Statistical methods in atomistic computer simulations

Prof. Michele Ceriotti, michele.ceriotti@epfl.ch

This course gives an overview of simulation techniques that are useful for the computational modeling of materials and molecules at the atomistic level. The student will learn about basic and advanced methods to evaluate thermodynamic averages by molecular dynamics, including accelerated sampling for the study of rare events, and non-linear dimensionality reduction to study structurally-complex systems.

Constant-temperature sampling in atomistic simulations

- Canonical averages and importance sampling
- Monte Carlo, detailed balance and the Metropolis algorithm
- Molecular dynamics, integrators, energy conservation
- Autocorrelation functions, correlation time and statistical efficiency

Thermostatting molecular dynamics

- Breaking energy conservation and getting into the canonical ensemble
- Global and local thermostats, deterministic and stochastic thermostats
- Langevin dynamics. Stochastic differential equations and sampling efficiency
- Colored-noise generalized Langevin dynamics

Rare events. Getting dynamics from ensemble averages

- Rare events and time-scale separation
- Transition-state theory on the potential energy surface
- Collective coordinates. Free energy and TST on the free-energy surface
- Beyond TST. Bennett-Chandler method, committor analysis

Reweighed sampling and adaptive biasing

- Reweighing a trajectory to get averages in a different ensemble
- Statistics of reweighing – sampling efficiency of weighted averages
- Umbrella sampling and metadynamics – basics, examples and caveats

Linear and non-linear dimensionality reduction

- Dimensionality reduction – coarse-graining the description of structurally complex systems
- Linear projections: principal component analysis; classical multidimensional scaling
- Non-linear dissimilarity measure: ISOMAP
- Sketch map: using proximity matching to describe atomistic problems

Prerequisites

Introductory knowledge of statistical mechanics and probability, some familiarity with a Unix/Linux environment. Some programming skills would be preferable.

Bibliography

M.P. Allen and D.J. Tildesley, Computer Simulation of Liquids
D. Frenkel and B. Smit, Understanding Molecular Simulation
M. Tuckerman, Statistical Mechanics: Theory and Molecular Simulation

License

1 Sampling in atomistic simulations

Here we will cover a particular set of problems that arise when one wants to model the behavior of a compound, a material, or a bio-molecule at the level of its atomic constituents. In all that follows we will assume 1. that the dynamics of the electronic degrees of freedom is completely decoupled from that of the nuclei, and that the electrons occupy the ground state at for each configuration $\mathbf{q}$ of the nuclei (Born-Oppenheimer approximation) 2. that the nuclei behave as classical, distinguishable particles, subject to the Hamiltonian

$$H(\mathbf{p},\mathbf{q}) = \sum_i \frac{\mathbf{p}_i^2}{2m_i} + V(\mathbf{q}),$$

m_i and $\mathbf{p}_i$ being the mass and the momentum of each nucleus, respectively[1]. It will be assumed that the *interaction* potential $V(\mathbf{q})$ between the atoms is well-understood, given by an empirical force field or obtained from first principles by solving the electronic structure problem with clamped nuclei.

We will focus on the problem of generating configurations of the atoms that are consistent with the given thermodynamic conditions, and computing (mostly) static thermodynamic properties that require averaging over such configurations. This is an issue that occurs very often in atomistic simulations, either because the equilibrium configuration of the atoms is not (fully) known experimentally, or because is not well defined (e.g. for liquids or amorphous materials), or because the property one wants to compute depends dramatically upon thermal fluctuations around the equilibrium geometry.

1.1 Sampling the canonical ensemble

Different thermodynamic ensembles are introduced by considering two thermodynamic parameters (energy, pressure, temperature, volume, ...) to be fixed, and to define the macroscopic state of the system. Here we will consider only the canonical (NVT) ensemble, in which temperature T, volume V and number of atoms N are assumed constant. This ensemble often corresponds to experimental conditions, and allows us to discuss most of sampling issues and the techniques to solve them. We will not discuss the derivation of the canonical ensemble, but just state that it implies that the probability of observing a configuration $(\mathbf{p},\mathbf{q})$ corresponds to

$$P(\mathbf{p},\mathbf{q}) = e^{-\beta H(\mathbf{p},\mathbf{q})}/Z, \qquad Z = \int \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q} e^{-\beta H(\mathbf{p},\mathbf{q})}, \tag{1.1}$$

where we have introduced the inverse temperature $\beta = 1/k_B T$, and the canonical partition function Z.

An important feature of the (classical) canonical ensemble – one that simplifies considerably analytical and numerical treatment – is that position and momentum are not correlated, so that the $\mathbf{p}$ and $\mathbf{q}$ parts of the partition function and of the probability distribution can be factored exactly, and treated separately

$$P(\mathbf{p},\mathbf{q}) = P(\mathbf{p}) \cdot P(\mathbf{q}) = \frac{e^{-\beta \sum_i \frac{\mathbf{p}_i^2}{2m_i}}}{\int \mathrm{d}\mathbf{p} e^{-\beta \sum_i \frac{\mathbf{p}_i^2}{2m_i}}} \cdot \frac{e^{-\beta V(\mathbf{q})}}{\int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})}}. \tag{1.2}$$

Note that $P(\mathbf{p})$ is just a multi-variate Gaussian, so the distribution of momenta is trivial and the normalization can be computed analytically. The difficulty is in determining the

[1] At times we will just use expressions such as $\mathbf{p}/m$ to mean the vector whose elements are $p_{i\alpha}/m_i$.

configurational part $P(\mathbf{q}) = e^{-\beta V(\mathbf{q})} / \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})}$, that depends on the potential and which is typically a very complicated function of the atomic coordinates.

Knowing $P(\mathbf{q})$ is important because the expectation value of any configuration-dependent property $A(\mathbf{q})$ (structure factors, average bond lengths, ...) can be computed as an integral over the configurational probability distribution:

$$\langle A \rangle = \int \mathrm{d}\mathbf{q}\, A(\mathbf{q})\; P(\mathbf{q}) = \frac{\int \mathrm{d}\mathbf{q}\, A(\mathbf{q})\; e^{-\beta V(\mathbf{q})}}{\int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})}}. \tag{1.3}$$

For simple, low-dimensional problems, computing an integral of the form (1.3) by some kind of quadrature (e.g. computing the value of the integrand on a grid of $\mathbf{q}$ points) is a sensible proposition, but it becomes completely impractical as the number of atoms increases: even with just two grid points per degree of freedom, the evaluation of the integrand on a grid requires 2^{3N} points.

1.2 Importance sampling and Monte Carlo methods

The exponential increase of the computational cost involved in an integration by quadrature can be in principle circumvented by using a Monte Carlo integration strategy. Without entering into the details of the general idea, one can generate a series of M random configurations $\mathbf{q}_i$, and approximate $\langle A \rangle$ as

$$\langle A \rangle \approx \frac{\sum_i A(\mathbf{q}_i)\, e^{-\beta V(\mathbf{q}_i)}}{\sum_i e^{-\beta V(\mathbf{q}_i)}}.$$

The problem with this approach is that typically $P(\mathbf{q})$ has a very erratic behavior, with sharp peaks over the regions of configuration space that correspond to a low value of the potential, and a minute, negligible value for the vast majority of configurations that would correspond to fairly random arrangements of atoms and would therefore have a very large value of the potential energy. Hence, one would need to generate a huge quantity of random configurations to pick any one with a non-negligible value of $e^{-\beta V(\mathbf{q})}$, not to mention to converge accurately the value of the integral [1].

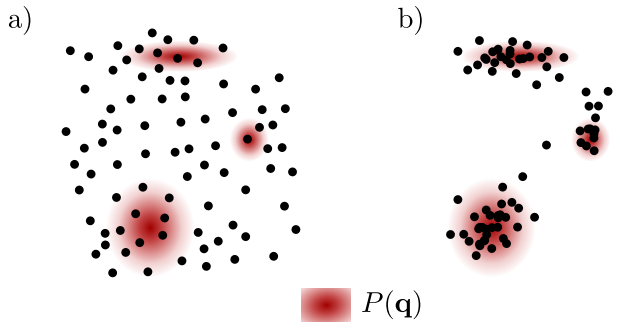


Figure 1: Panel (a) demonstrates randomly selected configuration points, while panel (b) demonstrates importance sampling relative to the probability distribution $P(\mathbf{q})$.

One possible solution to this problem is to generate a sequence of points that are precisely distribute according to $P(\mathbf{q})$ (a strategy known as *importance sampling*, see Figure 1). In that case, points would naturally concentrate in regions with a high value of $P(\mathbf{q})$, and few or no points would be present where the probability distribution has a negligible value. Expectation values could then be obtained from this sequence of points simply as

$\langle A \rangle \approx \frac{1}{M} \sum_i A(\mathbf{q}_i)$, since the exponential Boltzmann factor would be implicitly accounted for by the uneven distribution of samples. The problem is then how to generate a set of atomic configuration consistent with the canonical distribution.

In general, techniques to generate this canonically-distributed set of points proceed iteratively, by taking one point $\mathbf{q}_i$ and providing a rule to generate a new point $\mathbf{q}_{i+1} = \mathcal{P}(\mathbf{q}_i)$, in such a way that for any pair of points in the sequence $P(\mathbf{q})/P(\mathbf{q}') = e^{-\beta\left(V(\mathbf{q})-V(\mathbf{q}')\right)}$. In most cases the rule that generates a new configuration is not deterministic, but is characterized by a probability distribution $p(\mathbf{q} \to \mathbf{q}') = p(\mathbf{q}_{i+1} = \mathbf{q}' | \mathbf{q}_i = \mathbf{q})$. A necessary condition for this to be the case is that the canonical distribution itself is left invariant under the action of the operation that generates the sequence of points, i.e. that

$$\int \mathrm{d}\mathbf{q} P(\mathbf{q}) p(\mathbf{q} \to \mathbf{q}') = P(\mathbf{q}'). \tag{1.4}$$

A more stringent, sufficient condition that is however easier to prove in most cases is that of *detailed balance*, that relates the probabilities of performing a move $\mathbf{q} \to \mathbf{q}'$ and that of the reverse move $\mathbf{q}' \to \mathbf{q}$ to the relative probability of the initial and final configurations:

$$P(\mathbf{q}) p(\mathbf{q} \to \mathbf{q}') = P(\mathbf{q}') p(\mathbf{q}' \to \mathbf{q}). \tag{1.5}$$

It is easy to show that Eq. (1.5) implies Eq. (1.4), by integrating both sides over $\mathbf{q}$ and realizing that the probability of going *anywhere* starting from $\mathbf{q}'$ has to integrate to one.

Metropolis Monte Carlo A simple strategy to construct a transition rule that satisfies the detailed balance condition is to split the rule in a generation step and an accept/reject step, so that the overall probability of doing the $\mathbf{q} \to \mathbf{q}'$ move is the product of the generation and acceptance probabilities, $p(\mathbf{q} \to \mathbf{q}') = g(\mathbf{q} \to \mathbf{q}') a(\mathbf{q} \to \mathbf{q}')$. In the simplest scenario, the generation step is characterized by a symmetric probability of giving rise to the forward or backward move, $g(\mathbf{q} \to \mathbf{q}') = g(\mathbf{q}' \to \mathbf{q})$. For instance, one could add a Gaussian random number with mean zero and a small variance to the coordinates of a randomly-selected atom. Then, one has to decide whether to accept the new configuration, or reject it and return to the original configuration *that must be counted as a new sample*[2] in order to compute averages correctly.

A common approach to construct the acceptance probability so that the detailed balance is satisfied is to apply the Metropolis criterion [2], $a(\mathbf{q} \to \mathbf{q}') = \min(1, P(\mathbf{q}')/P(\mathbf{q}))$. In fact,

- if $P(\mathbf{q}') > P(\mathbf{q})$, $p(\mathbf{q} \to \mathbf{q}') = g(\mathbf{q} \to \mathbf{q}')$ and $p(\mathbf{q}' \to \mathbf{q}) = g(\mathbf{q}' \to \mathbf{q}) P(\mathbf{q})/P(\mathbf{q}')$ – that implies detailed balance provided that the generation probability is symmetric
- if $P(\mathbf{q}') < P(\mathbf{q})$, $p(\mathbf{q} \to \mathbf{q}') = g(\mathbf{q} \to \mathbf{q}') P(\mathbf{q}')/P(\mathbf{q})$ and $p(\mathbf{q}' \to \mathbf{q}) = g(\mathbf{q}' \to \mathbf{q})$ – that is again consistent with Eq. (1.5).

Note that what makes this approach doable is that the acceptance criterion only depends on the *ratio* of $P(\mathbf{q}')$ and $P(\mathbf{q})$. Looking back at Eq. (1.2), it is clear that the difficult part of $P(\mathbf{q})$ is the normalization, and that the ratio of the two probability is just $e^{-\beta\left(V(\mathbf{q}')-V(\mathbf{q})\right)}$, which is as easy to evaluate as the energy of the tentative new configuration.

One could then wonder what is the most efficient way to design an effective strategy for the moves. We will discuss later a quantitative approach to evaluate the statistical

[2] It is simple to convince oneself that in case of rejection the initial configuration must be counted one more time. Imagine a system that can exist in two discrete states, A and B, with $P(A) \gg P(B)$. The simplest generation rule is just to jump to the different state. Clearly, in order to be consistent with the ratio of the probabilities, most attempts to go from A to B should be rejected, whereas most reverse attempts should be accepted. However, if rejections did not contribute an extra sample to the averages, this procedure would just create an alternating sequence $ABABABAB\ldots$, which is clearly inconsistent with the relative probabilities of the two states.

efficiency of a sampling strategy. Now let us just hand-wavingly state that a good Monte Carlo step should make the system "move" a lot in configuration space. So, taking the simple example of moving one atom by adding a Gaussian random number with variance σ to its Cartesian coordinates, one would like to take large steps and so to pick a very large σ. However, if the atom is moved by a large amount, it is likely that in the process a chemical bond will be broken, or that the atom will end up in very close proximity of a second one. In both cases, the energy associated with the final configuration will be much larger than the initial one, and the ratio between the final and initial probabilities will be very close to zero, so that the move will almost invariably be rejected. In practice there is a trade-off between how large is the step taken and the probability it will be accepted.

This leads us to a more fundamental drawback of Monte Carlo methods. Consider a system composed of 100 atoms. Clearly, in order for the system to visit a configuration that bears no resemblance with the starting configuration, all the atoms will need to move. This means that one will need to perform 100 times more steps to generate a truly un-correlated configuration than if there was a single atom: unless the potential can be broken into local contributions, these 100 moves will imply the full cost of computing 100 times the system's potential energy. Now, one may think to move more than one atom at a time. To see why this would not really solve the problem, imagine having tuned the step size for a single-atom move so that the acceptance probability is on average 0.5. Now, create a copy of the system, that does not interact with the first copy, and move one atom by the same amount. The probability of accepting the combined move of the two systems is the square of the probability of moving just one, so the overall acceptance is reduced to 1/4. This though experiment demonstrates that Monte Carlo methods do not scale well with system size – as one needs to compute the energy for the whole system just to move a small subset of the particles. Clearly, this is not a problem in cases where one can compute inexpensively the change in energy determined by a local move, and can be more than compensated in cases in which one can generate smart moves that evolve the system across high free-energy barriers.

1.3 Molecular dynamics

Let's consider a different approach to generate a series of configurations consistent with the canonical ensemble. Let us consider what happens if we choose a configuration of position and momentum $(\mathbf{p}, \mathbf{q})$ that is consistent with Boltzmann statistics, and evolve it in time based on Hamilton's equations

$$\dot{\mathbf{q}} = \frac{\partial H}{\partial \mathbf{p}} = \frac{\mathbf{p}}{m}, \qquad \dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{q}} = -\frac{\partial V}{\partial \mathbf{q}}. \tag{1.6}$$

Integration of Eqs. (1.6) for a finite time step can be regarded as a discrete Monte Carlo step: this is sort of a peculiar rule to generate new configurations, in that it is deterministic and the acceptance is one. The initial conditions at time zero determine a trajectory $(\mathcal{P}_{\mathbf{p}}(\mathbf{p}, \mathbf{q}; t), \mathcal{P}_{\mathbf{q}}(\mathbf{p}, \mathbf{q}; t))$ in phase space, so that

$$p((\mathbf{p},\mathbf{q}) \to (\mathbf{p}', \mathbf{q}')) = \delta((\mathbf{p}', \mathbf{q}') - (\mathcal{P}_{\mathbf{p}}(\mathbf{p}, \mathbf{q}; t), \mathcal{P}_{\mathbf{q}}(\mathbf{p}, \mathbf{q}; t))).$$

Note that because of time-reversal symmetry of the momentum, this move does *not* satisfy the detailed balance condition Eq. (1.5) – loosely speaking, one would have to flip the sign of the final momentum in order to have a reverse move that brings one back to the initial $(\mathbf{p},\mathbf{q})$. One can prove [3] that in actuality MD satisfies a simple generalization of detailed balance. However, it is more instructive to show that a MD time step fulfills the more general necessary condition (1.4).

