Politecnico di Torino

Politecnico di Torino

Physics of Complex Systems
A.a. 2024/2025
Sessione di Laurea Aprile 2025

From Biased Molecular Simulations to Unbiased Free Energy Landscapes

Relatori:
Politecnico di Torino: Andrea Pagnani
IBPC Paris: Jérôme Hénin

Candidato:
Bersano Daniele

Abstract

Understanding the free energy landscape is a fundamental step for understanding the thermodynamics and kinetics of complex molecular processes such as phase transitions, conformational changes, and chemical reactions. Traditional molecular dynamics simulations, however, are often limited by the difficulty of sampling rare-event due to high free-energy barriers. To address these challenges, accelerated simulation techniques employing biasing forces have been developed, though they can pollute the estimation of the underlying unbiased free energy profiles.

In this work, we introduce a novel inference framework that reconstructs unbiased free energy landscapes directly from biased simulation data. Our approach employs an overdamped Langevin model and exploits a Bayesian maximum likelihood estimation strategy to accurately determine the drift and diffusion parameters governing the system's effective dynamics. The methodology is systematically validated using a series of benchmark toy models, including both one and two dimensional double-well potentials under unbiased and biased conditions. Results show that when an optimal collective variable is chosen, the framework successfully recovers the true free energy landscape; conversely, suboptimal projections lead to noticeable inaccuracies, underscoring the critical role of variable selection. This work not only enhances the efficiency of free energy estimation from biased simulations but also provides a robust tool for extracting detailed thermodynamic and kinetic insights, with potential applications across biophysics, chemistry, and materials science.

Keywords

Bayesian Inference, Stochastic Processes, Langevin Equation, Free Energy estimation, Molecular Dynamics

Contents

Chapter 1
Introduction

In the field of atomistic computer simulations, in particular concerning rare events in biochemistry such as phase transitions, conformational changes or chemical reactions, scientists are often interested in predicting the underlying free energy landscapes of the process under exam.

A well suited tool to perform this kind of analysis is Molecular Dynamics (MD) simulations since they are able to sample metastable states, transitions and fluctuations [10]. However, especially in activation processes where a high-energy barrier has to be overcome, the simulation might end up being stuck in a local energy minimum. In order to escape such local minima simply by thermodynamical fluctuation, it is necessary to run the simulation for an extremely long time. Hence, it is very challenging to ergodically sample the entire energy landscape basing solely on trajectory data, for this reason some accelerated simulation techniques such as umbrella sampling [12] are employed; these techniques allow one to artificially lower free energy barriers adding ad hoc potentials, increasing thermodynamical fluctuation sizes via temperature increase or to steer the direction of the simulation exploiting fictitious forces.

Unfortunately the free energy landscape is often a highly dimensional and rarely known a priori function, which brought the community to investigate and try to reconstruct such free energy landscapes along a few selected Collective Variables (CV), chosen to be the most representative for the problem under analysis.

Choosing an optimal collective variable to describe the state of the system is a hard task, for instance, one might appreciate a CV that takes clear different values in different phases, thus acting as a metric or order parameter [7] in the evolution of the simulation itself.

Concerning dynamical aspects [3] it is sufficient to require it to obey an effective Langevin equation (see sec(2.1)).

The team of Fabio Pietrucci, our collaborator at IMPMC, Sorbonne Université, laid the foundation of this work [8], [9]. They inferred the drift and diffusion terms of an effective overdamped Langevin equation starting from some Molecular dynamics trajectories and randomly changing this two parameters, accepting the new proposed drift and diffusion if those lead to an increase of the likelihood function (see Sec. 2.2) according to a Metropolis - Monte Carlo criterion.

Although effective, the above described method turns out to be very inefficient in terms of computational cost and most importantly it was applicable only in unbiased simula-

tions therefore not relying on the different acceleration techniques [10] available nowadays to sample bigger regions of phase space in a smaller amount of time.
To tackle this problem the following work will implement a new methodology to perform more efficiently the same estimation.
In particular the method used aims at updating the drift and force of the considered Langevin model (see Sec.2.1) in order to reach ever-increasing values of likelihood of the transitions observed in the data.
To do so we first calculate the Jacobian (gradient) of the likelihood with respect to the input parameters, then the value of the parameters themselves are updated according to a gradient descend algorithm of the likelihood profile, in contrast with the previous method [9] who relied on random changes in the force and diffusion parameters.
In addition to this framework the presence of biasing forces is considered, such forces are meant to accelerate the simulation, allowing for the exploration of further regions of phase space. The effect of such additional (biasing) forces translates into a correction term to the likelihood of the transition step. This consideration allows for the reconstruction of the underlying free energy landscape, provided that sufficient data is available in the explored region of the phase space.

Chapter 2
Theoretical Setting

The phase space of the system under analysis is $\Theta(\vec{x},\vec{p})$, with $\mathrm{Dim}\big(\Theta(\vec{x},\vec{p})\big) = 2N$, the Energy landscape is described by the function $U(\vec{x},\vec{p})$ and the time evolution of the dynamical variables $\vec{x} = (x_1, x_2, ..., x_N)$ and $\vec{p} = (p_1, p_2, ...p_N)$ is dictated by
Hamilton's Equations:

$$\begin{cases} \dot{x}_i = \frac{\partial U(\vec{x},\vec{p})}{\partial p_i} \\ \dot{p}_i = -\frac{\partial U(\vec{x},\vec{p})}{\partial x_i} \end{cases}$$

The partition function of the system $\mathcal{Z}$ is the following

$$\mathcal{Z} = e^{-\beta \mathcal{A}} = \int e^{-\beta U(\vec{x},\vec{p})} d\vec{x} d\vec{p}$$

with $\mathcal{A} = -\beta^{-1} \log(\mathcal{Z})$ being the corresponding Free Energy.
We are interested in a projection onto a different space $\Omega(\vec{q})$ of far lower dimension, namely $\mathrm{Dim}\big(\Omega(\vec{x},\vec{p})\big) = M << N$.
This is done to reconstruct the kinetics and thermodynamics of a process that is well-described by the set of collective variables $\vec{q} = \{q_1 \dots q_M\}$, chosen to be a function of the dynamical variables $\{\vec{x},\vec{p}\}$ themselves, i.e. $q_i = \xi_i(x_1...x_N, p_1...p_N)$.
In this M-dimensional manifold the relative Free Energy will be obtained integrating the Boltzmann weight, imposing $\vec{q} = \vec{\xi}(x_1...x_N, p_1...p_N))$ while computing the partition function, obtaining the relative equilibrium probability density

$$\rho_{eq}(\vec{q}) = \int e^{-\beta U(\vec{x},\vec{p})} \delta(\vec{q} - \vec{\xi}(x_1...x_N, p_1...p_N))) d\vec{x} d\vec{p}$$

Then the logarithm of this expression returns the desired Free Energy $A(\vec{q})$:

$$A(\vec{q}) = -\frac{1}{\beta} \log\left\{ \int e^{-\beta U(\vec{x},\vec{p})} \delta(\vec{q} - \vec{\xi}(x_1...x_N, p_1...p_N))) d\vec{x} d\vec{p} \right\}$$

The dynamics in this lower dimensional projected space is no more described Hamilton's equations but by an effective Langevin Equation [7] instead.

To recap, the main shift in perspective consist in :

$$\begin{array}{ccc} \Theta(\vec{x},\vec{p}) & \longrightarrow & \Omega(\vec{q}) \\ \text{High dimensional} & & \text{Low dimensional} \\ \text{Hamilton's Equations} & \longrightarrow & \text{Effective Langevin Model} \\ \text{(Deterministic)} & & \text{(Stochastic)} \end{array}$$

In the following only unidimensional Langevin equation is considered :

2.1 Langevin Models

Restricting ourselves to a single generalised coordinate q with associated momentum $p = m\dot{q}$ the set of possible Langevin equations one can get [8] is the following :

- **Generalised Langevin Equation**

$$\dot{p} = -\frac{\partial A(q)}{\partial q} - \int_0^\infty ds\Gamma(s)p(t-s) + R(t) \tag{2.1}$$

that, if written explicitly in the q variable, making use of the fact that $p = m\dot{q}$ it reads:

$$m\ddot{q} = -\frac{\partial A(q)}{\partial q} - m\int_0^\infty ds\Gamma(s)\dot{q}(t-s) + R(t)$$

where $A(q)$ is the free energy landscape in which the dynamics takes place and $\Gamma(s)$ is a memory kernel, a time dependent function describing the correlation of the velocity at time t with itself at a previous time $t-s$ and $R(t)$ is a random force, relate to the memory kernel according to the fluctuation-dissipation theorem $\langle R(0)R(t)\rangle = k_B T m\Gamma(t)$.

