EPJ manuscript No.
(will be inserted by the editor)

On two possible definitions of the free energy for collective variables

Carsten Hartmann[1,a], Juan C. Latorre[1,b], and Giovanni Ciccotti[2,3,c]

[1] Institut für Mathematik, Freie Universität Berlin, Germany
[2] Dipartimento di Fisica, Università di Roma "La Sapienza", Italy
[3] School of Physics, University College Dublin, Ireland

Abstract The aim of this mini-review article is to clarify the relation between two distinct formulations of the thermodynamic free energy for collective variables which can be found in the molecular dynamics literature. In doing so, we discuss the different ensemble concepts underlying the two definitions and reveal their relation to strong confinement (restraints) and molecular constraints. The latter analysis is based on a variant of Federer's coarea formula which can be regarded as a generalization of Fubini's theorem for iterated integrals to curvilinear coordinates and which implies the famous "blue moon" ensemble identity for computing conditional expectations using constrained simulations. For illustration we will present a few paradigmatic examples.

1 Introduction

Free energy is probably one of the most important quantities in analysing molecular systems [1]. If certain collective variables are given which monitor, e.g., transitions between molecular conformations, one can define a free energy associated with these collective variables as the logarithm of their probability density [2]. Free energy, as thus defined, encodes the statistical weights of the molecular conformations. Nevertheless, many choices of collective variables give rise to the same partitioning of state space into conformations, and one may likewise define a free energy as a function of the level sets of the collective variables. Other than the aforementioned free energy, this second free energy, which is intrinsically geometric in that it solely depends upon the foliation given by the collective variables, appears to be a common quantity for computing, e.g., transition rates between metastable states [3].

In this article we want to advocate the view that both free energies are, in their own right, physically sensible concepts to which the term *potential of mean force* applies, albeit the words "mean" and "force" may have different meanings in each case. Moreover the two free energies can be transformed into one another by a suitable reweighting procedure. There is a large body of literature arguing in favour of one or the other (e.g., see [4,5]) and sometimes confusing the two concepts; cf. the discussion in [6,7]. As this part of the literature has a large readership while the few articles

[a] e-mail: carsten.hartmann@fu-berlin.de
[b] e-mail: latorre@mi.fu-berlin.de
[c] e-mail: giovanni.ciccotti@ucd.ie

(e.g., [3,8]) that have clarified the distinction are much less known, we feel it is useful to give a pedagogical introduction to the problem and explain when it is appropriate to use one or the other definition. The two definitions are the theme of the next two sections; Section 2 is devoted to the probabilistic ansatz, whereas Section 3 deals with the more geometric approach. While picking either definition depends on the scope of application rather than being a question of right or wrong, the matter bears some resemblance with the classical problem of how to realise holonomic constraints [9,10,11]; cf. also [12, Secs. 3.5–3.8]. The relation between free energy, strong confinement and constraints is outlined in the Sections 2.1 and 4. Finally, the findings of the article are briefly summarised in Section 5.

This mini-review article contains results from [5,3,8]. Its new contribution consists in giving illustrative examples, the pedagogical use of the coarea formula (see Sec. 2.1) that connects the two definitions of free energy, and in pointing out the relation to strongly confined molecular systems.

2 Free energy as a probabilistic concept

Consider a system of particles assuming states $(x,p) \in \mathbb{R}^{2n}$ with total energy $H(x,p) = T(p) + V(x)$, where T is kinetic and V is potential energy. We suppose that the system is in contact with a heat bath and that its states follow the canonical probability distribution

$$\mu(x,p) = \frac{1}{Z}\exp(-\beta H(x,p))\,, \quad Z = \int_{\mathbb{R}^{2n}} \exp(-\beta H(x,p))\,dxdp \tag{2.1}$$

where we assume that the integral exists. Here $\beta = (k_B T)^{-1}$ denotes the inverse temperature with k_B being the Boltzmann constant. We further assume that $T, V \in C^1(\mathbb{R}^n)$ are bounded below and sufficiently confining so that the energy level sets

$$\Sigma(E) = \big\{(x,p) \in \mathbb{R}^{2n} \colon H(x,p) = E\big\}$$

are bounded for all regular values $E \in \mathbb{R}$ that $H = T + V$ takes. The normalization constant $Z < \infty$, the *partition function*, plays an important role, for the various macroscopic system properties can be derived from it, the most important one being the Helmholtz free energy that is the thermodynamic potential of the canonical ensemble [13].

To start off with an example, suppose we want to know the probability of observing states $(x,p) \in \mathbb{R}^{2n}$ having the energy $H(x,p) = E$. The probability density

$$\mu_H(E)\,dE = \mathbf{P}[H(x,p) \in [E, E+dE)],$$

is given by

$$\mu_H(E) = \frac{1}{Z}\int_{\mathbb{R}^{2n}} \exp(-\beta H(x,p))\delta(H(x,p) - E)\,dxdp\,,$$

where δ denotes Dirac's delta function (see also Sec. 2.1 below). Defining the quantity

$$\Omega(E) = \int_{\mathbb{R}^{2n}} \delta(H(x,p) - E)\,dxdp\,,$$

which is known as the *density of states* and which, by the lower bound and the growth conditions on the total energy H, is finite for all regular values of E, the density $\mu_H(E)$

can be recast as

$$\mu_H(E) = \frac{1}{Z}\Omega(E)\exp(-\beta E)$$
$$= \frac{1}{Z}\exp(-\beta F(E)),$$

with $F(E) = E - \beta^{-1}\ln\Omega(E)$. We call $F(E)$ the free energy of the macroscopic observable *energy*. Since the exponential function quickly decays as F grows, the major contribution to the integral comes from the energy value $E = E^*$ for which $F(E)$ attains a minimum. Hence, by construction, the minimiser $E^* = \operatorname{argmin} F(E)$ determines the most probable value of E in the canonical distribution.

The approach can be generalised to arbitrary macroscopic properties of the system, not just the energy. To this end, we let $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ be a sufficiently smooth macroscopic observable that is independent of the momenta p. The latter assumption implies that we may easily integrate out the kinetic energy part from the canonical distribution and replace (2.1) by the momentum-reduced probability density

$$\rho(x) = \frac{1}{\bar{Z}}\exp(-\beta V(x))\,, \quad \bar{Z} = \int_{\mathbb{R}^n}\exp(-\beta V(x)\,dx\,, \tag{2.2}$$

where we will drop the overbar for notational convenience and simply write $Z := \bar{Z}$. In the following we will speak of Φ as a *collective variable*.[1] Then, as before, the corresponding probability distribution

$$\rho_\Phi(\xi)\,d\xi = \mathbf{P}\left[\Phi(x) \in [\xi, \xi + d\xi)\right],$$

is found by marginalising out the states $x \in \mathbb{R}^n$, i.e.,

$$\rho_\Phi(\xi) = \frac{1}{Z}\int_{\mathbb{R}^n}\exp(-\beta V(x))\delta(\Phi(x) - \xi)\,dx\,.$$