Conservation of density and phase-space volume First, let us define the position of an atomistic system in phase space as the $6N$-dimensional vector $\mathbf{x} = (\mathbf{p}, \mathbf{q})$ that

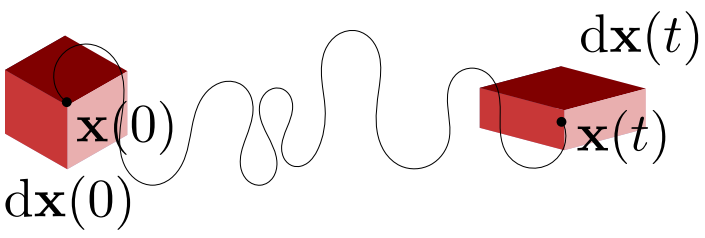


Figure 2: Evolution of a configuration in phase space and of the corresponding volume element along a molecular dynamics simulation.

combines position and momentum of all the N atoms. Then, it is easy to see that the MD trajectory conserves the probability distribution, a

$$\frac{dP}{dt} \propto e^{-\beta H} \frac{dH}{dt}, \qquad \frac{dH}{dt} = \frac{\partial H}{\partial \mathbf{p}} \cdot \dot{\mathbf{p}} + \frac{\partial H}{\partial \mathbf{q}} \cdot \dot{\mathbf{q}} = -\frac{\partial H}{\partial \mathbf{p}} \cdot \frac{\partial H}{\partial \mathbf{q}} + \frac{\partial H}{\partial \mathbf{q}} \cdot \frac{\partial H}{\partial \mathbf{p}} = 0. \tag{1.7}$$

In order to be able to perform the integral (1.4) one also needs to work out how the volume element $\mathrm{d}\mathbf{x}(0)$ is transformed to $\mathrm{d}\mathbf{x}(t)$. Here it is useful to imagine the evolution of the volume element as that of a swarm of trajectories starting off around $\mathbf{x}(0)$ – a picture that is very useful as it naturally links a description of dynamics in terms of trajectories in phase space to one that deals with the time evolution of a probability density. The change of variables $\mathbf{x}(0) \to \mathbf{x}(t)$ is associated with the Jacobian determinant [4]

$$J(t) = \det \mathbf{M}, \qquad M_{ij} = \frac{\partial x_i(t)}{\partial x_j(0)}.$$

Clearly, $M_{ij}(0) = \delta_{ij}$, so $J(0) = 1$. Then, the point is showing that $J'(t) = 0$, to prove that the volume of the phase space element is conserved by Hamiltonian dynamics. One can use Jacobi's formula[3] to get $J'(t) = J(t) \operatorname{Tr}\left(\mathbf{M}^{-1}\mathbf{M}'\right)$. Then, one can see that $\left(\mathbf{M}^{-1}\right)_{ij} = \partial x_i(0)/\partial x_j(t)$ since

$$\sum_k \frac{\partial x_i(0)}{\partial x_k(t)} \frac{\partial x_k(t)}{\partial x_j(0)} = \frac{\partial x_i(0)}{\partial x_j(0)} = \delta_{ij},$$

as the left-hand side is just a chain-rule sum. Considering also that $M'_{ij} = \partial \dot{x}_i(t)/\partial x_j(0)$,

$$\operatorname{Tr}\left(\mathbf{M}^{-1}\mathbf{M}'\right) = \sum_{ij} \frac{\partial x_i(0)}{\partial x_j(t)} \frac{\partial \dot{x}_j(t)}{\partial x_i(0)} = \sum_{ijk} \frac{\partial x_i(0)}{\partial x_j(t)} \frac{\partial x_k(t)}{\partial x_i(0)} \frac{\partial \dot{x}_j(t)}{\partial x_k(t)}.$$

The sum over i corresponds to $\mathbf{M}\mathbf{M}^{-1} = \mathbf{1}$, so

$$\operatorname{Tr}\left(\mathbf{M}^{-1}\mathbf{M}'\right) = \sum_k \frac{\partial \dot{x}_k(t)}{\partial x_k(t)} = \frac{\partial}{\partial \mathbf{p}} \cdot \dot{\mathbf{p}} + \frac{\partial}{\partial \mathbf{q}} \cdot \dot{\mathbf{q}} = \frac{\partial}{\partial \mathbf{p}} \cdot \left(-\frac{\partial H}{\partial \mathbf{q}}\right) + \frac{\partial}{\partial \mathbf{q}} \cdot \frac{\partial H}{\partial \mathbf{p}} = 0$$

The conservation of the Hamiltonian and of the phase space differential means that the probability distribution is conserved by the MD propagation, i.e. that the necessary condition for canonical sampling is satisfied:

$$\int \mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q} e^{-\beta H(\mathbf{p},\mathbf{q})} \delta\left(\left(\mathbf{p}', \mathbf{q}'\right) - \left(\mathcal{P}_{\mathbf{p}}(\mathbf{p}, \mathbf{q}; t), \mathcal{P}_{\mathbf{q}}(\mathbf{p}, \mathbf{q}; t)\right)\right) = e^{-\beta H\left(\mathbf{p}', \mathbf{q}'\right)}.$$

[3] Alternatively, starting from the definition of the determinant of a matrix $\mathbf{M}$ as the product of the eigenvalues μ_i, it is easy to see that $\ln \det \mathbf{M} = \sum_i \ln \mu_i = \operatorname{Tr} \ln \mathbf{M}$, so formally $\frac{d}{dt} e^{\operatorname{Tr} \ln \mathbf{M}} = e^{\operatorname{Tr} \ln \mathbf{M}} \operatorname{Tr}\left(\mathbf{M}^{-1} \frac{d}{dt} \mathbf{M}\right)$.

Velocity Verlet integrator Having established that integrating Hamilton's equations, let us now get on to how they can be integrated on a computer. Despite being relatively simple first-order differential equations, Eqs. (1.6) cannot be integrated analytically except for the simplest problems, so one has to resort to an approximate scheme to evolve the system along a MD trajectory. Many more or less complicated integrators (algorithms to perform evolve Eqs. (1.6) over a finite time step dt) have been used and proposed, but the simplest and in many ways effective integrator is probably the symmetric-split velocity Verlet algorithm [5, 6]. In this algorithm, the momentum $\mathbf{p}$ and the position $\mathbf{q}$ are propagated in turns according to a linearization of Hamilton's equation:

$$\begin{aligned} \mathbf{p} \leftarrow & \mathbf{p} - \frac{\partial V}{\partial \mathbf{q}} \frac{dt}{2} \\ \mathbf{q} \leftarrow & \mathbf{q} + \frac{\mathbf{p}}{m} dt \\ \mathbf{p} \leftarrow & \mathbf{p} - \frac{\partial V}{\partial \mathbf{q}} \frac{dt}{2}. \end{aligned} \tag{1.8}$$

Note that even though it appears that the force has to be computed twice in Eqs. 1.8, in practice the force computed at the end of one step can be re-used at the beginning of the following step. The reason why it is useful to split in two the propagation of the momentum is that in this form the finite-time velocity Verlet propagator is explicitly time-reversible and symplectic. This can be seen by writing explicitly the propagators for q and p – we will use a one-dimensional example to make notation less cumbersome:

$$\mathcal{P}_q\left(p, q; dt\right) = q' = q + \frac{p}{m} dt - V'\left(q\right) \frac{dt^2}{2m} \qquad \mathcal{P}_p\left(p, q; dt\right) = p' = p - V'\left(q\right) \frac{dt}{2} - V'\left(q'\right) \frac{dt}{2}. \tag{1.9}$$

It is simple to write the time-reversed step, starting from $(q', -p')$ and evolving for $-dt$. For instance, from Eqs. (1.9) one can obtain $V'\left(q'\right) \frac{dt}{2} = p - p' - V'\left(q\right) \frac{dt}{2}$ and $V'\left(q\right) \frac{dt}{2} = p + \frac{m}{dt}\left(q - q'\right)$. Using these it is easy to see that

$$\mathcal{P}_q\left(-p', q'; -dt\right) = q' + \frac{p'}{m} dt - V'\left(q'\right) \frac{dt^2}{2m} = q,$$

and that similarly $\mathcal{P}_p\left(-p', q'; -dt\right) = p$. In order to prove that the velocity Verlet step is also exactly symplectic it is sufficient to compute the elements of the Jacobian matrix from Eq. (1.9): $M_{qq} = 1 - V''\left(q\right) \frac{dt^2}{2m}$, $M_{qp} = \frac{dt}{m}$, $M_{pp} = 1 - V''\left(q'\right) \frac{dt^2}{2m}$, $M_{pq} = V''\left(q\right) \frac{dt}{2} - V''\left(q'\right) \frac{dt}{2} \left(1 - V''\left(q\right) \frac{dt^2}{2m}\right)$ and see that the determinant is $J = M_{qq} M_{pp} - M_{qp} M_{pq} = 1$.

Energy conservation Even though the velocity Verlet integrator fulfills exactly two of the properties Hamiltonian dynamics, yet it is not exact. The use of a finite time step entails necessarily an integration error, which – in a sufficiently large, chaotic system – will lead to the discrete trajectories to diverge exponentially from the trajectory that would be obtained by exact integration of the dynamics. In practice, provided that the time step is sufficiently small, the molecular dynamics trajectory is still sampling an ensemble which is extremely close to the target one, and also exhibit very similar dynamical properties. This is at times explained in terms of the existence of a "shadow Hamiltonian", which would generate precisely the trajectory that is obtained by finite time step integration, and that is very close to the actual Hamiltonian. The hypothetical existence of this shadow Hamiltonian explains why it is important to use an integration algorithm that fulfills the symmetries of Hamiltonian dynamics.

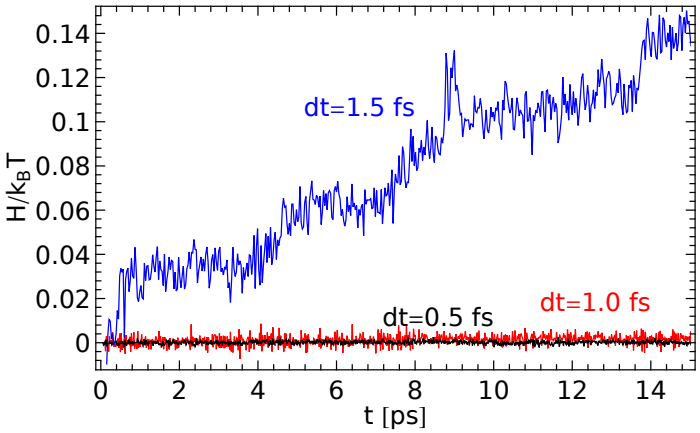


Figure 3: Energy conservation for a simulation of liquid water at room temperature, using a velocity Verlet integrator with three different values of the time step.

The most straightforward manifestation of integration errors is the fact that the total energy is not conserved along the MD trajectory. It is tedious but straightforward to write $V(q') + p'/2m$ in terms of a Taylor expansion around the initial value of q and p, finding that the leading error term in the expansion is $\mathcal{O}(dt^3)$. One can exploit the fact that energy conservation is violated because of the finite time step integration to monitor the accuracy of the trajectory – under the assumption that a simulation with poor energy conservation will contain sizable errors in average and dynamical observables. As shown in Figure 3, the for small values of the time step dt, the total energy fluctuates around a constant value, with fluctuations getting smaller as dt is decreased. For too large values, instead, H exhibits a systematic *drift* – a sign of a very substantial violation of energy conservation, that should be avoided. On a longer time scale, it is common to observe a drift even with reasonable choices of dt. In the case of constant-temperature simulations (that will be discussed in Chapter 2) a small drift is generally acceptable.

1.4 Ergodicity and autocorrelation functions

In order to compute ensemble average by generating a sequence of configurations, it is not sufficient that these configurations are distributed according to the probability distribution function of the ensemble. The trajectory $\{A_i\}$ must also satisfy an ergodic hypothesis, i.e. it must be true that

$$\lim_{M\to\infty} \frac{1}{M} \sum_i A(\mathbf{q}_i) = \int \mathrm{d}\mathbf{q}\, P(\mathbf{q})\, A(\mathbf{q})\,. \tag{1.10}$$

To see how a set of configurations that satisfy the requirements given in the previous sections could break the assumption (1.10), imagine a situation in which the configuration space is divided in two disconnected regions, so that transitions between any pair of points satisfy detailed balance, but there is zero probability of having a transition between the two regions. A single trajectory starting on one of the two areas would never visit the other half of configuration space, and hence the trajectory average would differ from the ensemble average.

This is all but an academic concern: in a practical case, one does not only require that Eq. (1.10) holds in the $M \to \infty$ limit, but also would like convergence to be *fast*, to evaluate averages accurately within the limited time available for the simulation. To see how to get a quantitative measure of the efficiency of sampling, consider obtaining a large

number of independent trajectories with M samples each, and compute the average of these independent averages:

$$\left\langle \frac{1}{M}\sum_i A_i \right\rangle = \frac{1}{M}\sum_i \langle A_i\rangle = \langle A\rangle\,.$$

Unsurprisingly, the average of the means is the target average value. Here we intend $\langle\cdot\rangle$ to represent an ensemble average, so there are no ergodicity concerns and the fact that $\langle A_i\rangle = \langle A\rangle$ follows from the fact that at each instant in time the samples are by hypothesis distributed according to the target distribution.

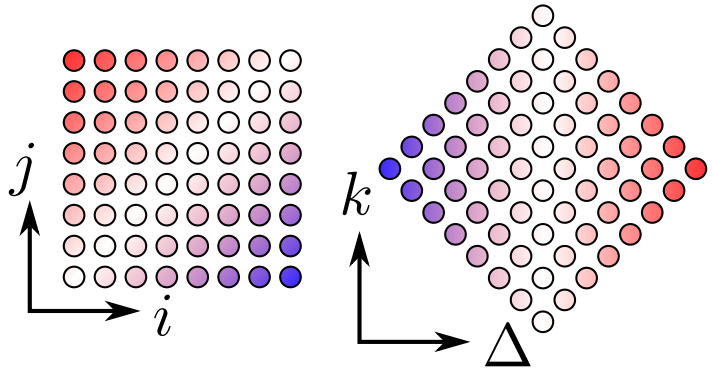


Figure 4: Change of summation variables to obtain a relation between the error in the mean and the correlation time.

Now, let us evaluate the *error in the mean*:

$$\epsilon_A^2(M) = \left\langle \left(\frac{1}{M}\sum_i A_i - \langle A\rangle \right)^2 \right\rangle = \frac{1}{M^2}\left\langle \sum_{i,j=0}^{M-1} (A_i - \langle A\rangle)(A_j - \langle A\rangle) \right\rangle.$$

One can start by re-arranging the summation so that the summation goes over the variables $\Delta = j - i$ and k (see Figure 4), that leads to

$$\begin{aligned}\epsilon_A^2(M) = \frac{1}{M^2}\Bigg\langle &\sum_{\Delta=-(M-1)}^{-1}\sum_{k=|\Delta|}^{M-1}(A_k - \langle A\rangle)(A_{k+\Delta} - \langle A\rangle) + \,. \\ &+ \sum_{\Delta=0}^{M-1}\sum_{k=0}^{M-1-|\Delta|}(A_k - \langle A\rangle)(A_{k+\Delta} - \langle A\rangle)\Bigg\rangle\end{aligned}$$

Then, the crux is assuming that the process that generates the trajectory is *stationary*, i.e. that the relation between two points in the sequence only depends on the difference between their position in the sequence and not on their "absolute" location within the sequence. So the dependence on k becomes immaterial when the ensemble average is brought inside the summation, and one gets

$$\begin{aligned}\epsilon_A^2(M) &= \frac{1}{M^2}\sum_{\Delta=-(M-1)}^{M-1}\langle (A_0 - \langle A\rangle)(A_\Delta - \langle A\rangle)\rangle\,(M - |\Delta|) = \\ &= \frac{\sigma^2(A)}{M}\sum_{\Delta=-(M-1)}^{M-1} c_{AA}(\Delta)\left(1 - \frac{|\Delta|}{M}\right),\end{aligned} \tag{1.11}$$

where we have introduced the autocorrelation function

$$c_{AA}(t) = \langle (A_0 - \langle A \rangle)(A_t - \langle A \rangle) \rangle / \sigma^2(A), \qquad \sigma^2(A) = \langle A^2 \rangle - \langle A \rangle^2 . \tag{1.12}$$

The autocorrelation function $c_{AA}(t)$ describes how quickly the trajectory loses memory of fluctuations away from the mean. It starts off at 1 for $t = 0$, and (except for cases with pathological behavior) decays to zero for $t \to \infty$. An analogous expression can be derived for a continuous trajectory, with the summation being replaced by an integral over time. Approximating the sum in (1.11) with the autocorrelation time $2\tau_A = \sum_{t=-\infty}^{\infty} c_{AA}(t)$ (which is the limit for $M \to \infty$), one obtains that $\epsilon_A^2(M) \approx \sigma^2(A)\, 2\tau_A / M$.

Hence, there is a very direct relation between how quickly the trajectory forgets about past fluctuations of an observable and how rapidly the error in the mean decreases. The autocorrelation time can therefore be taken as a rigorous measure of the ergodicity of a trajectory, and in general one would want to manipulate the sampling strategy to minimize τ_A for the observables of interest. It is worth however listing some caveats:

- different observables might have very different correlation times, so having a short correlation time for one observable does not guarantee that the trajectory is ergodic for any other observable;
- a single observable can exhibit multiple time scales – this is typical for instance of observables that sum many contributions (e.g the potential energy of a complex system). In these cases, one might be misled by the fast initial decay of $c_{AA}(t)$, and miss a long-time tail that can contribute a lot towards τ_A;
- computing τ_A from an actual simulation is not a trivial exercise – in general the simulation should be hundreds of times longer than τ_A itself;
- in the (many) cases in which the mean value of the observable $\langle A \rangle$ is not know exactly, and is computed from the same trajectory, the estimator[4] for $c_{AA}(t)$

$$c_{AA}(t) \approx \frac{1}{M-t} \frac{1}{\sigma_A^2} \sum_{k=0}^{M-1-t} (A_k - \langle A \rangle)(A_{k+t} - \langle A \rangle) \tag{1.13}$$

  is a biased estimator, and the resulting autocorrelation time will be under-estimated. One can again think to the case of two disconnected regions of phase space: computing c_{AA} from a trajectory that only visits one of the regions will under-estimate the fluctuations, and miss the fact that the estimate of c_{AA} should not decay to zero (since the mean obtained from the trajectory differs from $\langle A \rangle$ averaged over the totality of phase space). Whenever it is possible, one should verify the ergodicity of the trajectory computing correlation functions of observables whose mean value is analytically known – e.g. when it should be zero by symmetry.

[4] It is possible to compute the correlation function more efficiently by using fast Fourier transforms, see e.g. Ref. [1].

2 Constant temperature molecular dynamics

The molecular dynamics approach described in Section 1.3 samples (apart from finite time step errors) the constant *energy* microcanonical ensemble. In other terms, even though it does conserve the canonical ensemble, and so a collection of independent trajectories starting from uncorrelated points consistent with the finite-temperature Boltzmann distribution would yield correct averages, a *single* trajectory is highly non-ergodic (it does not allow fluctuations of H!), and there is no guarantee that it would yield averages consistent with the target experimental conditions. Fortunately, it is relatively simple to modify Hamilton equations (1.6) to allow for energy fluctuations, so as to obtain ergodic sampling of the canonical ensemble. These changes to Hamiltonian dynamics are generally referred to as *thermostats*.

2.1 Andersen thermostat

A very simple – and in many ways elegant – idea to to obtain ergodic trajectories from Hamiltonian dynamics exploits the factorization of the canonical distribution in position and momentum-dependent parts (1.2). Since the momentum distribution is just a multivariate Gaussian, it is easy to obtain at any time a new random value of the momentum consistent with the ensemble. So, the Andersen thermostat [7] simply amounts at performing segments of Hamiltonian dynamics, and re-sampling a new value of the momentum $\mathbf{p}$ from the target distribution every now and then – either at regular intervals or probabilistically. Every time the momentum is re-sampled from $\mathbf{p}$ to $\mathbf{p}'$ the total energy changes by $\mathbf{p}'^2/2m - \mathbf{p}^2/2m$, so that the trajectory can explore different constant-energy surfaces and ultimately achieve ergodic sampling. Note that the Andersen thermostat lends itself to a straightforward physical analogy: re-sampling the momentum is equivalent to the atoms in the system interacting at once with an ideal gas at the target temperature.

