beta} \left(\frac{1}{\alpha^2} - 1\right) \left(\frac{\partial^2 \phi_0}{\partial p^2}\right)^2 \mu(\mathrm{d}q\, \mathrm{d}p). \end{aligned} \tag{4.38}$$

We recognize that the first term is the expression of the effective diffusion coefficient for the Langevin equation, and we notice, by substitution of (4.37) in (4.38), that the ν^2 correction to the effective diffusion coefficient is zero when $\alpha = 1$. In this case it can be easily shown that the first nonzero correction is of order $\mathcal{O}(\nu^4)$.

4.4 Numerical experiments

In Sections 4.3.1 to 4.3.3, we established the following limits (in solid arrows):



One of the aims of this section is to verify numerically the limits $\gamma D_{\nu,\gamma} \to \overline{D}$ as $\gamma \to \infty$ and $D_{\nu,\gamma} \to D_\gamma$ as $\nu \to 0$, by employing a spectral method to approximate the solution to the Poisson equation (4.9). We consider the general, non-conformal case where the finite-dimensional Galerkin space, which we denote by V_N, does not necessarily contain only mean-zero functions with respect to μ. Following the ideas developed in [190], we use a saddle point formulation to obtain an approximation of the solution to Eq. (4.9):

$$\begin{cases} -\,\Pi_N\, \mathcal{L}\, \Pi_N \Phi_N + \alpha_N u_N = \Pi_N p, \\ \langle \Phi_N, u_N \rangle = 0, \end{cases} \tag{4.39}$$

where Π_N is the $L^2(\mu)$ projection operator on V_N, satisfying $\langle \Pi_N u - u, v_N \rangle = 0$ for all functions $v_N \in V_N$, $u_N = \Pi_N 1 / \|\Pi_N 1\| \in V_N$, and α_N is a Lagrange multiplier. In this chapter, $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote respectively the standard inner product and norm of $L^2(\mu)$. We will choose V_N as a subspace of $L^2(\mu)$ spanned by tensor products of appropriate one-dimensional functions constructed from trigonometric functions (in the q direction) and Hermite polynomials (in the p and $\mathbf{z}$ directions). In the case of harmonic noise, for example, we employ the basis functions

$$e_{i,j,k,l} = \left(\mathcal{Z}\, \mathrm{e}^{\beta\left(H(q,p) + |\mathbf{z}|^2\right)} \right)^{\frac{1}{2}} G_i(q)\, H_j(p)\, H_k(z_1)\, H_l(z_2), \qquad 0 \le i, j, k, l \le N.$$

where G_i are trigonometric functions,

$$G_i(q) = \begin{cases} 1/\sqrt{2\pi}, & \text{if } i = 0, \\ \sin\left(\frac{i+1}{2} q\right)/\sqrt{\pi}, & \text{if } i \text{ is odd}, \\ \cos\left(\frac{i}{2} q\right)/\sqrt{\pi}, & \text{if } i \text{ is even}, i > 0. \end{cases} \tag{4.40}$$

and H_j are rescaled normalized Hermite functions,

$$H_j(p) = \frac{1}{\sqrt{\sigma}} \psi_j\left(\frac{p}{\sigma}\right), \qquad \psi_j(p) := (2\pi)^{-\frac{1}{4}} \frac{(-1)^j}{\sqrt{j!}} \mathrm{e}^{\frac{p^2}{4}} \frac{\mathrm{d}^j}{\mathrm{d}p^j}\left(\mathrm{e}^{-\frac{p^2}{2}}\right). \tag{4.41}$$

The functions $(H_j)_{j \in \mathbf{N}}$ are orthonormal in $L^2(\mathbf{R})$ regardless of the value of σ, a scaling parameter that can be adjusted to better resolve Φ_N. We use the same number of basis function in every direction because, although Hermite series converge much slower than Fourier series as $N \to \infty$ when σ is fixed, their spatial resolution is comparable to that of Fourier series when σ is chosen appropriately, as demonstrated in [209]. More specifically, while $\mathcal{O}(k^2)$ Hermite modes are required in the former case to resolve k wavelengths of the Gaussian-windowed cosine function $(\cos(kx)\ \mathrm{e}^{-x^2/2})$, only $\mathcal{O}(k)$ are required when $\sigma \propto 1/\sqrt{N}$.

To solve the linear system associated with (4.39) and the basis functions (4.4), we use either a direct method or, when the computational cost and memory needed to solve Eq. (4.39) using a direct method quickly is prohibitive, the generalized minimal residual method (GMRES). The GMRES implementation we employed is the one included in the `sparse.linalg` module of *SciPy* [123]. Once Φ_N has been calculated, the effective diffusion coefficient is approximated by:

$$D_N \approx \langle \Phi_N, \Pi_N p \rangle, \tag{4.42}$$

where the projection $\Pi_N p$ is calculated by numerical quadrature. We note that, when V is a trigonometric function, the stiffness matrix associated with (4.39) and the basis functions (4.4) is sparse, so the matrix-vector product involved in GMRES can be performed efficiently.

4.4.1 Calculation of the velocity autocorrelation function

We now demonstrate how a Hermite–Galerkin discretization can be employed to calculate the velocity autocorrelation function and its Fourier transform, and we validate our numerical method by comparison with some of the results obtained in the early studies [119, 118]. In these, the authors use the method of continued matrix-valued fractions, together with analog simulations, in order to approximate the spectrum of the velocity autocorrelation function (VAF), for the same Markovian approximations of the GLE as the ones considered in this chapter. Though interesting in its own right, the VAF is also related to the effective diffusion coefficient by a formula of Green–Kubo type:

$$D_n = \int \int_0^\infty (\mathrm{e}^{\mathcal{L}t} p)\, p \, \mathrm{d}t \, \mathrm{d}\mu = \int_0^\infty \mathbf{E}(p(0)\, p(t)) \, \mathrm{d}t,$$

where $\mathbf{E}$ is the expectation with respect to product measure induced by the measure μ on the initial condition and the Wiener measure on the solution. To calculate the VAF numerically, we approximate the solution $\mathrm{e}^{\mathcal{L}t}\, p$ of the backward Kolmogorov equation by a Galerkin method, using the same finite-dimensional space as above:

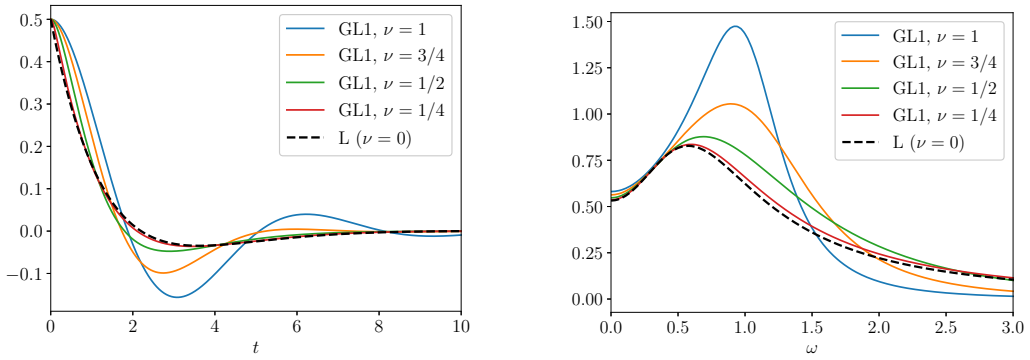
$$\begin{cases} \dfrac{\partial v_N}{\partial t} = (\Pi_N \mathcal{L} \Pi_N) v_N, \\ v_N(0) = \Pi_N p. \end{cases}$$

The Runge–Kutta 45 method is used for the integration in time of this system. The VAF can then be obtained by the inner product $\langle v_N(t), \Pi_N p \rangle$, and the associated spectral measure, i.e. the (purely real) Fourier transform of the VAF, is defined as

$$\hat{v}_N(\omega) := \int_{\mathbf{R}} \langle v_N(t), \Pi_N p \rangle \ \mathrm{e}^{-i\omega t} \ \mathrm{d}t. \tag{4.43}$$

Numerically, the signal v_N is available only at discrete times, so the Fourier transform in Eq. (4.43) must be approximated. Although dedicated methods exist [12], for our purposes it was sufficient to carry out this approximation by appending a large number of zero entries to the discrete representation of v_N, in order to increase the resolution in frequency space, and by using the usual discrete Fourier transform. Here this padding of the numerical solution with zeros is appropriate because we expect the VAF to decrease exponentially with time, see Section 4.5.

In Figs. 4.6a and 4.6b, the VAF and its spectrum are presented in the case of the GL1 dynamics for a range of values of ν. We note the presence, already highlighted in [118, 119], of a maximum of the spectrum close to the natural frequency of oscillations of the Hamiltonian dynamics ($\gamma = 0$) near the minimum of the potential (i.e., near $\omega = \sqrt{V''(0)} = \sqrt{2}/2$), a maximum that grows in magnitude as ν increases.

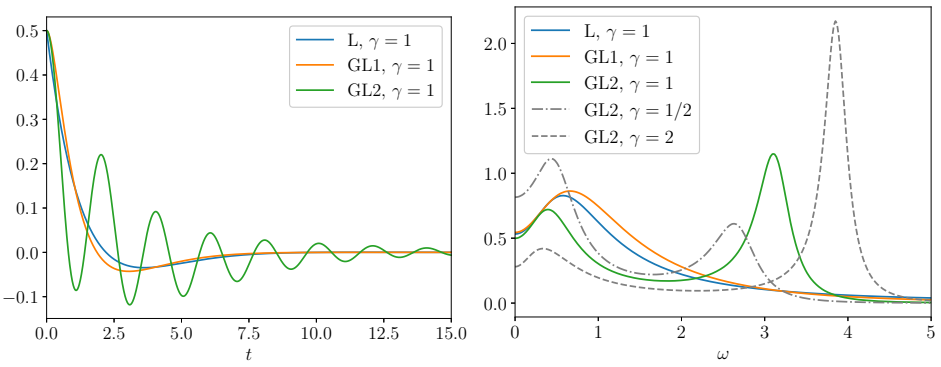


(a) VAF **(b)** Spectrum of the VAF

Figure 4.6: VAF and its spectrum for the Langevin and the GL1 dynamics, with the parameters $\gamma = 1$, $\beta = 2$, and $\nu \in \{1/4, 1/2, 3/4, 1\}$. In both figures, the curves corresponding to GLE with $\nu = 1/4$ are very close to those associated with the Langevin dynamics.

We now consider the model GL2. The corresponding results are presented in Figs. 4.7a and 4.7b. In contrast with the previous model, we note the presence of an additional peak near the frequency associated with the harmonic noise (see Eq. (4.6), $\omega \approx \alpha/\nu^2$), although for a larger value of ω. Numerically, we found that, as the value of the effective friction parameter γ goes to

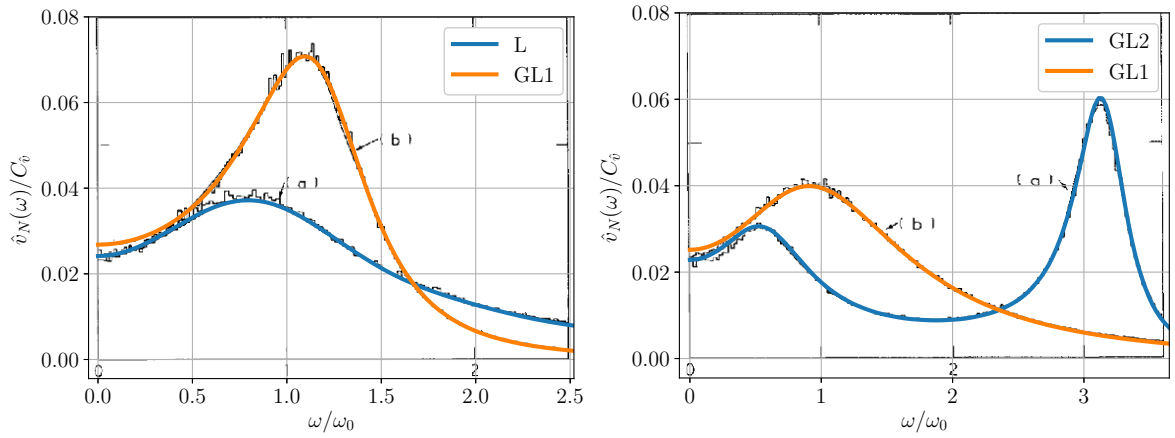
zero, i.e. as the coupling between p and $\mathbf{z}$ decreases, the position of the second peak moves closer to the frequency of the driving noise, and conversely when γ increases.



(a) VAF **(b)** Spectrum of the VAF

Figure 4.7: VAF and its spectrum for the Langevin, GL1 and GL2 dynamics, with the parameters $\gamma = 1$, $\beta = 2$, $\nu = 1/\sqrt{5}$, $\alpha = 1/\sqrt{5}$.

A more quantitative comparison with the numerical and experimental results obtained in [118, 119] is presented in Figs. 4.8a and 4.8b. The agreement between the results is very good.



(a) GL1: $\gamma = 2^{-1/2}$, $\nu = 2^{1/4}$ **(b)** GL2: $\gamma = 2^{-1/2}$, $\nu = 2^{1/4}5^{-1/2}$, $\alpha = 5^{-1/2}$

Figure 4.8: Spectrum of the VAF and comparison and with the results from [118, 119], which are displayed in the background of the figures, for the Langevin, GL1 and GL2 dynamics. Here $\omega_0 = \sqrt{2}/2$ is the natural frequency of oscillations of the Hamiltonian dynamics, and $C_{\hat{v}} = \sqrt{2}/(2\pi)^2$ is the scaling factor used in [118, 119] to obtain a dimensionless quantity.

4.4.2 Convergence of the method

In this subsection, we discuss the convergence of the effective diffusion diffusion D_N, as approximated by the Galerkin method (4.39) and Eq. (4.42), when the number N of basis functions tends to infinity. Our study of convergence in this subsection is exclusively empirical, and the problem of showing convergence rigorously will be addressed in future work. We remind the reader that, since the models we consider are hypocoercive, standard arguments based on a Lax–Milgram theorem and Céa's lemma, presented for instance in [2], are not sufficient for proving the well-posedness of the discrete linear system and obtaining error bounds.

The main obstacle to proving convergence at present is the absence of a discrete hypocoercivity estimate for the GLE, on which not only the convergence but also the well-posedness and the consistency of the numerical method depend. In [190], where the mixed Fourier/Hermite spectral method of which our method is an extension was introduced, the authors proved the hypocoercivity of the discretized generator of the Langevin equation in $L^2(\mu)$ by employing arguments from the hypocoercivity setting presented in [57, 58]. Generalizing these arguments to the GLE, however, is considerably more difficult and might even not be possible. In principle, proving the hypocoercivity of the discrete dynamics in $H^1(\mu)$ would be already constitute a step towards obtaining rigorous estimates, but our attempts have so far proved inconclusive as well, see Section 4.B.

Through numerical exploration, we observed that the saddle point Galerkin formulation (4.39) yields accurate results even for small values of N, provided that the friction coefficient γ is high enough. This is illustrated in Fig. 4.9 for the case of Langevin dynamics. For $\gamma \geq 1$, in particular, a value of the diffusion coefficient with accuracy of the order of one percent can be obtained with $N = 6$ basis functions. In the underdamped regime ($\gamma \ll 1$), however, the solution to the Poisson equation (4.9) is less regular and more basis functions are required to obtain an accurate solution. As is apparent from the figure, this difficulty can be alleviated by increasing the scaling coefficient σ, which enables a significant reduction of the error.

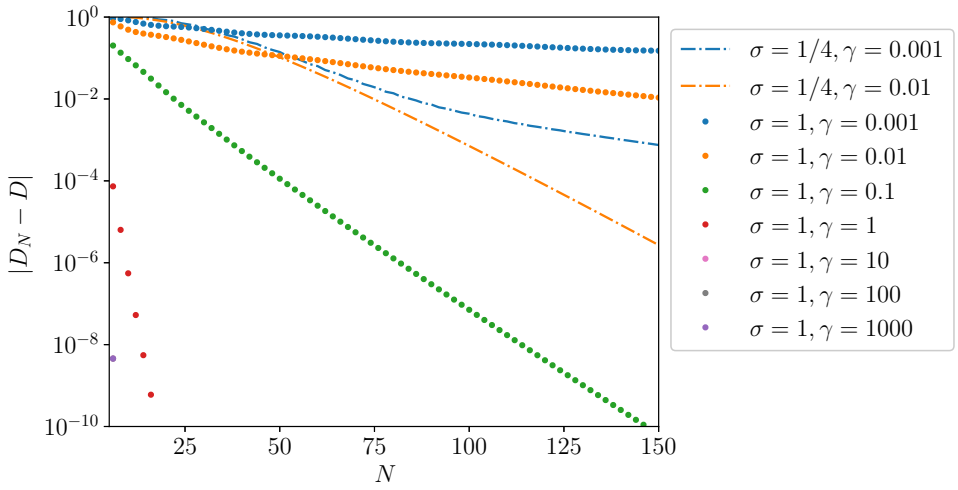


Figure 4.9: Numerical error on the effective diffusion coefficient associated with Langevin dynamics, when an equal number of basis functions is used in each direction. Precise solutions, used for estimating the error, were calculated using $N = 250$, and for the scaling parameter $\sigma = \frac{1}{4}$ in the cases $\gamma = 0.001, 0.01$ and $\sigma = 1$ otherwise. Notice that the data points corresponding to $\gamma = 10, 100, 1000$ are not visible on the graph; this is because the corresponding errors are all below 10^{-10}.

Such a convergence study is not possible for the models GL1 and GL2, as the time required for the computations would be prohibitive. When setting the tolerance of the GMRES method to 0.01, for example, the calculation of the solution to (4.39) when $\gamma = 0.001$ using $N = 100$ basis functions per direction takes over 4 days with an *Intel i7-3770* processor.

4.4.3 Effective diffusion

We examine the variation of the diffusion coefficient with respect to γ in Fig. 4.10. The parameters used in the simulations are presented in Table 4.1. The effective diffusion coefficient was computed for 100 values of γ evenly spaced on a logarithmic scale, and for each value of γ the numerical error was approximated by carrying out the computation with half the numbers of basis function in each direction. When the estimated relative error was over 1%, which occurred roughly when $\gamma \leq 10^{-2}$ for the model GL1 and $\gamma \leq 10^{-1}$ for the model GL2, the corresponding data point was considered inaccurate and was removed.

Model	Method when $\gamma < 1$	Method when $\gamma > 1$
L	Direct ($N = 250$, $\sigma^{-2} = 16$)	Direct ($N = 250$, $\sigma^{-2} = 16$)
GL1	GMRES ($N = 100$, $\sigma^{-2} = 9$, tol $= 10^{-3}$)	Direct ($N = 40$, $\sigma^{-2} = 3$)
GL2	GMRES ($N = 40$, $\sigma^{-2} = 6$, tol $= 10^{-3}$)	Direct ($N = 16$, $\sigma^{-2} = 2$)

Table 4.1: Numerical parameters used to generate the data presented in Fig. 4.10.

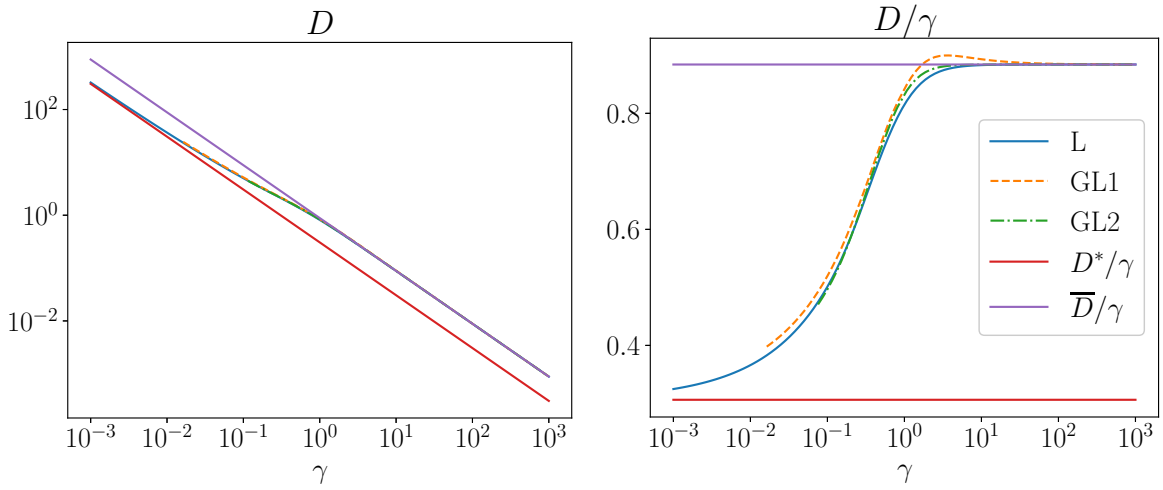


Figure 4.10: Diffusion coefficient as a function of γ, for the parameters $\nu = 1$ (for the models GL1 and GL2) and $\alpha = 1$ (for GL2). We observe that, for values of γ in the range $[1, 10]$, the GL1 diffusion coefficient is slightly larger than $\overline{D}/\gamma$.

We observe from Fig. 4.10 that the effective diffusion coefficient is of the same order of magnitude for the three models across the whole range of γ. We also notice that Eq. (4.16), which states that $D^* \leq \gamma D_\gamma \leq \overline{D}$ for the Langevin dynamics, does does not carry over to the GLE; indeed it is clear from the figure that, for γ close to 2, the effective diffusion coefficient for the model GL1 is strictly greater than $\overline{D}/\gamma$.

To conclude this section, we verify the convergence in the limit $\nu \to 0$, which we studied theoretically in Section 4.3.3. Figure 4.11 presents the evolution of the diffusion coefficient with ν for a range of values of α and fixed values of β and γ. As expected, we recover the effective diffusion coefficient of the model GL1 as $\alpha \to \infty$, and that of the Langevin dynamics as $\nu \to 0$. For $\alpha = 1$, the convergence to the limit as $\nu \to 0$ appears to be faster than for the other values of α, which is consistent with our observation in Section 4.3.3 that the $\mathcal{O}(\nu^2)$ term in the asymptotic expansion (4.38) vanishes in that case. In Fig. 4.12, we compare in a log-log plot the

observed convergence with respect to ν with predictions obtained by solving (4.36) numerically and calculating the effective diffusion using (4.38). We note that, since Eq. (4.36) is posed in two dimensions, the numerical solution of this equation is much cheaper than that of the 3D or 4D Poisson equation involved in the direct calculation of the effective diffusion coefficient associated with the GLE. We see from the figure that the asymptotic expansion, truncated after the first nonzero correction, is very accurate for ν up to about 0.3. When $\alpha = 1$, numerical errors appear for small ν, which explains the deviation from the theoretical behavior as $\mathcal{O}(\nu^4)$.

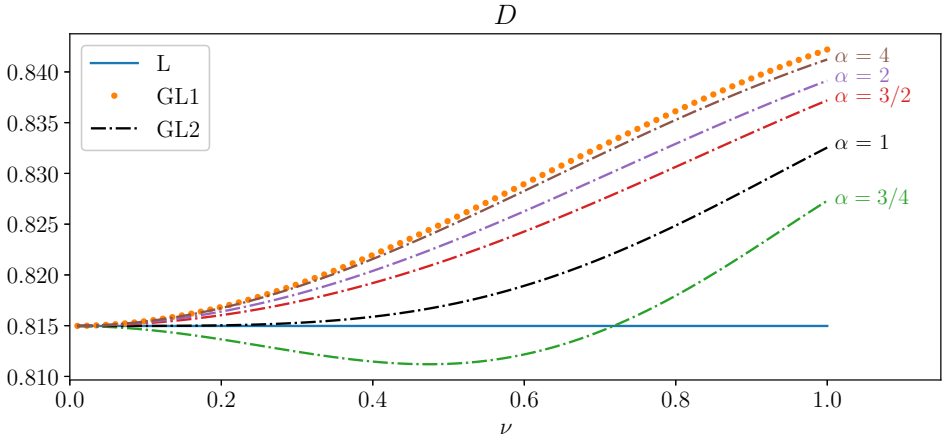


Figure 4.11: Effective diffusion coefficient against ν, whose square encodes the characteristic time of the autocorrelation function of the noise, for fixed values $\beta = \gamma = 1$.

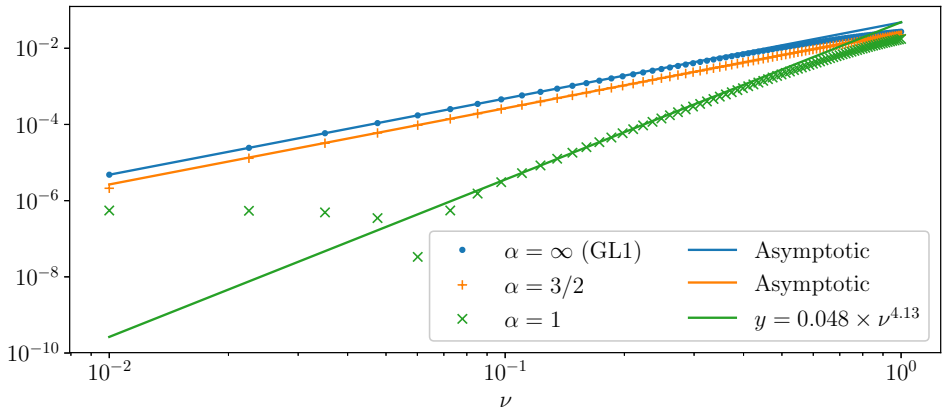


Figure 4.12: Deviation of the effective diffusion coefficient from its limiting value as $\nu \to 0$, and comparison with the asymptotic expansion (4.38). For $\alpha = 1$, the straight line was fitted using a subset of the data points excluding the data corresponding to small ν (for which numerical errors dominate) and to large ν (for which the asymptotic expansion is no longer accurate).

4.5 Convergence of the GLE dynamics to equilibrium

In [190], the authors show that the Dolbeault–Mouhot–Schmeiser (DMS) framework introduced in [57, 58] can be applied, almost immediately, in order to prove the hypocoercivity of the Langevin dynamics in $L^2(\mu)$. As mentioned above, generalizing the DMS approach to the GLE turns out to be considerably more difficult, and has not yet been achieved despite recent efforts

by ourselves and others. In this section, we will therefore study the convergence of the generalized Langevin dynamics to equilibrium by using the $H^1(\mu)$ approach formalized in [220].

In the calculations below we consider the model GL2, but similar calculations can be carried out for the other two models. Introducing the adjoint operator $\partial_{z_2}^* = \beta\, z_2 - \partial_{z_2}$ and rewriting the generator of the dynamics explicitly for the model GL2,

$$\begin{aligned} -\mathcal{L} &= \nu^{-2}\,\alpha^2\,\beta^{-1}\,\partial_{z_2}^*\,\partial_{z_2} \\ &\quad + \nu^{-2}\,\alpha(z_1\,\partial_{z_2} - z_2\,\partial_{z_1}) + \sqrt{\gamma}\,\nu^{-1}(p\,\partial_{z_1} - z_1\,\partial_p) + (\partial_q V\,\partial_p - p\,\partial_q) =: A^*A + B. \end{aligned} \tag{4.44}$$

In addition to the periodic setting, it will be useful to assume that q can be unbounded, so for generality we will denote the state space of the variable q by $\mathcal{X}$. The relevant commutators for the study of hypocoercivity are

$$C_0 := \nu\,\alpha^{-1}\,\beta^{1/2}\,A = \partial_{z_2},$$

$$C_1 := \nu^2\,\alpha^{-1}\,[C_0, B] = -\partial_{z_1}, \tag{4.45a}$$

$$C_2 := \nu\,\gamma^{-1/2}\,[C_1, B] + \alpha\,\nu^{-1}\,\gamma^{-1/2}\,C_0 = \partial_p, \tag{4.45b}$$

$$C_3 := [C_2, B] + \gamma^{1/2}\,\nu^{-1}\,C_1 = -\partial_q, \tag{4.45c}$$

$$[C_3, B] + V''(q)\,C_2 = 0. \tag{4.45d}$$

We are interested not only in whether the dynamics converge to the equilibrium but also in the estimation of rate of convergence, so we will carry out carefully the details of the proof of hypocoercivity, which in [220] was presented in an abstract setting. Let us define

$$\begin{aligned} (\!(h,h)\!) &= \|h\|^2 + a_0\,\|\partial_{z_2}h\|^2 + a_1\,\|\partial_{z_1}h\|^2 + a_2\,\|\partial_p h\|^2 + a_3\,\|\partial_q h\|^2 \\ &\quad - b_0\,\langle\partial_{z_2}h, \partial_{z_1}h\rangle - b_1\,\langle\partial_{z_1}h, \partial_p h\rangle - b_2\,\langle\partial_p h, \partial_q h\rangle\,. \end{aligned} \tag{4.46}$$

To prove hypocoercivity for the norm of $H^1(\mathcal{X}\times R^3, \mu)$, where μ is the invariant measure of the dynamics, we must show that it is possible to find coefficients $a_0, a_1, a_2, a_3, b_0, b_1, b_2$ such that, first, the weighted Sobolev norm is equivalent to the norm (4.46) and, second, coercivity holds for this modified norm, i.e. there exists $\lambda > 0$ such that $-(\!(h, \mathcal{L}h)\!) \geq \lambda\,(\!(h,h)\!)$, where the scalar product $(\!(\cdot,\cdot)\!)$ is obtained by polarization. By the Cauchy–Schwarz inequality, it is clear that

$$(\!(h,h)\!) - \|h\|^2 \geq \begin{pmatrix} \|\partial_{z_2}h\| \\ \|\partial_{z_1}h\| \\ \|\partial_p h\| \\ \|\partial_q h\| \end{pmatrix}^T \underbrace{\begin{pmatrix} a_0 & -b_0 & 0 & 0 \\ 0 & a_1 & -b_1 & 0 \\ 0 & 0 & a_2 & -b_2 \\ 0 & 0 & 0 & a_3 \end{pmatrix}}_{:=M_1} \begin{pmatrix} \|\partial_{z_2}h\| \\ \|\partial_{z_1}h\| \\ \|\partial_p h\| \\ \|\partial_q h\| \end{pmatrix}, \tag{4.47}$$

so for the first condition it is necessary and sufficient that M_1 be positive definite. Moving to the second condition, we calculate the action of the symmetric part of the generator on the pure

terms (the ones multiplying a_0, a_1, a_2, a_3):

$$\langle C_0 h, C_0(A^*A)h\rangle = \alpha^2\,\nu^{-2}\,\beta^{-1}\,\|C_0^2\,h\|^2 + \alpha^2\,\nu^{-2}\,\|C_0 h\|^2, \tag{4.48a}$$
$$\langle C_1 h, C_1(A^*A)h\rangle = \alpha^2\,\nu^{-2}\,\beta^{-1}\,\|C_0\,C_1\,h\|^2, \tag{4.48b}$$
$$\langle C_2 h, C_2(A^*A)h\rangle = \alpha^2\,\nu^{-2}\,\beta^{-1}\,\|C_0\,C_2\,h\|^2, \tag{4.48c}$$
$$\langle C_3 h, C_3(A^*A)h\rangle = \alpha^2\,\nu^{-2}\,\beta^{-1}\,\|C_0\,C_3\,h\|^2. \tag{4.48d}$$

where we took into account that all the C_i's commute and $[A, A^*] = \alpha^2\,\nu^{-2}$. The action of the antisymmetric part of the generator (B) on the pure terms is

$$\langle C_0 h, C_0 B h\rangle = \alpha\,\nu^{-2}\,\langle C_0 h, C_1 h\rangle\,, \tag{4.49a}$$
$$\langle C_1 h, C_1 B h\rangle = \gamma^{1/2}\,\nu^{-1}\,\langle C_1 h, C_2 h\rangle - \alpha\,\nu^{-2}\,\langle C_0 h, C_1 h\rangle\,, \tag{4.49b}$$
$$\langle C_2 h, C_2 B h\rangle = \langle C_2 h, C_3 h\rangle - \gamma^{1/2}\,\nu^{-1}\,\langle C_1 h, C_2 h\rangle\,, \tag{4.49c}$$
$$\langle C_3 h, C_3 B h\rangle = -\left\langle C_3 h, V''(q)\,C_2\,h\right\rangle. \tag{4.49d}$$

For the mixed terms (the ones multiplying b_0, b_1, b_2), we have

$$\langle C_0(A^*A)h, C_1 h\rangle + \langle C_0 h, C_1(A^*A)h\rangle = \alpha^2\,\nu^{-2}\,\left(\langle C_0 h, C_1 h\rangle + 2\,\beta^{-1}\,\left\langle C_0^2 h, C_0 C_1 h\right\rangle\right), \tag{4.50a}$$
$$\langle C_1(A^*A)h, C_2 h\rangle + \langle C_1 h, C_2(A^*A)h\rangle = 2\,\alpha^2\,\nu^{-2}\,\beta^{-1}\,\langle C_0 C_1 h, C_0 C_2 h\rangle\,, \tag{4.50b}$$
$$\langle C_2(A^*A)h, C_3 h\rangle + \langle C_2 h, C_3(A^*A)h\rangle = 2\,\alpha^2\,\nu^{-2}\,\beta^{-1}\,\langle C_0 C_2 h, C_0 C_3 h\rangle\,, \tag{4.50c}$$

and

$$\begin{aligned}\langle C_0 B h, C_1 h\rangle + \langle C_0 h, C_1 B h\rangle &= \alpha\,\nu^{-2}\,\|C_1 h\|^2 \\ &\quad + \sqrt{\gamma}\,\nu^{-1}\,\langle C_0 h, C_2 h\rangle - \alpha\,\nu^{-2}\,\|C_0 h\|^2,\end{aligned} \tag{4.51a}$$
$$\begin{aligned}\langle C_1 B h, C_2 h\rangle + \langle C_1 h, C_2 B h\rangle &= \gamma^{1/2}\,\nu^{-1}\,\|C_2 h\|^2 - \alpha\,\nu^{-2}\,\langle C_0 h, C_2 h\rangle \\ &\quad + \langle C_1 h, C_3 h\rangle - \gamma^{1/2}\,\nu^{-1}\|C_1 h\|^2\end{aligned} \tag{4.51b}$$
$$\begin{aligned}\langle C_2 B h, C_3 h\rangle + \langle C_2 h, C_3 B h\rangle &= \|C_3 h\|^2 - \gamma^{1/2}\,\nu^{-1}\,\langle C_1 h, C_3 h\rangle \\ &\quad - \left\langle V''(q) C_2 h, C_2 h\right\rangle.\end{aligned} \tag{4.51c}$$

Combining Eqs. (4.49) to (4.51) and using Cauchy–Schwarz inequality again, we obtain

$$-(\!(h, \mathcal{L}h)\!) \geq \begin{pmatrix}\|C_0\,C_0\,h\|\\ \|C_0\,C_1\,h\|\\ \|C_0\,C_2\,h\|\\ \|C_0\,C_3\,h\|\end{pmatrix}^T (\alpha^2\nu^{-2}\,\beta^{-1}\,M_1) \begin{pmatrix}\|C_0\,C_0\,h\|\\ \|C_0\,C_1\,h\|\\ \|C_0\,C_2\,h\|\\ \|C_0\,C_3\,h\|\end{pmatrix} + \begin{pmatrix}\|C_0\,h\|\\ \|C_1\,h\|\\ \|C_2\,h\|\\ \|C_3\,h\|\end{pmatrix}^T M_2 \begin{pmatrix}\|C_0\,h\|\\ \|C_1\,h\|\\ \|C_2\,h\|\\ \|C_3\,h\|\end{pmatrix},$$

where M_1 is the same matrix as in Eq. (4.47) and M_2 is equal to

$$M_2 = \begin{pmatrix} -\frac{b_0\alpha}{2\,\nu^2}+\frac{\alpha^2(a_0+1)}{\nu^2} & -\nu^{-2}\left|a_0\alpha-a_1\alpha+\frac{b_0\,\alpha^2}{2}\right| & -\left|\frac{b_0\sqrt{\gamma}}{2\nu}-\frac{b_1\alpha}{2\nu^2}\right| & 0 \\ 0 & \frac{b_0\alpha}{2\nu^2}-\frac{b_1\sqrt{\gamma}}{2\nu} & -\frac{\sqrt{\gamma}}{\nu}|a_1-a_2| & -\left|\frac{b_1}{2}-\frac{b_2\sqrt{\gamma}}{2\nu}\right| \\ 0 & 0 & \frac{b_1\sqrt{\gamma}}{2\nu}-\frac{b_2}{2}\max_q|V''(q)| & -\max_q|a_2-a_3V''(q)| \\ 0 & 0 & 0 & \frac{b_2}{2} \end{pmatrix}. \quad (4.52)$$

Theorem 24 in [220] shows that it is always possible to find parameters $a_0, a_1, a_2, a_3, b_0, b_1, b_2$ such that this matrix is positive-definite, leading to the coercivity of $-\mathcal{L}$ for the modified norm. From there one can conclude to the convergence to equilibrium by noticing that, if h satisfies the equation $\partial_t h = \mathcal{L}h$,

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\!(h,h)\!) = (\!(h,\mathcal{L}h)\!) &\le -c_2\sum_{i=0}^{3}\|C_ih\|^2 \\ &\le -c_2\left(\zeta\kappa\|h\|^2+(1-\zeta)\sum_{i=0}^{3}\|C_ih\|^2\right) \qquad (\kappa\text{-coercivity of } \textstyle\sum_{i=0}^{3}C_i^*C_i) \\ &\le -c_2\min\left(\zeta\kappa,\frac{1-\zeta}{C_1}\right)(\!(h,h)\!), \end{aligned}$$

where C_1 is the largest eigenvalue of the symmetric part of M_1 and c_2 is the smallest eigenvalue of the symmetric part of M_2. The sharpest bound is obtained when the arguments of the minimum are equal, leading by Grönwall's lemma to the convergence rate $c_2/(C_1+\kappa^{-1})$.

Rather than adopting the procedure outlined in [220] for the calculation of suitable coefficients a_i and b_i, the generality of which leads to a rather crude estimate, we obtain a lower bound for the rate of convergence by solving the following optimization problem:

$$\max_{\substack{a_0,a_1,a_2,a_3\\ b_0,b_1,b_2\\ c_2,C_1}} \frac{c_2}{C_1+\kappa^{-1}} \quad \text{subject to} \quad \begin{cases} \frac{1}{2}(M_2+M_2^T)-c_2I\succeq 0, \\ C_1I-\frac{1}{2}(M_1+M_1^T)\succeq 0, \end{cases} \quad (4.53)$$

where both M_1 and M_2 depend on the coefficients a_i and b_i. In practice we solved this problem by using the sequential least squares programming (SLSQP) iterative method as implemented the SciPy library [123].

To gain some insight into the extent to which the rate of convergence is underestimated by this approach, we consider the generalized Langevin dynamics confined by the quadratic potential $V(q) = q^2/2$ in an unbounded domain. In this case, Eq. (4.4) can be written in the general form

$$\dot{\mathbf{X}} = -\mathbf{D}\mathbf{X}\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\mathbf{W},$$

where $\mathbf{X}^T = (q, p, \mathbf{z}^T)$, $\mathbf{D}$ and $\boldsymbol{\sigma}$ are constant matrices, and $\mathbf{W}$ is a standard Brownian motion on $\mathbf{R}^{2+n}$. It is known [147, Section 9.3] that the corresponding generator $\mathcal{L}$ generates a compact strongly continuous semigroup in $L^p(\mu)$, and by extension[1] in $W^{1,p}(\mu)$, for any $p \in (1,\infty)$, and

[1] Proving rigorously that this extension is indeed possible would require generalizing Theorem 9.3.19 in [147] to the spaces $W^{1,p}(\mu)$, which should follow from Proposition 9.3.8 in the same book.

that the associated spectrum can be obtained explicitly in terms of the eigenvalues of the drift matrix $\mathbf{D}$ in Eq. (4.4):

$$\sigma(\mathcal{L}) = \left\{ -\sum_{\mu \in \sigma(\mathbf{D})} \mu\, k_\mu, \quad k_\mu \in \mathbf{N} \right\}.$$

By [10, Theorem 5.3], the spectral bound of the generator, i.e. the eigenvalue of largest real part of $\mathcal{L}$, coincides with the exponential growth bound of the semigroup. In the case of the Langevin dynamics, for example, the spectral bound (independent of β) equals

$$-\max \sigma(\mathcal{L}) = \begin{cases} \gamma/2, & \text{if } \gamma < 2, \\ \gamma - \sqrt{\gamma^2 - 4}, & \text{if } \gamma \geq 2. \end{cases}$$

In Fig. 4.13, left panel, we compare the rate of convergence obtained by maximization of Eq. (4.53) to the exact exponential growth bound obtained by direct calculation of the eigenvalues of the (matrix-valued) drift coefficient of the system. In the case of the Langevin dynamics, we observe that, although our method underestimates the convergence rate for intermediate values of γ, it captures the exact value in the underdamped and overdamped limits. For the other two models, the correct rate of variation of the exponential growth bound with respect to γ is recovered as $|\log \gamma| \to \infty$, but the value of the bound itself is underestimated by a constant factor in these limits, a factor less than about 3.

