

Transition state theory: Variational formulation, dynamical corrections, and error estimates

Eric Vanden-Eijnden^{a)}*Courant Institute of Mathematical Sciences, New York University, New York, New York 10012*Fabio A. Tal^{b)}*Instituto de Matemática e Estatística, Universidade de São Paulo, SP 06608-900, São Paulo, Brazil*

(Received 18 February 2005; accepted 9 September 2005; published online 7 November 2005)

Transition state theory (TST) is revisited, as well as evolutions upon TST such as variational TST in which the TST dividing surface is optimized so as to minimize the rate of recrossing through this surface and methods which aim at computing dynamical corrections to the TST transition rate constant. The theory is discussed from an original viewpoint. It is shown how to compute exactly the mean frequency of transition between two predefined sets which either partition phase space (as in TST) or are taken to be well-separated metastable sets corresponding to long-lived conformation states (as necessary to obtain the actual transition rate constants between these states). Exact and approximate criteria for the optimal TST dividing surface with minimum recrossing rate are derived. Some issues about the definition and meaning of the free energy in the context of TST are also discussed. Finally precise error estimates for the numerical procedure to evaluate the transmission coefficient κ_S of the TST dividing surface are given, and it is shown that the relative error on κ_S scales as $1/\sqrt{\kappa_S}$ when κ_S is small. This implies that dynamical corrections to the TST rate constant can be computed efficiently if and only if the TST dividing surface has a transmission coefficient κ_S which is not too small. In particular, the TST dividing surface must be optimized upon (for otherwise κ_S is generally very small), but this may not be sufficient to make the procedure numerically efficient (because the optimal dividing surface has maximum κ_S , but this coefficient may still be very small). © 2005 American Institute of Physics. [DOI: 10.1063/1.2102898]

I. INTRODUCTION

Dynamical systems often display a few long-lived preferred conformation states between which the systems only switch once in a while. Examples include chemical reactions, conformational changes of molecules, nucleation events in phase transition, etc. The reason is the existence of dynamical bottlenecks which confine the system for very long periods of time in some regions in phase space, usually referred to as metastable states. In many cases, the waiting times in the metastable states are very long compared to the typical molecular time scale. In these situations, the main objectives become the identification of the mechanisms by which the system hops from one metastable state to another (i.e., the identification of the dynamical bottlenecks between these states) and of the rate constants at which transitions between these states occur.

Transition state theory¹⁻³ (TST) (see also Refs. 4-6) is the earliest attempt in this direction. TST gives an exact expression for the mean frequency of transition between any two sets partitioning phase space. TST also gives the mean residency times in each of these sets. The inverses of the TST residency times give a first approximation of the transition rate constants between the metastable states. Unfortunately this approximation can be quite poor. The reason is

that the trajectory can cross many times the TST dividing surface in the course of one transition between the metastable states or even reach the dividing surface without actually making a transition between these states. As a result, the TST transition rate constants always overestimate the actual rate constants, sometimes significantly so.

Two complementary strategies have been proposed to improve TST. The first, which was originally suggested by Horiuti³ and is now referred to as variational TST,^{5,7,8} amounts to choosing the dividing surface which minimizes the TST rate constant. Since TST always overestimates the rate constant, this dividing surface will give the best estimate that TST can achieve.

The second strategy amounts to discounting crossing events of the dividing surface which are not related to actual transitions between the metastable states. This idea was first proposed by Keck⁹ (see also Ref. 10) and further developed by Bennett¹¹ and Chandler.¹² The dynamical corrections to TST are most conveniently expressed in terms of the transmission coefficient κ_S of the dividing surface S : κ_S is a number between zero and 1 which gives the ratio between the actual rate constant and the TST rate constant of the dividing surface. Since the actual rate constant is independent of the dividing surface, the choice of the dividing surface may seem irrelevant as long as one computes also the transmission coefficient of this surface. In practice, however, this choice is essential because a poorly chosen dividing surface has a very small transmission coefficient which is very dif-

^{a)}Electronic mail: eve2@cims.nyu.edu^{b)}Electronic mail: fabiot@ime.usp.br

difficult to estimate accurately. Therefore, the transmission coefficient κ_S should always be computed from the variational TST dividing surface which, by construction, has the largest κ_S .

So, do these improvements upon TST terminate all issues about the theory? Not quite. First while the idea behind variational TST has been applied to many situations, most of the time the optimization is done on a case to case basis in a nonsystematic way. As a result there is still a lack of a methodology which would be both general and practical to identify the variational TST dividing surface. In addition, many of the works on variational TST are based on a misconception, namely, that the variational TST dividing surface is the one that maximizes the free energy of an associated reaction coordinate. This assertion is incorrect for reasons that we elucidate below which involve some confusion about terminology (what is the free energy of a reaction coordinate?). In this paper we derive the equations for the variational TST dividing surface within specific classes, e.g., hyperplanes, and indicate how to design practical algorithms to identify this variational TST surface. At the same time, we elucidate the relation between variational TST and the free energy and discuss in detail some issues of terminology regarding the free energy.

Another issue concerns the practical validity of the procedure to compute dynamical corrections to TST due to the lack of *a priori* error estimate on this procedure. In this paper, we give a precise error estimate on the transmission coefficient κ_S of a dividing surface S in terms of the number of trajectories used to estimate κ_S . To do so, we formulate the question of computing the dynamical corrections in a nontraditional way. In the traditional approach based on correlation functions, fluctuation-dissipation theorem, and Onsager's relation, the actual rate constant of the transition is given by the plateau value of a time-dependent rate constant. It is important to realize, however, that this traditional approach only produces a well-defined rate constant in the limit of infinite separation between the molecular time scale and the transition time scale (only in this limit does a flat plateau exist). In any realistic situation, this separation of time scale is only approximate, and therefore the identification of the rate by the plateau becomes ambiguous. As a result it becomes difficult to distinguish between the errors in the rate constant due to finite separation of time scale and the ones due to finite-size sampling in the numerical procedure. Here we avoid this problem by defining the actual rate constant not as the plateau value of some time-dependent rate constant, but rather as the *exact* rate of transition between two predefined sets representing the metastable sets. This allows us to clearly separate the problem of identifying these sets from the problem of computing the rate of transition between these sets. This way, we can obtain a precise error estimate on the rate constant due solely to the statistical errors induced by finite-size sampling.

The remainder of this paper is organized as follows. In Sec. II we recall the conditions under which it is reasonable to identify rate constants for the transition between metastable sets and reduce the original dynamics to a Markov chain on these sets. In Sec. III, we give an original account

of TST in the context of Langevin dynamics. These are standard results, but we derive them in a way which makes easier the incorporation of dynamical corrections later on. In Sec. IV we revisit variational TST and give the general equations for the variational TST dividing surface within certain classes, e.g., hyperplanes. We also indicate how to design algorithms to identify this surface in practice. In Sec. V, we elucidate the relation between TST and the free energy of a reaction coordinate. In Sec. VI we discuss dynamical corrections to TST and show how to compute exactly the rate of transitions between two predefined sets in configuration space. We also give a new expression for the transmission coefficient κ_S of the dividing surface S in terms of an equilibrium ensemble average restricted to the dividing surface. In Sec. VII, we indicate how to compute the transmission coefficient κ_S in practice and obtain precise error estimates on the result in terms of the number of trajectories used to evaluate κ_S . Finally, some concluding remarks are given in Sec. VIII. This paper is a companion paper to the more mathematical reference.¹³

II. METASTABILITY AND EFFECTIVE DYNAMICS

Most of this paper is devoted to the problem of computing the mean frequency of transitions between two predefined sets in configuration space, see (12) below. In this section we recall why this quantity is relevant in the context of systems displaying metastability.