Defining a conserved quantity In Section 1.3, we discussed how the total energy can be used to monitor the accuracy of the integration of a molecular dynamics trajectory. It is useful to introduce a conserved quantity that can be used to the same aim in the presence of a thermostat. In the case of the Andersen thermostat – and of all the thermostat that will be discussed here – it is simple to do so. Each segment of Hamiltonian dynamics conserves the total energy, so the problem is keeping track of the changes in H that occur when the momentum is changed by application of the thermostat equations. In practice, one can accumulate $\Delta H \leftarrow \Delta H + \mathbf{p}^2/2m - \mathbf{p}'^2/2m$, where $\mathbf{p}$ and $\mathbf{p}'$ is the momentum just before and just after the step that correspond to the application of the thermostat. Then, $\tilde{H} = H + \Delta H$ would be perfectly conserved along the trajectory if the time step was infinitesimally small, and so one can monitor $\tilde{H}$ to assess the accuracy of the integration [8]. In practice, it is often the case that a considerably larger drift can be tolerated in the presence of a well-designed thermostat than with purely Hamiltonian dynamics.

2.2 Local and global, stochastic and deterministic thermostats

Before looking into the details of how it is possible to improve the effectiveness of a thermostat to perform ergodic sampling, it is perhaps useful to briefly review the different approaches that exist to manipulate Hamiltonian dynamics into performing canonical sampling. Andersen thermostat can be deemed to be the archetype of so-called *local* thermostats, that modify the velocities of individual degrees of freedom enforcing the

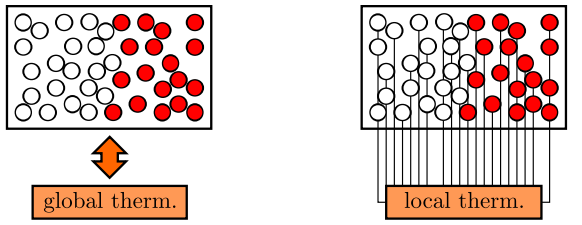


Figure 5: A schematic representation of the action of a global and a local thermostat on a system of particles which is initially strongly out of equilibrium.

prescribed $e^{-\beta p_{i\alpha}^2/2m_i}$ canonical distribution.

Even before the introduction of the Andersen thermostat, it was customary to periodically re-scale the velocities of the system in such a way that $\mathbf{p}^2/m = 3Nk_BT$. This method (velocity rescaling) and a version with smoothly varying velocities (Berendsen thermostat [9]) do not guarantee canonical sampling, but can be considered the model for *global* thermostats, that do not couple to individual degrees of freedom but to the total kinetic energy of the system, enforcing its distribution to be the one dictated by the canonical distribution:

$$P(K) \propto \int \mathrm{d}\mathbf{p} e^{-\beta \mathbf{p}^2/2m} \delta\left(K - \mathbf{p}^2/2m\right).$$

By transforming to spherical coordinates [1], and by transforming into an integral in $\mathrm{d}K$, one gets

$$P(K) \propto \int \mathrm{d}p\, p^{3N-1} e^{-\beta p^2/2m} \delta\left(K - p^2/2m\right) \propto K^{3N/2-1} e^{-\beta K}. \tag{2.1}$$

A simple and effective approach to perform velocity rescaling in a way that is consistent with (2.1) has been for instance discussed in Ref. [8]. The advantage of a global thermostat is that it typically introduces minor changes to the dynamical properties of the system, so that one can reliably evaluate time-dependent observables without having to run multiple constant-energy trajectories starting off different equilibrated configurations. The main problem of coupling the thermostat to the *total* kinetic energy of the system, is that one has to assume that internal equilibration will be reached quickly thanks to the inherent dynamics of the system. The problem is schematically represented in Figure 5: if one started off with half of the system frozen at zero temperature, and the other half at twice the target temperature, the overall kinetic energy would be consistent with (2.1), and a global thermostat would not accelerate internal equilibration. What is more, the resulting dynamics might exhibit long correlations for observables that depend on the energy partitioning among different degrees of freedom. In general, global thermostats tend to be very effective for internally ergodic systems such as liquids, but they can be problematic for solids, or in general for problems that contain weakly coupled subsystems.

For both local and global schemes, two radically different philosophies have been used to obtain a canonical distribution at a prescribed temperature. Thermostats such as Andersen that rely on (pseudo)random numbers are classified as *stochastic*. The thermostats that

[1] In the presence of multiple masses it is convenient to work in mass-scaled units $\tilde{p} = p/\sqrt{m}$, which only introduce a multiplicative factor in the integral.

will be discussed in this chapter belong to this category, which has as an advantage relative simplicity – since the complexity of describing an infinite, ergodic bath is modeled by obtaining uncorrelated variates from the appropriate distribution. However, stochastic thermostats did not have an associated conserved quantity, which made somewhat harder to assess the accuracy of finite-dt integration and of the resulting averages. Because of this, considerable effort has been devoted to the derivation of *deterministic* thermostatting schemes, which are more or less explicitly based on an extended-Lagrangian formulation, in which additional degrees of freedom are added to mimic the behavior of the bath. The first example of such a thermostat was devised by Nosé and Hoover [10, 11], based for a one-dimensional problem on equations of the following kind

$$\dot{q} = \frac{p}{m}, \quad \dot{p} = -\frac{\partial V}{\partial q} - p\frac{p_s}{Q}, \quad \dot{p}_s = \frac{p^2}{m} - k_B T, \quad \dot{s} = \frac{p_s}{Q}. \tag{2.2}$$

Here s and p_s are fictitious "position" and "momentum" of an additional degree of freedom of mass Q, whose dynamics is such that canonical sampling is enforced on the physical variables that are coupled to it. Nosé-Hoover thermostats exist in both a global and local form, the latter having the formally displeasing features of not being invariant with respect to a rigid rotation of the physical system. Furthermore, when applied to poorly ergodic systems such as a harmonic crystal, Nosé-Hoover thermostats do not help much enhancing the efficiency of exploration of phase space, and it is necessary to include further fictitious degrees of freedom, forming chains of thermostats [12] that require more complex integrators.

2.3 Langevin dynamics

A simple and elegant approach to achieve Boltzmann sampling in molecular dynamics is based on the Langevin equation [13]. Langevin dynamics was initially obtained as a model for Brownian motion, and consists in the introduction of a viscous friction and noisy force terms on top of Hamilton's equations. In one dimension,

$$\dot{q} = \frac{p}{m}, \quad \dot{p} = -\frac{\partial V}{\partial q} - \gamma p + \sqrt{2m\gamma/\beta}\xi, \quad \langle \xi(t)\,\xi(0)\rangle = \delta(t), \tag{2.3}$$

where $\beta = 1/k_B T$, γ is a friction term, and ξ is uncorrelated in time. It is not obvious to define the meaning of Eq. (2.3), particularly for what concerns the noisy force ξ, which varies discontinuously from time to time. In order to give a more precise meaning to the Langevin equation, it is useful to introduce a few concepts from the theory of random processes and stochastic differential equations [3].

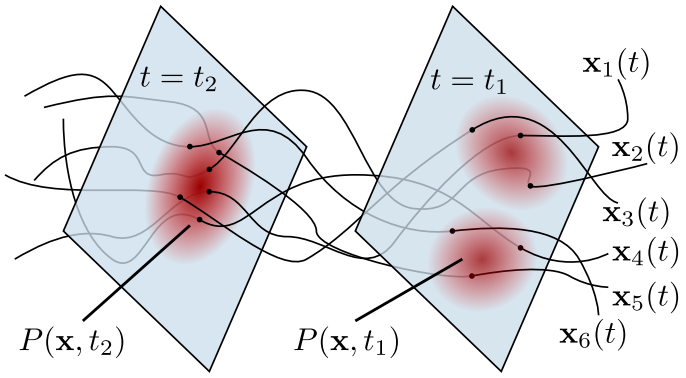
Random processes and the Fokker-Planck equation Consider a system whose state is described by the value of a vector $\mathbf{x}$ (e.g. the momentum and position $(\mathbf{p}, \mathbf{q})$), which can evolve in time according to an unknown law, possibly characterized by a degree of random behavior. We now assume that we have collected several realizations of this process. We will refer to each trajectory as a *sample path* $\mathbf{x}(t)$. We let Ω be the set of all such paths, and label each path according to some index ω. One can then describe the random process in terms of the distribution of the points in phase space at a given time, and hence construct a probability density[2] (figure 6)

$$P(\mathbf{x}, t) \propto \int \delta(\mathbf{x}_\omega(t) - \mathbf{x})\, \mathrm{d}\omega.$$

This probability, however, does not characterize the random process completely, since one only has knowledge on the "snapshots" of the collection of sample paths at different

[2]The integral here is just used to mean some "averaging" procedure to be performed over all the realizations of the random process.

Figure 6: A collection of sample paths for a random process. Also shown is how the one-time probability density $P(\mathbf{x}, t)$ can be constructed as the distribution of the points of all the sample paths at a given time.



times. No information regarding the identity of the paths in the different snapshots has been collected. One could compute the *joint* probability for a sample path to be at $\mathbf{x}_1$ at time t_1, *and* at $\mathbf{x}_2$ at time t_2,[3]

$$P(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2) \propto \int \delta(\mathbf{x}_\omega(t_1) - \mathbf{x}_1)\, \delta(\mathbf{x}_\omega(t_2) - \mathbf{x}_2)\, \mathrm{d}\omega. \tag{2.4}$$

One could then define a hierarchy of n-point probability densities. Fortunately, it is often justified to make a number of assumptions on the form of the joint probability (2.4) so as to bring it in a more treatable form. A first simplification requires the process to be *stationary* in time, i.e. that the two-times joint probabilities only depend on the time difference,

$$P(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2) = P(\mathbf{x}_1, t_1 - t_2; \mathbf{x}_2, 0). \tag{2.5}$$

Let us now introduce the *conditional probability* $P(\mathbf{x}_1, t_1 | \mathbf{x}_2, t_2)$, which is defined as the probability of the system being at $\mathbf{x}_1$ at a given time t_1, given that it was as $\mathbf{x}_2$ at time t_2. Its relation to the joint probability is

$$P(\mathbf{x}_1, t_1 | \mathbf{x}_2, t_2) = P(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2) / P(\mathbf{x}_2, t_2). \tag{2.6}$$

A random process is said do be Markovian if the joint conditional probability densities only depend on the most recent time frame, e.g.:

$$P(\mathbf{x}_1, t_1; \ldots; \mathbf{x}_k, t_k | \mathbf{x}_{k+1}, t_{k+1}; \ldots; \mathbf{x}_n, t_n) = P(\mathbf{x}_1, t_1; \ldots; \mathbf{x}_k, t_k | \mathbf{x}_{k+1}, t_{k+1}). \tag{2.7}$$

This ansatz means that at each time the model has no memory of the past history, and that further evolution is (probabilistically) determined uniquely by knowledge of the status at a given instant[4]. The description of the stochastic process is thus enormously simplified. Using the definition of conditional probability (2.6), one can write

$$\begin{aligned} P(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \mathbf{x}_3, t_3) =& P(\mathbf{x}_1, t_1 | \mathbf{x}_2, t_2; \mathbf{x}_3, t_3)\, P(\mathbf{x}_2, t_2; \mathbf{x}_3, t_3) = \\ =& P(\mathbf{x}_1, t_1 | \mathbf{x}_2, t_2)\, P(\mathbf{x}_2, t_2 | \mathbf{x}_3, t_3)\, P(\mathbf{x}_3, t_3), \end{aligned}$$

[3] We will assume times to be ordered according to $t_1 \geq t_2 \geq \ldots$.

[4] This might seem to be a very crude assumption, but it holds true at least approximately for a large number of physically-relevant problems.

i.e. any joint probability can be broken down to a product of the initial, single-time distribution and a series of conditional probabilities. If the process is also stationary, the conditional probability will depend only on the time difference, and hence its evolution is completely determined by the initial probability distribution and by the unique two-point conditional probability $P(\mathbf{x}, t|\mathbf{x}_0, 0)$.

One can see that under mild conditions on the form of $P(\mathbf{x}, t|\mathbf{x}_0, 0)$ (e.g. that it arises from a Markovian, stationary process, with continuous sample paths), the most general way to describe the time evolution of $P(\mathbf{x}, t|\mathbf{x}_0, 0)$ is given by the Fokker-Planck equation [14, 3]:

$$\begin{array}{ll} \frac{\partial}{\partial t} P(\mathbf{x}, t|\mathbf{x}_0, 0) = & -\sum_i \frac{\partial}{\partial x_i} \left[a_i(\mathbf{x}, t) P(\mathbf{x}, t|\mathbf{x}_0, 0)\right] + \quad \leftarrow \text{drift} \\ & +\frac{1}{2} \sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \left[D_{ij}(\mathbf{x}, t) P(\mathbf{x}, t|\mathbf{x}_0, 0)\right] \quad \leftarrow \text{diffusion} \end{array} \tag{2.8}$$

Showing the link between a Fokker-Planck equation and Langevin dynamics would require one to first define the meaning of a stochastic differential equation, which can be done for instance by Itō calculus (see e.g. Ref. [3]). Here we will consider the stochastic differential equation

$$\dot{\mathbf{x}} = \mathbf{a}(\mathbf{x}, t) + \mathbf{B}(\mathbf{x}, t)\boldsymbol{\xi}, \tag{2.9}$$

with $\mathbf{B}(\mathbf{x}, t)\mathbf{B}(\mathbf{x}, t)^T = \mathbf{D}(\mathbf{x}, t)$, to be just a short-hand for the associated Fokker-Planck equation (2.8).

Liouville equation While in the general case showing the connection between Eqs. 2.9 and 2.8 requires the use of stochastic calculus, the deterministic case for which $\mathbf{B} = 0$ can be discussed more easily. Consider an arbitrary test function $f(\mathbf{x})$, and define $\langle \square \rangle = \int d\mathbf{x}\, \square\, P(\mathbf{x}, t|\mathbf{x}_0, 0)$. From any given sample path, one can obtain that

$$\frac{\partial}{\partial t} f(\mathbf{x}) = \nabla f \cdot \dot{\mathbf{x}} = \nabla f(\mathbf{x}) \cdot \mathbf{a}(\mathbf{x}, t),$$

and

$$\begin{aligned} \left\langle \frac{\partial}{\partial t} f(\mathbf{x}) \right\rangle &= \sum_i \int d\mathbf{x}\, \frac{\partial f}{\partial x_i} P(\mathbf{x}, t|\mathbf{x}_0, 0)\, a_i(\mathbf{x}, t) = \\ &= -\int d\mathbf{x}\, f(\mathbf{x}) \sum_i \frac{\partial}{\partial x_i} \left[a_i(\mathbf{x}, t) P(\mathbf{x}, t|\mathbf{x}_0, 0)\right], \end{aligned} \tag{2.10}$$

integrating by parts and knowing that the boundary term must be zero if P is to be normalizable. At the same time, by exchanging the integral and the time derivative one gets

$$\frac{\partial}{\partial t} \langle f(\mathbf{x}) \rangle = \int d\mathbf{x}\, f(\mathbf{x}) \frac{\partial}{\partial t} P(\mathbf{x}, t|\mathbf{x}_0, 0). \tag{2.11}$$

Since the right hand sides of (2.10) and (2.11) are equal for any test function f, the drift part of the Fokker-Planck equation (2.8) follows. Taking the case of Hamiltonian dynamics, that has no explicit dependence on time, and that in this context can be formulated as

$$\frac{\partial}{\partial t}(\mathbf{p}, \mathbf{q}) = \mathbf{a}(\mathbf{p}, \mathbf{q}) = (-\nabla V(\mathbf{q}), \mathbf{p}/m),$$

one gets the Liouville formulation of classical mechanics in terms of a probability density of trajectories,

$$\frac{\partial}{\partial t} P((\mathbf{p}, \mathbf{q}), t|(\mathbf{p}_0, \mathbf{q}_0), 0) = \nabla V \cdot \nabla_{\mathbf{p}} P - \frac{\mathbf{p}}{m} \cdot \nabla_{\mathbf{q}} P.$$

Free-particle limit of the Langevin equation The free-particle limit of the Langevin equation can be easily integrated using its Fokker-Planck form. In one dimension, it reads just $\dot{p} = -\gamma p + \sqrt{2m\gamma/\beta}\xi$, that corresponds to the Fokker-Planck equation

$$\dot{P}\left(p,t|p_0,0\right) = \gamma\frac{\partial}{\partial p}\left(pP\right) + \frac{m\gamma}{\beta}\frac{\partial^2 P}{\partial p^2}. \tag{2.12}$$

First, it is easy to find the stationary probability by taking $\dot{P} = 0$. One integral can be done straight away, and the integration constant has to be zero for P to be positive definite. One is left to solve

$$\frac{\partial P}{\partial p} = -\frac{\beta}{m}pP \qquad \Rightarrow \qquad P\left(p\right) \propto e^{-\beta\frac{p^2}{2m}};$$

at equilibrium, the momenta of a system evolving under a free-particle Langevin equation are canonically distributed. Note that the friction γ does not enter the stationary solution, as it only governs the relaxation dynamics and not the equilibrium properties. The finite-time solution with a boundary condition $P\left(p,0|p_0,0\right) = \delta\left(p-p_0\right)$ is

$$P\left(p,t|p_0,0\right) \propto \exp -\frac{\beta}{2m}\frac{\left(p-p_0e^{-\gamma t}\right)^2}{1-e^{-2\gamma t}},$$

as it can be checked by direct substitution into (2.12). Note that this expression provides an explicit finite-time propagator to obtain a sequence of momenta consistent with the (free-particle) Langevin equation: starting from p_0, the sample path at time t is a Gaussian centered in $p_0e^{-\gamma t}$, with variance $\frac{m}{\beta}\left(1-e^{-2\gamma t}\right)$: given the initial momentum $p\left(0\right)$ one can obtain

$$p\left(t\right) = e^{-\gamma t}p\left(0\right) + \sqrt{m/\beta}\sqrt{1-e^{-2\gamma t}}\xi, \quad \left\langle\xi\right\rangle = 0, \left\langle\xi^2\right\rangle = 1,$$

where ξ is a Gaussian variate with zero mean and unit variance.

The harmonic oscillator As a slightly more complex example, let's now consider a harmonic oscillator. For simplicity, we will work in mass-scaled units, i.e. $p \leftarrow p/\sqrt{m}$ and $q \leftarrow q\sqrt{m}$. Then, the Langevin equation can be written in a matrix form

$$\frac{\partial}{\partial t}\begin{pmatrix} q \\ p \end{pmatrix} = -\begin{pmatrix} 0 & -1 \\ \omega^2 & \gamma \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{2\gamma/\beta} \end{pmatrix}\begin{pmatrix} 0 \\ \xi \end{pmatrix}$$

that corresponds to the Fokker-Planck equation

$$\frac{\partial}{\partial t}P\left(\left(p,q\right),t|\left(p_0,q_0\right),0\right) = \left[-p\frac{\partial}{\partial q} + \omega^2 q\frac{\partial}{\partial p} + \gamma p\frac{\partial}{\partial p} + \gamma\right]P + \frac{\gamma}{\beta}\frac{\partial^2 P}{\partial p^2}.$$

It is easy to check by direct substitution that the Boltzmann distribution for the oscillator $P\left(p,q\right) \propto \exp -\beta\left(\frac{1}{2}\omega^2q^2 + \frac{1}{2}p^2\right)$ is stationary.

