Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Classical statistical mechanics, equilibrium and non-equilibrium

Effective equations for reaction coordinates in polymer transport

Marco Baldovin[1,4], **Fabio Cecconi**[2] **and Angelo Vulpiani**[1,3]

[1] Department of Physics, Università 'Sapienza', Roma Piazzale A. Moro 5, I-00185, Italy
[2] CNR-Istituto dei Sistemi Complessi, Via Taurini 19, I-00185 Roma, Italy
[3] Centro Interdisciplinare 'B. Segre', Accademia dei Lincei, Via della Lungara 10, I-00165 Roma, Italy
E-mail: marco.baldovin@roma1.infn.it

Received 5 July 2019
Accepted for publication 22 October 2019
Published 10 January 2020

Online at stacks.iop.org/JSTAT/2020/013208
https://doi.org/10.1088/1742-5468/ab5368

Abstract. In the framework of the problem of finding proper reaction coordinates (RCs) for complex systems and their effective evolution equations, we consider the case study of a polymer chain in an external double-well potential, experiencing thermally activated dynamics. Langevin effective equations describing the macroscopic dynamics of the system can be inferred from data by using a data-driven approach, once a suitable set of RCs is chosen.

We show that, in this case, the validity of such choice depends on the stiffness of the polymer's bonds: if they are sufficiently rigid, we can employ a reduced description based only on the coordinate of the center of mass; whereas, if the stiffness reduces, the one-variable dynamics is no more Markovian and (at least) a second reaction coordinate has to be taken into account to achieve a realistic dynamical description in terms of memoryless Langevin equations.

Keywords: biopolymers, coarse-graining, transport properties

[4] Author to whom any correspondence should be addressed.

1742-5468/20/013208+19$33.00

Contents

1. Introduction

The study of many interesting phenomena often faces severe difficulties due to the presence of a large amount of degrees of freedom and of very different time scales. As important examples, we can mention protein dynamics and climate physics: the time scale of vibrations of covalent bonds is $O(10^{-12})$ s, while the protein folding time may range from milliseconds to even hours for the largest and most complex polypeptides [1–3]. In the case of climate dynamics the characteristic times may range from days (for the atmosphere) to $O(10^4)$ years (for the deep ocean and ice shields) [4]. In such classes of systems, numerical simulations are certainly a very powerful and useful tool to investigate the dynamics, but the enormous amount of information contained in each single trajectory can be considered somehow redundant if one is interested in a description of the processes occurring on a given range of time and spatial scales. Therefore, the opportunity to use computational methods cannot be seen as a 'panacea' able to explain everything, because the presence of high-dimensional phase-space strongly limits the possibilities to identify and extract a simple representation of the relevant processes.

The proper approach in multi-scale systems is the introduction of suitable effective equations describing the *slow dynamics* in terms of 'slow observables', generally referred to as 'reaction coordinates' (RC). RCs are a proper class of observables which are able to characterize, in a reduced way, the progress of a reaction in terms of a sequence of chemical events (or states) [5]. This methodology is rather useful both at practical level and from a conceptual point of view: effective equations are able to catch certain general features and to reveal dominant behaviors which could remain hidden in the fully detailed description [5, 6]. Let us also note that the use of the RCs, which compress (project) the multi-dimensional dynamics on a strongly reduced phase-space,

J. Stat. Mech. (2020) 013208

produces a drastic loss of information; but this loss is compensated by an immediate and compact picture of the possible regimes or collective behaviors taking place during the system evolution.

The problem of finding effective equations for multi-scale phenomena has a long history in Science, in particular in Mathematics and Physics: as prototype examples we can mention the averaging method in mechanics [7] and the Langevin equations for colloidal particles [8].

In few lucky cases, the effective equations can be derived from first principles. A remarkable example is the approach dating back to Smoluchowski to obtain, using kinetic theory, the Langevin equation for a heavy particle in a dilute gas of light particles [9]. Another important attempt was suggested by several authors in the 60's [10–13] and later by Zwanzig [14], which amounts to rigorously deriving Langevin equations for an heavy particle interacting with a chain of light harmonic oscillators.

In the study of the dynamical behavior of complex systems with a multi-scale structure, the first (and perhaps most difficult) step in the derivation of the effective equations, either from first principles or from data analysis, is the identification of suitable RCs. This task is far from being trivial and remains conceptually challenging. We can remind the caveat by Onsager and Machlup in their seminal work on fluctuations and irreversible processes [15]: *'how do you know you have taken enough variables, for it to be Markovian?'*

There are several systematic methods to partially answer the caveat by Onsager and Machlup. The most widely used is *principal component analysis* (PCA) [16], which searches for independent linear combinations of available observables with maximal variance. *Dynamic mode decomposition* (DMD) [17], *variational approach of conformation dynamics* (VAC) [18], *time-lagged independent component analysis* (TICA) [19] are some of the many, related, techniques used to project the evolution of the coordinates describing the full system into a smaller set of relevant RCs [20]. This is usually done by considering linear combinations of the original variables and exploiting the methods of linear algebra. In recent years, neural networks and deep learning techniques have been applied to enhance such algorithms; specifically, they can select combination of nonlinear functions (from libraries of possible candidates) to encode original data into the reduced RC space [21, 22].

For sure artificial intelligence methods [23] are useful tools in this perspective. However they should not be viewed as automatic or unsupervised protocols [24–27]: we cannot disregard the physical intuition and the (partial) knowledge of the studied systems for selecting the correct RCs and avoiding 'bad' choices, which could neglect or hide relevant phenomenologies occurring in the dynamics.

Even in the lucky case in which the proper RCs are already identified, and we know that the relevant features of the system can be modeled by Markovian evolution equations for these RCs only, the problem of finding the form of such equations can be non-trivial. Some methods attempt to extrapolate from data the most suitable drift and diffusivity terms [27–29] for memoryless generalized Langevin equations. This strategy has been successfully used to describe the slow dynamics in several contexts, such as turbulence [30, 31], granular media [32] and polymer physics [33, 34].

In this work, we revisit the Kramers' problem for a polymer in a double well [35, 36] by using the evolution equations of proper RCs characterizing the slow dynamics

J. Stat. Mech. (2020) 013208

of the chain. We model the polymer as a one-dimensional harmonic chain (the so-called Rouse chain).

The most natural RCs for our polymer are its center of mass Q and its end-to-end distance L. We try to derive their evolution equations with a data-driven approach: starting from long time series of data, we apply a well-established procedure to infer the right functional forms of the Langevin terms appearing in the dynamics. We then analyze the validity of such description by comparing the behavior of the inferred model to that of the original system.

The conceptual issues related to the usage of such data-based protocol are discussed. We shall see that the effective Langevin equations for such RCs depend crucially on the stiffness of the polymer bonds: if the polymer is sufficiently rigid, only the evolution of Q can be taken into account, while in the regime of low stiffness, we need to enlarge the phase space by considering also another coordinate (L in our case) to achieve a satisfactory picture of the jump dynamics over the barrier. In other words, this second coordinate allows the Markov property of the Langevin equations to be preserved.

The outline of the paper is as follows: in section 2 we describe the model; section 3 reports the results about the reconstruction of the Langevin equations obtained by the numerical extrapolation procedure, in the case of one RC (Q) and in that of two RCs (both Q and L); finally section 4 contains our conclusions and remarks.