We shall focus on systems in the *NVT* ensemble and assume that their dynamics can be modeled by the Langevin equation

$$\ddot{X} = -\nabla V(X) - \gamma M^{-1} \dot{X} + \sqrt{2\gamma\beta^{-1}} M^{-1/2} \eta(t). \quad (1)$$

[Here and below we use capital $X=X(t)$ to denote the trajectory solution of (1) and we use lower cases $(x, v) \in \mathbb{R}^n \times \mathbb{R}^n$ to identify a point in phase space.] The discussion below can be generalized to other dynamics consistent with the *NVT* ensemble (such as Nosé-Hoover, etc.) or to the *NVE* ensemble when the dynamics is governed by Hamilton's equation of motion $\ddot{X} = -\nabla V(X)$ (see Ref. 13). (1) is written in mass-weighted coordinate. $V(x)$ is the potential, $\beta=1/k_B T$ is the inverse temperature, γ is the friction coefficient, M is the diagonal mass matrix, and $\eta(t)$ is a white noise satisfying $\langle \eta(t) \otimes \eta(t') \rangle = \text{Id} \delta(t-t')$ (Id is the identity matrix). The dynamics in (1) is ergodic with respect to the Boltzmann-Gibbs probability density function

$$\rho(x, v) = Z_H^{-1} e^{-\beta H(x, v)}, \quad (2)$$

where $H=(1/2)|v|^2+V(x)$ is the Hamiltonian, $Z_H=(2\pi/\beta)^{n/2}Z$ is the partition function, and

$$Z = \int_{\mathbb{R}^n} e^{-\beta V(x)} dx \quad (3)$$

its configurational part. In particular, for any observable $A(x, v)$, we have

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T A(X(t), \dot{X}(t)) dt = \int_{\mathbb{R}^n \times \mathbb{R}^n} A(x, v) \rho(x, v) dx dv, \quad (4)$$

and for any observable $B(x, v, x', v')$ and $s \geq 0$, we have

$$\begin{aligned} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T B(X(t), \dot{X}(t), X(t+s), \dot{X}(t+s)) dt \\ = \int_{\mathbb{R}^n \times \mathbb{R}^n} \mathbf{E} B(x, v, X(s, x, v), \dot{X}(s, x, v)) \rho(x, v) dx dv, \end{aligned} \quad (5)$$

where $\mathbf{E}(\cdot)$ denotes expectation with respect to the noise η and $X(t, x, v)$ denotes the solution of (1) with the initial condition $(X(0, x, v), \dot{X}(0, x, v)) = (x, v)$.

We shall assume that (1) displays bistability, in the sense that the following two properties, (a) and (b), hold:

- (a) There exist two disjoint open sets in configuration space, $a \subset \mathbb{R}^n$ and $b \subset \mathbb{R}^n$, conventionally referred to as the reactant and product states, respectively. The volume of these sets may be small, but we assume that they concentrate almost all the probabilities, i.e., they are such that

$$N_a + N_b \approx 1, \quad (6)$$

where N_a and N_b are the equilibrium population densities in a and b :

$$N_a = Z^{-1} \int_a e^{-\beta V(x)} dx, \quad (7)$$

$$N_b = Z^{-1} \int_b e^{-\beta V(x)} dx.$$

(6) implies that a typical trajectory spends most of its time in a and b and a negligible portion of time in the buffer region $\mathbb{R}^n / (a \cup b)$ between these sets.

- (b) The waiting times between successive hopping events between a and b can effectively be described as independent random variables with a Poisson distribution. This will be the case, e.g., if these waiting times are so large compared to the molecular time scale than the system loses memory between transitions.

The validity of assumptions (a) and (b) can be assessed by analyzing the spectrum of the Fokker-Planck operator associated with (1) and establishing the existence of a spectral gap, see, e.g., Ref. 14. We shall not dwell on this issue here but rather focus on the implication of these assumptions. Note also that the situation with more than two metastable sets, say, $\{a_j\}_{j=1..n}$, can be considered as well by generalizing assumptions (a) and (b) and iterating the argument below on $a_1, \cup_{j \neq 1} a_j$, then $a_2, \cup_{j \neq 2} a_j$, etc.

Under assumptions (a) and (b) the dynamics of the system can effectively be reduced to that of a two-state Markov chain on the sets a and b . This means that it is possible to write down a closed equation which approximates the evolution of the instantaneous population densities in a and b :

$$\dot{n}_a(t) = \mathbf{E}(\chi_a \dot{X}(t)), \quad \dot{n}_b(t) = \mathbf{E}(\chi_b \dot{X}(t)), \quad (8)$$

where $\chi_D(x)$ denotes the indicator function of a set $D \subset \mathbb{R}^n$:

$$\chi_D(x) = \begin{cases} 1 & \text{if } x \in D \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Consistent with the Markov character of the effective dynamics, the evolution of $n_a(t)$ and $n_b(t)$ is governed by the master equation (justified below):

$$\begin{cases} \dot{n}_a(t) = -k_{ab}n_a(t) + k_{ba}n_b(t), \\ \dot{n}_b(t) = -k_{ba}n_b(t) + k_{ab}n_a(t) \end{cases}, \quad (10)$$

where the rates k_{ab} and k_{ba} are given by

$$k_{ab} = 2\nu/N_a, \quad k_{ba} = 2\nu/N_b. \quad (11)$$

Here N_a and N_b are the equilibrium densities in a and b defined in (7), and ν is the mean frequency of hopping between a and b , i.e.,

$$\nu = \lim_{T \rightarrow \infty} \frac{N_T^{ab}}{T}, \quad (12)$$

where N_T^{ab} denotes the number of times a trajectory hops between a and b in the time interval $[0, T]$. Note that, by ergodicity, the limit in (12) exists and does not depend on the specific trajectory.

To justify (11), notice that $t_a = k_{ab}^{-1}$ and $t_b = k_{ba}^{-1}$ are the average times between transitions from a to b and b to a , respectively. Since we know that the trajectory spends a fraction N_a of its time in a and N_b in b and on average makes ν transitions between the sets per unit of time, we must have $t_a = N_a / (2\nu)$ and $t_b = N_b / (2\nu)$. The Markov chain with the rates in (11) is the only two-state chain consistent with this property. Notice also that (10) is consistent with equilibrium since it implies that [using (6)]

$$\lim_{t \rightarrow \infty} n_a(t) = \frac{k_{ba}}{k_{ab} + k_{ba}} = \frac{N_a}{N_a + N_b} \approx N_a, \quad (13)$$

$$\lim_{t \rightarrow \infty} n_b(t) = \frac{k_{ab}}{k_{ab} + k_{ba}} = \frac{N_b}{N_a + N_b} \approx N_b.$$

The equilibrium densities N_a and N_b can be deduced from the free energy computed, e.g., by blue-moon sampling and thermodynamic integration,^{15,16} see Sec. V. The difficult part is to estimate the mean frequency of hopping ν , and this is the subject of most of the remainder of this paper. Notice that the viewpoint taken above where one computes the mean frequency of transition between predefined sets (the reactant state a and the product state b) is similar to the stable state picture (SSP) originally developed in Ref. 17 and 18 where the net fluxes between these sets were computed. Our formulas for the rate constants k_{ab} and k_{ba} given below are explicit exact expressions for these fluxes.

III. TRANSITION STATE THEORY

The theory goes back to Eyring,¹ Wigner,² and Horiuti³ who made the following observation. If an ergodic system is partitioned into two sets, the frequency of hopping between

these sets is given by the absolute value of the velocity normal to the dividing surface between the two partitioning sets averaged with respect to the equilibrium probability distribution (e.g., the Gibbs distribution) restricted to this surface. Let us recall why.