Working out the time-dependent solution without using stochastic calculus techniques is very tedious, so we will just state the results for the matrix generalization of the Langevin equation – a Ornstein-Uhlenbeck process

$$\dot{\mathbf{u}} = -\mathbf{A}\mathbf{u} + \mathbf{B}\boldsymbol{\xi}, \tag{2.13}$$

and the associated Fokker-Planck equation is

$$\frac{\partial}{\partial t}P\left(\mathbf{u},t|\mathbf{u}_0,0\right) = \sum_{ij}A_{ij}\frac{\partial}{\partial u_i}\left[u_jP\left(\mathbf{u},t|\mathbf{u}_0,0\right)\right] + \frac{1}{2}\sum_{ij}D_{ij}\frac{\partial^2}{\partial u_i\partial u_j}P\left(\mathbf{u},t|\mathbf{u}_0,0\right), \tag{2.14}$$

where $\mathbf{D} = \mathbf{B}\mathbf{B}^T$. The finite-time propagator can be shown to be [3]:

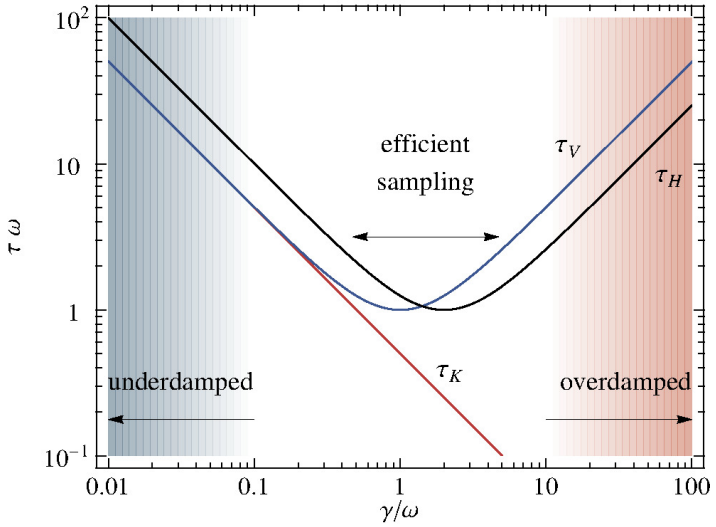
$$\mathbf{u}(t) = e^{-t\mathbf{A}}\mathbf{u}(0) + \sqrt{\mathbf{C} - e^{-t\mathbf{A}}\mathbf{C}e^{-t\mathbf{A}^T}}\boldsymbol{\xi}, \qquad \langle \xi_i \rangle = 0, \langle \xi_i \xi_j \rangle = \delta_{ij}, \tag{2.15}$$

where $\mathbf{C}$ is the static covariance matrix ($\mathbf{C} = \langle \mathbf{u}\mathbf{u}^T \rangle$) that satisfies $\mathbf{A}\mathbf{C} + \mathbf{C}\mathbf{A}^T = \mathbf{B}\mathbf{B}^T$.

Sampling efficiency for a harmonic oscillator Having an analytical expression 2.15 for the finite-time evolution of the stochastic differential equation for a Langevin harmonic oscillator means that in principle one can compute any static or dynamic quantity describing the stochastic dynamics. In particular, it is possible to evaluate autocorrelation times of different observables as a function of the frequency of the oscillator ω and the friction γ – so that the impact of the Langevin term on the ergodicity of sampling can be assessed quantitatively. In particular, one can get the autocorrelation time for the potential V, the kinetic energy K and the total energy H:

$$\tau_V = \frac{1}{\gamma} + \frac{\gamma}{\omega^2}, \quad \tau_K = \frac{1}{\gamma}, \quad \tau_H = \frac{2}{\gamma} + \frac{\gamma}{2\omega^2}. \tag{2.16}$$

Figure 7: Autocorrelation time for different observables for a harmonic oscillator of frequency ω, as a function of the Langevin friction γ. Both the friction and the autocorrelation times are expressed in terms of the intrinsic time scale of the oscillator.



Apart from the kinetic energy, for any value of γ these quantities grow as $1/\omega^2$ – which is reasonable since $1/\omega$ corresponds to a characteristic time scale for the dynamics of the oscillator. It is therefore more convenient to assess the efficiency using the a-dimensional quantity $\kappa = 2/\omega\tau$:

$$\kappa_V = 2\left(\frac{\omega}{\gamma} + \frac{\gamma}{\omega}\right)^{-1}, \quad \kappa_K = 2\frac{\gamma}{\omega}, \quad \kappa_H = 2\left(\frac{2\omega}{\gamma} + \frac{\gamma}{2\omega}\right)^{-1}.$$

As shown in Figure 7, there is an optimal range of frictions close to critical damping $\gamma = \omega$ for which the correlation time is minimum, and the dynamics is most ergodic. Lower values of the friction would yield under-damped dynamics, with the oscillator going back and forth with very slow changes in amplitude. Higher values lead to an over-damped regime, in which the dynamics becomes sluggish, and configuration space exploration

is greatly slowed down. Figure 7 can also be read keeping the friction constant and varying the frequency: oscillators with frequency much different from γ would be samples sub-optimally. This poses a problem when one wants to apply a Langevin thermostat to a real system, in which many different time scales will be present at the same time: one would need to use a different friction for each normal mode in the system, in order to obtain the most ergodic sampling.

2.4 Colored-noise generalized Langevin dynamics

Applying a different white-noise Langevin thermostat to different molecular coordinates require knowing the normal modes of the system, which is often impractical and computationally demanding. The question is therefore whether it is possible to obtain a thermostatting technique that *automatically* adapts to the different time scales, and enforces ergodic sampling on all of them. One possibility is to generalize Langevin dynamics by making it *non-Markovian*, i.e. by allowing a degree of memory of the past history of the dynamics to influence the evolution of the trajectories. Such an equation could be written (for a one-dimensional system and using mass-scaled coordinates)

$$\begin{aligned} \dot{q} =& p \\ \dot{p} =& -V'(q) - \int_{-\infty}^{t} K(t-s)\, p(s)\, \mathrm{d}s + \zeta, \end{aligned} \tag{2.17}$$

where $K(t)$ is a memory kernel describing the friction, and $\zeta(t)$ is a Gaussian random process whose time correlation function is $H(t) = \langle \zeta(t)\, \zeta(0) \rangle$. Analytical treatment of this kind of stochastic differential equations is even more complex than the Markovian case. However, it is possible to map this history-dependent dynamics onto a Markovian dynamics in an extended phase space. We supplement the dynamical variables (q, p) by a set of n additional momenta $\mathbf{s}$, which will be bilinearly coupled to the physical momentum p, so as to construct a Markovian Langevin equation:

$$\begin{aligned} \dot{q} =& p \\ \begin{pmatrix} \dot{p} \\ \dot{\mathbf{s}} \end{pmatrix} =& \begin{pmatrix} -V'(q) \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} a_{pp} & \mathbf{a}_p^T \\ \bar{\mathbf{a}}_p & \mathbf{A} \end{pmatrix} \begin{pmatrix} p \\ \mathbf{s} \end{pmatrix} + \begin{pmatrix} b_{pp} & \mathbf{b}_p^T \\ \bar{\mathbf{b}}_p & \mathbf{B} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi} \end{pmatrix}, \end{aligned} \tag{2.18}$$

where $\boldsymbol{\xi}$ is a vector of $n+1$, uncorrelated Gaussian numbers, i.e. $\langle \xi_i(t)\, \xi_j(0) \rangle = \delta_{ij} \delta(t)$.

Using a Mori-Zwanzig [15] formalism, one can show that the trajectories generated by (2.18) are equivalent to those generated by (2.17), with the memory kernel $K(t) = 2a_{pp}\delta(t) - \mathbf{a}_p^T e^{-|t|\mathbf{A}} \bar{\mathbf{a}}_p$. The details of this formalism are well beyond the scope of the present introduction. Figure (8) shows a simplified picture of how a non-Markovian dynamics can be represented by a Markovian dynamics in an extended phase space.

The crucial aspect of this generalized Langevin equation formalism is that Eqs. (2.18) closely resemble the Ornstein-Uhlenbeck process (2.13), and actually in the case of a harmonic potential represent precisely an Ornstein-Uhlenbeck process that can be solved analytically. One can compute sampling efficiencies $\kappa(\omega)$ as a function of the many parameters of the stochastic dynamics, that can be optimized to yield constant and high sampling efficiency over a range of frequencies that encompasses all of the frequencies that are relevant for the system being studied. Figure (9) demonstrates how the GLE parameters can be optimized to give efficient sampling for many different time scales at once – which is crucial whenever one needs to extract as much statistics as possible from a simulation of limited length [16, 17].

Integrating the (generalized) Langevin equation Despite their apparent complexity, integrating Eqs. (2.18) is not difficult. The idea is to modify the velocity Verlet

Figure 8: A schematic representation of how a Markovian trajectory in an extended phase space can represent a non-Markovian trajectory in a reduced-dimensional space. The two trajectories in the (q, p) subspace cannot clearly be obtained from a Markovian formulation, since they cross at (q_0, p_0): since a Markovian deterministic dynamics is uniquely determined by the starting conditions, there can be just one trajectory evolving forward in time from any given point of phase space. However, if the trajectories are considered as projections of a higher-dimensional dynamics, the two paths could actually correspond to different values of the additional parameter s, so that they could well be generated by a Markovian, deterministic dynamics in this (q, p, s) space.

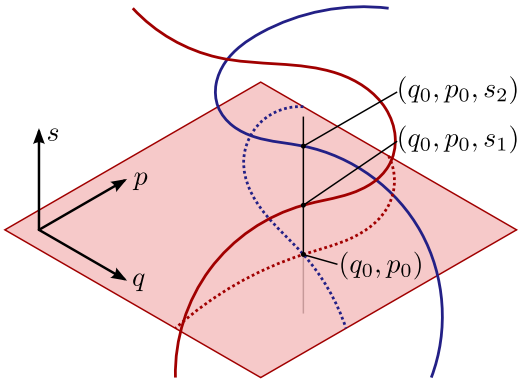
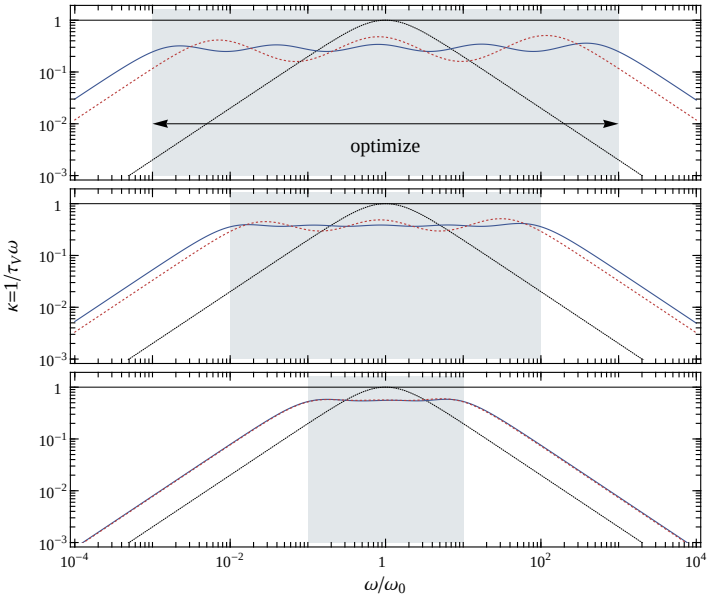


Figure 9: Sampling efficiency for a harmonic oscillator subject to a colored-noise thermostat that has been optimized to yield large, constant sampling efficiency over a broad range of frequencies. The panels, from bottom to top, contain the results fitted over frequency ranges spanning two, four and six orders of magnitude around $\omega = 1$ respectively. Blue, continuous lines correspond to matrices with $n = 4$, while the red, dashed lines are for $n = 2$.The curve for a white-noise Langevin thermostat is shown as reference, in black.



integrator (1.8) to include a propagator of the free-particle version of (2.18). This approach can be derived with a Trotter splitting of the Liouville operator, and preserves the time-reversal symmetry of velocity Verlet

$$\begin{aligned} p_{i\alpha} \leftarrow & \mathcal{P}\left[\left(p_{i\alpha}, \mathbf{s}_{i\alpha}\right), dt/2\right] \\ \mathbf{p} \leftarrow & \mathbf{p} - \frac{\partial V}{\partial \mathbf{q}} \frac{dt}{2} \\ \mathbf{q} \leftarrow & \mathbf{q} + \frac{\mathbf{p}}{m} dt \\ \mathbf{p} \leftarrow & \mathbf{p} - \frac{\partial V}{\partial \mathbf{q}} \frac{dt}{2}. \\ p_{i\alpha} \leftarrow & \mathcal{P}\left[\left(p_{i\alpha}, \mathbf{s}_{i\alpha}\right), dt/2\right] \end{aligned} \tag{2.19}$$

The exact finite-time propagator is applied to individual components of the momentum separately – there is a set of additional momenta $\mathbf{s}_{i\alpha}$ for each momentum component $p_{i\alpha}$ – and it corresponds to the propagator derived for the Ornstein-Uhlenbeck process (2.15):

$$\mathcal{P}\left[\left(p, \mathbf{s}\right), dt\right]^T = \mathbf{T}(dt)\left(p, \mathbf{s}\right)^T + \sqrt{m}\mathbf{S}(dt)\boldsymbol{\xi}^T \tag{2.20}$$

where $\boldsymbol{\xi}$ is a vector of $n+1$ uncorrelated Gaussian numbers, and the matrices $\mathbf{T}$ and $\mathbf{S}$ can be computed once, at the beginning of the simulation and for all degrees of freedom. The relations between $\mathbf{T}$, $\mathbf{S}$, and the inverse temperature $\beta = 1/k_B T$ is

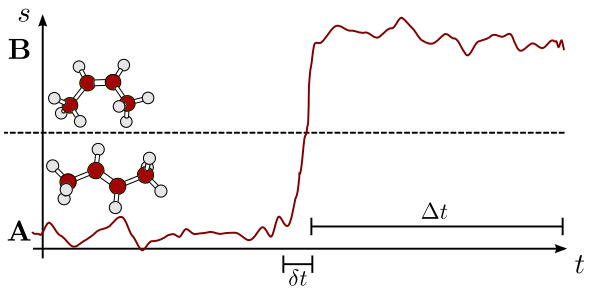
$$\mathbf{T} = e^{-dt\mathbf{A}_p}, \qquad \mathbf{S}\mathbf{S}^T = \frac{1}{\beta}\left[\mathbf{1} - e^{-dt\mathbf{A}_p} e^{-dt\mathbf{A}_p^T}\right] \qquad \mathbf{A}_p = \begin{pmatrix} a_{pp} & \mathbf{a}_p^T \\ \bar{\mathbf{a}}_p & \mathbf{A} \end{pmatrix}.$$

3 Rare events and transition state theory

It is often the case that atomistic simulations have to deal with systems which can exist in multiple (meta)stable states. This is for instance the case of compounds undergoing a chemical reaction, or materials whose functioning depends on the transition between different phases. This is a scenario that is particularly challenging to both molecular dynamics and Monte Carlo methods, because the stability of the various configurations means that they are long-lived, and that transitions between them happens only rarely.

A typical situation is schematically represented in Figure 10, that also explains why rare events are hard to deal with: the time scale of the transition δt is fast (and the time step needed to integrate Hamilton's equations is typically even shorter), but the time Δt one has to wait between transitions, observing uninteresting thermal fluctuations, can be several orders of magnitude longer. Often it is just impossible to extract information on the time scale of a transition by waiting for it to happen in a molecular dynamics trajectory, and one has to resort to different techniques that are based on a equilibrium picture of the reactive event.

Figure 10: Schematic representation of the time evolution of a molecule that can exist in two meta-stable states. The molecule spends most of the time oscillating around one of the stable conformations, and only rarely a transition between the two states take place. The ratio between the "waiting" time Δt and the time it takes to complete a successful transition δt can be huge.



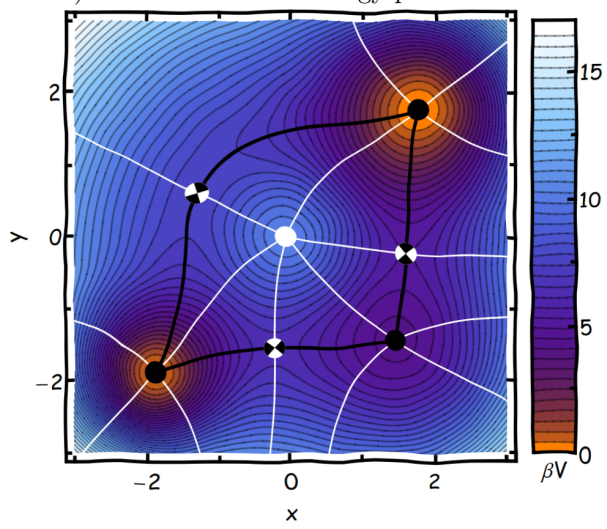
3.1 Potential energy landscapes

A first, essential task when studying transitions between stable configurations of a system is to identify the regions of configuration space that can be considered to be representative of the different meta-stable states. In many cases – particularly for the simpler problems – it is possible to partition configuration space based on the underlying potential energy function $V(\mathbf{q})$ Wales [18]. This approach is reasonable, because the Boltzmann distribution $e^{-\beta V(\mathbf{q})}$ will be peaked in regions with low values of the potential energy, and so knowing the minima of $V(\mathbf{q})$ can give insight on what are the most highly populated conformations of a molecule, and the most probable paths that might be taken to transform from one state to another.

One can characterize points on the potential energy surface (PES) based on a Taylor expansion around each point $\mathbf{q}_0$:

$$V(\mathbf{q}) \approx V(\mathbf{q}_0) + \mathbf{g}(\mathbf{q}_0) \cdot (\mathbf{q} - \mathbf{q}_0) + \frac{1}{2} (\mathbf{q} - \mathbf{q}_0)^T \mathbf{H}(\mathbf{q}_0) (\mathbf{q} - \mathbf{q}_0) + \mathcal{O}\left(|\mathbf{q} - \mathbf{q}_0|^3\right),$$

Figure 11: A cartoon of a 2D potential energy surface, with the critical points marked as large dots. Minima are marked as black dots, maxima as white dots, and first-order saddle points as checkered dots. The figure also displays a few steepest descent paths, three of which (in black) are also minimum energy paths.



where $g_i(\mathbf{q}) = \partial V(\mathbf{q})/\partial q_i$ is the gradient and $H_{ij}(\mathbf{q}) = \partial^2 V(\mathbf{q})/\partial q_i \partial q_j$ is the Hessian matrix. Critical points on the PES are those with $|\mathbf{g}| = 0$, and can be classified based on the eigenvalues of $\mathbf{H}$: points for which all the eigenvalues are positive are minima – that will typically correspond to maxima in the Boltzmann distribution, and hence to stable configurations – points for which all the eigenvalues are negative are local maxima, and points for which all the eigenvalues are positive except for m negative ones are saddle points of order m. Saddle points of order 1 are particularly important, as they often correspond to the point of least probability along a path connecting two minima, and play a crucial role in determining the rate at which transitions between stable states can happen.