Having verified that our methodology produces sensible results when $\mathcal{X} = \mathbf{R}$ and $V(q) = q^2/2$, we apply it to the case of a periodic potential. As V'' is no longer a constant function, we expect the crude bound $|V''| \leq \max_q |V''|$ in Eq. (4.52) to result in a loss of accuracy. The estimated rate of convergence for all three models is illustrated in Fig. 4.13, right panel. We observe that that the bound obtained is generally lower than in the quadratic case, and we notice that, for the Langevin dynamics, the estimated rate of convergence behaves asymptotically as $\mathcal{O}(\min(\gamma, \gamma^{-1}))$ as $|\log \gamma| \to \infty$, which is consistent with previous theoretical findings on the subject, for example in [190].

For the Langevin dynamics, it is also possible to approximate the actual rate of convergence with very good accuracy, although over a narrower range of γ, by discretizing $\mathcal{L}$ with the method of Section 4.4 and calculating numerically the eigenvalue of largest real part of the associated matrix. The results obtained using this procedure are also depicted in Fig. 4.13, and they reveal that the rate of convergence estimated by maximization of (4.53) is in fact about an order of magnitude below the true value. Though very accurate for the Langevin equation, this approach is unsuitable for the models GL1 and GL2, as the corresponding computational cost of estimating the relevant eigenvalue precisely is prohibitive, especially when $\gamma < 1$.

4.6 Conclusions

In this work, we studied quasi-Markovian approximations of the GLE, and we scrutinized in particular two finite-dimensional models of the noise: the scalar OU noise and the harmonic noise. The relation of these models to each other and to more general Markovian approximations

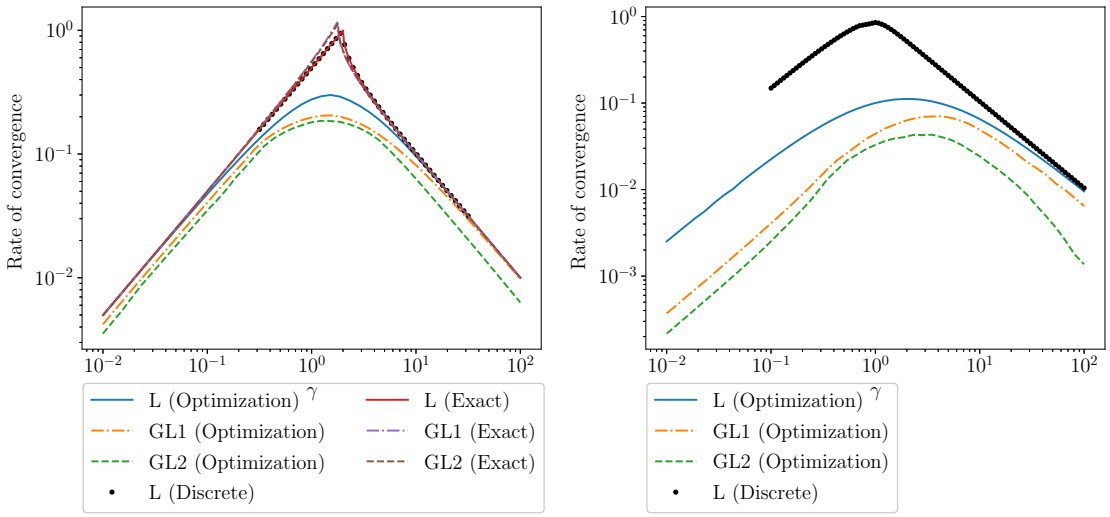


Figure 4.13: Convergence rate of the dynamics for the three models when $\mathcal{X} = \mathbf{R}, V(q) = q^2/2$ (left) and $\mathcal{X} = \mathbf{T}, V(q) = (1/2)(1 - \cos(q))$ (right), obtained by numerical solution of the maximization problem (4.53). In both cases, $\nu = \alpha^{-1} = 1/3$.

of the GLE was discussed and the asymptotic behavior of the associated effective diffusion coefficient was studied in different limits of physical relevance. We also employed an efficient Fourier/Hermite spectral method to verify most of our findings numerically, and in doing so we observed a very good agreement with previous numerical work [119, 118] on the subject.

Exciting questions remain open for future work. On the theoretical front, it is not clear whether an $L^2(\mu)$ hypocoercivity statement of the type proved in [57, 58] holds for the GLE. If this was the case, the need to employ hypoelliptic regularization in order to deduce convergence to equilibrium in the $L^2(\mu)$ norm, a program performed in [167], would be circumvented. In addition, it seems that even proving hypocoercivity of the discretized dynamics in the $H^1(\mu)$ norm is a difficult problem, as our unsuccessful attempt in Section 4.B proves. The availability of such a discrete hypocoercivity estimate would enable the calculation of bounds on the consistency error of the numerical method. On the numerical front, we believe that it would be possible to improve the accuracy and efficiency of the spectral method by using appropriate preconditioning, which would be especially valuable in the underdamped limit. We also intend to glean more insight into this limit by employing particle simulations in conjunction with the variance reduction technique based on control variates recently developed in [189].

4.7 Outlook: extension to two-dimensional periodic potentials

In future work, we will study the diffusion of a Brownian particle in a periodic potential in more than one dimension. Calculating the effective diffusion coefficient deterministically in this case is considerably more difficult than in the one-dimensional situation, because the Poisson equation is posed in twice the number of dimensions; considering only the Langevin dynamics for simplicity,

this equation reads

$$-\mathcal{L}u_j := -\sum_{i=1,2}\left(p_i\,\frac{\partial u_j}{\partial q_i} - \frac{\partial V}{\partial q_i}\,\frac{\partial}{\partial p_i} + \gamma\left(-p_i\,\frac{\partial u_j}{\partial p_i} + \beta^{-1}\,\frac{\partial^2 u_j}{\partial p_i}\right)\right) = p_j, \qquad j = 1,2, \quad (4.54)$$

where q_1, q_2 and p_1, p_2 are the position and momentum variables, respectively. The associated diffusion tensor is the following:

$$\mathbf{D} = -\int_{\mathbf{T}^2\times\mathbf{R}^2} (\mathcal{L}\mathbf{u}\otimes\mathbf{u})\ \mathrm{e}^{-\beta H(q_1,q_2,p_1,p_2)}\ \mathrm{d}q_1\,\mathrm{d}q_2\,\mathrm{d}p_1\,\mathrm{d}p_2,$$

where $H(q_1,q_2,p_1,p_2) = V(q_1,q_2) + p_1^2/2 + p_2^2/2$. It is simple to show that, when the potential is additively separable, i.e. when it can be decomposed as a sum of two one-dimensional functions as $V(\mathbf{q}) = V_1(q_1) + V_2(q_2)$, the tensor $\mathbf{D}$ is diagonal and the diagonal entries coincide with the one-dimensional effective diffusion coefficients for potentials V_1 and V_2, respectively.

Of more interest to us is the behavior of the effective diffusion tensor and, more specifically, the scaling of this tensor with respect to the friction coefficient γ in the underdamped limit, when the periodic potential is not additively separable. It is has been shown empirically, see for example [191, Chapter 4.2] and the references therein, that this scaling is different from that in the one-dimensional case (where we saw that $D_\gamma \approx D^*/\gamma$ to leading order) and that, furthermore, it depends on the degree of "non-separability" exhibited by the potential. In this respect, the underdamped limit is unlike the overdamped limit, in which it is possible to prove that the effective diffusion coefficient, divided by γ, always converges to the effective diffusion coefficient associated with the overdamped dynamics.

An example of such a non-separable potential is

$$V(q_1,q_2) = \frac{1}{2}\,(1-\cos(q_1)) + \frac{1}{2}\,(1-\cos(q_2)) + \delta\,\cos(q_1)\,\cos(q_2),$$

where here δ measures the degree of non-separability. By symmetry, the effective diffusion for this potential is a multiple of the identity matrix, i.e. $\mathbf{D} =: D_{\gamma,\delta}\,\mathbf{I}$. It is simple to adapt the numerical method presented in Section 4.4 in order to estimate $D_{\gamma,\delta}$; the only difference is that now we use the basis functions

$$e_{i,j,k,l} = \left(\mathcal{Z}\,\mathrm{e}^{\beta(H(q_1,q_2,p_1,p_2))}\right)^{\frac{1}{2}} G_i(q_1)\,G_j(q_2)\,H_k(p_1)\,H_l(p_2),$$

where $\{G_i\}_{i=0,1,\ldots}$ and $\{H_i\}_{i=0,1,\ldots}$ are as defined in Eqs. (4.40) and (4.41). Because the problem is high-dimensional and becomes singular as $\gamma \to 0$, however, we haven't yet been able to obtain accurate numerical results for γ significantly lower than about 0.1 using this approach. Some preliminary results, illustrating the convergence to the effective diffusion coefficient of the one-dimensional case as $\delta \to 0$, as well as the convergence to the overdamped limit as $\gamma \to \infty$, are presented in Fig. 4.14. We hope that future work on preconditioning for the Poisson equation (4.54), together with the addition of support for parallel computing in *Hermipy*, the *Python* library we developed for the numerical solution of PDEs using Fourier/Galerkin method, will enable us to further our investigation of the problem.

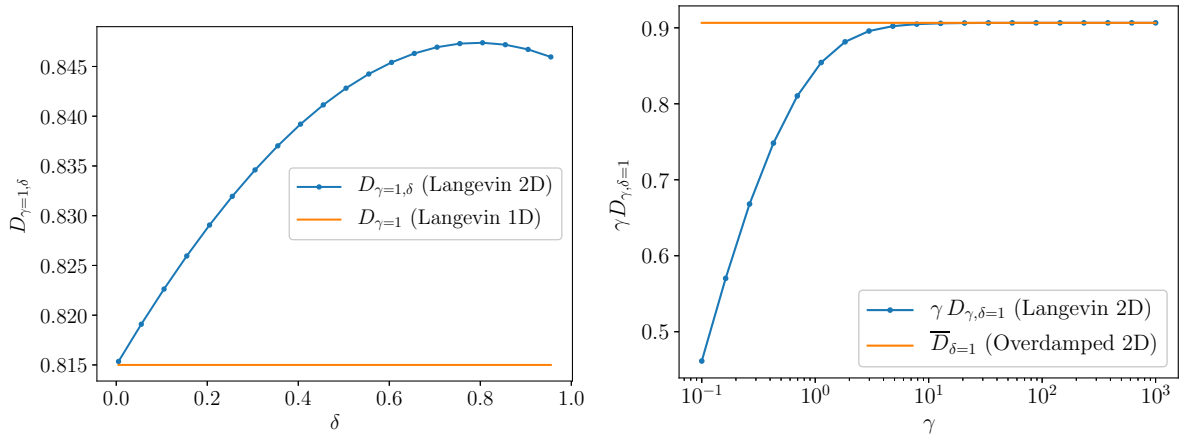


Figure 4.14: *Left*: convergence to the effective diffusion coefficient of the one-dimensional case as $\delta \to 0$. *Right*: convergence to the overdamped limit as $\gamma \to \infty$.

4.A Effective diffusion coefficient for model (4.1a)

Consider Eq. (4.1a) and define $X(t) = \varepsilon\, q(t/\varepsilon^2)$; then

$$\mathrm{d}X_t = -\frac{1}{\varepsilon}\, V'(X_t/\varepsilon)\,\mathrm{d}t + \sqrt{2\,\beta^{-1}}\,\mathrm{d}W'_t, \tag{4.55}$$

where $W'_t = \varepsilon\, W_{t/\varepsilon^2}$ is also a standard Brownian motion. The backward Kolmogorov equation associated with Eq. (4.55) is

$$\frac{\partial u}{\partial t} = -\frac{1}{\varepsilon}\, V'\left(\frac{x}{\varepsilon}\right)\, \frac{\partial u}{\partial x} + \beta^{-1}\, \frac{\partial^2 u}{\partial x^2}. \tag{4.56}$$

We assume that the solution to Eq. (4.56) can be expressed as

$$u(x,t) = u_0(x, x/\varepsilon, t) + \varepsilon\, u_1(x, x/\varepsilon, t) + \cdots. \tag{4.57}$$

We now introduce the auxiliary variable $y = x/\varepsilon$, which will be treated as independent from the variable x; see [175] for a justification of this assumption. Substituting the expansion (4.57) in Eq. (4.56) and gathering terms multiplying equal powers of ε, we obtain the equations

$$\mathcal{O}(\varepsilon^{-2}): \qquad \mathcal{L}_0 u_0 = 0, \tag{4.58a}$$

$$\mathcal{O}(\varepsilon^{-1}): \qquad \mathcal{L}_0 u_1 - V'(y)\, \frac{\partial u_0}{\partial x} + 2\,\beta^{-1}\, \frac{\partial^2 u_0}{\partial x \partial y} = 0, \tag{4.58b}$$

$$\mathcal{O}(1): \qquad \mathcal{L}_0 u_2 - V'(y)\, \frac{\partial u_1}{\partial x} + 2\,\beta^{-1}\, \frac{\partial^2 u_1}{\partial x \partial y} + \beta^{-1}\, \frac{\partial^2 u_0}{\partial x^2} = \frac{\partial u_0}{\partial t}, \tag{4.58c}$$

where $\mathcal{L}_0$ is defined by

$$\mathcal{L}_0 = -\frac{\partial V}{\partial q}(y)\, \frac{\partial}{\partial y} + \beta^{-1}\, \frac{\partial^2}{\partial y^2}.$$

From Eq. (4.58a), we deduce that u_0 depends only on x and t, and from Eq. (4.58b), we obtain $u_1(x,y,t) = \chi(y)\, \partial_x u_0(x,t) + u_1(x,t)$, where χ solves $-\mathcal{L}_0 \chi = -V'$. Finally, from Eq. (4.58c),

we obtain a backward Kolmogorov-type equation for u_0,

$$\frac{\partial u_0}{\partial t} = \beta^{-1} \frac{\partial^2 u_0}{\partial x^2} \int_{\mathbf{R}} \left(1 - \beta V' \chi + 2\chi'\right) \frac{\mathrm{e}^{-V(y)}}{\mathcal{Z}} \, \mathrm{d}y,$$

from which we can read off the effective diffusion coefficient. Taking into account that the function χ satisfies $-V' = -\mathcal{L}_0 \chi$, this expression can be further simplified to Eq. (4.11).

4.B Attempt at proving discrete hypocoercivity in $H^1(\mu)$

In this section we will study the hypocoercivity of the discrete dynamics for the $H^1(\mu)$ norm. Although the reasoning below applies to both the bounded and unbounded cases (with appropriate assumptions on V), we consider for simplicity only the bounded case with the periodic potential $(1/2)\,(1 - \cos(q))$. We also assume that the scaling coefficient for Hermite functions is $\sigma = 1$ and that the inverse temperature is $\beta = 1$. We use the notations Π_N^q, Π_N^p, $\Pi_N^{z_1}$, $\Pi_N^{z_2}$ to denote the projection operators on the finite-dimensional spaces in q, p, z_1, z_2, and Π_N to denote $\Pi_N^q \, \Pi_N^p \, \Pi_N^{z_1} \, \Pi_N^{z_2}$. Similarly to the continuous generator, the discrete generator, viewed as an operator on $V_N := \Pi_N \, H^1(\mathbf{T} \times \mathbf{R}^3, \mu)$, can be decomposed into a symmetric part and an antisymmetric part,

$$\begin{aligned} -\Pi_N \mathcal{L} &= \Pi_N A^* A + \Pi_N B, \\ &= (\Pi_N^{z_2} A)^* \, (\Pi_N^{z_2} A) + \Pi_N B, \\ &=: \hat{A}^* \hat{A} + \hat{B}. \end{aligned}$$

with A and B as in Eq. (4.44). It is useful to think of $\hat{A}$ and $\hat{B}$ as matrices, and of their $L^2\left(\mathbf{T} \times \mathbf{R}^3, \mu\right)$ adjoints as the transpose matrices. Let us also define $\hat{C}_i$, $i = 0, \ldots, 3$ as the operators ∂_{z_2}, $-\partial_{z_1}$, ∂_p, $-\partial_q$ on V_N, respectively. The major difference with the continuous case is that additional terms (in bold below) appear in the expressions of some commutators:

$$[\hat{C}_0, \hat{C}_0^*] = I - (\mathbf{N} + \mathbf{1}) \, \mathbf{\Pi_{N-1}^{z_2 \perp}}, \tag{4.59a}$$

$$[\hat{C}_1, \hat{C}_1^*] = I - (\mathbf{N} + \mathbf{1}) \, \mathbf{\Pi_{N-1}^{z_1 \perp}}, \tag{4.59b}$$

$$[\hat{C}_2, \hat{C}_2^*] = I - (\mathbf{N} + \mathbf{1}) \, \mathbf{\Pi_{N-1}^{p \perp}}, \tag{4.59c}$$

$$[\hat{C}_3, \hat{C}_3^*] = \Pi_N^q \, V''(q), \tag{4.59d}$$

where $\Pi_{N-1}^{z_2 \perp} := 1 - \Pi_{N-1}^{z_2}$. The expression of the commutator of $\hat{C}_3$ with its adjoint can be verified using the definition of the basis functions in the q direction, Eq. (4.40). Using Eq. (4.59), we obtain equations similar to Eqs. (4.45a) to (4.45d).

$$\begin{aligned} &\hat{C}_1 - (\mathbf{N} + \mathbf{1}) \, \mathbf{\Pi_{N-1}^{z_2 \perp}} \mathbf{\hat{C}_1} = \nu^2 \, \alpha^{-1} \, [\hat{C}_0, \hat{B}], \\ &\hat{C}_2 - (\mathbf{N} + \mathbf{1}) \, \mathbf{\Pi_{N-1}^{z_1 \perp}} \mathbf{\hat{C}_2} = \nu \, \gamma^{-1/2} \, [\hat{C}_1, \hat{B}] + \alpha \, \nu^{-1} \, \gamma^{-1/2} \, (\hat{C}_0 - (\mathbf{N} + \mathbf{1}) \, \mathbf{\Pi_{N-1}^{z_1 \perp}} \mathbf{\hat{C}_0}), \\ &\hat{C}_3 - (\mathbf{N} + \mathbf{1}) \, \mathbf{\Pi_{N-1}^{p \perp}} \mathbf{\hat{C}_3} = [\hat{C}_2, \hat{B}] + \gamma^{1/2} \, \nu^{-1} \, (C_1 - (\mathbf{N} + \mathbf{1}) \, \mathbf{\Pi_{N-1}^{p \perp}} \mathbf{\hat{C}_1}), \\ &0 = [\hat{C}_3, \hat{B}] + \Pi_N^q V''(q) \, \hat{C}_2. \end{aligned}$$

The extra terms are such that the assumptions of generic theorems presented in [220] are not satisfied, so a bespoke treatment is necessary. Using the identities

$$\Pi_{N-1}^{z_2\perp}\hat{C}_0 = 0, \quad \Pi_{N-1}^{z_1\perp}\hat{C}_1 = 0, \quad \Pi_{N-1}^{p\perp}\hat{C}_2 = 0, \tag{4.60}$$

and their adjoints, we can check that the discrete versions of Eqs. (4.48a) to (4.48d), Eqs. (4.49a) to (4.49c), and Eqs. (4.50b) and (4.50c) are exactly the same as in the continuous case. A natural manner of bounding the three remaining terms, Eqs. (4.51a) to (4.51c), would be to use the bounds

$$\begin{aligned}
(N+1)\,\Pi_{N-1}^{z_2\perp} &\leq \frac{N+1}{N}\,C_0^*C_0,\\
(N+1)\,\Pi_{N-1}^{z_1\perp} &\leq \frac{N+1}{N}\,C_1^*C_1,\\
(N+1)\,\Pi_{N-1}^{p\perp} &\leq \frac{N+1}{N}\,C_2^*C_2,
\end{aligned}$$

in the sense of positive self-adjoint operators. This approach, however, leads to additional negative terms that cannot be bounded by any of the controllable terms. Coloring in red the terms that we haven't managed to bound and denoting by "…" the expressions in the continuous case, we have

$$\begin{aligned}
&\left\langle \hat{C}_0\hat{B}h, \hat{C}_1h\right\rangle + \left\langle \hat{C}_0h, \hat{C}_1\hat{B}h\right\rangle = \dots\\
&\quad - \alpha\,\nu^{-2}(N+1)\left\langle \Pi_{N-1}^{z_2\perp}\hat{C}_1h, \hat{C}_1h\right\rangle - \sqrt{\gamma}\,\nu^{-1}(N+1)\left\langle \hat{C}_0h, \Pi_{N-1}^{z_1\perp}\hat{C}_2h\right\rangle\\
&\quad + \alpha\,\nu^{-2}\,(N+1)\left\langle \Pi_{N-1}^{z_1\perp}\hat{C}_0h, \hat{C}_0h\right\rangle,\\
&\left\langle \hat{C}_1\hat{B}h, \hat{C}_2h\right\rangle + \left\langle \hat{C}_1h, \hat{C}_2\hat{B}h\right\rangle = \dots\\
&\quad - \gamma^{1/2}\,\nu^{-1}(N+1)\left\langle \Pi_{N-1}^{z_1\perp}\hat{C}_2h, \hat{C}_2h\right\rangle + \alpha\,\nu^{-2}\,(N+1)\left\langle \Pi_{N-1}^{z_1\perp}\hat{C}_0h, \hat{C}_2h\right\rangle\\
&\quad + (N+1)\left\langle \hat{C}_1h, \Pi_{N-1}^{p\perp}\hat{C}_3h\right\rangle + \gamma^{1/2}\,\nu^{-1}\,(N+1)\left\langle \Pi_{N-1}^{p\perp}\hat{C}_1h, \hat{C}_1h\right\rangle,\\
&\left\langle \hat{C}_2\hat{B}h, \hat{C}_3h\right\rangle + \left\langle \hat{C}_2h, \hat{C}_3\hat{B}h\right\rangle = \dots\\
&\quad - (N+1)\left\langle \Pi_{N-1}^{p\perp}\hat{C}_3h, \hat{C}_3h\right\rangle + \gamma^{1/2}\,\nu^{-1}\,(N+1)\left\langle \Pi_{N-1}^{p\perp}\hat{C}_1h, \hat{C}_3h\right\rangle.
\end{aligned}$$

Although this approach does not seem to work for the models GL1 and GL2, it does enable concluding discrete hypocoercivity for the Langevin dynamics; in that case all the terms arising from the discretization can be bounded from above by terms of the type $\|C_0C_i\|^2$. This was proved by Julien Roussel in unpublished notes extending his work in [190].

Chapter 5

A linear, second-order, fully adaptive finite element method for phase-field modeling of wetting phenomena

Abstract

We propose a new numerical method to solve the Cahn–Hilliard equation coupled with non-linear wetting boundary conditions. We show that the method is mass-conservative and that the discrete solution satisfies a discrete energy law similar to the one satisfied by the exact solution. We perform several tests inspired by realistic situations to verify the accuracy and performance of the method: wetting of a chemically heterogeneous substrate in three dimensions, wetting-driven nucleation in a complex two-dimensional domain and three-dimensional diffusion through a porous medium.

5.1 Introduction . . . 190
5.2 Phase-field model for wetting phenomena . . . 192
5.2.1 Existence of a solution . . . 194
5.3 Numerical method . . . 195
5.3.1 OD2-W scheme . . . 195
5.4 Numerical results . . . 200
5.4.1 Equilibrium contact angle . . . 200
5.4.2 Convergence of the method . . . 201
5.4.3 Time-adaptation scheme . . . 203
5.4.4 Wetting in complex geometries and with heterogeneous substrates . . . 205
5.5 Extension to the Cahn–Hilliard–Navier–Stokes system . . . 212
5.6 Conclusions . . . 213
5.A Proof of Theorem 5.1 . . . 214

5.1 Introduction

Capillarity and wetting phenomena, driven primarily by interfacial forces, are ubiquitous in a wide spectrum of natural phenomena and technological applications. Examples range from the wetting of plant leaves by rainwater and insects walking on water to coating processes, inkjet printing, oil recovery and microfluidic devices; for reviews, see e.g. [52, 27]. From an historical point of view, two of the concepts essential to the understanding of capillarity and wetting were introduced and studied already in 1805: these are the Laplace pressure [136] and the Young–Dupré contact angle [224]. Later, following the work of Plateau on soap films [181], Poincaré [182] linked interfacial phenomena with the theory of minimal surfaces.

Wetting phenomena typically involve a fluid-fluid interface advancing or receding on a solid substrate and a contact line formed at the intersection between the interface and the substrate. The wetting properties of the substrate determine to a large extent the behavior of the fluids in the contact-line region, and in particular the contact angle at the three-phase conjunction, defined as the angle between the fluid-fluid interface and the tangent plane at the substrate. At equilibrium, this is precisely the Young–Dupré angle. When one of the two fluids moves against the other, the contact angle becomes a dynamic quantity, and when the problem is formulated in the framework of conventional hydrodynamics, the contact line motion relative to the solid boundary results in the notorious stress singularity there, as first noted in the pioneering studies by Moffat [157] and Huh and Scriven [117]. Since then there have been numerous analyses and discussions of the singularity over the years, see e.g. [214, 114, 48] and also recent studies in [201, 199] (with the latter one revisiting the classical Cox–Hocking matched asymptotic analysis and providing a correction to it). Recently, it was shown [163] that mesoscopic approaches such as dynamic density functional theory (DDFT) can fix this singularity behavior.

A popular model for interface dynamics is the Cahn–Hilliard (CH) equation [35, 36], which belongs to the class of phase-field and diffuse interface models. Originally proposed to model spinodal decomposition, the mechanism by which a binary mixture can separate to form two coexisting phases due to, e.g., a change of temperature [36], it has been used in a wide spectrum of different contexts since, such as solidification phenomena [46] and Saffman–Taylor instabilities in Hele–Shaw flows [110]. To account for wetting phenomena and contact lines on solid boundaries, the CH equation can be coupled to a wall boundary condition [34]. Such CH model has been employed successfully in various situations, including microfluidic devices [53, 54, 183, 222], flow in porous media [26], rheological systems [29], and patterning of thin polymer films [130]. Other potential applications include micro-separators [192], fuel cells [19] and CPU chip cooling based on electro-wetting [162]. Many of these applications are characterized by the presence of chemically heterogeneous substrates and/or complex geometries, which make their numerical simulation challenging.

The form of the wetting boundary condition is dictated by the form of the wall free energy. For liquid-gas problems linear forms have been adopted, e.g. in the pioneering study by Seppecher [194] and in [33, 223]. But a cubic is the lowest-order polynomial required such that the wall free energy can be minimized for the bulk densities, and it prevents the formation of boundary layers on the wall. Cubic forms have been adopted for binary fluid problems, e.g. in [121, 226], but

also for liquid-gas ones, see [202, 203]. The latter studies, in particular, showed asymptotically that a CH model can alleviate the contact line discontinuity without any additional physics (and at the same time completing but also correcting Seppecher's work). The detailed asymptotic analysis of the unification of binary-fluid CH models can be found in [200].

Various approaches have been proposed in the literature for the numerical solution of the CH equation. Because of the high order of the equation and its multiscale features (scale separation between interface size and the characteristic length), most existing time-stepping schemes are implicit or semi-implicit. Several of these schemes aim to satisfy discrete mass and energy laws in agreement with the underlying continuum model. Discretization in space can be achieved using finite-difference methods [81, 132], finite element methods [17, 67, 212], spectral methods [221], or pseudospectral methods [164]. In addition, the computation time can be reduced by applying adaptive mesh refinement [227, 205] and time-step adaptation [99].

Among the several linear schemes for the CH equation with homogeneous Neumann boundary conditions introduced in [98], the authors showed by means of numerical experiments that their second-order optimal dissipation scheme, referred to as OD2, is the most accurate and the one introducing the least numerical dissipation. In this work, we outline a numerical scheme that extends and appropriately generalizes OD2 as follows: (a) it includes a non-linear wetting boundary condition; (b) it adopts an efficient energy-based time-step adaptation strategy. In contrast with the time-adaptation scheme introduced in [99], where the time step is adapted to limit numerical dissipation, we base the time-step adaptation directly on the variation of free energy. With this method we are able to solve the CH system efficiently and systematically to capture wetting phenomena in both two- and three-dimensional (2D and 3D, respectively) settings, and in a wide range of situations, including confinement with complex geometry, chemical and topographical heterogeneities, or both.

Like the OD2 scheme on which it was based, the time-stepping scheme we propose is semi-implicit and linear. We show that it is mass-conservative and satisfies a discrete free-energy law with a numerical dissipation term of order 2 in time. Space discretization is achieved using a finite element method, leading to an unsymmetrical sparse linear system to solve at each iteration. In addition to adapting the time step as mentioned above, with the aim of increasing the resolution in time during fast phenomena, we use a mesh refinement strategy to capture interfaces precisely.

To test the efficiency of the proposed numerical scheme we consider several wetting problems as test cases. We first study relaxation towards equilibrium in two situations: the spreading of a sessile droplet and the coalescence of two sessile droplets on a flat, chemically homogeneous substrate. We then consider two-component systems in complex geometries delimited by chemically heterogeneous substrates in both 2D and 3D.

In Section 5.2, we introduce the CH system and the non-linear wetting boundary condition. In Section 5.3, we outline our numerical scheme and prove the associated conservation properties. In Section 5.4, we present the results of several numerical experiments. Conclusions and perspectives for future work are offered in Section 5.6.

5.2 Phase-field model for wetting phenomena

Throughout this study, $\Omega \subset \mathbf{R}^d$ denotes a d-dimensional domain, $\partial\Omega$ denotes its boundary with outward unit normal vector $\mathbf{n}$, Γ_S is the solid substrate and $\Gamma_G = \partial\Omega \setminus \Gamma_S$. The CH system we use to describe the dynamics of two immiscible fluids in contact with a solid substrate, is a free-energy-based model. The starting point is the introduction of a locally conserved field, denoted by $\phi : \Omega \to \mathbf{R}$, that plays the role of an order-parameter: two equilibrium values, say $+1$ and -1, represent the pure phases, and the interface is conventionally located at the points where $\phi = 0$ [35, 36]. We consider systems with a free energy given by

$$\begin{aligned} E(\phi) &:= E_m(\phi) + E_w(\phi) \\ &:= \int_\Omega \left(\frac{1}{\varepsilon} F_m(\phi) + \varepsilon \frac{|\nabla\phi|^2}{2} \right) \mathrm{d}\Omega + \int_{\partial\Omega} F_w(\phi)\, \mathrm{d}\sigma, \end{aligned}$$

where the two terms, E_m and E_w, represent the mixing and wall components of the free energy, respectively. Here $F_m(\phi) = \frac{1}{4}(\phi^2 - 1)^2$ and F_w is taken to be a cubic polynomial, following e.g. [202, 203]:

$$F_w(\phi) = \frac{\sqrt{2}}{2} \cos\theta(\mathbf{x}) \left(\frac{\phi^3}{3} - \phi \right), \tag{5.1}$$

where $\theta = \theta(\mathbf{x})$ is the equilibrium contact angle, which can depend on the spatial position $\mathbf{x}$. From the expression of the free energy, we calculate that, for a sufficiently smooth function $\psi : \Omega \to \mathbf{R}$:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} E(\phi + \alpha\,\psi)\big|_{\alpha=0} = \int_\Omega \left(\frac{1}{\varepsilon} f_m(\phi) - \varepsilon \Delta\phi \right) \psi \, \mathrm{d}\Omega + \int_{\partial\Omega} (f_w(\phi) + \varepsilon \nabla\phi \cdot \mathbf{n})\, \psi \, \mathrm{d}\sigma,$$

with $f_m = F_m'$ and $f_w = F_w'$, so the chemical potential is equal to

$$\mu := \frac{\delta E}{\delta \phi} = \frac{1}{\varepsilon} f_m(\phi) - \varepsilon\, \Delta\phi,$$

and the natural boundary condition associated with the surface energy is

$$\varepsilon \nabla\phi \cdot \mathbf{n} = -f_w(\phi) = \frac{\sqrt{2}}{2} \cos\theta(\mathbf{x})(1 - \phi^2). \tag{5.2}$$

We assume that the dynamics of the system is governed by the CH equation,

$$\frac{\partial \phi}{\partial t} = \nabla \cdot (b(\mathbf{x})\, \nabla\mu), \tag{5.3}$$

where $b(\mathbf{x})$ is a mobility parameter, assumed to be uniform hereafter. This leads to the following mass-conservation property:

$$\frac{\mathrm{d}}{\mathrm{d}t} M(\phi) := \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \phi \, \mathrm{d}\Omega = \int_{\partial\Omega} b\, \nabla\mu \cdot \mathbf{n} \, \mathrm{d}\sigma, \tag{5.4}$$

so the mass flux at the boundary can be specified using the condition $b\nabla\mu \cdot \mathbf{n} = \dot{m}(\mathbf{x})$, where $\dot{m}(\mathbf{x})$ is the desired mass flux. In particular, we will set $\dot{m}(\mathbf{x}) = 0$ at the solid boundary, Γ_S. In

summary, the equations we are solving in this study are the following:

$$\frac{\partial \phi}{\partial t} = \nabla \cdot (b \nabla \mu), \tag{5.5a}$$

$$\mu = \frac{1}{\varepsilon} f_m(\phi) - \varepsilon \Delta \phi \qquad \text{for } x \in \Omega,\ t \in (0, T], \tag{5.5b}$$

$$\varepsilon \nabla \phi \cdot \mathbf{n} = -f_w(\phi), \tag{5.5c}$$

$$b \nabla \mu \cdot \mathbf{n} = \dot{m}(\mathbf{x}) \qquad \text{for } x \in \partial\Omega,\ t \in (0, T]. \tag{5.5d}$$

In addition to the conservation of mass, Eqs. (5.5a) to (5.5d) imply the following energy-conservation law, involving the phase field and the chemical potential:

$$\frac{\mathrm{d}}{\mathrm{d}t} E(\phi(t)) = -\|\sqrt{b}\, \nabla \mu\|^2_{L^2(\Omega)} + \int_{\partial\Omega} \dot{m}\, \mu \, \mathrm{d}\sigma. \tag{5.6}$$

An advantage of the cubic surface energy (5.1) over other surface energy formulations (see [116] for a review of wetting boundary conditions for binary fluids) is that the well-known hyperbolic tangent profile is an equilibrium solution in more than 1 dimension. Specifically, the function

$$\phi(\mathbf{x}) = \tanh\left(\frac{\mathbf{x} \cdot \mathbf{u}}{\sqrt{2}\,\varepsilon}\right), \quad \text{where } \mathbf{u} = (\pm \sin\theta, \cos\theta)^T \tag{5.7}$$

is solution to the CH equation posed in the half plane $\{y \geq 0\}$ with the boundary condition (5.1) at $\{y = 0\}$ and constant $\theta(\mathbf{x}) = \theta$. A schematic representation of this solution and the corresponding fluid-fluid interface is given in Fig. 5.1.

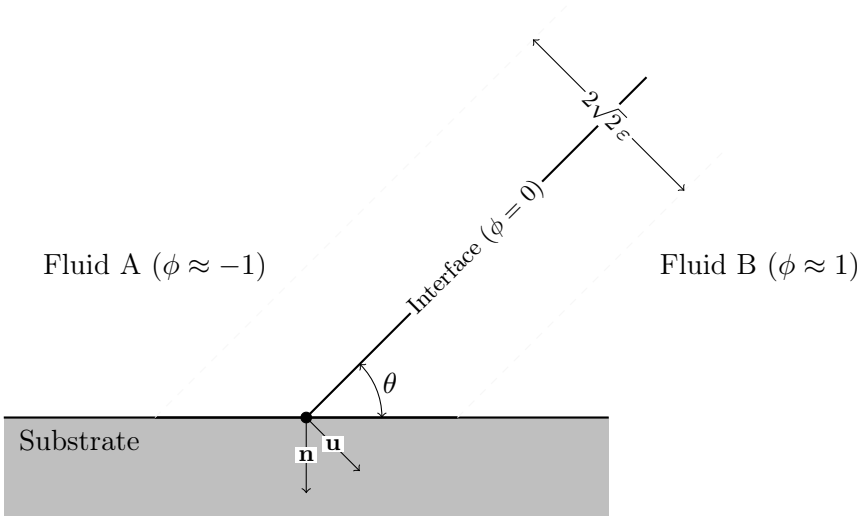


Figure 5.1: Schematic of the profile geometry of a fluid-fluid interface intersecting a solid boundary and illustration of the stationary solution (5.7).

A drawback of the cubic wall energy (5.1) is that the conservation of energy no longer seems to imply stability bounds for the solution, making it impossible to use the tools traditionally employed (see e.g. [68]) to prove the existence of a solution. Indeed, an application of the trace

inequality gives only that, under appropriate regularity assumptions on ϕ:

$$\begin{aligned}\|\phi^3\|_{L^1(\partial\Omega)} &\leq C\left(\|\phi^3\|_{L^1(\Omega)} + \|\nabla(\phi^3)\|_{L^1(\Omega)}\right)\\ &= C\left(\|\phi^3\|_{L^1(\Omega)} + 3\|\phi^2\,\nabla\phi\|_{L^1(\Omega)}\right)\\ &\leq C\left(\|\phi^3\|_{L^1(\Omega)} + \frac{3}{2\alpha}\|\phi^4\|_{L^1(\Omega)} + \frac{3\alpha}{2}\|\nabla\phi\|^2_{L^2(\Omega)}\right) \quad \forall\alpha>0,\end{aligned}$$

where we used Hölder's inequality and Young's inequality with a parameter. Therefore, the wall energy cannot be controlled by the mixing energy for arbitrary domains. This issue can be remedied by a simple modification of the wall energy outside of the physical range $[-1;1]$; instead of (5.1), we consider the following wall energy:

$$F_w^*(\phi) = \frac{\sqrt{2}}{2}\cos\theta(\mathbf{x}) \times \begin{cases} \left(\frac{2}{3} - (\phi+1)^2\right) & \text{if } \phi < -1;\\ \left(\frac{\phi^3-3\phi}{3}\right) & \text{if } \phi\in[-1,1];\\ \left(-\frac{2}{3} + (\phi-1)^2\right) & \text{if } \phi > 1,\end{cases} \tag{5.9}$$

This function is such that $F_w^*(\phi) = F_w(\phi)$ for $\phi\in[-1,1]$, $F_w^*\in C^2(\mathbf{R})$, and $(F_w^*)''$ is absolutely continuous, which makes it possible to prove the second order convergence of our time-stepping scheme, see Section 5.3. Another possibility would have been to choose constant values for F_w^* outside of the interval $[-1,1]$, but this would have lead to F_w^* being only $C^1(\mathbf{R})$, making it more difficult to show second order convergence theoretically. The weak formulation of Eqs. (5.5a) to (5.5d) with the modified wall energy (5.9) is as follows: find (ϕ,μ) such that

$$\phi\in L^\infty\left(0,T;H^1(\Omega)\right), \quad \frac{\partial\phi}{\partial t}\in L^2\left(0,T;(H^1(\Omega))'\right), \quad \mu\in L^2\left(0,T;H^1(\Omega)\right),$$

and the following variational formulation is satisfied:

$$\langle\partial_t\phi,\psi\rangle + \langle b\,\nabla\mu,\nabla\psi\rangle = \langle\dot m,\psi\rangle_{\partial\Omega} \qquad \forall\psi\in H^1(\Omega) \text{ and a.e. } t, \tag{5.10a}$$

$$\langle\mu,\nu\rangle = \varepsilon\,\langle\nabla\phi,\nabla\nu\rangle + \frac{1}{\varepsilon}\langle f_m(\phi),\nu\rangle + \langle f_w^*(\phi),\nu\rangle_{\partial\Omega} \qquad \forall\nu\in H^1(\Omega) \text{ and a.e. } t, \tag{5.10b}$$

with $f_w^* := (F_w^*)'$ and where $\langle\cdot,\cdot\rangle$, $(\cdot,\cdot)$ and $(\cdot,\cdot)_{\partial\Omega}$ denote respectively the duality pairing between $(H^1(\Omega))'$ and $H^1(\Omega)$, the standard inner product in $L^2(\Omega)$, and the standard inner product in $L^2(\partial\Omega)$. For simplicity of notations, the symbols F_w, f_w and E will refer in the rest of this chapter to F_w^*, f_w^*, and $E_m + \int_{\partial\Omega} F_w^*\,\mathrm{d}\sigma$, respectively.

5.2.1 Existence of a solution

We can show the following existence result for the weak formulation of the Cahn–Hilliard system with the modified boundary condition presented above, under appropriate regularity assumptions for the initial condition and the mass flux $\dot m$.

Theorem 5.1. *Assume that $\phi_0\in H^1(\Omega)$ and $\dot m\in C([0,T];L^2(\partial\Omega))$. Then there exists a pair of functions (ϕ,μ), with*

1. $\phi\in L^\infty(0,T;H^1(\Omega))\bigcap C([0,T];L^2(\Omega))$,

2. $\partial_t \phi \in L^2\left(0,T;(H^1(\Omega))'\right)$,

3. $\phi(0) = \phi_0$,

4. $\mu \in L^2\left(0,T;H^1(\Omega)\right)$,

that solve the variational formulation Eqs. (5.10a) *and* (5.10b).