Consider a system governed by (1) with two metastable sets a and b . Let us partition phase space into two open sets A and B containing a and b (i.e., $a \subset A$ and $b \subset B$) and a dividing surface S between these sets so that $A \cup B \cup S = \mathbb{R}^n$. It is convenient to parametrize S as the zero-level set of some scalar-valued dimensionless function $q(x)$, i.e.,

$$S \equiv \{x: q(x) = 0\}, \quad (14)$$

and to conventionally take $q(x) > 0$ if $x \in A$ and $q(x) < 0$ if $x \in B$. In terms of $q(x)$, the indicator function of A can be represented as

$$\chi_A(x) = H(q(x)), \quad (15)$$

where $H(z)$ is the Heaviside step function defined as $H(z) = 1$ if $z > 0$ and $H(z) = 0$ otherwise. Similarly, $\chi_B(x) = H \times (-q(x))$.

We define the mean residence time in A as

$$t_A = \lim_{T \rightarrow \infty} \frac{2}{N_T} T \sum_j t_A^j = \lim_{T \rightarrow \infty} \frac{2}{N_T} \int_0^T H(q(X(t))) dt, \quad (16)$$

and similarly for t_B , the mean residence time in B . Here t_A^j is the duration of the j th visit in A of a generic trajectory $X(t)$, the sum over j is taken over all visits in A up to time T , and N_T is the number of times the trajectory crosses the boundary S before time T . (16) can be written as

$$t_A = 2N_A/\nu_S^{\text{TST}}, \quad t_B = 2N_B/\nu_S^{\text{TST}}, \quad (17)$$

where

$$N_A = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T H(q(X(t))) dt = Z^{-1} \int_A e^{-\beta V(x)} dx \quad (18)$$

is the equilibrium population density in A [which by ergodicity is also the proportion of time that the trajectory $X(t)$ spends in A], $N_B = 1 - N_A$, and

$$\nu_S^{\text{TST}} = \lim_{T \rightarrow \infty} \frac{N_T}{T} \quad (19)$$

is the mean frequency of crossing the boundary S . This mean frequency can be estimated upon noting that the integral $N(T) = \int_0^T |(d/dt)H(q(X(t)))| dt$ precisely counts the number of times the trajectory $X(t)$ has crossed S during $[0, T]$. Therefore

$$\begin{aligned} \nu_S^{\text{TST}} &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \left| \frac{d}{dt} H(q(X(t))) \right| dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T |\dot{X}(t) \cdot \nabla q(X(t))| \delta(q(X(t))) dt \\ &= \int_{\mathbb{R}^n \times \mathbb{R}^n} |v \cdot \nabla q(x)| \delta(q(x)) \rho(x, v) dx dv, \end{aligned} \quad (20)$$

where $\rho(x, v)$ is the density in (2). Note that (20) is intrinsic

to the surface S [i.e., it does not depend on the explicit form of $q(x)$ except for the fact that $q(x) = 0$ on S] since it can be rewritten as

$$\nu_S^{\text{TST}} = \int_{\mathbb{R}^n} \int_S |\hat{n}(x) \cdot v| \rho(x, v) d\sigma(x) dv, \quad (21)$$

where $\hat{n}(x)$ is the unit normal to S at $x \in S$ and $d\sigma(x)$ is the surface element on S . The integration on v in (20) can be performed explicitly:

$$\begin{aligned} \nu_S^{\text{TST}} &= \sqrt{\frac{2}{\pi\beta}} Z^{-1} \int_{\mathbb{R}^n} |\nabla q(x)| \delta(q(x)) e^{-\beta V(x)} dx \\ &= \sqrt{\frac{2}{\pi\beta}} Z^{-1} \int_S e^{-\beta V(x)} d\sigma(x). \end{aligned} \quad (22)$$

Inserting (18) and (22) in (17) gives exact expressions for the mean residence times t_A and t_B which can be computed, e.g., by umbrella or blue-moon sampling and thermodynamic integration, see Sec. V.

Within TST, the mean transition frequency ν between a and b is approximated by ν_S^{TST} . Since $N_A \approx N_a$ and $N_B \approx N_b$ by (6), from (11) this approximation gives the following TST estimate for the rates k_{ab} and k_{ba} :

$$k_{ab}^{\text{TST}} = t_A^{-1}, \quad k_{ba}^{\text{TST}} = t_B^{-1}. \quad (23)$$

Unfortunately, these rates can be quite poor approximations of the actual rates. Indeed, the mean residency times t_A and t_B take in account *all* the crossings the trajectory makes from A to B or B to A irrespective on whether this trajectory does actually visit the metastable sets a and b between these crossings or does not. In other words, ν_S^{TST} always overestimates the actual transition frequency ν , and significantly so if the trajectory tends to recross S many times before entering a or b . This effect can be corrected, see Sec. VI. But even within the strict framework of TST, one can optimize the result of the theory by minimizing ν_S^{TST} , as explained next.

IV. VARIATIONAL TST

Since TST necessarily overestimates the transition frequency, $\nu_S^{\text{TST}} \geq \nu$, it is natural to look for the dividing surface S with minimum ν_S^{TST} . This idea was proposed by Horiuti³ and the surface with minimum ν_S^{TST} is now referred to as the variational TST dividing surface.^{5,7,8} We first derive the general equation for this surface in the context of the Langevin dynamics in (1) then discuss the possibility to derive practical algorithms by restricting the class of surfaces one optimizes upon.

A. The variational TST dividing surface

From (22), the variational TST dividing surface is the one that minimizes

$$I = \int_S e^{-\beta V} d\sigma(x) = \int_{\mathbb{R}^n} |\nabla q(x)| e^{-\beta V} \delta(q(x)) dx. \quad (24)$$

If S minimizes I , then I is left invariant to the first order under variations of S . It is equivalent but simpler to vary $q(x)$. Using

$$|\nabla(q(x) + \varepsilon\tilde{q}(x))| = |\nabla q| + \varepsilon\hat{n}(x) \cdot \nabla\tilde{q}(x) + O(\varepsilon^2), \quad (25)$$

where $\hat{n}(x) = \nabla q(x)/|\nabla q(x)|$ evaluated at $x \in S$ is the unit vector normal to S , and

$$\delta(q(x) + \varepsilon\tilde{q}(x)) = \delta(q(x)) + \varepsilon\delta'(q(x))\tilde{q}(x) + O(\varepsilon^2), \quad (26)$$

we deduce that $I[q + \varepsilon\tilde{q}] - I[q] = \varepsilon\tilde{I} + O(\varepsilon^2)$ with

$$\begin{aligned} \tilde{I} = & \int_{\mathbb{R}^n} \hat{n}(x) \cdot \nabla\tilde{q}(x) e^{-\beta V} \delta(q(x)) dx \\ & + \int_{\mathbb{R}^n} |\nabla q(x)| e^{-\beta V} \delta'(q(x)) \tilde{q}(x) dx. \end{aligned} \quad (27)$$

Integrating by parts the first integral at the right-hand side gives

$$\begin{aligned} \tilde{I} = & - \int_{\mathbb{R}^n} \nabla \cdot (\hat{n}(x) e^{-\beta V} \delta(q(x))) \tilde{q}(x) dx \\ & + \int_{\mathbb{R}^n} |\nabla q(x)| e^{-\beta V} \delta'(q(x)) \tilde{q}(x) dx. \end{aligned} \quad (28)$$