Calling

$$F_\Phi(\xi) = -\beta^{-1}\ln Z_\Phi(\xi)\,, \quad Z_\Phi(\xi) = \int_{\mathbb{R}^n}\exp(-\beta V(x))\delta(\Phi(x) - \xi)\,dx\,, \tag{2.3}$$

the *thermodynamic free energy associated with the collective variable* Φ, it readily follows that ρ_Φ has the form of the usual canonical probability density, namely

$$\rho_\Phi(\xi) = \frac{1}{Z}\exp(-\beta F_\Phi(\xi))\,. \tag{2.4}$$

2.1 The coarea formula

The delta function in (2.3) is a symbolic expression whose precise meaning is provided by the coarea formula: calling

$$\Sigma(\xi) = \{x \in \mathbb{R}^n\colon \Phi(x) = \xi\},$$

the level set of the function Φ for a regular value $\xi \in \mathbb{R}$, assuming that $|\nabla\Phi| \neq 0$ and denoting by $d\sigma$ the area element on $\Sigma(\xi)$, the coarea formula asserts that [14,15]

$$\int_{\mathbb{R}^n} f(x)\,dx = \int_{\mathbb{R}}\left(\int_{\Sigma(\xi')} f|\nabla\Phi|^{-1}d\sigma\right)d\xi', \tag{2.5}$$

[1] For the sake of simplicity, we consider only the case in which Φ is scalar.

for any integrable function $f\colon \mathbb{R}^n \to \mathbb{R}$. Hence,

$$\begin{aligned}\int_{\mathbb{R}^n} f(x)\delta(\Phi(x)-\xi)\,dx &= \int_{\mathbb{R}} \left(\int_{\Sigma(\xi')} f|\nabla\Phi|^{-1}\delta(\xi'-\xi)\,d\sigma \right) d\xi' \\ &= \int_{\Sigma(\xi)} f|\nabla\Phi|^{-1} d\sigma\,.\end{aligned} \tag{2.6}$$

A limiting procedure for the coarea formula using a bias potential. We may give a quick-and-dirty derivation of the identity (2.6) without resorting to (2.5). We suppose that the rightmost integral in (2.6) exists and define the bias potential

$$\varphi(x) = \frac{1}{2}\left(\Phi(x)-\xi\right)^2\,. \tag{2.7}$$

A "mollified" delta function can be defined as

$$\delta_\varepsilon(\Phi(x)-\xi) = \frac{1}{\sqrt{2\pi\varepsilon}} \exp\left(-\frac{1}{\varepsilon}\varphi(x)\right)\,, \quad \varepsilon > 0\,. \tag{2.8}$$

Then, by definition of the mollifier δ_ε,

$$\lim_{\varepsilon\to 0} \int_{\mathbb{R}^n} f(x)\delta_\varepsilon(\Phi(x)-\xi)\,dx = \int_{\mathbb{R}^n} f(x)\delta(\Phi(x)-\xi)\,dx, \tag{2.9}$$

for all suitable test functions $f\colon \mathbb{R}^n \to \mathbb{R}$. Obviously φ is strictly convex in the direction normal to the level set $\Sigma(\xi)$ and attains its minimum exactly on the level set (see Figure 1). This can be rephrased by saying that φ penalises deviations of $x \in \mathbb{R}^n$ from the level set. Taylor expanding φ about its minimum at $\Phi(x) = \xi$, noting that the function φ is even in $\Phi(x)-\xi$, we find

$$\varphi(x) = \frac{1}{2}\left(\nabla\Phi(\sigma)^T(x-\sigma)\right)^2 + \mathcal{O}(|x-\sigma|^4),$$

where

$$\sigma = \operatorname*{argmin}_{y\in\Sigma(\xi)} |x-y|$$

is the point closest to x that lies on $\Sigma(\xi)$ or, in other words, the orthogonal projection of x on $\Sigma(\xi)$; accordingly the surface element $d\sigma$ is understood as the restriction of the n-dimensional integration measure dx to the submanifold $\Sigma(\xi) \subset \mathbb{R}^n$. Since σ is unique whenever x is sufficiently close to $\Sigma(\xi)$ and φ penalises deviations from $\Sigma(\xi)$, it follows by Laplace's method that

$$\begin{aligned}&\int_{\mathbb{R}^n} \delta_\varepsilon(\Phi(x)-\xi) f(x)\,dx \\ &\qquad \approx \frac{1}{\sqrt{2\pi\varepsilon}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2\varepsilon}\left(\nabla\Phi(\sigma)^T(x-\sigma)\right)^2\right) f(x)\,dx,\end{aligned}$$

for all sufficiently small $\varepsilon > 0$. The Gaussian integral on the right-hand side can be carried out most easily using appropriate coordinates on $\Sigma(\xi)$ and in the direction normal to it. More precisely, we let $u(\sigma) = (u_1(\sigma), \ldots, u_{n-1}(\sigma))$ be a local coordinate system on $\Sigma(\xi)$. Its inverse $\sigma(u)$ is a local embedding that maps local coordinates $u \in \mathbb{R}^{n-1}$ to points $\sigma \in \Sigma(\xi)$. We further call $v = \pm|x-\sigma|$ the signed distance from

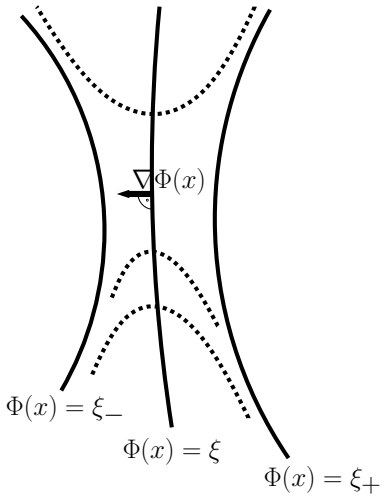

Figure 1. Geometry of level sets: the gully width of the confinement potential φ (dotted lines) in the direction of the normal is of the order $|\nabla\Phi|^{-2}$.

$\Sigma(\xi)$. The transformation from the coordinates x to the adapted coordinate system u, v is one-to-one for all points sufficiently close to $\Sigma(\xi)$ and is given by[2]

$$x(u, v) = \sigma(u) + v\, n(\sigma(u))\,.$$

where $n(\sigma)$ is the unit normal to $\Sigma(\xi)$ at σ. Transforming to the new coordinates and integrating over the normal direction, we find

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \int_{\mathbb{R}^n} \delta_\varepsilon(\Phi(x) - \xi) f(x)\, dx \\
&\quad = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\,|\nabla\Phi(\sigma(u))|^2\, v^2\right) f(\sigma(u))\, \sqrt{\det J(u)}\, du dv \\
&\quad = \int_{\Sigma(\xi)} f(\sigma(u)) |\nabla\Phi(\sigma(u))|^{-1}\, \sqrt{\det J(u)}\, du
\end{aligned}$$

where $J(u) = \nabla\sigma(u)^T \nabla\sigma(u)$ in the second and third line is the metric tensor that is induced by embedding $u \mapsto \sigma(u)$; for more details, see [16]. Finally noting that $\sqrt{\det J(u)}\, du$ is the coordinate expression of $d\sigma$, the last equation together with (2.9) implies the coarea formula (2.6).

2.2 Gauge dependence and the Fixman potential

An important lesson from our short derivation of the coarea formula is that the free energy $F_\Phi(\xi)$ is not invariant under transformations of the collective variable. Different confinement potentials, e.g., quartic instead of quadratic give rise to different gradient terms in (2.6), and accordingly the corresponding free energies will be different. That

[2] By a common, but harmless abuse of notation we do not distinguish between coordinates and coordinate maps, i.e., we write $x = x(u, v)$ and understand x as the Cartesian coordinate as well as the function that maps the local coordinates u, v to the Cartesian ones.