- **Standard Langevin Equation**

Restricting ourselves to the case of delta-correlated velocities $\Gamma(s) = \gamma\delta(s)$ implies $\langle R(0)R(t)\rangle = k_B T m\gamma\delta(t)$ which is fulfilled by a random force of the kind

$$R(t) = \sqrt{2k_B T m\gamma}\eta(t)$$

where $\eta(t)$ is a Gaussian white noise, for which $\langle\eta(t)\rangle = 0$ and $\langle\eta(0)\eta(t)\rangle = \delta(t)$. This leads to a memory-less (Markovian) equation called Standard or Underdamped Langevin Equation

$$m\ddot{q} = -\frac{\partial A(q)}{\partial q} - \gamma m\dot{q} + \sqrt{2k_B T m\gamma}\eta(t) \tag{2.2}$$

- **Overdamped Langevin Equation**

  If one were to consider the dynamics for underdamped motion on a time scale $\tau >> \frac{m}{\gamma}$ non equilibrium fluctuations are quickly damped, this entitles us to improperly consider $\ddot{q} \approx 0$ leading to the Overdamped Langevin Equation

$$0 \approx -\frac{\partial A(q)}{\partial q} - \gamma m \dot{q} + \sqrt{2 k_B T m \gamma}\eta(t)$$
$$\gamma m \dot{q} = -\frac{\partial A(q)}{\partial q} + \sqrt{2 k_B T m \gamma}\eta(t)$$

  Which if rearranged defining the diffusion coefficient $D = \frac{k_B T}{m\gamma}$ and reduced temperature $\beta = \frac{1}{k_B T}$ as

$$\dot{q} = -\beta D(q)\frac{\partial A(q)}{\partial q} + \sqrt{2D(q)}\eta(t) \tag{2.3}$$

  sensibly simplifies the mathematical structure being a first order differential equation.
  The idea of the following work is to develop a method to infer the correct free energy and diffusion profile of the reduced Langevin model starting from Molecular dynamics trajectories, focusing on the Overdamped case.

2.2 The Bayesian criterion

In order to construct the optimal Overdamped Langevin model to describe the dynamics in this lower-dimensional projection, we need to optimize the parameters of said model. We do so through a Bayesian approach.

2.2.1 Bayes' Theorem

Given a set of data $\vec{x} = \{x_i\}$ generated from a probability distribution of unknown parameters, the probability of $\vec{\theta} = \{\theta_i\}$ being the actual parameters of the distribution from which $\vec{x}$ is sampled is $p(\vec{\theta}|\vec{x})$ also known as *Posterior distribution*.
This probability is computed according to Bayes' Theorem as:

$$\underbrace{p(\vec{\theta}\,|\,\vec{x})}_{\text{Posterior}} = \frac{p(\vec{x}\,|\,\vec{\theta})p(\vec{\theta})}{p(\vec{x})} \propto \underbrace{p(\vec{x}\,|\,\vec{\theta})}_{\text{Likelihood}} \overbrace{p(\vec{\theta})}^{\text{Prior}}$$

The *Prior* encodes preliminary information about the parameters θ while the *Likelihood* gives the probability of observing the data $\vec{x}$ if the were to be generated by a distribution of parameters $\vec{\theta}$.
As it is often the case in bayesian estimation the goal of this approach is to recover the values of the parameters more apt to describe the observed data, in other words those who maximize the posterior probability.
Such quantity in called *Maximum A Posteriori (MAP)* estimate, consisting in :

$$\vec{\theta}_{\text{MAP}} = \underset{\vec{\theta}}{\text{argmax}}\; p(\vec{\theta}\,|\,\vec{x})$$

2.2.2 Maximum Likelihood Inference

In the case under analysis the data are a set of molecular dynamics trajectories $\vec{q}$ while $\vec{\theta}$ stands for a compact way of regrouping both the force field $F(q) = -\beta D(q)\frac{\partial A(q)}{\partial q}$ and position dependent diffusion $D(q)$ present in equation (2.3).
Under the assumption of the so called *"Uninformative Prior"* the latter is considered as constant reducing the problem of finding $\vec{\theta}_{MAP}$, thus maximizing the posterior distribution of parameters, to that of finding $\vec{\theta}_{MLE}$ the value of $\vec{\theta}$ maximizing $p(\vec{x}\,|\,\vec{\theta}) \equiv \mathcal{L}_{\vec{q}}(\vec{\theta})$, the Likelihood of the observed trajectory $\vec{q}$.

$$\theta_{MLE} = \underset{\vec{\theta}}{argmax}\ \mathcal{L}_{\vec{q}}(\vec{\theta})$$

The analytical shape of the likelihood function is not always known a priory but, restricting to the overdamped case, a convenient property comes into play.
In fact being the equation Markovian means that, if the simulation is in position q_i at the current timestep t_i, the probability of it hopping to position q_{i+1} at timestep t_{i+1} depends only the present position q_i!
Such probability $p_\theta(q_{i+1}, t_{i+1}|q_i, t_i)$ is often called *propagator*.
For given initial conditions (q_0, t_0), the Likelihood of the dataset $\vec{q} = (q_0, q_1, ..., q_N)$ is nothing but the product of the short time transition probability between consecutive trajectory points [9], namely:

$$\mathcal{L}_{\vec{q}}(\vec{\theta}) = \prod_{i=0}^{N-1} p_\theta(q_{i+1}, t_{i+1}|q_i, t_i) \tag{2.4}$$

from which follows that the Log-Likelihood of the trajectory $\mathfrak{L}_{\vec{q}}(\vec{\theta})$ is :

$$\mathfrak{L}_{\vec{q}}(\vec{\theta}) = \log\{\mathcal{L}_{\vec{q}}(\vec{\theta})\} = \sum_{i=0}^{N-1} \log\Big\{p_\theta(q_{i+1}, t_{i+1}|q_i, t_i)\Big\} \tag{2.5}$$

2.3 Addition of biasing forces

In order to go forth with the simulation in case the trajectory were to be stuck in a local free energy minimum, some additional biasing forces are employed to push the simulation onwards and allowing it to sample larger portions of the phase space.

2.3.1 Adiabatic Bias Molecular Dynamics

ABMD [6] is a simulation method in which an harmonic biasing potential is used to drive a system from an initial to a final position along a selected coordinate q employing an quadratic biasing potential centered in position q_{max} and thus a linear force:

$$V_{ABMD}(q) = \frac{1}{2}k(q - q_{max})^2$$
$$f_{ABMD}(q) = k(q_{max} - q)$$

The center of this potential, q_{max} is updated at each step before computing the biasing potential to make sure that $q_{max} \geqslant q$, ensuring a positive biasing fore along the q direction at every step.
Whenever the system moves further along the coordinate, the center of the harmonic potential follows this new position, effectively pushing the system towards the final position specified at the beginning. If instead it moves backwards it encounters the harmonic bias preventing it from moving further back, as depicted in Fig(2.1)

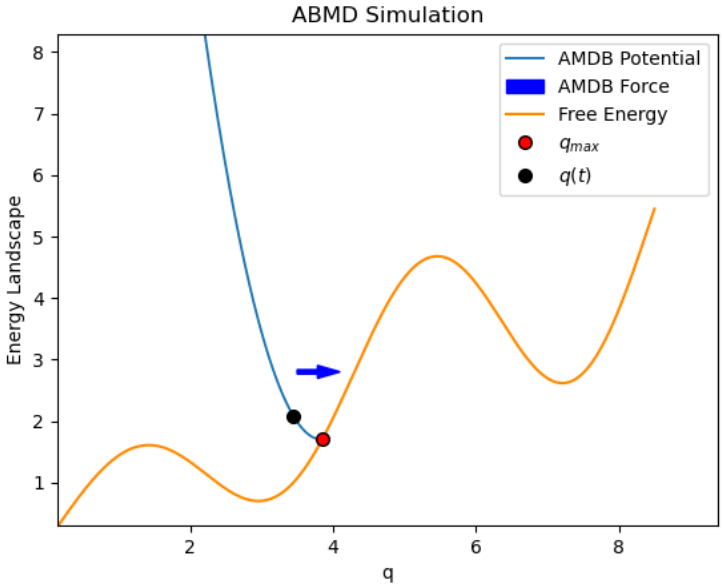


Figure 2.1: Example of particle moving in a 1D free energy landscape in the presence of ABMD bias pushing it to towards increasing values of $q(t)$

This method of course drives the simulation towards a user's specified direction which is somewhat "unnatural" therefore proper care needs to be employed in the computation of the likelihood function in order to correctly compensate for this phenomenon (Sec. 3.3).