Expanding the factor in the first integral at the right-hand side using $\hat{n}(x) \cdot \nabla \delta(q(x)) = \hat{n}(x) \cdot \nabla q(x) \delta'(q(x)) = |\nabla q(x)| \delta'(q(x))$ we arrive at

$$\tilde{I} = \int_{\mathbb{R}^n} (\beta \hat{n}(x) \cdot \nabla V - \kappa(x)) e^{-\beta V} \delta(q(x)) \tilde{q}(x) dx, \quad (29)$$

where $\kappa(x) = \nabla \cdot \hat{n}(x)$ evaluated at $x \in S$ is the Gauss curvature of S . Since $\tilde{q}(x)$ is arbitrary, the integrand in (29) must vanish in order that first variation of I vanishes. Because of the presence of the factor $\delta(q(x))$ this requirement is nontrivial only on the surface S where $q(x) = 0$, where it reads

$$0 = \beta \hat{n}(x) \cdot \nabla V - \kappa(x). \quad (30)$$

This equation for the variational TST dividing surface was first derived by Horiuti.³ As such it is too complicated to be practical, and various approximations of this equation will be considered in Secs. IV B and IV C. Note that since $\hat{n}(x)$ and $\kappa(x)$ are geometric quantities attached to the surface S , (30) is intrinsic to S as it should be. Notice also that taking $\beta \rightarrow \infty$, (30) formally reduces to $0 = \hat{n}(x) \cdot \nabla V$, which is satisfied by any separatrix associated with the potential $V(x)$ [i.e., any dividing surface such that $-\nabla V(x)$ is everywhere tangent to the surface].

B. Planar Ansatz

Optimizing S by solving (30) seems hardly possible in practice. More realistically, one can minimize the functional (24) over restricted classes of surfaces. The simplest choice is to assume that S is planar, in which case it can be parametrized as

$$0 = \hat{n} \cdot x - b, \quad (31)$$

where b is a scalar and \hat{n} is the unit normal to the plane. Jóhannesson and Jónsson¹⁹ were the first to suggest to optimize the TST dividing surface among planes but they gave

incorrect equations for b and \hat{n} (Ref. 20)—the correct equations are given in (33) below.

Using the planar Ansatz for S in (24) gives

$$I = \int_{\mathbb{R}^n} e^{-\beta V} \delta(\hat{n} \cdot x - b) dx, \quad (32)$$

and this functional has to be minimized over b and \hat{n} to find the variational TST dividing surface S within the class of surfaces specified by (31). The equations that \hat{n} and b must satisfy in order to minimize (32) can be derived by taking the first variation of (32) with respect to \hat{n} and b subject to the constraint that $|\hat{n}| = 1$. The calculation, similar to the one which led to (30), is given in the Appendix and yields

$$\begin{cases} 0 = \int_{\mathbb{R}^n} \hat{n} \cdot \nabla V e^{-\beta V} \delta(\hat{n} \cdot x - b) dx, \\ 0 = - \int_{\mathbb{R}^n} \hat{n} \cdot \nabla V e^{-\beta V} x^\perp \delta(\hat{n} \cdot x - b) dx, \end{cases} \quad (33)$$

where $x^\perp = x - (x \cdot \hat{n})\hat{n}$ is the in-plane projection of x .

(33) are satisfied at steady state by the solutions of

$$\begin{cases} \dot{b} = \langle \hat{n} \cdot \nabla V \rangle_P, \\ \dot{\hat{n}} = - \langle (\hat{n} \cdot \nabla V) x^\perp \rangle_P, \end{cases} \quad (34)$$

where $\langle \cdot \rangle_P$ denotes the conditional averaging in the plane specified by (31): for any $f(x)$,

$$\langle f(x) \rangle_P = \frac{\int_{\mathbb{R}^n} f(x) \delta(\hat{n} \cdot x - b) e^{-\beta V(x)} dx}{\int_{\mathbb{R}^n} \delta(\hat{n} \cdot x - b) e^{-\beta V(x)} dx}. \quad (35)$$

The averages at the right-hand sides of (34) can be evaluated using the blue-moon sampling technique.¹⁵ Therefore these equations provide a practical scheme to identify the variational TST dividing plane. A scheme of this sort but with different equations for b and \hat{n} was actually implemented in Ref. 19.

C. Other Ansätze

Ansätze other than (31) can be introduced, which, e.g., are consistent with some symmetry of the system or involve collective variables, say,

$$z = \phi(x) \in \mathbb{R}^m \quad (m \leq n), \quad (36)$$

which are physically motivated and thought to be adequate to parametrize the variational TST dividing surface. Assuming that the variational TST is planar in z (which in general does not define a plane in the original configuration space), this surface can be parametrized as

$$0 = \hat{v} \cdot \phi(x) - b, \quad (37)$$

where \hat{v} is a unit vector in \mathbb{R}^m , $\hat{v} \cdot \phi(x)$ denotes the scalar product in \mathbb{R}^m , and b is a scalar. In terms of this Ansatz, (24) reads

$$I = \int_{\mathbb{R}^n} e^{-\beta V} |\nabla g(x)| \delta(\hat{\nu} \cdot \phi(x) - b) dx, \quad (38)$$

where $g(x) = \nu \cdot \phi(x)$. It is shown in the Appendix that in order to minimize this functional, $\hat{\nu}$ and b must satisfy

$$\begin{cases} 0 = \langle |\nabla g(x)|^{-1} (\mathbf{g}(x) \cdot \nabla V - k_B T \nabla \cdot \mathbf{g}(x)) \rangle_{\bar{P}} \\ 0 = - \langle |\nabla g(x)|^{-1} (\phi^\perp \mathbf{g}(x) \cdot \nabla V - k_B T \mathbf{g}(x) \cdot \nabla \phi^\perp) \rangle_{\bar{P}} \\ \quad + k_B T \langle |\nabla g(x)|^{-1} \phi^\perp \nabla \cdot \mathbf{g}(x) \rangle_{\bar{P}}, \end{cases} \quad (39)$$

where $\mathbf{g}(x) = \nabla g(x) / |\nabla g(x)|$, $\phi^\perp = \phi - \hat{\nu}(\hat{\nu} \cdot \phi)$, and $\langle \cdot \rangle_{\bar{P}}$ denotes averaging on the surface where $0 = \hat{\nu} \cdot \phi(x) - b$, similar to (35).

V. REMARKS ON THE FREE ENERGY

In this section we clarify the common assertion that the optimal TST dividing surface is the one that maximizes the free energy. We show that this assertion is incorrect if one uses the standard free energy associated with a reaction coordinate and explain why. We also recall how to compute the free energy in practice and how to use it to evaluate the TST transition rate constant. Finally we discuss the meaning of another quantity which is sometimes confused with the free energy.

A. Free energy: Definition, evaluation, and interpretation

Let the scalar-valued function $q(x)$ be a reaction coordinate, i.e., assume that the level sets (or isosurfaces) $q(x) = z$ each define a dividing surface, the collection of which for $z \in \mathbb{R}$ foliates configuration space. The free energy (or potential of mean force) associated with $q(x)$ is defined as

$$F(z) = -k_B T \ln \left(Z^{-1} \int_{\mathbb{R}^n} e^{-\beta V(x)} \delta(q(x) - z) dx \right). \quad (40)$$

$e^{-\beta F(z)}$ is the marginal probability density function in the variable $q(x) = z$ of the equilibrium probability density $Z^{-1} e^{-\beta V(x)}$:

$$\int_{z_1}^{z_2} e^{-\beta F(z)} dz = \text{prob}(z_1 \leq q(x) \leq z_2). \quad (41)$$

In other words, to leading order in $\delta z \ll 1$, $e^{-\beta F(z)} \delta z$ gives the equilibrium probability to find the system in the slab $z \leq q(x) \leq z + \delta z$ around the surface S , where $q(x) = z$. Notice that the thickness of this slab is nonuniform and depends on the local value of $|\nabla q(x)|$ on S (the smaller $|\nabla q(x)|$, the thicker the slab is). This implies, in particular, that the integral in (40) is not intrinsic to the surface parametrized by $q(x) = z$. This has important implications in the context of TST.