2. Model and simple remarks

Let us consider the problem of a polymer crossing the barrier of a double-well energy profile, which is related to the transport of biomolecules accross nano-scale pores [37–42]. In many practical situations channels are so narrow that the transport dynamics of biopolymers and ions occurs on a single axis, thus, as a matter of fact, it can be considered one-dimensional [43–46]. In this crude approximation, the polymer is composed by a chain of N beads (point particles), interacting via nearest-neighbors forces and subjected to a thermal noise at temperature T.

The nanopore is portrayed as a region of the translocation axis where the polymer feels the effect of an energy barrier, which acts independently on each particle and separates the left-side and right-side of the pore [41, 47, 48], see figure 1. As customary, in this kind of phenomenology we can assume the evolution of each monomer to be accessible on time-scales long enough to neglect the effect of inertia. Accordingly, the polymer monomers are governed by the overdamped Langevin dynamics:

$$\begin{aligned}\gamma\dot{x}_1 &= -V'(x_1) + U'(x_2 - x_1) + \xi_1\\ \gamma\dot{x}_i &= -V'(x_i) + U'(x_{i+1} - x_i) - U'(x_i - x_{i-1}) + \xi_i\\ \gamma\dot{x}_N &= -V'(x_N) - U'(x_N - x_{N-1}) + \xi_N\end{aligned} \tag{1}$$

with $i = 2, \ldots, N-1$, where x_j is the position of the jth bead. V represents here the external potential, due to the nanopore action on the chain. U is the nearest-neighbor interparticle potential that is chosen to be an even and convex function of $x_i - x_{i-1} - \sigma$, where σ is the equilibrium distance between consecutive particles. Each ξ_i is a Gaussian

J. Stat. Mech. (2020) 013208

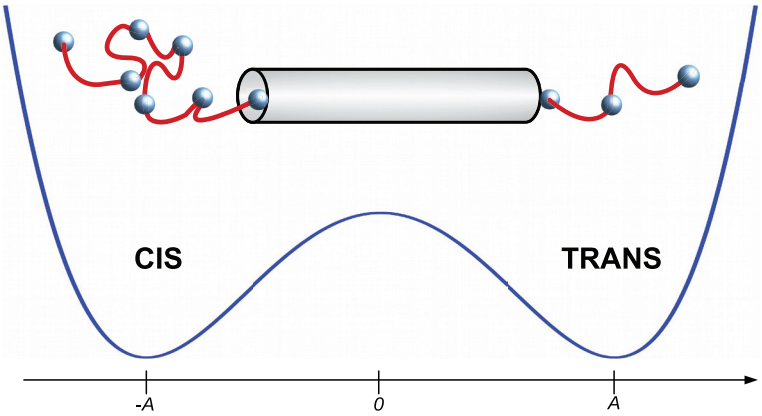

Figure 1. Cartoon of the translocation process of a polymer from CIS to TRANS side of a narrow nanopore. The double-well potential is a caricature of the two-state free energy landscape associated to the translocation.

noise, with average $\langle \xi_i(t) \rangle = 0$ and correlation $\langle \xi_i(0)\xi_j(t) \rangle = 2\gamma T \delta_{ij}\delta(t)$, where γ is a dimensional constant, that we will put equal to 1 in the following.

Let us notice, however, that our interest is not in the behavior of this particular model per se: we are indeed concerned with the general problem of finding proper RCs and their effective evolution equations from data. In this respect, model (1) should be meant as a 'generator' of time series. The analysis is not restricted to this specific system, and the same conceptual scheme could be applied to more complex models as well.

As discussed in the Introduction, in order to study the collective dynamics of the polymer we need to identify proper reaction coordinates that describe the state of the system, and then we have to infer effective equations for their evolution.

A natural choice seems to be the center of mass Q of the polymer, which roughly indicates the spatial position of the chain:

$$Q = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2}$$

Its dynamical equation is obtained by just summing up equation (1) for all the particles and dividing by N,

$$\dot{Q} = -\frac{1}{N}\sum_{i=1}^{N} V'(x_i) + \sqrt{\frac{2T}{N}}\,\eta_Q \tag{3}$$

where the reciprocal elimination of internal forces has been taken into account, as well as the mutual independence of the noises $\{\xi_i\}$ that combine into a delta-correlated Gaussian noise with zero mean and such that $\langle \eta_Q(0)\eta_Q(t) \rangle = \delta(t)$.

By posing $x_i = Q + u_i$, equation (3) can be recast as

$$\dot{Q} = -\frac{1}{N}\sum_{i} V'(Q + u_i) + \sqrt{\frac{2T}{N}}\,\eta_Q. \tag{4}$$

The above equation is formally exact, but it is not very useful in this form, since it depends on all the u_i terms.

The roughest approximation that can be done to achieve a closed form for equation (4) is to assume that the force term due to the external potential can be written as a (possibly complicated) function of Q only. If this is the case, a one-variable, memoryless model of the kind

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = F(Q) + \sqrt{2D}\,\eta_Q, \tag{M1}$$

should catch the relevant features of the macroscopic evolution of the system, where $D = T/N$. Assuming that the above approximation holds, the specific form of $F(Q)$ can be inferred from data, as will be discussed in the following.

Let us notice that equation (M1) describes a Markovian stochastic process for the variable Q, and it is expected to give a reasonable approximation of the real dynamics only when the knowledge of Q suffices to determine the macroscopic state of the system. For instance, equation (M1) gives a good approximation of the real dynamics in the limit of high-rigidity chain, as it will discussed in the following for a particular case.

In general, however, the above one-variable model will not be valid, meaning that it will not be possible to find any form of $F(Q)$ able to reproduce the dynamical properties of the original system in an accurate way. This is due to the fact that the dynamics of Q, in general, is not Markovian: in order to achieve a satisfactory coarse-grained description, one possibility is to modify equation (M1) by introducing memory-dependent terms, which in some cases can be found analytically by mean of projection methods [14, 49, 50]. To avoid such dependence on memory kernels, which are often difficult to manipulate, the only possibility is to search for (at least) a second RC of the system, such that the vector composed by Q and this new variable obeys a Markovian dynamics.

The additional RC can be individuated through the methods briefly mentioned in the Introduction, or it can be suggested by physical intuition. For instance, one can study a simplified version of the considered problem in order to understand what are the relevant variables. In our case, if the bond fluctuations are large enough, it is reasonable that the elongation $L = x_N - x_1$ have a role in the macroscopic dynamics. In appendix A we discuss our system under a strong approximation, which allows a simple analytical treatment: in this context the role of L is transparent. We can then guess an effective model of the form:

$$\begin{cases} \frac{\mathrm{d}Q}{\mathrm{d}t} = F_Q(Q,L) + \sqrt{2D_Q}\,\eta_Q \\ \\ \frac{\mathrm{d}L}{\mathrm{d}t} = F_L(Q,L) + \sqrt{2D_L}\,\eta_L. \end{cases} \tag{M2}$$

Again, assuming that the dynamics of (Q, L) is fairly described by a Markov process, the best choices for F_Q, F_L, D_Q, D_L can be found with data-driven approaches as the one used in this paper. However, one has then to verify that the chosen RCs are actually 'valid' macroscopic variables, i.e. that the coarse-grained dynamics (M2) reproduces the macroscopic features of the original system.