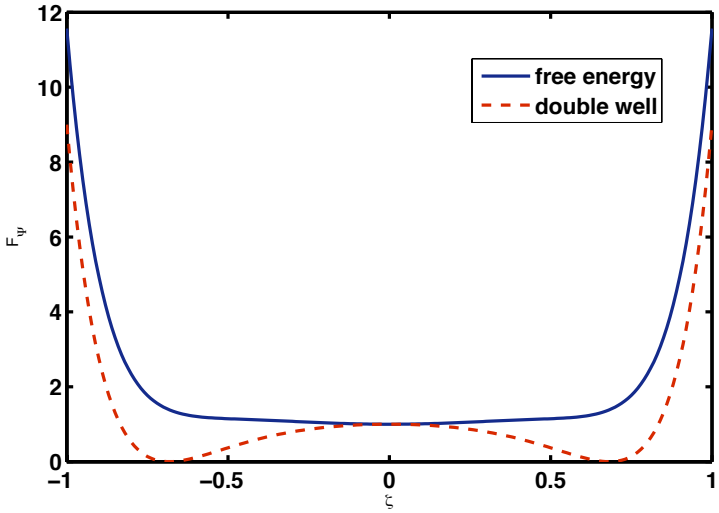

Figure 2. Gauge dependence of the free energy: the plot shows the bistable free energy $F_\Phi(\xi) = (\xi^2 - 1)^2$ after a transformation $\Phi \mapsto \Phi^3 + \Phi$ (solid line); for comparison the dashed line shows the double well potential $V(\xi) = (\xi^2 - 1)^2$ as a function of the transformed variable $\zeta = \xi^3 + \xi$, i.e., without the logarithmic gauge term.

is, in general $F_\Phi(\xi) \neq F_\Psi(\zeta)$, even though $\Phi(x) = \xi$ and $\Psi(x) = \zeta$ correspond to the same microstates. An extreme case in which a maximum of F_Φ is turned into a minimum of F_Ψ will be discussed in Example 1 below.

The gauge dependence of the free energy has been discussed elsewhere in detail, e.g., [3, Sec. 7] or [5, Sec. V], but several features are noteworthy, so we feel it is useful to add a few remarks. To this end let $h\colon \mathbb{R} \to \mathbb{R}$ be a transformation with $h' > 0$ which transforms Φ into $\Psi = h(\Phi)$. If we set $\zeta = h(\xi)$ then

$$\Sigma(\xi) = \{x \in \mathbb{R}^n \colon \Psi(x) = \zeta\},$$

is the same surface as the one defined by $\Phi(x) = \xi$. However the corresponding free energies differ: by chain rule, $\nabla\Psi(x) = h'(\Phi)\nabla\Phi(x)$ which, together with the coarea formula (2.6), entails that

$$F_\Psi(\zeta) = F_\Phi(\xi) + \beta^{-1} \ln h'(\xi)\,. \tag{2.10}$$

The transformation behaviour should not come as a surprise if we keep in mind that $\exp(-\beta F_\Phi)$ is the marginal distribution of the collective variable Φ. However, it has the consequence that critical points of the free energy landscape (e.g., local minima or saddle points) have no coordinate independent meaning. We illustrate the gauge dependence in the following simple example.

Example 1 *Suppose $x \in \mathbb{R}$. Let $\Phi(x) = x$ and $\Psi(x) = x^3 + x$ be two collective variables that are related by the transformation $h(w) = w^3 + w$ that has the property that $h'(w) \neq 0$ for all $w \in \mathbb{R}$. Equation (2.10) now reads*

$$F_\Psi(\zeta) = F_\Phi(r_\zeta) + \beta^{-1} \ln(3r_\zeta^2 + 1)\,,$$

where r_ζ is the real root of the cubic equation $r^3 + r = \zeta$. Now suppose that $F_\Phi(\xi) = (\xi^2 - 1)^2$ is a bistable energy function with two minima at $\xi = \pm 1$ and a local maximum

at $\xi = 0$. However the transformed free energy F_Ψ has a global minimum at zero, but no further minima as can be seen in Figure 2). For comparison, the dashed line in the figure shows the double well potential $V(\xi) = (\xi^2 - 1)^2$ in the transformed variable $\zeta = \xi^3 + \xi$, ignoring the logarithmic gauge term in equation (2.10).

2.3 Reversible work and the potential of mean force

The thermodynamic potential *free energy* is also known as the *potential of mean force*. Noting that

$$\begin{aligned}\nabla\delta(\Phi(x) - \xi) &= \nabla\Phi(x)\delta'(\Phi(x) - \xi) \\ &= -\nabla\Phi(x)\frac{\partial}{\partial\xi}\delta(\Phi(x) - \xi)\end{aligned}$$

which entails the formal identity (that can be made precise by replacing δ by its mollifier δ_ε)

$$\frac{\partial}{\partial\xi}\delta(\Phi(x) - \xi) = -|\nabla\Phi(x)|^{-2}\nabla\Phi(x) \cdot \nabla\delta(\Phi(x) - \xi)\,, \tag{2.11}$$

we can differentiate $F_\Phi(\xi)$ with respect to ξ and integrate by parts, by which we find

$$F'_\Phi(\xi) = \frac{1}{Z_\Phi(\xi)}\int_{\mathbb{R}^n}\left(\frac{\nabla\Phi\cdot\nabla V}{|\nabla\Phi|^2} - \beta^{-1}\nabla\cdot\left(\frac{\nabla\Phi}{|\nabla\Phi|^2}\right)\right)\exp(-\beta V(x))\delta(\Phi(x) - \xi)\,dx\,.$$

If we recall that the states $x \in \mathbb{R}^n$ of our system are random with probability distribution $\rho(x) = \exp(-\beta V(x))/Z$, it follows that $\rho(x)\delta(\Phi(x)-\xi)$ is the joint distribution of x and Φ. Moreover since $\rho_\Phi = Z_\Phi/Z$ is the distribution of Φ, we recognise

$$\mathbf{P}\left[x \in [y, y + dy)\,|\,\Phi(x) = \xi\right] = \frac{1}{Z_\Phi(\xi)}\exp(-\beta V(y))\delta(\Phi(y) - \xi)\,dy$$

as the conditional probability distribution of x given that $\Phi(x) = \xi$. Hence the right-hand side in the expression for F'_Φ turns out to be the conditional expectation

$$F'_\Phi(\xi) = \mathbf{E}\left[\frac{\nabla\Phi\cdot\nabla V}{|\nabla\Phi|^2} - \beta^{-1}\nabla\cdot\left(\frac{\nabla\Phi}{|\nabla\Phi|^2}\right)\,\Big|\,\Phi(x) = \xi\right]. \tag{2.12}$$

Calling

$$f_\Phi(x) = \left(\frac{\nabla\Phi\cdot\nabla V}{|\nabla\Phi|^2} - \beta^{-1}\nabla\cdot\left(\frac{\nabla\Phi}{|\nabla\Phi|^2}\right)\right)\nabla\Phi,$$

the thermodynamic force in the direction of Φ, we find that $f_\Phi = f_\Phi^{(1)} + f_\Phi^{(2)}$ involves essentially two different contributions, the first of which is gauge invariant and a second one that is not. We start with the first one: letting $n(x)$ denote the unit normal to Σ at x, we see that

$$f_\Phi^{(1)}(x) = \big(n(x)\cdot\nabla V(x) - \beta^{-1}\nabla\cdot n(x)\big)\,n(x),$$

depends only on the unit normal that is independent of how the collective variable Φ is chosen. The divergence term $\nabla\cdot n$ is known as the mean curvature; roughly speaking, the mean curvature is the normal derivative of the surface element of $\Sigma(\xi) \subset \mathbb{R}^n$ and

