Efficient Computation, Sensitivity and Error Analysis of Committor Probabilities for Complex Dynamical Processes

Jan-Hendrik Prinz,[1,2,*] Martin Held,[3,†] Jeremy C. Smith,[3,‡] and Frank Noé[3,§]

[1]*DFG Research Center Matheon, FU Berlin,*
Arnimallee 6, 14195 Berlin, Germany

[2]*IWR, University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany*

[3]*UT/ORNL Center for Molecular Biophysics,*
Oak Ridge National Laboratory P.O.Box 2008 Oak Ridge TN 37831-6164, USA

Abstract

In many fields of physics, chemistry and biology the characterization of dynamical processes between states or species is of fundamental interest. The central mathematical function in such situations is the committor probability - a generalized reaction coordinate that measures the progress of the process of interest as the probability of proceeding towards the target state rather than relapsing to the source state. Here, we present methodology for the efficient computation of committor probabilities for large-scale systems, such as, for example simuations of biomolecular folding. A method is derived for computing the committor for discrete state spaces using eigenvectors with expressions for the sensitivity and a Bayesian error model for the committor. The concepts are illustrated on two examples of diffusive dynamics with a very large number of states: a two-dimensional model potential with three minima, and a three-dimensional model representing protein-ligand binding. The method can finally be used to compute committor probabilities including error estimations for medium and large system sizes allowing access to the apparatus of transition path theory and its applications.

INTRODUCTION

The essential features of dynamical systems can often be understood in terms of the transitions between substates of special interest. This is particularly true where the dynamics itself points to the substates being metastable. Examples of this include protein folding or misfolding [19, 21], molecular association [26], chemical reactions [7], phase transitions in spin systems [12, 20, 22] or liquids [27], climate systems [9] and trend changes in financial systems [10]. In many cases, characterizing the dynamics between two substates $\mathcal{A}, \mathcal{B}$ of configurational space Ω provides a satisfactory picture of the process (e.g. in protein folding $\mathcal{A}$ being unfolded and $\mathcal{B}$ native [19]), whereas in other cases the simultaneous consideration of multiple substates is necessary.

It is now widely recognized that the committor probability, also called splitting probability or probability of folding in some contexts, is the central mathematical object that allows intersubstate processes to be characterized [2–6, 11, 13, 14]. The committor $q(x)$ is a state function that provides the probability at any state $x \in \Omega$ of next moving to $\mathcal{B}$ next rather than to $\mathcal{A}$ under the action of the system dynamics. By definition $q(x) = 0$ for $x \in \mathcal{A}$ and $q(x) = 1$ for $x \in \mathcal{B}$. The committor thus defines a dynamical reaction coordinate, which has the advantage over *ad hoc* reaction coordinates that it does not bring the danger of concealing relevant dynamics of the system. In the present work, we investigate how the committor probability can be efficiently computed for large-scale systems and study its sensitivity as well as its uncertainty in cases where the full dynamics has been inferred from a finite set of observations.

We concentrate here on dynamical systems which can be modeled as Markov processes between a finite (but possibly large) number m of discrete states. This includes systems which are discrete and Markovian by definition, such as spin glasses and on-lattice Go models [29] or resulting from a space discretization of a continuous generator or propagator [23]. In the latter case, the spacial discretization will cause the discretized system to be no longer exactly Markovian. The unintentionally introduced memory can in principle be described by the Mori-Zwanzig projection formalism of the full-dimensional dynamics onto a basis set defining the discrete states [16, 31, 32], but, from a numerical point of view the error made by using a Markov model in the discrete state space can in principle be rendered as small as desired by using a fine enough discretization, a small

enough time resolution [23], or, alternatively, embedding the dynamics in an extended discrete state space as proposed in Ref. [28].

The system dynamics is then described by a discrete-time transition matrix $T(\tau) \in \mathbb{R}^{m\times m}$, giving rise to the Chapman-Kolmogorov equation,

$$\mathbf{p}(k\tau) = \mathbf{p}(0)\mathbf{T}^k(\tau), \tag{1}$$

which propagates state probabilities $p \in \mathbb{R}^m$ in time, or by the rate matrix $\mathbf{K} \in \mathbb{R}^{m\times m}$ and the corresponding master equation

$$\frac{d\mathbf{p}(t)}{dt} = \mathbf{p}(t)\mathbf{K}$$

with the formal solution:

$$\mathbf{p}(t) = \mathbf{p}(0)\exp(t\mathbf{K}) \tag{2}$$

yielding the formal relationship

$$\mathbf{T}(\tau) = \exp\left(\tau\mathbf{K}\right).$$

A large number of studies treat the estimation of $\mathbf{T}$ or $\mathbf{K}$ from observation data in cases where they are not defined by the model itself or can be derived from the discretization of a continuous operator, but this estimation problem is not further considered here.

Given such a dynamical model, let us examine a number of aspects of the system dynamics that can be accessed *via* the committor probability. Firstly, all sets of constant committor probability in the state space Ω

$$I(q^*) = \{x \in \Omega \mid q(x) = q^*\},\ \forall q^* \in [0,1] \tag{3}$$

are hypersurfaces that partition the state space into the two disjoint subsets $I_{\mathcal{A}}(q^*) = \{x \in \Omega \mid q(x) < q^*\}$ with $\mathcal{A} \subset I_{\mathcal{A}}(q^*)$, $\forall q^* > 0$ and $I_{\mathcal{B}}(q^*) = \{x \in \Omega \mid q(x) > q^*\}$ with $\mathcal{B} \subset I_{\mathcal{B}}(q^*)$, $\forall q^* < 1$. The committor is thus a measure for the progress of a process or reaction, i.e. it is the ideal reaction coordinate for the process $\mathcal{A} \to \mathcal{B}$ [2, 6, 13]. Of special interest in this context is the isocommittor surface $I(0.5)$, which can be interpreted as the transition state ensemble in protein folding theory [21].