In the remaining part of this paper, we will try to implement this program in a specific case. In particular we will show that, as expected, the stiffness of the polymer plays an important role in the choice of the right set of RCs.

J. Stat. Mech. (2020) 013208

3. Polymer in a double well

We consider now the case in which the external potential $V(x)$ in equation (1) is a double-well. This simple model allows us to study some properties of thermally activated barrier crossing as, for instance, the dependence of the jump rate r on the physical parameters.

The general problem of activated dynamics has been extensively studied since the seminal works by Kramers, and the reaction-rate theory provides many analytic methods to compute jump times in different contexts (see [51] and reference therein). Important results have been derived also for polymeric chains [35, 36, 42]. We want to stress that our aim here is not to improve such results: we are interested in analyzing the ability of a data-driven approach to reconstruct an activated dynamics. In particular, we will focus on the role of the chain stiffness on the activated dynamics and, more importantly, on its relevance for the description in terms of the RCs.

The external potential reads, in this case,

$$V(x) = \frac{B^2}{4}(x^2 - A^2)^2 \tag{5}$$

where A and B are suitable constants. The typical dynamics of the center of mass, $Q(t)$, is therefore characterized by jumps over the barrier separating the two minima of the potential (two-state model). For the interaction potential we choose the form $U(r) = K(r - \sigma)^2/2$.

In the following, we will always consider the limit $N\sigma \simeq A$, i.e. the case in which the equilibrium length of the polymer is comparable to the half distance between the well minima: it is reasonable to expect that in these conditions the value of the bond rigidity affects the qualitative behavior of the chain in a relevant way.

Our aim is to show that for high values of K the model described by equation (M1) suffices to reproduce the quantitative macroscopic behaviour of the system; whereas, as soon as K becomes comparable to B^2A^2/σ, the evolution of Q is no more Markovian and any attempt to describe it through model (M1) is doomed to fail. However, if the phase-space is expanded by including a suitable additional RC, it is still possible that the evolution of the new RCs vector turns out to be Markovian, so that a dynamical description based on equation (M2) can be accurate enough.

The validity of such scenario can be tested by using the data-driven approach mentioned in the Introduction and detailed in appendix B. We first perform numerical simulations of the whole system by using a Stochastic Runge–Kutta algorithm [52] and measuring the relevant RCs of the system at every time step. As a first attempt, from long time-series of such data we build an effective stochastic equation for Q only, in the form of equation (M1); then we apply the extrapolation procedure to the dynamics of the two-dimensional vector (Q, L), obtaining an M2-like model. The 'goodness' of M1 and M2 is tested by measuring the Kramers' transition times of the reconstructed models and comparing the corresponding jump rates to the original ones.

3.1. 1-variable model

Before applying the mentioned extrapolation method to infer numerically the functional form of the terms appearing in equation (M1), let us derive analytically an

J. Stat. Mech. (2020) 013208

effective equation for Q for the high-stiffness limit, $K \gg B^2A^2/\sigma$. In this case we can assume that the position of two consecutive beads is fixed and equal to σ. Due to the simple form of the external potential $V(x)$, the drift term in equation (4) can be exactly computed in this case:

$$\begin{aligned} -\frac{1}{N}\sum_{i=1}^{N} V'(Q+u_i) &= -\frac{B^2}{N}\sum_{i=1}^{N}\left[(Q+u_i)^3 - A^2(Q+u_i)\right] \\ &= -\frac{B^2}{N}\sum_{i=1}^{N}\left[3Qu_i^2 + Q^3 - A^2Q\right] \end{aligned} \tag{6}$$

where we have used the fact that, due to the rigidity of the polymer, $\sum_i u_i^3 = \sum_i u_i = 0$.

Now we substitute the explicit expression for the relative positions of the polymer beads, $u_i = (2\mathrm{i} - N - 1)\sigma/2$, and after straightforward algebra we get

$$\begin{aligned} -\frac{1}{N}\sum_{i=1}^{N} V'(x_i) &= -B^2Q\left(Q^2 - A^2 + \frac{\sigma^2}{4}(N+1)(N-1)\right) \\ &= -B^2Q\left(Q^2 - A_{\mathrm{eff}}^2\right) \end{aligned} \tag{7}$$

with

$$A_{\mathrm{eff}} = A\sqrt{1 - \frac{L^2(N+1)}{4A^2(N-1)}}, \tag{8}$$

where we have used the definition of the polymer length for the rigid case, $L = (N-1)\sigma$. Let us notice that the above drift corresponds to an effective potential

$$V_{\mathrm{eff}}(Q) = \frac{B^2}{4}(Q^2 - A_{\mathrm{eff}}^2)^2, \tag{9}$$

i.e. a 'rescaled' version of the original external potential (5).

We can now use the theory of escape times [53] to estimate the jump rate r for the effective potential (9). The average waiting time between two consecutive jumps can be computed through the formula [53]

$$\begin{aligned} \tau &= \frac{1}{D}\int_{-A}^{A} \mathrm{d}y\, \mathrm{e}^{V(y)/D} \int_{-\infty}^{y} \mathrm{d}z\, \mathrm{e}^{-V(z)/D} \\ &= \frac{1}{D}\int_{-A}^{A} \mathrm{d}y\, \mathrm{e}^{\frac{B^2}{4D}\left[y^4 - 2A_{\mathrm{eff}}^2 y^2\right]} \int_{-\infty}^{y} \mathrm{d}z\, \mathrm{e}^{-\frac{B^2}{4D}\left[z^4 - 2A_{\mathrm{eff}}^2 z^2\right]} \end{aligned} \tag{10}$$

where $D = T/N$ is the diffusivity associated to Q. The jump rate r is then found as

$$r = \frac{1}{\tau}. \tag{11}$$

The above equations, which are only valid in the rigid-rod limit, will be a useful touchstone to evaluate the level of accuracy of the model inferred numerically.

We apply now the extrapolation procedure mentioned in the Introduction and detailed in appendix B to infer the right functional forms for the terms of the Langevin equation (M1), assuming that the dynamics of Q is Markovian and a 1-variable

J. Stat. Mech. (2020) 013208

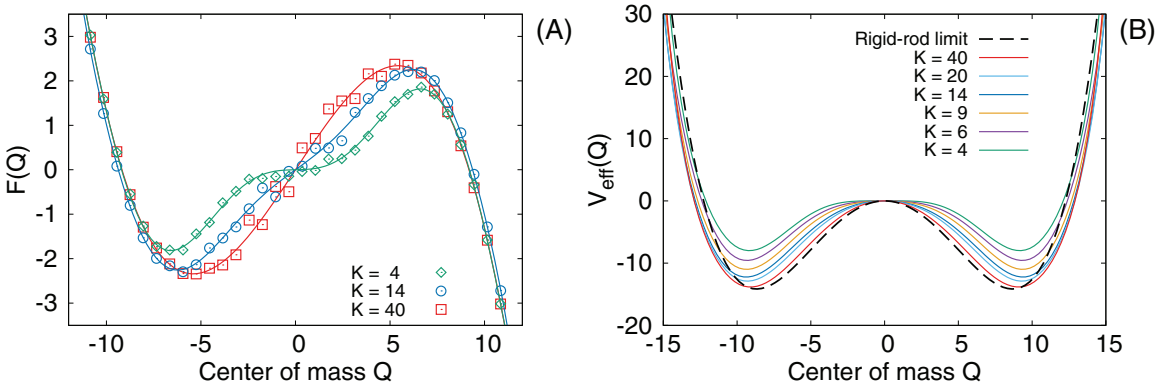

Figure 2. (A): drift term $F(Q)$ of model (M1) as reconstructed from data (points) and fitted with a 9th degree odd polynomial (solid lines), for three different values of K. (B): effective potential obtained by integration of $F(Q)$. Parameters for the simulations of the complete system: $A = 10$, $B = 0.1$, $T = 30$, $\sigma = 1$, $N = 10$, using a time-step $\mathrm{d}t = 10^{-5}$. Simulations on model (M1) have been run with a time-step $\mathrm{d}t = 10^{-4}$.