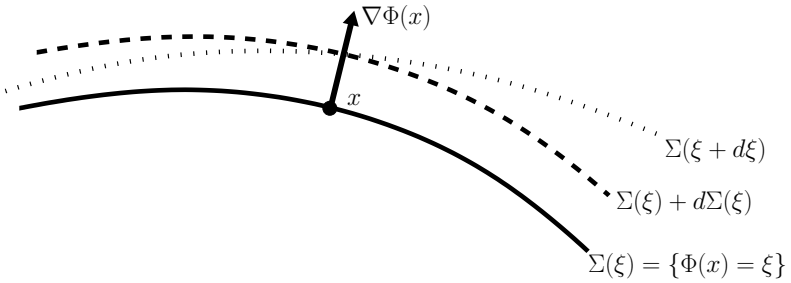

Figure 3. The distance between $\Sigma(\xi)$ and its parallel surface $\Sigma(\xi) + d\Sigma(\xi)$ is constant, which is not true for $\Sigma(\xi + d\xi)$ defined by $\Phi(x) = \xi + d\xi$ that is not parallel to $\Sigma(\xi)$.

hence represents inertial forces that are induced by the variation of the local area element [17]. For the the second term, we have

$$f_\Phi^{(2)}(x) = \left(\beta^{-1} n(x) \cdot \nabla \ln |\nabla\Phi(x)|\right) n(x)\,,$$

which is the directional derivative of the Fixman potential along the normal, hence a pure ambient-space term; it accounts for the density of states in the neighbourhood of the macrostate $\{\Phi(x) = \xi\}$ as is illustrated in Figure 1 (compare also the discussion in Section 2.2).

3 Another definition of free energy

We have seen that the free energy F_Φ is the potential of the mean (thermodynamic) force. As such it encodes the statistical properties of a system in thermal equilibrium. But can it also describe the effective dynamics of a molecular system as is commonly asserted (e.g., see [18,19,20,21])?

An obvious drawback of F_Φ is that it is not gauge invariant in the sense that it is not invariant under a change of coordinates, so that identical microscopic states may be assigned to different potentials of mean force. Moreover, as a consequence of (2.10), the derivative of F_Φ does not transform like a proper gradient which entails that critical points, e.g., local minima or saddle points (i.e., candidates for transition states) are not preserved under a change of coordinates (cf. the discussion in [3]).

3.1 Geometric definition of free energy

An obvious way to fix the gauge problem is to replace the delta function in (2.3) by a proper surface measure, i.e., to define a free energy of the form

$$G(\xi) = -\beta^{-1} \ln M(\xi)\,, \quad M(\xi) = \int_{\Sigma(\xi)} \exp(-\beta V)\, d\sigma\,. \tag{3.1}$$

In view of the coarea formula (2.6), the obvious difference between F and G lies in the Fixman potential: by changing the potential according to $V \mapsto V \pm \beta^{-1} \ln |\nabla\Phi|^{-1}$ we can switch back and forth between the two definitions. As the free energy thus defined only depends on the ensemble of microstates, i.e., on the particular level set $x \in \Sigma(\xi)$, we shall term it *geometric free energy* (cf. [8]).

In analogy with the marginal property (2.4), we notice that $\exp(-\beta G(\xi))/Z$ is the probability density of the surface $\Sigma(\xi)$. If we denote by $\Sigma(\xi) + d\Sigma(\xi)$ the parallel surface to $\Sigma(\xi)$ that is $d\xi$ away from it, and let $Q(\Sigma(\xi), d\Sigma(\xi))$ be the half-open set of points lying between $\Sigma(\xi)$ and $\Sigma(\xi) + d\Sigma(\xi)$, then (see Fig. 3)

$$\frac{\exp(-\beta G(\xi))}{Z}\, d\xi = \mathbf{P}\left[x \in Q(\Sigma(\xi), d\Sigma(\xi))\right],$$

is the probability of finding states in the infinitesimal slab $Q(\Sigma(\xi), d\Sigma(\xi))$. By the coarea formula (2.6), equation (3.1) is equivalent to

$$G(\xi) = -\beta^{-1} \ln \int_{\mathbb{R}^n} \exp(-\beta V(x)) |\nabla \Phi(x)| \delta(\Phi(x) - \xi)\, dx\,. \tag{3.2}$$

Therefore we can compute the derivative of G just as before using (2.11). This yields

$$G'(\xi) = \frac{1}{M(\xi)} \int_{\Sigma(\xi)} \left(\frac{\nabla \Phi \cdot \nabla V}{|\nabla \Phi|^2} - \beta^{-1} |\nabla \Phi|^{-1} \nabla \cdot \left(\frac{\nabla \Phi}{|\nabla \Phi|} \right) \right) \exp(-\beta V)\, d\sigma\,,$$

which should be compared to (2.12). Similarly to the previous section, we recognise the right-hand side of the equation as the expectation with respect to the (gauge-invariant) canonical distribution $\rho(x)$ conditional on $x \in \Sigma(\xi)$. We write this as

$$G'(\xi) = \mathbf{E}_\Sigma \left[\frac{\nabla \Phi \cdot \nabla V}{|\nabla \Phi|^2} - \beta^{-1} |\nabla \Phi|^{-1} \nabla \cdot \left(\frac{\nabla \Phi}{|\nabla \Phi|} \right) \,\middle|\, \xi \right]. \tag{3.3}$$

Keeping the terminology of the previous section, the corresponding force

$$g_\Phi(x) = \left(\frac{\nabla \Phi \cdot \nabla V}{|\nabla \Phi|^2} - -\beta^{-1} |\nabla \Phi|^{-1} \nabla \cdot \left(\frac{\nabla \Phi}{|\nabla \Phi|} \right) \right) \nabla \Phi,$$

in the direction of Φ can be seen to be gauge invariant; indeed,

$$g_\Phi(x) = \big(n(x) \cdot \nabla V(x) - \beta^{-1} \nabla \cdot n(x) \big)\, n(x)$$

equals the gauge invariant part $f_\Phi^{(1)}$ in (2.12). Note that G' transforms like a proper gradient field.

The reader may be tempted to think that the geometric free energy G rather than the Helmholtz free energy F_Φ is the effective potential governing the *dynamics* of a collective variable, because its derivative G' behaves like a force under transformations of the collective variable, whereas F'_Φ does not. As we will show below, G is related to the transition rates between metastable states. There are other instances in which G turns out to be the effective potential that drives the dynamics of a collective variable, e.g., when the collective variable is adiabatically separated from the remaining degrees of freedom [22] or in the optimal prediction of Hamiltonian systems [23,16].