Once the committor has been computed, the change of any state variable $a(x)$ along the $\mathcal{A} \to \mathcal{B}$ process may be monitored by projecting onto this reaction coordinate using

$$a(q^*) = \mathbb{E}(a(x) \mid x \in I(q^*)) = \frac{\int_{x \in I(q^*)} dx\, \pi(x)\, a(x)}{\int_{x \in I(q^*)} dx\, \pi(x)}, \tag{4}$$

with $\pi(x)$ proportional to the statistical weight of state x and also the stationary distribution of state x, if this exists. In the latter case, one can define a dimensionless potential of mean force (PMF) along the $\mathcal{A} \to \mathcal{B}$ process given by

$$F(q^*) = -\log \frac{\int_{x \in I(q^*)} dx\, \pi(x)}{\int_{x \in \Omega} dx\, \pi(x)}. \tag{5}$$

The transport properties from $\mathcal{A}$ to $\mathcal{B}$ can be be computed *via* transition path theory (TPT) [15, 30]. In particular, the reactive flux f_{ij} between two states i and j is given by

$$f_{ij} = \pi_i q_i^- k_{ij} q_j^+ \tag{6}$$

for rate matrices [15], or

$$f_{ij}(\tau) = \pi_i q_i^- T_{ij}(\tau) q_j^+ \tag{7}$$

if the transition probability matrix is used [19]. Here, q^- is the backward committor which is the probability that of the two states $\mathcal{A}$ has been visited last and not $\mathcal{B}$ which

is equal to $1 - q^+$ if the dynamics is reversible. The reactive flux f_{ij} is proportional to the probability that a reactive trajectory, that is, a trajectory directly connecting $\mathcal{A}$ and $\mathcal{B}$, passes through the transition $i \to j$. The net transport through $i \to j$ is given by

$$f_{ij}^+ = \max\{f_{ij} - f_{ji}, 0\}, \tag{8}$$

which defines a network flow out of $\mathcal{A}$ and into $\mathcal{B}$ that can be decomposed into a set of $\mathcal{A} \to \mathcal{B}$ reaction pathways along with their probabilities [15, 19, 30].

Finally the $\mathcal{A} \to \mathcal{B}$ reaction rate is given by [19]:

$$k_{AB} = \frac{\sum_{i \in A, j \notin A} f_{aj}^+}{\sum_{i \in \Omega} q_i^-}. \tag{9}$$

Given the fundamental relevance of the committor probability in the characterization of dynamical processes, it is important to be able to compute $q(x)$ efficiently, and also to understand its sensitivity to perturbations, especially in cases where the system dynamics can be computed only approximately, e.g., by some sampling scheme such as

molecular dynamics simulations or Monte-Carlo dynamics. The remainder of the paper will concentrate on these numerical questions together with the illustration of the methods process on a simple 2-dimensional energy surface with metastable states and on a 3D model reminiscent of protein-ligand association.

COMMITTOR EQUATIONS

The committor is defined as the probability of reaching state $\mathcal{B}$ before state $\mathcal{A}$ is visited and thus corresponds to the result of a hypothetical experiment which starts a large number of Monte Carlo simulations in state s and measures $\boldsymbol{q}_s$ as the fraction of simulations, that reach $\mathcal{B}$ first.

Transition Matrix

We first derive the committor equations *via* the hitting times $h^{\mathcal{A}}$ of given subsets, which corresponds to the average number of steps a stochastic process needs to reach a set $\mathcal{A}$, if started at state x. Let $h^{\mathcal{A}}$ be the hitting time of set $\mathcal{A} \subset \Omega$: $h^{\mathcal{A}} : \Omega \to \mathbb{N} \cup \{\infty\}$ and X_i a time-discrete trajectory $X_i : i \to \Omega$ with initial starting point $X_0 = x$ given by

$$h^{A}(\omega) = \inf\{n > 0 : X_n(\omega) \in \mathcal{A}\},$$

Now consider the committor probability, q_i^+ pertaining to two sets A and B, which is the probability that, starting in state i, the system goes to $\mathcal{B}$ next rather than to $\mathcal{A}$ using the hitting times h

$$q_i^+ = \mathbb{P}_i(h^{\mathcal{B}} < h^{\mathcal{A}}) \equiv \mathbb{P}_i(h^{\mathcal{B}} < h^{\mathcal{A}})$$

where $\mathbb{P}_i$ indicates the probability for all trajectories X, that originate in state i.

In order to compute q_i^+, we use a recursive relation in the committor between connected points in configurational space Ω which states that the committor probability of a state $i \notin \mathcal{A} \cup \mathcal{B}$ is given by the sum of all products of the probabilities of reaching a neighboring state j given from the transition probability T_{ij} and the committor probability at state j while for states $\mathcal{A}$ and $\mathcal{B}$ we set the given solution to be in correspondence with the boundary conditions

$$q_i^+ = \begin{cases} 0 & \text{if } i \in \mathcal{A} \\ 1 & \text{if } i \in \mathcal{B} \\ \sum_j T_{ij} q_j^+ & \text{if } i \notin \mathcal{A}, \mathcal{B} \end{cases} . \tag{10}$$

The backward committor probability, q_i^- is defined respectively as the probability that being in state i, the system was in $\mathcal{B}$ last rather than in $\mathcal{A}$. In order to obtain the backward committor, we use the backward propagator

$$T_{ij}^- := \frac{\pi_j}{\pi_i} T_{ji}$$

which contains the probabilities that if the system is in state i then it came from state j. Proceeding in analogy to the forward committor we get

$$q_i^- = \begin{cases} 0 & \text{if } i \in \mathcal{A} \\ 1 & \text{if } i \in \mathcal{B} \\ \sum_{j \in I} T_{ij}^- q_j^- & \text{if } i \notin \mathcal{A}, \mathcal{B} \end{cases} .$$

For reversible dynamics the forward and backward propagators are equal , $T_{ij} = \frac{\pi_j}{\pi_i} T_{ji} = T_{ij}^-$ from which it follows immediately that

$$q_i^- = 1 - q_i^+ .$$

Rate Matrix

Given the rate matrix $K \in \mathbb{R}^{m \times m}$, we can use a similar arguments as for the time discrete case and derive expressions for the committor

$$\begin{aligned} q_i^+ &= 0 \quad \text{if } i \in A \\ q_i^+ &= 1 \quad \text{if } i \in B \\ \sum_{j \in I} K_{ij} q_j^+ &= 0 \quad \text{if } i \notin A \cup B. \end{aligned}$$

and the same equations hold also for the backward committor. A proof is given in the Appendix.

Transforming between Rate and Transition Matrices

It turns out that there is a simple way to transform rate matrices into transition matrices and *vice versa* that leaves the committor probabilities unchanged. This transformation is useful when a method is available to compute the committor from the transition matrices, but not for rate matrices, or *vice versa*.

Theorem 1. *Let $T(K) \in \mathbb{R}^{m\times m}$ be a stochastic matrix and $K \in \mathbb{R}^{m\times m}$ be a rate matrix related by the transformation:*

$$\mathbf{T}(\mathbf{K}) = \frac{c}{||\mathbf{K}||_\infty}\mathbf{K} + Id, \quad 0 < c < 1. \tag{11}$$

with $||\mathbf{K}||_\infty$ being the maximum norm representing the largest entry in the rate matrix. Then $\mathbf{T}(\mathbf{K})$ and $\mathbf{K}$ have the same committor probabilities for any choice of $\mathcal{A}$, $\mathcal{B}$.