Proof. See Section 5.A. □

5.3 Numerical method

In this section we introduce a new time-stepping scheme to solve the CH equation (5.3) with the non-linear wetting boundary condition (5.2), which is a generalization of the optimal dissipation scheme of order 2, OD2, developed in [98]. We decided to extend this particular scheme because the authors of [98] showed that, among all the linear schemes they proposed, it is the most accurate and the least dissipative. And in selected test cases, they showed that for a large enough time step, it is the only scheme that leads to the correct equilibrium solution. We refer to our scheme as OD2-W, with W denoting wetting, and show that it leads to a consistent discrete energy law.

We also develop a new adaptive time-stepping strategy which, combined with adaptation in space, leads to a fully adaptive finite element method. An excellent introduction to the finite element method and corresponding mixed formulations can be found in [43] and to mesh generation and adaptive refinement in [77].

5.3.1 OD2-W scheme

In this section, we assume for simplicity that $\dot{m} = 0$ and that θ is uniform on $\partial\Omega$. We denote by Δt the time step, and by ϕ^n and $\mu^{n+1/\alpha}$ the numerical approximations of ϕ and μ at times t^n and $t^n + \frac{1}{\alpha}\Delta t$, respectively. To define a discretization in time of the CH system appropriate for wetting phenomena, we follow the approach proposed in [98] to design an optimal dissipation scheme, and consider the following generic implicit-explicit numerical scheme: given $\phi^n \in H^1(\Omega)$, find $(\phi^{n+1}, \mu^{n+\frac{1}{\alpha}}) \in H^1(\Omega) \times H^1(\Omega)$ such that,

$$\left\langle \delta_t \phi^{n+1}, \psi \right\rangle + \left\langle b\, \nabla \mu^{n+\frac{1}{\alpha}}, \nabla \psi \right\rangle = 0 \qquad \forall \psi \in H^1(\Omega), \tag{5.11a}$$

$$\begin{aligned} \left\langle \mu^{n+\frac{1}{\alpha}}, \nu \right\rangle &= \varepsilon \left\langle \nabla \phi^{n+\frac{1}{\alpha}+\beta}, \nabla \nu \right\rangle \\ &\quad + \frac{1}{\varepsilon} \left\langle \hat{f}_m(\phi^n, \phi^{n+1}), \nu \right\rangle + \left\langle \hat{f}_w(\phi^n, \phi^{n+1}), \nu \right\rangle_{\partial\Omega} \qquad \forall \nu \in H^1(\Omega). \end{aligned} \tag{5.11b}$$

In these expressions, $\hat{f}_m, \hat{f}_w$ are functions to be specified, linear in their second argument. The parameter $\alpha \in \{1, 2\}$ determines the accuracy of the numerical scheme, and the parameter $\beta \in [0, 1 - 1/\alpha]$ controls the numerical diffusion. The function $\phi^{n+\frac{1}{\alpha}+\beta}$ is defined by linear interpolation between ϕ^n and ϕ^{n+1},

$$\phi^{n+\frac{1}{\alpha}+\beta} := \left(1 - \frac{1}{\alpha} - \beta\right) \phi^n + \left(\frac{1}{\alpha} + \beta\right) \phi^{n+1},$$

and $\delta_t\phi^{n+1}$ is the approximation of the time derivative of ϕ given by

$$\frac{\phi^{n+1}-\phi^n}{\Delta t}.$$

In most numerical experiments presented in this chapter, we consider the case $(\alpha,\beta)=(2,0)$ (OD2-W), but we note that other usual choices include $(\alpha,\beta)=(1,0)$ (OD1-W) and $(\alpha,\beta)=(2,\mathcal{O}(\Delta t))$ (OD2mod-W). By taking $\psi=\mu^{n+\frac{1}{\alpha}}$ and $\nu=\delta_t\phi^{n+1}$ in (5.11), we obtain

$$\frac{E(\phi^{n+1})-E(\phi^n)}{\Delta t}=-\|\sqrt{b}\,\nabla\mu^{n+\frac{1}{\alpha}}\|^2_{L^2(\Omega)}-ND(\phi^n,\phi^{n+1}),\quad n=0,1,\dots, \tag{5.12}$$

where $ND(\phi^n,\phi^{n+1})$, representing the non-physical numerical dissipation introduced by the time-stepping scheme, can be broken down in three parts:

$$ND(\phi^n,\phi^{n+1})=\varepsilon ND_{\text{philic}}(\phi^n,\phi^{n+1})+\frac{1}{\varepsilon}ND_{\text{phobic}}(\phi^n,\phi^{n+1})+ND_{\text{wall}}(\phi^n,\phi^{n+1}),$$

with

$$ND_{\text{philic}}(\phi^n,\phi^{n+1})=\Delta t\int_\Omega\left[\frac{1}{\alpha}-\frac{1}{2}+\beta\right]\left|\nabla\delta_t\phi^{n+1}\right|^2\,\mathrm{d}\Omega, \tag{5.13a}$$

$$ND_{\text{phobic}}(\phi^n,\phi^{n+1})=\int_\Omega\left(\hat{f}_m(\phi^n,\phi^{n+1})\,\delta_t\phi^{n+1}-\frac{1}{\Delta t}(F_m(\phi^{n+1})-F_m(\phi^n))\right)\mathrm{d}\Omega, \tag{5.13b}$$

$$ND_{\text{wall}}(\phi^n,\phi^{n+1})=\int_{\partial\Omega}\left(\hat{f}_w(\phi^n,\phi^{n+1})\,\delta_t\phi^{n+1}-\frac{1}{\Delta t}(F_w(\phi^{n+1})-F_w(\phi^n))\right)\mathrm{d}\sigma. \tag{5.13c}$$

Notice that the philic dissipation is always nonnegative, with $ND_{\text{philic}}(\cdot,\cdot)=0$ if $(\alpha,\beta)=(2,0)$ (OD2-W), $ND_{\text{philic}}(\cdot,\cdot)=\mathcal{O}(\Delta t^2)$ if $(\alpha,\beta)=(2,\mathcal{O}(\Delta t))$ (OD2mod-W), and $ND_{\text{philic}}(\cdot,\cdot)=\mathcal{O}(\Delta t)$ if $(\alpha,\beta)=(1,0)$ (OD1-W). The two other terms can be expanded using Taylor's formula, taking into account that F_m is a polynomial of degree 4 and using the integral form of the remainder:

$$\begin{aligned}ND_{\text{phobic}}(\phi^n,\phi^{n+1})=\int_\Omega\delta_t\phi^{n+1}\Big(&\hat{f}_m(\phi^n,\phi^{n+1})-f_m(\phi^n)-\frac{1}{2}\,\Delta t\,f_m'(\phi^n)\,\delta_t\phi^{n+1}\\&-\frac{1}{6}\,\Delta t^2\,f_m''(\phi^n)\,(\delta_t\phi^{n+1})^2-\frac{1}{24}\,\Delta t^3\,f_m'''(\phi^n)\,(\delta_t\phi^{n+1})^3\Big)\,\mathrm{d}\Omega,\end{aligned} \tag{5.14a}$$

$$\begin{aligned}ND_{\text{wall}}(\phi^n,\phi^{n+1})=\int_{\partial\Omega}\delta_t\phi^{n+1}\Big(&\hat{f}_w(\phi^n,\phi^{n+1})-f_w(\phi^n)-\frac{1}{2}\,\Delta t\,f_w'(\phi^n)\,\delta_t\phi^{n+1}\Big)\\&-\frac{1}{2\,\Delta t}\int_{\phi^n}^{\phi^{n+1}}f_w''(\phi)\,(\phi-\phi^n)^2\,\mathrm{d}\phi\,\mathrm{d}\sigma.\end{aligned} \tag{5.14b}$$

This suggests the following choices for the functions $\hat{f}_m$ and $\hat{f}_w$:

$$\begin{aligned}\hat{f}_m(\phi^n,\phi^{n+1}) &= f_m(\phi^n) + \frac{1}{2}\,\Delta t\, f_m'(\phi^n)\,\delta_t\phi^{n+1}\\ &= \left[\frac{3}{2}(\phi^n)^2\phi^{n+1} - \frac{1}{2}(\phi^n)^3 - \frac{\phi^n+\phi^{n+1}}{2}\right], \qquad (5.15a)\\ \hat{f}_w(\phi^n,\phi^{n+1}) &= f_w(\phi^n) + \frac{1}{2}\,\Delta t\, f_w'(\phi^n)\,\delta_t\phi^{n+1}\\ &= -\frac{\sqrt{2}}{2}\cos(\theta)\times\begin{cases}2+\phi^n+\phi^{n+1} & \text{if } \phi^n<-1;\\ 1-\phi^n\phi^{n+1} & \text{if } \phi^n\in[-1,1]\\ 2-\phi^n-\phi^{n+1} & \text{if } \phi^n>1,\end{cases} \qquad (5.15b)\\ &= -\frac{\sqrt{2}}{2}\cos(\theta)\times\left(1+(1-\phi^n)\wedge 0+(1+\phi^n)\wedge 0-(-1\vee\phi^n\wedge 1)\,\phi^{n+1}\right),\end{aligned}$$

where the last expression is convenient for programming purposes. We note that this methodology to derive a second-order scheme can be applied *mutatis mutandis* when using the unmodified wall energy (5.1), although we haven't been able to prove that the weak formulation admits a solution in that case. Doing so leads to $\hat{f}_w(\phi^n,\phi^{n+1}) = -(\sqrt{2}/2)\,\cos\theta\,(1-\phi^n\,\phi^{n+1})$, which coincides with (5.15b) when $\phi^n\in[-1,1]$. In either case, we have the following property:

Property 5.1. Assume that $\alpha = 2$ and $\beta = 0$. Then the numerical dissipation term in Eq. (5.12) is such that

$$|ND(\phi(t^n),\phi(t^{n+1}))| \le C\Delta t^2,$$

with $C := \left(C_1\,\|\phi\|_{C([0,T],L^\infty(\Omega))}\,\|\partial_t\phi\|^3_{C([0,T],L^3(\Omega))} + C_2\,\|\partial_t\phi\,\|^3_{C([0,T],L^3(\partial\Omega))}\right)$, provided that all the terms in the definition of C are well-defined.

Proof. In [98], the authors show that:

$$\begin{aligned}&ND_{\text{philic}}(\cdot,\cdot) = ND_{\text{philic}}(\phi^n,\phi^{n+1}) = ND_{\text{philic}}(\phi(t^n),\phi(t^{n+1})) = 0;\\ &\left|ND_{\text{phobic}}(\phi(t^n),\phi(t^{n+1}))\right| \le \Delta t^2\,\left(C_1\,\|\phi\|_{C([0,T],L^\infty(\Omega))}\,\|\partial_t\phi\|^3_{C([0,T],L^3(\Omega))}\right).\end{aligned}$$

For the wall term, we obtain from Eqs. (5.14b) and (5.15b):

$$\begin{aligned}|ND_{\text{wall}}(\phi(t_n),\phi(t_{n+1})| &\le C_2\,\Delta t^2\int_{\partial\Omega}\left|\delta_t\phi^{n+1}\right|^3\mathrm{d}\sigma\\ &\le C_2\,\Delta t^2\,\|\partial_t\phi\|^3_{C([0,T],L^3(\partial\Omega))}.\end{aligned}$$

□

In addition to the energy law (5.12), the numerical scheme (5.11) satisfies a discrete version of the conservation law (5.4) presented in Section 5.2, which can be seen by choosing $\psi = 1$ in Eq. (5.11a).

Property 5.2. The numerical solution satisfies the following mass conservation law:

$$\int_\Omega \phi^n\,\mathrm{d}\Omega = \int_\Omega \phi^0\,\mathrm{d}\Omega \quad \text{for } n = 0,1,2,\ldots$$

Space discretization and adaptive mesh refinement

Our approach for mesh adaptation is based on a method proposed in [107], and implemented through the FreeFem++ functions *adaptmesh* (in 2D) and *mshmet* (in 3D). The idea of the method is to define a metric on the computational domain based on the solution at the current time step, and to use for the next time step a mesh that is uniform in that metric. The metric we consider corresponds to the following metric tensor, depending only on the phase field ϕ:

$$G(\mathbf{x}) = R(\mathbf{x})\,\mathrm{diag}(\tilde{\lambda}_i(\mathbf{x}))\,R(\mathbf{x})^T, \quad \tilde{\lambda}_i(\mathbf{x}) = \min\left(\max\left(\frac{1}{\gamma}|\lambda_i(\mathbf{x})|, \frac{1}{h_{\max}^2}\right), \frac{1}{h_{\min}^2}\right),$$

where $(\lambda_i(\mathbf{x}))_{i=1}^d$ are the eigenvalues of the Hessian of ϕ at $\mathbf{x}$, $R(\mathbf{x})$ is the matrix containing the associated orthonormal eigenvectors, and $\gamma > 0$ is a parameter controlling the interpolation error. A standard algorithm of Delaunay type is used to generate a mesh that is equilateral and uniform with characteristic length 1 in that metric. This mesh definition ensures that the interpolation error of the phase field is roughly equi-distributed over the parts of the domain where $h_{\max}^{-2} \leq \frac{1}{\gamma}\max_{i\in\{1,2,\dots,d\}}|\lambda_i| \leq h_{\min}^{-2}$.

In most of the simulations presented in the next section, we set $h_{\min}$ to a value lower than or equal to $\varepsilon/5$, to ensure that enough mesh points are available for the discretization of the interface region in its normal direction, and $h_{\max}$ to a value small enough that a good approximation of the chemical potential is possible. For 3D simulations, however, choosing $h_{\min} \leq \varepsilon/5$ when ε is of the order of 0.01 leads to a prohibitive computational cost; in these cases we have thus used a less precise value, as specified in the relevant sections.

For a given mesh $\mathcal{T} = \bigcup_{i=1}^{N_T} T_i$, we use the standard finite element space

$$V_h = \{\phi \in C(\Omega) : \phi_{|T_i} \in P_\rho \text{ for } i = 1, \dots, N_T\},$$

with P_ρ the space of polynomials of degree ρ. In the numerical experiments below, we used both quadratic elements ($\rho = 2$) and linear ones ($\rho = 1$). Space discretization is achieved by replacing $H^1(\Omega)$ by V_h in the variational formulation (5.11), leading to a sparse unsymmetric linear system at each iteration.

Time step adaptation

Here we assume that $\dot{m} = 0$ in the boundary condition (5.5d). From Eqs. (5.4) and (5.6), this implies that $M(\phi)$ is constant in time and $E(\phi)$ decreases. Numerical exploration suggests that large free-energy variations are usually caused by topological changes of interfaces, corresponding to physical phenomena such as the coalescence of droplets. Since capturing such phenomena precisely is crucial to the accuracy of the solution, we propose an adaptive strategy that aims to limit the variation of free energy at each time step. We adapt the time step based on the dissipation of free energy,

$$\Delta E^{n+1} := -\Delta t^{n+1}\, b\, \|\nabla \mu^{n+\frac{1}{\alpha}}\|_{L^2}^2,$$

which is equal to $E(\phi^{n+1}) - E(\phi^n)$ up to numerical dissipation. Here $\Delta t^{n+1} := t^{n+1} - t^n$. Six parameters enter in our time-adaptation scheme:

- $\Delta t_{\min}, \Delta t_{\max}$: the time steps below which we stop refining and beyond which we stop coarsening, respectively.
- $\Delta E_{\min}$: the variation of free energy below which we increase the time step at the next iteration.
- $\Delta E_{\max}$: the variation of free energy beyond which we refine the time step and recalculate the numerical solution.
- $f > 1$: the factor by which the time step is multiplied or divided at each adaptation.
- $0 < \zeta \ll 1$: a small parameter controlling the maximum amount by which the free energy is allowed to increase at each time step. In the numerical experiments presented in Section 5.4, we always take $\zeta = 0.01$.

Algorithm 1: Time step adaptation

Data: $\Delta t_{\min}, \Delta t_{\max}, \Delta E_{\min}, \Delta E_{\max}, f, \phi^n, \Delta t^{n+1}$

Compute a solution (ϕ^*, μ^*) of (5.11) using time step Δt^{n+1} ;

Compute $|\Delta^* E| := \Delta t^{n+1}\, b\, \|\nabla \mu^*\|^2_{L^2}$;

if $(|\Delta^* E| > \Delta E_{\max} \wedge \Delta t^n > \Delta t_{\min}) \vee (E(\phi^*) - E(\phi^n) > \zeta\, \Delta E_{\max})$ **then**

  Set $\Delta t^{n+1} = \frac{\Delta t^{n+1}}{f}$ and go back to 1;

else

  $\phi^{n+1} = \phi^*$;

  if $(|\Delta^* E| < \Delta E_{\min} \wedge \Delta t^{n+1} < \Delta t_{\max})$ **then**

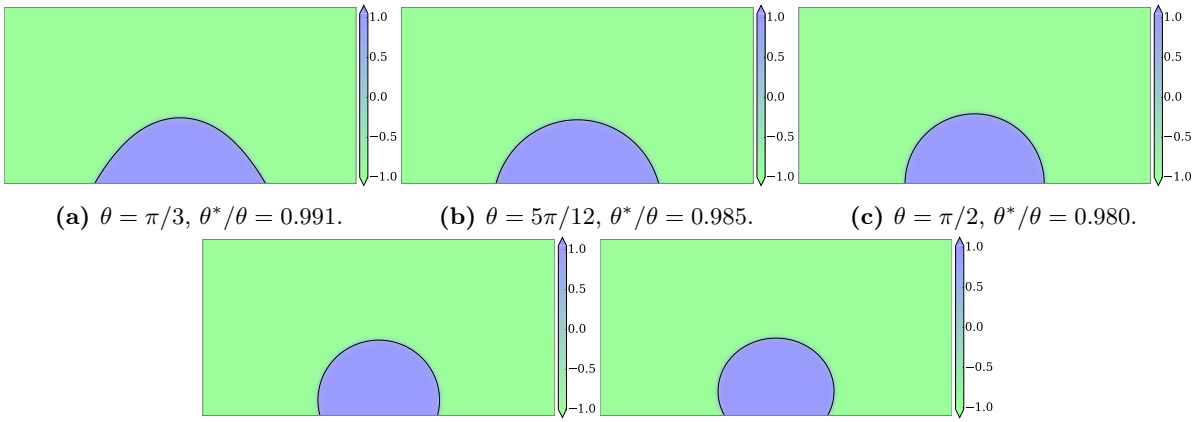
    $\Delta t^{n+2} = f \Delta t^{n+1}$;

$n = n + 1$ and go back to 1.

The condition $(E(\phi^*) - E(\phi^n)) > \zeta\, \Delta E_{\max}$ serves to guarantee that the method does not blow up. The choice of a nonzero right-hand side is motivated by the fact that, when the system is close to equilibrium, it can happen that $E(\phi^*) > E(\phi^n)$. This is because, in contrast with the sign of $ND_{\text{philic}}(\phi^n, \phi^{n+1})$, which is always positive or zero according to Eq. (5.13), the signs of $ND_{\text{phobic}}(\phi^n, \phi^{n+1})$ and $ND_{\text{wall}}(\phi^n, \phi^{n+1})$ are in general unknown.

In the numerical experiments presented in Section 5.4, we chose $\Delta t_{\min} = 0$. Since the numerical dissipation term scales as Δt^2 for the method OD2-W, the inequality $E(\phi^{n+1}) \leq E(\phi^n) + \zeta\, \Delta E_{\max}$ will always hold for Δt small enough, so the refinement process is guaranteed to terminate at each iteration.

If the condition $\dot{m} = 0$ is not satisfied, then the adaptation scheme we proposed, being based on limiting the decrease and preventing the increase (beyond a very small tolerance equal to $\zeta \Delta E_{\max}$) of free energy at each time step, is not applicable. Indeed, by Eq. (5.6), a nonzero mass flux can cause a positive variation of the total free energy, which our scheme would be unable to distinguish from a nonphysical increase of energy caused by numerical errors. In this nonzero flux situation, the adaptation scheme presented in [99], which is based on limiting the numerical dissipation, could be used instead (with suitable modifications to accommodate for the presence of a flux).



(a) $\theta = \pi/3$, $\theta^*/\theta = 0.991$. **(b)** $\theta = 5\pi/12$, $\theta^*/\theta = 0.985$. **(c)** $\theta = \pi/2$, $\theta^*/\theta = 0.980$.

(d) $\theta = 7\pi/12$, $\theta^*/\theta = 0.979$. **(e)** $\theta = 2\pi/3$, $\theta^*/\theta = 0.978$.

Figure 5.2: Equilibrium position of a droplet on substrates with different wetting properties. In all cases, θ denotes the angle appearing in boundary condition (5.16) and θ^* denotes the angle calculated numerically. Blue corresponds to $\phi = 1$ and green to $\phi = -1$.

5.4 Numerical results

The new numerical method is applied on a number of test cases. We have used *FreeFem++* [106] for the implementation off the finite element method and 2D mesh adaptation, *umfpack* [50] for the linear solver, *mshmet* [76] and *tetgen* [198] for the mesh adaptation in 3D, and *gmsh* [88] for the description of the geometry, post-processing and 3D visualization. In Section 5.4.1 we check that the numerical scheme leads to the correct equilibrium solution in the simple case of a droplet spreading on a philic or phobic substrate. In Section 5.4.2 we study the convergence of the method with respect to the time step and the mesh size, when a uniform mesh and a constant time step are used. In Section 5.4.2 we illustrate the time-adaptation scheme in the case of two droplets coalescing on a substrate. Finally, Section 5.4.4 demonstrates the ability of the numerical scheme to scrutinize wetting phenomena in more complicated geometries, and in the presence of heterogeneous substrates. The code used for the simulations is available on line, see [217].

5.4.1 Equilibrium contact angle

We consider a 2D sessile droplet on a flat substrate where we impose the no-flux condition and the wetting condition (5.2) with the modified wall energy (5.9) and uniform contact angle θ:

$$\nabla\mu \cdot \mathbf{n} = 0, \quad \varepsilon\nabla\phi \cdot \mathbf{n} = -f_w(\phi) \tag{5.16}$$

Our aim in this section is to check that our method is able to accurately capture the imposed contact angle, θ. Figure 5.2 shows the equilibrium position of a droplet for different values of θ, for $b = 1$ and $\varepsilon = 5 \times 10^{-3}$. In all cases we used the scheme OD2-W, with quadratic basis functions and adaptation in space only using the parameters $h_{\max} = 10\, h_{\min} = 0.01$, and we computed the contact angle of the $\phi = 0$ isoline at the substrate. A very good agreement is achieved between the imposed equilibrium contact angle and the observed numerical one.

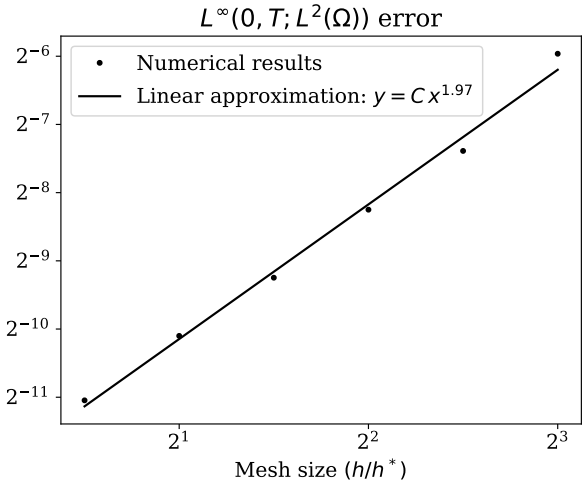


Figure 5.3: Convergence of OD2-W with respect to the mesh size, without mesh adaptation. In this case, h corresponds to the uniform mesh size, given as input to the mesh generator of *gmsh*, and $h^* = 0.01$.

5.4.2 Convergence of the method

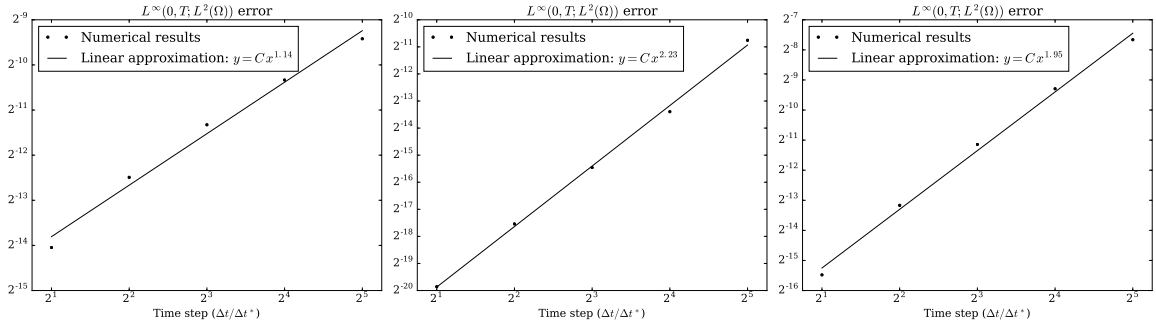
Here, we study the convergence of the method when both time step and mesh size decrease. The problem we considered to that purpose is the coalescence of two adjacent sessile droplets as they spread on a flat substrate. For the simulation, we used the initial condition

$$\phi(x,0) = 1 - \tanh\left(\frac{\sqrt{(x-x_1)^2+y^2}-r}{\sqrt{2}\varepsilon}\right) - \tanh\left(\frac{\sqrt{(x-x_2)^2+y^2}-r}{\sqrt{2}\varepsilon}\right), \tag{5.17}$$

in the domain $[0,2]\times[0,0.5]$, with $x_1 = 0.65$, $x_2 = 1.35$, $r = 0.25$, and at the boundary we imposed a uniform contact angle, $\theta = \pi/4$, using the wall energy (5.9). Only linear elements were used.

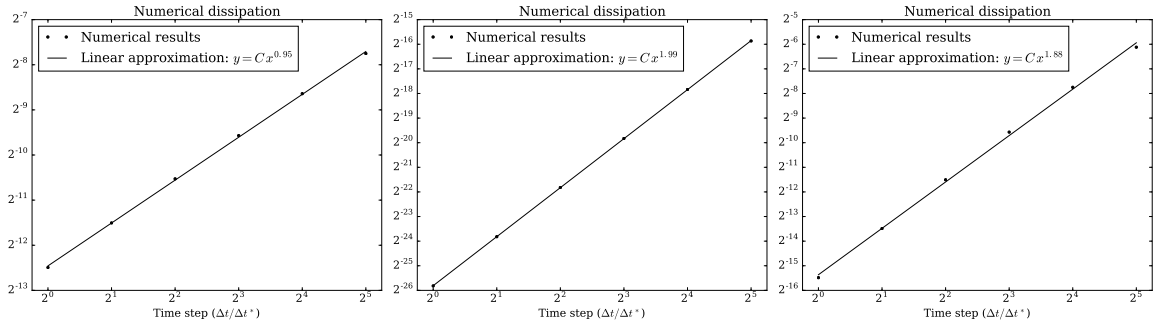
For the convergence as $h \to 0$, we solved the problem numerically for several values of h, without mesh adaptation and for $\varepsilon = 0.1$, so that enough data points could be generated at a reasonable numerical cost. Since the exact solution to the CH equation in this case is not known analytically, we calculated the error by comparison of the numerical solutions to the solution obtained with the smallest value of h. Results are presented in Fig. 5.3. As we can see, the observed convergence rate is almost equal to 2, which is the optimal rate in the case of linear basis functions.

Now we address the convergence with respect to the time step. For this case, we used the parameters $\varepsilon = 0.1$, $b = 10^4$, and the minimum time step we considered was $\Delta t^* := 0.00665$. In Fig. 5.4, we present convergence curves for OD1-W, OD2-W, and OD2mod-W. We note that the convergence rates are close to the expected ones, and that the use of OD2 gives significantly more accurate results than the other two methods. In Fig. 5.5, the total numerical dissipation produced by the numerical schemes is presented. Here too, numerical results agree with the theoretical results of Section 5.3.



(a) OD1-W **(b)** OD2-W **(c)** OD2mod-W with $\beta = 10\,\Delta t$

Figure 5.4: Convergence of the numerical method with respect to the time step, without mesh adaptation. In the case of OD1-W, the rate of convergence is close to the expected value of 1. In the other two cases, the rate of convergence is close to the expected value of 2.



(a) OD1-W **(b)** OD2-W **(c)** OD2mod-W with $\beta = 10\,\Delta t$

Figure 5.5: Total numerical dissipation generated by the numerical schemes in the simulation used to produce Fig. 5.4. OD2-W is by far the scheme producing the least numerical dissipation, even for relatively large time steps. OD2mod-W, on the other hand, introduces significant numerical dissipation for large time steps, owing to the large value of β that was chosen for the simulation, but is much less dissipative than OD1-W for small time steps.

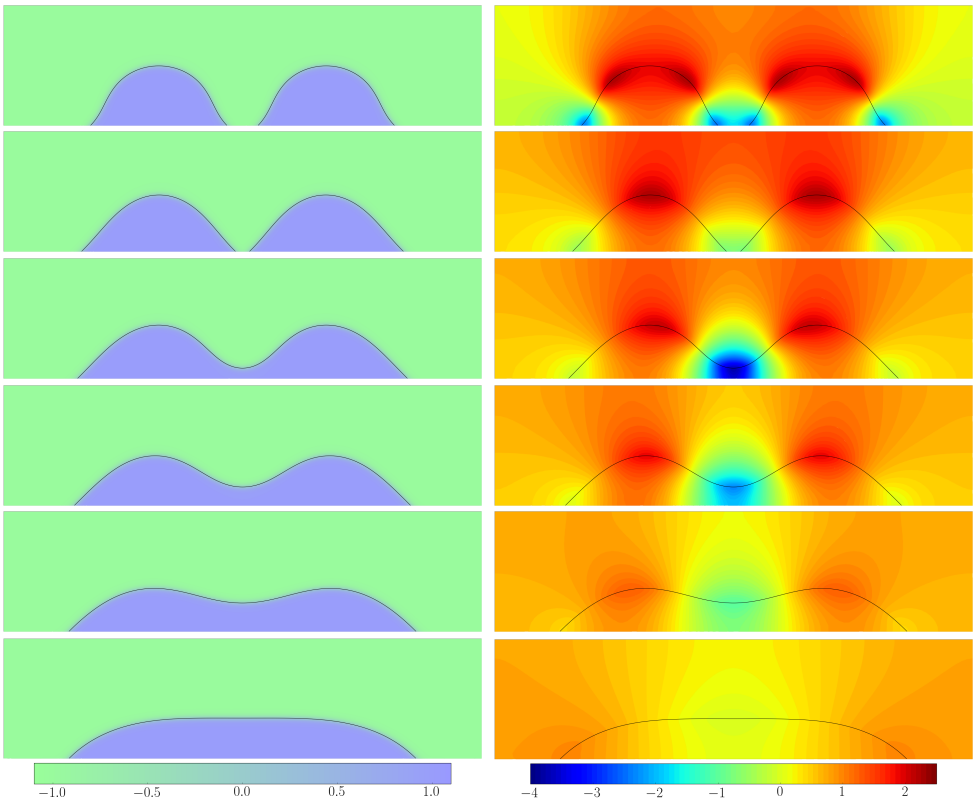


Figure 5.6: Phase field and chemical potential during the coalescence of two sessile droplets on a hydrophilic substrate for $\theta = \pi/4$. The snapshots correspond to iterations 500, 1000, 1500, 2000, 2500, and 3000, which correspond to times 2.26, 19.55, 34.13, 49.03, 83.20 and 273.64. Blue color represents phase $\phi = 1$ and green phase $\phi = -1$.

5.4.3 Time-adaptation scheme

In this section, we examine the performance of the adaptive time-stepping scheme in the case of two droplets evolving on a chemically homogeneous substrate. We start from the situation where $\phi = -1$ everywhere except in two half-circles, of radius $r = 0.25$ and centred at $(0.65, 0)$ and $(1.35, 0)$, where $\phi = 1$. We used quadratic basis functions and the following parameters: $b = 10^{-4}$, $\varepsilon = 0.01$, $f = \sqrt{2}$, $\Delta t_0 = 0.02$, $\Delta t_{\min} = 0$, $\Delta t_{\max} = 16\Delta t_0$, $\Delta E_{\min} = 0.0001$, $\Delta E_{\max} = 0.0002$, $h_{\max} = 0.05$ $h_{\min} = 0.001$, and for θ we considered three values: $\pi/4$, $\pi/2$, $3\pi/4$.

Snapshots of the phase field and of the chemical potential at different times of the simulation are presented in Figs. 5.6 and 5.7 for the case $\theta = \pi/4$ and $\theta = 3\pi/4$, respectively. The case $\theta = \pi/2$ is less interesting because, in view of the initial condition, the droplets remain essentially motionless throughout the simulation; we do not present snapshots of the solution in that case.

The evolution of the time step, of the number of recalculations, and of the free energies is displayed in Fig. 5.8. In all three cases, the time step is refined several times at the first iteration, which can be explained as follows. On one hand, this refinement originates from the discontinuity of the initial condition at the interface between the two fluids. On the other hand, in the cases $\theta = \pi/4$ and $\theta = 3\pi/4$, it is also a consequence of the fact that the initial condition does not satisfy the wetting boundary condition. Since the initial angle between the interface and the substrate is equal $\pi/2$, the number of recalculations performed at the first iteration is higher for

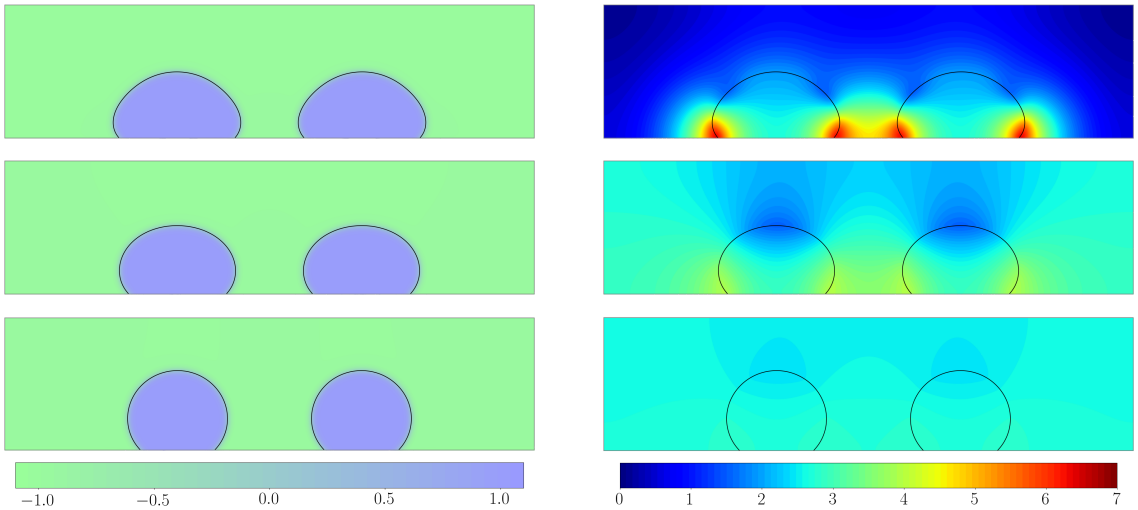


Figure 5.7: Phase field and chemical potential when the contact angle is set to $3\pi/4$. The snapshots correspond to iterations 500, 1000, 1500, which correspond to times 1.47, 17.32, and 115.71.

$\theta = \pi/4, 3\pi/4$ than for $\theta = \pi/2$. After the initial refinement, the time step steadily increases to its maximum allowed value for $\theta = \pi/2$ and $\theta = 3\pi/4$, but when $\theta = \pi/4$ a second refinement occurs to capture the coalescence of the droplets.

In this latter case, we observe, simultaneous with the second refinement of the time step, an increase in the rate of dissipation of free energy. After the formation of a new stable interface, the total free energy continues to decrease, but more slowly, as a new droplet, formed by the merging of the two original droplets, moves towards its equilibrium position. We clearly identify the coalescence time by looking at the singularity in the curve corresponding to the mixing energy. This energy increases before coalescence, as the interfaces are being stretched, and it decreases steadily after. The wall energy, on the other hand, decreases at first and increases in the later stage of the simulation. As prescribed by algorithm 1, the time step detects the variations of free energy; it decreases when the rate of variation of the total free energy increases, and conversely.

For comparison purposes, we also included in Figs. 5.8d to 5.8f data corresponding to the case where a fixed time step is used for the simulations presented in this section. There does not currently exist any result with conditions on the time step that ensure the stability of OD2, and we haven't been able to show stability results for OD2-W either. In practice, we observed that the time step required to ensure stability of OD2-W with the set of parameters we use in this test case would lead to a very high computational cost. We point out that, contrary to what we expected, the time step required to achieve stable integration in time with the modified wall energy (5.9), which we use in this chapter, seems to be generally smaller than with the cubic formulation (5.1). To keep the computational cost at a reasonable level, we carried out the simulations with a fixed time step using the method OD1-W, the greater stability of which enabled us to choose $\Delta t = 0.02$. In Figs. 5.8d to 5.8f, we see that, for the same contact angle, the curves corresponding to a fixed and an adaptive time step are almost undistinguishable. The agreement is also very good at the level of the phase field and chemical potential, although we do not present snapshots of the solutions obtained with a fixed time step.

The CPU times corresponding to the three contact angles considered are presented in Table 5.1.

Without adaptation, the simulations take significantly longer to run, which is consistent with the fact that more iterations (20000) were necessary to reach the final time. In addition, among the simulations that used an adaptive time-step, the difference between the CPU times is also significant, with the case $\theta = \pi/4$ taking more than twice as long as the case $\theta = \pi/2$.

Contact angle	Adaptive time step	Fixed time step
$\pi/4$	44:15:17	130:16:16
$\pi/2$	21:38:38	128:35:20
$3\pi/4$	31:12:18	122:33:28

Table 5.1: CPU times (hh:mm:ss) using an Intel i7-3770 processor for the simulations presented in Section 5.4.3 (two droplets on a substrate), with or without time-step adaptation. The method OD2-W was used for the simulations with an adaptive time step, and the method OD1-W was used for the simulations with a fixed time step. In both cases, an adaptive mesh was used, with the parameter $h_{\min}$ equal to $\varepsilon/10 = 0.001$.

5.4.4 Wetting in complex geometries and with heterogeneous substrates

We now present the results of numerical experiments in more complicated and realistic settings, in both 2D and 3D systems. The results presented in this section were all obtained using linear basis functions.

3D droplet on a chemically heterogeneous substrate

We study the dynamics of a 3D sessile droplet on a flat substrate with chemical heterogeneities, i.e. the contact angle has a spatial dependence now, say $\theta = \theta(x, y)$. This situation typically arises in electro-wetting settings [145]. It is widely accepted that the droplet shape can be controlled using patterned substrates, e.g. [219, 197], that may also be modeled efficiently using a space varying contact angle [219]. We consider chemical heterogeneities on the substrate of the form

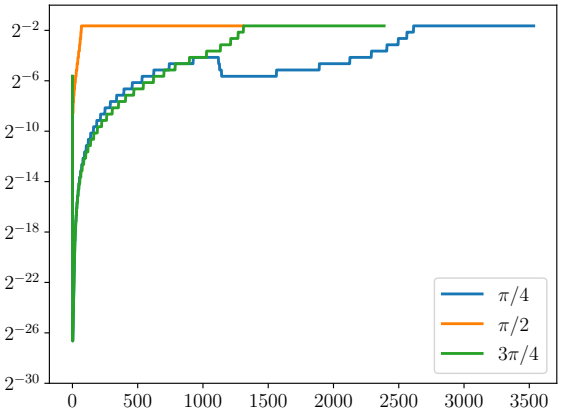
$$\theta(x, y) = \theta_0 + a\cos(f_x \pi x)\cos(f_y \pi y), \tag{5.18}$$

with $\theta_0 = \frac{\pi}{2}$ the mean contact angle, $a = \frac{\pi}{6}$ the amplitude, and $f_x = f_y = 4$ the frequencies in x and y directions, respectively. As initial condition we take a droplet of base radius $r_0 = 0.24$ centered at $\mathbf{x_0} = (0.5, 0.5, 0)$. The initial values of the phase field are given as

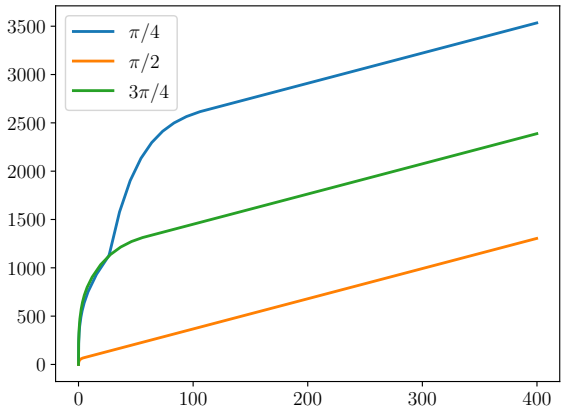
$$\phi_0(\mathbf{x}) = -\tanh\left(\frac{\|\mathbf{x} - \mathbf{x_0}\| - r_0}{\sqrt{2}\varepsilon}\right). \tag{5.19}$$

Results are displayed in Fig. 5.9. The droplet, initially spherical, spreads on the hydrophilic regions of the substrate, and retracts from the hydrophobic patches. While we do not present any quantitative analysis of the error in this case, we note that the wetting behaviour agrees qualitatively with what one might expect intuitively from our understanding of wetting phenomena. While it progresses towards equilibrium, the droplet adopts a diamond-like shape.