Chapter 3
Simulation and Analysis Methods

3.1 The F.O.L.I.E. Module

In order to validate this inferential method some benchmark analysis of toy models has been produced as well as the underlying simulation of the trajectories themselves.
To carry out this task we co-developed further the F.O.L.I.E. project, a python module aimed at performing inference analysis of molecular dynamics trajectories, the relative code is available on the Folie GitHub Repository.[1]

The main features present in the package are:

- **Model of Overdamped Langevin Dynamics**:
  Several models of possible implementation of Overdamped Langevin equations are present starting from the parent python class `Overdamped`, followed by the implementation of particular cases such as `BrownianMotion` and `OrnsteinUhlenbeck`, the underdamped case is still under development.

- **Transition densities**
  The different methods aimed at approximating the form of the propagator described in sec.(3.2) are implemented. They require as input a Model object whose force and diffusion will be used to compute the mean, variance, and, in case of the Elerian transition probability density, the additional parameters to obtain the relative likelihood.
  These probability densities are later fed to the Likelihood estimator object who will recover the adequate drift and diffusion parameters.

- **Estimation**
  Given as input the probability densities used to compute the likelihood of the observed trajectories, an estimator object is created.
  The class of estimator objects playing a central role in performing the addressed task is the `LikelihoodEstimator` class, within it, the MLE estimators are recovered by making use of the `optimize.minimize()` method from the `scipy` library applied to $-\mathfrak{L}(\vec{q}|\theta)$ the negative of the log-likelihood function eq.(2.5).

- **Functions**
  The `Function` class has been implemented in order to specify for instance the

analytical form of the force and diffusion function of a given `Overdamped` object, or more generally to specify parameters of the different class objects defined above. Such `Function` objects are in fact often needed to build the appropriate transition densities and likelihood estimators.

- **Simulation**
  In addition to the principal purpose of training the maximum likelihood estimator for a given set of input trajectories, the F.O.L.I.E. module also allows to simulate trajectories in the first place.
  To do that one first needs to specify the model of the Langevin equation guiding the evolution of the system through a suitable `Overdamped` object. Then this passed to the `Simulator` (or possibly `BiasedSimulator`) class specifying the integration timestep employed. Eventually, the dataset is generated via the method `Simulator.run()`.
  The latter tool is the one employed to generate the trajectories analyzed in Chap(4).

3.2 Models of statistical inference

Now to go further with the estimation of the free energy profile it is useful to have at one's disposal a suited approximation for the propagator $p_\theta(q_{i+1}, t_{i+1}|q_i, t_i)$ so to compute explicitly Eq. (2.5)
The starting point to achieve this is using Itô Stochastic Differential Equation (SDE) formalism, where the infinitesimal increment of the position at time t namely $dq(t)$ is computed as:

$$dq_t = F(q_t)dt + \sigma(q_t, t)dW_t \tag{3.1}$$

to rewrite Eq.(2.3)

$$dq_t = -\beta D(q_t)\frac{\partial A(q_t)}{\partial q}dt + \sqrt{2D(q_t)}dW_t \tag{3.2}$$

where q_t and dW_t refer to $q(t)$ and $dW(t)$ respectively, the latter being the increment of the Wiener process, a random variable normally distributed with zero mean and variance equal to dt.

By comparing the two it is evident that $F(q_t) = -\beta D(q_t)\frac{\partial A(q_t)}{\partial q}$ and $\sigma(q_t) = \sqrt{2D(q_t)}$ are respectively the deterministic and stochastic contributions to the evolution of process q_t.
The advance respect to the literature methods considered in this work consisted in applying to the system some external biasing force $f^{bias}(q_t)$, influencing the drift term of the evolution equation according to

$$\widetilde{F}(q_t) = -\beta D(q_t)\frac{\partial A(q_t)}{\partial q} + f^{bias}(q_t, t) \tag{3.3}$$

In the following, four different methods to compute the likelihood of the transition are presented.
To help them converge faster each one of them is fed as initial condition a first guess on the mean and variance of the transition probability.
These two are estimated by performing a spline regression of the data at each step and

empirically calculating the deviation from said regression curve.
This procedure is equivalent to an empirical estimation of the first two coefficients in the Kramers-Moyal [11] expansion of the master equation governing the evolution of the equilibrium distribution.

In the following we illustrate the four methods considered to approximate the likelihood of transition (propagator) of the generic SDE Eq.(3.1), keeping in mind that for what concerns our analysis:

$$F(q_t) = -\beta D(q_t)\frac{\partial A(q_t)}{\partial q}$$
$$\sigma(q_t) = \sqrt{2D(q_t)}$$

Moreover it is assumed that the analyzed trajectories evolve at constant temperature, therefore we set $\beta = 1$ for convenience.

3.2.1 Euler Method

The Euler Method [5] consists in the approximation of Eq. (3.1) with the succession $\{q_i\}$ satisfying the iterative scheme :

$$q_{i+1} = q_i + F(q_i)\Delta t + \sigma(q_i)\sqrt{\Delta t}G \tag{3.4}$$

where $\Delta t = t_{i+1} - t_i$ is the time increment and ΔW_t has been replaced by $\sqrt{\Delta t}G$ being $G \sim \mathcal{N}(0,1)$ a standard normally distributed random variable.
From Eq.(3.4) it is clear that this increment is purely deterministic except for G.
This implies that the propagator at time t_i is a Gaussian random variable with mean ϕ_i

$$\phi_i = F(q_i)\Delta t = -\left[D(q_i)\frac{\partial A(q_i)}{\partial q} + f_{bias}(q_i)\right]\Delta t$$

and variance μ_i

$$\mu_i = \sigma^2(q_i)\Delta t = 2D(q_i)\Delta t$$

namely

$$p_\theta(q_{i+1}, t_{i+1}|q_i, t_i) = \frac{1}{\sqrt{2\pi\mu_i}}e^{-\frac{(q_{i+1}-q_i-\phi_i)^2}{2\mu_i}}$$

Thanks to this result we can now compute the log-likelihood (2.5) of the trajectory as :

$$\mathfrak{L}(\vec{q}|\theta) = \sum_{i=0}^{N-1}\frac{1}{2}\log(2\pi\mu_i) + \sum_{i=0}^{N-1}\frac{(q_{i+1}-q_i-\phi_i)^2}{2\mu_i}$$

3.2.2 Elerian method

The Elerian method performs the same passages as the Euler one but takes as starting point the Milstein discretization of the SDE [9].

$$q_{i+1} = q_i + \left[F(q_i) + \frac{1}{2}\sigma(q_i)\sigma'(q_i)\right]\Delta t + \sigma(q_i)\sqrt{\Delta t}G + \frac{1}{2}\sigma(q_i)\sigma'(q_i)G^2\Delta t \tag{3.5}$$

The latter consists in including a second-order term derived from applying the Itô formula to the Itô-Taylor expansion [2] of the stochastic differential equation; here $\sigma'(q_i)$ stands for $\frac{d\sigma}{dq}\Big|_{q_i}$.
Naturally Eq(3.5) falls into Eq.(3.4) for constant diffusion function.