The choice of c assures, that $T_{ij} \in [0, 1]$ and with the row sum of zero for rate matrices $\mathbf{T}(\mathbf{K})$ is a stochastic matrix. Scaling of matrices by a constant factor does not change the eigenvectors, as does the addition of a multiple of the identity matrix. Both operations, however, change the eigenvalues, which can be seen by writing down the expression for the characteristic polynomial, thus $\mathbf{T}^K$ inherits the same eigenvectors as $\mathbf{K}$, but with different eigenvalues. Note that although this transformation will leave the committor invariant, it will affect other dynamical properties of the matrix. In particular, $\mathbf{T}(\mathbf{K})$ will not reproduce the dynamical behaviour of the rate matrix on any but infinite timescales.

Numerical Solution

The committor equations above can be solved with any linear systems solver. When the system is very large and sparse, a sparse linear systems solver may still be able to handle them efficiently. An alternative approach to computing the committor probability from $\mathbf{K}$ has been proposed in [12]. However, this approach requires the $\mathbf{K}$-matrix to be inverted, which effectively limits its applicability to systems of $\leq 10^4$ states.

EIGENVECTOR FORMULATION

An alternative view is obtained when formulating the committor problem in terms of the dominant eigenvectors of either $\mathbf{K}$ or $\mathbf{T}(\tau)$. This is useful from a numerical point of view, because efficient solvers, such as the Power method or Krylov subspace methods, exist for dominant Eigenvectors. Moreover, it is useful from a physical standpoint as it allows the committor to be understood in terms of the slowest relaxation process of the system.

An approach to approximate $q(x)$ in terms of the second eigenvector of $\mathbf{K}$ or $\mathbf{T}(\tau)$ has been proposed in [1]. This approach is valid only if the second eigenvector is similar to the $\mathcal{A} \to \mathcal{B}$ committor and the second and third eigenvalues are well separated. In molecular processes, this is often referred to as "two-state" process, where there exists one slow process that is clearly separated from all other processes in terms of timescales. In the following, we will derive equations that allow the committors to be computed exactly in terms of its dominant eigenvectors for any Markovian system.

$\mathcal{A} \to \mathcal{B}$ Committor

We construct the transition matrix $\hat{\mathbf{T}}$ with absorbing states $\mathcal{A}$ and $\mathcal{B}$ from $\mathbf{T}$ by

$$\hat{T}_{ij} = \begin{cases} T_{ij} & i \notin \mathcal{A} \cup \mathcal{B}, j \in X \\ 1 & i \in \mathcal{A} \cup \mathcal{B}, j = i \\ 0 & i \in \mathcal{A} \cup \mathcal{B}, j \neq i \end{cases} . \tag{12}$$

and then define a transition matrix $\hat{\mathbf{T}}^{\infty}$ that transports any distribution infinitely into the future:

$$\hat{\mathbf{T}}^{\infty} = \lim_{n \to \infty} \hat{T}^n \tag{13}$$

and thus directly into either state $\mathcal{A}$ or $\mathcal{B}$. Thus the committor is given by

$$q_s = \sum_{k \in \mathcal{B}} ((\boldsymbol{e}^s)^T \hat{\mathbf{T}}^{\infty})_k = \sum_{k \in \mathcal{B}} \hat{T}^{\infty}_{sk} \tag{14}$$

In the following we will show that $\hat{T}^{\infty}$ and thus $\boldsymbol{q}$ are computationally fast and robust to derive.

Without loss of generality, we treat here the case where the sets $\mathcal{A} = \{a\}$ and $\mathcal{B} = \{b\}$ consist of only one state each. In cases where the sets are larger, they can simply be aggregated into a single state in the definition of $\hat{\mathbf{T}}$ and finally be diagonalized, obtaining:

$$\hat{\mathbf{T}}^{\infty} = \mathbf{R} \cdot \lim_{n\to\infty} \operatorname{diag}(\lambda_1^n, \lambda_2^n, \ldots, \lambda_N^n) \cdot \mathbf{R}^{-1}\cdot, \tag{15}$$

with $\mathbf{R} := [\boldsymbol{r}_1, ..., \boldsymbol{r}_N]$ being the matrix of right eigenvectors of $\hat{T}$ and λ_i are the corresponding eigenvalues, sorted from the largest to the smallest modulus of the eigenvalue. It follows from the Perron-Frobenius theorem that there exist exactly two left[1] eigenvectors with eigenvalue one, $\boldsymbol{e}^a$ and $\boldsymbol{e}^b$. The modulus of all other eigenvalues is strictly smaller than one. As a result,

$$\lim_{n\to\infty} |\lambda_i^n| = 0, \forall \lambda_i < 1, \tag{16}$$

and thus

$$\hat{\mathbf{T}}^{\infty} = \mathbf{R} \cdot \operatorname{diag}(1, 1, 0, \ldots, 0) \cdot \mathbf{R}^{-1}. \tag{17}$$

We now define $\mathbf{L} := \mathbf{R}^{-1}$ to be the inverse of the eigenvector matrix. $\mathbf{L}$ is a matrix of left eigenvectors since all of its rows fulfil the requirement for a left eigenvector with the same diagonalized eigenvalue matrix

$$\mathbf{L} \cdot \hat{T}^{\infty} = \mathbf{L}\,\Lambda. \tag{18}$$

This means that once we have the basis of left eigenvectors that equal $\mathbf{R}^{-1}$ we can avoid the expensive calculation of the inverse. Although general this is no advantage, in the present case the left eigenvectors of $\hat{\mathbf{T}}^{\infty}$ take a particularly simple form. We choose the following representation by row vectors for $\mathbf{L} := [\boldsymbol{l}_1, ..., \boldsymbol{l}_N]^T$ and get

$$\hat{\mathbf{T}}^{\infty} = [\boldsymbol{r}_1, ..., \boldsymbol{r}_N] \cdot \operatorname{diag}(1, 1, 0, \ldots, 0) \cdot [\boldsymbol{l}_1, ..., \boldsymbol{l}_N]^T \tag{19}$$

$$= [\boldsymbol{r}_1, \boldsymbol{r}_2] \cdot [\boldsymbol{l}_1, \boldsymbol{l}_2]^T. \tag{20}$$

As mentioned before the left eigenvectors to the eigenvalue of one are a linear combination of $\boldsymbol{e}^a$ and $\boldsymbol{e}^b$

$$[\boldsymbol{l}_1, \boldsymbol{l}_2]^T = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} [\boldsymbol{e}^a, \boldsymbol{e}^b]^T. \tag{21}$$

[1] The number of left and right eigenvectors to the same eigenvalue are equal.