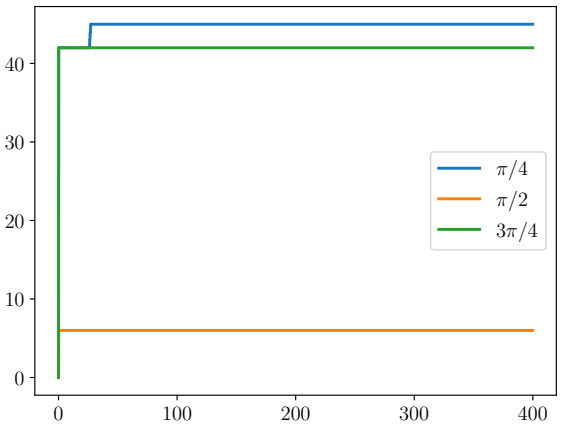
For this test case, we used the method OD2-W with adaptation in space and time. The parameters used were the following: $b = 10^4$, $\varepsilon = 0.02$, $h_{\max} = 10\, h_{\min} = 0.1$, $\Delta t_0 = 0.0016$,



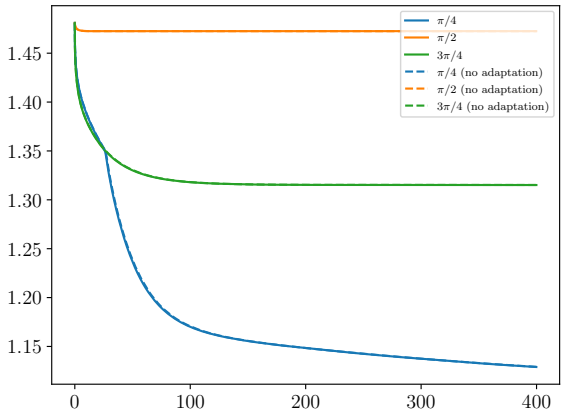
(a) Time step vs iteration.



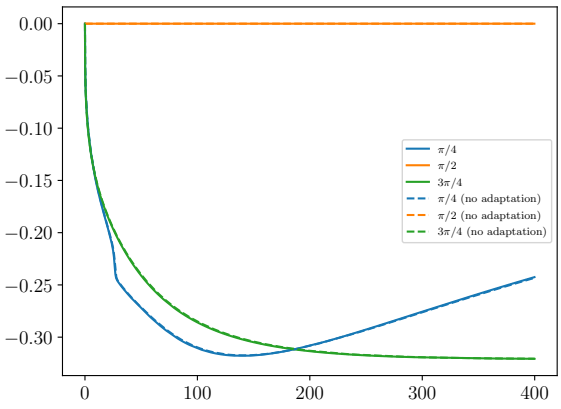
(b) Iteration vs time.



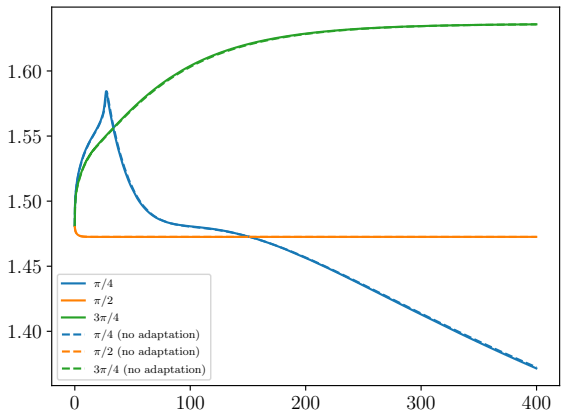
(c) # Recalculations vs time.



(d) Total free energy vs time.



(e) Wall free energy vs time.



(f) Mixing free energy vs time.

Figure 5.8: Simulation data for the numerical experiments presented in Section 5.4.3 (two droplets on a substrate), when using the adaptive time-stepping scheme 1. As expected, the total free energy decreases with time for all three values of the contact angle. In the case $\theta = \pi/4$, we note a peak in the mixing component of the free energy and a refinement of the time step at the coalescence time. See Table 5.1 for the associated computational costs; we notice that the simulations with an adaptive time step are about 3 times faster or more.

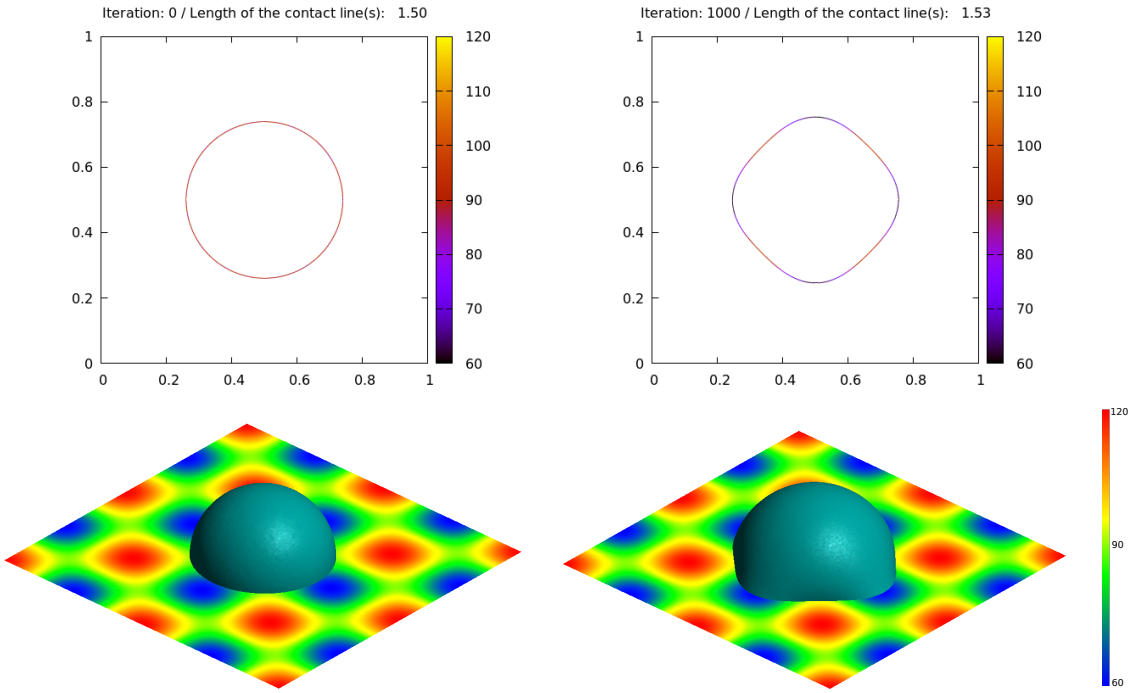


Figure 5.9: Evolution of the contact line (top) and the isosurface $\phi = 0$ (bottom) of the phase field, on a chemically heterogeneous substrate with a contact angle defined by (5.18). The balance of the wall component and the mixing component of the free energy determines the motion of the drop. The field represented in the top figures is the value of the contact angle at the triple line. In the bottom figures, the field represented on the plane representing the substrate is the imposed contact angle. Interestingly, the heterogeneities of the substrate cause the length of the contact line to increase.

$\Delta t_{\min} = 0$, $\Delta t_{\max} = 16\,\Delta t_0$, $f = \sqrt{2}$, $\Delta E_{\max} = 10\,\Delta E_{\min} = 0.0001$. With these parameters, the time step was refined only at the beginning of the simulation, which is consistent with the absence of coalescence events in this case. There were 24 recalculations at the first time step, corresponding to a refinement of the time step by a factor $f^{24} = 4096$.

Diffusion in a 3D porous medium

Here we consider a binary fluid in a model porous medium consisting of a cube filled with spheres. The cube has edges of length 1, and the spheres have radius 0.1 and are located at positions $(1.5, 1.5, 1.5) + 2\Delta(i, j, k)$ with $\Delta = 1/7$ and $i, j, k \in \{0, 1, 2\}$. We take all the substrates to be neutral, i.e. $\theta = \frac{\pi}{2}$, and the initial condition is the same as used before, defined by Eq. (5.19). In addition, we include an inflow boundary condition at the bottom of the cube to represent a pore where liquid can be pumped in. The radius of this pore is 0.1 and is located at $(0.5, 0.5, 0)$. This boundary condition can be incorporated by imposing

$$\nabla\mu \cdot \mathbf{n} = -10, \quad \phi = 1, \tag{5.20}$$

which models the situation when the component $\phi = 1$ is pumped into the domain. Under these conditions, we study how the flow is affected by the geometry of the domain. Our results are depicted in Figs. 5.10 and 5.11.

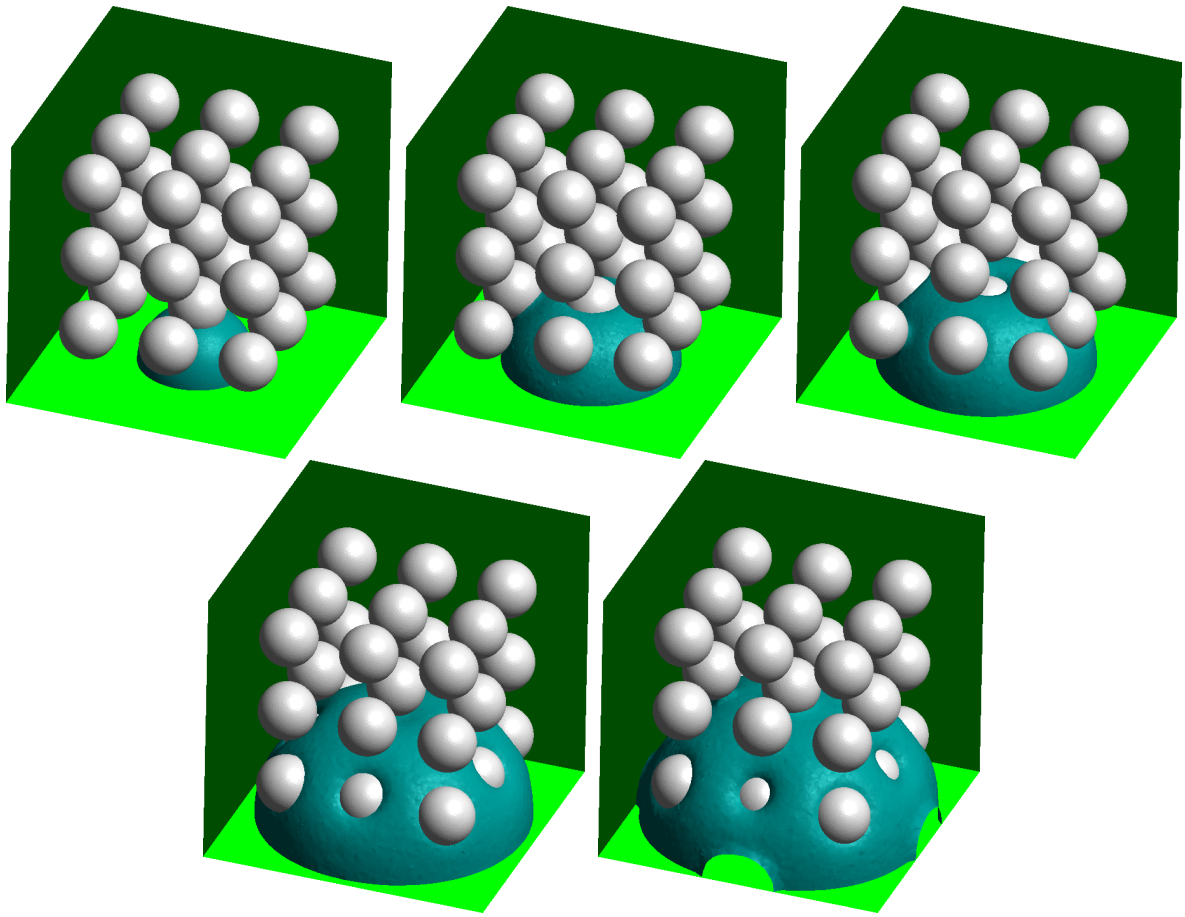
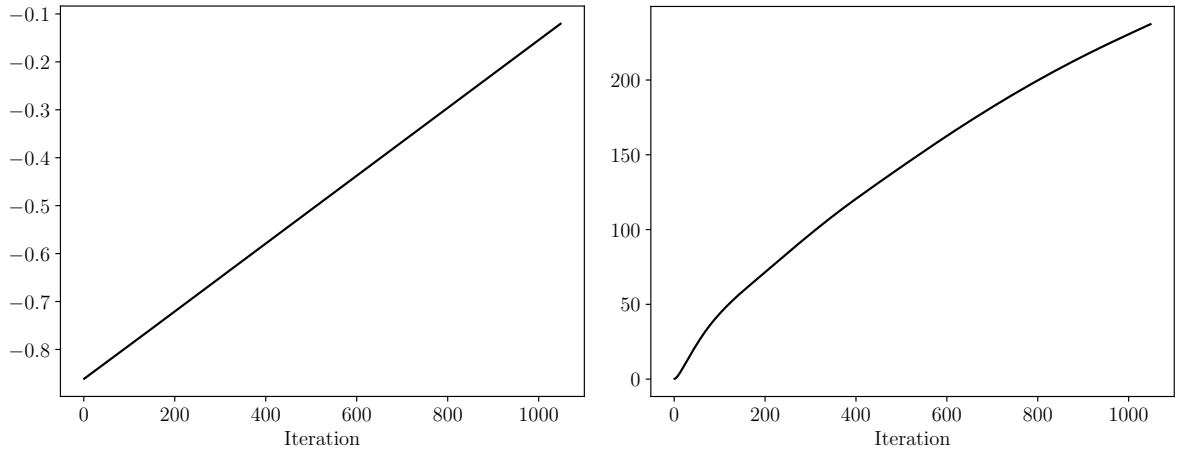


Figure 5.10: Evolution of the isosurface $\phi = 0$ of the phase field when a constant flux is imposed at the bottom boundary; The pictures correspond to iterations 0, 200, 400, 800 and 1000. Note that, because of the neutral boundary condition imposed at the spheres, the isosurface tends to stay normal to them as long as they are not completely covered.

The imposed contact angle at the spheres is $\pi/2$, forcing the isosurface to stay normal to the spheres as long as these are not completely covered. Because of the boundary condition (5.20), the mass increases linearly, and the free energy increases, in agreement with Eqs. (5.4) and (5.6). This case study demonstrates the ability of our method to easily tackle complex geometries. The parameters used for this test case are the same as in Section 5.4.4, except that we employed the fixed time step $\Delta t = 0.001$.

Nucleation processes with complex boundaries

The last problem we study is the process of phase separation in a domain with complex boundary characterized by different length scales. Specifically, we consider a domain defined by the coastline of the two islands that form the United Kingdom and Ireland. Starting from a satellite black and white picture, we extracted the isolines that define the contour of the different islands, which we passed to the FreeFem++ mesh generator to obtain a triangular mesh (for this, we based our code on a FreeFem++ example for the Leman lake). At the boundary we consider the contact angles $\theta = \pi/4, \pi/2, 3\pi/4$, and we assume that the phase field is initially set to a random value at each grid point, drawn from a random normal distribution with variance 0.1. A fixed mesh



(a) Mass, Eq. (5.4). **(b)** Free energy, Eq. (5.6).

Figure 5.11: Evolution of $M(\phi)$ and $E_m(\phi)$ as a function of time in the case of a 3D flow through a porous medium (Figure 5.10). In this case, the mass increases linearly because we impose a constant mass inflow at the pore. The free energy increases as well, because the size of the interface grows, in agreement with both the mass and energy laws (5.4) and (5.6).

was used for this simulation, and the parameters used were $b = 1000$, $\varepsilon = 0.02$, $\Delta E_{\min} = 0.02$, $\Delta E_{\max} = 0.04$, $f = \sqrt{2}$, $\Delta t_{\min} = 0$, $\Delta t_{\max} = 1$.

The evolution of the phase field and of the chemical potential in the case $\theta = \pi/4$, obtained with the adaptive time-stepping scheme 1, is presented in Fig. 5.12. For each of the contact angles considered, we also ran a simulation with the fixed time step $\Delta t = 0.01$, using the method OD1-W instead of OD2-W to benefit from the stabilizing effect introduced by the philic numerical dissipation of OD1-W. We note in particular that OD2-W is unstable for the selected value of Δt, with oscillations appearing in the energy curves from the first iterations, and that the time step would have to be reduced significantly to ensure stability. The final configurations (time 500) are presented in Fig. 5.13 for the three contact angles considered. We observe that the final configurations are different depending on whether or not an adaptive time step is used, which can be attributed to the high sensitivity of the solution to perturbations of the initial condition chosen for this test case; the areas where separation of the phases first occurs is influenced by numerical errors in the early stages of the simulation.

Simulation data are presented in Fig. 5.14. With an adaptive time-stepping scheme, it appears from Fig. 5.14 (a) that, overall, the time step increases steadily as the frequency of coalescence events decreases. At specific times, the time step decreases slightly in order to accurately capture the evolution. As expected, the total free energy has a roughly constant negative slope when plotted against the iteration number. Here too, we observe a small discrepancy between the fixed and adaptive cases, which is consistent with differences observed at the final time in Fig. 5.13.

The CPU times corresponding to the simulations presented in this section are displayed in 5.2. For the parameters selected, the adaptive time-stepping scheme leads to a lower computational cost. This test demonstrates the advantage of using a finite element approach, as it would have been very complicated to solve the CH equation in the geometries we consider here with e.g. a spectral method or finite differences.

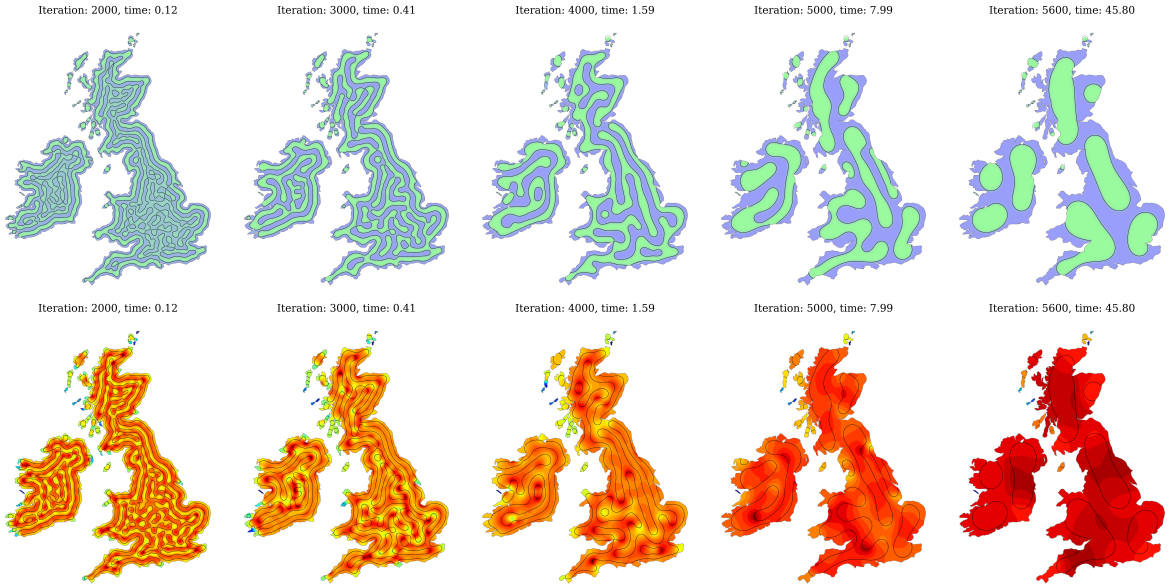
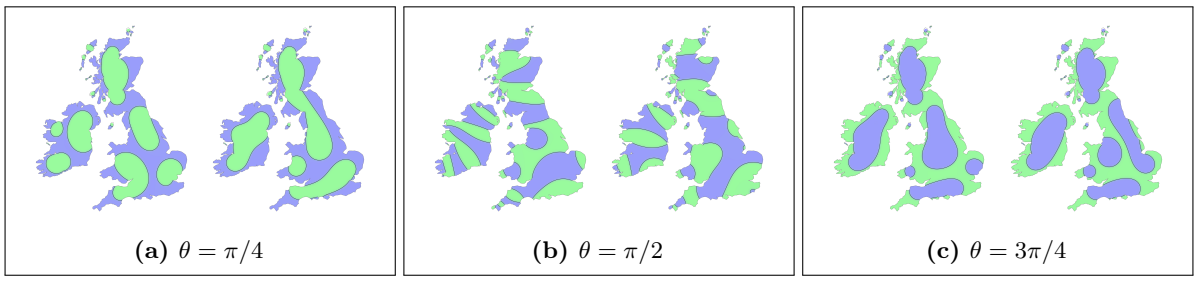
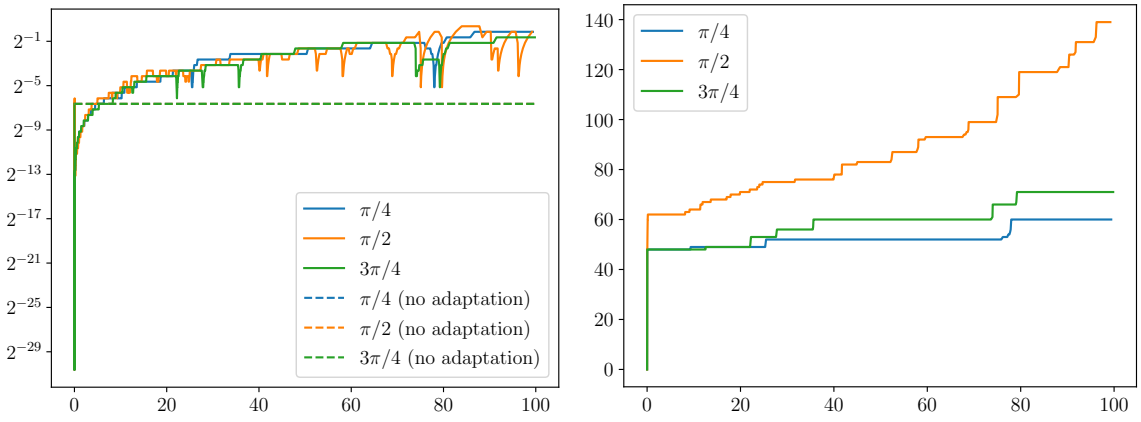


Figure 5.12: Evolution of the phase field and chemical potential for the nucleation in a domain with complex boundaries, when starting from a random distribution. As before, blue corresponds to $\phi = 1$ and green to $\phi = -1$. The contact angle imposed at the boundaries is $\theta = \frac{\pi}{4}$.



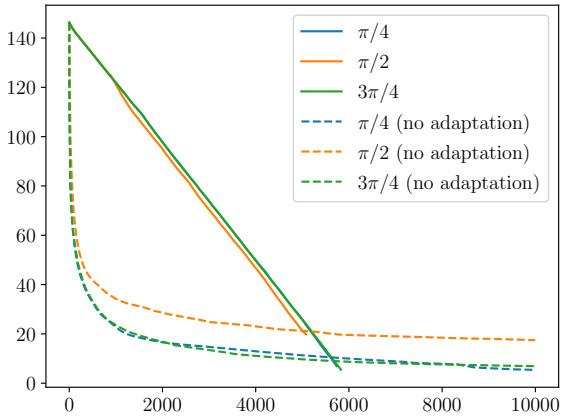
(a) $\theta = \pi/4$ **(b)** $\theta = \pi/2$ **(c)** $\theta = 3\pi/4$

Figure 5.13: Comparison of the solutions at the final time, with (left) and without (right) time-step adaptation. At the initial time, the phase field is set to a random value at each grid point, drawn from a random normal distribution with variance 0.1. Although one could expect the solutions for $\theta = \pi/4$ and $\theta = 3\pi/4$ to differ only by a sign, this is not the case. There is also a significant difference between the solutions obtained with and without time-step adaptation. These differences can be explained by the sensitivity of the evolution to perturbations of the initial condition and to numerical errors in the early stage of the simulation.

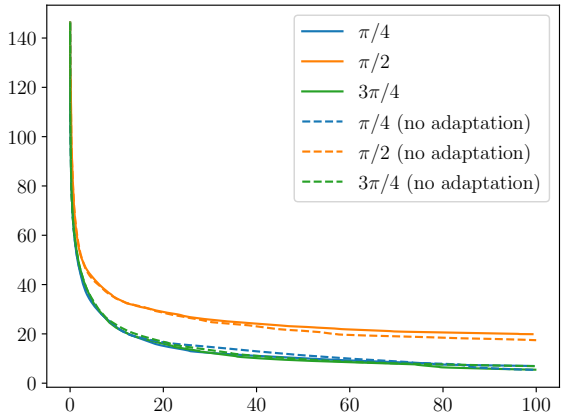


(a) Time step (algorithm 1) vs time

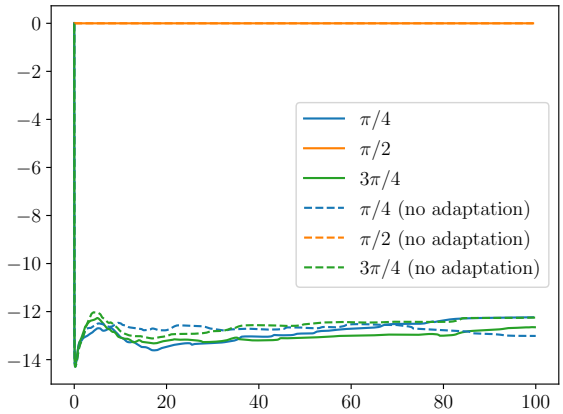
(b) # of recalculations vs time.



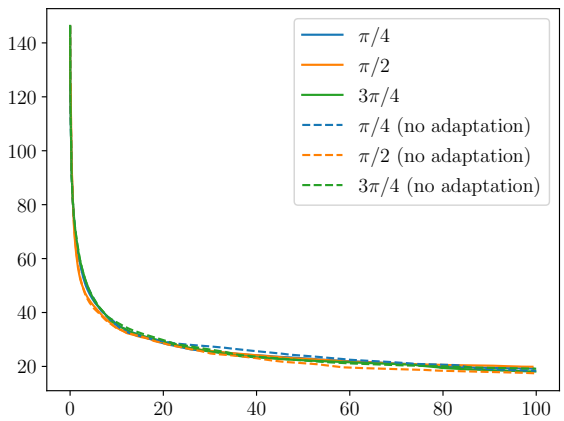
(c) Total energy vs iteration.



(d) Total energy vs time.



(e) Wall energy vs time.



(f) Mixing energy vs time.

Figure 5.14: Simulation data for the numerical experiments presented in Section 5.4.4 (nucleation in a geometry with complex boundaries). Overall, the time step increases steadily when the adaptive time-stepping scheme is used, which is consistent with the decreasing frequency of coalescence events. The time step is refined at times to ensure that the incremental decrease of free energy at each iteration is approximately constant.

Contact angle	Adaptive time step	Fixed time step
$\pi/4$	36:13:15	65:48:44
$\pi/2$	32:58:01	65:24:10
$3\pi/4$	36:43:13	67:00:50

Table 5.2: CPU times (hh:mm:ss) using an Intel i7-3770 processor for the simulations presented in Section 5.4.4 (nucleation in a domain with complex boundaries), with or without time-step adaptation. The method OD2-W was used for the simulations with an adaptive time step, and the method OD1-W was used for the simulations with a fixed time step. In all cases, we used a fixed mesh with mesh size $h = 0.01$ (the size of the domain is roughly 5 by 5) and $P1$ elements, leading to 181587 unknowns. The parameter ε was set to 0.02.

5.5 Extension to the Cahn–Hilliard–Navier–Stokes system

So far we considered a regime in which the contact line motion is controlled by diffusive interfacial fluxes, or in other words, we considered a large diffusivity limit, where any possible advection effects are neglected. To account for advective effects the Cahn–Hilliard must be appropriately modified to include an advection term and they must be coupled to the Navier–Stokes equations [8, 194, 121, 122, 129, 222]. Such generalizations are indeed possible within the numerical scheme we proposed in Section 5.3. In this section, we present a possible approach to this problem in the hope that it will foster future research on the subject.

One possible manner of extending our numerical method in the presence of advection is to use a splitting strategy based on solving the Cahn–Hilliard (CH) and the Navier–Stokes (NS) equations separately at each time step. We will assume for simplicity that the computational domain is surrounded by a solid substrate, with a homogeneous Dirichlet boundary condition for the velocity, and we will denote by ϕ^n, μ^n, $\mathbf{v}^n$ and p^n the numerical approximations of ϕ, μ, $\mathbf{v}$ and p at time $n\,\Delta t$, respectively. In the first step of the method we propose, the velocity field is considered to be fixed and the following variational formulation, similar to Eqs. (5.11a) and (5.11b), is employed to advance the phase field and chemical potential according to the CH equation with advection

$$
\begin{aligned}
&\left\langle \overline{\delta}_t \phi^{n+1}, \psi \right\rangle + \left\langle b\, \nabla \mu^{n+\frac{1}{\alpha}}, \nabla \psi \right\rangle = 0 && \forall \psi \in H^1(\Omega),\\
&\left\langle \mu^{n+\frac{1}{\alpha}}, \nu \right\rangle = \varepsilon \left\langle \nabla \phi^{n+\frac{1}{\alpha}+\beta}, \nabla \nu \right\rangle \\
&\qquad\qquad + \frac{1}{\varepsilon} \left\langle \hat{f}_m(\phi^n, \phi^{n+1}), \nu \right\rangle + \left\langle \hat{f}_w(\phi^n, \phi^{n+1}), \nu \right\rangle_{\partial\Omega} && \forall \nu \in H^1(\Omega),
\end{aligned}
$$

where $\overline{\delta}_t \phi^{n+1}$ is the approximation of the material derivative given by

$$
\frac{\phi^{n+1} - (\phi^n - \Delta t\, \mathbf{v}^n \cdot \nabla \phi^n)}{\Delta t},
$$

and the terms $\hat{f}_m(\phi^n, \phi^{n+1})$ and $\hat{f}_w(\phi^n, \phi^{n+1})$ are the same functions, linear in their second argument, as in Eqs. (5.15a) and (5.15b). The second step is to solve the NS system with the phase field and chemical potential fixed, for which the original first-order projection scheme proposed by Alexander Chorin [42] can be employed, after appropriate generalization to include the capillary forces. For a review of projection methods and of the related numerical issues, see

e.g. [66] and the references therein. The projection method can itself be divided in three steps: first, an auxiliary velocity field is obtained by solving

$$\frac{\mathbf{v}^* - (\mathbf{v}^n - \Delta t\, \mathbf{v}^n \cdot \nabla \mathbf{v}^n)}{\Delta t} = \frac{1}{\mathrm{Re}}\, \Delta \mathbf{v}^* + \frac{1}{\lambda}\, \mu^{n+1} \nabla \phi^{n+1}$$

with homogeneous Dirichlet boundary condition, i.e. $\mathbf{v}^* = 0$ on Ω, and with the same notations as in Section 1.5.4. Next, the pressure is calculated from the divergence of $\mathbf{v}^*$:

$$\Delta p^{n+1} = \frac{1}{\Delta t} \nabla \cdot \mathbf{v}^*, \tag{5.22}$$

and finally, the velocity field is updated in such a way that $\mathbf{v}^{n+1}$ satisfies the incompressibility (divergence-free) condition:

$$\frac{\mathbf{v}^{n+1} - \mathbf{v}^*}{\Delta t} = -\nabla p^{n+1}.$$

Indeed we verify using Eq. (5.22) that

$$\nabla \cdot \mathbf{v}^{n+1} = \nabla \cdot \mathbf{v}^* - \Delta t\, \Delta p^{n+1} = 0.$$

Although a thorough discussion of the projection method is beyond the scope of this work, we remark that the choice of the boundary condition for Eq. (5.22) is a delicate one, as demonstrated in [66]. The most widely used approach in the literature is to impose the homogeneous Neumann boundary condition $\nabla p^{n+1} \cdot \mathbf{n} = 0$ on $\partial\Omega$. This guarantees that the normal velocity $\mathbf{v}^{n+1} \cdot \mathbf{n}$ is 0 at the boundary. The numerical method presented above can be generalized to systems with open boundaries, at the cost of modifying the functional spaces in the variational formulations appropriately.

5.6 Conclusions

We have proposed a new, fast and reliable numerical method to solve the CH equation with a wetting boundary condition. Our method is a generalization of the OD2 scheme introduced in [98], which considered only the homogeneous condition $\nabla \phi \cdot \mathbf{n} = 0$. In addition, we have designed a new time-step adaptation algorithm, leading to a scheme that is adaptive both in space and time, and we have shown that this scheme is mass-conservative and satisfies a consistent discrete energy law.

We checked the validity of the proposed numerical scheme with several examples. First we considered the relaxation towards equilibrium of a sessile droplet and the coalescence of two sessile droplets on flat, chemically homogeneous substrates; then we considered several multiphase systems in complex geometries or surrounded by chemically heterogeneous substrates.

Compared to finite differences or spectral approaches, the method introduced here has the advantage that it can be used without modification with complex geometries. Furthermore, the numerical scheme we have proposed can easily be extended to include at least two additional features. First, a linear, energy-stable, second-order scheme could be developed for the three-component CH model with wetting boundary conditions, building on the work of [30, 31]. And

second, different models of the free energy and of the viscosity, such as the ones reviewed in Section 1.5.2, could be considered.

5.A Proof of Theorem 5.1

Before presenting the proof, we recall a particular Sobolev embedding for smooth bounded domains; see e.g. [70, Chap. 5]. Let $d \geq 2$, $\emptyset \neq \Omega \subset \mathbf{R}^d$ be open with C^1 boundary, and assume that $q < \infty$ if $d = 2$ or $q < p^* := 2d/(d-2)$ if $d > 2$. Then the following embedding is compact

$$H^1(\Omega) \hookrightarrow L^q(\Omega).$$

We also recall two other well-known compactness results; see e.g. [144]. Let X, Y, Z be Banach spaces with a compact embedding $X \subset Y$ and a continuous embedding $Y \subset Z$. Then the following embeddings are compact:

$$\{u \in L^2(0,T;X) \,|\, \frac{\partial u}{\partial t} \in L^2(0,T;Z)\} \hookrightarrow L^2(0,T;Y), \tag{5.23a}$$

$$\{u \in L^\infty(0,T;X) \,|\, \frac{\partial u}{\partial t} \in L^2(0,T;Z)\} \hookrightarrow C([0,T],Y). \tag{5.23b}$$

Proof. Without loss of generality, we assume that the mobility, b, is equal to 1. In the spirit of [68, Theorem 2], we apply a Faedo–Galerkin approximation. Let $\{\varphi_n\}_{n\in\mathbf{N}}$ and $\{\lambda_n\}_{n\in\mathbf{N}}$ denote the eigenfunctions and eigenvalues of the Laplace operator with a homogeneous Neumann boundary condition, i.e.

$$\begin{cases} -\Delta\varphi_n = \lambda_n \varphi_n & \text{in } \Omega, \\ \nabla\varphi_n \cdot \mathbf{n} = 0 & \text{in } \partial\Omega, \end{cases}$$

normalized such that

$$\int_\Omega \varphi_n \varphi_m \,\mathrm{d}\Omega = \delta_{mn}.$$

We assume without loss of generality that $\lambda_1 = 0$. To build an approximation of the solution to Eqs. (5.10a) and (5.10b) in the finite-dimensional space $S_N := \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$, we consider the following ansatz,

$$\phi^N(t) = \sum_{n=1}^N a_n^N(t)\,\varphi_n, \qquad \mu^N(t) = \sum_{n=1}^N b_n^N(t)\,\varphi_n,$$

and the variational formulation

$$\langle \partial_t \phi^N, \bar\phi\rangle + \langle \nabla\mu^N, \nabla\bar\phi\rangle = \langle \dot m, \bar\phi\rangle_{\partial\Omega} \qquad \forall \bar\phi \in S_N, \tag{5.24a}$$

$$\langle \mu^N, \bar\mu\rangle = \varepsilon \langle \nabla\phi^N, \nabla\bar\mu\rangle + \frac{1}{\varepsilon}\langle f_m(\phi^N), \bar\mu\rangle + \langle f_w(\phi^N), \bar\mu\rangle_{\partial\Omega} \qquad \forall \bar\mu \in S_N, \tag{5.24b}$$

$$\langle \phi^N(0), \bar\phi\rangle = \langle \phi_0, \bar\phi\rangle \qquad \forall \bar\phi \in S_N. \tag{5.24c}$$

To this formulation corresponds the following system of ordinary differential equations, with

unknown functions $\{a_n^N\}_{n=1}^N$ and $\{b_n^N\}_{n=1}^N$:

$$(a_n^N)' = \lambda_n\, b_n^N + \langle \dot{m}, \varphi_n \rangle_{\partial\Omega}\,, \tag{5.25a}$$

$$b_n^N = \varepsilon\, \lambda_n\, a_n^N + \frac{1}{\varepsilon} \left\langle f_m\left(\textstyle\sum_{i=1}^N a_i^N\, \varphi_i\right), \varphi_n \right\rangle + \left\langle f_w\left(\textstyle\sum_{i=1}^N a_i^N\, \varphi_i\right), \varphi_n \right\rangle_{\partial\Omega}, \tag{5.25b}$$

$$a_n^N(0) = \langle \phi_0, \varphi_n \rangle\,, \tag{5.25c}$$

for $n = 1, \dots, N$. Local existence and uniqueness of a solution to this system of equations is guaranteed by the fact that that the right-hand side of (5.25a) depends continuously on the coefficients $\{a_n^N\}_{n=1}^N$. To show the existence of a global solution, we will use the *a priori* estimate presented in the following lemma.

Lemma 5.2. *Assume that* $|F_w(\phi)| \le C(1 + |\phi|^2)$. *Then the solution* (ϕ^N, μ^N) *to Eqs.* (5.24a) *to* (5.24c) *satisfies*

$$\frac{1}{2} \int_\Omega \left(\frac{1}{2}\, \varepsilon\, \left|\nabla \phi^N\right|^2 + \frac{1}{\varepsilon}\, F_m(\phi^N) \right) \mathrm{d}\Omega + \frac{1}{2} \int_{\Omega_T} \left|\nabla \mu^N\right|^2 \le C, \tag{5.26}$$

where C *is independent of* N *and* $\Omega_T := \Omega \times (0, T)$.

Proof. Setting $\bar{\phi} = \mu^N$, $\bar{\mu} = \partial_t \phi^N$ in Eqs. (5.24a) and (5.24b) and subtracting leads to the equation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \left[E_m(\phi^N) + E_w(\phi^N) \right] :=& \frac{\mathrm{d}}{\mathrm{d}t} \left[\int_\Omega \frac{1}{2}\, \varepsilon\, \left|\nabla \phi^N\right|^2 + \frac{1}{\varepsilon}\, F_m(\phi^N)\, \mathrm{d}\Omega + \int_{\partial\Omega} F_w(\phi^N)\, \mathrm{d}\sigma \right] \\ =& - \int_\Omega \left|\nabla \mu^N\right|^2 \mathrm{d}\Omega + \int_{\partial\Omega} \dot{m}\, \mu^N\, \mathrm{d}\sigma. \end{aligned} \tag{5.27}$$

Using a trace inequality, Hölder's inequality, and Young's inequality with a parameter, we have, for all $u \in H^1(\Omega)$,

$$\begin{aligned} \|F_w(u)\|_{L^1(\partial\Omega)} &\le C \|1 + |u|^2\,\|_{L^1(\partial\Omega)} \\ &\le C\, \left(1 + \|u^2\|_{L^1(\Omega)} + \|\nabla(u^2)\|_{L^1(\Omega)}\right) \\ &= C\, \left(1 + \|u^2\|_{L^1(\Omega)} + 2\|u\, \nabla u\|_{L^1(\Omega)}\right) \\ &\le C\, \left(1 + \|u^2\|_{L^1(\Omega)} + \frac{1}{\alpha}\|u^2\|_{L^1(\Omega)} + \alpha \|\nabla u\|_{L^2(\Omega)}^2\right) \qquad \forall \alpha > 0. \end{aligned}$$

Now we use the simple fact that, for any $\beta > 0$ and $0 \le s \le t$, the inequality $|x|^s \le \beta^s + \beta^{s-t}\, |x|^t$ holds true for all $x \in \mathbf{R}$, to obtain

$$\|F_w(u)\|_{L^1(\partial\Omega)} \le C + \frac{1}{2} E_m(u) \tag{5.28}$$

for a constant C independent of u.