The associated transition density for this model [5] is

$$p_\theta(q_{i+1}, t_{i+1}|q_i, t_i) = \frac{z^{-1/2}\cosh(\sqrt{Cz})}{|K|\sqrt{2\pi}} e^{-\frac{C+z}{2}}$$

where

$$K = \frac{D'(q_i)}{2}\Delta t \qquad B = -\frac{2D(q_i)}{D'(q_i)} + q_i + F(q_i)\Delta t - K$$

$$z = \frac{q_{i+1} - B}{K} \qquad C = \frac{2D}{D'(q_i)^2\Delta t}$$

3.2.3 Kessler method

Instead of focusing on approximations of the method of integration of SDE who produced the trajectories, as done in the previous two methods, Kessler proposed to approximate directly the mean and variance of the transition density through a higher-order Ito-Taylor expansion [2].
The result is a gaussian probability measure with mean ϕ_i and variance μ_i [5] :

$$\mu_i = q_i + F(q_i)\Delta t + \left(F(q_i)F'(q_i) + \frac{1}{2}\sigma^2(q_i)F''(q_i)\right)\frac{\Delta t^2}{2} \tag{3.6}$$

$$\phi_i = q_i^2 + \left(2F(q_i)q_i + \sigma^2(q_i)\right)\Delta t + \Big\{2F(q_i)\left(F'(q_i)q_i + F(q_i) + \sigma(q_i)\sigma'(q_i)\right) + \\ + \sigma^2(q_i)\left(F''(q_i)q_i + 2F(q_i) + \sigma^2(q_i) + \sigma(q_i)\sigma''(q_i)\right)\Big\}\frac{\Delta t^2}{2} - \phi_i^2$$

3.2.4 Drozdov method

The time evolution of the ensemble density $\rho(q,t)$ associated to our Langevin model is dictated by an appropriate Fokker-Planck equation

$$\frac{\partial\rho(q,t)}{\partial t} = \frac{\partial}{\partial q}\left(De^{-A(q)}\frac{\partial}{\partial q}e^{A(q)}\rho(q,t)\right) = \mathscr{G}^\dagger\rho(q,t) \tag{3.7}$$

where $\mathscr{G}^\dagger$ is the Fokker-Plank operator [11] defined in Eq.(3.7).
The short time propagator $p_\theta(q_{i+1}, t_i + \Delta t|q_i, t_i)$ takes the form [8]

$$p_\theta(q_{i+1}, t_i + \Delta t|q_i, t_i) = e^{\mathscr{G}^\dagger\Delta t}\delta(q_{i+1} - q_i) = [1 + \mathscr{G}^\dagger\Delta t + \frac{1}{2}(\mathscr{G}^\dagger)^2\Delta t^2 + \dots]\delta(q_{i+1} - q_i) \tag{3.8}$$

Drozdov considered a second order expansion of (3.8) and retrieved an expression for the propagator making use of generating functions [4].
The result is again a Gaussian with mean ϕ_i and variance μ_i given by the following expressions:

$$\phi_i = F(q_i)\Delta t + \frac{1}{2}(F(q_i)F'(q_i) + D(q_i)F''(q_i))\Delta t^2$$

$$\mu_i = \sigma(q_i)\Delta t + \left\{F(q_i)\sigma(q_i)\sigma'(q_i) + \sigma^2(q_i)\left[F'(q_i) + \frac{1}{2}\left(\sigma(q_i)\sigma''(q_i) + (\sigma'(q_i))^2\right)\right]\right\}\Delta t^2$$

3.3 Reconstructing the Free Energy Landscape

At this stage we find ourselves with four possible approximations for the shape of the Likelihood function. Thus the maximum likelihood estimators for the parameters of the Overdamped Langevin equation, namely $F_{MLE}(q)$ and $D_{MLE}(q) = \frac{1}{2}\sigma^2_{MLE}(q)$ are retrieved. To do so we make use of the `scipy.optimize.minimize()` method, providing also :

$$\nabla_{\vec{\theta}}\left(-\mathfrak{L}_{\vec{q}}(F, D)\right) = \begin{pmatrix} -\dfrac{\partial \mathfrak{L}_{\vec{q}}(F, D)}{\partial F} \\ -\dfrac{\partial \mathfrak{L}_{\vec{q}}(F, D)}{\partial D} \end{pmatrix}$$

the gradient of the negative Log-likelihood with respect to $F(q)$ and $D(q)$.
It is important to remark that this procedure is run multiple times,one for each approximation considered (Euler, Elerian, Kessler, Drozdov).
Equipped now with said estimators $F_{MLE}(q)$ and $D_{MLE}(q)$, one can in principle generate new trajectories with the MLE effective Overdamped Langevin model

$$\dot{q} = -F_{MLE}(q) + \sqrt{2D_{MLE}(q)}\eta(t)$$

simply by integrating the corresponding stochastic differential equation which, in Itô form, reads:

$$dq_t = F_{MLE}(q)dt + \sqrt{2D_{MLE}(q)}dW_t \tag{3.9}$$

This piece of information is used at the same time to infer the Free Energy Landscape underlying the dynamics. In particular we focus on the Milstein Integrator Eq.(3.5), being more accurate in case of position-dependent diffusion.
Then we consider the average displacement between the position at timestep i and at the subsequent timestep $i+1$ over the realization of the noise.

$$\langle \Delta q_i \rangle = \langle q_{i+1} - q_i \rangle$$

Writing equation (3.5) explicitly in terms of $A(q), D(q)$, and its derivative $D'(q) = \frac{dD}{dq}$, the drift $F(q)$ and diffusion $\sigma(q)$ read:

$$F(q_i) = -D(q_i)\frac{\partial A(q_i)}{\partial q} + \frac{1}{2}D'(q_i)$$
$$\sigma(q) = \sqrt{2D(q)}$$

giving for $\langle q_i \rangle$ the following expression:

$$\langle \Delta q_i \rangle = \left[-D(q_i) \frac{\partial A(q_i)}{\partial q} + \frac{1}{2} D'(q_i) \right] \Delta t + \sigma(q_i) \underbrace{\langle \Delta W_i \rangle}_{=0} + \frac{1}{2} D'(q_i) \underbrace{\langle \Delta W_i^2 \rangle}_{=\Delta t} =$$

$$\langle \Delta q_i \rangle = \left[-D(q_i) \frac{\partial A(q_i)}{\partial q} + D'(q_i) \right] \Delta t \tag{3.10}$$

Performing the same computation to the (SDE) with parameters of maximum likelihood (3.9) one obtains

$$\langle dq_t \rangle = F_{MLE}(q) dt + \sqrt{2 D_{MLE}(q)} \underbrace{\langle dW_t \rangle}_{=0}$$

$$\langle dq_t \rangle = F_{MLE}(q) dt \tag{3.11}$$

Therefore approximating $\langle dq_t \rangle \approx \langle \Delta q_i \rangle$ and $dt \approx \Delta t$ in the two equations above enables us to recover an the estimation for the free energy profile.
In fact,

$$\langle dq_t \rangle \approx \langle \Delta q_i \rangle \qquad \Longrightarrow \qquad F_{MLE}(q) dt \approx \left[-D(q_i) \frac{\partial A(q_i)}{\partial q} + \frac{1}{2} D'(q_i) \right] \Delta t$$

$$\frac{\partial A_{MLE}(q)}{\partial q} = - \frac{F_{MLE}(q) - D'_{MLE}(q)}{D_{MLE}(q)}$$

which, when integrated from an initial value q_0, leads to

$$A_{MLE}(q) = - \int_{q_0}^{q} \frac{F_{MLE}(q^\star) - D'_{MLE}(q^\star)}{D_{MLE}(q^\star)} \, dq^\star \tag{3.12}$$

up to an additive constant.

3.3.1 Reconstructing the Free Energy Landscape from biased trajectories

A further step in this analysis, proposed for the first time, has been to include the effect of an additional bias in the drift term as specified in Eq.(3.3).
Doing so implies that at each step in computing the likelihood, instead of computing the MLE for $F(q_i)$, it computes the one relative to:

$$\widetilde{F}_{MLE}(q_i) = F_{MLE}(q_i) + D_{MLE}(q_i) f_i^{bias}(q_i)$$

One key observation about it is to underline the fact that it is necessary to correct for this term at each instant. In fact for different timesteps to which it might correspond the same position value $q_i = q_j$ for $i \neq j$ the value of biasing force is not necessary the same $f_i^{bias}(q_i) \neq f_j^{bias}(q_j)$!
It is straightforward to recover the appropriate estimator for the unbiased drift by inverting this equation: $F_{MLE}(q_i) = \widetilde{F}_{MLE}(q_i) - D(q_i) f_i^{bias}(q_i)$.
Plugging this back to Eq.(3.12) returns once again the Maximum Likelihood Estimator for the free energy landscape.