description in the form of model (M1) holds. We find that D (not shown here) is always almost constant and equal to T/N, as it would be expected if the process was Markovian, while the drift shows a more complex shape (figure 2(A)); we fit the data by a 9th degree, odd polynomial, then we integrate the resulting function in order to get an effective potential, which is reported in figure 2(B) for several values of K. In the large-K limit, as expected, we recover a quartic effective potential: terms of higher order become relevant when the bond stiffness is low, and their effect is to flatten the potential barrier between the two wells.

As mentioned above, the validity of equation (M1) relies on the assumption that the evolution of Q is Markovian, which has to be checked. First, one can define and measure the following quantity:

$$\zeta(t) = \dot{Q}(t) - F(Q(t)), \tag{12}$$

which represents the 'noise' of equation (M1), if the dynamics of Q is Markovian. We can compare the autocorrelation time of $\zeta(t)$ and verify that it is much shorter than any characteristic time-scale of the dynamics of Q. In our case, $\zeta(t)$ always decorrelates on the scale of the time-step of the integration algorithm, $\mathrm{d}t$ (see figure 3(A)).

This first check assures that there is a clear time-scale separation between the dynamics of the center of mass and its 'noise'. However, this does not imply that the original dynamics of Q has to be Markovian: in order to check that, we also have to verify the consistency with the original dynamics. If the one-variable description is able to catch the relevant features of the whole system, we can conclude, *a posteriori*, that the evolution of Q was Markovian also in the complete dynamics; if not, a different description has to be taken into account.

Figure 3(B) shows, for several values of the rigidity, the jump rates measured in the original dynamics and those observed in the reconstructed model, using a standard stochastic integration algorithm (the one discussed in [54], up to order $\mathrm{d}t^{3/2}$). In the high-K limit the simple rigid-rod approximation (10) holds, there is no dependence on K and the agreement between the jump rates of M1 and of system (1) is excellent. As

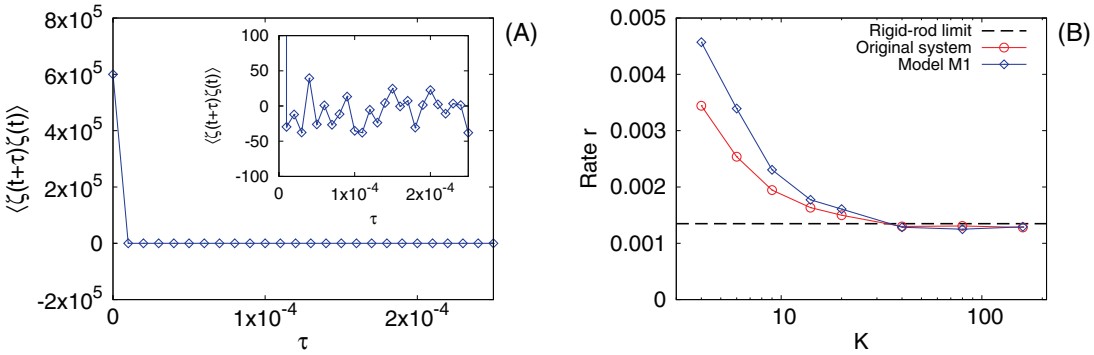

Figure 3. Checking the validity of model (M1). (A) Autocorrelation $\langle\zeta(t)\zeta(t+\tau)\rangle$ of the 'noise' term (12) versus τ. The inset is a zoom. It is apparent the fast decay with τ that can be considered effectively a delta-correlation, since the time-step of the integration algorithm generating the dynamics is $\mathrm{d}t = 10^{-5}$. Here, $K = 14$. (B): jump rates as functions of K in the original system (red circles) and in simulations of the reconstructed M1 (blue diamonds). The rigid-rod approximation (dashed line) is reported as reference. Parameters as in figure 2.

the polymer becomes softer, even if a significant improvement on equation (10) can be observed, the relative error between M1 and the true dynamics exceeds 30%: this is a clear hint that a 1-variable description, even if inferred directly from data, cannot reproduce all the relevant features of the dynamics. This is due to the fact that our implicit assumption on the Markovianity of the process is wrong.

3.2. 2-variables model

The failure of model (M1) for small values of K, revealed by the discrepancies between the reconstructed and the original jump rate, suggests to go beyond a single variable description in order to achieve satisfactory results. As discussed in section 2, a reasonable attempt to recover a Markovian dynamics is to consider the elongation of the polymer, L, as a second RC for our model, and we postulate the validity of an evolution equation of the form (M2).

The requirement of a variable accounting for the elongation of the polymer can be easily understood by looking at figure 4 reporting three scatter plots of the original dynamics in the (Q, L) plane, for different bond rigidity, where $L = x_N - x_1$. When K is high, and the system is well approximated by the rigid-rod model, the region of the phase space explored by the dynamics is a thin strip around the equilibrium value $L \simeq (N-1)\sigma$. As soon as the rigidity condition is relaxed, and the system is allowed to vary its length in a significant way, a two-lobe distribution takes place: L tends to be smaller than the rest length of the chain when the polymer occupies one of the two minima of the double-well potential, while it significantly increases during the transition across the barrier. This particular shape of the scatter plot indicates that the typical pathways in the space (Q, L) include a non-negligible deformation in L, which can be straightforwardly interpreted as follows: when the rigidity K is low, the transition across the barriers of the polymer occurs with a concomitant stretching of the bonds, presumably those that instantaneously lay on top the barrier. As a consequence,

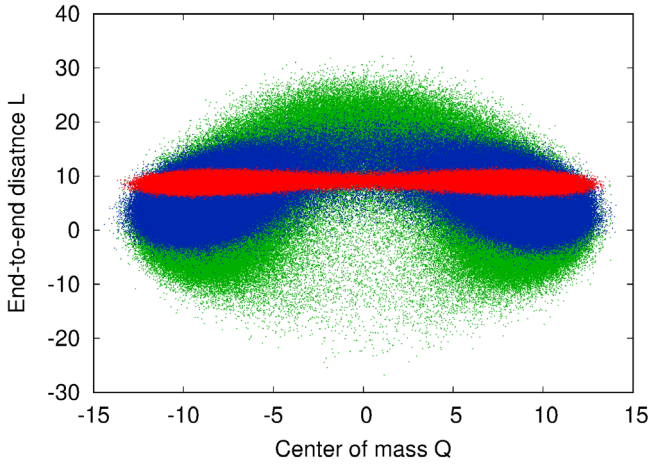

Figure 4. Scatter plot of the end-to-end distance L versus Q (total integration time: 10^6). Green dots: $K = 4$; blue dots: $K = 20$; red dots: $K = 600$. Other parameters as in figure 2.

any Markovian effective description involving only the center of mass is completely insufficient to fairly approximate the dynamics of the system.