In addition, using a trace inequality, Poincaré inequality, and (5.24b) with $\bar{\mu} = 1$,

$$\begin{aligned}\left|\int_{\partial\Omega} \dot{m}\,\mu^N \,\mathrm{d}\sigma\right| &\le \left|\int_{\partial\Omega} \dot{m}\left(\mu^N - \frac{1}{|\Omega|}\int_\Omega \mu^N \,\mathrm{d}\Omega\right)\mathrm{d}\sigma\right| + \left|\frac{1}{|\Omega|}\int_{\partial\Omega}\dot{m}\,\mathrm{d}\sigma\int_\Omega \mu^N\,\mathrm{d}\Omega\right| \\ &\le C\,\|\dot{m}\|_{L^2(\partial\Omega)}\|\nabla\mu^N\|_{L^2(\Omega)} + \frac{1}{|\Omega|}\int_{\partial\Omega}|\dot{m}|\,\mathrm{d}\sigma\left|\frac{1}{\varepsilon}\left\langle f_m(\phi^N),1\right\rangle + \left\langle f_w(\phi^N),1\right\rangle_{\partial\Omega}\right| \\ &\le C\,\|\dot{m}\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\|\nabla\mu^N\|^2_{L^2(\Omega)} + C\,\|\dot{m}\|_{L^2(\partial\Omega)}\left(E_m(\phi^N)+1\right). \end{aligned} \tag{5.29}$$

Integrating (5.27) in time, and rearranging using Eqs. (5.28) and (5.29),

$$\begin{aligned}\frac{1}{2}E_m(\phi^N(t)) + \frac{1}{2}\int_{\Omega_t}\left|\nabla\mu^N\right|^2 &\le C + \frac{3}{2}E_m(\phi^N(0)) + C\int_{\partial\Omega_T}|\dot{m}|^2 + C\int_0^t \|\dot{m}\|_{L^2(\partial\Omega)}\,E_m(\phi^N)\,\mathrm{d}s, \\ &\le C + C\int_0^t \|\dot{m}\|_{L^2(\partial\Omega)}\,E_m(\phi^N(s))\,\mathrm{d}s\end{aligned}$$

where we used the notations Ω_t and $\partial\Omega_t$, $t > 0$, to denote $\Omega\times(0,t)$ and $\partial\Omega\times(0,t)$, respectively. The last inequality holds by the assumptions that $\phi_0 \in H^1(\Omega)$ and $\dot{m} \in C([0,T];L^2(\partial\Omega))$. Using a Grönwall inequality, we have Eq. (5.26). □

By integration by parts of the first term in Eq. (5.26), we obtain $\sum_{n=1}^N \lambda_n (a_n^N)^2 < C$. This result, together with the inequality

$$a_1^N(t) = a_1^N(0) + \int_0^T\int_{\partial\Omega}\dot{m}\,\mathrm{d}\sigma \le a_1^N(0) + C\,\|\dot{m}\|_{C([0,T];L^2(\partial\Omega)}$$

implied by Equation (5.25a) and the fact that $\lambda_1 = 0$, show that the coefficients $\left\{a_n^N\right\}_{n=1}^N$ do not blow up, and by Eq. (5.25b) neither do the coefficients $\left\{b_n^N\right\}_{n=1}^N$, implying global existence.

In addition to (5.26), we have the usual estimate on $\partial_t\phi^N$: denoting by Π^N the $L^2(\Omega)$ projection on S_N, for all $\psi \in L^2\left(0,T;H^1(\Omega)\right)$ the following holds:

$$\begin{aligned}\left|\int_{\Omega_T}\frac{\partial\phi^N}{\partial t}\,\psi\right| &= \left|\int_{\Omega_T}\frac{\partial\phi^N}{\partial t}\left(\Pi^N\psi\right)\right| \\ &= \left|\int_{\Omega_T}\nabla\mu^N\cdot\nabla(\Pi^N\psi)\right| + \left|\int_{\partial\Omega_T}\dot{m}\left(\Pi^N\psi\right)\right| \\ &\le \left(\int_{\Omega_T}\left|\nabla\mu^N\right|^2\right)^{\frac{1}{2}}\left(\int_{\Omega_T}\left|\nabla(\Pi^N\psi)\right|^2\right)^{\frac{1}{2}} + \left(\int_{\partial\Omega_T}|\dot{m}|^2\right)^{\frac{1}{2}}\left(\int_{\partial\Omega_T}\left|\Pi^N\psi\right|^2\right)^{\frac{1}{2}} \\ &\le C\|\psi\|_{L^2(0,T;H^1(\Omega))}. \end{aligned} \tag{5.30}$$

This shows that $\|\partial_t\phi^N\|_{L^2(0,T;(H^1(\Omega))')} \le C$.

Let p be such that the embedding $H^1(\Omega)\subset L^p(\Omega)$ is compact, i.e., by Rellich–Kondrachov theorem, $p < \infty$ if $d = 1$ or $d = 2$, and $p < \frac{2d}{d-2}$ if $d > 2$. Using Eqs. (5.26) and (5.30), we can apply results (5.23a) and (5.23b) to our case, with $X = H^1(\Omega)$, $Y = L^p(\Omega)$ and $Z = (H^1(\Omega))'$,

to conclude that there exists a subsequence such that

$$\phi^N \to \phi \quad \text{weak-* in } L^\infty\left(0,T;H^1(\Omega)\right), \tag{5.31a}$$
$$\frac{\partial \phi^N}{\partial t} \to \frac{\partial \phi}{\partial t} \quad \text{weakly in } L^2\left(0,T;(H^1(\Omega))'\right), \tag{5.31b}$$
$$\phi^N \to \phi \quad \text{strongly in } C([0,T],L^p(\Omega)), \tag{5.31c}$$
$$\phi^N \to \phi \quad \text{strongly in } L^2\left(0,T;L^p\left(\Omega\right)\right). \tag{5.31d}$$

when $N \to \infty$. In addition, note that since ϕ^N is bounded in $L^\infty\left(0,T;L^2\left(\partial\Omega\right)\right)$, there is a subsequence such that $\phi^N \to v$ weak-* in $L^\infty\left(0,T;L^2\left(\partial\Omega\right)\right)$ for some function v in that space, and thus also $\phi^N \to v$ weakly in the coarser $L^2\left(0,T;L^2\left(\partial\Omega\right)\right)$. But also $\phi^N \to \phi$ weakly in $L^2\left(0,T;L^2\left(\partial\Omega\right)\right)$, because $\phi^N \to \phi$ weakly in $L^2\left(0,T;H^1(\Omega)\right)$ and by continuity of the trace operator (indeed, an operator between Hilbert spaces that is continuous in the strong topologies, is continuous in the weak ones too), so $v = \phi$. The same reasoning can be applied to $f_w(\phi^N)$, taking into account that $f_w : u \to f_w(u)$ is continuous on $L^2\left(0,T;L^2\left(\partial\Omega\right)\right)$, to conclude

$$\begin{aligned} \phi^N &\to \phi \quad \text{weak-* in } L^\infty\left(0,T;L^2\left(\partial\Omega\right)\right), \\ f_w(\phi^N) &\to f_w(\phi) \quad \text{weak-* in } L^\infty\left(0,T;L^2\left(\partial\Omega\right)\right). \end{aligned} \tag{5.32}$$

Regarding the chemical potential, testing (5.24b) with $\bar{\mu} = 1$ implies that:

$$\int_\Omega \mu \,\mathrm{d}\Omega = \int_\Omega \frac{1}{\varepsilon} f_m(\phi^N)\,\mathrm{d}\Omega + \int_{\partial\Omega} f_w(\phi^N)\,\mathrm{d}\sigma,$$

which, together with the energy estimate (5.26), implies that μ^N is bounded in $L^2\left(0,T;H^1(\Omega)\right)$, leading to the existence of a further subsequence such that

$$\mu^N \to \mu \quad \text{weakly in } L^2\left(0,T;H^1(\Omega)\right). \tag{5.33}$$

Proceeding in a standard fashion, we consider an integer M and arbitrary functions $\phi^M, \mu^M \in C([0,T],H^1(\Omega))$ such that

$$\phi^M = \sum_{n=1}^{M} \bar{a}_n(t)\,\varphi_n, \qquad \mu^M = \sum_{n=1}^{M} \bar{b}_n(t)\,\varphi_n,$$

with $\{\bar{a}_n\}_{n=1}^M$, $\{\bar{b}_n\}_{n=1}^M$ smooth functions. Using ϕ^M and μ^M as test functions in Eqs. (5.24a) and (5.24b), integrating in time, taking the limit $N \to \infty$, and using the convergence results given in Eqs. (5.31a) to (5.31d), (5.32), and (5.33), we obtain

$$\begin{aligned} &\int_0^T \left\langle \partial_t \phi, \phi^M \right\rangle \mathrm{d}t + \int_0^T \left\langle \nabla\mu, \nabla\phi^M \right\rangle \mathrm{d}t = \int_0^T \left\langle \dot{m}, \phi^M \right\rangle_{\partial\Omega} \mathrm{d}t, \\ &\int_0^T \left\langle \mu, \mu^M \right\rangle \mathrm{d}t = \int_0^T \varepsilon \left\langle \nabla\phi, \nabla\mu^M \right\rangle \mathrm{d}t + \int_0^T \frac{1}{\varepsilon} \left\langle f_m(\phi), \mu^M \right\rangle \mathrm{d}t + \int_0^T \left\langle f_w(\phi), \mu^M \right\rangle_{\partial\Omega} \mathrm{d}t, \end{aligned}$$

from which we conclude using a standard density argument. □

Chapter 6

Software contributions

6.1 A Python library for the Hermite spectral method 219
6.1.1 Library settings 221
6.1.2 Calculation of the discrete Hermite transform 222
6.1.3 Discretization of operators 224
6.1.4 Examples 225
6.1.5 Outlook 228
6.2 Finite-element framework for Cahn–Hilliard equation 229

In this chapter, we present the software contributions that emerged from the research in this thesis. Section 6.1 introduces and describes the essential features of *Hermipy*, a *Python* library initially designed to automate the Hermite Galerkin method and later extended to add support for trigonometric basis functions in a periodic setting. Section 6.2 then briefly describes the finite element framework we built on top of *FreeFem++* [106] to efficiently solve the Cahn–Hilliard (CH) and Cahn–Hilliard–Navier–Stokes (CHNS) equations.

6.1 A Python library for the Hermite spectral method

While working on Chapter 3, we developed *Hermipy*, a thin Python library that allows automating most of the operations involved in implementing a Hermite spectral method. The library uses the data structures provided by the *NumPy* library for linear algebra, and it also depends on *SymPy* for symbolic calculations. Computationally intensive parts of the computations, such as Hermite transforms, are handed over to a C++ compiled component, and *Boost* is used for efficient exchange of data with Python. The code is freely available [216], under the conditions of the GNU General Public License, version 3. Installation instructions are available on the project page; for users of the *Arch Linux* distribution, the package can be installed via the *Arch User Repository*, where the package is hosted under the name *python-hermipy-git*. The choice of the *Python* language was motivated by the following reasons:

Software freedom : The reference implementation of *Python* is open source, with the important

Language	files	blank	comment	code
Python	40	1606	1537	5361
C++	19	463	401	2408
Markdown	2	122	0	593
C/C++ Header	15	167	304	526
CMake	1	15	17	50
make	3	11	7	35
Scheme	1	1	0	34
SUM:	82	2385	2266	9013

Table 6.1: Lines of code of *Hermipy*, obtained by using the `cloc` utility with the command `cloc $(git ls-files)` in the *Git* repository of the code.

practical implication that the ability to run a *Python* program distributed under a free licence is not dependent upon the continual existence of a third-party organization.

Ecosystem : There exist a number of mature *Python* libraries for computational mathematics, such as NumPy, SciPy and SymPy, which make the *Python* language particularly well-suited for the development of mathematical software. More recently, some work has also been invested in the development of the Numba library, which brings just-in-time (JIT) compilation to *Python* and provides scope for writing more efficient code than was possible previously.

Extensibility : It is possible to write *C* or *C++* extensions for *Python*, which can be used to improve the efficiency of a *Python* program when appropriate. With Boost.Python and its recent NumPy extension, the exchange of data between *Python* and *C++* or *C* is straightforward and efficient.

When the development of *Hermipy* began, the only *Python* library known to the author for automating spectral methods in *Python* was *shenfun*, which exposes an interface similar to that of FEniCS, a collection of free software components that enables automating the finite element method. The focus of *shenfun* is on solving PDEs in bounded boxes, by means of tensor products of Fourier, Chebyshev or Legendre bases. In contrast, *Hermipy* aims to provide the tools for automating spectral methods in unbounded domains, using bases of Hermite polynomials or Hermite functions.

Hermipy is developed in the open in the hope that it will be useful to other people and foster future research. As part of the development of the library, unit tests were written to ensure the good functionality of the different features. Every time new code is *pushed* to the *Git* repository of the code, a Jenkins automation server runs the test suite and performs a coverage analysis that assesses the quality of the tests. This development process ensures that bugs introduced by new code are discovered early. Some statistics on *Hermipy*, including a detailed breakdown of the lines of code between programming languages, as well as information on the test coverage of the module, are presented in Tables 6.1 and 6.2.

Module name	# statements	# miss	coverage
`__init__.py`	10	0	100
`cache.py`	117	20	83
`core.py`	226	18	92
`equations.py`	212	97	54
`function.py`	142	38	73
`lib.py`	16	0	100
`operator.py`	101	40	60
`position.py`	97	19	80
`quad.py`	377	184	51
`series.py`	173	90	48
`varf.py`	179	104	42
TOTAL	1753	675	61

Table 6.2: Test coverage of the submodules comprised in *Hermipy*, obtained using the `coverage` utility, `coverage run -source=hermipy -m unittest discover -v -f tests`.

6.1.1 Library settings

A dictionary of options can be modified to configure *Hermipy*. The default options that apply when the module is loaded are the following:

```
settings = {
        'cache': False,
        'cachedir': '.cache',
        'debug': False,
        'sparse': False,
        'tensorize': True,
        'trails': False,
        }
```

Changing the value of an option is achieved by modifying the corresponding dictionary entry. For example, the code below loads *Hermipy*, enables the caching feature, and sets the cache directory to `/tmp/hermipy-cache/`:

```
import hermipy as hm
hm.settings['cache'] = True
hm.settings['cachedir'] = '/tmp/hermipy-cache'
```

The available settings are described below.

cache* and *cachedir Setting `cache` to `True` enables the caching feature of *Hermipy*; with this option, the results of all the low-level calls to the C++ compiled component are saved for later reuse. These include in particular the results of Hermite transforms and of the computation of stiffness matrices. This feature works by hashing the arguments of function calls to test for availability of the result in the cache, and by using the *NumPy* functions `save` and `load` to write to and read from the cache. It is most useful when used in combination with the `tensorize` option.

tensorize In many applications, the Hermite or Fourier transform associated to functions in several dimensions can be obtained from lower dimensional transforms by tensorization,

reducing the computational cost significantly. Likewise, it is often possible to take advantage of the fact that many operators are tensorizable to speed up the computations of the corresponding stiffness matrices. By setting `tensorize` to `True`, *Hermipy* will automatically split functions and operators to make the computations more efficient.

sparse Setting `sparse` to `True` instructs *Hermipy* to use sparse matrices for the representation of stiffness matrices. This is useful only when the operator being discretized admits a sparse representation in the space of basis functions being considered, i.e. for operators with polynomial coefficients when using Hermite series, or operators with trigonometric coefficients when using Fourier series.

debug Setting `debug` to `True` enables the display of debug information when using *Hermipy*. This is useful mainly for debugging purposes by contributors.

trails Setting `trails` to `True` enables the display of information related to the function calls performed by *Hermipy*. A potential use of this feature is to identify the computational bottleneck of a simulation code, or to monitor the progress of computationally demanding functions. Like `debug`, this option does not influence the algorithms used for computations.

6.1.2 Calculation of the discrete Hermite transform

While it is in theory possible to calculate the Gauss–Hermite transform of a function and its inverse in $\mathcal{O}(d \log^2 d)$ operations [59, 139], where d is the number of Hermite polynomials used for the approximation, there does not yet in practice exist a stable method to do so. The approach we use is therefore less efficient, with a computational cost scaling as $\mathcal{O}(d^2)$ in one dimension.

We consider for simplicity an n-dimensional quadrature obtained by tensorization of one-dimensional quadratures of the type described in Appendix A.4, indexed by i and with nodes $(x_{i,j})_{j=1}^{N_i}$ and weights $(w_{i,j})_{j=1}^{N_i}$, as well as an index set $\mathcal{I}_A$ to support the Hermite transform. In order for the Hermite transform to be invertible, the index set $\mathcal{I}_A$ must have the same dimensions as the n-dimensional product quadrature, but other choices are possible if one abandons this requirement; one might for example wish to compute the Hermite coefficients of a function corresponding to all the multi-indices α such that $|\alpha|_s \leq d$, for some degree d and some norm s that determines the shape of the associated index set. Possible choices include $s = 1$ and $s = \infty$, which correspond respectively to triangle and square index sets when $n = 2$. We use the notation $\mathcal{I}_B$ for the index set associated with the quadrature, $\mathcal{I}_B := \{\beta \in \mathbf{N}^n : \beta_i \in \{1, 2, \cdots, N_i\}\}$ and, for $\beta \in \mathcal{I}_B$, we write $x_\beta := (x_{1,\beta_1}, x_{2,\beta_2}, \ldots, x_{n,\beta_n})$. The procedure that *Hermipy* uses to calculate Hermite transforms works in two steps:

- First, for each dimension i, the one-dimensional Hermite polynomials of degree less than or equal to $d_i := \max_{\alpha \in \mathcal{I}_A} \alpha_i$ are evaluated at the nodes of the $i-$th one-dimensional quadrature. To that purpose, the 3-term recursion relation (A.16) is used, which reduces the cost of this operation at each grid point to $\mathcal{O}(d_i)$. The total number of floating point operations required for this step thus scales as $\mathcal{O}(\sum_{i=1}^n N_i d_i)$.

- Second, for all $\alpha \in \mathcal{I}_A$ and $\beta \in \mathcal{I}_B$ the following products are evaluated:

$$p_{\alpha,\beta} = \prod_{i=1}^{n} (\hat{w}_{i,\beta_i}) \prod_{i=1}^{n} H_{\alpha_i}(x_{i,\beta_i}), \tag{6.1}$$

  where $\hat{w}_{i,j} = w_{i,j}$ in the case of the forward transform and $\hat{w}_{i,j} = 1$ in the case of the backward transform. We note that, if non-normalized Hermite polynomials had been used, including non-unit weights in both the forward and backward transforms would have been required. Carrying out the calculation of these products constitutes the computationally expensive part of the discrete Hermite transform. In the case of a full index set (i.e. when $\mathcal{I}_A$ has the same number of points as $\mathcal{I}_B$ along each direction), for example, the number of such products scales as $\mathcal{O}(\prod_{i=1}^{n} N_i^2)$. From these, the forward and backward discrete transforms are calculated as follows:

$$f_\alpha = \sum_{\beta \in \mathcal{I}_B} p_{\alpha,\beta}\, f(x_\beta), \qquad \text{(Forward transform)} \tag{6.2}$$

$$f(x_\beta) = \sum_{\alpha \in \mathcal{I}_A} p_{\alpha,\beta}\, f_\alpha. \qquad \text{(Backward transform)} \tag{6.3}$$

Often in applications, the function of which the Hermite transform is calculated can be expressed as a product (or a sum of products) of lower dimensional functions. In these cases, the discrete Hermite transform can be obtained by appropriate tensorization of lower dimensional transforms. Suppose for example that

$$f(x, y, z) = g(x, z)\, f(y), \tag{6.4}$$

and that the index set where we evaluate the Hermite transform, which we denote by $\mathcal{I}_{x,y,z}$, can be expressed as a subset of the product of two index sets of lower dimension, in the sense that that there exist $\mathcal{I}_{x,z}$ and $\mathcal{I}_y$ such that

$$\mathcal{I}_{x,y,z} \subset \left\{(\alpha_1, \alpha_2, \alpha_3) \in \mathbf{N}^3 : \alpha_2 \in \mathcal{I}_y \text{ and } (\alpha_1, \alpha_3) \in \mathcal{I}_{x,z}\right\}. \tag{6.5}$$

Then, denoting by $\mathcal{T}[\mathcal{I}]$ the discrete Hermite transform operator over the index set $\mathcal{I}$, the transform of f can be obtained from those of g and h:

$$\mathcal{T}[\mathcal{I}_{x,y,z}] f(\alpha_1, \alpha_2, \alpha_3) = \mathcal{T}[\mathcal{I}_{x,z}] g(\alpha_1, \alpha_3)\, \mathcal{T}[\mathcal{I}_y] h(\alpha_2) \quad \forall (\alpha_1, \alpha_2, \alpha_3) \in \mathcal{I}_{x,y,z}. \tag{6.6}$$

This procedure is applied seamlessly by *Hermipy* when the input of the transform can be expressed as a sum of products and the option `tensorize` is set to `True`, leading to a significant speedup compared to a naive computation. In these cases, the cost of computing the transforms of the factors is often negligible in comparison with the cost of the tensorization expressed in (6.6), which is still proportional to the cardinality of $\mathcal{I}_{x,y,z}$.

In addition to the usual discrete Hermite transform, *Hermipy* implements the generalized transform defined in Eq. (A.21), making it suitable for Galerkin methods based on not only Hermite polynomials but also Hermite functions and generalizations of these. When specified at the creation of a Gauss-Hermite quadrature, the multiplier function $\mathrm{e}^{-U/2}$ is taken into account in

all function calls to the discrete Hermite transform. This is made possible by the object-oriented programming paradigm available in Python, which also enables storing, along with the coefficients of a Hermite series, the parameters used for the transform, including the multiplier function, the scaling factor, and the supporting index set.

6.1.3 Discretization of operators

The second operation that is ubiquitous in the implementation of Hermite spectral methods, and indeed more generally in Galerkin methods, is the discretization of operators. A design goal of our library was to make this operation as user-friendly and efficient as possible. Regarding this former objective, *Hermipy* makes extensive use of the *SymPy* library, which provides an elegant and fully fledged language for operating on symbolic functions and operators. With regard to efficiency, *Hermipy* takes advantage of the often sparse nature of the discretized operators to speed up calculations and store data more efficiently. It also uses tensorization of lower-dimensional operators whenever possible.

Let us consider a general linear operator of order s over n-dimensional functions, with possibly varying coefficients,

$$\mathcal{L} = \sum_{|\gamma|_1 \leq s} c_\gamma(x) \frac{\partial^{|\gamma|_1}}{\partial x^\gamma},$$

as well as a basis $(H^\alpha)_{\alpha \in \mathbf{N}^n}$ of Hermite polynomials and a multiplier function $\mathrm{e}^{-U/2}$ to allow for different weights, and let us assume for simplicity that no spatial scaling of the Hermite basis is employed. Discretizing this operator requires the calculation of the inner products

$$L_{\alpha\,\beta} := \left\langle \mathcal{L}(\mathrm{e}^{-U/2}\, H_\alpha), \mathrm{e}^{-U/2}\, H_\beta \right\rangle_{g\,\mathrm{e}^U} = \left\langle (\mathrm{e}^{U/2}\, \mathcal{L}\, \mathrm{e}^{-U/2})\, H_\alpha, H_\beta \right\rangle_g \tag{6.7}$$

where g is the multidimensional Gaussian weight associated with the basis of Hermite polynomials. The weight $g\,\mathrm{e}^U$ is the natural choice for Galerkin methods given the multiplier function $\mathrm{e}^{-U/2}$; it ensures that the discretization of the identity operator is the identity matrix. The expression on the right-hand side of (6.7) is computationally more convenient because it enables working with the usual Hermite polynomials. Its calculation requires an expression of the mapped operator $\mathrm{e}^{U/2}\, \mathcal{L}\, \mathrm{e}^{-U/2}$, which our library performs automatically using the tools provided by *SymPy*.

We assume now, without loss of generality, that $U = 0$ (i.e. $\mathrm{e}^{-U/2} = 1$), and we remark that, in the same manner as how a partial differential operator with trigonometric coefficient admits a sparse representation in Fourier space, a partial differential operator with polynomial coefficients admits a sparse representation in Hermite space. It is important, in these cases, that the multi-indices α supporting the Hermite transform be ordered in such a way that the bandwidth of the resulting matrix is close to its minimum possible value. How this is achieved depends on the problem; with a triangular index set, one could for example use an order, i.e. a bijective function between $\{1, 2, \ldots, |\mathcal{I}_A|\}$ and the set $\mathcal{I}_A$ of multi-indices, such that the degree $|\alpha|_1$ of a multi-index is a non-decreasing function of the linear index. An advantage of this ordering is that obtaining the discretization of the operator on a smaller index set becomes trivial, which is very convenient e.g. when performing convergence studies. Likewise, it is natural

to order the multi-indices of a cubic index set by their $\|\cdot\|_\infty$ norm. In addition to these two choices, *Hermipy* provides several other different orderings multi-indices to choose from, such as the lexicographic ordering and one suited for approximations on the hyperbolic cross.

Regarding tensorization, the strategy applied for the discretization of operators is similar in spirit to that described in Section 6.1.2 for the Hermite transform; the operator $\mathcal{L}$ is broken down into its constitutive parts as much as possible, which are then discretized separately and assembled by tensorization to obtain the matrix representation of $\mathcal{L}$. Suppose for instance that $n = 2$ and that $\mathcal{L}$ contains a term of the form

$$f_1(x_1)\, f_2(x_2)\, \frac{\partial^2}{\partial x_1 \partial x_2}.$$

Then the corresponding contribution to $L_{\alpha\beta}$, see Eq. (6.7), is equal to

$$\begin{aligned}\left\langle f_1(x_1)\, f_2(x_2)\, \frac{\partial^2 H_\alpha}{\partial x_1 \partial x_2}, H_\beta \right\rangle_g &= \left\langle f_1\, H'_{\alpha_1}, H_{\beta_1} \right\rangle_{g_1} \times \left\langle f_2\, H'_{\alpha_2}, H_{\beta_2} \right\rangle_{g_2} \\ &= \sqrt{\alpha_1 \alpha_2} \times \left\langle f_1\, H_{\alpha_1 - 1}, H_{\beta_1} \right\rangle_{g_1} \times \left\langle f_2\, H_{\alpha_2 - 1}, H_{\beta_2} \right\rangle_{g_2},\end{aligned}$$

where we used the recursion relation (A.13) in the appendix, and where g_1, g_2 are the marginal densities of g. Since the potentials considered in Chapters 3 and 4 are all polynomial functions, this procedure was applied for all the simulations presented, but we note that *Hermipy* is also able to discretize operators that can't be decomposed in this way. It remains to explain the calculation of the products $\langle f_1 H_i, H_j \rangle_{g_1}$ (the same reasoning applies *mutatis mutandis* in higher dimensions). Rewriting this product as $\langle f_1, H_i\, H_j \rangle_{g_1}$ and using an explicit formula for the triple products of Hermite polynomials, see e.g. [210], this product can be computed from the Hermite transform of f_1. Since $H_i H_j$ is a polynomial of degree at most $2d$, where d is the maximum degree of the polynomials used for the approximation, the Hermite transform must be known up to degree $2d$ to ensure the exact computation of $\langle f_1, H_i\, H_j \rangle_{g_1}$ for $i, j \leq d$.

To conclude this subsection, we reiterate that *Hermipy* exposes an option to automatically save the results of computationally intensive parts of the calculation for later reuse. This is especially useful for parameter exploration – often one has to run the same simulation many times with mostly similar parameter sets, and in these cases the automatic caching functionality ensures that only the necessary parts are recalculated.

6.1.4 Examples

To conclude this section, we present two example *Python* scripts making use of *Hermipy*. The first solves the steady-state linear Fokker–Planck equation

$$\frac{\partial}{\partial x}\left(V'\rho + \theta(x-m)\rho - \frac{1}{\varepsilon}\sqrt{2\beta^{-1}}\,\eta\,\rho\right) + \frac{1}{\varepsilon^2}\frac{\partial}{\partial \eta}\left(\eta\,\rho + \frac{\partial \rho}{\partial \eta}\right) = 0, \tag{6.9}$$

where $\theta, m, \beta, \varepsilon$ are parameters and $\rho(x, \eta)$ is the unknown function. This equation corresponds to the McKean–Vlasov equation studied in Chapter 3 when the mean position is taken as a parameter (m) and the noise driving the particles is a scalar OU process. For the Galerkin

method we will use the following basis functions:

$$e_{ij}(x,\eta) = \exp\left(-\frac{1}{2}\left(\beta V(x) + \frac{x^2}{2\sigma_x^2}\right)\right) H_i\left(\frac{x}{\sigma_x}\right) \times \exp\left(-\frac{1}{2}\left(\frac{\eta^2}{2} + \frac{\eta^2}{2\sigma_\eta^2}\right)\right) H_j\left(\frac{\eta}{\sigma_\eta}\right),$$

where $\{H_i\}_{i=0,1,\dots}$ are the Hermite polynomials orthonormal for the standard Gaussian weight. Below the variables `y`, `sx`, `sy` correspond to η, σ_x^2 and σ_η^2 respectively. Knowing this, the code should be self-explanatory. We note in particular that the argument `factor`, used in the constructor of the quadrature, corresponds to the weight between the basis functions e_{ij} and the rescaled Hermite polynomials.

```
import hermipy as hm
import sympy as sym

# Set library options
hm.settings['sparse'] = True
hm.settings['tensorize'] = True

x, y = sym.symbols('x y', real=True)
f = sym.Function('f')(x, y)

# Set parameters for the Fokker-Planck equation
theta, beta, epsilon, m = .2, 1, .1, .2
Vp, Vy = x**4/4 - x**2/2 + theta*(x-m)**2/2, y**2/2

# Construct the Fokker-Planck operator
flux_x = - (Vp.diff(x)*f - sym.sqrt(1/beta)*y*f/epsilon)
flux_y = - (1/epsilon**2) * (f*Vy.diff(y) + f.diff(y))
fp_operator = - flux_x.diff(x) - flux_y.diff(y)

d = 50              # Number of basis functions in each direction
sx, sy = .05, .05   # Scaling of basis functions along x and y

# Set multiplier function
# (= ratio between basis functions and scaled Hermite polynomials)
factor = sym.exp(-x*x/sx/4 - y*y/sy/4) * sym.exp(- beta*Vp/2 - Vy/2)

# Define a quadrature
quad = hm.Quad.gauss_hermite(
        n_points=(2*d + 1),      # Number of quadrature points
        dim=2,                   # Dimension of the problem
        mean=[0, 0],             # Optional translation
        cov=[[sx, 0], [0, sy]],  # Scaling of basis functions
        factor=factor)           # Multiplier function

# Discretize the operator using (d + 1)**2 basis functions
discrete_operator = quad.discretize_op(fp_operator, d, index_set='cube')

# Calculate the eigenvector corresponding to the stationary solution
[e_val], [e_vec] = discrete_operator.eigs(k=1, which='LR')

# Plot the solution
```

```
quad.plot(e_vec)
```

Listing 6.1: *Python* script that uses *Hermipy* to solve the stationary linear Fokker–Planck equation associated with the McKean–Vlasov equation studied in Chapter 3. A similar script was used to generate numerically the self-consistency map, which maps m to the first moment of the solution to Eq. (6.9).

The second example we present is slightly more involved, and it is a simplified version of the code used to calculate the effective diffusion coefficient in Chapter 4. The script variables `x`, `y`, `z`, `w`, `a`, `b`, `g`, `n` correspond to $q, p, z_1, z_2, \alpha, \beta, \gamma, \nu$ in the notations of Chapter 4, respectively. Here we use the usual Hermite polynomials in the p, z_1 and z_2 directions. In the q direction, we use a Fourier expansion with the factor $\frac{\mathrm{e}^{\beta V/2}}{\mathcal{Z}_q}$, for a normalization constant $\mathcal{Z}_q$, to guarantee orthogonality in $L^2\left(\mathbf{R}, \mathrm{e}^{-\beta V}\right)$. In this case too, the code should be fairly self-explanatory. We note that, although it is convenient to think of them as arrays of numbers containing Fourier/Hermite coefficients, `rhs` and `solution` are in fact objects of the *Hermipy* `series` type, which contains not only an array of Fourier/Hermite coefficients but also additional information, such as the index set and ordering used for the discretization and the weight function of the generalized Hermite transform. The operator `*`, when acting on `series` objects as in the last line below, returns the inner product of its arguments in the appropriate weighted space.

```
# Import necessary modules
import math
import sympy as sym
import hermipy as hm
import scipy.integrate as integrate

# Set Hermipy settings
hm.settings['cache'] = True
hm.settings['tensorize'] = True
hm.settings['sparse'] = True

# Variables, corresponding to (q, p, z1, z2), and unknown function
x, y, z, w = sym.symbols('x y z w')
f = sym.Function('f')(x, y, z, w)

# Vector containing auxiliary variables
z_vec = sym.Matrix([z, w])

# Periodic potential
V = (1/2)*(1 - sym.cos(x))

# Physical parameters
b, a, n, g = 1, 1, 1, 1

# Vectors lambda and matrix A in the auxiliary processes
lamb = sym.Matrix([sym.sqrt(g)/n, 0])
A = sym.Matrix([[0, -a/n**2  ], [a/n**2, a**2/n**2]])

dfdz = sym.Matrix([f.diff(v) for v in z_vec])
ddfdzdz = sym.Matrix([dfdz.diff(v).T for v in z_vec])

# Backward Kolmogorov operator
```

```
backward = \
    y*f.diff(x) - V.diff(x)*f.diff(y) \
    + (lamb.dot(z_vec))*f.diff(y) \
    - y*lamb.dot(dfdz) \
    - (A*z_vec).dot(dfdz) \
    + (1/b)*(A*ddfdzdz).trace()

# Parameters for quadrature
kwargs = {'degree': 10, 'index_set': 'cube'}
npoints = kwargs['degree']*2 + 1

# Factor for Fourier series in q direction
zx = integrate.quad(sym.lambdify(x, sym.exp(-b*V)), -math.pi, math.pi)[0]
factor_x = math.sqrt(zx) * sym.exp(b/2*V)

# Define quadratures
qx = hm.Quad.fourier(npoints, dirs=[0], factor=factor_x)
qy = hm.Quad.gauss_hermite(npoints, dirs=[1], cov=[[1]], factor=1)
qz = hm.Quad.gauss_hermite(npoints, dirs=[2], cov=[[1]], factor=1)
qw = hm.Quad.gauss_hermite(npoints, dirs=[3], cov=[[1]], factor=1)

# Calculate the tensor product of the quadratures
quad = qx*qy*qz*qw

# Matrix discretizing operator
discretized_op = quad.discretize_op(backward, **kwargs)

# Right-hand side
rhs = quad.transform('y', **kwargs)

# Vector in the kernel
remove = quad.transform(1, **kwargs)

# Calculate solution
solution = (-discretized_op).solve(rhs, remove_vec=remove)

# Calculate the diffusion coefficient
diffusion = float(solution*rhs)
```

Listing 6.2: *Python* script using *Hermipy* to calculate the effective diffusion coefficient associated with the GL2 dynamics in a periodic potential, see Chapter 4.

We note that, in the two examples presented above, the entries of the stiffness matrices discretizing the operators could be calculated analytically using recursion relations of Hermite polynomials and standard trigonometric identities. However, this approach becomes very laborious and error-prone when scaling factors or non-cubic index sets are involved.

6.1.5 Outlook

So far in this chapter, we introduced the *Python* library that we developed for the automation of the Hermite/Fourier Galerkin method. We summarized the procedure followed by the library

to compute Hermite transforms and discretize operators, and we also showed, through explicit examples, how it can be used in practical applications. A more thorough documentation, that we hope to continue populating as the application programming interface (API) of *Hermipy* stabilizes, is available at `https://hermipy.readthedocs.io`.

Although the current capabilities of *Hermipy* were more than sufficient for the problems addressed in this thesis, several improvements are possible, pertaining to both efficiency and functionality.

In terms of efficiency, *Hermipy* could be improved by making use of a third-party library for the discrete Fourier transform, such as *fftw* [78], as opposed to implementing it from scratch. In addition, the wide access to computing clusters available today could be leveraged by implementing parallelization via the message passing interface (MPI), a feature already present in many of the popular frameworks for the automation of Galerkin methods.

Regarding functionality, the library could be extended in several manners: first, support could be added for additional families of orthogonal polynomials, such as Laguerre, Chebyshev and Jacobi polynomials. Since support for Chebyshev and Jacobi polynomials is already present in *shenfun* [160], a collaboration could be considered, although this would first require homogenizing the programming interfaces. Second, the library could be extended to handle complex numbers, which would be useful in a variety of scenarios, for instance to solve the classical Schödinger equations and related problems. Finally, *Hermipy* could provide the infrastructure for using Smolyak-like composite quadratures, which would alleviate the so-called *curse of dimensionality* affecting the efficiency of the Galerkin spectral methods in high dimensions. Some work in this direction has already been undertaken during the development of *Hermipy*, but additional efforts are required in order to ensure good performance and provide an interface similar to the one available for simple, tensor-product quadratures.

After these improvements are incorporated in *Hermipy*, our plan is to submit the package to the *Journal of Open Source Software*.

6.2 Finite-element framework for Cahn–Hilliard equation

In this section, we describe briefly the finite element framework we developed for the Cahn–Hilliard (CH) and Cahn–Hilliard–Navier–Stokes (CHNS) equation. Our code is freely available online [217], and it is also licensed under the GNU General Public License, version 3.

Our code was developed not only to produce the simulation results in Chapter 5, but also with the long term goal of assisting the investigation of microfluidic engineering applications, such as multiphase flows in microchannels and microseparators [192]. Essentially, it patches together a number of tools, in particular *FreeeFem++*, *gmsh*, *GNU Make* and *Python* scripts, in order to make the finite element analysis (FEA) of problems related to phase-field modeling as convenient as possible for researchers. It also addresses a lack of support for modularization of the *FreeFem++* language that makes it difficult to decouple parts of the code with different functions, such as those corresponding to the geometry, the initial and boundary conditions, the numerical method, and the visualization. Assume for example that one would like to carry out a FEA of three different problems, differing only by the associated initial and boundary conditions.

A natural way of achieving this in a high level language would be to create separate files with the initial and boundary conditions of each problem, and to include one of them, based on the value of a command line flag, in a larger file with the bulk of the FEA. This would enable the execution of a command along the lines of `FreeFem++ -i PROBLEM_FILE fea.pde` to perform the analysis for the initial and boundary conditions specified in `PROBLEM_FILE`. Achieving this in *FreeFem++* is surprisingly difficult; although the *FreeFem++* language exposes an `include` command, that command behaves similarly to the familiar *C/C++* `#include` preprocessor directive, i.e. it is preprocessed before the compilation stage and therefore cannot be used to include files defined in *FreeFem++* variables.

To remedy this issue and avoid code duplication, our code explicitly calls the *C* preprocessor, `cpp`, to assemble a standalone *FreeFem++* script from separate modules containing the problem definition as well as the solver configuration. This is achieved automatically via *GNU Make*, which serves as the interface between the users and the program. The specification of the modules to use, as well as the rest of the configuration for the test case, such as the geometry and some post-processing options, is centralized in a file consisting only of preprocessing directives. The code below, for example, was used for the study of the coalescence of two droplets on a substrate as presented in Section 5.4.3. The file `simple-square.geo`, referenced in the code, is a *gmsh* file describing a simple square geometry , and `problem.pde` is a *FreeFem++* file that contains the specification of the initial and boundary conditions.

```
// General
#define DIMENSION 2

// Postprocessing
#define VIEW GITROOT/sources/views/2D.geo

// Geometry
#define GEOMETRY_LX 2
#define GEOMETRY_LY 0.5
#define GEOMETRY GITROOT/sources/geometries/square/simple-square.geo

// File containing the initial and boundary conditions
#define PROBLEM_CONF HERE/problem.pde

// Plots
#define PLOT_FLAGS -e png -p -s 100 -C

/***********
*  Solver  *
************/

// Finite element space
#define SOLVER_POLYNOMIAL_ORDER 2

// Time step and number of iterations
#define SOLVER_NITER 1e6
#define SOLVER_TMAX 400
#define SOLVER_DT 1e-3
```

```
// Numerical methtod
#define SOLVER_METHOD OD2
#define SOLVER_BOUNDARY_CONDITION CUBIC

// Dimensionless numbers
#define SOLVER_PE 1e4
#define SOLVER_CN 0.01

// Mesh adaptation
#define SOLVER_MESH_ADAPTATION
#define SOLVER_MESH_ADAPTATION_HMIN 0.001
#define SOLVER_MESH_ADAPTATION_HMAX 0.01
```

Listing 6.3: Header file used in Section 5.4.3 to configure the simulation of the coalescence of two droplets on a substrate.

In addition to assembling the components required for the simulation of a phase-field problem specified as above, the framework we developed is able to automatically create an isolated environments for each problem. This is especially convenient when running a simulation several times with different sets of parameters, as it enables a clean separation between the data generated by different simulations and it allows these to be run simultaneously.

Our code also attempts to provide more advanced visualization capabilities than is possible using built-in *FreeFem++* functions. We wrote a number of custom scripts to streamline the exchange of data between *FreeFem++* and *gmsh*, as well as several *Python* scripts using *Matplotlib* in order to produce quality figures of simulation results.

With regard to numerical methods, our code implements the time-stepping schemes for the Cahn–Hilliard equation presented in Chapter 5, as well the Lagrange multiplier schemes LM1 and LM2 introduced in [98]. It also implements the time-step adaptation strategy we presented in Section 5.3.1 and the one introduced in [99]. For problems with advection, i.e. for the Cahn–Hilliard–Navier–Stokes equations, our code offers support for the splitting method described in Section 5.5, which relies on a projection scheme for the Navier–Stokes part of the system. At each time step, the energies (mixing free energy, wall free energy, and kinetic energy) are calculated and saved to data files.