Suppose that the reaction coordinate is consistent with the variational TST dividing surface, i.e., this surface can be parametrized as $q(x) = 0$ as in Sec. III. Even when this is the case, in general $Z^{-1} e^{-\beta F(0)} \neq \min_{\mathcal{I}} J$, where J is the object function of variational TST defined in (24) [the difference between the integrals in (24) and (40) being the presence or

the absence of the factor $|\nabla q(x)|$ —the latter makes (24) intrinsic to S , whereas (40) is not for the reasons we just explained]. As a result, $F(z)$ usually reaches its maximum at $z_* \neq 0$, meaning that the variational TST surface does not maximize the free energy of the reaction coordinate $q(x)$ in general.

Of course, this does not mean that $F(z)$ is not useful within TST. In particular, using (22) and assuming that the TST dividing surface S is parametrized as $q(x) = 0$ as in (14), it is easy to see that ν_S^{TST} can be expressed in terms of $F(z)$ as

$$\nu_S^{\text{TST}} = \sqrt{\frac{2}{\pi\beta}} \langle |\nabla q(x)| \rangle_{q(x)=0} e^{-\beta F(0)}, \quad (42)$$

where $\langle |\nabla q(x)| \rangle_{q(x)=z}$ is the average value of $|\nabla q(x)|$ on the surface $q(x) = z$ —i.e., on S when $z = 0$. Similarly, assuming that A lies in the region where $q(x) > 0$ and B in the one where $q(x) < 0$, N_A and N_B can be expressed in terms of $F(z)$ as

$$N_A = \int_0^\infty e^{-\beta F(z)} dz, \quad N_B = \int_{-\infty}^0 e^{-\beta F(z)} dz. \quad (43)$$

These formulas are especially interesting if one recalls that $F(z)$ is rather easy to evaluate.²¹ By taking the derivative of both sides of (40) with respect to z and integrating the right-hand side by part using the identity

$$\frac{d}{dz} \delta(q(x) - z) = -\mathbf{q}(x) \cdot \nabla \delta(q(x) - z), \quad (44)$$

where $\mathbf{q}(x) = \nabla q(x) / |\nabla q(x)|^2$, one arrives at the following expression for the mean force $F'(z)$:

$$F'(z) = \langle \mathbf{q}(x) \cdot \nabla V(x) - k_B T \nabla \cdot \mathbf{q}(x) \rangle_{q(x)=z}, \quad (45)$$

where $\langle \cdot \rangle_{q(x)=z}$ denotes conditional averaging on the surface $q(x) = z$: for any function $f(x)$,

$$\langle f(x) \rangle_{q(x)=z} = \frac{\int_{\mathbb{R}^n} f(x) \delta(q(x) - z) e^{-\beta V(x)} dx}{\int_{\mathbb{R}^n} \delta(q(x) - z) e^{-\beta V(x)} dx}. \quad (46)$$

The average in (45) can be computed, e.g., by blue-moon sampling,¹⁵ and $F(z)$ can be obtained from $F'(z)$ by thermodynamic integration.¹⁶ This gives $F(z)$ up to a constant of integration which can be determined from the following normalization condition which follows from (41):

$$\int_{\mathbb{R}} e^{-\beta F(z)} dz = 1. \quad (47)$$

B. An issue of terminology

In view of the above, the following quantity may seem more relevant than $F(z)$ in the context of TST:

$$G(z) = -k_B T \ln \left(\lambda Z^{-1} \int_{\mathbb{R}^n} e^{-\beta V(x)} |\nabla q(x)| \delta(q(x) - z) dx \right), \quad (48)$$

where λ is an arbitrary length scale introduced for dimensional consistency. If $q(x)=0$ parametrize the variational TST surface, then $G(z)$ [unlike $F(z)$] is maximum at $z=0$. It is also easy to see that the TST rate constant can be expressed in terms of $G(z)$ as

$$\nu_S^{\text{TST}} = \sqrt{\frac{2}{\pi\beta}} \lambda^{-1} e^{-\beta G(0)}$$

since $G(z)$ is related to $F(z)$ as

$$e^{-\beta G(z)} = \lambda \langle |\nabla q(x)| \rangle_{q(x)=z} e^{-\beta F(z)}. \quad (49)$$

(Notice that this shows that $F(z) \neq G(z)$ if $\langle |\nabla q(x)| \rangle_{q(x)=z}$ is nonconstant in z , which is the case if the level sets $q(x)=z$ are nonplanar.)

So, should $G(z)$ rather than $F(z)$ be called the free energy, as asserted in Ref. 22? We argue that it should not. First, unlike $e^{-\beta F(z)}$, $e^{-\beta G(z)}$ is not a probability density in the variable z , i.e., $\int_{z_1}^{z_2} e^{-\beta G(z)} dz \neq \text{prob}(z_1 \leq q(x) \leq z_2)$ in general. This is obvious from the definition of $G(z)$ and consistent with the fact that $e^{-\beta G(z)}$ gives the probability density of the surface S parametrized by $q(x)=z$. In other words, the equilibrium probability to find the system in a slab of uniform thickness $d > 0$ around the surface $S \equiv \{x: q(x)=z\}$ is for small d given by

$$d\lambda e^{-\beta G(z)} + O(d^2). \quad (50)$$

Second, the statement that $G(z)$ [unlike $F(z)$] is such that it reaches its maximum at the variational TST surface is somewhat misleading. Indeed, if G were to be used to identify the variational TST surface, then this function should be defined for all possible dividing surfaces S and not only the ones that can be parametrized as the level sets (or isosurfaces) of a given reaction coordinate $q(x)$. In other words, instead of (48), one should define

$$G(S) = -k_B T \ln \left(\lambda Z^{-1} \int_S e^{-\beta V(x)} d\sigma(x) \right), \quad (51)$$

where S is any possible dividing surfaces S and therefore $G(S)$ has to be thought of as a functional and not a simple function as $F(z)$. Since $e^{-\beta G(S)} = \lambda Z^{-1} I$, where I is the object function of variational TST defined in (24), then indeed we have that the variational TST surface maximizes $G(S)$. But $G(S)$ is so much more complicated than $F(z)$ that it cannot be identified for all possible dividing surfaces S in practice. For these reasons F and not G is the free energy of the reaction coordinate $q(x)$.

VI. DYNAMICAL CORRECTIONS

Here we discuss how to account for the effect of recrossing and correct the TST transition frequency by means of the transmission coefficient of the TST dividing surface. In the usual approach, due to Keck⁹ and further developed by Bennett and Chandler,^{11,12} one estimates a time-dependent rate

and looks for its quasiplateau value at times large compared to the molecular time scale but small compared to the time scale $1/(k_{ab}+k_{ba})$ over which it eventually decays to zero. The existence of a plateau relies on the existence of two metastable sets a and b , as defined in Sec. II. Indeed, it is precisely because the trajectory will eventually get committed to either a or b and stays there for a very long time before making another transition to the other set that the time-dependent rate saturates on a time scale slow compared to the molecular time scale but fast compared to $1/(k_{ab}+k_{ba})$.