Exploiting the fact that $\hat{\mathbf{T}}^{\infty}$ is still a stochastic matrix and thus has a constant right Perron Eigenvector, we can choose without loss of generality $\mathbf{1} := \boldsymbol{r}_1 = (1, \ldots, 1)$. Thus, only one second linear independent right eigenvector $\boldsymbol{r}_2$ needs to be computed:

$$\hat{\mathbf{T}}^{\infty} = [\mathbf{1}, \boldsymbol{r}_2] \cdot \mathbf{S} \cdot [\boldsymbol{e}^a, \boldsymbol{e}^b]^T, \tag{22}$$

Our goal was to compute the committor using (14) which leads us to the following relation for $\boldsymbol{q}^A$ and $\boldsymbol{q}^B$ respectively

$$[\boldsymbol{q}^A, \boldsymbol{q}^B] = \hat{\mathbf{T}}^{\infty} \cdot [\boldsymbol{e}^a, \boldsymbol{e}^b] \tag{23}$$

$$= [\mathbf{1}, r^2] \cdot \mathbf{S} \tag{24}$$

Thus we have shown that the committor is a linear combination of the right eigenvectors of $\hat{\mathbf{T}}^{\infty}$. To compute the mixing matrix $\mathbf{S}$ we make use of the fact that the solution is known already, by definition, for the two states $\mathcal{A}$ and $\mathcal{B}$:

$$\begin{pmatrix} q_k^A \\ q_k^B \end{pmatrix} = \begin{pmatrix} \delta_{ak} \\ \delta_{bk} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_k r_k^2 \end{pmatrix} \cdot \mathbf{S}, \quad k \in \{a, b\} \tag{25}$$

Writing this as a matrix equation leads to

$$\mathbf{S} = \begin{pmatrix} 1 & r_a^2 \\ 1 & r_b^2 \end{pmatrix}^{-1}, \tag{26}$$

yielding the solutions

$$[\boldsymbol{q}^U, \boldsymbol{q}^F] = [1, \boldsymbol{r}_2] \cdot \begin{pmatrix} (\mathbf{1})_a & (r^2)_a \\ (\mathbf{1})_b & (r^2)_b \end{pmatrix}^{-1} \tag{27}$$

$$q_i = (\boldsymbol{e}_i \hat{\mathbf{T}}^{\infty})_f = \frac{(r^2)_i - (r^2)_a}{(r^2)_b - (r^2)_a}. \tag{28}$$

Finally we have avoided the inversion of the matrix $\mathbf{R}$ required in Eq. (17) and instead reduced the effort to computing one largest right eigenvalue.

Based on Eq. (28), the committor probability can be easily computed for large sparse transition matrices using e.g. the Power method. When, instead, the system dynamics is specified in terms of the rate matrix, this computation can be performed by using the transformation (11). In the case of the Power method for solving for r_2, the parameter c should be as large as possible, but smaller than one. This will maximize the rate of convergence, since it maximizes the relative gap between the Perron-Eigenvalues and the next smaller eigenvalues.

Extension to multiple states

In many applications, it is desirable to compute more than one committor probability. Consider a system for which a number $M \geq 2$ of core sets have been defined, and for which at each state we wish to evaluate the probability that the system dynamics will hit the core i rather than any other core. This defines a set of M committors, $[\boldsymbol{q}^{y_1}, \ldots, \boldsymbol{q}^{y_M}]$, where $\boldsymbol{q}^{y_i}$ indicates the vector of committor probabilities of going to core i next rather than any other core, and each row sums up to 1 ($\boldsymbol{q}^{y_1} + \ldots + \boldsymbol{q}^{y_M} = 1$), thus forming a membership probability.

To solve this general case, all states $[Y_1, \ldots, Y_M]$ are made absorbing in the transition matrix, and a basis for all eigenvectors of the eigenvalue of one is computed. The parameters for the eigenvectors can then be computed using a simple matrix inversion in analogy to the two state case by

$$[\boldsymbol{q}^{y_1}, \ldots, \boldsymbol{q}^{y_M}] = [\mathbf{1}, \ldots, \boldsymbol{r}^M] \cdot \begin{pmatrix} \mathbf{1}_{y_1} & \cdots & (\mathbf{r_x})_{y_1} \\ \vdots & \ddots & \vdots \\ \mathbf{1}_{y_M} & \cdots & (\boldsymbol{r}^M)_{y_M} \end{pmatrix}^{-1}, \tag{29}$$

where $y_1, \ldots, y_M$ are the states, $r_2, \ldots, r_x$ the eigenvectors of the eigenvalue of one and $\mathbf{1}$ is again the constant right Perron-Eigenvector.

SENSITIVITY AND UNCERTAINTY

We now characterize the sensitivity of the committor $\boldsymbol{q}$ to changes in the transition matrix given by $\frac{\partial q_i}{\partial T_{ab}}$ and also examine how the sensitivity leads to a first-order estimate of the uncertainty of the committor $\delta\boldsymbol{q}$ in cases where the transition matrix $\mathbf{T}$ is not exactly known, but is for example estimated from simulation data such as from molecular dynamics [24, 25].

Sensitivity analysis

We are interested in $\frac{\partial \boldsymbol{q}}{\partial T_{ab}}$, *i.e.* the sensitivity of the committor with respect to perturbations in the transition matrix and define $\hat{\mathbf{A}} := \hat{\mathbf{T}} - \mathrm{Id}$, so that the null space of $\hat{\mathbf{A}}$ is the

space spanned by the eigenvectors to the eigenvalue of $\lambda_1 = \lambda_2 = 1$, *i.e.*

$$\hat{\mathbf{A}}\boldsymbol{q} = 0. \tag{30}$$

First, we start with the derivative of (30) with respect to T_{ab}

$$\frac{\partial \hat{\mathbf{A}}\boldsymbol{q}}{\partial T_{ab}} = \hat{\mathbf{A}} \cdot \frac{\partial \boldsymbol{q}}{\partial T_{ab}} + \frac{\partial \hat{\mathbf{A}}}{\partial T_{ab}} \cdot \boldsymbol{q} = 0. \tag{31}$$

and make the convention that all derivatives are taken at T_{ab}, if not specified otherwise. Since $\hat{\mathbf{A}}$ does not have full rank and its inverse is not defined, so that we use

$$\frac{\partial \hat{A}_{ij}}{\partial T_{ab}} = \frac{\partial \hat{T}_{ij}}{\partial T_{ab}} = \begin{cases} \delta_{ia}\delta_{jb} & i \notin \mathcal{A}, \mathcal{B} \\ 0 & i \in \mathcal{A}, \mathcal{B} \end{cases} \tag{32}$$

and then rewrite this (31) as

$$\sum_k \hat{A}_{ik} \cdot \frac{\partial q_k}{\partial T_{ab}} = -q_b \begin{cases} \delta_{ia} & i \notin \mathcal{A}, \mathcal{B} \\ 0 & i \in \mathcal{A}, \mathcal{B} \end{cases}.$$