Another important concept in analyzing the PES is that of steepest descent paths – trajectories that join any point $\mathbf{q}$ to a local minimum that are at each point tangent to the gradient. The steepest descent path that starts from a point $\mathbf{q}_0$ satisfies the conditions

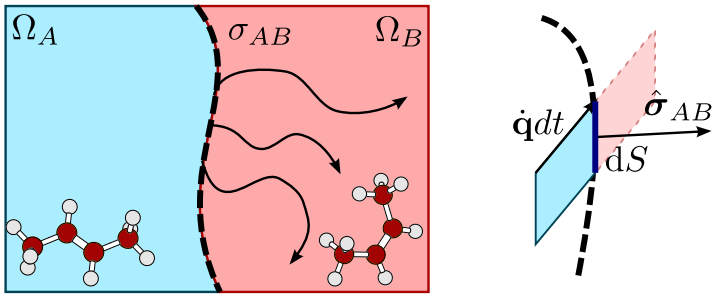
$$\mathbf{r}_{\mathbf{q}_0}(0) = \mathbf{q}_0, \qquad \dot{\mathbf{r}}_{\mathbf{q}_0}(s) = -\mathbf{g}\left(\mathbf{r}_{\mathbf{q}_0}(s)\right) / \left|\mathbf{g}\left(\mathbf{r}_{\mathbf{q}_0}(s)\right)\right|.$$

Each steepest-descent path eventually arrives at a local minimum, and so steepest descent paths provide a natural criterion to partition configuration space in a set of adjacent regions: each local minimum $\bar{\mathbf{q}}$ corresponds to a region $\Omega_{\bar{\mathbf{q}}}$ that contains all the points $\mathbf{q}_0$ whose steepest descent path converges onto $\bar{\mathbf{q}}$: $\Omega_{\bar{\mathbf{q}}} = \{\mathbf{q} : \lim_{s\to\infty} r_{\mathbf{q}}(s) = \bar{\mathbf{q}}\}$. Steepest descent paths that join a first-order saddle point to a minimum are also known as minimum energy paths, and are thought to be representative of the sequence of configurations that correspond to a reaction pathways, since they are at all points the configurations with maximum probability in all directions apart from the direction leading to the saddle point dividing the two regions of configuration space (see Figure 11).

3.2 Transition state theory

Having given an idea of how one could use the potential energy surface to identify regions of configuration space that correspond to meta-stable conformations of the system being studied, let us see how one can obtain information on the time scale of the transformation

Figure 12: A simplified representation of the partitioning of configuration space that is used to define the transition state theory. The right-hand side explains how the reactive flux can be defined in terms of an integral over the dividing surface, within the transition state approximation.



between a reactants state A and a products state B based (almost) exclusively on equilibrium, time-independent information. Let us define the characteristic function of the reactants region

$$\theta_A(\mathbf{q}) = \begin{cases} 1 & \mathbf{q} \in \Omega_A \\ 0 & \mathbf{q} \notin \Omega_A \end{cases}$$

The microscopic transition rate k_{AB} for the $A \to B$ reaction can be defined based on the equilibrium concentration of the reactants x_A and the probability that a configuration starting in A will be found outside of A in an infinitesimal time interval dt, $P_{AB}(dt) = k_{AB} x_A dt$ (see Figure 12). The equilibrium concentration of the reactants can be defined as a configurational ensemble average

$$x_A = \langle \theta_A(\mathbf{q}) \rangle = \frac{\int \mathrm{d}\mathbf{q}\, \theta_A(\mathbf{q})\, e^{-\beta V(\mathbf{q})}}{\int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})}},$$

while the transition probability can be obtained by considering the trajectories starting close enough to the dividing surface σ_{AB} and having a velocity with the appropriate orientation relative to the surface normal $\hat{\boldsymbol{\sigma}}_{AB}(\mathbf{q})$:

$$P_{AB}^{\mathrm{TST}}(dt) = \frac{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int_{\sigma_{AB}} \mathrm{d}S\, e^{-\beta V(\mathbf{q})} \left[\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q}) \mathbf{M}^{-1}\mathbf{p} dt\right] \theta\left(\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q}) \mathbf{M}^{-1}\mathbf{p}\right)}{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})}},$$

where $\mathbf{M}$ is a diagonal matrix having the masses corresponding to individual degrees of freedom on the diagonal, and $\theta(x)$ is the Heaviside step function, and is meant to select trajectories that are oriented from A to B (see Figure 12). Note that here we are assuming that all the trajectories that cross the dividing surface in the infinitesimal time dt and are oriented towards the products will give rise to a long-lived product specie. This assumption gives rise to a transition state theory expression of the rate, that combines the definition of x_A and P_{AB} to give a definition of the rate in terms of ensemble averages.

$$k_{AB}^{\mathrm{TST}} = \frac{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int_{\sigma_{AB}} \mathrm{d}S\, e^{-\beta V(\mathbf{q})} \left[\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q}) \mathbf{M}^{-1}\mathbf{p}\right] \theta\left(\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q}) \mathbf{M}^{-1}\mathbf{p}\right)}{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int \mathrm{d}\mathbf{q}\, \theta_A(\mathbf{q})\, e^{-\beta V(\mathbf{q})}} \tag{3.1}$$

This expression can be simplified further by integrating out the $\mathbf{p}$ dependence, computing

$$\frac{\int \mathrm{d}\mathbf{p} e^{-\beta \frac{1}{2}\mathbf{p}^T \mathbf{M}^{-1}\mathbf{p}} \left[\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q}) \mathbf{M}^{-1}\mathbf{p}\right] \theta\left(\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q}) \mathbf{M}^{-1}\mathbf{p}\right)}{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})}}.$$

One first goes into mass-scaled coordinates $\tilde{\mathbf{p}} \leftarrow \mathbf{M}^{-1/2}\mathbf{p}$to get

$$\frac{\sqrt{\det \mathbf{M}} \int \mathrm{d}\tilde{\mathbf{p}} e^{-\beta \frac{1}{2}\tilde{\mathbf{p}}^T\tilde{\mathbf{p}}} \left[\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}\tilde{\mathbf{p}}\right] \theta\left(\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}\tilde{\mathbf{p}}\right)}{\sqrt{\det \mathbf{M}} \int \mathrm{d}\tilde{\mathbf{p}} e^{-\beta \frac{1}{2}\tilde{\mathbf{p}}^T\tilde{\mathbf{p}}}};$$

then, one has to realize that $e^{-\beta \frac{1}{2}\tilde{\mathbf{p}}^T\tilde{\mathbf{p}}}$ is spherically symmetric, that $\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}$ is just a vector of length $\left|\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}\right|$ along which the scaled momentum is being projected. So, one can integrate out the $3N-1$ orthogonal directions (that cancel out with the denominator) and just consider explicitly the integrals along the direction of $\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}$:

$$\left|\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}\right| \int_0^\infty \tilde{p} e^{-\frac{1}{2}\beta\tilde{p}^2} \mathrm{d}\tilde{p} \Big/ \int_0^\infty \tilde{p} e^{-\frac{1}{2}\beta\tilde{p}^2} \mathrm{d}\tilde{p} = \sqrt{\frac{1}{2\pi\beta}} \left|\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}\right|,$$

which leaves the configurational average

$$k_{AB}^{\mathrm{TST}} = \sqrt{\frac{1}{2\pi\beta}} \frac{\int_{\sigma_{AB}} \mathrm{d}S\, e^{-\beta V(\mathbf{q})} \left|\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}\right|}{\int \mathrm{d}\mathbf{q}\, \theta_A(\mathbf{q})\, e^{-\beta V(\mathbf{q})}} \tag{3.2}$$

Harmonic transition state theory The transition state theory rate (3.1) is not very useful without a working definition of the reactants region and the dividing surface. One could use an analysis of the PES to define the region assigned to the minimum corresponding to the reactants, along the lines discussed above, and also define the dividing surface σ_{AB} as the (hyper)surface that passes through the saddle point between the two minima and is orthogonal to the minimum energy path. It would be very hard to parameterize this hypersurface, not to mention the complexity of evaluating high-dimensional integrals that was the original motivation for using importance sampling or molecular dynamics. However, one can work out a simple expression for k_{AB}^{TST} by performing a harmonic expansion of the potential around the minimum A, and around the transition state – assuming that there is just one saddle point along the dividing surface between the reactants and products. The denominator is easily evaluated by expanding $V(\mathbf{q}) \approx V_A + \frac{1}{2}(\mathbf{q}-\mathbf{q}_A)^T \mathbf{H}_A (\mathbf{q}-\mathbf{q}_A)$, and assuming that the Boltzmann probability for this approximate quadratic form vanishes outside Ω_A:

$$\int \mathrm{d}\mathbf{q}\, \theta_A(\mathbf{q})\, e^{-\beta V(\mathbf{q})} \approx e^{-\beta V_A} \int \mathrm{d}\mathbf{q}\, e^{-\beta \frac{1}{2}\mathbf{q}^T\mathbf{H}_A\mathbf{q}} = \frac{e^{-\beta V_A}}{\sqrt{\det \mathbf{M}}} \int \mathrm{d}\tilde{\mathbf{q}}\, e^{-\beta \frac{1}{2}\tilde{\mathbf{q}}^T\mathbf{D}_A\tilde{\mathbf{q}}},$$

where we have changed the reference frame to be centered around $\mathbf{q}_A$, have transformed to mass-scaled coordinates and have introduced the dynamical matrix $(\mathbf{D}_A)_{i\alpha j\alpha'} = (\mathbf{H}_A)_{i\alpha j\alpha'} \Big/ \sqrt{m_i m_j}$. Then, the eigenvalues of $\mathbf{D}_A$ are the squared frequencies of the system's normal modes in the A minimum, $\left(\omega_i^A\right)^2$, and the integral yields $(2\pi/\beta)^{3N/2} \Big/ \prod_{k=1}^{3N} \omega_k^A$.

The integral over the dividing surface can be evaluated in a similar way, noting that the dividing surface can be approximated as the hyperplane that is orthogonal to the single negative eigenvalue of the Hessian matrix $\mathbf{H}_{AB}$ at the saddle point. Transforming in mass-scaled coordinates, expanding in normal modes coordinates and integrating over the whole space while excluding the negative eigenvector direction one gets

$$\begin{aligned}\int_{\sigma_{AB}} \mathrm{d}S\, e^{-\beta V(\mathbf{q})} \left|\hat{\boldsymbol{\sigma}}_{AB}^T(\mathbf{q})\mathbf{M}^{-1/2}\right| \approx & \frac{e^{-\beta V_{AB}}}{\sqrt{\det \mathbf{M}}} \int_{\sigma_{AB}} \mathrm{d}\tilde{S}\, e^{-\beta \frac{1}{2}\tilde{\mathbf{q}}^T\mathbf{D}_{AB}\tilde{\mathbf{q}}} = \\ = & \frac{e^{-\beta V_{AB}}}{\sqrt{\det \mathbf{M}}} \left(\frac{2\pi}{\beta}\right)^{(3N-1)/2} \frac{1}{\prod_{k=1}^{3N-1} \omega_k^{AB}}.\end{aligned}$$

Combining all the pieces together, one gets the familiar expression for the reaction rate in terms of an exponential term that depends on the energy barrier and a pre-factor with the units of a frequency:

$$k_{AB}^{\mathrm{hTST}} = \frac{1}{2\pi} e^{-\beta(V_{AB}-V_A)} \frac{\prod_{k=1}^{3N} \omega_k^A}{\prod_{k=1}^{3N-1} \omega_k^{AB}} = \nu^\star e^{-\beta \Delta V} \tag{3.3}$$

3.3 Collective variables and free energy surface

At times it is just not practical to perform a thorough investigation of the potential energy surface, when studying disordered systems, liquids, or reactions taking place in these disordered environment. In these cases there can be an enormous number of shallow local minima, which are nearly impossible to map, and which make it hard to identify properly what is the reactant region or the transition state. In these cases, it is useful to introduce a coarse-grained description of the system in terms of one or more *collective variables*, order parameters that are meant to differentiate between stable states and reaction pathways. Simple collective variables to describe reactive events might be bond lengths, coordination numbers, order parameters that can differentiate between ordered and disordered environments, etc. Having defined a set of collective variables $\mathbf{s}(\mathbf{q})$, it is customary to introduce the concept of a *free energy*

$$F(\mathbf{s}) = -\frac{1}{\beta} \ln \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} \delta(\mathbf{s} - \mathbf{s}(\mathbf{q})). \tag{3.4}$$

The free energy is to the collective coordinates $\mathbf{s}$ what the potential is to the position $\mathbf{q}$, as the canonical probability density of finding a configuration which takes the collective variables value $\mathbf{s}$ is $P(\mathbf{s}) \propto e^{-\beta F(\mathbf{s})}$.

For simplicity, here we will discuss the case where one can find a single collective variable $s(\mathbf{q})$ that can distinguish between reactants and products for a reaction of interest, i.e. $s(\mathbf{q}) = 0$ gives the dividing surface, $s(\mathbf{q}) > 0$ identifies the products and $s(\mathbf{q}) < 0$ the reactants. In this context, the transition-state rate (3.1) can be written as

$$k_{AB}^{\mathrm{TST}} = \frac{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} \delta(s(\mathbf{q})) \left[\nabla s(\mathbf{q})^T \mathbf{M}^{-1} \mathbf{p}\right] \theta\left(\nabla s(\mathbf{q})^T \mathbf{M}^{-1} \mathbf{p}\right)}{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} \theta(-s(\mathbf{q}))}$$

where θ is the Heaviside step function. Let us show an alternative treatment of the momentum integral, based on a Fourier representation of the Heaviside function

$$\theta(x) = \frac{1}{2} - P \int_{-\infty}^{\infty} \frac{e^{-\mathrm{i}xt}}{2\pi \mathrm{i} t} \mathrm{d}t.$$

The constant term in the Fourier representation of θ integrates to zero because $K(\mathbf{p})$ is even in $\mathbf{p}$ and $\left[\nabla s(\mathbf{q})^T \mathbf{M}^{-1} \mathbf{p}\right]$ is odd. We are left with

$$-P \int_{-\infty}^{\infty} \frac{1}{2\pi \mathrm{i} t} \int e^{-\mathrm{i}t\left[\nabla s(\mathbf{q})^T \mathbf{M}^{-1} \mathbf{p}\right]} e^{-\beta \frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}} \left[\nabla s(\mathbf{q})^T \mathbf{M}^{-1} \mathbf{p}\right] \mathrm{d}\mathbf{p} \mathrm{d}t.$$

The integral in $\mathbf{p}$ reads can be written by spelling out the scalar product between $\nabla s(\mathbf{q})$ and $\dot{\mathbf{q}}$:

$$\sum_{i\alpha} \frac{\nabla_{i\alpha} s(\mathbf{q})}{m_i} \int p_{i\alpha} e^{-\mathrm{i}t \frac{\nabla_{i\alpha} s(\mathbf{q})}{m_i} p_{i\alpha}} e^{-\beta \frac{p_{i\alpha}^2}{2m_i}} \mathrm{d}p_{i\alpha} \prod_{j\alpha' \neq i\alpha} \int e^{\mathrm{i}t \frac{\nabla_{j\alpha'} s(\mathbf{q})}{m_j} p_{j\alpha'}} e^{-\beta \frac{p_{j\alpha'}^2}{2m_j}} \mathrm{d}p_{j\alpha'} =$$

$$= \sum_{i\alpha} \frac{-\mathrm{i}t}{\beta m_i} \sqrt{\frac{2\pi m_i}{\beta}} \nabla_{i\alpha} s(\mathbf{q})^2 e^{-\frac{\nabla_{i\alpha} s(\mathbf{q})^2 t^2}{2\beta m_i}} \prod_{j\alpha' \neq i\alpha} \sqrt{\frac{2\pi m_j}{\beta}} e^{-\frac{\nabla_{j\alpha'} s(\mathbf{q})^2 t^2}{2\beta m_j}}.$$

where we have completed the square in the integral in $\mathrm{d}p_{i\alpha}$. Let us drop the $\sqrt{2\pi m_j/\beta}$ terms, as they are canceled by the $\mathbf{p}$ integral at the denominator. We are left to compute a term that corresponds to the flux through the dividing surface at $\mathbf{q}$, that accounts also for the distortion of the metric that arises because of the use of s to define the different regions of the system

$$\frac{-1}{2\pi \mathrm{i}t}\frac{-\mathrm{i}t}{\beta m_i}\nabla s\left(\mathbf{q}\right)^T\mathbf{M}^{-1}\nabla s\left(\mathbf{q}\right)\int_{-\infty}^{\infty}e^{-\frac{t^2}{2\beta}\nabla s(\mathbf{q})^T\mathbf{M}^{-1}\nabla s(\mathbf{q})}\mathrm{d}t=\sqrt{\frac{\nabla s\left(\mathbf{q}\right)^T\mathbf{M}^{-1}\nabla s\left(\mathbf{q}\right)}{2\pi\beta}}=\phi\left(\mathbf{q}\right).$$

This leaves an expression equivalent to (3.2):

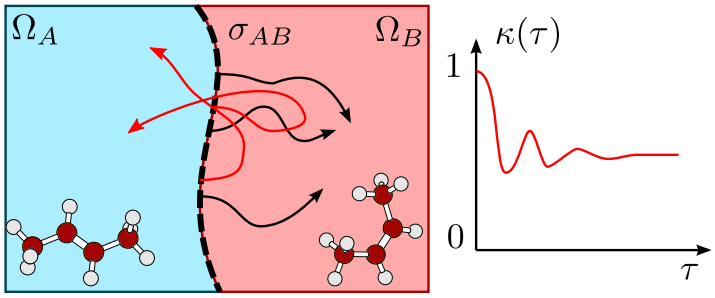
$$k_{AB}^{\mathrm{TST}}\left(t\right)=\frac{\int\mathrm{d}\mathbf{q}e^{-\beta V(\mathbf{q})}\delta\left(s\left(\mathbf{q}\right)\right)\phi\left(\mathbf{q}\right)}{\int\mathrm{d}\mathbf{q}e^{-\beta V(\mathbf{q})}\theta\left(-s\left(\mathbf{q}\right)\right)},\tag{3.5}$$

which only involves equilibrium averages. This can also be rewritten in a form that shows more clearly the definition in terms of the free energy (3.4):

$$k_{AB}^{\mathrm{TST}}\left(t\right)=\frac{e^{-\beta F(0)}}{\int\mathrm{d}s\,e^{-\beta F(s)}\theta\left(-s\right)}\phi\left(s\right),\qquad\phi\left(s\right)=\frac{\int\mathrm{d}\mathbf{q}e^{-\beta V(\mathbf{q})}\delta\left(s\left(\mathbf{q}\right)\right)\phi\left(\mathbf{q}\right)}{\int\mathrm{d}\mathbf{q}e^{-\beta V(\mathbf{q})}\delta\left(s\left(\mathbf{q}\right)\right)}.\tag{3.6}$$

The crucial step to evaluate the rate is therefore to define an effective collective coordinate and to compute the free energy surface relative to it. The flux is just the mean value of $\phi\left(\mathbf{q}\right)$ at the transition state, and can be evaluated by averaging the values of $\phi\left(\mathbf{q}\right)$ for the configurations observed during the trajectory that have $\left|s\left(\mathbf{q}\right)\right|<\epsilon$.