On the other hand, the standard procedure to compute dynamical corrections does not use explicitly the definition of the metastable sets a and b . Our approach is different since we are interested in the exact mean frequency of transition between the two sets a and b defined in (12). Having such an exact expression for the rate is useful since it will allow us to give precise error estimate on the numerical procedure to estimate this frequency, see Sec. VII B. In fact, we show next that this exact mean frequency can be expressed as

$$\nu = \lim_{T \rightarrow \infty} \frac{2}{T} \int_0^T \chi_{\bar{a}}(X(t_{as}^+(t))) \chi_{\bar{b}}(X(t_{ab}^-(t))) \frac{d}{dt} H(q(X(t))) dt. \quad (52)$$

(A similar expression has been proposed in Ref. 23.) Here $\chi_{\bar{a}}(x)$ and $\chi_{\bar{b}}(x)$ are the indicator functions of the closure of the metastable sets a and b , respectively; $t_{as}^+(t)$ is the first time after t that the trajectory enters either \bar{a} or S :

$$t_{as}^+(t) = \text{minimum time } t' > t \text{ such that } X(t') \in \bar{a} \cup S; \quad (53)$$

and $t_{ab}^-(t)$ is the last time before t that the trajectory left either \bar{a} or \bar{b} :

$$t_{ab}^-(t) = \text{maximum time } t' < t \text{ such that } X(t') \in \bar{a} \cup \bar{b}; \quad (54)$$

These times are unambiguously defined when the trajectory is on S at time t , which is all what matters in (52) because the factor $(d/dt)H(q(X(t)))$ is only nonzero when t is such that $X(t) \in S$.

A. Justification of (52)

To justify (52) notice that integrating the factor $(d/dt)H(q(X(t)))$ alone would count all the transitions from B to A positively and from A to B negatively and hence results in a net cancellation after each pairs of crossing-recrossing. However, since $X(t_{as}^+(t)) \in \bar{a} \cup S$ and $X(t_{ab}^-(t)) \in \bar{a} \cup \bar{b}$ by definition, we have

$$\chi_{\bar{a}}(X(t_{as}^+(t))) \chi_{\bar{b}}(X(t_{ab}^-(t))) = 1 \quad (55)$$

if and only if $X(t_{as}^+(t)) \in \bar{a}$ and $X(t_{ab}^-(t)) \in \bar{b}$, and zero otherwise. It follows that the integral in (52) only counts those crossings such that the trajectory will subsequently go to a before returning to S and was in b rather than a before reaching S . In other words, the integral in (52) only counts the last crossing of S during an actual transition from b to a , as

required to obtain the actual frequency ν (the factor of 2 accounts for the fact that there are twice as many transition between a and b as from b to a).

B. Ensemble versus time averaging

Here we show that (52) can be expressed as

$$\begin{aligned} \nu &= 2 \int_{\mathbb{R}^n \times \mathbb{R}^n} (v \cdot \nabla q(x))_+ \rho(x, v) \xi_{a|S}(x, v) \xi_{b|a}(x, -v) \\ &\quad \times \delta(q(x)) dx dv = 2 \int_{\mathbb{R}^n} \int_S (v \cdot \hat{n}(x))_+ \rho(x, v) \\ &\quad \times \xi_{a|S}(x, v) \xi_{b|a}(x, -v) d\sigma(x) dv, \end{aligned} \quad (56)$$

where $(z)_+ = \max(z, 0)$ and we defined

$$\xi_{a|S}(x, v) = \text{probability to reach } a \text{ before } S \text{ starting from } (x, v), \quad (57)$$

and

$$\xi_{b|a}(x, v) = \text{probability to reach } b \text{ before } a \text{ starting from } (x, v), \quad (58)$$

To derive (56), note that by ergodicity (52) is²⁴

$$\nu = 2 \int_{\mathbb{R}^n \times \mathbb{R}^n} v \cdot \nabla q(x) G(x, v) \rho(x, v) \delta(q(x)) dx v, \quad (59)$$

where

$$\begin{aligned} G(x, v) &= \mathbf{E}(\chi_{\bar{a}}(X(\tau_{aS}^+(x, v), x, v))) \\ &\quad \times \mathbf{E}(\chi_{\bar{b}}(X(\tau_{ab}^-(x, v), x, v))). \end{aligned} \quad (60)$$

Here $\mathbf{E}(\cdot)$ denotes expectation with respect to the noise η , $X(t, x, v)$ is the solution of (1) with initial condition $(X(0), \dot{X}(0)) = (x, v)$, $\tau_{aS}^+(x, v)$ is the first time after $t=0$ that this trajectory enters either \bar{a} or S :

$$\tau_{aS}^+(x, v) = \text{minimum } t > 0 \text{ such that } X(t, x, v) \in \bar{a} \cup S, \quad (61)$$

and $\tau_{ab}^-(x, v)$ is the last time before $t=0$ that this trajectory left either \bar{a} or \bar{b} :

$$\tau_{ab}^-(x, v) = \text{maximum } t < 0 \text{ such that } X(t, x, v) \in \bar{a} \cup \bar{b}. \quad (62)$$

Note that $\tau_{aS}^+(x, v)$ and $\tau_{ab}^-(x, v)$ are random quantity due to the noise η , hence the expectation over η in (60). However, due to the Markov character of the dynamics $\tau_{aS}^+(x, v)$ and $\tau_{ab}^-(x, v)$ are statistically independent, hence the factorization of the expectations in (60).

Clearly, the first expectation in (60) can be expressed as

$$\mathbf{E}(\chi_{\bar{a}}(X(\tau_{aS}^+(x, v), x, v))) = \xi_{a|S}(x, v). \quad (63)$$

To express the second expectation in (60), notice that $X(t, x, v)$ is statistically equivalent to $X(-t, x, -v)$, and therefore $\tau_{ab}^-(x, v)$ is statistically equivalent to $-\tau_{ab}^+(x, -v)$. It follows that

$$\begin{aligned} \mathbf{E}(\chi_{\bar{b}}(X(\tau_{ab}^-(x, v), x, v))) &= \mathbf{E}(\chi_{\bar{b}}(X(\tau_{ab}^+(x, -v), x, -v))) \\ &= \xi_{b|a}(x, -v). \end{aligned} \quad (64)$$

Combining (59), (60), (63), and (64) and noting that $\xi_{a|S}(x, v)$ is nonzero only when $v \cdot \nabla q(x) > 0$, we arrive at (56).

Notice that it is readily apparent from (56) that $0 \leq \nu \leq \nu^{\text{TST}}$ since ν^{TST} can also be expressed as

$$\nu^{\text{TST}} = 2Z_H^{-1} \int_{\mathbb{R}^n \times \mathbb{R}^n} (v \cdot \nabla q(x))_+ e^{-\beta H(x, v)} \delta(q(x)) dx dv, \quad (65)$$

and $0 \leq \xi_{a|S}(x, v) \xi_{b|a}(x, -v) \leq 1$.