Future development work on our code will focus mainly on the addition of support for parallel computations, for both the calculation of the stiffness matrices and the solution of linear systems.

Chapter 7

Conclusions

In this thesis, we studied several problems in the field of multiscale modeling and simulation. This chapter summarizes the results obtained and it presents perspectives for future work.

In Chapter 2, we developed a novel numerical method for the solution of multiscale SDEs. We showed, both theoretically and by means of extensive numerical experiments, that the method converges spectrally under appropriate conditions, and we demonstrated its applicability to systems of SDEs arising from the discretization of multiscale stochastic PDEs posed in simple domains. The key ingredient for the development of our method was the relation, rigorously established in [170, 169], between the effective drift and diffusion coefficients in the limit of large scale separation and the solution to an appropriate Poisson equation involving the generator of the fast dynamics. Given this relation, the coefficients of the effective SDE can be approximated through the solution of an elliptic PDE, which we accomplished by using a spectral method based on Hermite functions. Also important to the analysis and efficient numerical implementation of the method was the unitary transformation leading to a Schrödinger-type equation, which provided access to powerful functional-analytic results and enabled the theoretical derivation of convergence results. Although we restricted our attention to a situation where the fast variables of the multiscale systems of SDEs are confined by a potential in $\mathbf{R}^n$, it method can easily be adapted for application in a periodic setting, by simply using Fourier series instead of Hermite series. Our approach works particularly well when the dimension of the problem is relatively low but we point out that, being based on the deterministic numerical solution of a PDE, it is less efficient than probabilistic methods for high-dimensional problems. Some exciting questions in this line of research remain open for future work. One could investigate, for example, whether sparse quadratures [86, 41] and sparse spectral representations, such as the so-called hyperbolic cross approximation [61], could be used to alleviate the curse of dimensionality affecting the spectral method. One could also study preconditioning techniques with the aim of reducing the computational cost of solving the Poisson equation.

In Chapter 3, we studied the mean-field limit of a system of particles interacting through their mean (i.e., via a Curie–Weiss interaction) in a confining potential. In contrast with the setting adopted in [94], the thermal noise driving the particles was assumed to be colored, i.e. not delta-correlated in time. Passing formally to the mean-field limit for this model, we obtained a nonlinear Fokker–Planck equation, known in the literature as a McKean–Vlasov equation, for

the 1-particle distribution function. This mean-field equation differs from that in the white noise case in two important aspects: first, it is posed in a higher-dimensional phase space, with the number of extra dimensions depending on the model of the noise, and second, it does not admit an explicit one-parameter family of (possible) stationary solutions. In view of this latter difficulty, the study of phase transitions requires the explicit solution of the mean-field equation via a numerical method for PDEs, which we accomplished by using a Hermite spectral method based on our work in Chapter 2, but generalized appropriately to handle the quadratic nonlinearity. We assessed the accuracy of our method, when applied to the time-dependent McKean–Vlasov equation, by comparison of our numerical results with those obtained through the energy-decreasing, positivity-preserving finite volume scheme introduced in [37], on one hand, and with a known exact solution in a simple setting, however. In both cases the agreement was excellent and spectral convergence of our method was observed. Overall, it emerged from our study that, at the level of the bifurcation diagram of the first moment of the solution with respect to the temperature, the effect of colored noise depends on the specific model of the noise; for noise with a distribution that is symmetric around 0, the topology of the bifurcation diagram is the same as in the white noise case, only with a different critical temperature. For noise with a non-symmetric distribution, on the other hand, the two branches in the bifurcation diagram are disconnected, indicating that the system always reaches the same equilibrium on slowly decreasing the temperature. These findings were verified in the small correlation time regime using an alternative approach based on singular perturbation theory for Markov processes, and in doing so we extended the results presented in [115]. To the extent that it was computationally affordable, we also simulated the interacting particle system by Monte–Carlo simulation, which produced results in good agreement with the two other methods. An interesting subject for future research would be the rigorous study of the convergence of our method for the McKean–Vlasov equation in its full generality. It is likely that obtaining results in that direction would require the derivation of new *a priori* estimates for this equation, which could then be combined with the techniques in [100].

In Chapter 4, we studied the generalized Langevin equation (GLE) in a periodic potential. Our main aim in this part was to study the dependence of the effective diffusion coefficient on the parameters of the equation, particularly with respect to the memory of the system and the friction coefficient (γ). Here by friction coefficient we mean, by a slight abuse of terminology, the friction coefficient of the limiting Langevin equation as the amount of memory in the system goes to zero. Employing tools from multiscale analysis, we showed theoretically that the effective diffusion coefficient scales as $\mathcal{O}(\gamma^{-1})$ in the limits of both $\gamma \to 0$ (underdamped) and $\gamma \to \infty$ (overdamped), with a leading coefficient that can be calculated explicitly. To calculate the effective diffusion coefficient over a wide range of parameters, we developed a Fourier/Hermite spectral method akin to that introduced in [190] for the Langevin dynamics. With this method, we confirmed and complemented early results obtained in [119, 118] for two specific models of the noise: scalar Ornstein–Uhlenbeck and harmonic. Alongside the study of effective diffusion in a periodic potential, we investigated the rate of convergence of the GLE to equilibrium using the hypocoercivity techniques from [220], and we proposed a simple method, based on the solution of an optimization problem, for obtaining an explicit lower bound on the rate of convergence. We

hope that two important open problems related to this work, both pertaining to hypocoercivity, will be addressed in future research. It would be interesting, first, to further examine whether the $L^2(\mu)$ hypocoercivity framework presented in [57] can be extended to the GLE, and, second, to investigate whether either $H^1(\mu)$ or $L^2(\mu)$ hypocoercivity can be proved at the discrete level. Obtaining results in this latter direction would enable obtaining rigorous estimates for the consistency error of the spectral method, by extending the techniques presented in [190].

In Chapter 5, our focus was on the development of a numerical method for phase field models with wetting phenomena. Building on the work presented in [98], we developed a class of linear schemes for the Cahn–Hilliard equation with wetting boundary condition. The schemes we proposed, of order 1 (OD1-W) or 2 (OD2-W, OD2mod-W) in time, satisfy discrete mass and energy conservation laws consistent with those of the continuous equation. We obtained explicit formulas for the artificial energy dissipation they introduce, and we showed formally that, in the particular case of the method OD1-W, the sign of the numerical dissipation can be controlled when the time step is small enough. Although it is in principle possible to use the time-stepping schemes we proposed with different methods of discretization in space, in our work we used exclusively a conforming, mesh-adaptive finite element method, which is well-suited for complex domains and real-life applications. Through careful numerical experiments, we verified the convergence rates of the resulting fully discrete methods with respect to both the time step and the mesh size. We also introduced a novel time-step adaptation algorithm that is based on limiting the variation of free energy at each time step, and we showed that this algorithm enables a significant reduction of the computational time for two paradigmatic model systems. Future work based on our research could proceed in several directions. First, our scheme could be extended to three-component phase-field models with wetting, for example by using the bulk free energy proposed in [30]. It could also be generalized, at least to some extent, to non-constant or degenerate mobilities. The form of the mobility $b(\phi) = 1 - \phi^2$ is of particular interest, as it has been studied extensively for Cahn–Hilliard models without wetting, see e.g. [17, 18], and it is considered more physical than constant mobility; it guarantees that the phase field is confined to the interval $[-1, 1]$ and it prevents diffusive transport of the phase field away from the interface. Besides the question of uniqueness of a solution to the associated degenerate Cahn–Hilliard equation, which even without wetting has remained elusive for over two decades, such a form of the mobility presents multiple challenges at both the theoretical and the numerical levels, and it would likely require reworking our discretization and time-stepping methods. Finally, we believe that our preliminary work on the development of a numerical method for the Cahn–Hilliard–Navier–Stokes equations with wetting could be further developed and employed for the design and analysis of microfluidics devices, such as the capillary microseparators presented in [192, 82].

The research in this thesis also lead to the development of two software projects. First, we wrote a thin Python library, *Hermipy* [216], that enables automating most of the operations involved in implementing a Hermite spectral method. The library uses the data structures provided by the NumPy library for linear algebra [123], and it also depends on *SymPy* [155] for symbolic calculations. Computationally intensive parts of the computations, such as Hermite transforms, are handed over to a *C++* compiled component, and *Boost* is used for efficient exchange of data with *Python*. Second, we developed a high-level framework [217] for the numerical

solution to the Cahn–Hilliard and the Cahn–Hilliard–Navier–Stokes equations, implementing a number of semi-implicit schemes in *FreeFem++* [106] as well as the time-step adaptation strategies introduced in [99] and [11]. The framework offers users the means to rapidly define the initial and boundary conditions of the problem, run a particular numerical method with appropriate parameters, and post-process the simulation data to generate figures and graphs. Both two and three-dimensional settings are supported, and the geometries specified by files generated by the *Gmsh* mesh-generator [88].

Bibliography

[1] A. Abdulle and G. A. Pavliotis. Numerical methods for stochastic partial differential equations with multiple scales. *J. Comput. Phys.*, 231(6):2482–2497, Mar. 2012. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2011.11.039. URL: http://dx.doi.org/10.1016/j.jcp.2011.11.039.

[2] A. Abdulle, G. A. Pavliotis, and U. Vaes. Spectral methods for multiscale stochastic differential equations. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):720–761, 2017. ISSN: 2166-2525. DOI: 10.1137/16M1094117. URL: https://doi.org/10.1137/16M1094117.

[3] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden. The heterogeneous multiscale method. *Acta Numer.*, 21:1–87, 2012. ISSN: 0962-4929.

[4] S. Agmon. *Lectures on exponential decay of solutions of second-order elliptic equations: bounds on eigenfunctions of N-body Schrödinger operators*, volume 29 of *Mathematical Notes*. Princeton University Press, Princeton, NJ; University of Tokyo Press, Tokyo, 1982, page 118. ISBN: 0-691-08318-5.

[5] N. D. Alikakos, P. W. Bates, and X. Chen. Convergence of the Cahn-Hilliard equation to the Hele-Shaw model. *Arch. Rational Mech. Anal.*, 128(2):165–205, 1994. ISSN: 0003-9527. DOI: 10.1007/BF00375025. URL: https://doi.org/10.1007/BF00375025.

[6] S. M. Allen and J. W. Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metal. Mater.*, 27(6):1085–1095, 1979.

[7] E. L. Allgower and K. Georg. *Introduction to numerical continuation methods*, volume 45 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003, pages xxvi+388. ISBN: 0-89871-544-X. DOI: 10.1137/1.9780898719154. URL: https://doi.org/10.1137/1.9780898719154.

[8] D. M. Anderson, G. B. McFadden, and A. A. Wheeler. Diffuse-interface methods in fluid mechanics. *Annu. Rev. Fluid Mech.*, 30:139, 1998.

[9] G. E. Andrews, R. Askey, and R. Roy. *Special functions*, volume 71 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1999, pages xvi+664. ISBN: 0-521-62321-9; 0-521-78988-5. DOI: 10.1017/CBO9781107325937. URL: http://dx.doi.org/10.1017/CBO9781107325937.

[10] W. Arendt, C. J. K. Batty, M. Hieber, and F. Neubrander. *Vector-valued Laplace transforms and Cauchy problems*, volume 96 of *Monographs in Mathematics*. Birkhäuser Verlag,

Basel, 2001, pages xii+523. ISBN: 3-7643-6549-8. DOI: 10.1007/978-3-0348-5075-9. URL: https://doi.org/10.1007/978-3-0348-5075-9.

[11] B. Aymard, U. Vaes, M. Pradas, and S. Kalliadasis. A linear, second-order, energy stable, fully adaptive finite element method for phase-field modeling of wetting phenomena. *J. Comput. Phys. X*, 2:100010, 2019. ISSN: 2590-0552. DOI: https://doi.org/10.1016/j.jcpx.2019.100010. URL: http://www.sciencedirect.com/science/article/pii/S2590055219300265.

[12] D. H. Bailey and P. N. Swarztrauber. A fast method for the numerical evaluation of continuous Fourier and Laplace transforms. *SIAM J. Sci. Comput.*, 15(5):1105–1110, 1994. ISSN: 1064-8275. DOI: 10.1137/0915067. URL: https://doi.org/10.1137/0915067.

[13] N. Bain and D. Bartolo. Critical mingling and universal correlations in model binary active liquids. *Nat. Commun.*, 8:15969, 2017.

[14] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Commun. Probab.*, 13:60–66, 2008. ISSN: 1083-589X. DOI: 10.1214/ECP.v13-1352. URL: https://doi.org/10.1214/ECP.v13-1352.

[15] R. Balescu. *Statistical dynamics. Matter out of equilibrium.* Imperial College Press, London, 1997.

[16] A. Barone and G. Paterno. *Physics and applications of the Josephson effect.* Wiley, New York, 1982.

[17] J. W. Barrett, J. F. Blowey, and H. Garcke. Finite element approximation of the Cahn-Hilliard equation with degenerate mobility. *SIAM J. Numer. Anal.*, 37(1):286–318, 1999. ISSN: 0036-1429. DOI: 10.1137/S0036142997331669. URL: https://doi.org/10.1137/S0036142997331669.

[18] J. W. Barrett, R. Nürnberg, and V. Styles. Finite element approximation of a phase field model for void electromigration. *SIAM J. Numer. Anal.*, 42(2):738–772, 2004. ISSN: 0036-1429. DOI: 10.1137/S0036142902413421. URL: https://doi.org/10.1137/S0036142902413421.

[19] J. Benziger, J. Nehlsen, D. Blackwell, T. Brennan, and J. Itescu. Water flow in the gas diffusion layer of PEM fuel cells. *J. Memb. Sci.*, 261(1):98–106, 2005.

[20] A. L. Bertozzi, S. Esedoglu, and A. Gillette. Inpainting of binary images using the Cahn–Hilliard equation. *IEEE Transactions on image processing*, 16(1):285–291, 2007.

[21] J. Binney and S. Tremaine. *Galactic dynamics*, volume 20. Princeton University Press, 2011.

[22] G. Blankenship and G. C. Papanicolaou. Stability and control of stochastic systems with wide-band noise disturbances. I. *SIAM J. Appl. Math.*, 34(3):437–476, 1978. ISSN: 0036-1399. DOI: 10.1137/0134036. URL: https://doi.org/10.1137/0134036.

[23] D. Blömker, M. Hairer, and G. A. Pavliotis. Multiscale analysis for stochastic partial differential equations with quadratic nonlinearities. *Nonlinearity*, 20(7):1721–1744, 2007. ISSN: 0951-7715. DOI: 10.1088/0951-7715/20/7/009. URL: http://dx.doi.org/10.1088/0951-7715/20/7/009.

[24] J. F. Blowey and C. M. Elliott. The Cahn-Hilliard gradient theory for phase separation with nonsmooth free energy. I. Mathematical analysis. *European J. Appl. Math.*, 2(3):233–280, 1991. ISSN: 0956-7925. DOI: 10.1017/S095679250000053X. URL: https://doi.org/10.1017/S095679250000053X.

[25] V. I. Bogachev, N. V. Krylov, M. Röckner, and S. V. Shaposhnikov. *Fokker-Planck-Kolmogorov equations*, volume 207 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2015, pages xii+479. ISBN: 978-1-4704-2558-6. DOI: 10.1090/surv/207. URL: http://dx.doi.org/10.1090/surv/207.

[26] I. Bogdanov, S. Jardel, A. Turki, and A. Kamp. Pore-scale phase field model of two-phase flow in porous medium. *Proc. COMSOL Conf.*, 2010.

[27] D. Bonn, J. Eggers, J. Indekeu, J. Meunier, and E. Rolley. Wetting and spreading. *Rev. Mod. Phys.*, 81:739–805, 2009.

[28] V. Bonnaillie-Noël, J. A. Carrillo, T. Goudon, and G. A. Pavliotis. Efficient numerical calculation of drift and diffusion coefficients in the diffusion approximation of kinetic equations. *IMA J. Numer. Anal.*, 36(4):1536–1569, 2016. ISSN: 0272-4979. DOI: 10.1093/imanum/drv066. URL: https://doi.org/10.1093/imanum/drv066.

[29] F. Boyer, L. Chupin, and P. Fabrie. Numerical study of viscoelastic mixtures through a Cahn-Hilliard flow model. *Eur. J. Mech. B Fluids*, 23(5):759–780, 2004. ISSN: 0997-7546. DOI: 10.1016/j.euromechflu.2004.03.001. URL: https://doi.org/10.1016/j.euromechflu.2004.03.001.

[30] F. Boyer and C. Lapuerta. Study of a three component Cahn-Hilliard flow model. *M2AN Math. Model. Numer. Anal.*, 40(4):653–687, 2006. ISSN: 0764-583X. DOI: 10.1051/m2an:2006028. URL: https://doi.org/10.1051/m2an:2006028.

[31] F. Boyer and S. Minjeaud. Numerical schemes for a three component Cahn-Hilliard model. *ESAIM Math. Model. Numer. Anal.*, 45(4):697–738, 2011. ISSN: 0764-583X. DOI: 10.1051/m2an/2010072. URL: https://doi.org/10.1051/m2an/2010072.

[32] C.-E. Bréhier. Analysis of an HMM time-discretization scheme for a system of stochastic PDEs. *SIAM J. Numer. Anal.*, 51(2):1185–1210, 2013. ISSN: 0036-1429. DOI: 10.1137/110853078. URL: http://dx.doi.org/10.1137/110853078.

[33] A. J. Briant, A. J. Wagner, and J. M. Yeomans. Lattice Boltzmann simulations of contact line motion. I. Liquid-gas systems. *Phys. Rev. E*, 69(3):031602, 2004.

[34] J. W. Cahn. Critical point wetting. *J. Chem. Phys.*, 66:3667–3672, 1977.

[35] J. W. Cahn. On spinodal decomposition. *Acta Metal. Mater.*, 9(9):795–801, 1961.

[36] J. W. Cahn and J. E. Hilliard. Free energy of a nonuniform system. I. Interfacial free energy. *J. Chem. Phys.*, 28(2):258–267, 1958.

[37] J. A. Carrillo, A. Chertock, and Y. Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Commun. Comput. Phys.*, 17(1):233–258, 2015. ISSN: 1815-2406. DOI: 10.4208/cicp.160214.010814a. URL: https://doi.org/10.4208/cicp.160214.010814a.

[38] J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil. Particle, kinetic, and hydrodynamic models of swarming. In *Mathematical modeling of collective behavior in socio-economic and life sciences*, Model. Simul. Sci. Eng. Technol. Pages 297–336. Birkhäuser Boston, Inc., Boston, MA, 2010. DOI: doi:10.1007/978-0-8176-4946-3_12. URL: http://dx.doi.org/10.1007/978-0-8176-4946-3_12.

[39] J. S. Chang and G. Cooper. A practical difference scheme for Fokker-Planck equations. *J. Comput. Phys.*, 6(1):1–16, 1970.

[40] Y.-C. Chang, T. Y. Hou, B. Merriman, and S. Osher. A level set formulation of eulerian interface capturing methods for incompressible fluid flows. *J. Comput. Phys.*, 124(2):449–464, 1996.

[41] P. Chen. Sparse quadrature for high-dimensional integration with Gaussian measure. *ESAIM Math. Model. Numer. Anal.*, 52(2):631–657, 2018. ISSN: 0764-583X.

[42] A. J. Chorin. Numerical solution of the Navier-Stokes equations. *Math. Comp.*, 22:745–762, 1968. ISSN: 0025-5718. DOI: 10.2307/2004575. URL: https://doi.org/10.2307/2004575.

[43] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978, pages xix+530. ISBN: 0-444-85028-7. Studies in Mathematics and its Applications, Vol. 4.

[44] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Model. Simul.*, 7(2):842–864, 2008. ISSN: 1540-3459. DOI: 10.1137/070696325. URL: http://dx.doi.org/10.1137/070696325.

[45] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006. ISSN: 1063-5203. DOI: 10.1016/j.acha.2006.04.006. URL: http://dx.doi.org/10.1016/j.acha.2006.04.006.

[46] J. B. Collins and H. Levine. Diffuse interface model of diffusion-limited crystal growth. *Phys. Rev. B*, 31:6229, 1985.

[47] C. Cowan. *The Cahn-Hilliard equation as a gradient flow*. Master's thesis, Department of Mathematics, Simon Fraser University, 2005.

[48] R. G. Cox. The dynamics of the spreading of liquids on a solid surface. Part 1. Viscous flow. *J. Fluid Mech.*, 357:249–278, 1986.

[49] C. Cuthbertson, G. A. Pavliotis, A. Rafailidis, and P. Wiberg. Asymptotic analysis for foreign exchange derivatives with stochastic volatility. *Int. J. Theor. Appl. Finance*, 13(7):1131–1147, 2010. ISSN: 0219-0249. DOI: 10.1142/S0219024910006145. URL: http://dx.doi.org/10.1142/S0219024910006145.

[50] T. A. Davis. UMFPACK User Guide, 2011.

[51] D. A. Dawson. Critical dynamics and fluctuations for a mean-field model of cooperative behavior. *J. Statist. Phys.*, 31(1):29–85, 1983. ISSN: 0022-4715. DOI: 10.1007/BF01010922. URL: https://doi.org/10.1007/BF01010922.

[52] P.-G. De Gennes. Wetting: statics and dynamics. *Rev. Mod. Phys.*, 57(3):827, 1985.

[53] M. De Menech. Modeling of droplet breakup in a microfluidic T-shaped junction with a phase-field model. *Phys. Rev. E*, 73(3):031505, 2006.

[54] M. De Menech, P. Garstecki, F. Jousse, and H. A. Stone. Transition from squeezing to dripping in a microfluidic T-shaped junction. *J. Fluid Mech.*, 595:141–161, 2008.

[55] A. Dhooge, W. Govaerts, Y. A. Kuznetsov, W. Mestrom, A. M. Riet, and B. Sautois. MATCONT and CL MATCONT: Continuation toolboxes in MATLAB, 2006.

[56] H. Ding and P. D. M. Spelt. Wetting condition in diffuse interface simulations of contact line motion. *Phys. Rev. E*, 75(4):046708, 2007.

[57] J. Dolbeault, C. Mouhot, and C. Schmeiser. Hypocoercivity for kinetic equations with linear relaxation terms. *C. R. Math. Acad. Sci. Paris*, 347(9-10):511–516, 2009. ISSN: 1631-073X. DOI: 10.1016/j.crma.2009.02.025. URL: https://doi.org/10.1016/j.crma.2009.02.025.

[58] J. Dolbeault, C. Mouhot, and C. Schmeiser. Hypocoercivity for linear kinetic equations conserving mass. *Trans. Amer. Math. Soc.*, 367(6):3807–3828, 2015. ISSN: 0002-9947. DOI: 10.1090/S0002-9947-2015-06012-7. URL: https://doi.org/10.1090/S0002-9947-2015-06012-7.

[59] J. R. Driscoll, D. M. Healy Jr., and D. N. Rockmore. Fast discrete polynomial transforms with applications to data analysis for distance transitive graphs. *SIAM J. Comput.*, 26(4):1066–1099, 1997. ISSN: 0097-5397. DOI: 10.1137/S0097539792240121. URL: https://doi.org/10.1137/S0097539792240121.

[60] A. B. Duncan, S. Kalliadasis, G. A. Pavliotis, and M. Pradas. Noise-induced transitions in rugged energy landscapes. *Phys. Rev. E*, 94:032107, 3, Sept. 2016. DOI: 10.1103/PhysRevE.94.032107. URL: http://link.aps.org/doi/10.1103/PhysRevE.94.032107.

[61] D. Dũng, V. N. Temlyakov, and T. Ullrich. Hyperbolic Cross Approximation. *arXiv e-prints*:arXiv:1601.03978, Jan. 2016. arXiv: 1601.03978 [math.NA].

[62] M. H. Duong and G. A. Pavliotis. Mean field limits for non-Markovian interacting particles: convergence to equilibrium, GENERIC formalism, asymptotic limits and phase transitions. *Commun. Math. Sci.*, 16(8):2199–2230, 2018. ISSN: 1539-6746. DOI: 10.4310/CMS.2018.v16.n8.a7. URL: http://dx.doi.org/10.4310/CMS.2018.v16.n8.a7.

[63] A. Durmus, A. Eberle, A. Guillin, and R. Zimmer. An elementary approach to uniform in time propagation of chaos. *ArXiv e-prints*, May 2018. arXiv: 1805.11387 [math.PR].

[64] W. E. *Principles of multiscale modeling*. Cambridge University Press, Cambridge, 2011, pages xviii+466. ISBN: 978-1-107-09654-7.

[65] W. E, D. Liu, and E. Vanden-Eijnden. Analysis of multiscale methods for stochastic differential equations. *Comm. Pure Appl. Math.*, 58(11):1544–1585, 2005. ISSN: 0010-3640. DOI: 10.1002/cpa.20088. URL: http://dx.doi.org/10.1002/cpa.20088.

[66] W. E and J.-G. Liu. Projection method. I. Convergence and numerical boundary layers. *SIAM J. Numer. Anal.*, 32(4):1017–1057, 1995. ISSN: 0036-1429. DOI: 10.1137/0732047. URL: https://doi.org/10.1137/0732047.

[67] C. M. Elliott and S. Larsson. Error estimates with smooth and nonsmooth data for a finite element method for the Cahn-Hilliard equation. *Math. Comp.*, 58(198):603–630, S33–S36, 1992. ISSN: 0025-5718. DOI: 10.2307/2153205. URL: https://doi.org/10.2307/2153205.

[68] C. M. Elliott and H. Garcke. On the Cahn-Hilliard equation with degenerate mobility. *SIAM J. Math. Anal.*, 27(2):404–423, 1996. ISSN: 0036-1410. DOI: 10.1137/S0036141094267662. URL: https://doi.org/10.1137/S0036141094267662.

[69] C. M. Elliott and Z. Songmu. On the Cahn-Hilliard equation. *Arch. Rational Mech. Anal.*, 96(4):339–357, 1986. ISSN: 0003-9527. DOI: 10.1007/BF00251803. URL: https://doi.org/10.1007/BF00251803.

[70] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010, pages xxii+749. ISBN: 978-0-8218-4974-3. DOI: 10.1090/gsm/019. URL: http://dx.doi.org/10.1090/gsm/019.

[71] D. J. Eyre. Unconditionally gradient stable time marching the Cahn-Hilliard equation. In *Computational and mathematical models of microstructural evolution (San Francisco, CA, 1998)*. Volume 529, Mater. Res. Soc. Sympos. Proc. Pages 39–46. MRS, Warrendale, PA, 1998. DOI: 10.1557/PROC-529-39. URL: https://doi.org/10.1557/PROC-529-39.

[72] F. Farkhooi and W. Stannat. A complete mean-field theory for dynamics of binary recurrent neural networks. *ArXiv e-prints*, Jan. 2017. arXiv: 1701.07128 [physics.bio-ph].

[73] P. C. Fife. Models for phase separation and their mathematics. *Electron. J. Differential Equations*:No. 48, 26, 2000. ISSN: 1072-6691.

[74] J. C. M. Fok, B. Guo, and T. Tang. Combined Hermite spectral-finite difference method for the Fokker-Planck equation. *Math. Comp.*, 71(240):1497–1528, 2002. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-01-01365-5. URL: http://dx.doi.org/10.1090/S0025-5718-01-01365-5.

[75] T. D. Frank. *Nonlinear Fokker-Planck equations*. Springer Series in Synergetics. Springer-Verlag, Berlin, 2005, pages xii+404. ISBN: 3-540-21264-7. Fundamentals and applications.

[76] P. J. Frey. Mshmet 2.3. https://www.ljll.math.upmc.fr/frey/software.html, 2016.

[77] P. J. Frey and P.-L. George. *Mesh generation*. ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2008, page 848. ISBN: 978-1-84821-029-5. DOI: 10.1002/9780470611166. URL: https://doi.org/10.1002/9780470611166. Application to finite elements.

[78] M. Frigo and S. G. Johnson. Fftw: an adaptive software architecture for the fft. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 3, pages 1381–1384. IEEE, 1998.

[79] G. Froyland, G. A. Gottwald, and A. Hammerlindl. A computational method to extract macroscopic variables and their dynamics in multiscale systems. *SIAM J. Appl. Dyn. Syst.*, 13(4):1816–1846, 2014. ISSN: 1536-0040. DOI: 10.1137/130943637. URL: http://dx.doi.org/10.1137/130943637.

[80] G. Froyland, G. A. Gottwald, and A. Hammerlindl. A trajectory-free framework for analysing multiscale systems. *Phys. D*, 328/329:34–43, 2016. ISSN: 0167-2789. DOI: 10.1016/j.physd.2016.04.010. URL: http://dx.doi.org/10.1016/j.physd.2016.04.010.

[81] D. Furihata. A stable and conservative finite difference scheme for the Cahn-Hilliard equation. *Numer. Math.*, 87(4):675–699, 2001.

[82] W. A. Gaakeer, M. H. J. M. De Croon, J. Van Der Schaaf, and J. C. Schouten. Liquid–liquid slug flow separation in a slit shaped micro device. *Chem. Eng. J.*, 207:440–444, 2012.

[83] J. Gagelman and H. Yserentant. A spectral method for Schrödinger equations with smooth confinement potentials. *Numer. Math.*, 122(2):383–398, 2012. ISSN: 0029-599X. DOI: 10.1007/s00211-012-0458-8. URL: http://dx.doi.org/10.1007/s00211-012-0458-8.

[84] J. Garnier, G. Papanicolaou, and T.-W. Yang. Consensus convergence with stochastic effects. *Vietnam J. Math.*, 45(1-2):51–75, 2017. ISSN: 2305-221X. DOI: 10.1007/s10013-016-0190-2. URL: https://doi.org/10.1007/s10013-016-0190-2.

[85] J. Garnier, G. Papanicolaou, and T.-W. Yang. Large deviations for a mean field model of systemic risk. *SIAM J. Financial Math.*, 4(1):151–184, 2013. ISSN: 1945-497X. DOI: 10.1137/12087387X. URL: https://doi.org/10.1137/12087387X.

[86] A. Genz and B. D. Keister. Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight. *J. Comput. Appl. Math.*, 71(2):299–309, 1996. ISSN: 0377-0427. DOI: 10.1016/0377-0427(95)00232-4. URL: https://doi.org/10.1016/0377-0427(95)00232-4.

[87] T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3-4):209–232, 1998. ISSN: 1017-1398. DOI: 10.1023/A:1019129717644. URL: http://dx.doi.org/10.1023/A:1019129717644.

[88] C. Geuzaine and J.-F. Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *Internat. J. Numer. Methods Engrg.*, 79(11):1309–1331, 2009. ISSN: 0029-5981. DOI: 10.1002/nme.2579. URL: https://doi.org/10.1002/nme.2579.

[89] J. Glimm, J. W. Grove, X. L. Li, K. Shyue, Y. Zeng, and Q. Zhang. Three-dimensional front tracking. *SIAM J. Sci. Comput.*, 19(3):703–727, 1998.

[90] B. D. Goddard, A. Nold, N. Savva, G. A. Pavliotis, and S. K. General dynamical density functional theory for classical fluids. *Phys. Rev. Lett.*, 109:120603, 12, Sept. 2012. DOI: 10.1103/PhysRevLett.109.120603. URL: https://link.aps.org/doi/10.1103/PhysRevLett.109.120603.

[91] B. D. Goddard, G. A. Pavliotis, and S. Kalliadasis. The overdamped limit of dynamic density functional theory: rigorous results. *Multiscale Model. Simul.*, 10(2):633–663, 2012. ISSN: 1540-3459. DOI: 10.1137/110844659. URL: https://doi.org/10.1137/110844659.

[92] R. Gomer. Diffusion of adsorbates on metal surfaces. *Rep. Prog. Phys.*, 53(7):917, 1990.

[93] S. N. Gomes, S. Kalliadasis, G. A. Pavliotis, and P. Yatsyshin. Dynamics of the Desai-Zwanzig model in multiwell and random energy landscapes. *Phys. Rev. E*, 99:032109, 3, Mar. 2019. DOI: 10.1103/PhysRevE.99.032109. URL: https://link.aps.org/doi/10.1103/PhysRevE.99.032109.

[94] S. N. Gomes and G. A. Pavliotis. Mean field limits for interacting diffusions in a two-scale potential. *J. Nonlinear Sci.*, 28(3):905–941, 2018. ISSN: 0938-8974. DOI: 10.1007/s00332-017-9433-y. URL: https://doi.org/10.1007/s00332-017-9433-y.

[95] S. N. Gomes, G. A. Pavliotis, and U. Vaes. Mean-field limits for interacting diffusions with colored noise, 2019, in preparation.

[96] D. Gottlieb and S. A. Orszag. *Numerical analysis of spectral methods: theory and applications*. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1977, pages v+172. CBMS-NSF Regional Conference Series in Applied Mathematics, No. 26.

[97] D. Gueyffier, J. Li, A. Nadim, R. Scardovelli, and S. Zaleski. Volume-of-fluid interface tracking with smoothed surface stress methods for three-dimensional flows. *J. Comput. Phys.*, 152(2):423–456, 1999.

[98] F. Guillén-González and G. Tierra. On linear schemes for a Cahn-Hilliard diffuse interface model. *J. Comput. Phys.*, 234:140–171, 2013. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2012.09.020. URL: https://doi.org/10.1016/j.jcp.2012.09.020.

[99] F. Guillén-González and G. Tierra. Second order schemes and time-step adaptivity for Allen-Cahn and Cahn-Hilliard models. *Comput. Math. Appl.*, 68(8):821–846, 2014. ISSN: 0898-1221. DOI: 10.1016/j.camwa.2014.07.014. URL: https://doi.org/10.1016/j.camwa.2014.07.014.

[100] B.-Y. Guo. Error estimation of Hermite spectral method for nonlinear partial differential equations. *Math. Comp.*, 68(227):1067–1078, 1999. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-99-01059-5. URL: https://doi.org/10.1090/S0025-5718-99-01059-5.

[101] M. E. Gurtin, D. Polignone, and J. Viñals. Two-phase binary fluids and immiscible fluids described by an order parameter. *Math. Models Methods Appl. Sci.*, 6(06):815–831, 1996.

[102] M. Hairer and J. C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*. Volume 63, Progr. Probab. Pages 109–117. Birkhäuser/Springer Basel AG, Basel, 2011. DOI: 10.1007/978-3-0348-0021-1_7. URL: https://doi.org/10.1007/978-3-0348-0021-1_7.

[103] M. Hairer and G. A. Pavliotis. Periodic homogenization for hypoelliptic diffusions. *J. Statist. Phys.*, 117(1-2):261–279, 2004. ISSN: 0022-4715. DOI: 10.1023/B:JOSS.0000044055.59822.20. URL: https://doi.org/10.1023/B:JOSS.0000044055.59822.20.

[104] P. Hänggi. Generalized langevin equations: a useful tool for the perplexed modeller of nonequilibrium fluctuations? In *Stochastic dynamics*, pages 15–22. Springer, 1997.

[105] P. Hanggi and P. Jung. Colored Noise in Dynamical Systems. *Adv. in Chem. Phys.*, 89:239–326, Jan. 1995.

[106] F. Hecht. New development in FreeFem++. *J. of Numer. Math.*, 20(3-4):251–266, 2012.

[107] F. Hecht and B. Mohammadi. Mesh adaption by metric control for multi-scale phenomena and turbulence. In *35th Aerospace Sciences Meeting and Exhibit*, page 859, 1997.

[108] B. Helffer and F. Nier. *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*, volume 1862 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2005, pages x+209. ISBN: 3-540-24200-7. DOI: 10.1007/b104762. URL: http://dx.doi.org/10.1007/b104762.

[109] F. Hérau. Short and long time behavior of the Fokker–Planck equation in a confining potential and applications. *J. Funct. Anal.*, 244(1):95–118, 2007.

[110] A. Hernández-Machado, J. Soriano, A. Lacasta, M. Rodríguez, L. Ramírez-Piscina, and J. Ortín. Interface roughening in Hele-Shaw flows with quenched disorder: experimental and theoretical results. *Europhys. Lett.*, 55:194, 2001.

[111] D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.*, 43(3):525–546, 2001. ISSN: 0036-1445. DOI: 10.1137/S0036144500378302. URL: https://doi.org/10.1137/S0036144500378302.

[112] D. J. Higham, X. Mao, and A. M. Stuart. Strong convergence of Euler-type methods for nonlinear stochastic differential equations. *SIAM J. Numer. Anal.*, 40(3):1041–1063, 2002. ISSN: 0036-1429. DOI: 10.1137/S0036142901389530. URL: https://doi.org/10.1137/S0036142901389530.

[113] P. D. Hislop and I. M. Sigal. *Introduction to spectral theory*, volume 113 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1996, pages x+337. ISBN: 0-387-94501-6. DOI: 10.1007/978-1-4612-0741-2. URL: http://dx.doi.org/10.1007/978-1-4612-0741-2. With applications to Schrödinger operators.

[114] L. M. Hocking. The spreading of a thin drop by gravity and capillarity. *Q. J. Mech. Appl. Maths*, 36:55–69, 1983.

[115] W. Horsthemke and R. Lefever. Noise-induced transitions: theory and application in physics, chemistry, and biology. *Springer Series in Synergetics, Springer-Verlag, Berlin*, 15, 1984.

[116] J.-J. Huang, H. Huang, and X. Wang. Wetting boundary conditions in numerical simulation of binary fluids by using phase-field method: some comparative studies and new development. *Internat. J. Numer. Methods Fluids*, 77(3):123–158, 2015. ISSN: 0271-2091. DOI: 10.1002/fld.3975. URL: https://doi.org/10.1002/fld.3975.

[117] C. Huh and L. E. Scriven. Hydrodynamic model of steady movement of a solid/liquid/fluid contact line. *J. Coll. Interface Sci.*, 35:85–101, 1971.

[118] A. Igarashi, P. V. E. McClintock, and N. G. Stocks. Velocity spectrum for non-Markovian Brownian motion in a periodic potential. *J. Statist. Phys.*, 66(3-4):1059–1070, 1992. ISSN: 0022-4715. DOI: 10.1007/BF01055716. URL: https://doi.org/10.1007/BF01055716.

[119] A. Igarashi and T. Munakata. Non-Markovian Brownian motion in a periodic potential. *J. Phys. Soc. Jpn.*, 57(7):2439–2447, 1988.

[120] D. Jacqmin. Calculation of two-phase Navier–Stokes flows using phase-field modeling. *Journal of Computational Physics*, 155(1):96–127, 1999.

[121] D. Jacqmin. Contact-line dynamics of a diffuse fluid interface. *J. Fluid Mech.*, 402:57–88, 2000.

[122] D. Jasnow and J. Viñals. Coarse-grained description of thermo-capillary flow. *Phys. Fluids*, 8:660, 1995.

[123] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: open source scientific tools for Python, 2001–2018. URL: http://www.scipy.org/.

[124] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998. ISSN: 0036-1410. DOI: 10.1137/S0036141096303359. URL: https://doi.org/10.1137/S0036141096303359.

[125] V. Kaarnioja. *Smolyak quadrature*. Master's thesis, Department of Mathematics and Statistics, University of Helsinky, 2013.

[126] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991, pages xxiv+470. ISBN: 0-387-97655-8. DOI: 10.1007/978-1-4612-0949-2. URL: http://dx.doi.org/10.1007/978-1-4612-0949-2.

[127] D. Kay, V. Styles, and R. Welford. Finite element approximation of a Cahn-Hilliard-Navier-Stokes system. *Interfaces Free Bound.*, 10(1):15–43, 2008. ISSN: 1463-9963. DOI: 10.4171/IFB/178. URL: https://doi.org/10.4171/IFB/178.

[128] I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. Kevrekidis, O. Runborg, and C. Theodoropoulos. Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level analysis. *Commun. Math. Sci.*, 1(4):715–762, 2003. ISSN: 1539-6746. URL: http://projecteuclid.org/euclid.cms/1119655353.

[129] V. V. Khatavkar, P. D. Anderson, and H. E. H. Meijer. Capillary spreading of a droplet in the partially wetting regime using a diffuse-interface model. *J. Fluid Mech.*, 572:367, 2007.

[130] D. Kim and W. Lu. Three-dimensional model of electrostatically induced pattern formation in thin polymer films. *Phys. Rev. B*, 73(3):035206, 2006.

[131] J. Kim. Phase-field models for multi-component fluid flows. *Communications in Computational Physics*, 12(03):613–661, 2012.

[132] J. Kim and L. Hyun Geun. Accurate contact angle boundary conditions for the Cahn-Hilliard equations. *Comput. Fluids*, 44:178–186, 2012.

[133] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992, pages xxxvi+632. ISBN: 3-540-54062-8. DOI: 10.1007/978-3-662-12616-5. URL: http://dx.doi.org/10.1007/978-3-662-12616-5.