Following again the strategy discussed in appendix B, we provide numerical values for F_Q, D_Q, F_L and D_L in the (Q, L) space, which have to be fitted using suitable functional forms. Due to the symmetries of the system, $F_Q(Q,L)$ has to be odd with respect to the variable Q, while $F_L(Q,L)$ should be even. Figure 5 shows the results obtained by fitting the following polynomial form:

$$\begin{aligned} F_Q(Q,L) &= Q\left[c_{10}^{(Q)} + c_{12}^{(Q)}L^2 + c_{13}^{(Q)}L^3\right] \\ &\quad + Q^3\left[c_{30}^{(Q)} + c_{32}^{(Q)}L^2 + c_{33}^{(Q)}L^3\right] \\ F_L(Q,L) &= c_{00}^{(L)} + c_{01}^{(L)}L + c_{03}^{(L)}L^3 + c_{21}^{(L)}Q^2L. \end{aligned} \tag{13}$$

The agreement between the actual data and the proposed functional form is good enough to hope that the guessed model catches the most relevant features of the dynamics. The diffusivity terms D_Q and D_L are again fitted by constant functions. Once model (M2) is determined, we can check the reliability of its stochastic evolution by a direct comparison with the original dynamics.

A first, important benchmark is given by the ability of the model to reproduce the static properties of the system, namely the joint probability distribution in the (Q, L) space. This test is reported in figure 6 for different values of K, showing a reasonable qualitative agreement even in the non-trivial case of low bond stiffness: in particular, the stretching occurring when the polymer crosses the barrier is clearly reproduced. The improvement of our effective description when also L is taken into account can be fully appreciated by looking at dynamical observables as the jump rate r. Figure 7 displays the relative errors between the values of r obtained in the reconstructed models M1, M2 and in the original dynamics. As already discussed, the 1-variable model fails when the polymer is soft, while the accuracy of M2 does not seem to be affected in this limit. Let us notice, on the other hand, that for $K \gg A^2B^2/\sigma$ the reconstructed model (M2) is less reliable than the 1-variable version: this is probably a consequence of the

J. Stat. Mech. (2020) 013208

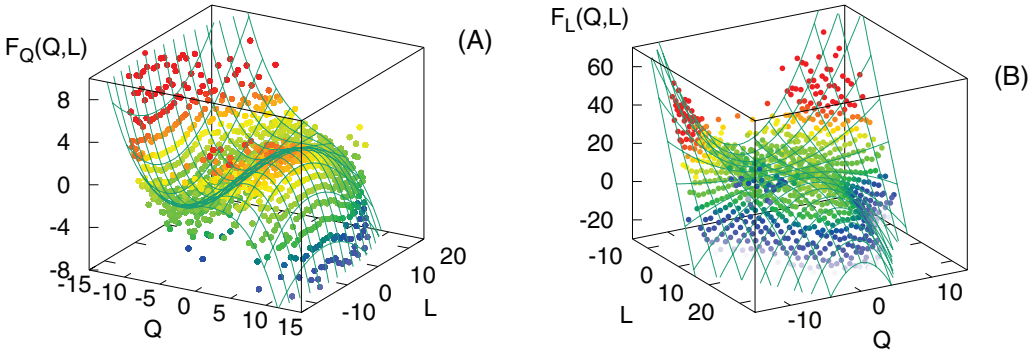

Figure 5. Reconstructed drift terms of Q and L in the model (M2), case $K = 4$. Points are extrapolated from data (see appendix B); the surface is obtained by fitting the polynomial (13). Other parameters are as in figure 2.

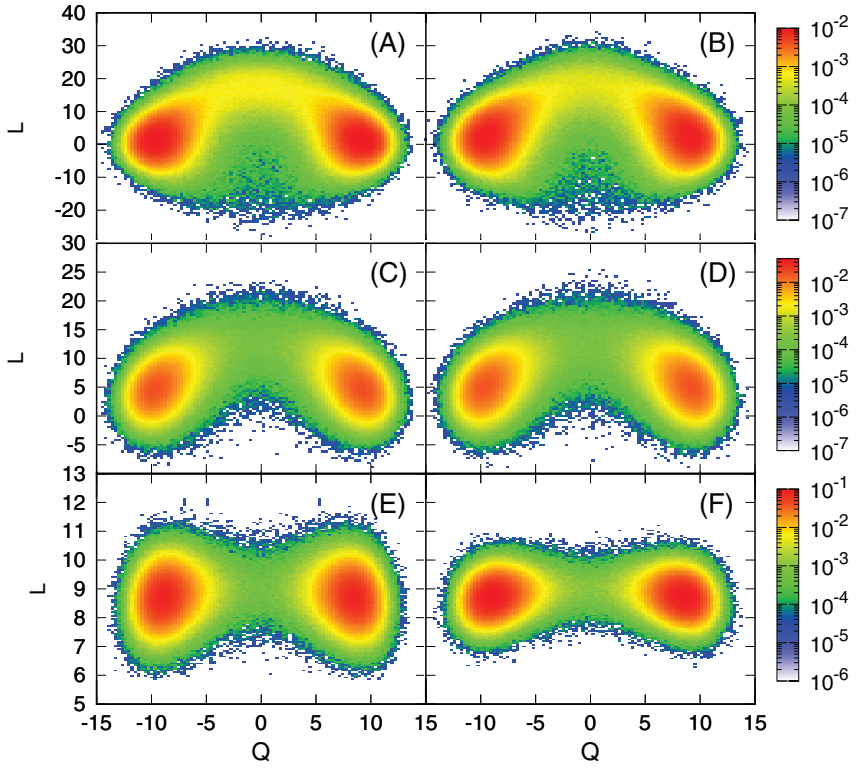

Figure 6. Pdfs in the 2-variable phase space of the original system, first column, and in the reconstructed 2-variable model (M2), second column. Stiffness: $K = 4$ (top), $K = 20$ (center), $K = 600$ (bottom). Other parameters are the same as figure 2.

larger number of parameters involved, which leads to a lower degree of precision on their determination with the discussed method.

3.3. Remarks on the structure of the effective equations

The procedure for the reconstruction of a model in the form (M2) that we used in the previous subsection is based on a 'dynamical' analysis, which builds the coefficients of

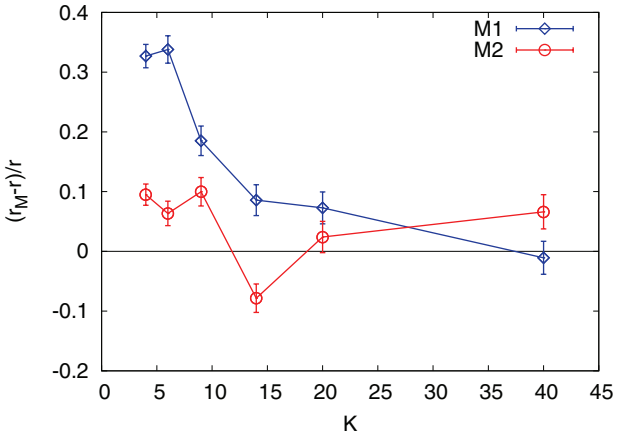

Figure 7. Relative errors of the jump rates r in the two reconstructed models, varying K. For other parameters, see caption of figure 2.

the guessed stochastic model by looking at the time evolution of suitable observables of the original system (see appendix B). One could wonder whether such procedure is really needed in order to get a realistic description of the studied process; for example, in analogy with many statistical mechanical problems, one may expect that the following recipe works:

1. Measure the stationary p.d.f. $\rho(Q, L)$ from long time series of data;
2. Deduce an effective 2-variable 'potential', also known as 'potential of mean force' in chemical and biophysical contexts:

$$W_S(Q, L) = -\log[\rho(Q, L)]. \tag{14}$$

3. Define $F_Q \equiv -c\partial_Q W_S$ and $F_L \equiv -c\partial_L W_S$, where the constant c has to be determined.

Let us notice that in many practical situations the multidimensional free-energy landscapes obtained by simulations or experiments are assumed to be generated by a system with a gradient (or gradient-like) structure and, using this hypothesis, Langer's formula [51, 55] is applied to derive the transition rates over saddles.