C. Transmission coefficient

It is convenient to define the transmission coefficient κ_S of the surface S as the ratio

$$\kappa_S = \nu / \nu_S^{\text{TST}}. \quad (66)$$

From (21) and (56) one sees that the transmission coefficient κ_S can be expressed as the average:

$$\kappa_S = \langle \langle \xi_{a|S}(x, v) \xi_{b|a}(x, -v) \rangle \rangle, \quad (67)$$

where $\langle \langle \cdot \rangle \rangle$ denotes the expectation with respect to the probability distribution

$$(v \cdot \nabla q(x))_+ e^{-\beta H} \delta(q(x)) dx \equiv (v \cdot \hat{n}(x))_+ e^{-\beta H} d\sigma(x) \quad (68)$$

properly normalized, i.e.,

$$\langle \langle f(x, v) \rangle \rangle = \frac{\int_{\mathbb{R}^n \times \mathbb{R}^n} (v \cdot \nabla q(x))_+ f(x, v) e^{-\beta H(x, v)} \delta(q(x)) dx dv}{\int_{\mathbb{R}^n \times \mathbb{R}^n} (v \cdot \nabla q(x))_+ e^{-\beta H(x, v)} \delta(q(x)) dx dv}. \quad (69)$$

Sampling with respect to the distribution in (68) can be done, e.g., by some type of blue-moon sampling in the surface where $q(x)=0$. For instance, we have

$$\langle \langle f(x, v) \rangle \rangle = \frac{\langle (v \cdot \nabla q(x))_+ f(x, v) \rangle_{q(x)=0}}{\langle (v \cdot \nabla q(x))_+ \rangle_{q(x)=0}}, \quad (70)$$

where

$$\langle f(x, v) \rangle_{q(x)=0} = \frac{\int_{\mathbb{R}^n \times \mathbb{R}^n} f(x, v) e^{-\beta H(x, v)} \delta(q(x)) dx dv}{\int_{\mathbb{R}^n \times \mathbb{R}^n} e^{-\beta H(x, v)} \delta(q(x)) dx dv}. \quad (71)$$

If f depends only on x , $f(x, v) = f(x)$, this last expectation reduces to the one in (46). Using $\langle \langle \cdot \rangle \rangle$ instead of $\langle \cdot \rangle_{q(x)=0}$ is not essential, but it will prove useful in the error estimate given in Sec. VII B.

Since ν is independent of S whereas ν^{TST} is not, the variational TST dividing surface which minimizes ν_S^{TST} also maximizes κ_S . Also, from (67), we obviously have that $0 \leq \kappa_S \leq 1$.

VII. PRACTICAL IMPLEMENTATION AND ERROR ESTIMATE

Here we discuss how to evaluate κ_S in practice and estimate the error in the numerical procedure. We show that the transmission coefficient can be estimated efficiently if and only if it is not too small. In other words, it is necessary to optimize the dividing surface as discussed in Sec. III (a poorly chosen surface with a very small transmission coefficient leads to large errors). But unfortunately, even the variational TST dividing surface may have a very small transmission coefficient, in which case the numerical procedure will not be efficient.

For a complementary discussion, see Ref. 25.

A. Practical implementation

From (67), κ_S can be estimated as

$$\langle\langle \xi_{a|S}(x, v) \xi_{b|a}(x, -v) \rangle\rangle \approx \frac{1}{R} \sum_{i=1}^R \chi_{a|S}^i \chi_{b|a}^i, \quad (72)$$

where

$$\begin{aligned} \chi_{a|S}^i &= \chi_{\bar{a}}(X(\tau_{aS}^+(x_i, v_i), x_i, v_i)), \\ \chi_{b|a}^i &= \chi_{\bar{b}}(X(\tau_{ab}^+(x_i, -v_i), x_i, -v_i)). \end{aligned} \quad (73)$$

Here $\{x_i, v_i\}_{i=1, \dots, R}$ are R -independent initial conditions on S drawn from the distribution in (68) properly normalized; $X(t, x_i, v_i)$ and $\tilde{X}(t, x_i, -v_i)$ are two independent trajectories generated from (x_i, v_i) and $(x_i, -v_i)$, respectively. The factor $\chi_{\bar{a}}(X(\tau_{aS}^+(x_i, v_i), x_i, v_i))$ is 1 if the i th trajectory generated from (x_i, v_i) reaches a before to S and 0 otherwise. The factor $\chi_{\bar{b}}(X(\tau_{ab}^+(x_i, -v_i), x_i, -v_i))$ is 1 if the i th trajectory generated from $(x_i, -v_i)$ reaches b before a and 0 otherwise.

B. Error estimate

Next, we estimate the mean-square relative error made in the estimate (72). Since $\chi_{a|S}^i \chi_{b|a}^i$ is either 0 or 1, we have

$$\mathbf{E}\langle\langle (\chi_{a|S}^i \chi_{b|a}^i)^2 \rangle\rangle = \mathbf{E}\langle\langle \chi_{a|S}^i \chi_{b|a}^i \rangle\rangle \equiv \kappa_S, \quad (74)$$

and hence

$$\text{var}(\chi_{a|S}^i \chi_{b|a}^i) = \kappa_S(1 - \kappa_S). \quad (75)$$

As a result,

$$\text{relative error on } \kappa_S \text{ from (72)} = \sqrt{\frac{1 - \kappa_S}{R\kappa_S}}. \quad (76)$$

This estimate indicates that the numerical procedure described is optimal when κ_S is close to 1. Unfortunately, the situation is not as good when κ_S is small. In this case, in order to compute κ_S within relative error tolerance λ , we must initiate of the order of $R = O(\lambda^{-2} \kappa_S^{-1})$ trajectories from S . Also it can be checked that this estimate does not improve if for each x_i on S , more than one v_i is drawn, or if one uses many realizations of the noise for each pair (x_i, v_i) : overall, the number of trajectories needed remains $R = O(\lambda^{-2} \kappa_S^{-1})$ for small κ_S .

The situation when κ_S is small is quite similar to what one encounters if, say, one wants to average the indicator function of $[0, \delta]$, $\chi_{[0, \delta]}(x)$, with $0 < \delta \ll 1$ over a random variable x which is uniformly distributed in $[0, 1]$. The estimator $(1/R) \sum_{i=1}^R \chi_{[0, \delta]}(x_i)$, where x_i are independent variables uniformly distributed in $[0, 1]$, is a very bad estimator of the average when $\delta \ll 1$. Indeed most of the x_i 's will lie outside $[0, \delta]$ (they have a probability $1 - \delta$ to do so) and hence will not contribute to improve the statistics of the average. In this example, the right strategy is to recognize the shape of the function $\chi_{[0, \delta]}(x)$ and use importance sampling accordingly. Unfortunately, in the case of κ_S , the function $\xi_{a|S}(x, v) \xi_{b|a}(x, -v)$ plays the role of $\chi_{[0, \delta]}(x)$ in the example, and we have no idea what the shape of this function is since it depends not only on the equilibrium probability distribution on S , which we know, but also on the behavior of the trajectories out of S , which we do not know *a priori*. As a result, we have no other option than shooting in the dark on S using (72), and when $\kappa_S \ll 1$ this only gives the bad error estimate in (76).

Finally, it should be pointed out that the variational TST dividing surface which minimizes recrossing and hence maximizes κ_S is also, according to the error estimate in (76), the best surface to compute the dynamical corrections.

VIII. CONCLUDING REMARKS

Summarizing, this paper contains three main new results. First, we have clarified the relation between the dividing TST surface and the free energies associated with two different observables: a reaction coordinate and the family of all possible dividing surfaces. The two free energies are different. The free energy of a reaction coordinate is the more practical object, but the free energy of all dividing surfaces is the one which is maximum at the variational TST dividing surface. Second, we have proposed a new procedure to compute the dynamical corrections upon TST based on the exact calculation of the rate between two predefined sets. This is different from the usual procedure based on the identification of a plateau value for a time-dependent rate constant, and allow us to give an *a priori* error estimate on the transmission coefficient computed by our method. Third, we have derived a new and systematic way to identify the variational TST dividing surface within certain classes, e.g., hyperplanes. The equations for this surface involve expectations which can be computed, e.g., by blue-moon sampling and therefore may be used as practical tools to identify the variational TST surface.