[134] B. Krauskopf, H. M. Osinga, and J. Galán-Vioque, editors. *Numerical continuation methods for dynamical systems*. Understanding Complex Systems. Springer, Dordrecht, 2007, pages xx+399. ISBN: 978-1-4020-6355-8. DOI: 10.1007/978-1-4020-6356-5. URL: https://doi.org/10.1007/978-1-4020-6356-5.

[135] R. Kupferman. Fractional kinetics in Kac-Zwanzig heat bath models. *J. Statist. Phys.*, 114(1-2):291–326, 2004. ISSN: 0022-4715. DOI: 10.1023/B:JOSS.0000003113.22621.f0. URL: https://doi.org/10.1023/B:JOSS.0000003113.22621.f0.

[136] P. S. Laplace. *Traité de mécanique céleste*, volume 1. Crapelet, 1799.

[137] J. C. Latorre, G. A. Pavliotis, and P. R. Kramer. Corrections to Einstein's relation for Brownian motion in a tilted periodic potential. *J. Stat. Phys.*, 150(4):776–803, 2013. ISSN: 0022-4715. DOI: 10.1007/s10955-013-0692-1. URL: http://dx.doi.org/10.1007/s10955-013-0692-1.

[138] D. Lee, J.-Y. Huh, D. Jeong, J. Shin, A. Yun, and J. Kim. Physical, mathematical, and numerical derivations of the Cahn–Hilliard equation. *Comput. Mater. Sci.*, 81:216–225, 2014.

[139] G. Leibon, D. N. Rockmore, W. Park, R. Taintor, and G. S. Chirikjian. A fast Hermite transform. *Theoret. Comput. Sci.*, 409(2):211–228, 2008. ISSN: 0304-3975. DOI: 10.1016/j.tcs.2008.09.010. URL: https://doi.org/10.1016/j.tcs.2008.09.010.

[140] B. Leimkuhler, C. Matthews, and G. Stoltz. The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA J. Numer. Anal.*, 36(1):13–79, 2016. ISSN: 0272-4979. DOI: 10.1093/imanum/dru056. URL: https://doi.org/10.1093/imanum/dru056.

[141] T. Lelièvre, M. Rousset, and G. Stoltz. *Free energy computations*. Imperial College Press, London, 2010, pages xiv+458. ISBN: 978-1-84816-247-1; 1-84816-247-2. DOI: 10.1142/9781848162488. URL: http://dx.doi.org/10.1142/9781848162488. A mathematical perspective.

[142] T. Lelièvre and G. Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numer.*, 25:681–880, 2016. ISSN: 0962-4929. DOI: 10.1017/S0962492916000039. URL: http://dx.doi.org/10.1017/S0962492916000039.

[143] T. Li, A. Abdulle, and W. E. Effectiveness of implicit methods for stiff stochastic differential equations. *Commun. Comput. Phys.*, 3(2):295–307, 2008. ISSN: 1815-2406.

[144] J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod; Gauthier-Villars, Paris, 1969, pages xx+554.

[145] G. Lippmann. *Relations entre les phénomènes électriques et capillaires*. PhD thesis, Gauthier-Villars, 1875.

[146] A. Logg, K.-A. Mardal, and G. Wells. *Automated solution of differential equations by the finite element method: The FEniCS book*, volume 84. Springer Science & Business Media, 2012.

[147] L. Lorenzi and M. Bertoldi. *Analytical methods for Markov semigroups*, volume 283 of *Pure and Applied Mathematics (Boca Raton)*. Chapman & Hall/CRC, Boca Raton, FL, 2007, pages xxxii+526. ISBN: 978-1-58488-659-4; 1-58488-659-5.

[148] J. Lu, Y. Lu, and J. Nolen. Scaling limit of the Stein variational gradient descent part I: the mean field regime. *ArXiv e-prints*, May 2018. arXiv: 1805.04035 [math.AP].

[149] E. Luçon and W. Stannat. Transition from Gaussian to non-Gaussian fluctuations for mean-field diffusions in spatial interaction. *Ann. Appl. Probab.*, 26(6):3840–3909, 2016. ISSN: 1050-5164. DOI: 10.1214/16-AAP1194. URL: https://doi.org/10.1214/16-AAP1194.

[150] A. Magnus. Analyse numérique 2. *Institut de Mathématique Pure et Appliquée, Université Catholique de Louvain*, 2009.

[151] A. J. Majda, I. Timofeyev, and E. Vanden-Eijnden. A mathematical framework for stochastic climate models. *Comm. Pure Appl. Math.*, 54(8):891–974, 2001. ISSN: 0010-3640. DOI: 10.1002/cpa.1014. URL: https://doi.org/10.1002/cpa.1014.

[152] P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis. *Mat. Contemp.*, 19:1–29, 2000. ISSN: 0103-9059. VI Workshop on Partial Differential Equations, Part II (Rio de Janeiro, 1999).

[153] N. Martzel and C. Aslangul. Mean-field treatment of the many-body Fokker-Planck equation. *J. Phys. A*, 34(50):11225–11240, 2001. ISSN: 0305-4470. DOI: 10.1088/0305-4470/34/50/305. URL: http://dx.doi.org/10.1088/0305-4470/34/50/305.

[154] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.*, 101(2):185–232, 2002. ISSN: 0304-4149. DOI: 10.1016/S0304-4149(02)00150-3. URL: https://doi.org/10.1016/S0304-4149(02)00150-3.

[155] A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, Š. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz. Sympy: symbolic computing in Python. *PeerJ Computer Science*, 3:e103, Jan. 2017. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.103. URL: https://doi.org/10.7717/peerj-cs.103.

[156] G. N. Milstein and M. V. Tretyakov. *Stochastic numerics for mathematical physics*. Scientific Computation. Springer-Verlag, Berlin, 2004, pages xx+594. ISBN: 3-540-21110-1. DOI: 10.1007/978-3-662-10063-9. URL: http://dx.doi.org/10.1007/978-3-662-10063-9.

[157] H. K. Moffatt. Viscous and resistive eddies near a sharp corner. *J. Fluid Mech.*, 18(1):1–18, 1964.

[158] P. Monmarché. Long-time behaviour and propagation of chaos for mean field kinetic particles. *Stochastic Processes and their Applications*, 127(6):1721–1737, 2017. ISSN: 0304-4149. DOI: https://doi.org/10.1016/j.spa.2016.10.003. URL: http://www.sciencedirect.com/science/article/pii/S0304414916301788.

[159] H. Mori. A continued-fraction representation of the time-correlation functions. *Progr. Theoret. Phys.*, 34:399–416, 1965. ISSN: 0033-068X. DOI: 10.1143/PTP.34.399. URL: https://doi.org/10.1143/PTP.34.399.

[160] M. Mortensen. Shenfun - automating the spectral galerkin method. In B. H. Skallerud and H. I. Andersson, editors, *MekIT'17 - Ninth national conference on Computational Mechanics*, pages 273–298. International Center for Numerical Methods in Engineering (CIMNE), 2017.

[161] S. Motsch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Rev.*, 56(4):577–621, 2014. ISSN: 0036-1445. DOI: 10.1137/120901866. URL: https://doi.org/10.1137/120901866.

[162] M. M. Nahar, G. S. Bindiganavane, J. Nikapitiya, and H. Moon. Numerical modeling of 3D electrowetting droplet actuation and cooling of a hotspot. *Proc. COMSOL conf.*, 2015.

[163] A. Nold. *From the nano- to the macroscale – Bridging scales for the moving contact line problem*. PhD thesis, Imperial College London, 2016.

[164] A. Nold, B. D. Goddard, P. Yatsyshin, N. Savva, and S. Kalliadasis. Pseudospectral methods for density functional theory in bounded and unbounded domains. *J. Comput. Phys.*, 334:639–664, 2017. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2016.12.023. URL: https://doi.org/10.1016/j.jcp.2016.12.023.

[165] K. Oelschläger. A martingale approach to the law of large numbers for weakly interacting stochastic processes. *Ann. Probab.*, 12(2):458–479, 1984. ISSN: 0091-1798. URL: HTTP://LINKS.JSTOR.ORG/SICI?SICI=0091-1798(198405)12:2%3C458:AMATTL%3E2.0.CO;2-H&origin=MSN.

[166] S. Osher and R. P. Fedkiw. Level set methods: an overview and some recent results. *J. Comput. Phys.*, 169(2):463–502, 2001.

[167] M. Ottobre. *Asymptotic analysis for Markovian models in non-equilibrium statistical mechanics*. PhD thesis, Department of Mathematics, Imperial College London, 2012.

[168] M. Ottobre and G. A. Pavliotis. Asymptotic analysis for the generalized Langevin equation. *Nonlinearity*, 24(5):1629–1653, 2011. ISSN: 0951-7715. DOI: 10.1088/0951-7715/24/5/013. URL: https://doi.org/10.1088/0951-7715/24/5/013.

[169] E. Pardoux and A. Y. Veretennikov. On Poisson equation and diffusion approximation. II. *Ann. Probab.*, 31(3):1166–1192, 2003. ISSN: 0091-1798. DOI: 10.1214/aop/1055425774. URL: http://dx.doi.org/10.1214/aop/1055425774.

[170] E. Pardoux and A. Y. Veretennikov. On the Poisson equation and diffusion approximation. I. *Ann. Probab.*, 29(3):1061–1085, 2001. ISSN: 0091-1798. DOI: 10.1214/aop/1015345596. URL: http://dx.doi.org/10.1214/aop/1015345596.

[171] E. Pardoux and A. Y. Veretennikov. On the Poisson equation and diffusion approximation. III. *Ann. Probab.*, 33(3):1111–1133, 2005. ISSN: 0091-1798. DOI: 10.1214/009117905000000062. URL: http://dx.doi.org/10.1214/009117905000000062.

[172] G. A. Pavliotis. A multiscale approach to Brownian motors. *Physics Letters A*, 344(5):331–345, 2005.

[173] G. A. Pavliotis. Asymptotic analysis of the Green-Kubo formula. *IMA J. Appl. Math.*, 75(6):951–967, 2010. ISSN: 0272-4960. DOI: 10.1093/imamat/hxq039. URL: http://dx.doi.org/10.1093/imamat/hxq039.

[174] G. A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014, pages xiv+339. ISBN: 978-1-4939-1322-0; 978-1-4939-1323-7. DOI: 10.1007/978-1-4939-1323-7. URL: http://dx.doi.org/10.1007/978-1-4939-1323-7. Diffusion processes, the Fokker-Planck and Langevin equations.

[175] G. A. Pavliotis and A. M. Stuart. *Multiscale methods*, volume 53 of *Texts in Applied Mathematics*. Springer, New York, 2008, pages xviii+307. ISBN: 978-0-387-73828-4. Averaging and homogenization.

[176] G. A. Pavliotis, A. M. Stuart, and K. C. Zygalakis. Homogenization for inertial particles in a random flow. *Commun. Math. Sci.*, 5(3):507–531, 2007. ISSN: 1539-6746. URL: http://projecteuclid.org/euclid.cms/1188405666.

[177] G. A. Pavliotis and U. Vaes. The generalized Langevin equation in a periodic potential, 2019, in preparation.

[178] G. A. Pavliotis and A. Vogiannou. Diffusive transport in periodic potentials: underdamped dynamics. *Fluct. Noise Lett.*, 8(2):L155–L173, 2008. ISSN: 0219-4775. DOI: 10.1142/S0219477508004453. URL: https://doi.org/10.1142/S0219477508004453.

[179] R. L. Pego. Front migration in the nonlinear Cahn-Hilliard equation. *Proc. Roy. Soc. London Ser. A*, 422(1863):261–278, 1989. ISSN: 0962-8444.

[180] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(1):183–204, 2017. ISSN: 0218-2025. DOI: 10.1142/S0218202517400061. URL: https://doi.org/10.1142/S0218202517400061.

[181] J. Plateau. *Experimental and theoretical statics of liquids subjected to molecular forces only*. Paris: Gauthiers-Villars, 1873.

[182] H. Poincaré. *Capillarité*. Georges Carré, Paris, 1895.

[183] M. Queralt-Martín, M. Pradas, R. Rodríguez-Trujillo, M. Arundell, E. Corvera-Poiré, and A. Hernández-Machado. Pinning and avalanches in hydrophobic microchannels. *Phys. Rev. Lett.*, 106:194501, 2011.

[184] M. Reed and B. Simon. *Methods of modern mathematical physics. IV. Analysis of operators*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1978, pages xv+396. ISBN: 0-12-585004-2.

[185] P. Reimann. Brownian motors: noisy transport far from equilibrium. *Phys. Rep.*, 361(2-4):57–265, 2002. ISSN: 0370-1573. DOI: 10.1016/S0370-1573(01)00081-3. URL: https://doi.org/10.1016/S0370-1573(01)00081-3.

[186] L. Rey-Bellet. Open classical systems. In *Open quantum systems. II.* Volume 1881, Lecture Notes in Math. Pages 41–78. Springer, Berlin, 2006. DOI: 10.1007/3-540-33966-3_2. URL: https://doi.org/10.1007/3-540-33966-3_2.

[187] H. Risken. *The Fokker-Planck equation*, volume 18 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, second edition, 1989, pages xiv+472. ISBN: 3-540-50498-2. DOI: 10.1007/978-3-642-61544-3. URL: https://doi.org/10.1007/978-3-642-61544-3. Methods of solution and applications.

[188] G. M. Rotskoff and E. Vanden-Eijnden. Neural Networks as Interacting Particle Systems: Asymptotic Convexity of the Loss Landscape and Universal Scaling of the Approximation Error. *ArXiv e-prints*, May 2018. arXiv: 1805.00915 [stat.ML].

[189] J. Roussel and G. Stoltz. A perturbative approach to control variates in molecular dynamics. *arXiv e-prints*:arXiv:1712.08022, arXiv:1712.08022, Dec. 2017. arXiv: 1712.08022 [math.NA].

[190] J. Roussel and G. Stoltz. Spectral methods for Langevin dynamics and associated error estimates. *ESAIM: Math. Model. Numer. Anal.*, 52(3):1051–1083, 2018.

[191] J. Roussel. *Analyse théorique et numérique de dynamiques non-réversibles en physique statistique computationnelle*. PhD thesis, Université Paris-Est, École doctorale MSTIC, 2018.

[192] M. Roydhouse, M. Pradas, N. Al-Rifai, B. Azizi, E. Cao, S. Kalliadasis, and A. Gavriilidis. Operating ranges of gas–liquid capillary microseparators: experiments and theory. *Chem. Eng. Sci.*, 114:30–39, 2014.

[193] Z. Schuss. *Theory and applications of stochastic processes*, volume 170 of *Applied Mathematical Sciences*. Springer, New York, 2010, pages xviii+468. ISBN: 978-1-4419-1604-4. DOI: 10.1007/978-1-4419-1605-1. URL: https://doi.org/10.1007/978-1-4419-1605-1. An analytical approach.

[194] P. Seppecher. Moving contact lines in the Cahn-Hilliard theory. *Int. J. Eng. Sci.*, 34:977–992, 1996.

[195] J. Shen, T. Tang, and L.-L. Wang. *Spectral methods*, volume 41 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2011, pages xvi+470. ISBN: 978-3-540-71040-0. DOI: 10.1007/978-3-540-71041-7. URL: https://doi.org/10.1007/978-3-540-71041-7. Algorithms, analysis and applications.

[196] M. Shiino. Dynamical behavior of stochastic systems of infinitely many coupled nonlinear oscillators exhibiting phase transitions of mean-field type: h theorem on asymptotic approach to equilibrium and critical slowing down of order-parameter fluctuations. *Phys. Rev. A*, 36:2393–2412, 5, Sept. 1987. DOI: 10.1103/PhysRevA.36.2393. URL: https://link.aps.org/doi/10.1103/PhysRevA.36.2393.

[197] P. Shuhua, P. Bat-El, H. Hao, M. Helmuth, and Z. Xuehua. Morphological transformation of surface femtodroplets upon dissolution. *J. Phys. Chem. Lett.*, 8:584–590, 2017.

[198] H. Si. Tetgen, a Delaunay-based quality tetrahedral mesh generator. *ACM Trans. Math. Softw.*, 41(2):11, 2015.

[199] D. N. Sibley, A. Nold, and S. Kalliadasis. The asymptotics of the moving contact line: cracking an old nut. *J. Fluid Mech.*, 764:445–462, 2015.

[200] D. N. Sibley, A. Nold, and S. Kalliadasis. Unifying binary fluid diffuse-interface models in the sharp-interface limit. *J. Fluid Mech.*, 736:5–43, 2013.

[201] D. N. Sibley, A. Nold, N. Savva, and S. Kalliadasis. A comparison of slip, disjoining pressure, and interface formation models for contact line motion through asymptotic analysis of thin two-dimensional droplet spreading. *J. Eng. Math.*, 94:19–41, 2015.

[202] D. N. Sibley, A. Nold, N. Savva, and S. Kalliadasis. On the moving contact line singularity: Asymptotics of a diffuse-interface model. *Eur. Phys. J. E*, 36:26, 2013.

[203] D. N. Sibley, A. Nold, N. Savva, and S. Kalliadasis. The contact line behavior of solid-liquid-gas diffuse-interface models. *Phys. Fluids*, 25:092111, 2013.

[204] J. Sirignano and K. Spiliopoulos. Mean Field Analysis of Neural Networks. *ArXiv e-prints*, May 2018. arXiv: 1805.01053 [math.PR].

[205] R. H. Stogner, G. F. Carey, and B. T. Murray. Approximation of Cahn-Hilliard diffuse interface models using parallel adaptive mesh refinement and coarsening with C^1 elements. *Internat. J. Numer. Methods Engrg.*, 76(5):636–661, 2008. ISSN: 0029-5981. DOI: 10.1002/nme.2337. URL: https://doi.org/10.1002/nme.2337.

[206] M. Sussman, P. Smereka, and S. Osher. A level set approach for computing solutions to incompressible two-phase flow. *J. Comput. Phys.*, 114(1):146–159, 1994.

[207] A.-S. Sznitman. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*. Volume 1464, Lecture Notes in Math. Pages 165–251. Springer, Berlin, 1991. DOI: 10.1007/BFb0085169. URL: https://doi.org/10.1007/BFb0085169.

[208] D. Talay. Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme. *Markov Process. Related Fields*, 8(2):163–198, 2002. ISSN: 1024-2953. Inhomogeneous random systems (Cergy-Pontoise, 2001).

[209] T. Tang. The Hermite spectral method for Gaussian-type functions. *SIAM J. Sci. Comput.*, 14(3):594–606, 1993. ISSN: 1064-8275. DOI: 10.1137/0914038. URL: https://doi.org/10.1137/0914038.

[210] J. Teuwen. A cornucopia of Hermite polynomials, 2012.

[211] G. Tryggvason, B. Bunner, A. Esmaeeli, D. Juric, N. Al-Rawahi, W. Tauber, J. Han, S. Nas, and Y.-J. Jan. A front-tracking method for the computations of multiphase flow. *J. Comput. Phys.*, 169(2):708–759, 2001.

[212] E. Tucker. *Finite Element Approximations of a Phase Field Model, based on the Cahn-Hilliard Equation in the Presence of an Electric Field and Kinetics*. PhD thesis, Imperial College London, 2013.

[213] J. Tugaut. Phase transitions of McKean-Vlasov processes in double-wells landscape. *Stochastics*, 86(2):257–284, 2014. ISSN: 1744-2508. DOI: 10.1080/17442508.2013.775287. URL: https://doi.org/10.1080/17442508.2013.775287.

[214] E. B. V. Dussan and S. H. Davis. On the motion of a fluid-fluid interface along a solid surface. *J. Fluid Mech.*, 65:71–95, 1974.

[215] U. Vaes. *Numerical solution of multiscale stochastic equations*. Master's thesis, Department of Mathematics, Imperial College London, 2014.

[216] U. Vaes. Python library for the Hermite spectral method. https://github.com/urbainvaes/hermipy, 2018.

[217] U. Vaes and B. Aymard. FreeFem++ solver for the Cahn-Hilliard equation. https://github.com/urbainvaes/cahn-hilliard, 2018.

[218] E. Vanden-Eijnden. Numerical techniques for multi-scale dynamical systems with stochastic effects. *Commun. Math. Sci.*, 1(2):385–391, 2003.

[219] R. Vellingiri, N. Savva, and S. Kalliadasis. Droplet spreading on chemically heterogeneous substrates. *Phys. Rev. E*, 84:036305, 2011.

[220] C. Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950):iv+141, 2009. ISSN: 0065-9266. DOI: 10.1090/S0065-9266-09-00567-5. URL: https://doi.org/10.1090/S0065-9266-09-00567-5.

[221] Z. Weng, S. Zhai, and X. Feng. A Fourier spectral method for fractional-in-space Cahn-Hilliard equation. *Appl. Math. Model.*, 42:462–477, 2017. ISSN: 0307-904X. DOI: 10.1016/j.apm.2016.10.035. URL: https://doi.org/10.1016/j.apm.2016.10.035.

[222] C. Wylock, M. Pradas, B. Haut, P. Colinet, and S. Kalliadasis. Disorder-induced hysteresis and nonlocality of contact line motion in chemically heterogeneous microchannels. *Phys. Fluids*, 24:032108, 2012.

[223] X. Xu and T. Qian. Contact line motion in confined liquid-gas systems: Slip versus phase transition. *J. Chem. Phys.*, 133:204704, 2010.

[224] T. Young. An essay on the cohesion of fluids. *Philos. Trans. Royal Soc.*, 95:65–87, 1805.

[225] H. Yserentant. *Regularity and approximability of electronic wave functions*, volume 2000 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2010, pages viii+182. ISBN: 978-3-642-12247-7. DOI: 10.1007/978-3-642-12248-4. URL: http://dx.doi.org/10.1007/978-3-642-12248-4.

[226] P. Yue, C. Zhou, and J. J. Feng. Sharp-interface limit of the Cahn-Hilliard model for moving contact lines. *J. Fluid Mech.*, 645:279–294, 2010. ISSN: 0022-1120. DOI: 10.1017/S0022112009992679. URL: https://doi.org/10.1017/S0022112009992679.

[227] P. Yue, C. Zhou, J. J. Feng, C. F. Ollivier-Gooch, and H. H. Hu. Phase-field simulations of interfacial dynamics in viscoelastic fluids using finite elements with adaptive meshing. *J. Comput. Phys.*, 219(1):47–67, 2006. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2006.03.016. URL: https://doi.org/10.1016/j.jcp.2006.03.016.

Appendix A

Background material

Throughout this thesis, we use some standard results on stochastic differential equations, weighted Sobolev spaces, and Hermite polynomials. In this appendix we will state and briefly discuss the relevant results.

A.1 Stochastic differential equations

In this appendix, we recall definitions and results about stochastic differential equations (SDEs). We consider the following autonomous equation,

$$\mathrm{d}Y_t = \mathbf{b}(Y_t)\,\mathrm{d}t + \boldsymbol{\sigma}(Y_t)\,\mathrm{d}W_t, \quad Y_0 = y_0, \tag{A.1}$$

with $\mathbf{b} : \mathbf{R}^n \to \mathbf{R}^n$ a smooth vector-valued function, $\boldsymbol{\sigma} : \mathbf{R}^n \to \mathbf{R}^{n\times d}$ a smooth matrix-valued function, W_t a standard d-dimensional Brownian motion, see [174, Definition 1.8], and y_0 a random variable on $\mathbf{R}^n$ independent of W_t. In this thesis, we interpret Eq. (A.1) in the Itô sense, and adopt the standard definition of a strong solution, see e.g. [174, Section 3.3].

Definition A.1 (Strong solution)**.** We call a stochastic process Y_t on $\mathbf{R}^n$ a solution to the SDE (A.1) on $t \in [0, T]$ if

- Y_t is almost surely continuous and adapted to the filtration $\mathcal{F}_t$ generated by W_t.
- $\mathbf{b}(Y_\cdot) \in L^1\left((0,T);\mathbf{R}^n\right)$ and $\boldsymbol{\sigma}(Y_\cdot) \in L^2\left((0,T;\mathbf{R}^n)\right)$ almost surely.
- For all $t \in [0, T]$, the equation

$$Y_t = y + \int_0^t \mathbf{b}(Y_t)\,dt + \int_0^t \boldsymbol{\sigma}(Y_t)\,dW_t,$$

  where the second integral is an Itô integral, defined e.g. in [174, Section 3.2], holds with probability 1.

The solution is said to be unique if the paths of any two solutions Y_t and Z_t coincide almost surely:

$$\mathbf{P}(Y_t = Z_t,\ \forall t \in [0,T]) = 1.$$

We assume hereafter that there exists a unique solution to Eq. (A.1) above. To study the properties of the solution Y_t, a powerful approach is to examine the *infinitesimal generator* of the Markov process Y_t solving (A.1), which enters in the *backward Kolmogorov equation*, as well as the formal $L^2(\mathbf{R}^n)$ adjoint of the infinitesimal generator, which enters in the *Fokker–Planck equation*. The infinitesimal generator $\mathcal{L}$ associated to Eq. (A.1), often referred to in this text as just the *generator*, is the second-order partial differential operator defined by

$$\mathcal{L}v = \mathbf{b}(y) \cdot \nabla v + \frac{1}{2}\,\mathbf{A}(y) : \nabla\nabla v,$$

where $\mathbf{A}(y) = \boldsymbol{\sigma}(y)\boldsymbol{\sigma}(y)^T$. The use of the generator enables an elegant formulation of Itô's formula, which provides information on the evolution in time of scalar functions of Y_t.

Lemma A.2 (Itô's formula)**.** *Consider $V(\,\cdot\,) \in C^2(\mathbf{R}^n;\mathbf{R})$, and assume that $\mathbf{b}$ and $\boldsymbol{\sigma}$ are Lipschitz continuous and that the initial condition is such that $\mathbf{E}\,|y_0|^2 < \infty$. Then $\phi_t := \phi(Y_t)$ satisfies the following SDE:*

$$\mathrm{d}\phi_t = \mathcal{L}\phi(Y_t) + \nabla\phi(Y_t) \cdot \boldsymbol{\sigma}(Y_t)\,\mathrm{d}W_t. \tag{A.2}$$

Denoting by $\mathbf{E}_y$ the expectation with respect to the Brownian paths when the initial condition is $y_0 = y$, and assuming that ϕ is smooth and compactly supported, we observe by taking expectations of both sides in Eq. (A.2) that

$$\mathbf{E}_y\phi_t - \phi(y) = \int_0^t \mathbf{E}_y\left[\mathcal{L}\phi(Y_t)\right]\,\mathrm{d}t, \tag{A.3}$$

by the martingale property of the stochastic integral. This implies, at least formally, that $\mathcal{L}$ is indeed the generator of the Markov process Y_t, in the sense that

$$\mathcal{L}\phi = \lim_{t\to 0}\frac{\mathbf{E}_y\phi(Y_t) - \phi(y)}{t},$$

and it suggests, by differentiation of (A.3) with respect to t, that $v(y,t) := \mathbf{E}_y(\phi_t)$ satisfies

$$\frac{\partial v}{\partial t} = \mathcal{L}v,$$

with the initial condition $v(\cdot,0) = \phi(\cdot)$. This result connects the generator to the associated SDE, Eq. (A.1). In fact we have the following result, presented in [175, Theorem 6.6].

Proposition A.3 (Backward Kolmogorov equation)**.** *Assume that $\phi(\,\cdot\,)$ is chosen sufficiently smooth so that there exists a unique classical solution $v(y,t) \in C^{2,1}(\mathbf{R}^n \times (0,T);\mathbf{R}) \cap C(\mathbf{R}^n \times [0,T];\mathbf{R})$ to the* backward Kolmogorov equation

$$\begin{aligned} \frac{\partial v}{\partial t} &= \mathcal{L}v \qquad && (y,t) \in (\mathbf{R}^n, (0,T)), \\ v &= \phi && (y,t) \in \mathbf{R}^n \times \{0\}. \end{aligned} \tag{A.4}$$

Then v is given by $\mathbf{E}_y(\phi_t) = \mathbf{E}(\phi(Y_t)|Y_0 = y)$.

$$v(y,t) = \mathbf{E}(\phi(Y_t)|Y_0 = y),$$

where the expectation is with respect to the Brownian paths.

The formal $L^2(\mathbf{R}^n)$-adjoint of the generator, such that $\int_{\mathbf{R}^n} \mathcal{L}u\, v \,\mathrm{d}y = \int_{\mathbf{R}^n} u\, \mathcal{L}^* v \,\mathrm{d}y$ for all smooth, compactly supported u, v, reads

$$\mathcal{L}^* = -\nabla \cdot \left(\mathbf{b} \cdot + \frac{1}{2}\, \mathbf{A}\, \nabla \cdot \right).$$

It is connected to the law of Y_t by the following result, [112, Theorem 6.8].

Proposition A.4 (Fokker–Planck equation)**.** *Consider Eq. (A.1), suppose that the initial condition y_0 has a probability density $\rho_0(y)$ with respect to the Lebesgue measure, and assume that the law of Y_t also has a density $\rho(y,t) \in C^{2,1}(\mathbf{R}^n \times (0,T); \mathbf{R}) \cap C(\mathbf{R}^n \times [0,T]; \mathbf{R})$. Then $\rho(y,t)$ satisfies the Fokker–Planck equation, also called the forward Kolmogorov equation:*

$$\begin{aligned} \frac{\partial \rho}{\partial t} &= \mathcal{L}^* v \quad \text{for } (y,t) \in (\mathbf{R}^n, (0,T)), \\ \rho &= \rho_0 \quad \text{for } (y,t) \in \mathbf{R}^n \times \{0\}. \end{aligned} \tag{A.5}$$

In view of this result, $\mathbf{J} = \mathbf{b}\,\rho + \frac{1}{2}\,\mathbf{A}\,\nabla\rho$ is a *probability flux*. If $\rho(y)$ denotes a steady-state solution of the Fokker–Planck equation, then the corresponding probability flux is divergence-free. In two dimensions, it is often interesting to visualize the field lines of this flux, which can be calculated via the so-called *stream function*: if the components of the flux, denoted by j_1 and j_2, are twice continuously differentiable, then it is possible to show (for example by Helmoltz decomposition) that there exists a *stream function* $\psi(y_1, y_2)$ such that

$$j_1 = \frac{\partial \psi}{\partial y_2}, \qquad j_2 = -\frac{\partial \psi}{\partial y_1}.$$

From these equations, it is clear that the contour lines of ψ correspond to the field lines of the probability flux, and we notice that ψ satisfies

$$-\Delta\psi = \frac{\partial j_2}{\partial y_1} - \frac{\partial j_1}{\partial y_2}, \tag{A.6}$$

which can be solved numerically using standard methods. Equation (A.6) needs to be supplemented by appropriate boundary conditions. In the context of the Fokker–Planck equations in an unbounded domain, the right-hand side of Eq. (A.6) often decreases to zero exponentially fast, and the adequate condition is therefore that $\lim_{|y|\to\infty} \psi(y) = 0$ (or any other constant, because the stream function is defined only up to an additive constant).

In some cases, the solution to the Fokker–Planck equation (A.5) converges, as $t \to \infty$, to a stationary probability density $\rho^\infty(y)$. The corresponding probability measure is called an *invariant measure* of the system, and it is the stochastic counterpart of an equilibrium point for deterministic dynamics.

An important concept for the study of the long-time behavior of solutions to SDEs is the concept of *ergodicity*, which is related to the existence and uniqueness of an invariant measure and to the convergence to this measure as $t \to \infty$. A possible definition of ergodicity, often used in the physics and engineering literature, is that the time average of an observable along trajectories converges, as the final time of the averaging window tends to infinity, to the expected value of the observable with respect to the unique invariant measure of the system. One setting in which ergodicity can be formulated and proved relatively easily is when the state space of the variable Y_t is the torus $\mathbf{T}^n$; in this case we have the following theorem, [175, Theorem 6.16].

Theorem A.5 (Ergodicity result)**.** *Assume that $\mathcal{L}, \mathcal{L}^*$ are equipped with periodic boundary conditions, and that the following uniform ellipticity condition is satisfied:*

$$\exists \gamma : \quad \langle \mathbf{A}(y)\,\xi, \xi \rangle > \gamma\,|\xi|^2 \quad \forall y \in \mathbf{T}^n, \xi \in \mathbf{R}^n. \tag{A.7}$$

Then the following statements hold true:

- $\mathcal{N}(\mathcal{L}) = \mathrm{span}\{1\}$*;*
- $\mathcal{N}(\mathcal{L}^*) = \mathrm{span}\{\rho^\infty\}$*, with ρ^∞ such that $\inf_{z \in \mathbf{T}^n} \rho^\infty(z) > 0$;*
- $\forall \phi \in C(\mathbf{T}^n), \exists C, \lambda > 0$ *such that the solution $v(t)$ of (A.4) with ϕ as initial data satisfies*

$$\| v(t) - \left(\int_{\mathbf{T}^n} \phi(y) \rho^\infty(y)\,\mathrm{d}y \right) \|_\infty \leq C\,\mathrm{e}^{-\lambda t} \quad \forall t > 0;$$

- $\forall y_0 \in \mathbf{T}^n, \exists C, \lambda > 0$ *such that the solution $\rho(t)$ of (A.5) with a Dirac mass in y_0 as initial data satisfies*

$$\| \rho(t) - \rho^\infty \|_1 \leq C\,\mathrm{e}^{-\lambda t} \quad \forall t > 0;$$

- *For all $\phi \in C(\mathbf{T}^n)$*

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \phi(Y_t)\,\mathrm{d}t = \int_{\mathbf{T}^n} \phi(y) \rho^\infty(y)\,\mathrm{d}y \quad a.s.$$

A.2 Weighted Sobolev Spaces

In this section, we recall a few results about weighted Sobolev spaces, required for the analysis presented in Chapters 2 and 3. For more details on this topic, see [83, 225, 25, 147]. Throughout the appendix, V denotes a smooth confining potential whose derivatives are all bounded above by a polynomial and such that $\rho := \mathrm{e}^{-V}$ is normalized.

Definition A.6. The weighted space $L^2(\mathbf{R}^n, \rho)$ is defined as

$$L^2(\mathbf{R}^n, \rho) = \left\{ u \text{ measurable } : \int_{\mathbf{R}^n} u^2\,\rho\,\mathrm{d}y < \infty \right\}.$$

It is a Hilbert space for the inner product given by:

$$\langle u, v \rangle_\rho = \int_{\mathbf{R}^n} u\,v\,\rho\,\mathrm{d}y.$$

Definition A.7. The weighted Sobolev space $H^s(\mathbf{R}^n, \rho)$, with $s \in \mathbf{N}$, is defined as

$$H^s(\mathbf{R}^n, \rho) = \left\{ u \in L^2(\mathbf{R}^n, \rho) : \partial^\alpha u \in L^2(\mathbf{R}^n, \rho) \ \forall\, |\alpha| \leq s \right\}.$$

It is a Hilbert space for the inner product

$$\langle u, v \rangle_{s,\rho} = \sum_{|\alpha| \leq s} \langle \partial^\alpha u, \partial^\alpha v \rangle_\rho$$

Note that $H^0(\mathbf{R}^n, \rho) = L^2(\mathbf{R}^n, \rho)$, and the notation for the associated inner product and norm will be shortened to $\langle \cdot, \cdot \rangle_\rho$ and $\| \cdot \|_\rho$, respectively. It can be shown that $C_c^\infty(\mathbf{R}^n)$ is dense in the weighted Sobolev space $H^s(\mathbf{R}^n, \rho)$.

Proposition A.8. *For any $s \in \mathbf{N}$, the smooth, compactly supported functions $C_c^\infty(\mathbf{R}^n)$ form a dense subspace of the weighted Sobolev space $H^s(\mathbf{R}^n, \rho)$.*

Proof. Let $f \in H^s(\mathbf{R}^n, \rho)$ and consider a smooth cutoff function $\eta(\cdot) \in C_c^\infty(\mathbf{R}^n)$ equal to 1 in $B(0,1)$, the open ball of radius 1 centered at the origin. By dominated convergence and the formula for the weak derivative of the product of a distribution with a smooth function, the functions $f_n(x) := \eta(x/n)\, f(x)$, $n = 1, 2, \ldots$, are such that $f_n \to f$ in $H^s(\mathbf{R}^n, \rho)$ when $n \to \infty$. Since $\rho(\cdot)$ is smooth and always positive, and since the function $f_n(\cdot)$ is compactly supported for any $n \in \mathbf{N}$, the inclusion $f_n \in H^s(\mathbf{R}^n, \rho)$ implies that also $f_n \in H^s(\mathbf{R}^n)$, where we note the absence of a weight. Therefore, by density of $C_c^\infty(\mathbf{R}^n)$ in $H^s(\mathbf{R}^n)$, there exists $f_{n,m} \in C_c^\infty(\mathbf{R}^n)$, $m = 1, 2, \ldots$, converging to f_n in $H^s(\mathbf{R}^n)$ when $m \to \infty$ and such that $\operatorname{supp} f_{n,m} \subset K_n$, where K_n is a fixed (with respect to m) compact set. But then $f_{n,m} \to f_n$ also in $H^s(\mathbf{R}^n, \rho)$, after which the conclusion follows easily. □

A case of particular interest in this thesis is that of the quadratic potential $V(x) = \frac{1}{2}\, x^T x$, for which we have the following important bound (we present only the one-dimensional version):

Proposition A.9. *Let $g(x) := \mathrm{e}^{-x^2/2}$. Then for any $u \in H^1(\mathbf{R}, g)$, we have*

$$\|xu\|_g \leq 2\, \|u\|_{1,g}. \tag{A.8}$$

Proof. Our proof we present is based (very loosely) on [195, Lemma B.6]. Using integration by parts, we readily verify that, if $u_1 \in C^1(\mathbf{R})$ and $u_2 \in C_c^1(\mathbf{R})$, then

$$\int_{\mathbf{R}} u_1(x) \left(x - \frac{\mathrm{d}}{\mathrm{d}x} \right) u_2(x)\, g(x)\, \mathrm{d}x = \int_{\mathbf{R}} u_2(x)\, u_1'(x)\, g(x)\, \mathrm{d}x. \tag{A.9}$$

Taking $u_1 = u_n$ and $u_2 = x u_n$ for $u_n \in C_c^1(\mathbf{R})$, rearranging the terms, and using Young's inequality, we obtain

$$\begin{aligned} \int_{\mathbf{R}} \big(x\, u_n(x)\big)^2\, g(x)\, \mathrm{d}x &= \int_{\mathbf{R}} 2\, x\, u_n(x)\, u_n'(x)\, g(x)\, \mathrm{d}x + \int_{\mathbf{R}} \big(u_n(x)\big)^2\, g(x)\, \mathrm{d}x \\ &\leq \frac{1}{2} \int_{\mathbf{R}} \big(x\, u_n(x)\big)^2\, g(x)\, \mathrm{d}x + 2 \int_{\mathbf{R}} \big(u_n'(x)\big)^2\, g(x)\, \mathrm{d}x + \int_{\mathbf{R}} \big(u_n(x)\big)^2\, g(x)\, \mathrm{d}x, \end{aligned}$$

and thus

$$\|xu_n\|_g^2 \leq 4\|u_n\|_{1,g}^2. \tag{A.10}$$

Now let $u_n \in C_c^1(\mathbf{R})$, $n = 1, 2, \ldots$ converge to u in $H^1(\mathbf{R}, g)$ and almost everywhere (finding such a sequence is possible by Proposition A.8 and in view of the fact that the weight is positive for all x). By Eq. (A.10), $v_n = x\sqrt{g}\,u_n$ is Cauchy in $L^2(\mathbf{R})$, and therefore by completeness there exists a subsequence converging to $v \in L^2(\mathbf{R})$ in $L^2(\mathbf{R})$ and almost everywhere. We deduce that $v = x\sqrt{g}\,u$, and taking the limit $n \to \infty$ in Eq. (A.10) we conclude to Eq. (A.8). □

Corollary A.10 (Adjunction relation)**.** *Equation* (A.9) *holds for general* $u_1, u_2 \in H^1(\mathbf{R}, g)$.

Remark A.1. Repeating the reasoning in the proof of Proposition A.9 for a general weight of the form $\rho = \mathrm{e}^{-V}$, for a smooth confining potential $V(\cdot)$ over $\mathbf{R}$, we obtain instead of Eq. (A.8)

$$\int_{\mathbf{R}} \left(\frac{1}{2}|V'(x)|^2 - V''(x)\right) \big(u(x)\big)^2 \rho(x)\,\mathrm{d}x \leq 4\int_{\mathbf{R}} \big(u'(x)\big)^2 \rho(x)\,\mathrm{d}x.$$

As we shall see below, the quantity multiplying $\big(u(x)\big)^2$ in the first integral plays an important role, and in particular in appears in a standard condition for the Poincaré inequality. ⊘

For a positive self-adjoint operator $-\mathcal{L}$ (including a negative sign might feel slightly unnatural at this point, but it leads to simpler formulas in Chapter 2), we also define the following generalized Sobolev spaces.

Definition A.11. Given $s \in \mathbf{N}$ and a positive selfadjoint operator $-\mathcal{L}$ on a Hilbert space H of functions on $\mathbf{R}^n$, we define $H^s(\mathbf{R}^n, \mathcal{L})$ as the space obtained by completion of $C_c^\infty(\mathbf{R}^n)$ for the inner product:

$$\langle u, v\rangle_{s,\mathcal{L}} = \sum_{i=0}^{s} \left\langle (-\mathcal{L})^i u, v\right\rangle_H.$$

The associated norm will be denoted by $\|\cdot\|_{s,\mathcal{L}}$.