The above procedure is a completely 'static' analysis, because it involves only quantities measured in equilibrium conditions. Leaving apart the problem of finding the right multiplicative constant c and the noise terms, this approach has a major issue: it is not sure, *a priori*, whether the dynamics of the chosen reaction coordinates can be described by a potential, even if the complete system actually can; in other words, there is in general no reason to expect that the model is a gradient system in the reduced phase space.

The knowledge of W_S alone is not sufficient to obtain the drift; additional informations on the dynamics need to be taken into account. A possibility is the theoretical approach discussed in [56], which also involves the transport coefficients. Our numerical method, instead, exploits the dynamical information by computing suitable conditioned moments of the RCs, as discussed in appendix B.

J. Stat. Mech. (2020) 013208

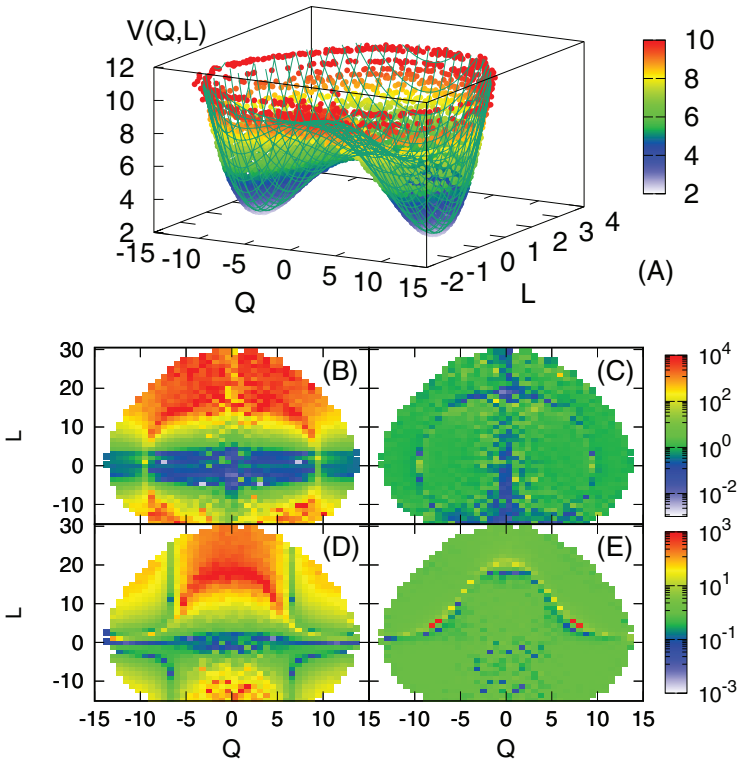

Figure 8. Static analysis. (A): effective potential $W_S(Q,L)$ from the empirical pdf in the (Q, L) space. Panels below compare the drift computed with this static approach, ∇W_s, and the one reconstructed with our dynamical approach, $(F_Q^{(2)}, F_L^{(2)})$, to the measured coefficients F_Q, F_L. (B): $|\partial_Q W_s/F_Q|$. (C): $|F_Q^{(2)}/F_Q|$. (D): $|\partial_L W_s/F_L|$. (E): $|F_L^{(2)}/F_L|$.

In order to show that in our case the simplifying assumption of a gradient structure for the (Q, L) dynamics is wrong, let us determine W_S using equation (14). In figure 8(A) we fit such 'potential' with a 2-variables polynomial (4th order in both Q and L), getting a nice superposition.

If the 2-variable system were gradient, $F_Q(Q, L)$ should be equal to $c\partial_Q W_S$, for some value of c; therefore the ratio $\partial_Q W_S/F_Q$ should be constant all over the phase-space. Figure 8(B) shows instead that, in our case, such function strongly depends on Q and L. Just for comparison, figure 8(C) displays the ratio

$$\epsilon = F_Q^{(2)}/F_Q$$

where $F_Q^{(2)}$ represents here the expression from the fitting (13): not surprisingly, this function is constant and equal to one almost everywhere, meaning only that we have made sensible choices for the polynomial functions to use in the fit. In figures 8(D) and (E) the same comparison is done for the drift of the end-to-end distance L.

This simple check clearly shows that the simplified 2-variables system is not gradient, and therefore a static analysis is not sufficient to infer reasonable effective equations for its evolution.

J. Stat. Mech. (2020) 013208

4. Conclusions

We studied the dynamics of a 1-dim polymer whose monomers are subjected to a double-well external potential. Under certain conditions, the phenomenology is characterized by thermally activated barrier crossing (classical Kramers' problem). We addressed the issue of describing this complex high-dimensional dynamics in terms of few suitable observables, the reaction coordinates (RCs). In our system the center of mass, Q, and the end-to-end distance, L, are the most natural candidates.

These RCs evolve according to effective Langevin equations that are generally difficult to be derived via a systematic procedure. The proper reconstruction of the effective stochastic equations for Q and L is achieved via a data-driven numerical method that extrapolates the drift and diffusion terms from a long trajectory of the original system. Let us stress that this method allows us to find nonlinear terms for the reconstructed Langevin equation; this is an important difference, e.g. from the standard Mori-Zwanzig approach, where the complexity of the problem is shifted into the shape of the memory kernels [49]. The reliability of the reconstructed dynamics is tested on the way it fairly reproduces some essential properties of the original dynamics, such as a correct estimate of the jump rate over the barrier.

From our study it emerges that the description level in terms of RCs strongly depends on the bond stiffness K. More precisely: if the bonds are rigid enough, we are allowed to consider the evolution of Q only, given by model (M1), to fully characterize the jump dynamics. However, when K decreases, the internal motion of the chain cannot be neglected and also the dynamics of L has to be considered, so that a satisfactory description can only be obtained in the plane (Q, L) through the model (M2). On a more mathematical perspective, lowering K can be regarded as a loss of Markovianity of the one-variable description. A second coordinate is necessary to recover the Markov property.

Our work shows how subtle is the procedure of reducing the dynamics of a many-dimensional system to a low-dimensional model, even in the simplest cases where the physical intuition leaves little ambiguity to the choice of the RCs. In fact, the choice of an RC can be correct in certain regimes but not sufficient in others. Specifically in our case, we can only know *a posteriori* that it is the stiffness of the polymer to determine whether one or two RCs are needed.