Since $\frac{\partial q_k}{\partial T_{ab}} = 0$ for $k \in \mathcal{A}, \mathcal{B}$ we can exclude these from the calculation and define a reduced inverse $\tilde{\mathbf{A}}^{-1}$ given by

$$\tilde{\mathbf{A}}^{-1} = \begin{pmatrix} 0 & & \cdots & & 0 \\ & \begin{bmatrix} T_{22}-1 & \cdots & T_{2,M-1} \\ \vdots & \ddots & \vdots \\ T_{M-1,2} & \cdots & T_{M-1,M-1}-1 \end{bmatrix}^{-1} & & \\ \vdots & & & & \vdots \\ 0 & & \cdots & & 0 \end{pmatrix}$$

which is inverted only on the subset of states neither in $\mathcal{A}$ or $\mathcal{B}$, and the remaining transitions are set to zero, thus assuring the correct boundary conditions for $\frac{\partial q_i}{\partial T_{ab}}$. This yields the sensitivity matrix S^a_{ib} defined by

$$S^a_{ib} := \frac{\partial q_i}{\partial T_{ab}} = -\sum_l \tilde{A}^{-1}_{il}\delta_{la}q_b = -\tilde{A}^{-1}_{ia}q_b. \tag{33}$$

Uncertainty / Sampling Error of the Committor

Let us now consider the case where the transition matrix $\mathbf{T}$ is not known exactly but is instead sampled by a finite number of observations, as is the case, for example, in molecular dynamics simulations. [17, 18, 24, 25]. We will be interested in the question of how the uncertainty involved in this finite sampling translates into uncertainty of the committor. Let $Z \in \mathbb{R}^{m\times m}$ be a count matrix with Z_{ij} being the number of independently observed transitions from state i to state j. The likelihood of transition matrices pertaining to this observation is given by:

$$\mathbb{P}(\mathbf{C} \mid \mathbf{T}) = \prod_{i,j} T_{ij}^{Z_{ij}}$$

When restricting the Prior distribution to the conjugate Dirichlet prior, the posterior distribution can be expressed as:

$$\mathbb{P}(\mathbf{T} \mid \mathbf{C}) \propto \mathbb{P}(\mathbf{T})\mathbb{P}(\mathbf{C} \mid \mathbf{T}) = \prod_{i,j} T_{ij}^{B_{ij}+Z_{ij}} = \prod_{i,j} T_{ij}^{C_{ij}}$$

where B_{ij} are prior counts. Comparing to the Dirichlet distribution

$$\prod_i \prod_j T_{ij}^{\alpha_{ij}-1} = \prod_i Dir(\boldsymbol{\alpha}_i)$$

with $\boldsymbol{\alpha}_i := \{\alpha_{i1}, \ldots, \alpha_{iM}\}$ results in the equivalence

$$\alpha_{ij} = C_{ij} + 1 = B_{ij} + Z_{ij} + 1. \tag{34}$$

The maximum likelihood transition matrix $\hat{T}_{ij}$ is given by

$$\hat{T}_{ij} = \frac{Z_{ij}}{\sum_k Z_{ik}}$$

and the mean of the posterior distributions $\bar{T}_{ij}$ by:

$$\bar{T}_{ij} = \frac{\alpha_{ij}}{\sum_k \alpha_{ik}} = \frac{B_{ij} + Z_{ij} + 1}{\sum_k (B_{ik} + Z_{ik} + 1)}$$

and both are equivalent for the Null prior $B_{ij} = -1$. Eq. 34 shows that the prior can be regarded as counts additional to the actual observed counts Z_{ij}. Thus, to obtain an

expectation based mainly on observations the number of real observations Z_{ij} must be larger than the number of prior counts

$$\sum_k Z_{ik} \gg \sum_k (B_{ik} + 1).$$

This forces us to be careful about the choice of the prior, which, in principle, compensates for the lack of information in states with few of in none observed transitions.

One choice is the Null prior, which adds no additional counts and thus the mean and maximum of the posterior probability distribution are equal. Another choice is a uniform prior probability distribution $\mathbb{P}(T) \propto 1 \Leftrightarrow B_{ij} = 0$, which will prove inadequate in the cases we consider, since $\sum_k Z_{ik} \gg m$. A further choice might be to distribute one additional count per state by $B_{ij} = 1/m$ and thus request $\sum_k Z_{ik} \gg 1$. Yet another approach is to use a prior that has counts restricted to a certain subset of elements. We will address this issue again in the application section.

As we have shown before, the probability distribution can be written as a product of independent Dirichlet distributions for each state. Hence, the covariance between entries in the transition matrix is zero between elements from different rows and we can define a set of reduced covariance matrices Σ^i_{ab} for each state or equivalently row in the transition matrix i separately by the expression

$$\Sigma^i_{ab} := \mathrm{Cov}\left(T_{ia}, T_{ib}\right) = \frac{\alpha_{ia}\left(\alpha_i \delta_{ab} - \alpha_{ib}\right)}{\alpha_i^2(\alpha_i + 1)}. \tag{35}$$

This leads finally to an expression for the standard deviation of each entry of the transition matrix

$$\delta T_{ia} = \sqrt{\mathrm{Cov}\left(T_{ia}, T_{ia}\right)} = \sqrt{\frac{\alpha_{ia}(\alpha_i - \alpha_{ia})}{\alpha_i^2(\alpha_i + 1)}}$$

with

$$\alpha_i := \sum_{j=1}^{m} \alpha_{ij}.$$

A simple and often used approach for propagating the uncertainty in $\mathbf{T}$ to the uncertainty of the committor (or any other property derived from $\mathbf{T}$), is to sample the posterior

distribution of transition matrices and compute the committor for each sample of this distribution [18, 25]. However, this procedure involves sampling itself and thus uncertainty in the estimation of the uncertainty, which may be undesirable in situations where the uncertainty estimation is conducted repeatedly, *e.g.* within an adaptive sampling scheme [24, 25].