3.2 Effective dynamics in the free energy landscape

In general the question as to whether G or F_Φ drives some coarse-grained dynamics does not allow for a straightforward answer, the reason being that there is neither a unique coarse-graining procedure nor a unique way of "measuring dynamics". To understand this point, suppose that the microscopic dynamics are governed by the following overdamped Langevin equation

$$\dot{x} = -\nabla V(x) + \sqrt{2\sigma}\dot{w}\,, \quad x(0) = x\,, \tag{3.4}$$

that generates dynamics that are ergodic with respect to the canonical ensemble. Here $\dot{w}$ denotes n-dimensional Gaussian white noise with coefficient $\sigma = k_B T$ (i.e., w is a Brownian motion). Given the solution $x(t)$ and using Itô's formula [24, Sec. 4.2]), it follows that $\Phi(x(t))$ evolves according to the equation

$$\dot{\Phi}(x(t)) = -\nabla\Phi(x(t)) \cdot \nabla V(x(t)) + \sigma\Delta\Phi(x(t)) + \sqrt{2\sigma}\nabla\Phi(x(t)) \cdot \dot{w}(t)\,. \tag{3.5}$$

Clearly, the evolution equation for $\xi(t) = \Phi(x(t))$ depends on $x(t)$ and we may ask for an appropriate closure scheme to find an equation for ξ alone. Since ξ is a stochastic process, there are various ways to define a process, say, z such that

$$z(t) \approx \xi(t)\,.$$

One possible choice that has been proposed in [25] is to construct $z(t)$ so that its probability distribution at time $t \geq 0$ coincides with the probability distribution of $\xi(t) = \Phi(x(t))$. Although appealing, such a closure requires that the exact time-dependent marginal distribution of $\Phi(x(t))$ is known which typically is not the case. Another possibility is to replace the right hand side of (3.5) by its conditional average $\mathbf{E}[\cdot|\Phi(x) = z]$ which essentially amounts to computing the least squares approximation of the forces acting on Φ (see, e.g., [26]). This possibility, which has been analysed in [27], yields a stochastic differential equation for z of the form,

$$\dot{z} = b(z) + a(z)\dot{\eta}, \tag{3.6}$$

where $\dot{\eta}$ is one-dimensional Gaussian white noise. The coefficients are the average force

$$b(z) = \mathbf{E}\left[-\nabla\Phi(x) \cdot \nabla V(x) + \sigma\Delta\Phi(x)\big|\Phi(x) = z\right]$$

and the average variance

$$(a(z))^2 = 2\sigma\mathbf{E}\left[|\nabla\Phi(x)|^2\big|\Phi(x) = z\right]\,.$$

By construction of the coarse-graining scheme using conditional expectations, the invariant distribution of (3.6) is the marginal distribution $\rho_\Phi \propto \exp(-\beta F_\Phi)$, but in general the average force b will be different from F'_Φ or G' as given by the equations (2.12) or (3.3); for a multidimensional collective variables it need not even be the derivative of any potential, i.e., it will not be a gradient field.[3]

3.3 Transition state theory

In the derivation of transition rates for thermostatted Hamiltonian systems, G comes out naturally (cf. [29,30]): let $A \subset \mathbb{R}^n$ be an open set that is a subset of the configuration space with a sufficiently smooth boundary and define $A^c = \mathbb{R}^n \setminus A$ to be its (closed) complement in $\mathbb{R}^n$. Further assume that the dynamics are given by the Langevin equation,

$$\begin{aligned}\dot{x} &= v,\\ \dot{v} &= -\nabla V(x) - \gamma v + \sigma\dot{w}\,,\end{aligned}$$

with $\gamma \in \mathbb{R}^{n\times n}$ being symmetric, positive definite and satisfying the fluctuation dissipation relation $2\gamma = \beta\sigma\sigma^T$ for some invertible noise matrix $\sigma \in \mathbb{R}^{n\times n}$. As before,

[3] If Φ is scalar, it is in fact possible to do a nonlinear change of variables (sometimes called a "volatility transformation"), such that $a = \sqrt{2\sigma}$ becomes constant and, as a result, $b = -\nabla F$. This, however, does not contradict any of the previous statements (cf. also [3,28]).

$\dot{w}$ denotes an n-dimensional Gaussian white noise process. Now let $(x, v)\colon [0,\infty) \to \mathbb{R}^n \times \mathbb{R}^n$ be a generic trajectory, assuming the potential $V(x)$ to be sufficiently confining, so that trajectories cannot escape to infinity. Letting N_T denote the number of times that $x(t)$ has crossed the boundary ∂A of A for $t < T$, the transition rate is defined as half the mean frequency of crossing ∂A, i.e.,

$$k_A = \lim_{T\to\infty} \frac{N_T}{2T}\,.$$

To obtain a handier expression for the rate, we follow the derivation given in [5]. For this purpose we represent the boundary of A as the zero-level set,

$$\partial A = \{x \in \mathbb{R}^n\colon \varphi(x) = 0\},$$

of a suitable smooth function $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ with the agreement that $\varphi(x) > 0$ when $x \in A$ and $\varphi(x) \le 0$ otherwise. We further denote by χ_A the indicator function of the set A and by $d\mathcal{S}(x)$ the surface element of its boundary, ∂A. Using that $\chi_A(x) = \theta(\varphi(x))$ where $\theta\colon \mathbb{R} \to \{0, 1\}$ is the Heaviside step function, the rate can be recast as

$$\begin{aligned} k_A &= \lim_{T\to\infty} \frac{1}{2T} \int_0^T \left| \frac{d}{dt}\chi_A(x(t)) \right| dt \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_0^T \left| \frac{d}{dt}\theta(\varphi(x(t))) \right| dt \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_0^T |\dot{x}(t)\cdot\nabla\varphi(x(t))|\; \delta(\varphi(x(t)))\, dt\,, \end{aligned}$$

where in accordance with the considerations in Section 2.1, the relation $\theta'(\varphi) = \delta(\varphi)$ should be understood in the sense of sharp-interface limit of a mollified step function. If we call $\mu(x, v) = \rho(x)\eta(v)$ with $\eta = \mathcal{N}(0, \beta^{-1}I)$ the joint equilibrium distribution of (x, v), then ergodicity of the Langevin dynamics entails that the rate is equal to

$$\begin{aligned} k_A &= \frac{1}{2} \int_{\mathbb{R}^n\times\mathbb{R}^n} |v\cdot\nabla\varphi(x)|\; \delta(\varphi(x))\, \tilde{\rho}(x, v) dx dv \\ &= \frac{1}{2} \int_{\partial A\times\mathbb{R}^n} |v\cdot n(x)| \tilde{\rho}(x, v)\, d\mathcal{S}(x) dv \\ &= \frac{1}{Z}\sqrt{\frac{2}{\pi\beta}} \int_{\partial A} \exp(-\beta V)\, d\mathcal{S}\,, \end{aligned} \tag{3.7}$$

where we have employed the coarea formula (2.6) in the second line, and the third line has followed from integrating the Gaussian density $\eta(v)$ over the velocities normal to the dividing surface. The remaining integral is basically the partition function M of the geometric free energy $G = -\beta^{-1}\ln M$ evaluated at the boundary of A.[4]

Remark 1 *The thus defined transition rate is exact as far as the transitions between A and A^c are concerned. However in many cases one is rather interested in the transition rates between two disjoint sets A and $B \subset A^c$, in which case k_A may be a poor approximation of transition rate k_{AB} between A and B, because a typical*

[4] Within the Bennett-Chandler approach to transition state theory (e.g., see [31]), the rate is typically expressed in terms of the other free energy, F_Φ. Gauge-invariance of the rate then requires normalisation by a factor $\mathbf{E}[|\nabla\Phi|\,|\,\Phi = \xi]$ which yields an expression that is equivalent to (3.7); cf. also (4.1) below or [3, Eqn. (48)].

trajectory that leaves A will probably re-enter A, before going to B. To get a better estimate of k_{AB}, it is therefore desirable to find an optimal dividing surface over all hypersurfaces $\Sigma \subset \mathbb{R}^n$ separating A from B that minimises k_{AB}, and it can be shown that an optimal dividing surfaces must be a stationary point of $G = G[\varphi]$, where $G[\varphi]$ is understood as a functional of the level set function φ that specifies the dividing surface [32]; cf. also [5] and the references given there.