Chapter 4
Toy models

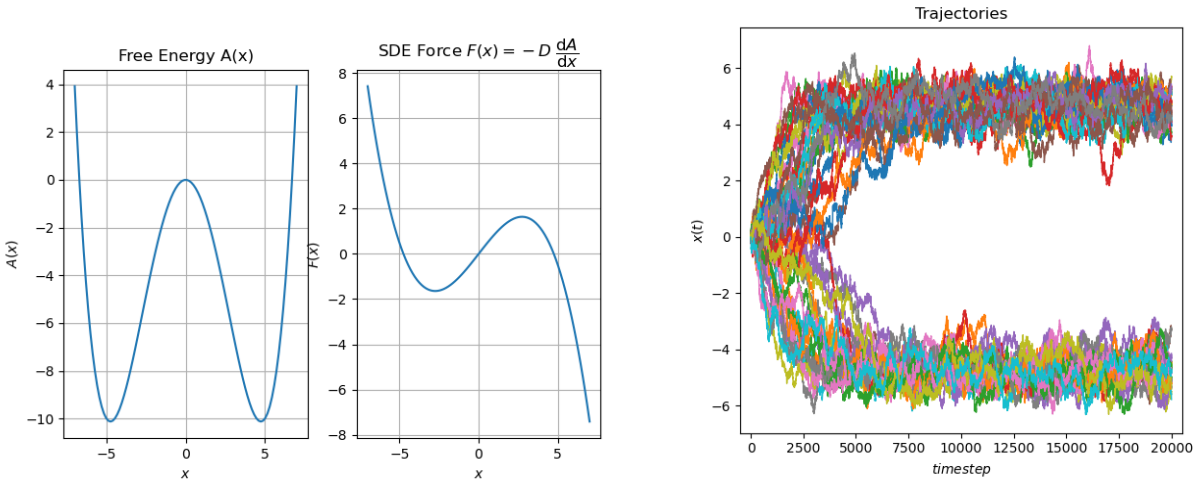
In the current section the numerical simulations and estimation of relative free energy profile is carried out so to asses the quality of the methodology employed.
In particular a couple of benchmark validation *"Toy-model"* systems cosisting of a 1D and 2D double well Free Energy profiles have been used.

4.1 Unbiased 1D Double Well

The first physically significant system considered was the one dimensional symmetric double well potential along the x axis, where the free energy is obtained using a 4^{th} order polynomial.

$$A(x) = \gamma_0 x^4 + \gamma_1 x^2 \tag{4.1}$$

with $\gamma_0 > 0$ and $\gamma_1 < 0$.
In this section the chosen coordinate for the training is the only one present $q = x$.
To run this simulation we initialized 50 copies of the system in $x = 0$ at $t = 0$ and, later let evolve with the Euler integrator Eq.(3.4) for 20 000 steps.
The chosen diffusion constant is $D = \frac{1}{2}$ and no additional external bias is applied. The result is an ensemble of trajectories moving in this double well free energy profile leading to approximately half of them falling into the right minimum and the other half into the left one, located at $x = \pm\sqrt{-\frac{\gamma_1}{2\gamma_0}}$; with our choice of parameters, $x = \pm\sqrt{\frac{45}{2}}$, as shown in Fig.(4.1).
This result does not come as a surprise due to the symmetry of the model considered.



(a) Free Energy and correspondent SDE Force of the simulation (b) Time evolution of the position of all the 50 trajectories

Figure 4.1: Numerical simulation of 50 trajectories evolving according to Eq.(3.2) in force field represented in Fig.(4.1a) for 20 000 timesteps,each of duration $\Delta t = 10^{-3}$

4.1.1 Estimation of Free energy and diffusion constants

Once having generated enough trajectories in the previous section it is now time to run an estimation following the maximum likelihood approach to test the accuracy of the method employing all the estimators described in Chapter(3.2).
The estimated force and diffusion fitted through the use of piece-wise polynomial (splines) function are plotted in Fig.(4.2).
This class of functions has been chosen to achieve more generality and, as it is possible to see from Fig.(4.2), they well fit both the polynomial behaviour of the force and the constant nature of the diffusion function.

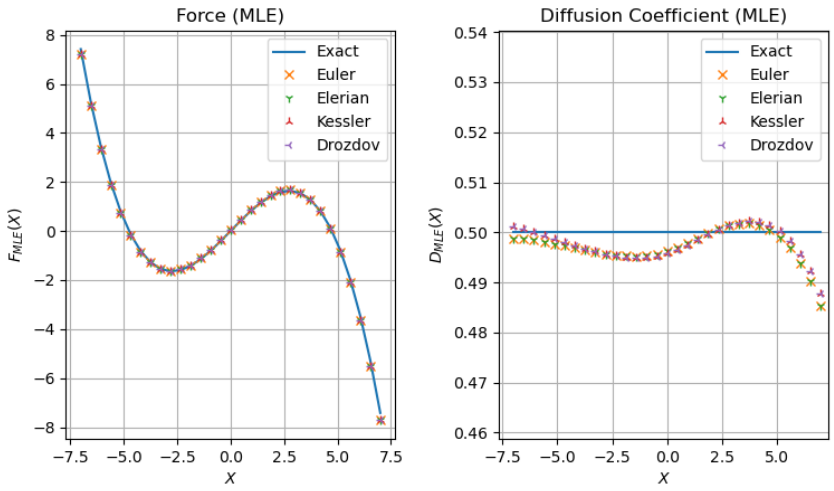


Figure 4.2: Estimation of the maximum likelihood force and diffusion functions according to the estimators described in Ch(3.2)

Having at our disposal the information regarding this two important parameters we can then recover the underlying free energy profile numerically integrating Eq.(3.12).

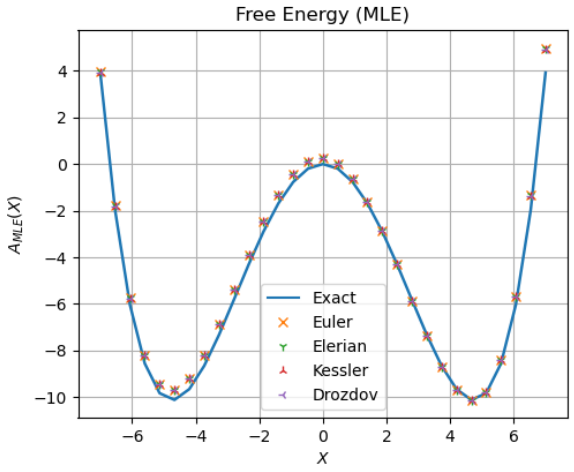


Figure 4.3: Caption

This benchmark system, although quite trivial, sets a first step in the validation of the maximum likelihood method giving good reason to hope that it will remain well-behaved also in case of the application of an external bias.

4.2 Biased 1D Double Well

The same analysis carried out for the one dimensional double well is now performed for the same system but under the application of an external Adiabatic Bias described in section (2.3.1) along the coordinate $q \equiv x$.

$$f^{bias}(q) = k(q - q_{MAX}) \tag{4.2}$$

This time the ensemble of trajectories has been initialised in the left of one of the two energy minima, precisely at $x = -6$, and it is later pushed towards increasing value x by the external bias.