Acknowledgments

This work is part of MIUR-PRIN2017 *Coarse-grained description for non-equilibrium systems and transport phenomena (CO-NEST).*

Appendix A. The homogeneous chain approximation

In this appendix we discuss the 'homogeneous chain approximation' (HCA), which amounts to assuming that the distances between nearest-neighbor particles, at each time t, are all equal:

$$x_{i+1}(t) - x_i(t) = x_{j+1}(t) - x_j(t), \quad \forall i, j. \tag{A.1}$$

Under such approximation we can derive closed analytical expressions for both models (M1) and (M2); HCA is very unphysical, and it only holds true if the polymer bonds are almost rigid. However, a simple analysis of this limit can give us some insight on the general case. Even under such simplifying hypotheses, as soon as K is large enough, model (M1) is not sufficient to describe correctly the dynamics, and a (M2)-type model is needed.

First, let us consider the case of high intra-chain forces acting on each monomer. Specifically, the bonds are much stronger than: i) the external forces due to the action of the potential $V(x)$ and ii) those induced by the thermal fluctuations. Under conditions i) and ii) our polymer reduces to a rigid rod with fixed distances among its elements, i.e.

$$x_{i+1}(t) - x_i(t) = \text{const} \quad \forall i. \tag{A.2}$$

Equation (4) can be written as

$$\dot{Q} = -\int \mathrm{d}u \rho(u) V'(Q+u) + \sqrt{\frac{2T}{N}}\,\eta_Q \tag{A.3}$$

where we have introduced the density, $\rho(u) = 1/N \sum_{i=1}^{N} \delta(u - u_i)$, to pass from the sum to an integral. Here

$$\rho(u) \simeq \frac{1}{L}\Theta(u^2 - L^2/4), \tag{A.4}$$

where we also took the $N \gg 1$ limit. The total length of the polymer is constant, $L = (N-1)\sigma$.

In this simple case, we can straightforwardly apply the fundamental theorem of calculus and get:

$$\dot{Q} = \frac{1}{L}[V(Q - L/2) - V(Q + L/2)] + \sqrt{\frac{2T}{N}}\eta_Q. \tag{A.5}$$

Notice that the above equation is of the form of model (M1).

Let us consider now the idealized situation in which the inter-particle distances do fluctuate, but the approximation (A.5) still holds because the polymer undergoes homogeneous deformations. In this limit, the end-to-end distance L defined as

$$L = x_N - x_1 \tag{A.6}$$

is no longer constant and equation (A.5) should be complemented by a new equation describing the dynamics of L. Now L is a necessary second RC, so the phase space is enlarged to the plane (Q, L).

Under this hypothesis, we can derive a second equation, by writing $\dot{L} = \dot{x}_N - \dot{x}_1$ from equation (1) and then approximating each $x_i - x_{i-1} \simeq L/(N-1)$. The final result for the drift expressions is

$$F_Q(Q,L) = \frac{1}{L}\left[V\left(Q - \frac{L}{2}\right) - V\left(Q + \frac{L}{2}\right)\right]$$

$$F_L(Q,L) = -2U'\left(\frac{L}{N-1}\right) + V'\left(Q - \frac{L}{2}\right) - V'\left(Q + \frac{L}{2}\right). \tag{A.7}$$

The above phenomenological discussions suggest that there are two regimes depending on the polymer stiffness:

- Stiff chain: the polymer dynamics can be characterized by a single reaction coordinate Q, for which a closed evolution equation can be found;
- Soft chain: also a second variable, for instance the end-to-end distance L, is needed to close the evolution equations.

Notice that in the above discussion the introduction of L as a second RC emerges quite naturally. This can be a clue on the relevant variable to choose also in the more general case where assumption (A.1) does not hold.

Appendix B. Extrapolating Langevin equations from data

In this appendix we briefly recall the basic aspects of the extrapolation procedure that we use to infer the parameters of effective Langevin equations from long-time series of data (in this case, produced by numerical simulations). An extensive discussion of the method can be found in [27, 28] and reference therein. See also [32] for a case in which the study of a multi-dimensional system is considered.

Let us assume that each component of the vector variable $\mathbf{X}$ obeys the following Langevin's equation:

$$\dot{X}_i = F_i(\mathbf{X}) + \sqrt{2D_i(\mathbf{X})}\xi_i \tag{B.1}$$

where each ξ_i is a white noise with unitary variance. For the sake of simplicity, here we assume that the coefficients F_i and D_i do not depend on time. It can be proved [57] that the following relations hold:

$$\begin{aligned} F_i(\mathbf{x}) &= \lim_{\Delta t\to 0} \frac{1}{\Delta t}\langle \Delta X_i | \mathbf{X}(0) = \mathbf{x}\rangle \\ D_i(\mathbf{x}) &= \lim_{\Delta t\to 0} \frac{1}{2\Delta t}\langle (\Delta X_i - F_i(\mathbf{x})\Delta t)^2 \, | \mathbf{X}(0) = \mathbf{x}\rangle \end{aligned} \tag{B.2}$$

where $\Delta X_i = X_i(\Delta t) - X_i(0)$. Due to the stationarity of the process, we can compute the ensemble averages (B.2) as temporal averages over long-time series of data.

The $\Delta t \to 0$ limit has to be interpreted in a proper physical way: for every real phenomenon, a stochastic description holds only for some not-too-small time scales. It is customary to define a typical time-scale of the considered problem, usually referred to as the 'Markov–Einstein time' τ_{ME}, such that the Langevin equations hold true only if one considers time scales larger than τ_{ME}. In order to get a reasonable esteem of the above limits, a good strategy is that of plotting the conditioned moments on the rhs of equation (B.2) as functions of Δt, then to individuate a sufficiently regular region that allows for a $\Delta t \to 0$ extrapolation. In our case, however, the separation of time-scales is a consequence of the fast decorrelation times of the quantity $\zeta(t)$ defined by equation (12), which is in turn due to the overdamped nature of the original dynamics we are considering. As a consequence, it is sufficient to take $\Delta t \gg \mathrm{d}t$, where $\mathrm{d}t$ is the time-step of the integration algorithm. Let us notice once again that such time-scale

J. Stat. Mech. (2020) 013208

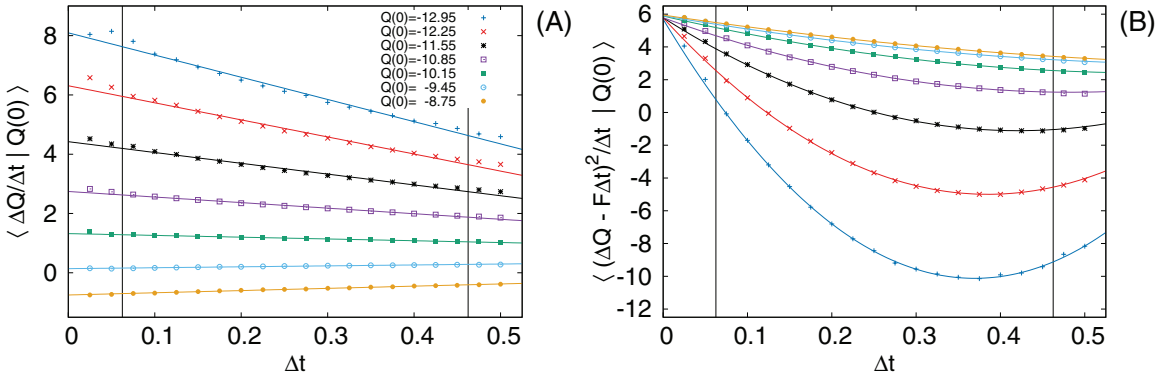

Figure B1. Extrapolation of the drift (panel A) and diffusivity (panel B) from data. The quantities in the rhs of equation (B.2) are fitted with a low-order polynomial; only the regions within the vertical bars are considered for the fit. The $\Delta t \to 0$ limit is then taken as the vertical intercept of the fitted functions with the y-axis. Here $K = 4$ (other parameters as in figure 2).

separation does not imply the Markovianity of the considered process, and the validity of such approximation can be only checked *a posteriori*.