Remark A.2. By density of the smooth, compactly support functions in $H^1(\mathbf{R}^n, \rho)$ (Proposition A.8), we observe, using integration by parts, that $H^1(\mathbf{R}^n, \rho) = H^1(\mathbf{R}^n, \mathcal{L})$, where $-\mathcal{L}$ is the positive selfadjoint operator on $L^2(\mathbf{R}^n, \rho)$ defined by $\mathcal{L} = \Delta - \nabla V \cdot \nabla$. Considering the one-dimensional case for simplicity, and denoting by $\mathrm{d}_x^* := V' - \mathrm{d}_x$ the $L^2(\mathbf{R}, \rho)$ adjoint of the derivative operator d_x, we can write $\mathcal{L} = -\mathrm{d}_x^*\mathrm{d}_x$ and so, for any $u \in C_c^\infty(\mathbf{R})$,

$$\left\langle (-\mathcal{L})^2 u, u\right\rangle_\rho = \left\langle u'', u''\right\rangle_\rho + \left\langle [\mathrm{d}_x, \mathrm{d}_x^*]u', u'\right\rangle_\rho,$$

where $[\mathrm{d}_x, \mathrm{d}_x^*] = V''$ denotes the commutator between d_x and d_x^*. In the case of a quadratic potential the commutator is constant, so $H^2(\mathbf{R}, \rho) = H^2(\mathbf{R}, \mathcal{L})$ and, by extension, $H^s(\mathbf{R}, \rho) = H^s(\mathbf{R}, \mathcal{L})$ for any $s \in \mathbf{N}$. For more complicated potentials the equivalence of the definitions does not necessarily hold. ⊘

For the rest of this appendix, we make the additional assumption that the confining potential V satisfies the following:

$$\lim_{|y|\to\infty} \left(\frac{1}{4}|\nabla V|^2 - \frac{1}{2}\Delta V\right) = \infty \quad \text{and} \quad \lim_{|y|\to\infty} |\nabla V| = \infty. \tag{A.11}$$

With this, the following compactness result holds.

Proposition A.12. *Assume that (A.11) holds. Then the embedding $H^1(\mathbf{R}^n, \rho) \subset L^2(\mathbf{R}^n, \rho)$ is compact, and the measure ρ satisfies Poincaré inequality:*

$$\int_{\mathbf{R}^n} (u - \bar{u})^2 \, \rho \, \mathrm{d}y \leq C \int_{\mathbf{R}^n} |\nabla u|^2 \, \rho \, \mathrm{d}y \quad \forall u \in H^1(\mathbf{R}^n, \rho),$$

where $\bar{u} = \int_{\mathbf{R}^n} u \, \rho \, \mathrm{d}y$.

Proof. See [147], sec. 8.5, p. 216. □

Remark A.3. Alternative conditions on the potential that ensure that the corresponding Gibbs measure satisfies a Poincaré inequality are presented in [142, Theorem 2.5]. See also [14] an L^1 version of the Poincaré inequality and the associated proof. ⊘

Now we consider the unitary transformation $\mathrm{e}^{-V/2} : L^2(\mathbf{R}^n, \rho) \to L^2(\mathbf{R}^n)$, and characterize the spaces obtained by applying this mapping to the weighted Sobolev spaces.

Proposition A.13. *The multiplication operator $\mathrm{e}^{-V/2}$ is a unitary transformation from $H^s(\mathbf{R}^n, \mathcal{L})$ to $H^s(\mathbf{R}^n, \mathcal{H})$, where $-\mathcal{H}$ is the positive selfadjoint operator on $L^2(\mathbf{R}^n)$ defined by*

$$-\mathcal{H} = \mathrm{e}^{-V/2} \, \mathcal{L} \, \mathrm{e}^{V/2} = -\Delta + \left(\frac{|\nabla V|^2}{4} - \frac{\Delta V}{2} \right) =: -\Delta + W.$$

Proof. Since $(-\mathcal{H})^i = \mathrm{e}^{-V/2} (-\mathcal{L})^i \, \mathrm{e}^{V/2}$, $\langle u, v \rangle_{s,\mathcal{L}} = \left\langle \mathrm{e}^{-V/2} u, \mathrm{e}^{-V/2} v \right\rangle_{s,\mathcal{H}}$ for any $u, v \in C_c^\infty(\mathbf{R}^n)$ and any exponent $i \in \mathbf{N}$, from which the result follows by density. □

The space $H^1(\mathbf{R}^n, \mathcal{H})$, for $\mathcal{H}$ defined as above, is of particular relevance in Chapter 2. This space can be equivalently defined by

$$H^1(\mathbf{R}^n, \mathcal{H}) = \left\{ u \in H^1(\mathbf{R}^n) : \int_{\mathbf{R}^n} |W| u^2 \, \mathrm{d}y < \infty \right\},$$

and, for $u \in H^1(\mathbf{R}^n, \mathcal{H})$,

$$\|u\|_{1,\mathcal{H}}^2 = \int_{\mathbf{R}^n} u^2 + |\nabla u|^2 \, \mathrm{d}y + \int_{\mathbf{R}^n} W u^2 \, \mathrm{d}y.$$

A.3 Hermite polynomials and Hermite functions

In this appendix, we recall some results on Hermite polynomials and Hermite functions. These results will be useful for the presentation of the numerical methods and the derivation of the associated theoretical results in Chapters 2 and 3. Chapters 2 and 3, as well as for the implementation of the *Python* library for the Hermite spectral method [216] described in Chapter 6.

A.3.1 Hermite polynomials and Hermite functions in 1 dimension

Denoting by $g(\,\cdot\,)$ the unit Gaussian weight,

$$g(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

the associated orthonormal Hermite polynomials are defined by

$$H_n(x) = \frac{(-1)^n}{\sqrt{n!}} \exp\left(\frac{x^2}{2}\right) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \left(\exp\left(-\frac{x^2}{2}\right)\right), \qquad n = 0, 1, \ldots \tag{A.12}$$

These polynomials satisfy the recursion relation:

$$\begin{aligned} H_{n+1}(x) &= -\sqrt{\frac{1}{n+1}} \exp\left(\frac{x^2}{2}\right) \frac{\mathrm{d}}{\mathrm{d}x} \left(\exp\left(-\frac{x^2}{2}\right) H_n(x)\right), && \text{(Definition)} \\ &= \sqrt{\frac{1}{n+1}} \left(x - \frac{\mathrm{d}}{\mathrm{d}x}\right) H_n(x). && \text{(Leibniz formula)} \end{aligned} \tag{A.13}$$

The operator $\mathrm{d}_x^* := x - \mathrm{d}_x$ on the right-hand side of Eq. (A.13) is the formal $L^2(\mathbf{R}; g)$ adjoint of d_x, see Eq. (A.9). Using this recursion relation, we can verify that the polynomials defined by Eq. (A.12) are indeed orthonormal with respect to the weight $g(\,\cdot\,)$: assuming without loss of generality that $j \leq i$ and using the adjunction identity (A.9) repeatedly, we obtain

$$\int_{\mathbf{R}} H_i(x)\, H_j(x)\, g(x)\, \mathrm{d}x = \sqrt{\frac{1}{i!}} \int_{\mathbf{R}} H_j^{(i)}(x)\, g(x)\, dx = \delta_{ij},$$

where the calculation in the case $i = j$ can be carried out by keeping track of the term of highest degree. There are several other standard definitions of the Hermite polynomials, notably the so-called "physicists' Hermite polynomials" and the "probabilists' Hermite polynomials", but in this thesis we work exclusively with the normalized versions, as defined by Eq. (A.12) in one dimension. In fact, it is possible to show that the Hermite polynomials (A.12) form a complete orthonormal basis of $L^2(\mathbf{R}; g(\,\cdot\,))$, i.e. for every $u \in L^2(\mathbf{R}; g)$,

$$\sum_{i=0}^{d} H_i \, \langle u, H_i \rangle_g \to u \quad \text{in } L^2(\mathbf{R}, g) \text{ as } d \to \infty.$$

A proof is presented in e.g. [9]. By continuity of the norm, a corollary of the completeness is that the $L^2(\mathbf{R}; g)$ norm of a function can be expressed in terms of the coefficients of its expansion in Hermite polynomials:

$$\|f\|^2_{L^2(\mathbf{R};g)} = \sum_{i=1}^{\infty} \langle f, H_i \rangle_g^{\,2}.$$

We will call *Hermite transform* the operator:

$$\begin{aligned} \mathcal{T} : L^2(\mathbf{R}, g) &\to \ell^2 \\ u &\mapsto \left(\langle u, H_0 \rangle_g, \langle u, H_1 \rangle_g, \ldots \right). \end{aligned} \tag{A.14}$$

Combining orthonormality with the recursion relation (A.13) and the adjunction relation encapsulated in Corollary A.10, we obtain two other recursion relations (the second follows from the first and Eq. (A.13)):

$$H'_{n+1} = \sum_{i=0}^{n} H_i \int_{\mathbf{R}} H'_{n+1}(x)\, H_i(x)\, g(x)\, \mathrm{d}x = \sqrt{n+1}\, H_n; \tag{A.15}$$

$$H_{n+1} = \sqrt{\frac{1}{n+1}}\, x\, H_n - \sqrt{\frac{n}{n+1}} H_{n-1}. \tag{A.16}$$

Let us now assume that $f \in H^n(\mathbf{R}, g)$, and let us look at the Hermite expansion of $f^{(m)}$, $m \leq n$. Using Corollary A.10 repeatedly again, we obtain:

$$\begin{aligned} \int_{\mathbf{R}} f^{(m)}(x)\, H_i(x)\, g(x)\, \mathrm{d}\, x &= \int_{\mathbf{R}} f(x)\, (\mathrm{d}_x^*)^m H_i(x)\, g(x)\, \mathrm{d}x \\ &= \sqrt{\frac{(i+m)!}{i!}} \int_{\mathbf{R}} f(x)\, H_{i+m}(x)\, g(x)\, \mathrm{d}x. \end{aligned} \tag{A.17}$$

Note in particular that if $f \in H^n(\mathbf{R}, g)$ for all $n \geq 0$, the coefficients of the associated expansion in Hermite polynomials decrease in absolute value to zero faster than any inverse polynomial.

Characterization of the weighted Sobolev spaces

Equation (A.17) provides us with an alternative but equivalent manner of defining the weighted Sobolev spaces $H^p(\mathbf{R}, g)$, initially defined in Definition A.7:

Proposition A.14. *The weighted Sobolev space $H^p(\mathbf{R}, g)$ can equivalently be defined as*

$$H^p(\mathbf{R}, g) = \left\{ u \in L^2(\mathbf{R}, g) : \sum_{i=0}^{\infty} i^p\, u_i^2 < \infty \right\}, \tag{A.18}$$

where u_i are the coefficients of the expansion of u in Hermite polynomials.

Proof. if $u \in H^p(\mathbf{R}, g)$ in the sense of Definition A.7, then using Eq. (A.17) we have that

$$\sum_{i=p}^{\infty} \frac{i!}{(i-p)!} |u_i|^2 = \sum_{i=0}^{\infty} \frac{(i+p)!}{i!} |u_{i+p}|^2 < \infty.$$

For i large enough,

$$i^p \leq 2 \left(\frac{i!}{(i-p)!} \right),$$

which implies that the coefficients satisfy the requirement of definition (A.18). Assume now that $\sum_{i=0}^{\infty} i^p\, u_i^2 < \infty$, i.e. $u \in H^p(\mathbf{R}, g)$ in the sense of Eq. (A.18), and let us show that

$$v = \sum_{i=0}^{\infty} u_{i+p} \sqrt{\frac{(i+p)!}{i!}}\, H_i(x),$$

which is convergent in $L^2(\mathbf{R}, g)$ by assumption, is the p-th weak derivative of u. To this end,

take $\varphi \in C_c^\infty(\mathbf{R})$, and define $\psi(x) \in C_c^\infty(\mathbf{R})$ as $\psi(x) = \varphi(x)/g(x)$. Then

$$\begin{aligned}(-1)^p \int_{\mathbf{R}} u\,\varphi^{(p)}\,\mathrm{d}x &= \int_{\mathbf{R}} u(x)\,(\mathrm{d}_x^*)^p \psi(x)\,g(x)\,\mathrm{d}x \\ &= \int_{\mathbf{R}} u(x) \left(\sum_{i=p}^{\infty} \psi_{i-p} \sqrt{\frac{i!}{(i-p)!}}\, H_i(x) \right) g(x)\,\mathrm{d}x \\ &= \sum_{i=0}^{\infty} \sqrt{\frac{(i+p)!}{i!}}\, u_{i+p}\,\psi_i \\ &= \int_{\mathbf{R}} \left(\sum_{i=0}^{\infty} u_{i+p} \sqrt{\frac{(i+p)!}{i!}}\, H_i(x) \right) \psi(x)\,g(x)\,\mathrm{d}x = \int_{\mathbf{R}} v\,\phi\,\mathrm{d}x.\end{aligned}$$

where we used the continuity of the weighted $L^2(\mathbf{R}, g)$ inner product and the fact that the series in the last equation is convergent in $L^2(\mathbf{R}, g)$ by assumption, and with ψ_i, $i = 0, 1, \ldots$ the Hermite coefficients of $\psi(\cdot)$. □

By combining Eqs. (A.13) and (A.15), we see that Hermite polynomials are the eigenfunctions of a second-order operator:

$$-\mathcal{L}\,H_i := \mathrm{d}_x^* \frac{\mathrm{d}}{\mathrm{d}x} H_i := \left(x - \frac{\mathrm{d}}{\mathrm{d}x} \right) \frac{\mathrm{d}}{\mathrm{d}x} H_i = i\,H_i, \tag{A.19}$$

which is essential to proving approximation results. We note that the eigenvalues grow linearly, which explains the square root in the rate of convergence in the results presented below. By the equivalence of Definition A.7 and Definition A.11 for the unit Gaussian weight (see Remark A.2), Proposition A.14 also implies that

$$H^p(\mathbf{R}, \mathcal{L}) = \left\{ u \in L^2(\mathbf{R}, g) : \sum_{i=0}^{\infty} i^p\, u_i^2 < \infty \right\}.$$

This set equality can be shown independently of Proposition A.14, which is interesting in its own right and leads to an explicit expression of the norm $\|\cdot\|_{\mathcal{L},p}$. Another advantage of this approach is that it carries over easily to the multidimensional case.

Proposition A.15. *For $u \in H^p(\mathbf{R}, \mathcal{L})$,*

$$\|u\|_{p,\mathcal{L}}^2 = \sum_{i=0}^{\infty} (1 + \lambda_i + \lambda_i^2 + \cdots + \lambda_i^p)\, u_i^2, \quad \textit{where} \quad u_i = \langle u, H_i \rangle_g \quad \textit{and} \quad \lambda_i = i. \tag{A.20}$$

In addition $u \in H^p(\mathbf{R}, \mathcal{L})$ if and only if the sum on the right-hand side converges.

Proof. Let $\mathcal{A} = \sum_{i=0}^{p} (-\mathcal{L})^i$ and $\mu_i = 1 + \lambda_i + \lambda_i^2 + \cdots + \lambda_i^p$, and consider first that $u \in H^p(\mathbf{R}, \mathcal{L})$. By definition, there exists a sequence u^n in $C_c^\infty(\mathbf{R})$, $n \in \mathbf{N}$, that is Cauchy in $H^p(\mathbf{R}, g)$. Since it holds that $\mathcal{A}u^n \in C_c^\infty(\mathbf{R})$ also,

$$\|u^n\|_{p,\mathcal{L}}^2 = \langle \mathcal{A}u^n, u^n \rangle_g = \sum_{i=1}^{\infty} u_i^n\, \langle \mathcal{A}u^n, H_i \rangle_g = \sum_{i=1}^{\infty} u_i^n\, \langle u^n, \mathcal{A}H_i \rangle_g = \sum_{i=1}^{\infty} \mu_i (u_i^n)^2,$$

by Eq. (A.19), and where $u_i^n := \langle u^n, H_i \rangle_g$. Since the sequence u^n is Cauchy in $H^p(\mathbf{R}, \mathcal{L})$, the sequence $(\sqrt{\mu_0}\, u_0^n, \sqrt{\mu_1}\, u_1^n, \dots)$, $n \in \mathbf{N}$, is Cauchy in ℓ^2, and therefore by completeness we conclude to Eq. (A.20).

It remains to show that if the series is convergent, then $u \in H^p(\mathbf{R}, \mathcal{L})$. To this end, it is sufficient to verify that the sequence $u^n = \sum_{i=1}^n \langle u, H_i \rangle_g H_i \in C^\infty(\mathbf{R}) \cap H^p(\mathbf{R}, \mathcal{L})$, for $n = 1, 2, \dots$, is Cauchy in the weighted Sobolev space $H^p(\mathbf{R}, \mathcal{L})$. But this is clear, because by Eq. (A.19)

$$\|u^{n+m} - u^n\|_{p,\mathcal{L}}^2 = \sum_{i=n+1}^{n+m} \mu_i\, u_i^2,$$

which goes to zero for n large enough by assumption. □

Approximation result

Denoting by Π_d the $L^2(\mathbf{R}, g)$ projection operator on $\operatorname{span}\{H_0, H_1, \dots, H_d\}$, the following theorem is a consequence of Eq. (A.17); see [195] for details.

Theorem A.16 (Approximation by Hermite polynomials)**.** *For any $u \in H^m(\mathbf{R}, g)$ with $0 \le m \le d+1$, the following inequality holds:*

$$\left\| \frac{\mathrm{d}^\ell}{\mathrm{d}x^\ell}(\Pi_d u - u) \right\|_g \le \sqrt{\frac{(d-m+1)!}{(d-\ell+1)!}}\, \|u^{(m)}\|_g, \qquad \ell = 0, \dots, m.$$

Proof. Using the notation $u_i = \langle u, H_i \rangle_g$, we have by Eq. (A.17)

$$\begin{aligned} \left\| \frac{\mathrm{d}^\ell}{\mathrm{d}x^\ell}(\Pi_d u - u) \right\|_g^2 &= \sum_{i=d+1}^{n} \frac{i!}{(i-l)!}\, u_i^2, \\ &\le \sup_{i \ge d+1} \left\{ \frac{(i-m)!}{(i-l)!} \right\} \sum_{i=d+1}^{n} \frac{i!}{(i-m)!}\, u_i^2, \\ &= \frac{(d-m+1)!}{(d-\ell+1)!}\, \|u^{(m)}\|_g^2, \end{aligned}$$

which concludes the proof. We note that, for large d, the error scales approximately as $d^{\frac{l-m}{2}}$. □

Generalized Hermite functions

The Hermite polynomials, as defined by Eq. (A.12), are suitable for the approximation of functions with respect to the norm of $L^2(\mathbf{R}, g)$, which assigns a significant weight only to the region around $x = 0$. For the approximation with respect to the flat $L^2(\mathbf{R})$ norm, or with respect to other norms that penalize growth as $x \to \infty$, such as the weighted $L^2\left(\mathbf{R}, e^{x^2/2}\right)$ norm, one can use the basis functions $(\mathrm{e}^{-U/2}\, H_i)_{i=0}^\infty$ for some function U, which constitute an orthonormal basis of $L^2\left(\mathbf{R}^n, \mathrm{e}^U\, g\right)$. For $u \in L^2\left(\mathbf{R}, \mathrm{e}^U\, g\right)$, we define the generalized Hermite transform associated

with the factor $\mathrm{e}^{-U/2}$, which we denote by $\mathcal{T}_U$

$$\begin{aligned} \mathcal{T}_U : L^2\left(\mathbf{R}, \mathrm{e}^U\, g\right) &\to \ell^2 \\ u &\mapsto \left(\left\langle u, \mathrm{e}^{-U/2}\, H_0 \right\rangle_{\mathrm{e}^U\, g}, \left\langle u, \mathrm{e}^{-U/2}\, H_1 \right\rangle_{\mathrm{e}^U\, g}, \dots \right). \end{aligned} \tag{A.21}$$

Note that $\mathcal{T}_U(u) = \mathcal{T}(\mathrm{e}^{U/2}\, u)$, i.e. $T_U(u)$ is the usual Hermite transform of $\mathrm{e}^{U/2}\, u$. This formalism enables us to treat in a unified manner the case of Hermite polynomials ($U = 0$), of Hermite functions ($\mathrm{e}^{-U/2} = \sqrt{g}$), as well as other useful cases. As an example of why such generality can be useful, it has been shown in [74] that the choice $\mathrm{e}^{-U/2} = g$ leads to basis functions, referred to as generalized Hermite functions in that paper, that can be used to design an efficient numerical method for the solution of a particular Fokker–Planck equation. By the property (A.19), we see that $(\mathrm{e}^{-U/2}\, H_i)_{i=0}^{\infty}$ are the eigenfunctions of the operator:

$$u \mapsto (\mathrm{e}^{-U/2}\, \mathcal{L}\, \mathrm{e}^{U/2})u,$$

which is of Schrödinger type when $\mathrm{e}^{-U} = g$, see [174]. Introducing the notations $\Pi_d^U := \mathrm{e}^{-U/2}\, \Pi_d\, \mathrm{e}^{U/2}$ and $\partial_x^U := \partial_x + U'/2 = \mathrm{e}^{-U/2}\, \partial_x(\mathrm{e}^{U/2}\, \cdot)$, we have the following immediate corollary of Theorem A.16.

Theorem A.17 (Approximation by generalized Hermite functions)**.** *For any u such that $(\partial_x^U)^m u \in L^2\left(\mathbf{R}, \mathrm{e}^U\, g\right)$ with $0 \le m \le d+1$,*

$$\|(\partial_x^U)^\ell(\Pi_d^U u - u)\|_{\mathrm{e}^U\, g} \le \sqrt{\frac{(d-m+1)!}{(d-\ell+1)!}}\, \|(\partial_x^U)^m u\|_{\mathrm{e}^U\, g}, \qquad \ell = 0, \dots, m.$$

In the case $\mathrm{e}^{-U} = g$ (orthonormal Hermite functions in $L^2(\mathbf{R})$), one can prove a similar statement with the usual derivative instead of ∂_x^U in the left-hand side, see [195, Theorem 7.14].

Scaling factor

In addition to the function $\mathrm{e}^{-U/2}$ multiplying the Hermite polynomials in the definition of basis functions, it is usual in numerical simulation to introduce a scaling factor, which can be chosen appropriately depending on how localized the function to be approximated is. For given parameters $\mu, \sigma \in \mathbf{R}$, we will employ the more general notation $g(\,\cdot\,; \mu, \sigma^2) := (1/\sigma)\, g((\,\cdot\, - \mu)/\sigma)$ to denote the Gaussian density with mean μ and standard deviation σ, and $H_n(x; \mu, \sigma^2) = H_n((x-\mu)/\sigma)$ to denote the associated Hermite polynomials. By a change of variable, we observe that these polynomials are orthonormal with respect to the weight function $g(\,\cdot\,; \mu, \sigma^2)$:

$$\int_{\mathbf{R}} H_i(x; \mu, \sigma^2)\, H_j(x; \mu, \sigma^2)\, g(x; \mu, \sigma^2)\, \mathrm{d}x = \int_{\mathbf{R}} H_i(x)\, H_j(x)\, g(x)\, \mathrm{d}x = \delta_{ij}. \tag{A.22}$$

When $\mu = 0$, we will sometimes use the notations $g_\sigma(x) := g(x; 0, \sigma^2)$ and $H_i^\sigma(x) := H_i(x; 0, \sigma)$ for conciseness. Although we do not present them explicitly, approximation results similar to Theorems A.16 and A.17 can be proved in the presence of this scaling factor, with the only difference being the presence of additional constant factors on the right-hand side.

The importance of choosing this scaling factor in a suitable manner was demonstrated rigorously in [209], where the author shows that, for Gaussian-type functions, the optimal value of σ should depend on the number of basis function used with a dependence of the form $1/\sqrt{d}$. This can be justified intuitively by taking into account that, on the one hand, the behavior of Hermite functions as $x \to \infty$ is well-understood, with the final inflection point occurring at $x \propto \sqrt{d}\,\sigma$ (in fact, $x = \sqrt{4d+2}\,\sigma$) and a rapid decrease to 0 beyond that point [139], and, on the other hand, scaled Hermite functions are the eigenfunctions of the Fourier transform operator, up to dilations/contractions. Since a contraction in real space results in a dilation in Fourier space, by choosing a scaling factor that decreases with d one effectively favors exploration in Fourier space. In particular, choosing $\sigma \propto 1/\sqrt{d}$ leads to a situation where the position of the final inflection point of the Hermite function of highest order remains approximately constant (while the position of the final inflection point of its Fourier transform grows linearly), so it is crucial in that case to ensure that $\sqrt{d}\,\sigma$ is large enough to ensure that the basis functions cover the support of the target function.

A.3.2 Hermite polynomials and Hermite functions in n dimensions

The Hermite polynomials can be naturally extended to the multidimensional case. For $\mu \in \mathbf{R}^n$ and a symmetric positive definite matrix $\Sigma \in \mathbf{R}^{n\times n}$, consider the Gaussian density $g_{\mu,\Sigma}$ of mean μ and covariance matrix Σ. Let D and Q be diagonal and orthogonal matrices such that $\Sigma = QDQ^T$, and note $S = QD^{1/2}$, such that $\Sigma = SS^T$. The polynomials defined by

$$\mathbf{H}_\alpha(y;\mu,\Sigma) \;=\; \mathbf{H}^*_\alpha(S^{-1}(y-\mu)), \qquad \text{with } \alpha \in \mathbf{N}^n \text{ and } \mathbf{H}^*_\alpha(z) = \prod\nolimits_{k=1}^{n} H_{\alpha_k}(z_k), \tag{A.23}$$

form a complete orthonormal basis of $L^2(\mathbf{R}^n, g_{\mu,\Sigma})$. Note that the Hermite polynomial corresponding to a multi-index α depends on the orthogonal matrix Q chosen. When μ and Σ are clear from the context, we will sometimes omit them to simplify the notation. In addition to forming a complete orthonormal basis, the Hermite polynomials defined by Eq. (A.23) are the eigenfunctions of the Ornstein–Ulhenbeck operator

$$-\,\mathcal{L}_{\mu,\Sigma} = \Sigma^{-1}(y-\mu)\cdot\nabla - \Delta, \tag{A.24}$$

which is the operator $\mathcal{L}$ from Appendix A.2 for $\rho = g_{\mu,\Sigma}$. The operator $-\mathcal{L}_{\mu,\Sigma}$ is positive and symmetric on $L^2(\mathbf{R}^n, g_{\mu,\Sigma})$: writing $\mathcal{L}_{\mu,\Sigma}\cdot = g^{-1}_{\mu,\Sigma}\nabla\cdot(g_{\mu,\Sigma}\nabla\cdot)$, we have

$$\int_{\mathbf{R}^n} (\mathcal{L}_{\mu,\Sigma}u)\, v\, g_{\mu,\Sigma}\,\mathrm{d}x = \int_{\mathbf{R}^n} (\mathcal{L}_{\mu,\Sigma}v)\, u\, g_{\mu,\Sigma}\,\mathrm{d}x = \int_{\mathbf{R}^n} \nabla u\cdot\nabla v\, g_{\mu,\Sigma}\,\mathrm{d}x \tag{A.25}$$

for all sufficiently regular functions u and v.

Proposition A.18. *The Hermite polynomials $\mathbf{H}_\alpha(y;\mu,\Sigma)$, $\alpha \in \mathbf{N}^n$, are the eigenfunctions of the operator $-\mathcal{L}_{\mu,\Sigma}$ defined in Eq. (A.24). The eigenvalue associated to $\mathbf{H}_\alpha(y;\mu,\Sigma)$ is given by*

$$\lambda_\alpha = \sum_{i=1}^{n} \alpha_i \lambda_i, \tag{A.26}$$

where $\{\lambda_i\}_{i=1}^n$ *are the diagonal elements of* D^{-1}.

Proof. Let us introduce, for an orthogonal matrix $P \in \mathbf{R}^{n\times n}$ and a vector $u \in \mathbf{R}^n$, the rotation operator $\mathcal{R}_P$ and the translation operator $\mathcal{T}_u$ defined respectively by $\mathcal{R}_P f(\cdot) = f(P^T\cdot)$ and $\mathcal{T}_u f(\cdot) = f(\cdot - u)$. With these notations, note that $\mathcal{L}_{\mu,\Sigma} = \mathcal{T}_\mu \circ \mathcal{R}_Q \circ \mathcal{L}_{0,D} \circ \mathcal{R}_{Q^T} \circ \mathcal{T}_{-\mu}$ and $\mathbf{H}_\alpha(\cdot;\mu,\Sigma) = \mathcal{T}_\mu \circ \mathcal{R}_Q \mathbf{H}_\alpha(\cdot;0,D)$, so it is sufficient to show that $\mathbf{H}_\alpha(\cdot;0,D)$ is an eigenfunction of $\mathcal{L}_{0,D}$. Writing $\mathcal{L}_{0,D}$ as a sum of one-dimensional operators, we see that it is enough to prove this in one dimension, which follows from a Eq. (A.19). □

We have the following result, which can be shown similarly to Proposition A.15.

Proposition A.19. *For* $u \in H^s(\mathbf{R}^n, \mathcal{L}_{\mu,\Sigma})$,

$$\|u\|^2_{s,\mathcal{L}_{\mu,\Sigma}} = \sum_{\alpha\in\mathbf{N}^n} (1 + \lambda_\alpha + \lambda_\alpha^2 + \cdots + \lambda_\alpha^s) c_\alpha^2, \qquad \textit{where} \qquad c_\alpha = \langle u, \mathbf{H}_\alpha\rangle_{g_{\mu,\Sigma}}.$$

In addition $u \in H^s(\mathbf{R}^n, \mathcal{L}_{\mu,\Sigma})$ *if and only if the sum on the right-hand side converges.*

It follows that Hermite polynomials are also orthogonal in $H^s(\mathbf{R}^n, \mathcal{L}_{\mu,\Sigma})$, and, using Proposition A.18, that the norm of $\mathbf{H}_\alpha(y;\mu,\Sigma)$ in that space is given by $(1 + \lambda_\alpha + \lambda_\alpha^2 + \cdots + \lambda_\alpha^s)^{1/2}$. In the more general, multidimensional setting considered in this subsection, it is more convenient, in order to establish approximation results, to use the spaces $H^s(\mathbf{R}^n, \mathcal{L}_{\mu,\Sigma})$ than those defined by Definition A.7, but it should be kept in mind that the two definitions are equivalent for Gaussian weights. Noting that $1 + r + r^2 + \cdots + r^s \le \mathrm{e}^{\frac{1}{r}}\, r^s$ for $r > 0$ and that $\lambda_\alpha \to \infty$ when $|\alpha| \to \infty$, this implies that, for a function $u \in H^s(\mathbf{R}^n, \mathcal{L}_{\mu,\Sigma})$ and $u_\alpha = \langle u, \mathbf{H}_\alpha(\cdot;\mu,\Sigma)\rangle_{g_{\mu,\Sigma}}$,

$$\left(u_0^2 + \sum_{|\alpha|>0} \lambda_\alpha^s u_\alpha^2\right) \le \|u\|^2_{s,\mathcal{L}_{\mu,\Sigma}} \le L \left(u_0^2 + \sum_{|\alpha|>0} \lambda_\alpha^s u_\alpha^2\right), \qquad \text{where } L = \max_{|\alpha|>0} \mathrm{e}^{\frac{1}{\lambda_\alpha}}. \tag{A.27}$$

Hermite polynomials have good approximation properties for smooth functions in $L^2(\mathbf{R}^n, g_{\mu,\Sigma})$. In Proposition A.20, we denote by $\Pi_d \cdot : L^2(\mathbf{R}^n, g_{\mu,\Sigma}) \to \mathbf{P}(d)$ the $L^2(\mathbf{R}^n, g_{\mu,\Sigma})$ projection operator on the space of polynomials of degree less than or equal to d.

Proposition A.20 (Approximation by polynomials in weighted spaces)**.** *Let* Σ *be a symmetric positive definite matrix, and suppose that* $u \in H^s(\mathbf{R}^n, \mathcal{L}_{\mu,\Sigma})$. *Then*

$$\|u - \Pi_d u\|_{r,\mathcal{L}_{\mu,\Sigma}} \le C(\Sigma, r, s)\,(d+1)^{-\frac{(s-r)}{2}}\, \|u\|_{s,\mathcal{L}_{\mu,\Sigma}},$$

for $r \in \mathbf{N}$ *such that* $0 \le r \le s$.

Proof. From (A.27), we have that

$$\|u - \Pi_d u\|^2_{r,\mathcal{L}_{\mu,\Sigma}} \le L \sum_{|\alpha|>d} \lambda_\alpha^r u_\alpha^2 \le L\, M^{r-s} \sum_{\alpha\in\mathbf{N}^n} \lambda_\alpha^s u_\alpha^2 \le LM^{r-s} \|u\|^2_{s,\mathcal{L}_{\mu,\Sigma}},$$

with $u_\alpha = \langle u, \mathbf{H}_\alpha(\cdot;\mu,\Sigma)\rangle_{g_{\mu,\Sigma}}$ and $M = \min_{|\alpha|>d} \lambda_\alpha$. Since $\lambda_\alpha > C(\Sigma)\,|\alpha|$, the conclusion follows. See [83, Theorem 3.1] for more details. □

From the Hermite polynomials, we define the Hermite functions $\mathbf{h}_\alpha(y;\mu,\Sigma)$ by:

$$\mathbf{h}_\alpha(y;\mu,\Sigma) = \sqrt{g_{\mu,\Sigma}}\,\mathbf{H}_\alpha(y;\mu,\Sigma) \quad \text{for } \alpha \in \mathbf{N}^n. \tag{A.28}$$

They form a complete orthonormal basis of $L^2(\mathbf{R}^n)$. It is possible to define generalized Hermite functions, with a factor possibly different from $g_{\mu,\Sigma}$, in the same manner as in the previous section, but for simplicity we will limit ourselves to the definition above. Since they are obtained from Hermite polynomials by a multiplication by $\sqrt{g_{\mu,\Sigma}}$, the Hermite functions (A.28) satisfy:

Proposition A.21. *Given $\mu \in \mathbf{R}^n$ and $\Sigma \in \mathbf{R}^{n\times n}$ positive definite, the Hermite functions $\mathbf{h}_\alpha(y;\mu,\Sigma)$ are the eigenfunctions of the operator:*

$$-\mathcal{H}_{\mu,\Sigma} = (g_{\mu,\Sigma})^{\frac{1}{2}}\,(-\mathcal{L}_{\mu,\Sigma})\,(g_{\mu,\Sigma})^{-\frac{1}{2}} = -\Delta + \left(\frac{(y-\mu)^T\Sigma^{-2}(y-\mu)}{4} - \frac{\operatorname{tr}\Sigma^{-1}}{2}\right),$$

with the same eigenvalues as in (A.26).

They also inherit the good approximation properties of Hermite polynomials expressed in Proposition A.20. We will denote by $\hat{\Pi}_d$ the $L^2(\mathbf{R}^n)$ projection operator on $\sqrt{g_{\mu,\Sigma}}\,\mathbf{P}(d)$, i.e.

$$\hat{\Pi}_d u = \sum_{|\alpha|\le d} \langle u, h_\alpha(\cdot;\mu,\Sigma)\rangle\; h_\alpha(\cdot;\mu,\Sigma).$$

Corollary A.22 (Approximation by Hermite functions in flat space)**.** *Let $\mu \in \mathbf{R}^n$ and $\Sigma \in \mathbf{R}^{n\times n}$ be a symmetric positive definite matrix, and suppose that $u \in H^s(\mathbf{R}^n, \mathcal{H}_{\mu,\Sigma})$. Then*

$$\|u - \hat{\Pi}_d u\|_{r,\mathcal{H}_{\mu,\Sigma}} \le C(\Sigma,r,s)\,(d+1)^{-\frac{(s-r)}{2}}\,\|u\|_{s,\mathcal{H}_{\mu,\Sigma}},$$

for any $r \in \mathbf{N}$ such that $0 \le r \le s$.

A.4 Discrete transforms and interpolation error

In practice, calculating the Hermite transform numerically requires the introduction of a quadrature. In this section we summarize the standard construction of the Gauss-Hermite quadrature and the associated error estimates. This section is based on [150, Chapter 3].

Introducing $a_d := \sqrt{d}$ and writing the recursion relation (A.16) in matrix form, we obtain

$$\begin{pmatrix} 0 & a_1 & & & & \\ a_1 & 0 & a_2 & & & \\ & a_2 & 0 & a_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{d-2} & 0 & a_{d-1} \\ & & & & a_{d-1} & 0 \end{pmatrix} \begin{pmatrix} H_0(x) \\ H_1(x) \\ H_2(x) \\ \vdots \\ H_{d-2}(x) \\ H_{d-1}(x) \end{pmatrix} = x \begin{pmatrix} H_0(x) \\ H_1(x) \\ H_2(x) \\ \vdots \\ H_{d-2}(x) \\ H_{d-1}(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ -a_d\,H_d(x) \end{pmatrix}.$$

Writing this equation as $\mathbf{T}_d\mathbf{H}_d(x) = x\mathbf{H}_d(x) + \mathbf{R}(x)$ for short, it is clear that the roots $x_{i,d}$ of $H_d(x)$ are the eigenvalues of $\mathbf{T}_d$. By symmetry of $\mathbf{T}_d$ and the fact that all the roots are simple,

the corresponding eigenvectors $\mathbf{H}_d(x_{i,d})$ are orthogonal, and therefore the columns vectors

$$\mathbf{v}_d^i := \frac{1}{\sqrt{\sum_{j=0}^{d-1} H_j^2(x_{i,d})}}\,\mathbf{H}_d(x_{i,d}), \qquad i = 1, 2, \dots, d,$$

form an orthonormal basis of $\mathbf{R}^d$. Collecting these columns in an orthogonal matrix $\mathbf{Q}$, and expressing the fact that any two lines ℓ_1 and ℓ_2 of $\mathbf{Q}$ are orthonormal too, we obtain

$$\sum_{i=1}^{d} \frac{H_{\ell_1}(x_{i,d})\, H_{\ell_2}(x_{i,d})}{\sum_{j=0}^{d-1} H_j^2(x_{i,d})} = \delta_{\ell_1 \ell_2}, \qquad 0 \le \ell_1, \ell_2 \le d-1. \tag{A.29}$$

Equation (A.29) is the basis for the definition of the quadrature nodes and weights. Introducing the discrete inner product

$$\langle u, v \rangle_{d-1} = \sum_{i=1}^{d} w_{i,d}\, u(x_{i,d})\, v(x_{i,d}), \qquad w_{i,d} = \frac{1}{\sum_{j=0}^{d} H_j^2(x_{i,d})}, \tag{A.30}$$

it is clear that $\langle u, v \rangle_{d-1} = \langle u, v \rangle_g$ provided that $u, v \in \mathbf{P}(d-1)$. In fact, since the quadrature points $x_{i,d}$ are the zeros of $H_d(\cdot)$, Eq. (A.30) holds under the more general condition that $uv \in \mathbf{P}(2d-1)$. See [195, Theorem 7.3] for a precise statement and alternative expressions of the weights.

Equation (A.30) enables the definition of the *discrete Hermite transform*:

$$\begin{aligned} \hat{\mathcal{T}}_d : C(\mathbf{R}) &\to \mathbf{R}^{d+1} \\ u &\mapsto \big(\langle u, H_0 \rangle_d, \langle u, H_1 \rangle_d, \dots, \langle u, H_d \rangle_d \big). \end{aligned}$$

We will denote by $\hat{u}_{i,d}$ the coefficients of the discrete Hermite transform of u calculated with $d+1$ nodes. It is clear that any polynomial $u \in \mathbf{P}(d)$ can be recovered from its discrete Hermite coefficients, $u = \sum_{i=0}^{d} \hat{u}_{i,d}\, H_i(x)$. Consequently, for a general $u \in C(\mathbf{R})$, the polynomial $\sum_{i=0}^{d} \hat{u}_{i,d}\, H_i(x)$ interpolates $u(\cdot)$ over the quadrature points $x_{i,d+1}$, $i = 1, \dots, d+1$. In this thesis, we always ensure that the error originating from the use of a numerical quadrature is negligible in comparison with the approximation error, which is controlled by Theorem A.16. For completeness, however, we include below a bound for the interpolation error; see [195, Theorem 7.17] for more details.

Theorem A.23 (Hermite interpolation error)**.** *For any $u \in H^m(\mathbf{R}, g)$ with $0 \le m \le d+1$, the following inequality holds:*

$$\left\| \frac{\mathrm{d}^\ell}{\mathrm{d}x^\ell} (\Pi_d u - u) \right\|_g \le C\, d^{\frac{1}{6} + \frac{l-m}{2}}\, \|u^{(m)}\|_g, \qquad \ell = 0, \dots, m,$$

where C is a constant independent of d or u.

Similarly to the approximation result A.16, Theorem A.23 can be extended to (possibly generalized) Hermite functions and to the multidimensional case, see e.g. [195, Theorem 7.17], but we will not present the associated error bounds here.