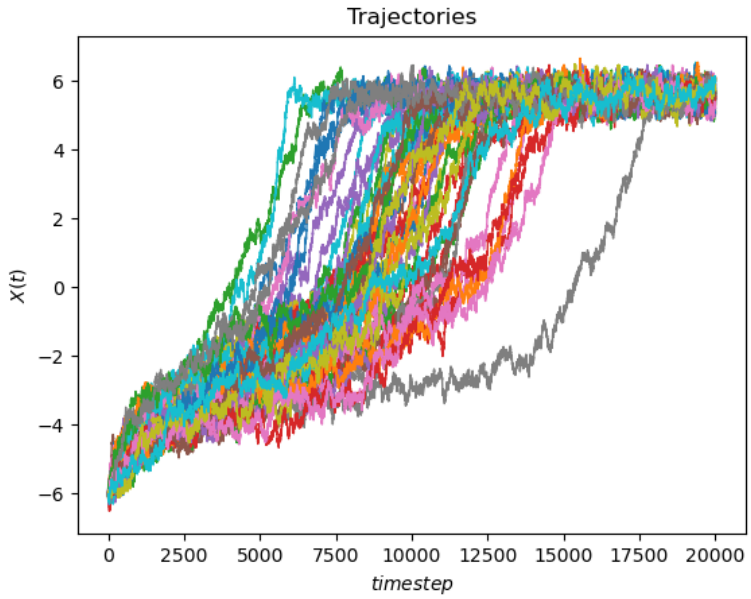


Figure 4.4: Numerical simulation of 50 trajectories evolving according to Eq.(3.2) in force field represented in Fig.(4.1a) for 20 000 timesteps,each of duration $\Delta t = 10^{-3}$ under the application of the Adiabatic bias with $k = 10$

It is noticeable from the picture above that every trajectory underwent a transition from one energy minimum to the other in a quite sharp manner by the time the simulation ended.
Such transition, given the initialisation point, would have required an enormous amount of time to happen ergodically! Therefore that is why it is so important to have at one's disposal a way to estimate the free energy profile even from biased numerical simulations.

4.2.1 Estimation of Free energy and diffusion constants

The same estimation carried out for the unbiased case is now performed for the biased data accounting for the bias which was known at each step, according to Eq.(3.3)

Again the estimated force and diffusion are fitted with splines function and plotted in the figure below

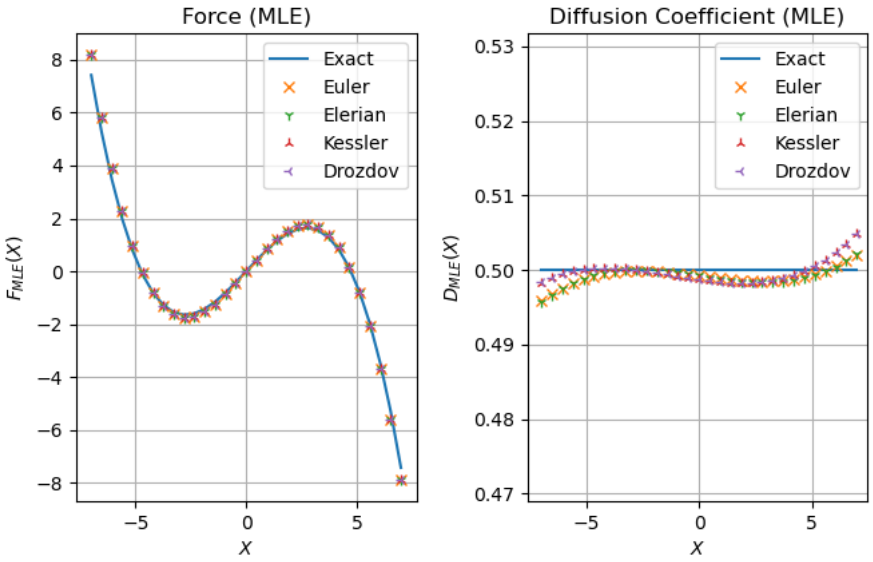


Figure 4.5: Estimation of the maximum likelihood force and diffusion functions according to the estimators described in Ch(3.2) accounting for the presence of the external bias

From which the underlying free energy profile is recovered.

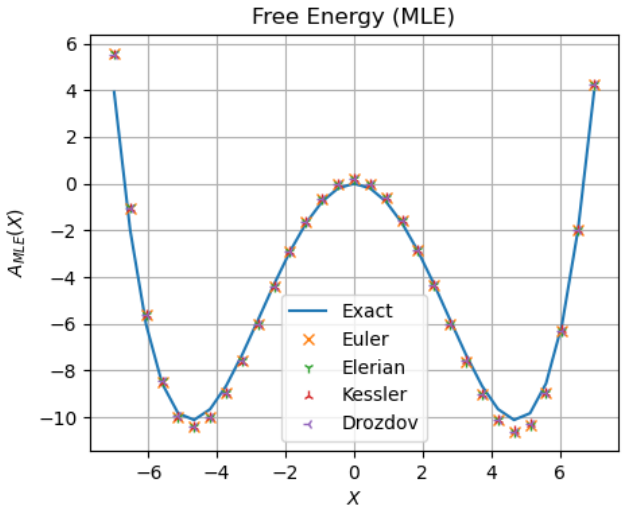
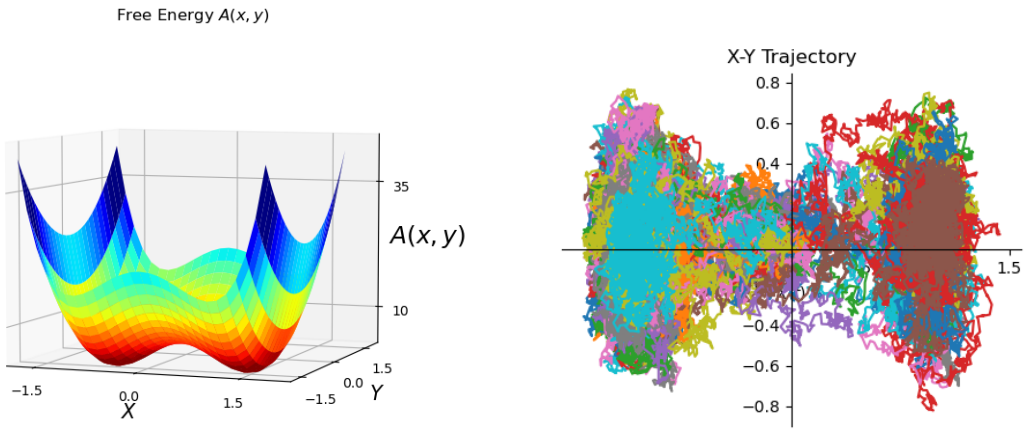


Figure 4.6: Caption

The recovered free energy well describes the original profile also in this biased situation, showing how efficiently the proposed method performs in this controlled environment.

4.3 2D Double well

Moving to a scenario closer to a realistic employment of the method we examine the dynamic of a bi-dimensional system evolving in a free energy landscape modelled by a 2D double well potential and subsequently project the dynamics to a one dimensional collective variable in an attempt to fit the reduced dynamics with a one-dimensional overdamped Langevin equation.
In addition to this we added a biasing force along said coordinate and follow the maximum likelihood principle to recover the projected free energy landscape.



(a) 2D double well potential

(b) 2D dynamics

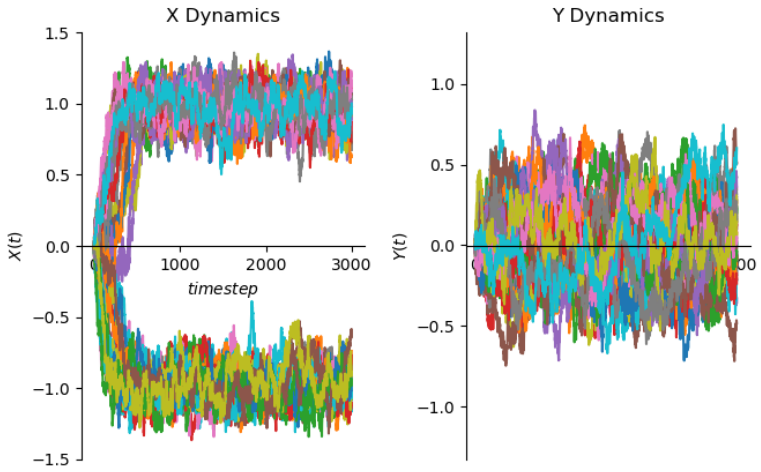
Figure 4.7: Numerical simulation of 50 trajectories evolving in the free energy landscape represented in Fig.(4.7a) for 3000 timesteps, each of duration $\Delta t = 5 \times 10^{-4}$. Each one of the trajectories has been initialized in the origin $(x, y) = (0, 0)$

The analytical shape of the potential used is decoupled in the two variables x and y :