An alternative is to propagate the covariance from the transition matrix elements linearly to the covariance in the committor using the computed sensitivity S^i_{ab} by

$$\begin{aligned}\mathrm{Cov}\left(q_a, q_d\right) = &\sum_{i,b,c=1}^{m} S^i_{ab}\Sigma^i_{bc}\left(S^T\right)^i_{cd} \\ &\sum_{i=1}^{m}\left(\tilde{\mathbf{A}}^{-1}\right)_{ai}\left(\tilde{\mathbf{A}}^{-1}\right)_{di}\sum_{b,c=1}^{m} q_b \frac{\alpha_{ib}\left(\alpha_i\delta_{bc}-\alpha_{ic}\right)}{\alpha_i^2(\alpha_i+1)}q_c\end{aligned}$$

and finally we can compute the variance in the elements of the committor by

$$\begin{aligned}\delta^2 q_a &= \mathrm{Cov}\left(q_a, q_a\right) &(36)\\ &= \sum_{i=1}^{m}\frac{1}{\alpha_i^2(\alpha_i+1)}\left(\tilde{\mathbf{A}}^{-1}\right)^2_{ai}\left(\alpha_i\sum_{b=1}^{m} q_b\alpha_{ib}q_b - \left(\sum_{b=1}^{m} q_b\alpha_{ib}\right)\left(\sum_{c=1}^{m}\alpha_{ic}q_c\right)\right). &(37)\end{aligned}$$

A complete derivation can be found in the Appendix. Clearly, the variance can be separated into contributions from each state i and we define a uncertainty contribution vector w_i by the norm of the single contributions

$$w_i = \left|\left|\frac{1}{\alpha_i^2(\alpha_i+1)}\left(\tilde{\mathbf{A}}^{-1}_{ai}\right)^2\left(\alpha_i\sum_{b=1}^{m} q_b\alpha_{ib}q_b - \left(\sum_{b=1}^{m} q_b\alpha_{ib}\right)\left(\sum_{c=1}^{m}\alpha_{ic}q_c\right)\right)\right|\right|_2 \qquad (38)$$

which can then be used in order to direct new simulations that are most promising in reducing the error [24].

APPLICATIONS

Diffusion in a 2D Three-Well Potential

To illustrate an application of the above equations we use a simple model of a particle diffusing in a two-dimensional potential with three wells (Fig. 1), partitioned into a

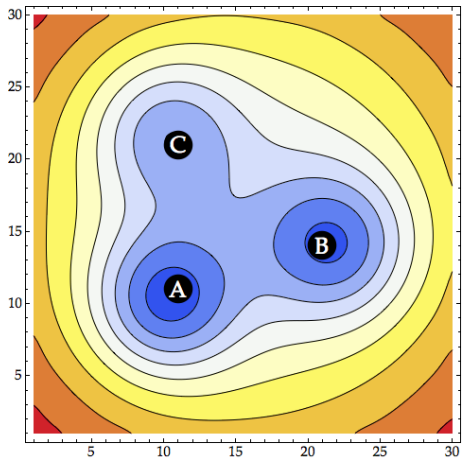

Figure 1. Energy Landscape for diffusion in a 2D potential with 3 basins discretized into a grid of 30x30 bins. The minima in each basin are indicated by the letters A, B and C. Blue indicates low energies, red high energies.

grid of $m = 30 \cdot 30 = 900$. The minima and their associated regions of configurational space will be referred to as A, B and C. Transition probabilities are defined based on the potential energies U_i on each gridpoint using a Metropolis acceptance criterion given by

$$T_{ij} = \frac{\mathbb{P}\left(i \to j\right)}{\sum_k \mathbb{P}\left(i \to k\right)} = \frac{\min\left(1, \exp\left(-\beta\left(U_j - U_i\right)\right)\right)}{\sum_k \min\left(1, \exp\left(-\beta\left(U_k - U_i\right)\right)\right)}, \tag{39}$$

with $\beta = 1$, which has the correct invariant distribution $\pi_i \propto \exp\left(-\beta U_i\right)$. Only transitions between horizontal or vertical neighboring microstates are allowed, resulting in a maximum of five nonzero entries per row in the 900x900 transition matrix. This matrix is used as the reference for the dynamics of the system. The committor from state A to C, as given in Eq. (28), is shown in Fig. 2.

To investigate the dependence of the committor and its uncertainty on the actual number of observations and the chosen prior probability distribution, we computed the expected number of observed transitions in an equilibrium simulation as

$$\bar{Z}_{ij} = L\ \pi_i\ \hat{T}_{ij},$$

which is the product of the total number of simulation steps L , the invariant density of a state π_i and the true transition probabilities $\hat{T}_{ij}$. Four different types of prior distributions are considered here (see Tab. I).

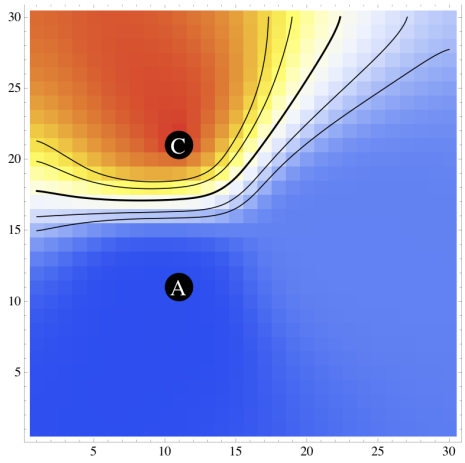

Figure 2. 2D three-well model: Committor from state A to C computed directly from the reference transition probability matrix T_{ij}.

Prior	B_{ij}
Null Prior	-1
$1/m$ Prior	$1/m - 1$
Neighbor Prior	$\begin{cases} 0 & \text{if } (i,j) \text{ neighbors} \\ -1 & \text{else} \end{cases}$
Uniform Prior	0

Table I. Prior probability distributions used for the 2D example

The committors computed for different simulation lengths $L = \{10^1, 10^3, 10^5, 10^7\}$ and all prior sets except the null prior are presented in Fig. 3. The null prior was omitted since in this case the committor does not depend on the simulation length L and equals the exact committor (Fig. 2). It is important to note that this equivalence is only true on average and not for every possible simulation outcome. The influence of the full uniform prior is so strong that the computed committor differs from the true committor vastly even for $L = 10^7$. The other two priors behave similarly while the neighbor prior has the general advantage over the null prior that it always provides a transition matrix that can numerically be evaluated.