4 Blue Moon ensemble

The difference between F_Φ and G highlights another important aspect that we have partially addressed in Section 2.1. In the seminal work [33], Fixman addressed the problem of how to compute so-called unbiased averages for polymeric fluids models that are subject to holonomic constraints.[5] The difference between the two free energies can be understood in the same fashion. To this end recall the definition (2.3) of the free energy F_Φ, which, using the coarea formula (2.5), can be recast as

$$\begin{aligned} F_\Phi(\xi) &= -\beta^{-1} \ln \int_{\mathbb{R}} \exp(-\beta V(x)) \delta(\Phi(x) - \xi)\, dx \\ &= -\beta^{-1} \ln \int_{\Sigma(\xi)} \exp(-\beta V) |\nabla \Phi|^{-1} d\sigma \\ &= G(\xi) - \beta^{-1} \ln \left(\frac{1}{M(\xi)} \int_{\Sigma(\xi)} \exp(-\beta V) |\nabla \Phi|^{-1} d\sigma \right), \end{aligned}$$

where the last equality is a straight consequence of the definition of G, equation (3.1). The term inside the logarithm can be understood as an expected value with respect to the Boltzmann distribution *constrained* to the submanifold $\Sigma(\xi)$ where $M(\xi)$ simply normalises the level set probability density to one. If we call

$$\mathbf{E}_\Sigma[f|\xi] = \frac{1}{M(\xi)} \int_{\Sigma(\xi)} f \exp(-\beta V)\, d\sigma$$

the expectation of an observable f with respect to the constrained Boltzmann distribution, the above equation turns into (cf. [4, Eqn. 31])

$$F_\Phi(\xi) = G(\xi) - \beta^{-1} \ln \mathbf{E}_\Sigma \left[|\nabla \Phi|^{-1} \,|\, \xi \right] . \tag{4.1}$$

The relation between the constrained expectation $\mathbf{E}_\Sigma[\cdot|\xi]$ thus defined and the conditional expectation $\mathbf{E}[\cdot|\Phi(x) = \xi]$, and its calculation using holonomic constraints is the topic of the next subsection.

4.1 Conditional probabilities and holonomic constraints

Suppose we want to compute the expected value of an observable f with respect to the Boltzmann distribution ρ. By the total law of expectation, the expectation can be recast as

$$\int_{\mathbb{R}^n} f(x)\rho(x)\, dx = \int_{\mathbb{R}} \mathbf{E}_\Phi[f|\xi] \exp(-\beta F_\Phi(\xi))\, d\xi\,, \tag{4.2}$$

[5] More precisely, Fixman compared the equilibrium distribution of a polymer chain with fixed bond lengths to one with stiff, but flexible bonds. The phrase "unbiased" expresses his viewpoint that the constrained model were merely an approximation to the more realistic, flexible model. Even without adopting Fixman's (unjustified) viewpoint, the mathematical problem of understanding how constraints affect statistical distributions remains meaningful.

where $\mathbf{E}_\Phi[f|\xi]$ is shorthand for the conditional expectation

$$\mathbf{E}[f(x)|\Phi(x)=\xi] = \frac{1}{Z(\xi)}\int_{\mathbb{R}} f(x)\exp(-\beta V(x))\delta(\Phi(x)-\xi)\,dx\,. \tag{4.3}$$

In many circumstances, the numerical evaluation of the total equilibrium distribution ρ is costly, while the conditional expectations $\mathbf{E}_\Phi[\cdot|\xi]$ are relatively easy to obtain (e.g., if Φ is an extremely metastable direction or spreads over regions of very low probability). Equation (4.2) states that, if just F_Φ is available, then the total expectation can be computed as a weighted average of the conditional expectations, which is known as thermodynamic integration (e.g., see [34,35,2])

Given a method that allows for sampling $\rho \propto \exp(-\beta V)$, say, a Monte-Carlo scheme, the obvious strategy for computing $\mathbf{E}_\Phi[\cdot|\xi]$ would be to let the sampler sample only the subspace $\Sigma(\xi) = \{x \in \mathbb{R}^n : \Phi(x) = \xi\}$. For a large class of sampling schemes such as hybrid Monte-Carlo [36], over- or underdamped Langevin dynamics [37,38] this can be done by imposing a holonomic constraint $\Phi(x) = \xi$ on the system. Strictly speaking a *constraint* is a set of admissible states $x \in \Sigma(\xi)$ and is intrinsic in that it always defines the same set of states, no matter whether it is defined by an algebraic equation $\Phi(x) = \xi$ or by any other means. Hence, under suitable assumptions, any constrained dynamics that are ergodic with respect to the Boltzmann distribution will sample the constrained expectation $\mathbf{E}_\Sigma[\cdot|\xi]$ rather than the conditional one.

As we have argued using the coarea formula (2.5)–(2.6), conditional and marginal distributions are not intrinsic to the constraint surface, whereas $\mathbf{E}_\Sigma[\cdot|\xi]$ is. We will call the problem of sampling the conditional distribution using constrained dynamics (or fake-dynamics such as Monte-Carlo) the *bias problem* [39]. For second-order differential equations, the bias problem has its origin in the hidden constraint $\dot\Phi(x) = 0$ that any system with velocities satisfies in addition to $\Phi(x) = \xi$. The bias can be removed by a suitable reweighting strategy as has been discussed in the *blue moon* article [40]. The bias problem, however, remains if the dynamics is purely configurational, e.g., in case of overdamped Brownian motion. To understand this, we recast (4.3) as

$$\begin{aligned}\mathbf{E}_\Phi[f|\xi] &= \frac{1}{Z(\xi)}\int_{\mathbb{R}} f(x)\exp(-\beta V(x))\delta(\Phi(x)-\xi)\,dx \\ &= \left(\int_{\Sigma(\xi)} |\nabla\Phi|^{-1}\exp(-\beta V)\,d\sigma\right)^{-1}\int_{\Sigma(\xi)} f|\nabla\Phi|^{-1}\exp(-\beta V)\,d\sigma\end{aligned}$$

where we have employed the coarea formula (2.6) to go from the first to the second line. Inserting $1 = M(\xi)/M(\xi)$, we immediately find that the fundamental equality

$$\mathbf{E}_\Phi[f|\xi] = \frac{\mathbf{E}_\Sigma\left[f|\nabla\Phi|^{-1}\,|\,\xi\right]}{\mathbf{E}_\Sigma\left[|\nabla\Phi|^{-1}\,|\,\xi\right]}\,, \tag{4.4}$$

relates constrained and conditional expectations and which is known in the literature by the name of *Fixman theorem* [33] or *Blue Moon formula* [40]. Since it is a direct consequence of the coarea formula, this relation holds true no matter whether the system involves momenta or not.[6]

[6] If the kinetic energy of a system that involves momenta carries a nontrivial mass matrix M, then the Euclidean norm $|\cdot|$ in the weight $|\nabla\Phi|$ has to be replaced by the Riemannian metric $|z|_{M^{-1}} = \sqrt{z\cdot M^{-1}z}$. For details, we refer to the textbook [38] or [41] in this issue.