In figure B1 we show a typical case of reconstruction of the drift and diffusivity terms of model (M1), for several values of the center of mass.

References

[1] Fersht A 1999 *Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding* 3rd edn (San Francisco, CA: W.H. Freeman & Co.)
[2] Finkelstein A and Ptitsyn O 2016 *Protein Physics: a Course of Lectures* (Amsterdam: Elsevier)
[3] Ghélis C 2012 *Protein Folding* (New York: Academic)
[4] Majda A and Wang X 2006 *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows* (Cambridge: Cambridge University Press)
[5] Zhang W, Hartmann C and Schütte C 2017 *Faraday Discuss.* **195** 365–94
[6] Banushkina P V and Krivov S V 2016 *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **6** 748–63
[7] Arnold I 1989 *Mathematical Methods of Classical Mechanics* (Berlin: Springer)
[8] Castiglione P, Falcioni M, Lesne A and Vulpiani A 2008 *Chaos and Coarse Graining in Statistical Mechanics* (Cambridge: Cambridge University Press)
[9] Cecconi F, Cencini M and Vulpiani A 2007 *J. Stat. Mech.* P12001
[10] Rubin R J 1960 *J. Math. Phys.* **1** 309–18
[11] Turner R E 1960 *Physica* **26** 269–73
[12] Mazur P and Braun E 1964 *Physica* **30** 1973–88
[13] Ford G W, Kac M and Mazur P 1965 *J. Math. Phys.* **6** 504–15
[14] Zwanzig R 1973 *J. Stat. Phys.* **9** 215–20
[15] Onsager L and Machlup S 1953 *Phys. Rev.* **91** 1505
[16] Jolliffe I T and Cadima J 2016 *Phil. Trans. R. Soc.* A **374** 20150202
[17] Tu J H 2014 *J. Comput. Dyn.* **1** 391
[18] Noé F and Nuske F 2013 *Multiscale Mod. Simul.* **11** 635–55
[19] Molgedey L and Schuster H G 1994 *Phys. Rev. Lett.* **72** 3634–7
[20] Klus S, Nüske F, Koltai P, Wu H, Kevrekidis I, Schütte C and Noé F 2018 *J. Nonlinear Sci.* **28** 985–1010
[21] Wehmeyer C and Noé F 2018 *J. Chem. Phys.* **148** 241703
[22] Brunton S L, Proctor J L and Kutz J N 2016 *Proc. Natl Acad. Sci.* **113** 3932–7
[23] Russell S J and Norvig P 2016 *Artificial Intelligence: a Modern Approach* (Malaysia: Pearson Education Limited)
[24] Mézard M 2018 *Europhys. News* **49** 26–9
[25] Hosni H and Vulpiani A 2018 *Phil. Technol.* **31** 557–69

[26] Buchanan M 2019 *Nat. Phys.* **15** 304
[27] Baldovin M, Cecconi F, Cencini M, Puglisi A and Vulpiani A 2018 *Entropy* **20** 807
[28] Kleinhans D, Friedrich R, Nawroth A and Peinke J 2005 *Phys. Lett.* A **346** 42–6
[29] Baldovin M, Puglisi A and Vulpiani A 2018 *J. Stat. Mech.* 043207
[30] Renner C, Peinke J and Friedrich R 2001 *J. Fluid Mech.* **433** 383–409
[31] Peinke J, Tabar M and Wächter M 2019 *Annu. Rev. Condens. Matter Phys.* **10** 107–32
[32] Baldovin M, Puglisi A and Vulpiani A 2019 *PLoS One* **14** 1–16
[33] Hummer G 2005 *New J. Phys.* **7** 34
[34] Micheletti C, Bussi G and Laio A 2008 *J. Chem. Phys.* **129** 074105
[35] Jun Park P and Sung W 1999 *J. Chem. Phys.* **111** 5259–66
[36] Sebastian K and Paul A K 2000 *Phys. Rev.* E **62** 927–39
[37] Bezrukov S M, Vodyanoy I and Parsegian V A 1994 *Nature* **370** 279
[38] Kasianowicz J J, Brandin E, Branton D and Deamer D W 1996 *Proc. Natl Acad. Sci.* **93** 13770–3
[39] Meller A, Nivon L and Branton D 2001 *Phys. Rev. Lett.* **86** 3435
[40] Huopaniemi I, Luo K, Ala-Nissila T and Ying S C 2006 *J. Chem. Phys.* **125** 124901
[41] Ammenti A, Cecconi F, Marini Bettolo Marconi U and Vulpiani A 2009 *J. Phys. Chem.* B **113** 10348–56
[42] Sebastian K L and Debnath A 2006 *J. Phys.: Condens. Matter* **18** S283–96
[43] Cressiot B, Oukhaled A, Patriarche G, Pastoriza-Gallego M, Betton J M, Auvray L, Muthukumar M, Bacri L and Pelta J 2012 *ACS Nano* **6** 6236–43
[44] Lubensky D K and Nelson D R 1999 *Biophys. J.* **77** 1824–38
[45] Ansalone P, Chinappi M, Rondoni L and Cecconi F 2015 *J. Chem. Phys.* **143** 154109
[46] Jalalinejad A, Bassereh H, Salari V, Ala-Nissila T and Giacometti A 2018 *J. Phys.: Condens. Matter* **30** 415101
[47] Muthukumar M 2010 *J. Chem. Phys.* **132** 195101
[48] Polson J M, Hassanabad M F and McCaffrey A 2013 *J. Chem. Phys.* **138** 024906
[49] Zwanzig R 1961 *Phys. Rev.* **124** 983–92
[50] Grabert H, Hänggi P and Talkner P 1980 *J. Stat. Phys.* **22** 537–52
[51] Hänggi P, Talkner P and Borkovec M 1990 *Rev. Mod. Phys.* **62** 251–341
[52] Honeycutt R L 1992 *Phys. Rev.* A **45** 600–3
[53] Gardiner C 2009 *Stochastic Methods: a Handbook for the Natural and Social Sciences* (Berlin: Springer)
[54] Mannella R and Palleschi V 1989 *Phys. Rev.* A **40** 3381–6
[55] Langer J 1969 *Ann. Phys.* **54** 258–75
[56] Grabert H 2006 *Projection Operator Techniques in Nonequilibrium Statistical Mechanics* vol 95 (Berlin: Springer)
[57] Friedrich R, Peinke J, Sahimi M and Reza Rahimi Tabar M 2011 *Phys. Rep.* **506** 87–162

J. Stat. Mech. (2020) 013208