Eq. 36 gives the expression for the uncertainty in the computed average committor from a given number of observations. For the same set of total observations L and all

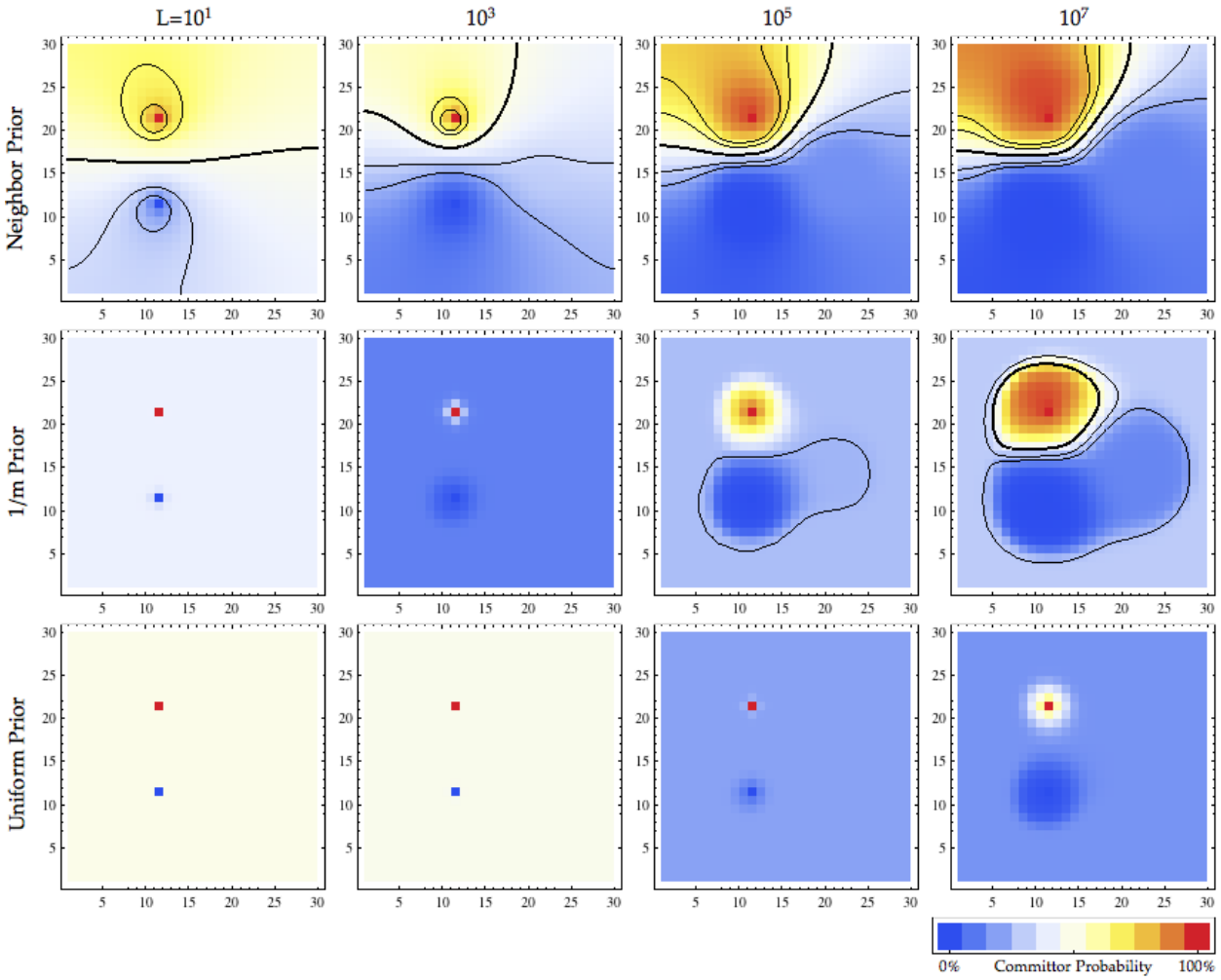

Figure 3. 2D three-well model: Committor from state A to C computed for different prior choices (rows: neighbor prior, $1/m$ prior, full uniform) and simulation lengths (columns: $L = \{10^1, 10^3, 10^5, 10^7\}$). Isocommittor surfaces for $q = \{0.25, 0.33, 0.5, 0.67, 0.75\}$ are given in black.

priors in Table I the covariance was computed and is shown in Fig. 4. The main uncertainty is always greatest in the transition region, and depends strongly on the choice of the prior, especially when few observations have been made.

Fig. 6 show the difference in the predicted committors compared to the reference committor given in Fig. 2. The quality of the average predicted committor depends mainly on the amount of prior information put into the predictions: Priors with few information (null prior, neighbor prior) allow for better predictions, while priors with much information (1/m prior, uniform prior) give worse committor prediction, but are less sensitive to perturbations in the transition matrix elements which is depicted in Fig.