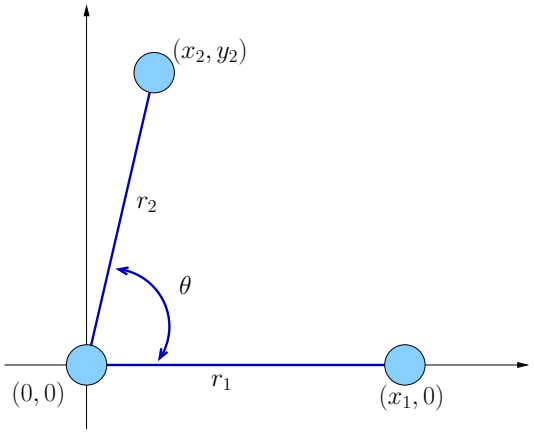

Figure 4. Internal coordinates for the planar triatomic molecule.

Fixman potential reloaded. The Blue Moon Formula (4.4), but also equation (4.1) which relates the two free energies F_Φ and G simply express the fact that the underlying statistical ensembles can be transformed into one another by augmenting the potential according to $V \mapsto V + \beta^{-1} \ln |\nabla \Phi|$. The origin of the Fixman correction $\beta^{-1} \ln |\nabla \Phi|$ is basically the same as in Sections 2.1–2.2, where the correction was revealed as the result of strong confinement. This can be rephrased by saying that the Fixman potential mimics unconstrained equilibrium dynamics, even though the system is constrained [10,11,42]. In this sense, the Fixman potential measures the difference between the two probability measures $\mathbf{P}_\Phi = \mathbf{P}\left[\Phi(x) \in [\xi, \xi + d\xi)\right]$ and $\mathbf{P}_\Sigma = \mathbf{P}\left[x \in Q(\Sigma(\xi), d\Sigma(\xi))\right]$ that correspond to the free energies F_Φ and G, respectively. It is therefore referred to as an *entropic* correction.[7]

The following example is a variant of a classical example from [10] and illustrates that the Fixman potential in (4.1) can arise from a very subtle interplay between the geometry of the macrostate $\{\Phi(x) = \xi\}$ and the potential energy landscape V.

Example 2 *Consider a molecule consisting of three identical particles, as in Figure 4. We shall assume that the particles are confined to the xy-plane and that the interactions between them are invariant under rigid translations and rotations. We further assume that the interaction potential $V = V_\varepsilon$ has the form*

$$V_\varepsilon = U(\theta) + \frac{1}{2\varepsilon} \sum_{i=1}^{2} K_i(\theta) \left(r_i - r_{i,0}(\theta)\right)^2 , \tag{4.5}$$

where $0 < \varepsilon \ll 1$ which indicates that the bond lengths r_1, r_2 are stiff. Here Cartesian and internal (i.e., polar) coordinates are related by the coordinate transformation

$$\begin{aligned} \theta &= \arctan(x_2/y_2) \pm \pi \\ r_1 &= x_1 \\ r_2 &= \sqrt{x_2^2 + y_2^2} \, . \end{aligned}$$

In polar coordinates the Boltzmann distribution (2.2) reads

$$d\rho_\varepsilon(\theta, r_1, r_2) = \frac{\exp\left(-\beta V_\varepsilon(\theta, r_1, r_2)\right)}{Z_\varepsilon} \, r_2 \, dr_1 dr_2 d\theta \, .$$

[7] For related results on the realization of holonomic constraints by strong confinement of deterministic (i.e., microcanonical) Hamiltonian systems the reader is referred to [9,43].

We let $\Phi = \theta$ be our collective variable, and we want to compute F_θ and G for small ε. To this end recall that both geometric and conditional free energies are defined as

$$G_\varepsilon(\xi) = -\beta^{-1} \log M_\varepsilon(\xi)\,, \quad F_{\theta,\varepsilon}(\xi) = -\beta^{-1} \log Z_{\theta,\varepsilon}(\xi), \tag{4.6}$$

with the geometric partition function

$$M_\varepsilon(\xi) = \int_0^\infty \int_0^\infty \exp\left(-\beta V_\varepsilon(\xi, r_1, r_2)\right) dr_1 dr_2, \tag{4.7}$$

and the conditional one[8]

$$Z_{\theta,\varepsilon}(\xi) = \int_0^\infty \int_0^\infty \exp\left(-\beta V_\varepsilon(\xi, r_1, r_2)\right) r_2\, dr_1 dr_2\,. \tag{4.8}$$

Let us analyse the conditional partition function $Z_{\theta,\varepsilon}$. To begin with, we argue that the domain of integration can be extended to the whole real line since the extra contribution makes up only a term $\mathcal{O}\left(\exp\left(-1/\varepsilon\right)\right)$. Indeed, applying Laplace's method to the integral over r_2 results in

$$\begin{aligned}
&\int_0^\infty \exp\left(-\frac{\beta}{2\varepsilon} K_2(\xi)\left(r_2 - r_{2,0}(\xi)\right)^2\right) r_2\, dr_2 \\
&\quad \approx \int_{-\infty}^\infty \exp\left(-\frac{\beta}{2\varepsilon} K_2(\xi)\left(r_2 - r_{2,0}(\xi)\right)^2\right) r_2\, dr_2 \\
&\quad = \sqrt{\frac{2\pi\varepsilon}{\beta}}\, \frac{r_{2,0}(\xi)}{\sqrt{K_2(\xi)}}.
\end{aligned}$$

The integration over r_1 is another Gaussian integral. Hence

$$Z_{\theta,\varepsilon}(\xi) \approx r_{2,0}(\xi) M_\varepsilon(\xi)$$

to lowest order in ε where M_ε is given by

$$M_\varepsilon(\xi) = \frac{2\pi\varepsilon}{\beta}\, \frac{\exp(-\beta U(\xi))}{\sqrt{K_1(\xi) K_2(\xi)}}.$$

Clearly, both $\ln M_\varepsilon$ and $\ln Z_{\theta,\varepsilon}$ approach $-\infty$ as $\varepsilon \to 0$. Yet the difference between the two free energies (4.6) is well defined, viz.,

$$\lim_{\varepsilon\to 0} \left(F_{\theta,\varepsilon}(\xi) - G_\varepsilon(\xi)\right) = -\beta^{-1} \ln r_{2,0}(\xi)\,.$$

The right-hand side of the last equation is precisely the Fixman correction in (4.1),

$$-\beta^{-1} \ln \mathbf{E}_\Sigma[|\nabla\theta|^{-1}|\xi] = -\beta^{-1} \ln r_{2,0}(\xi) + \mathcal{O}(\sqrt{\varepsilon})\,.$$

It is interesting to note that the Fixman potential is not a function of $r_{1,0}(\xi)$ and $r_{2,0}(\xi)$, but only a function of the latter. This is intriguing because the three particles are indistinguishable; also from a geometric perspective, the internal state spaces of the two exterior particles are identical copies of the positive real line. This confirms that F_Φ does not have an intrinsic meaning, for had we chosen a different gauge (e.g., with the top particle in Fig. 4 sitting on the y-axis), the Fixman potential would have become a function of $r_{1,0}(\theta)$ rather than $r_{2,0}(\theta)$.