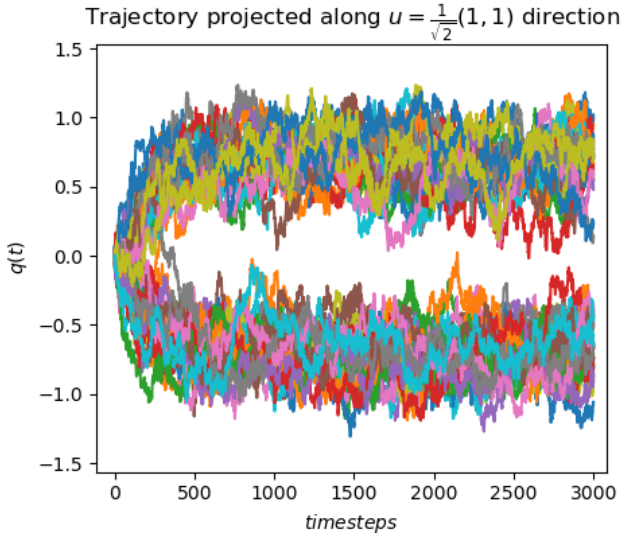
$$A(x,y) = \nu(x) + \mu(y) \qquad where \qquad \begin{cases} \nu(x) = a(x^2-1)^2 \\ \mu(y) = \frac{1}{2}by^2 \end{cases}$$

therefore the two Cartesian components evolve following two independent overdamped Langevin equations in a free energy profile dictated by $\nu(x)$ and $\mu(y)$ respectively.

Instead of focusing along these two directions whose behavior was known, we decided to project the given trajectories $\vec{X} = (x, y)$ along a collective variable q whose direction $(1, 1)$ is chosen to be the bisectrix of the 1^{st} and 3^{rd} quadrant i.e. $q = \vec{X} \cdot \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ equivalent to a rotation the reference frame by an angle $\theta = \pi/4$.



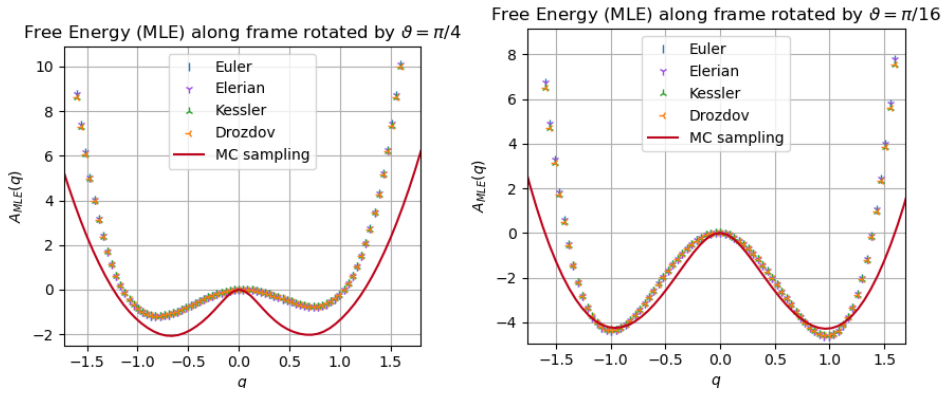
(a) Cartesian components of 2d the trajectories



(b) Projection of the 2D trajectories along the bisectrix of 1^{st} and 3^{rd} quadrant

Figure 4.8

Running an estimation along this collective variable leads to a result that seems to underestimate the depth of the free energy minima.



(a) Free energy along q tilted by $\theta = \pi/4$ (b) Free energy along q tilted by $\theta = \pi/16$

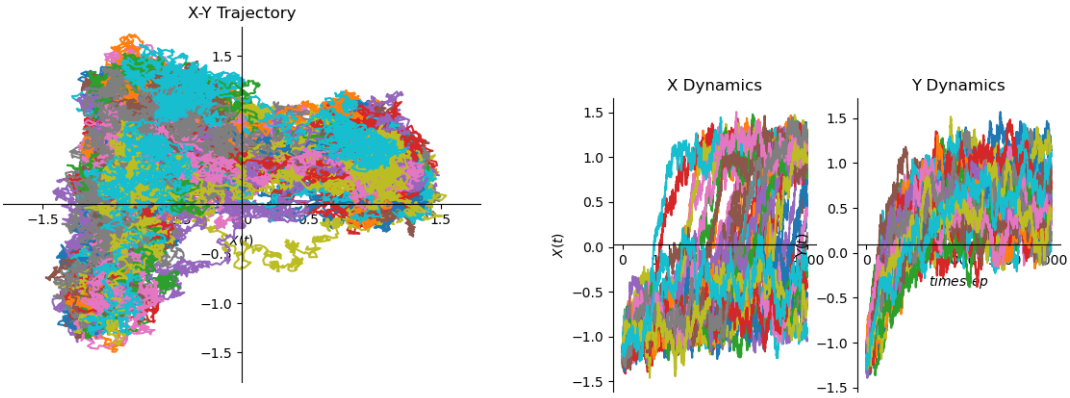
Figure 4.9: Estimated free energy landscape along $q = (x\cos(\pi/4), y\sin(\pi/4))$, and $q = (x\cos(\pi/16), y\sin(\pi/16))$ the reference red curve is obtained by Monte Carlo sampling uniformly in the interesting area and reweighting the samples along q with their Boltzmann's weight $\exp\left[-U(x,y)\big|_{q=(x\cos(\pi/4), y\sin(\pi/4))}\right]$

If instead we employ a rotated frame by a different angle the estimated free energy gets progressively more accurate when compared with the MC sampling the closer the rotated frame is to one of the two original axis $x : (\theta = 0)$ or $y : (\theta = \pi/2)$ case, where the estimation recovers the $\nu(x)$ and $\mu(y)$ functions respectively.

4.3.1 Application of bias

However while one runs a Molecular Dynamics simulation in a unknown Free Energy landscape with the help of biasing forces, often the chosen coordinate might not be the optimal one.
The remaining of this section is aimed at attempting to reconstruct the free energy of the system along the collective variable $q = x\cos(\theta) + y\sin(\theta)$ in presence of adiabatic bias on the same coordinate and thus run an estimation to fit the dynamics in this one-dimensional manifold.

In this case the trajectories are initialized in $(x_0, y_0) = (-1.2, -1.2)$ with angle $\theta = \frac{\pi}{4}$ i.e. $q_0 = -\frac{2.4}{\sqrt{2}}$, therefore the adiabatic bias pushed the simulation toward the 1^{st} quadrant.

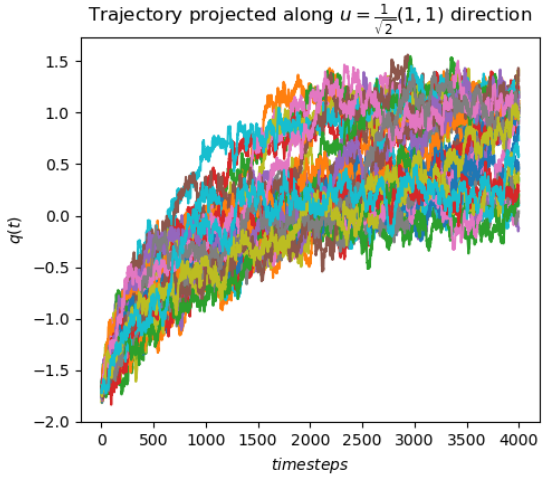


(a) 2D (biased) dynamics

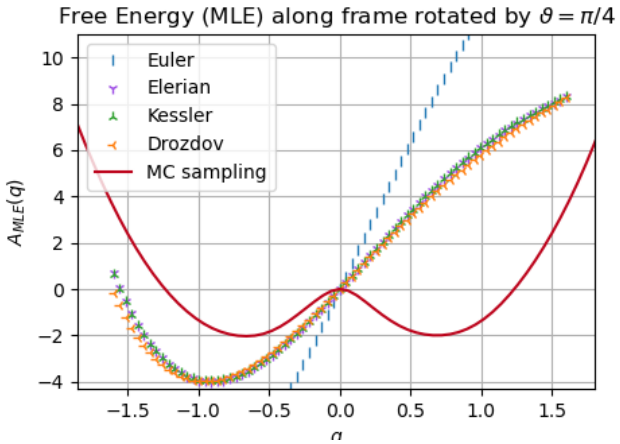
(b) Cartesian components of 2D the trajectories