Figure 13: Trajectories that re-cross the dividing surface reduce the overall reaction rate, and are not accounted for within transition state theory. Bennett-Chandler rate theory allows to express the correction in terms of a transmission coefficient, that starts off at 1 (the transition-state theory is the $\tau\to 0^+$ limit of Bennett-Chandler rate theory) and levels off to a plateau value after an intermediate transient.



3.4 Bennett-Chandler rate theory

In introducing transition state theory we have observed that an assumption was being made that all the trajectories starting at the A/B dividing surface, with the velocity oriented towards the products, would have ended up in a successful reactive event, i.e. they would have stayed in B for a very long time (Figure 13). This is usually a good assumption for simple systems that can be treated within the harmonic transition state theory approximation, but it can fail dramatically for more complex systems, and when the transition is described in terms of imperfect collective variables. In these cases, trajectories may re-cross the dividing surface and go back to the reactants region: one should define

a time-dependent rate that describes the probability that a trajectory initiated at the dividing surface will be in the product region at a later time τ:

$$k_{AB}(\tau) = \frac{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} \delta(s(\mathbf{q})) \left[\nabla s(\mathbf{q})^T \mathbf{M}^{-1}\mathbf{p}\right] \theta(\mathcal{P}_{\mathbf{q}}((\mathbf{p},\mathbf{q}),\tau))}{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} \theta(-s(\mathbf{q}))}.$$

It is actually convenient to re-write this expression as $k_{AB}(\tau) = k_{AB}^{\mathrm{TST}} \kappa(\tau)$, the product of the transition-state theory rate (3.5) and the transmission coefficient

$$\kappa(\tau) = \frac{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} \delta(s(\mathbf{q})) \left[\nabla s(\mathbf{q})^T \mathbf{M}^{-1}\mathbf{p}\right] \theta(\mathcal{P}_{\mathbf{q}}((\mathbf{p},\mathbf{q}),\tau))}{\int \mathrm{d}\mathbf{p} e^{-\beta K(\mathbf{p})} \int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})} \delta(s(\mathbf{q})) \phi(\mathbf{q})}. \quad (3.7)$$

In fact, Eq. (3.7) can be easily computed by choosing n starting configurations $\mathbf{q}_0^{(i)}$ among those that are within ϵ of the dividing surface (these are typically obtained in the process of computing the transition-state rate), and *shooting* m short trajectories out of those starting configurations, by generating random momenta $\mathbf{p}_0^{(j)}$ in the forward direction, and counting the fraction of trajectories that are in the product region at each time τ:

$$\kappa(\tau) \approx \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\nabla s\left(\mathbf{q}_0^{(i)}\right)^T \mathbf{M}^{-1}\mathbf{p}_0^{(j)}\right] \theta\left(\mathcal{P}_{\mathbf{q}}\left(\left(\mathbf{p}_0^{(j)}, \mathbf{q}_0^{(i)}\right), \tau\right)\right).$$

Partitioning the rate in these two components makes it possible to perform enhanced sampling techniques to evaluate k_{AB}^{TST} and to collect initial configurations from the transition state region, and then perform only short constant-energy trajectories to compute the correction due to recrossings. The typical behavior of the transmission coefficient with time is shown in Figure 13: by construction, $\lim_{\tau\to 0^+} \kappa(\tau) = 1$, and after a transient period the transmission coefficient converges to a plateau value. This is the value that typically should be used to define a macroscopic rate. In principle, the Bennett-Chandler formalism should give the same overall plateau value of the rate constant independent on the precise choice of the dividing surface, since the transmission coefficient corrects for the recrossings due to a poor choice of collective variable to describe the transition. In practice, evaluating $\kappa(\tau)$ accurately gets harder and harder as the plateau value decreases to zero, and so the efficiency of the calculation depends crucially on the choice of collective variables.

Committor analysis The Bennett-Chandler formalism also provides a very robust method to assess the quality of a collective variable in terms of being able to describe the $A \to B$ transition, by an analysis of the distribution of the committor in the transition state ensemble. The committor function is defined for each configuration $\bar{\mathbf{q}}$ as the fraction of trajectories that start at $\bar{\mathbf{q}}$ and end up in the product region after a time interval τ:

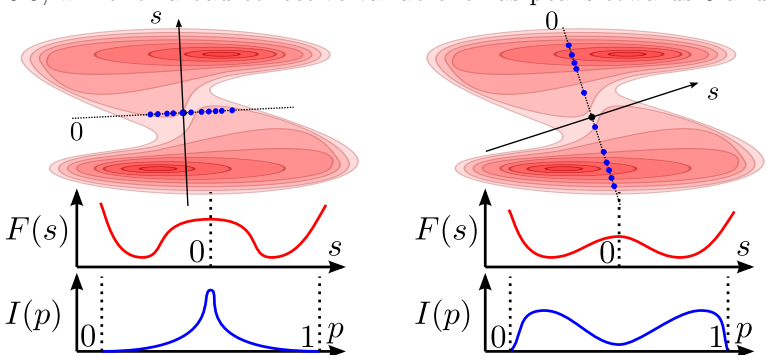
$$p_B(\bar{\mathbf{q}}, \tau) = \frac{\langle \delta(\mathbf{q} - \bar{\mathbf{q}}) \theta_B(\mathcal{P}_{\mathbf{q}}((\mathbf{p},\mathbf{q}),\tau)) \rangle}{\langle \delta(\mathbf{q} - \bar{\mathbf{q}}) \rangle} \quad (3.8)$$

The committor distribution relative to a collective variable s is then defined as the histogram of the values of the committor for configurations that have the specified value of $s(\mathbf{q})$:

$$I_B(s, p) = \frac{\langle \delta[p - p_B(\mathbf{q},\tau)] \delta(s - s(\mathbf{q})) \rangle}{\langle \delta(s - s(\mathbf{q})) \rangle}.$$

One can then look at the histogram of committors for the transition state ensemble, $I_B(0, p)$ in our one-dimensional collective variable example. Ideally, points at the transition

Figure 14: Schematic representation of a good (left) and bad (right) choice of a collective variable to describe a transition on a two-dimensional potential energy surface. Blue dots represent sample configurations from the transition-state ensemble. With a good collective variable, the histogram of committors $I(p)$ for the transition state ensemble is peaked around $p = 0.5$, while for a bad collective variable it has peaks towards 0 and 1.



state should have a 50% probability to evolve towards the products or to go back into the reactants region, so the histogram of committors should be peaked at 0.5. In fact, the committor (3.8) could in principle be used to define an ideal (but impractical) collective variable, where the transition state ensemble corresponds by construction to the configurations with $p_B(\bar{\mathbf{q}}, \tau) = 0.5$. In the case of a bad collective coordinate (right panel of Figure (14)) points at the transition state ensemble have a large probability of recrossing back into the reactants region, and $I_B(0, p)$ has peaks for $p \approx 0$ and $p \approx 1$. Note that this can happen even though the collective variable does a decent job differentiating reactants and products region, because it fails at describing the transition in the high-potential region in the vicinity of the "true" transition state.

4 Reweighing and biased sampling

The basic sampling methods discussed in Chapter 1, as well as the enhanced thermostatting techniques introduced in Chapter 2 make it possible to explore phase space efficiently in the vicinity of a (meta)stable configuration, but typically do not help much when it comes to the study of rare events. The rate theory introduced in Chapter 3 makes it possible to express dynamical properties such as transition probabilities in terms of ensemble averages, but gives no clue as to how one could evaluate more efficiently averages that depend on configurations that are only visited rarely. One could try to modify the ensemble that is being sampled, i.e. to sample according to a modified probability distribution $\tilde{P}(\mathbf{q})$ in which transition-state configurations have a much more significant weight than they have in the original Boltzmann distribution $P(\mathbf{q})$. The question is then whether one can recover ensemble averages relative to P by sampling performed based on $\tilde{P}$ [19, 20]. It is very easy to see that if one wants to compute the mean of an observable $A(\mathbf{q})$ relative to P, it is sufficient to compute

$$\left\langle A\frac{P}{\tilde{P}}\right\rangle_{\tilde{P}} = \int \mathrm{d}\mathbf{q}\, A(\mathbf{q})\, \frac{P(\mathbf{q})}{\tilde{P}(\mathbf{q})}\tilde{P}(\mathbf{q}) = \int \mathrm{d}\mathbf{q}\, A(\mathbf{q})\, P(\mathbf{q}) = \langle A\rangle_P .$$

Note that a very important point is that in the right-hand side the average is corrected by a weight factor that is the *ratio* between the two probability distributions. This means that in practice one does not need to know the normalization of P and $\tilde{P}$, which would be exceedingly difficult to obtain. In practice, one can just obtain a sequence of configurations $\{\mathbf{q}_i\}$ distributed according to $\tilde{P}$ by importance sampling (be it Monte Carlo or thermostatted molecular dynamics), and compute

$$\langle A\rangle_P = \lim_{M\to\infty} \frac{\sum_i A(\mathbf{q}_i)\, P(\mathbf{q}_i)/\tilde{P}(\mathbf{q}_i)}{\sum_i P(\mathbf{q}_i)/\tilde{P}(\mathbf{q}_i)} = \frac{\left\langle A\,P/\tilde{P}\right\rangle_{\tilde{P}}}{\left\langle P/\tilde{P}\right\rangle_{\tilde{P}}}. \tag{4.1}$$

For instance, one could perform a simulation at a higher temperature (lower inverse temperature $\tilde{\beta}$), at which the system diffuses faster over free energy barriers, and obtain statistics at a the desired inverse temperature β. The expression for the reweighed average,

$$\langle A\rangle_\beta \approx \frac{\sum_i A(\mathbf{q}_i)\, e^{-(\beta-\tilde{\beta})V(\mathbf{q}_i)}}{\sum_i e^{-(\beta-\tilde{\beta})V(\mathbf{q}_i)}}, \tag{4.2}$$

is deceptively simple, as will be discussed in the next section. A very similar expression appears when one samples using a potential energy function $\tilde{V}$, and wants to compute averages relative to V:

$$\langle A\rangle_V \approx \frac{\sum_i A(\mathbf{q}_i)\, e^{-\beta\left(V(\mathbf{q}_i)-\tilde{V}(\mathbf{q}_i)\right)}}{\sum_i e^{-\beta\left(V(\mathbf{q}_i)-\tilde{V}(\mathbf{q}_i)\right)}}.$$

This is useful, for instance, when one wants to assess how observables are changed by small modifications to the model (e.g. changing the partial charge on an atom), or when one samples using an inexpensive potential to obtain truly uncorrelated configurations, and then uses these to compute averages consistent with a more expensive potential.

4.1 Statistics of reweighing

It is quite easy to convince oneself that, although formally correct, Eq. (4.1) cannot be universally applicable. If it were, one could take an ideal gas as a reference and

compute inexpensively averages for complex electronic structure calculations. Intuitively, the problem is that in a paradoxical case like this there would be no single configuration from the ideal gas that resembles a reasonable arrangement of covalently-bound atoms, and so the average would be computed over structures with completely negligible weights in the target ensemble. It is possible to give a much more quantitative spin to this analysis, framing the problem in terms of the statistical efficiency of sampling. Similarly to what we did in Section 1.4, the point is performing a large number of independent simulations and computing the mean the fluctuations of the averages, to evaluate statistical and (possibly) systematic errors.

One should start by considering in the abstract the joint probability distribution of the observable A and the weight $W = P/\tilde{P}$, $p(A, W)$. Then, one assumes that all the samples are perfectly uncorrelated (as time correlation efficiency can be dealt with in terms of the autocorrelation time, as discussed in Section 1.4), and considers

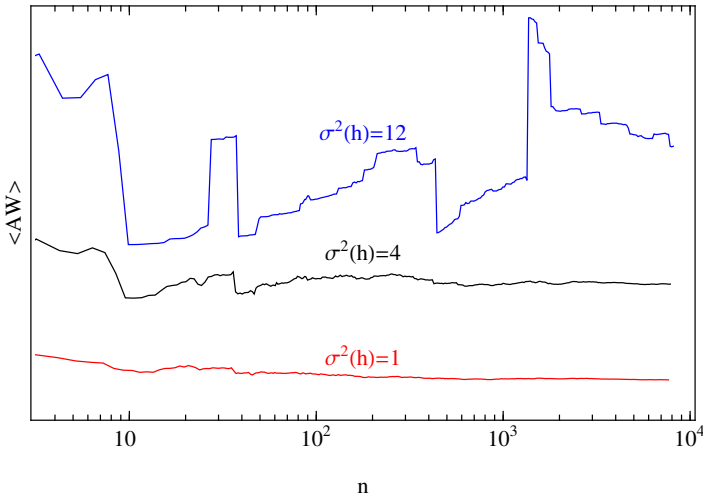
$$\left\langle \bar{A}_n \right\rangle = \left\langle \frac{\sum_i A_i W_i}{\sum_i W_i} \right\rangle = n \left\langle \frac{A\,W}{W + \sum_{i=2}^{n} W_i} \right\rangle , \tag{4.3}$$

since all the samples are equivalent and so one can single one out in the sum. Then, one proceeds by using the moments in the distribution of W to compute the moments of the *sum* of $n-1$ variates, that are uncorrelated from both W and A. The procedure is far from trivial [21], but one can eventually obtain the asymptotic expression

$$\left\langle \bar{A}_n \right\rangle \approx \frac{\langle AW \rangle}{\langle W \rangle} + \frac{\langle AW \rangle \left\langle W^2 \right\rangle - \langle W \rangle \left\langle AW^2 \right\rangle}{n \left\langle W^3 \right\rangle}.$$

This shows that if A and W are correlated the reweighed mean of a finite number of samples is a biased estimator of the true value $\langle AW \rangle / \langle W \rangle$, i.e. that it bears a systematic error that decreases with the number of sample points. To convince oneself that (4.3) is a biased estimator it suffices to consider that $\left\langle \bar{A}_1 \right\rangle = \langle A \rangle \neq \langle AW \rangle / \langle W \rangle$, since the weight at the numerator and denominator cancel out.

Figure 15: Schematic representation of cumulative reweighed averages for cases with low (red), medium (black) and large (blue) variance of the difference Hamiltonian.



What is more, the systematic error term is associated with a large statistical error. What makes it hard to converge reweighed average is that the weight is often the exponential of a

difference Hamiltonian $H = -\ln W$, which is obtained from a sum of random (or chaotic) terms, and that in many cases has quasi-Gaussian statistics. Hence, the weight has a log-normal distribution, i.e. the distribution of the exponential of a Gaussian variate. The log-normal distribution has a pathological behavior, with very large tails, which means that the distribution of W will be plagued by very large outliers. This becomes apparent when one computes cumulative averages out a trajectory, i.e. the mean of the first n terms in the trajectory as a function of n. Pathological cases are apparent because the cumulative average shows large jumps whenever a new outlier with a large weight appears in the data set. In the worst cases, these discontinuities are still evident after hundreds of thousands of uncorrelated samples.

Under the assumptions that W is precisely log-normal, and that $p(A, H)$ is a multivariate Gaussian with A having variance $\sigma^2(A)$, $H = -\ln W$ having mean zero and variance $\sigma^2(H)$, and the cross-correlation between A and H being $\langle AH \rangle$, one can evaluate the asymptotic expressions

$$\left\langle \bar{A}_n \right\rangle \approx \langle A \rangle - \langle AH \rangle + \langle AH \rangle \frac{e^{\sigma^2(H)}}{n} \tag{4.4}$$

$$\sigma^2\left(\bar{A}_n\right) \approx \left(\sigma^2(A) + \langle AH \rangle^2\right) \frac{e^{\sigma^2(H)}}{n}. \tag{4.5}$$

Both the systematic error and the statistical variance of the mean, decrease asymptotically (for a large number of uncorrelated samples) with n, but the prefactor grows *exponentially* with the variance of the difference Hamiltonian. These estimates are based on the assumption of a Gaussian distribution of the difference Hamiltonian, and are often somewhat pessimistic. However, every time one performs a simulation that involves some kind of reweighing, it is essential to evaluate $\sigma^2(\ln W)$, and to perform more in-depth checks every time the fluctuations are larger than a few units. It might very well be the case that the enhanced sampling or the computational savings obtained by using a reweighing scheme are more than offset by the loss in statistical efficiency due to the pathological behavior of the distribution of W.