[8] In terms of the Cartesian coordinates $q = (x_1, x_2, y_2)$ and up to conventional additive constants $\pm\pi$, the reaction coordinate reads $\theta(q) = \arctan(x_2/y_2)$ which implies $|\nabla\theta|^{-1} = r_2$ and explains the extra "r_2" in the expression for $Z_{\theta,\varepsilon}$.

5 Conclusions

In this article we have reviewed two distinct free energy concepts for collective variables that circulate in the literature. The more common probabilistic definition in terms of marginal probabilities encodes the complete statistics of a system in thermal equilibrium. The second, geometric definition is related to the probability distribution of the state space foliation defined by the collective variable and is covariant under transformations of its parametrization. Although lacking a straightforward interpretation in terms of marginal or conditional probabilities, it appears to be the natural quantity for estimating, e.g., the transition rates between metastable states. We wish to stress that choosing either of the two definitions is not a religious question, but rather a question of the scope of application; both are free energies in their own right. Moreover they are linked by a simple transformation of the potential energy function.

Acknowledgement

This tutorial review and its authors owe a significant debt of gratitude to Eric Vanden-Eijnden who first pointed out the problem of the two possible free energy definitions and who taught us many of the concepts discussed here.

References

1. P. Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, 93:2395–2417, 1993.
2. D. Frenkel and B Smit. *Understanding Molecular Dynamics: From Algorithms to Applications*. Academic Press, London, 2002.
3. W. E and E. Vanden-Eijnden. Metastability, conformation dynamics, and transition pathways in complex systems. In S. Attinger and P. Koumoutsakos, editors, *Multiscale, Modelling, and Simulation*, pages 35–68. Springer, Berlin, 2004.
4. G.K. Schenter, B.C. Garrett, and D.G. Truhlar. Generalized transition state theory in terms of the potential of mean force. *J. Chem. Phys.*, 119(12):5828–5833, 2003.
5. E. Vanden-Eijnden and F.A. Tal. Transition state theory: Variational formulation, dynamical corrections, and error estimates. *J. Chem. Phys.*, 123:184103, 2005.
6. T. Mülders, P. Krüger, W. Swegat, and L. Schlitter. Free energy as the potential of mean constraint force. *J. Chem. Phys.*, 104(12):4869–4870, 1996.
7. W.K. den Otter and W.J. Briels. The calculation of free-energy differences by constrained molecular-dynamics simulations. *J. Chem. Phys.*, 109(11):4139–4146, 1998.
8. C. Hartmann and Ch. Schütte. A comment on two disctinct notions of free energy. *Physica D*, 228(1):59–63, 2007.
9. H. Rubin and P. Ungar. Motion under a strong constraining force. *Comm. Pure Appl. Math.*, 10:65–87, 1957.
10. N.G. van Kampen and J.J. Lodder. Constraints. *Am. J. Phys.*, 52(5):419–424, 1984.
11. E.J. Hinch. Brownian motion with stiff bonds and rigid constraints. *J. Fluid. Mech.*, 271:219–234, 1994.
12. G. Gallavotti. *The Elements of Mechanics*. Springer, New York, 1983.
13. R. Kubo, H. Ichimura, T. Usui, and N. Hashitsume. *Statistical Mechanics*. North-Holland, Amsterdam, 1965.
14. H. Federer. *Geometric Measure Theory*. Springer, Berlin, 1969.
15. L.C. Evans and R.F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.
16. C. Hartmann. *Model Reduction in Classical Molecular Dynamics*. PhD Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 2007.

17. K. Ecker. *Regularity Theory for Mean Curvature Flow*, volume 75 of *Progress in nonlinear differential equations and their applications*. Birkhäuser, Boston, 2004.
18. G. Hummer and I.G. Kevrekidis. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.*, 118:10762–10773, 2003.
19. S. Yang, J.N. Onuchic, and H. Levine. Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.*, 125:054910, 2006.
20. R. Hegger and G. Stock. Multidimensional langevin modeling of biomolecular dynamics. *J. Chem. Phys.*, 130(3):034106, 2009.
21. Robert B. Best and Gerhard Hummer. Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. USA*, 107(3):1088–1093, 2010.
22. C. Hartmann. Balanced model reduction of partially observed langevin equations: an averaging principle. *Math. Comput. Model. Dyn. Syst. (to appear)*, 2011.
23. A.J. Chorin, O.H. Hald, and R. Kupferman. Optimal prediction with memory. *Physica D*, 166:239–257, 2002.
24. B. Øksendal. *Stochastic differential equations : an introduction with applications*. Springer, Berlin, 2003.
25. I. Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an Ito differential. *Probab. Th. Rel. Fields*, 71(4):501–516, 1986.
26. A.J. Chorin and O.H. Hald. *Stochastic Tools in Mathematics and Science*. Springer, New York, 2006.
27. F. Legoll and T. Lelièvre. Effective dynamics using conditional expectations. *Nonlinearity*, 23:2131–2163, 2010.
28. L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.*, 125:024106, 2006.
29. H. Eyring. The activated complex in chemical reactions. *J. Chem. Phys.*, 3:107–115, 1935.
30. E.P. Wigner. Calculation of the rate of elementary association reactions. *J. Chem. Phys.*, 5:720–725, 1937.
31. D. Chandler. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem Phys.*, 68:2959–2970, 1978.
32. J. Horiuti. On the statistical mechanical treatment of the absolute rate of chemical reaction. *Bull. Chem. Soc. Jpn.*, 13(1):210–216, 1938.
33. M. Fixman. Classical statistical mechanics of constraints: a theorem and applications to polymers. *Proc. Natl. Acad. Sci. USA*, 71:3050–3053, 1974.
34. J.G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3:300–313, 1935.
35. T.P. Straatsma and J. A. McCammon. Multiconfiguration thermodynamic integration. *J. Chem. Phys.*, 95:1175–1118, 1991.
36. C. Hartmann. An ergodic sampling scheme for constrained Hamiltonian systems with applications to molecular dynamics. *J. Stat. Phys.*, 130(4):687–712, 2008.
37. G. Ciccotti, T. Lelièvre, and E. Vanden-Eijnden. Projection of diffusions on submanifolds: Application to mean force computation. *Commun. Pure Appl. Math.*, 61(3):371–408, 2008.
38. T. Lelièvre, M. Rousset, and G. Stoltz. *Free Energy Computations: A Mathematical Perspective*. Imperial College Press, London, 2010.
39. G. Ciccotti, R. Kapral, and E. Vanden-Eijnden. Blue moon sampling, vectorial reaction coordinates, and unbiased constrained dynamics. *ChemPhysChem*, 6(9):1809–1814, 2005.
40. E.A. Carter, G. Ciccotti, J.T. Hynes, and R. Kapral. Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.*, 156(5):472–477, 1989.
41. J. Walter, C. Hartmann, and J. Maddocks. Ambient space formulations and statistical mechanics of holonomically constrained Lagrangian systems. *Eur. Phys. J. ST*. This issue.

42. S. Reich. Smoothed dynamics of highly oscillatory Hamiltonian systems. *Physica D*, 89:28–42, 1995.
43. F.A. Bornemann. *Homogenization of Singularly Perturbed Mechanical Systems*, volume 1687 of *Lecture Notes in Mathematics*. Springer, Berlin, 1998.