It was also shown that the dynamical corrections can be computed efficiently if and only if the dividing surface has a transmission coefficient which is not too small. In situations when even the transmission coefficient of the variational TST dividing surface is small, computing the dynamical corrections becomes prohibitively expensive since it requires to use a number of trajectories which is inversely proportional to the transmission coefficient. In these situations, different approaches have to be employed to describe metastability, such as transition path sampling,²⁶ or the finite temperature string method.²⁷

ACKNOWLEDGMENTS

The ideas discussed here matured during the workshop on “Conformational Dynamics in Complex Systems,” organized CECAM, July 12–23, 2004. One of the authors (E.V.E.) wish to thank all the participants. We are particularly grateful to Giovanni Ciccotti for his careful reading of this paper and his many suggestions to improve it, and to Christoph Dellago and Daan Frenkel for discussions about the free energy. This work was partially supported by NSF Grant Nos. DMS01-01439, DMS02-09959, and DMS02-39625, and ONR Grant No. N00014-04-1-0565.

APPENDIX: DERIVATION OF (33) and (39)

To derive (33), we start from (32) to which we had a Lagrange multiplier term to enforce the constraint $|\hat{n}|=1$:

$$I_\lambda = \int_{\mathbb{R}^n} e^{-\beta V} \delta(\hat{n} \cdot x - b) dx + \lambda |\hat{n}|^2, \quad (\text{A1})$$

where λ is the Lagrange multiplier to be determined later. The Euler-Lagrange equations are obtained by requiring that the derivatives of (A1) with respect to b and \hat{n} are zero. Starting with b , we obtain

$$\begin{aligned} 0 &= \frac{\partial I_\lambda}{\partial b} = - \int_{\mathbb{R}^n} e^{-\beta V} \delta'(\hat{n} \cdot x - b) dx \\ &= - \int_{\mathbb{R}^n} e^{-\beta V} \hat{n} \cdot \nabla \delta(\hat{n} \cdot x - b) dx \\ &= \beta \int_{\mathbb{R}^n} \hat{n} \cdot \nabla V e^{-\beta V} \delta(\hat{n} \cdot x - b) dx, \end{aligned} \quad (\text{A2})$$

where we used $\delta'(\hat{n} \cdot x - b) = \hat{n} \cdot \nabla \delta(\hat{n} \cdot x - b)$ and integrated by parts. (A2) gives the first equation in (33). Differentiating (A1) with respect to \hat{n} gives

$$\begin{aligned} 0 &= \nabla_{\hat{n}} I_\lambda = \int_{\mathbb{R}^n} e^{-\beta V} x \delta'(\hat{n} \cdot x - b) dx + 2\lambda \hat{n} \\ &= \int_{\mathbb{R}^n} e^{-\beta V} x \hat{n} \cdot \nabla \delta(\hat{n} \cdot x - b) dx + 2\lambda \hat{n} \\ &= -\beta \int_{\mathbb{R}^n} x \hat{n} \cdot \nabla V e^{-\beta V} \delta(\hat{n} \cdot x - b) dx \\ &\quad + \hat{n} \int_{\mathbb{R}^n} e^{-\beta V} \delta(\hat{n} \cdot x - b) dx + 2\lambda \hat{n}. \end{aligned} \quad (\text{A3})$$

Multiplying this equation by \hat{n} using $|\hat{n}|=1$ and solving in λ give

$$\begin{aligned} \lambda &= \frac{1}{2} \beta \int_{\mathbb{R}^n} \hat{n} \cdot x \hat{n} \cdot \nabla V e^{-\beta V} \delta(\hat{n} \cdot x - b) dx \\ &\quad - \frac{1}{2} \int_{\mathbb{R}^n} e^{-\beta V} \delta(\hat{n} \cdot x - b) dx. \end{aligned} \quad (\text{A4})$$

This choice of λ guarantees that the constraint $|\hat{n}|=1$ is satisfied, and inserting (A4) into (A3) gives the second equation in (33).

The derivation of (39) from (38) follows similar steps except that it uses the identity $\delta'(\hat{v} \cdot \phi(x) - b) = \mathbf{g}(x) \cdot \nabla \delta(\hat{v} \cdot \phi(x) - b)$.

¹H. Eyring, *J. Chem. Phys.* **3**, 107 (1935).

²E. Wigner, *Trans. Faraday Soc.* **34**, 29 (1938).

³J. Horiiuti, *Bull. Chem. Soc. Jpn.* **13**, 210 (1938).

⁴J. B. Anderson, *Adv. Chem. Phys.* **91**, 381 (1995).

⁵D. G. Truhlar, B. C. Garrett, and S. J. Klippenstein, *J. Phys. Chem.* **100**, 12771 (1996).

⁶J. E. Straub, in *Computational Biochemistry and Biophysics*, edited by, O. M. Becker, A. D. MacKerell, Jr., B. Roux, and M. Watanabe (Marcel Dekker, New York, 2001), p. 199.

⁷P. Pechukas, *Annu. Rev. Phys. Chem.* **32**, 159 (1981).

⁸D. G. Truhlar and B. C. Garrett, *Annu. Rev. Phys. Chem.* **35**, 159 (1984).

⁹J. C. Keck, *Discuss. Faraday Soc.* **33**, 173 (1962).

¹⁰T. Yamamoto, *J. Chem. Phys.* **33**, 281 (1960).

¹¹C. H. Bennett, in *Algorithms for Chemical Computation*, ACS Symposium Series Vol. 46, edited by A. S. Nowick and J. J. Burton (Washington, DC, 1977), p. 63.

¹²D. Chandler, *J. Chem. Phys.* **68**, 2959 (1978).

¹³F. Tal and E. Vanden-Eijnden, *Nonlinearity* (submitted).

¹⁴W. Huisinga, S. Meyn, and Ch. Schütte, *Ann. Appl. Probab.* **14**, 419 (2004).

¹⁵E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral, *Chem. Phys. Lett.* **156**, 472 (1989).

¹⁶D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithm to Applications*, 2nd ed. (Elsevier, New York, 2001).

¹⁷S. H. Northrup and J. T. Hynes, *J. Chem. Phys.* **73**, 2700 (1980).

¹⁸R. F. Grote and J. T. Hynes, *J. Chem. Phys.* **73**, 2715 (1980).

¹⁹G. H. Jóhannesson and H. Jónsson, *J. Chem. Phys.* **115**, 9644 (2001).

²⁰It seems that the origin of the error in Ref. 19 is related to the misconception about the free energy that we discuss in Sec. V (see also Ref. 21).

²¹G. Ciccotti, R. Kapral, and E. Vanden-Eijnden, *ChemPhysChem* **6**, 1809 (2005).

²²G. K. Schenter, B. C. Garrett, and D. G. Truhlar, *J. Chem. Phys.* **119**, 5828 (2003).

²³T. S. van Erp and P. G. Bolhuis, *J. Comput. Phys.* **205**, 157 (2005).

²⁴(56) use the strong Markov property and the fact that $\tau_{as}^+(x, v)$ and $\tau_{ab}^-(x, v)$ are stopping times for an introduction to these concepts see, e.g., R. Durrett, *Stochastic Calculus* (CRC, Boca, Raton, FL, 1996).

²⁵M. J. Ruiz-Montero, D. Frenkel, and J. J. Brey, *Mol. Phys.* **90**, 925 (1997).

²⁶C. Dellago, P. G. Bolhuis, and P. L. Geissler, *Adv. Chem. Phys.* **123**, 1 (2002).

²⁷W. E. W. Ren and E. Vanden-Eijnden, *J. Phys. Chem. B* **109**, 6688 (2005).