Figure 4.10: Numerical simulation of 50 trajectories evolving in the free energy landscape represented in Fig.(4.7a) for 3000 timesteps,each of duration $\Delta t = 5 \times 10^{-4}$. Here the biasing constant is $k = 25$ and trajectories were initialized $(x_0, y_0) = (-1.2, -1.2)$

This simulation well shows how the systems first initialized in the left basin of attraction $x = -1$, under the effect of the bias, is forced to perform a transition to the right one in $x = +1$.
Training the overdamped model upon these data while keeping track of the biasing force that has been used, lead to the following free energy

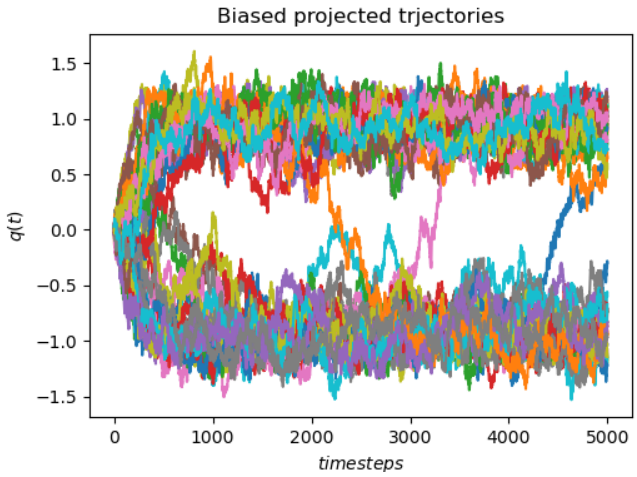


(a) Projection of the 2D trajectories along the coordinate q=(xcos(π/4),ysin(π/4))

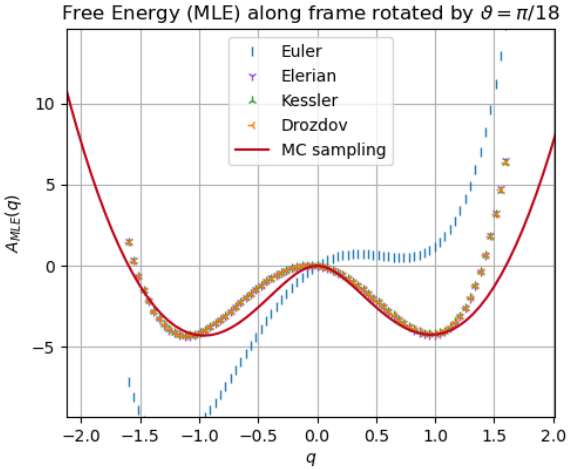


(b) Free energy profile relative to the projected trajectory

In this case the method does not seem to recover the appropriate free energy landscape, a far better job is done when considering a situation not too far from the optimal coordinate case $\theta = 0$.
For instance the same analysis using the not optimal $q = (\cos(\pi/18), \sin(\pi/18))$ exhibits a far better adherence to the Monte Carlo reference profile.



(c) Projection of the 2D trajectories along the coordinate q=(xcos(π/18),ysin(π/18))



(d) Free energy profile relative to the projected trajectory

4.3.2 Estimation of error

To provide a qualitative assessment of the error, we calculated the maximum likelihood estimation (MLE) for the free energy using four different datasets and computed the root mean square deviation (RMSD) of the four estimations for the free energy landscape.

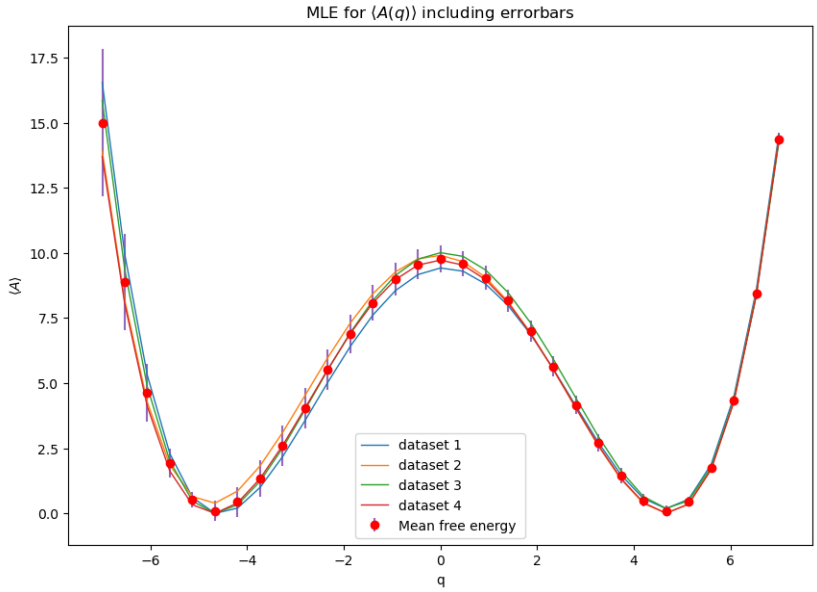


Figure 4.11: Estimation of the same free energy profile using four independent datasets, here plotted with errorbars = 2 × RMSD

Unfortunately during the period of this internship, there was not enough time to pursue a deeper analysis in this direction.

Chapter 5
Conclusions

The method proves itself to work satisfyingly well along good projection coordinates, recovering the underlying free energy profile even in the case of biased dynamics.
However, in the case of a poorly chosen collective variable, it showed to work approximately well in certain conditions.

For instance, in the case of the 2D Biased Double Well projected along the tilted coordinate, the result is not so far from the reference even though it moves further from it the more the rotation angle approaches the bisectrix of first and third quadrant.
The reason behind this can be attributed to the fact that when projecting, many points with different (x, y) values will fall to the same q (CV) value, thus when collecting samples from a given q, they might come all from a similar portion of configuration space leaving other important regions, belonging to the same collective variable value, unexplored.

In addition to this, another possible explanation can be traced back to the assumption about the dynamics itself. For instance it is possible that when projecting along a badly chosen coordinate, a more appropriate choice of the fitting model might be a *Non-Markovian* Langevin equation such as Eq.(2.1), opening for a far broader range of flexibility in modeling the experimental data.

References

[1] Langevinmodel/folie: Finding optimal langevin inferred equations, *GitHub*, https://github.com/langevinmodel/folie. (Cited on page 11)

[2] Orucova Buyukoz Gulsen Bayram Mustafa, Partal Tugcem. Numerical methods for simulation of stochastic differential equations. *Advances in Difference Equations*, 2018. (Cited on page 14)

[3] Alexander Berezhkovskii and Attila Szabo. One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. *The Journal of Chemical Physics*, 122(1):014503, 12 2004. (Cited on page 4)

[4] Alexander N. Drozdov. High-accuracy discrete path integral solutions for stochastic processes with noninvertible diffusion matrices. *Phys. Rev. E*, 55:2496–2508, Mar 1997. (Cited on page 15)

[5] Stefano M. Iacus. *Simulation and Inference for Stochastic Differential Equations: With R Examples (Springer Series in Statistics)*. Springer Publishing Company, Incorporated, 1 edition, 2008. (Cited on pages 13 and 14)

[6] Massimo Marchi and Pietro Ballone. Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems. *The Journal of Chemical Physics*, 110(8):3697–3702, 02 1999. (Cited on page 9)

[7] Roberto Meloni, Carlo Camilloni, and Guido Tiana. Properties of low-dimensional collective variables in the molecular dynamics of biopolymers. *Phys. Rev. E*, 94:052406, Nov 2016. (Cited on pages 4 and 6)

[8] Karen Palacio Rodriguez. *Development of predictive approaches for biomolecular association kinetics*. Theses, Sorbonne Université, September 2022. (Cited on pages 4, 7, and 14)

[9] Karen Palacio-Rodriguez and Fabio Pietrucci. Free energy landscapes, diffusion coefficients, and kinetic rates from transition paths. *Journal of Chemical Theory and Computation*, 18(8):4639âĂŞ4648, July 2022. (Cited on pages 4, 5, 9, and 13)

[10] Fabio Pietrucci. Strategies for the exploration of free energy landscapes: unity in diversity and challenges ahead. *Reviews in Physics*, 2017. (Cited on pages 4 and 5)

[11] Hannes Risken. *Fokker-Planck Equation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996. (Cited on pages 13 and 14)

[12] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977. (Cited on page 4)