Combining multiple simulations Say we have several simulations that are produced by n different probability distributions $P_k(\mathbf{q})$. They may have different temperatures, different biases, etc. A typical example are replica exchange simulations[22]. We want to compute a unique average for an observable, combining the data from the different runs. For a change, we will consider the case where one wants to compute a histogram for a property $s = s(\mathbf{q})$, relative to a target distribution $P(\mathbf{q})$, that reads

$$h(s) = \int \mathrm{d}P(\mathbf{q})\, \delta(s - s(\mathbf{q})).$$

Each simulation generates a histogram consistent with $P_k(\mathbf{q})$

$$h_k(s) = \int \mathrm{d}P_k(\mathbf{q})\, \delta(s - s(\mathbf{q})) = \langle \delta(s - s(\mathbf{q})) \rangle_k$$

where $\langle \cdot \rangle_k$ is meant to indicate an average over the trajectory generated by the k-th simulation. From each simulation we can also recover an estimate of $h(s)$ by reweighing the simulations,

$$\bar{h}_k(s) = \frac{\left\langle \frac{P(\mathbf{q})}{P_k(\mathbf{q})} \delta(s - s(\mathbf{q})) \right\rangle_k}{\langle P(\mathbf{q}) / P_k(\mathbf{q}) \rangle_k},$$

The idea is then to combine the different histograms with coefficients that sum up to one, $\bar{h}(s) = \sum_i c_k \bar{h}_k(s)$, choosing the coefficients so as to minimize the error in the estimate

of $h(s)$. The philosophy is very similar to the Weighted Histogram Analysis Method (WHAM) [23], but we will use Eq. (4.5) to simplify considerably the treatment, and to extend it to different cases than when one is computing the histogram of the collective variable that is used in the bias. Let us assume that the "intrinsic" quality of sampling of all the trajectories is equivalent, so that $\epsilon_k^2(s) \approx \frac{\epsilon^2(s)}{N_k}$ where N_k is the number of uncorrelated samples in each trajectory (if the correlation time is known, $N_k \approx t_k/\tau_k$, where t_k are the length and the correlation time of the k-th simulation – otherwise one might just use the number of samples, even though this might be a rather crude approximation). Also, let us assume that $s(\mathbf{q})$ and the weight functions $w_k(\mathbf{q}) = P(\mathbf{q})/P_k(\mathbf{q})$ are weakly correlated, so that we can just ignore $\langle AH \rangle$ in Eq. (4.5) and estimate the error in the k-th reweighed histogram as

$$\bar{\epsilon}_k^2(s) \approx \frac{\epsilon^2(s)}{N_k} \exp\left[\left\langle (\log W_k(\mathbf{q}))^2 \right\rangle_k - \langle \log W_k(\mathbf{q}) \rangle_k^2\right],$$

independent on the value of s. One could adapt this expression using a more general expression for the error, that does not assume Gaussian statistics or uncorrelated weights and observables. The total square error in $\bar{h}$ is $\bar{\epsilon}(s) = \frac{1}{n}\sum c_k^2 \bar{\epsilon}_k^2(s)$. This can be minimized with a Lagrange multiplier procedure, and eventually one gets

$$c_k = \frac{1/\bar{\epsilon}_k^2(s)}{\sum_k 1/\bar{\epsilon}_k^2(s)} = \frac{N_k \exp\left[\langle \log W_k(\mathbf{q}) \rangle_k^2 - \left\langle (\log W_k(\mathbf{q}))^2 \right\rangle_k\right]}{\sum_k N_k \exp\left[\langle \log W_k(\mathbf{q}) \rangle_k^2 - \left\langle (\log W_k(\mathbf{q}))^2 \right\rangle_k\right]},$$

giving a prescription to combine the averages computed from different reweighed simulations.

4.2 Biasing in collective variable space

In order to compute reaction rates in a complex system, it is first necessary to define good collective variables $\mathbf{s}$ and to compute the free energy (3.4) relative to them, so as to be able to evaluate the transition-state theory rate (3.6). Since however the probability of being at the transition state is extremely small (the very reason why the underlying transition is a rare event!), it is often hard or basically impossible to compute $F(\mathbf{s})$ by conventional, canonical sampling. One possibility is to raise the temperature of the system, that would increase the probability of reaching the high-energy portions of the free energy surface (FES). However, this is often a bad idea: the log-weight for the temperature reweighing (4.2) is $-\left(\beta - \tilde{\beta}\right) V(\mathbf{q})$, and the fluctuations will be proportional to $\left\langle V(\mathbf{q})^2 \right\rangle$, that grows linearly with the system size. No matter how small is the temperature difference, as soon as the system gets large enough the pathological distribution of the weights will make sampling dramatically inefficient. The profound reason for this problem is that the bias depends on all the particle coordinates independently, which leads to extensive growth of the log-weight with system size, even though most of the degrees of freedom are basically spectators and do not need to change for the reaction to take place (think for instance at a bond-breaking reaction, or at a transition in solution).

To focus the sampling on exploring the free energy relative to a set of collective variables $\mathbf{s}$, it is therefore tempting to use a bias potential that depends on the values of the collective variables themselves, i.e. sample relative to $\tilde{V}(\mathbf{q}) = V(\mathbf{q}) + B(\mathbf{s}(\mathbf{q}))$. It is easy to see that the ensemble distorted by the bias can be reweighed with a term $e^{\beta B(\mathbf{s})}$, for which one can hope to obtain less pathological statistics thanks to the fact that $\mathbf{s}$ is typically of much lower dimensionality than $\mathbf{q}$, and that fluctuations will not therefore depend strongly on the size of the system. It is particularly instructive to see how the free

energy is changed by the biasing procedure:

$$\tilde{F}(\mathbf{s}) = -\frac{1}{\beta}\ln\int \mathrm{d}\mathbf{q} e^{-\beta[V(\mathbf{q})+B(\mathbf{s}(\mathbf{q}))]}\delta(\mathbf{s}-\mathbf{s}(\mathbf{q})) = B(\mathbf{s}) + F(\mathbf{s}), \tag{4.6}$$

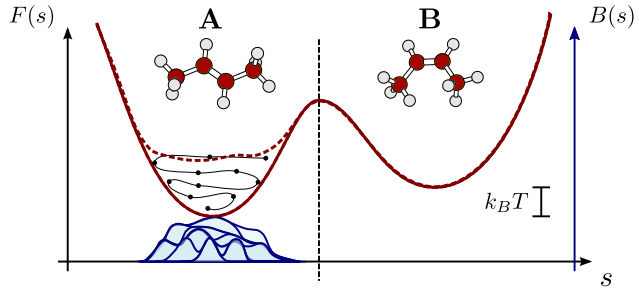
where we exploited the fact that $e^{-\beta B(\mathbf{s})}$ can be brought outside the integral, as the δ function selects only configurations with $\mathbf{s}(\mathbf{q}) = \mathbf{s}$.

Hence, the effect of the bias is to modify the free energy by the value of the bias itself. Eq. (4.6) means that if one chose a bias that cancels completely the free energy $B(\mathbf{s}) = -F(\mathbf{s})$, $\tilde{F}(\mathbf{s})$ would be a constant, and all the values of the collective variables would be equally likely in the simulation. Biasing a simulation based on the same collective coordinates used to estimate a reaction rate has the additional advantage that the flux term $\phi(\mathbf{s})$ that enters Eq. (3.6) could be evaluated from the biased simulation without the need of reweighing, since

$$\tilde{\phi}(s) = \frac{\int \mathrm{d}\mathbf{q} e^{-\beta[V(\mathbf{q})+B(s(\mathbf{q}))]}\delta(s(\mathbf{q}))\phi(\mathbf{q})}{\int \mathrm{d}\mathbf{q} e^{-\beta[V(\mathbf{q})+B(s(\mathbf{q}))]}\delta(s(\mathbf{q}))} = \frac{e^{-\beta B(s)}\int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})}\delta(s(\mathbf{q}))\phi(\mathbf{q})}{e^{-\beta B(s)}\int \mathrm{d}\mathbf{q} e^{-\beta V(\mathbf{q})}\delta(s(\mathbf{q}))} = \phi(s),$$

where we used again the fact that the bias is written as a function of the same collective variable that selects the transition state.

Figure 16: Schematic representation of the metadynamics bias accumulated after a short trajectory. At each instant in time, the bias is built from an accumulation of repulsive Gaussians centered around previously-visited locations.



4.3 Metadynamics

As attractive as it might sound, the idea of performing umbrella sampling in collective variable space to flatten the free energy profile poses the considerable problem of choosing the bias potential $B(\mathbf{s})$. One would want to pick it to be the negative of the free energy, but this would require knowing the FES in the first place. To get out of this circular problem, one needs to construct the bias in an adaptive way, exploiting information from the simulation to refine on the fly the estimate of the free energy, obtain a better bias and hence more efficient sampling. A relatively simple implementation of this idea is offered by the local elevation [24] and metadynamics [25] methods, in which a repulsive bias is built from a superposition of repulsive Gaussians of height w and width σ, that are centered on positions in collective variables space that had been visited before along the trajectory (Figure 16):

$$B(\mathbf{q},t) = B(\mathbf{s}(\mathbf{q}),t) = w\sum_{i\Delta t<t}\exp -\frac{(\mathbf{s}(\mathbf{q}) - \mathbf{s}(\mathbf{q}(i\Delta t)))^2}{2\sigma^2}. \tag{4.7}$$

The repulsive bias acts as "computational sand", discouraging the system from indulging in regions of collective variable space that had been visited before, and forcing it to explore

new configurations and traverse transition state regions. What is more, one can show that under mild conditions, the non-equilibrium dynamics generated by the time-dependent bias (4.7) is such that for a sufficiently long simulation $B(\mathbf{s}, t \to \infty) \to -F(\mathbf{s})$, so that a converged metadynamics trajectory gives access to the free energy of the system [26]. The accuracy of metadynamics rests on a quasi-equilibrium assumption (that sampling at any time is consistent with the biased ensemble at that time), and therefore small deposition rates $w/\Delta t$ should be used to obtain reliable estimates of the free energy. The converged bias can be used for an umbrella sampling calculation to evaluate properties of the system, but there is typically a negligible error associated with re-using also the non-equilibrium section of the calculation to perform reweighed sampling [27].

Well-tempered metadynamics One should use some care when inferring an estimate of the free energy from the negative of the metadynamics bias. For instance, the resolution in collective variable space is limited by the choice of σ, and the fact that discrete hills are being added leads to a residual error of the order of w in the free energy profile. A strategy to improve the accuracy of the profile – also by making the quasi-equilibrium assumption more accurately fulfilled – is to reduce progressively the deposition rate as the simulation progresses. So-called "well-tempered" metadynamics [28] does so in a way that adaptively depends on the magnitude of the bias accumulated at each time

$$B(\mathbf{s}(\mathbf{q}), t) = w \sum_{i\Delta t < t} e^{-B(\mathbf{s}(\mathbf{q}(i\Delta t)))/k_B \Delta T} \exp - \frac{\left(\mathbf{s}(\mathbf{q}) - \mathbf{s}(\mathbf{q}(i\Delta t))\right)^2}{2\sigma^2}. \tag{4.8}$$

As the bias increases, the exponential term reduces the height of the new hills being added. Consider now the limit as $t \to \infty$, when the height of the hills is infinitesimal and one can consider how the bias changes in time based on the probability distribution of being in a given state $\mathbf{q}$ (which in turns depend on the bias):

$$\dot{B}(\mathbf{s}, t) = P(\mathbf{s}) \frac{w}{\Delta t} e^{-B(\mathbf{s},t)/k_B \Delta T} \propto e^{-[F(\mathbf{s}) + B(\mathbf{s},t)]/k_B T} e^{-B(\mathbf{s},t)/k_B \Delta T}.$$

In order for the change in bias to be a constant – meaning that its change in time does not alter its profile or the probability distribution – the bias at the exponent must cancel the free-energy, which leads to $B(\mathbf{s}, t)(1/T + 1/\Delta T) = -F(\mathbf{s})/T$, implying that

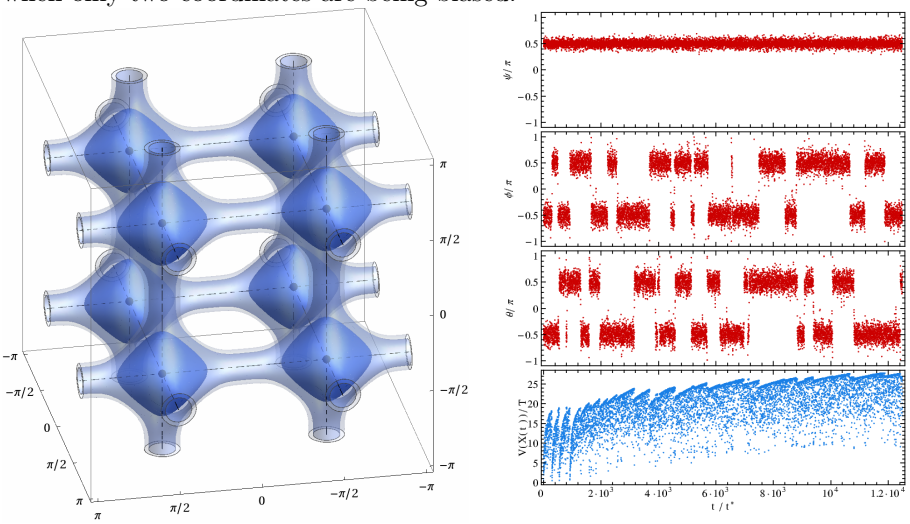
$$B(\mathbf{s}, t \to \infty) \to -F(\mathbf{s}) \frac{\Delta T}{\Delta T + T}.$$

Besides allowing for smooth, systematic convergence, the addition of a bias-dependent damping to the deposition rate makes it possible to interpolate between a $\Delta T \gg T$ limit, at which the bias compensates fully for the free energy, and a $\Delta T = 0$ limit at which no bias is added. Intermediate values make it possible to increase the probability of sampling the transition states without flattening completely the FES, which would make the system spend a lot of time in unphysically high-energy regions.

Dimensionality of sampling and hidden variables This far we have not commented much on the importance of the choice of collective variables $\mathbf{s}$. Just as in the case of rate theory, the choice of "good" collective variables is crucial for the efficiency and accuracy of metadynamics. As it is often the case, there are two conflicting goals to aim for. On one hand, $\mathbf{s}$ should be a low(d)-dimensional description of the transition, ideally one to three-dimensional. This comes for two reasons: on one hand, a bias in a high-dimensional space would exhibit near-Gaussian fluctiations, and lead to poor statistical efficiency of the reweighing procedure, as discussed in Section 4.1. Also, each Gaussian hill covers a "volume" of phase space of the order of σ^d. If a basin of attraction on the FES has an extension of the order of Δs in each dimension, the number of hills that must be deposited

to compensate the basin have to be at least $(\Delta s/\sigma)^d$. Since σ has to be smaller than Δs to have sufficient resolution to identify the basin, the number of hills grows exponentially with dimensionality.

Figure 17: Well-tempered metadynamics trajectory on a 3D potential energy surface $V(\theta, \phi, \psi) = \exp\left[3\left(3 - \sin^4\theta - \sin^4\phi - \sin^4\psi\right)\right] - 1$, periodic in three dimensions. The right panel shows the trajectory of the three angles and the instantaneous value of the bias when only two coordinates are being biased.



On the other hand, one has to be wary of the possibility that the description of the configuration space given by a low-dimensional set of collective variables is incomplete, and do not capture the full complexity of the reactivity of the system. This might be reflected in the fact that the rate estimated from a transition state theory expression (3.6) requires a major correction for re-crossing trajectories, but also might reduce dramatically the efficiency of sampling of a metadynamics trajectory. Figure 17 gives a simple example of the effect of neglecting important degrees of freedom in the description of configuration space. The sampling along the "hidden variable" which is not being biased is not accelerated, and only a partial exploration of available phase space can be obtained. While this is clearly an extreme example, it highlights the fact that the choice of collective variables is perhaps the most important step in performing efficient enhanced sampling applied to the study of rare events.

5 Dimensionality reduction

In many ways, obtaining a low-dimensional, coarse-grained description of the configurations that are accessible at a given temperature is the key to unravel fully the power of atomistic simulations. A simplified model of the complex behavior of a system composed of thousands (or millions!) of degrees of freedom is necessary to present an intuitive picture of the physics of the problem, as well as to extract dynamical and non-equilibrium properties from a equilibrium simulation, and finally to accelerate sampling by umbrella sampling or metadynamics.

In complex materials, biomolecules and in general systems in which the order parameter that underlies a given transition is not obvious, one needs to perform time-consuming trial-and-error tests to determine good reaction coordinates that can capture the essence of the underlying physics. One is then led to wonder whether it is possible to perform a computational analysis that can automatically infer the best collective coordinates to describe a reaction starting from the simulation data itself. As discussed in Chapter 3, the committor is an ideal reaction coordinate to describe a rare event, but it requires a preliminary definition of the stable states, and very extensive preliminary sampling. In practice, one needs a procedure that can be performed on preliminary, incomplete sampling of the free energy surface, that is sufficiently predictive to describe states that were missing in the initial data set, and that is robust enough to deal with sparse, noisy data.

The general idea is to start from a high-dimensional description of the structure being studied (e.g. the set of all distances between pair of atoms in a molecule, coordination numbers, torsions in the backbone of a polymer, etc.), and find the projections of N high, D-dimensional vectors $\{X_i\}$ onto a lower, d-dimensional set of points $\{x_i\}$. Having performed this preliminary mapping one can proceed to compute an out-of-sample embedding function that associates a projection $x = f(X)$ to each high-dimensional vector X.

A large number of techniques have been developed in the machine-learning community to perform dimensionality reduction. Here we will focus on simple linear projections (principal component analysis and classical multidimensional scaling) [29], and a couple of more advanced non-linear projections (ISOMAP [30] and sketch-map [31]). Other techniques we will not cover here include Laplacian eigenmaps [32], locally-linear embedding [33], Hessian eigenmaps [34], or diffusion maps [35].

5.1 Linear projections

The simplest approach to taking the projection of a set of D-dimensional vectors into a lower d-dimensional space would be to take a linear projection, i.e. take $x_i = \mathbf{P}^T X_i$, where $\mathbf{P}$ is a $D \times d$ matrix, and x_i and X_i are taken to be column vectors. The question is then how to choose the projector in such a way that the embedding is "best" in some metric.

Principal component analysis (PCA) Let us assume that the columns of $\mathbf{P}$ are orthonormal, so that $\mathbf{P}^T\mathbf{P} = \mathbf{1}_d$ (where $\mathbf{1}_n$ is the n-dimensional identity matrix). Then, $\tilde{X}_i = \mathbf{P}x_i$ would be the D-dimensional vector that can be reconstructed based on the limited information that remains in x_i. We can then try to figure out how to choose $\mathbf{P}$ so that the mean squared distance

$$\mathcal{S} = \frac{1}{N}\sum_i \left|X_i - \tilde{X}_i\right|^2$$

is minimum. To do so, assume without loss of generality that the X_i had been shifted rigidly to have zero mean, so that $\mathbf{C} = \sum_i X_i X_i^T/N$ is the covariance matrix of the high-dimensional data set, and consider that $\left|X_i - \tilde{X}_i\right|^2 = \mathrm{Tr}\left(X_i - \tilde{X}_i\right)\left(X_i - \tilde{X}_i\right)^T$. Then,

$$\begin{aligned} S &= \mathrm{Tr}\,\frac{1}{N}\sum_i \left(X_i - \tilde{X}_i\right)\left(X_i - \tilde{X}_i\right)^T = \\ &= \mathrm{Tr}\,\frac{1}{N}\sum_i \left[X_i X_i^T - \mathbf{P}\mathbf{P}^T X_i X_i^T - X_i X_i^T \mathbf{P}\mathbf{P}^T + \mathbf{P}\mathbf{P}^T X_i X_i^T \mathbf{P}\mathbf{P}^T\right] = \\ &= 2\left(\mathrm{Tr}\,\mathbf{C} - \mathrm{Tr}\,\mathbf{P}^T\mathbf{C}\mathbf{P}\right) \end{aligned}$$

where we used the fact that $\mathbf{P}^T\mathbf{P} = \mathbf{1}_d$ and that circular permutations of the factors in a product of matrices do not change the trace. Now, $\mathrm{Tr}\,\mathbf{C}$ is the sum of the eigenvalues of the covariance matrix. If $\mathbf{P}$ was a one-dimensional projector, then the best choice to reduce $\mathcal{S}$ is clearly to choose $\mathbf{P}$ to be the eigenvector associated with the largest eigenvalue c_1 of $\mathbf{C}$, as then $\mathrm{Tr}\,\mathbf{P}^T\mathbf{C}\mathbf{P} = c_1$ would remove the largest possible component from $\mathrm{Tr}\,\mathbf{C}$. Then it is easy to see that if one wanted to add a second column to the projector $\mathbf{P}$, the best choice would be to pick the eigenvector associated with the second largest eigenvalue, and so on.

Multidimensional scaling (MDS) Let us try a different approach. Consider a measure of similarity between the data points $S_{ij} = |X_i - X_j|$ – we take it to be the Euclidean distance, but any other metric could be used. A reasonable requirement for a projection to be a faithful representation of the distribution of points is that the similarity between the high-dimensional points is preserved after the projection, i.e. that a suitable *stress* function

$$\chi^2 = \sum_{ij} \left(S_{ij} - |x_i - x_j|\right)^2 \tag{5.1}$$

is minimized with respect to the positions of the embedded points $\{x_i\}$. Note that at this stage this is not a linear projection, and that the procedure involves an iterative minimization to find a (possibly only local) minimum of χ^2.

While Eq. (5.1) is extremely general, in that one could easily use any sort of measure of similarity in the definition of S_{ij}, assuming that the underlying metric is just the Euclidean distance makes it possible formulate the problem so that it can be tackled as an eigenvalue decomposition. In particular, we exploit the fact that Euclidean distance matrices (the matrices of squared Euclidean distances between vectors $S_{ij} = |X_i - X_j|^2$) can be shown to be in biunivocal relation with positive semidefinite matrices of the form $\mathbf{X}\mathbf{X}^T$. Specifically, let $\mathbf{X}$ be the $N \times D$ matrix having the points X_i as rows. Define the $N \times N$ centering matrix $\mathbf{H}$ such that $H_{ij} = \delta_{ij} - \frac{1}{N}$. It is easy to see that applying this matrix to the data set removes the center of mass $\bar{X} = \frac{1}{n}\sum_i X_i$ from each one of the points (i.e. $(\mathbf{H}\mathbf{X})_i = \left(X_i - \bar{X}\right)^T$) and that $\mathbf{H}^T\mathbf{H} = \mathbf{H}$. Now, let's see that one can write an inner product form $\mathbf{B} = (\mathbf{H}\mathbf{X})(\mathbf{H}\mathbf{X})^T$ as $\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{S}\mathbf{H}$. It is a bit tedious but straightforward to write out

$$\begin{aligned} -\frac{1}{2}\left(\mathbf{H}\mathbf{S}\mathbf{H}\right)_{ij} &= -\frac{1}{2}\sum_{kk'}\left(\delta_{ik} - \frac{1}{n}\right)\left[\left(X_k - X_{k'}\right)^T\left(X_k - X_{k'}\right)\right]\left(\delta_{jk'} - \frac{1}{n}\right) = \\ &= -\frac{1}{2}\sum_{kk'}\left(\delta_{ik} - \frac{1}{n}\right)\left(X_k^T X_k + X_{k'}^T X_{k'} - 2X_k^T X_{k'}\right)\left(\delta_{jk'} - \frac{1}{n}\right) = \\ &= X_i^T X_j + \bar{X}^T\bar{X} - \bar{X}^T X_j - X_i^T\bar{X}. \end{aligned}$$

But this is just the same as

$$\left[(\mathbf{HX})\,(\mathbf{HX})^T\right]_{ij} = \left(X_i - \bar{X}\right)^T \left(X_j - \bar{X}\right) = X_i^T X_j + \bar{X}^T\bar{X} - \bar{X}^T X_j - X_i^T \bar{X}.$$

So, after having centered the data set – which certainly will not change the relative similarity between reference configurations – $\mathbf{XX}^T$ is the positive semidefinite matrix that corresponds to the matrix of squared Euclidean distances $\mathbf{S}$.

Then, one can proceed to approximate $\mathbf{X}$ taking the singular value decomposition of $\mathbf{B}$, picking the d largest eigenvalues υ_i and associated eigenvectors U_i, and set $(x_j)_i = \sqrt{\upsilon_i}\,(U_i)_j$, i.e. taking the elements of the eigenvectors to be the coordinates of the low-dimensional projections of the X_i. It can be shown that $\mathbf{xx}^T$ is the matrix of rank d that best approximates $\mathbf{XX}^T$ in the Frobenius norm[1]. So, classical MDS can be understood as the process of mapping the Euclidean distance matrix to a positive semidefinite matrix, approximating it by a singular value decomposition, and then interpreting the resulting matrix as the set of points whose mutual distances approximate the high-dimensional distance matrix.

Interestingly, in this form classical MDS is completely equivalent to PCA. To see this, note that the covariance matrix of the high-dimensional data set can be written as $\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{HX}$. In the following, we will assume to have centered preliminarily the data set, so that $n\mathbf{C} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{B} = \mathbf{XX}^T$. In PCA, we take the eigenvectors of $\mathbf{C}$ associated with the largest eigenvalues of $\mathbf{C}$, γ_i, take them to be the columns in a projector $\mathbf{P}$ and compute the low-dimensional embeddings as $x_i = \mathbf{P}^T X_i$. In terms of a matrix $\mathbf{X}$ written in the same notation we introduced for MDS, this can be written as $\mathbf{x}^T = \mathbf{P}^T\mathbf{X}^T$, i.e. $\mathbf{x} = \mathbf{XP}$. First, let's see that the eigenvalues of $n\mathbf{C}$, υ_i, are also eigenvalues of $\mathbf{B}$, and that if u_i is the eigenvector of $n\mathbf{C}$ associated with υ_i, $\mathbf{X}u_i$ is the corresponding eigenvector of $\mathbf{B}$. To see this, just consider that

$$\mathbf{B}\mathbf{X}u_i = \left(\mathbf{XX}^T\right)\mathbf{X}u_i = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)u_i = \upsilon_i\mathbf{X}u_i$$

This automatically implies that the low-dimensional projections obtained from PCA differ from those obtained by MDS at most by a scaling of the coordinates – since the coordinates of PCA projections are of the form $\mathbf{X}u_i$, and the coordinates of MDS projections are eigenvectors of $\mathbf{B}$. it remains to prove that the scaling by $\sqrt{\upsilon_i}$ is what is needed to make the projections identical. Considering normalized eigenvectors, $U_i^T U_i = 1$,whereas $u_i^T\mathbf{X}^T\mathbf{X}u_i = \upsilon_i u_i^T u_i = \upsilon_i$, which proves that indeed $\sqrt{\upsilon_i}$ is the correct scaling.

5.2 Non-linear dimensionality reduction: ISOMAP

PCA is by construction a linear projection, that can only provide a faithful representation of the high-dimensional data set if the vectors lie (at least approximately) in a linear subspace. MDS (at least in the non-iterative version) is equivalent to PCA and therefore has the same limitations. In many cases the low-dimensional structure of the data set is only locally linear, but cannot be represented globally in a low-dimensional linear subspace. This is for instance the case of a curved (hyper)surface embedded in a high-dimensional space (see Figure 18). Among the possible approaches to improve a linear dimensionality-reduction algorithm, the simplest is perhaps to work on the definition of a non-linear metric that can capture the curved structure of the data set, and then just proceed with classical MDS.

This is the approach taken by ISOMAP [30], that uses local connectivity information to evaluate geodesic distances over the curved low-dimensional structure of the data set,

[1] Recall that $\|\mathbf{M}\|_F^2 = \sum_{ij} M_{ij}^2 = \mathrm{Tr}\,\mathbf{MM}^T$. When $\mathbf{M}$ is real and symmetric and can be decomposed as $\mathbf{M} = \mathbf{O}\boldsymbol{\mu}\mathbf{O}^T$, where $\boldsymbol{\mu}$ is the diagonal matrix of the eigenvalues of $\mathbf{M}$, $\|\mathbf{M}\|_F^2 = \sum_i \mu_i^2$, as it can be readily seen by considering that $\mathbf{O}^T\mathbf{O} = \mathbf{1}$ and the circular invariance of the trace.

Figure 18: Example of a distribution of points in three dimensions that is *locally* 2D, but globally curved, and that cannot therefore be treated by a linear projection method.

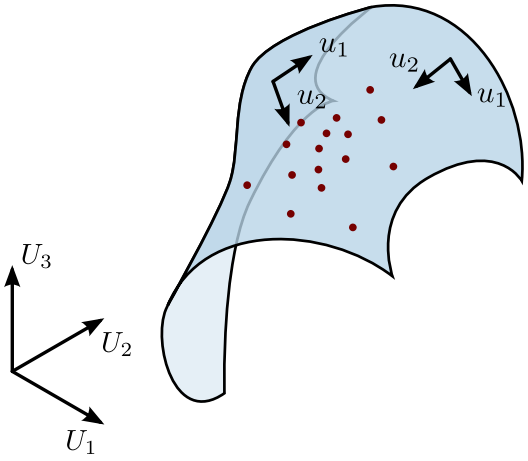
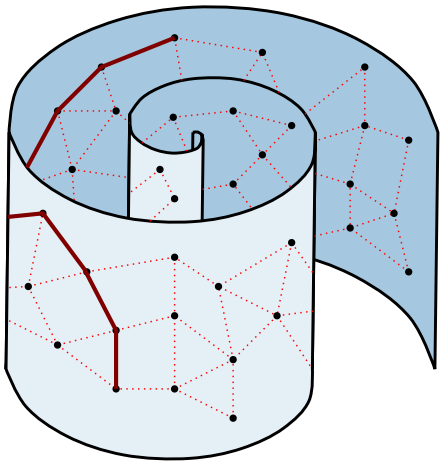


Figure 19: Definition of curved geodesic distances based on the local connectivity of the data set.



and then applies a spectral decomposition to the (centered) matrix of squared distances – effectively applying the linear version of multi-dimensional scaling. The algorithm is as follows:

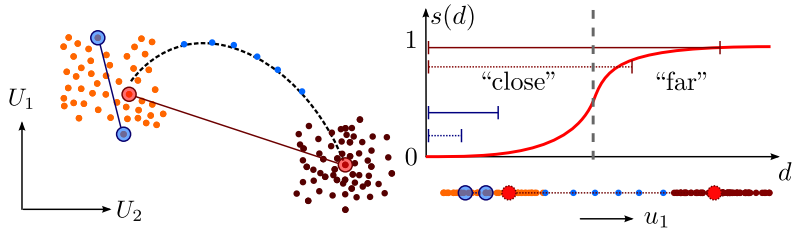
1. Define a neighborhood of each data point. The definition of neighborhoods is the only free parameter of ISOMAP. One can either select the k nearest neighbor of each sample, or take all the samples within a distance ϵ as neighbors.
2. If $d_{ij} = |X_i - X_j|$ if X_j is a neighbor of X_i, or ∞ otherwise, the distance between any pairs of points is then defined as the shortest path between the two nodes (see Figure 19)
$$d(X_i, X_j) = \min_{k_1, k_2, \ldots k_m} d_{ik_1} + \sum_{\alpha} d_{k_\alpha k_{\alpha+1}} + d_{k_m j}$$
3. Apply MDS to the matrix of squared (geodesic) distances to obtain the low-dimensional projections.

5.3 Sketch-map

When analyzing atomistic simulations, one rarely is dealing with a locally data distribution of the kind ISOMAP is best suited at. Atoms at finite temperature experience thermal fluctuations along all directions, which often translates in near-Gaussian, multidimensional fluctuations when describing the system with a high-dimensional set of collective variables.

A MDS scheme would then face the impossible task of rendering the set of distances between points scattered in D dimensions by distributing the projections in a much lower dimensional space, which would probably affect the accuracy at which meaningful information about the reciprocal similarity of different metastable states can be represented.

Figure 20: Thermal fluctuations mean that the distances within a the region of phase space that can be associated with a meta-stable state cannot be represented faithfully in low dimension. Applying a sigmoid function to the distances defines a coarse-grained measure of proximity, where distances within the same basin are transformed to zero, and distances between points that are far apart saturate at one.



Sketch-map is a non-linear dimensionality reduction scheme, loosely based on multi-dimensional scaling, that starts by acknowledging the issue of high-dimensional thermal fluctuations, and modifies the objective function of (iterative) MDS (5.1) by introducing sigmoid cut off functions to transform both the high-dimensional similarities and the Euclidean distances that are taken to represent the similarity between projected points [31]:

$$\chi^2 = \sum_{i,j=1}^{N} \left[s(|X_i - X_j|) - s(|x_i - x_j|) \right]^2 . \tag{5.2}$$

The function $s(d)$ transforms to zero distances that are characteristic of fluctuations within the same meta-stable state, and to one the distances between configurations that

are completely unrelated. Since the transformation is applied in both the high and low-dimensional spaces, this is equivalent to requiring that points that are close together stay close in the projected space, and configurations that are far apart from each other are projected in separate regions. This is a much simpler task than matching distance, and the iterative optimization (which has to be used since Eq. (5.2) cannot easily be expressed as an eigenvalue problem) can focus on representing correctly the connectivity between nearby basins, which is probably the most important requirement to obtain a meaningful representation of the configuration space of a compound at the atomic scale. The iterative minimization of Eq. (5.2) is not trivial, and so it is important to start from a good selection of reference configurations. Then, the projection of most other data points can be obtained based on these reference configurations.

Out-of-sample embedding The ultimate goal of a (non-linear) dimensionality reduction routine is to obtain a mapping from the D-dimensional to the d-dimensional space, $x(X)$. PCA makes this step obvious, since the embedding is a linear projection of the high-dimensional data set and any additional configuration can be dealt with using the same projector, i.e. $x = \mathbf{P}^T X$. Non-linear methods are somewhat trickier. The idea is to use the high-dimensional reference points X_i as milestones, inferring the position of X relative to them and then using this information to position the projection x relative to the projections x_i of the reference configurations. One simple approach to do so uses a weighed combination of the x_is [36]

$$x(X) = \frac{\sum_i e^{-|X-X_i|/\lambda} x_i}{\sum_i e^{-|X-X_i|/\lambda}},$$

where λ specifies a characteristic length scale for the distance between neighboring reference points. This has the advantage of simplicity, but has some limitations. The projection is bound to lie within the convex hull defined by the x_is, and one can see that configurations that are far away from all of the references, the projection tends to approach the center of mass of the set of projections, which is clearly an artifact. In practice, this kind of approach works when the references cover densely all of the accessible configuration space, which is rarely the case when dealing with complex systems that require accelerated sampling.

Another possibility is to use the same iterative minimization framework as for Eq. (5.2), and define the projection as

$$x(X) = \arg\min_x \chi^2(x, X), \qquad \chi^2(x, X) = \sum_{i=1}^{N} \left[s(|X - X_i|) - s(|x - x_i|) \right]^2.$$

This definition is much more robust in cases where there is poor sampling of configuration space [37], since all the references can be used to get information on the proximity of the out-of-sample point, and to find an embedding that reproduces at best this information in low dimension. For instance, a configuration that is very different from all of the reference points will be projected somewhere on the outer rim of the region occupied by the x_is, and not in the center of the data set.

Bibliography

[1] M P Allen and D J Tildesley. *Computer simulation of liquids*. Oxford University Press, USA, 1990.

[2] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[3] C W Gardiner. *Handbook of Stochastic Methods*. Springer, Berlin, third edition, 2003.

[4] Mark Tuckerman. *Statistical Mechanics and Molecular Simulations*. Oxford University Press, 2008.

[5] Loup Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, 159(1):98–103, July 1967.

[6] M. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97(3):1990, 1992.

[7] Hans C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72(4):2384–2393, 1980.

[8] G Bussi, D Donadio, and M Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):14101, 2007.

[9] H J C Berendsen, J P M Postma, W F Van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684, 1984.

[10] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81(1):511, 1984.

[11] W G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[12] G J Martyna, M E Tuckerman, and M L Klein. Nosé-Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.*, 97(4):2635, 1992.

[13] P Langevin. The theory of Brownian movement. *CR Acad. Sci*, 146:530, 1908.

[14] H Risken. *The Fokker-Planck Equation*. Springer, Berlin, 1996.

[15] Robert Zwanzig. *Nonequilibrium statistical mechanics*. Oxford University Press, New York, 2001.

[16] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. Langevin Equation with Colored Noise for Constant-Temperature Molecular Dynamics Simulations. *Phys. Rev. Lett.*, 102(2):020601, January 2009.

[17] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. Colored-Noise Thermostats à la Carte. *J. Chem. Theory Comput.*, 6(4):1170–1180, April 2010.

[18] David Wales. *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

[19] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[20] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, February 1977.

[21] Michele Ceriotti, Guy a. R. Brain, Oliver Riordan, and David E. Manolopoulos. The inefficiency of re-weighted sampling and the curse of system size in high-order path integration. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 468(2137):2–17, September 2011.

[22] David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7(23):3910, December 2005.

[23] S Kumar, J M Rosenberg, D Bouzida, R H Swendsen, P A Kollman, and John M Rosenbergl. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[24] T Huber, a E Torda, and W F van Gunsteren. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J. Comput. Aided. Mol. Des.*, 8(6):695–708, December 1994.

[25] A Laio and M Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA*, 99(20):12562–12566, 2002.

[26] G Bussi, A Laio, and M Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.*, 96(9):90601, 2006.

[27] M Bonomi, A Barducci, and M Parrinello. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *J. Comput. Chem.*, 30(11):1615–1621, 2009.

[28] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.*, 100(2):20603, 2008.

[29] Trevor F Cox and Michael A A Cox. *Multidimensional scaling*. CRC Press, 2010.

[30] J B Tenenbaum, V de Silva, and J C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (80-.).*, 290(5500):2319–2323, December 2000.

[31] Michele Ceriotti, Gareth a Tribello, and Michele Parrinello. From the Cover: Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. USA U. S. A.*, 108(32):13023–8, August 2011.

[32] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.*, 15(6):1373–1396, June 2003.

[33] S T Roweis and L K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science (80-.).*, 290(5500):2323–6, December 2000.

[34] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA*, 100(10):5591–5596, April 2003.

[35] Andrew L Ferguson, Athanassios Z Panagiotopoulos, Pablo G Debenedetti, and Ioannis G Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc. Natl. Acad. Sci. USA U. S. A.*, 107(31): 13597–602, August 2010.

[36] Vojtěch Spiwok and Blanka Králová. Metadynamics in the conformational space nonlinearly dimensionally reduced by Isomap. *J. Chem. Phys.*, 135(22):224504, December 2011.

[37] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *J. Chem. Theory Comput.*, 9(3):1521–1532, March 2013.