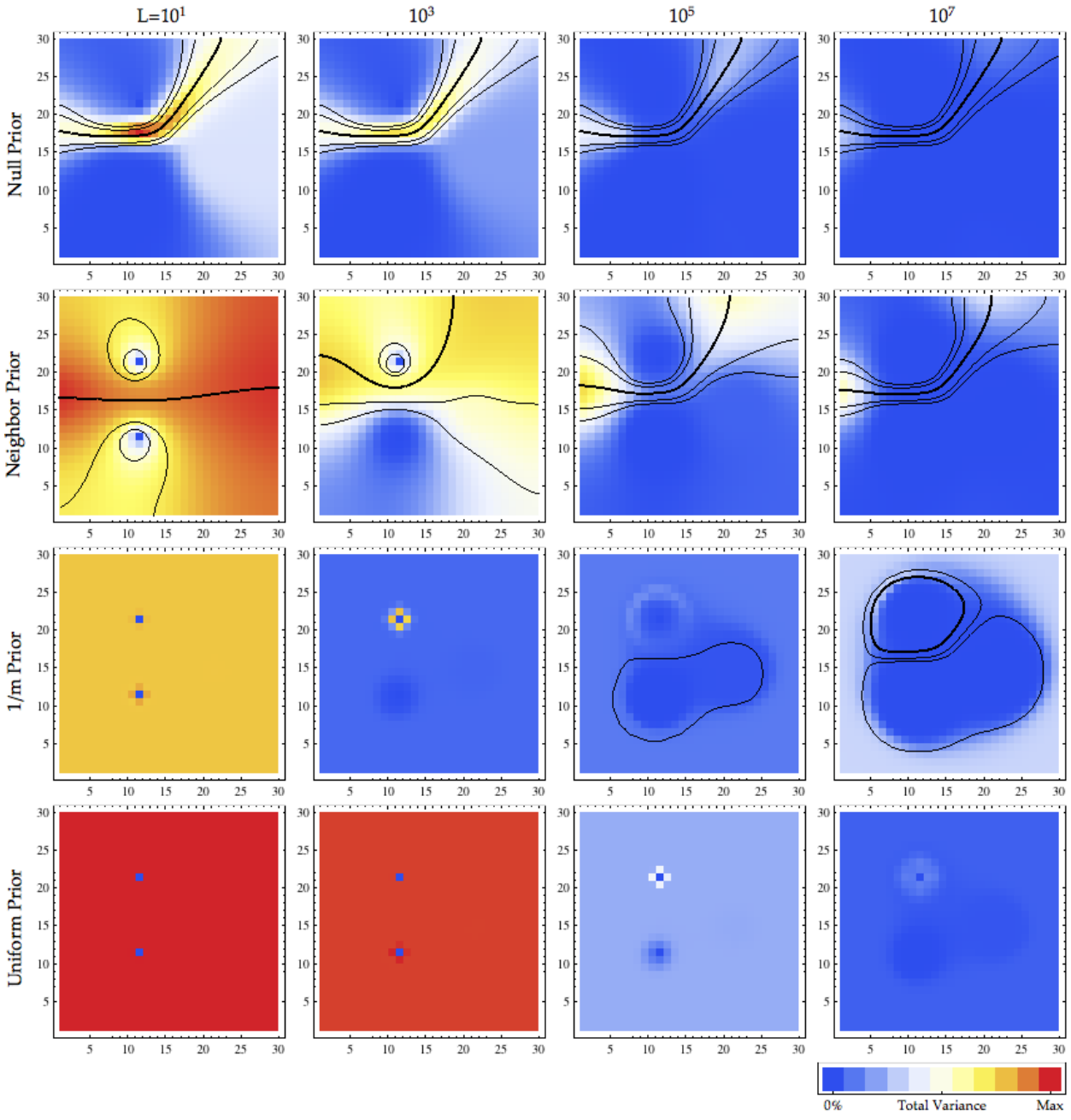

Figure 4. 2D three-well model: Total variance in the entries of the committor probability $\mathrm{Cov}\,(q_i, q_i)$ in Eq. (36) from state A to C for different prior choices (rows: null prior, neighbor prior, $1/m$ prior, full uniform) and simulation lengths (columns: $L = \{10^1, 10^3, 10^5, 10^7\}$). Iso-committor surfaces from Fig. (3) shown in black. Blue indicates no variance, red indicates high change. All plots have been normalized per row, so absolute comparison between plots is possible only for the same prior configuration. The related absolute error development is given in Fig. (5). States A and C are fixed by definition, thus at this points the change vanishes. The highest variation is found in the transition region, the size of which depends strongly on the prior information. With increasing simulation length, the error in the low energy states reduces fastest.

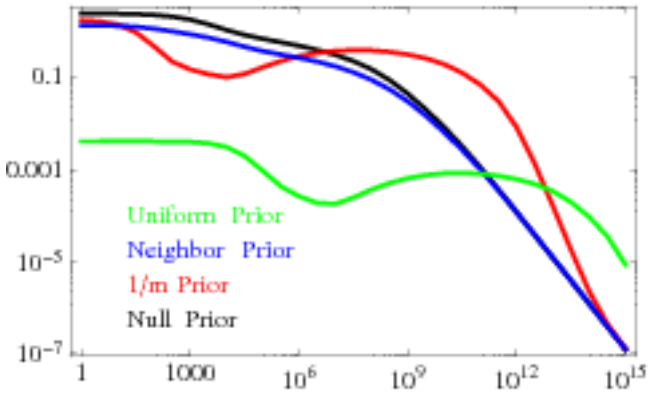

Figure 5. 2D three-well model: Theoretical average uncertainty in the estimated committor for different prior probability distributions (null prior, neighbor prior, $1/m$ prior, full uniform) versus simulation length L. The initial erratic behaviour of the $1/m$ prior and uniform prior is caused by a wrong committor prediction due to the high impact of these prior, when only few transitions have been observed.

5. The fact, that the $1/m$ prior and the uniform prior have a smaller uncertainties for small simulation lengths is due to the wrong committor predictions discussed before which are less sensitive to changes in the transition matrix elements. This behavior changes once the simulation length is long enough for the estimated committors to be similar.

The effects of differences in the prior probabilities are also visible in the contribution to the uncertainty from each state i by w_i in Eq. (38) as shown in Fig. 7. In general the main contributions to the uncertainty is located in states inside the transition region. For small simulation lengths L the contribution is more widely distributed and mainly in regions that have also a significant equilibrium probability. With increasing simulation time, the uncertainty contributing states shift towards the outer perimeter of the energy landscape, where the uncertainty remains mostly unchanged since these parts of phase space are hardly visited at all.

The net flux for the system as given by Eq. (8) is shown in Fig. (8). The opacity of the arrows indicates the intensity of the flux in the direction of the arrow. The main fraction of the flux traverses the barrier between A and C, while a minor fraction travels over state B.

Finally, the 3-state committor, given by Eq. (29) was computed for states A, B and C (s. Fig. 9), thus partitioning the configurational space into three subsets divided by the main

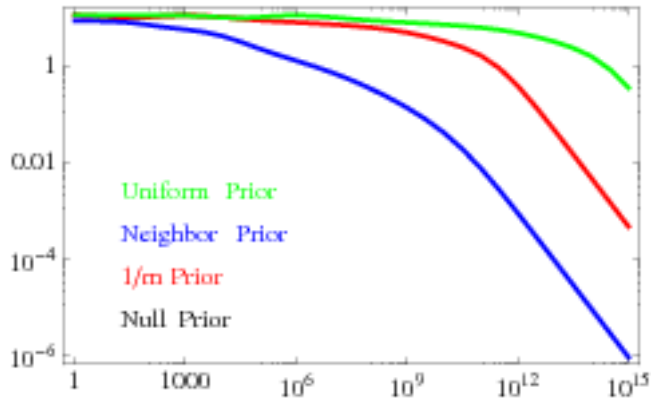

Figure 6. 2D three-well model: Norm of the difference of the computed committor for different prior probability distributions (null prior, neighbor prior, $1/m$ prior, full uniform) versus simulation length L. The uniform estimation is about six orders of magnitude slower in convergence since the amount of prior information is also about six orders of magnitude larger compared to the other priors.

barriers. In this manner the multistate committor can be used to partition the configurational space into subsets, that are dynamically close to one state of a set of predefined states which can be regarded as cluster centers.

3D MODEL

The method is now further examined on a simple model system that mimics diffusional protein:ligand association. For this, a 3-dimensional potential was defined by a potential function U

$$U(x) = \sum_i \frac{b_i}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\bar{x}_i)}{2\sigma_i^2}\right)$$

as the sum of five 3D-Gaussian functions as an exemplary electrical field in which the ligand diffuses (for parameters see table II).

The potential was coarse-grained on a grid with a total of $m = 100 \cdot 100 \cdot 100 = 10^6$ states in the range of $[-1,1] \times [-1,1] \times [-1,1]$. The dynamics was modeled as a diffusional process under the influence of the potential as in the previous 2D case (see Eq. (39)). Figure (10) shows equipotential surfaces for a set of 19 exponentially spaced values

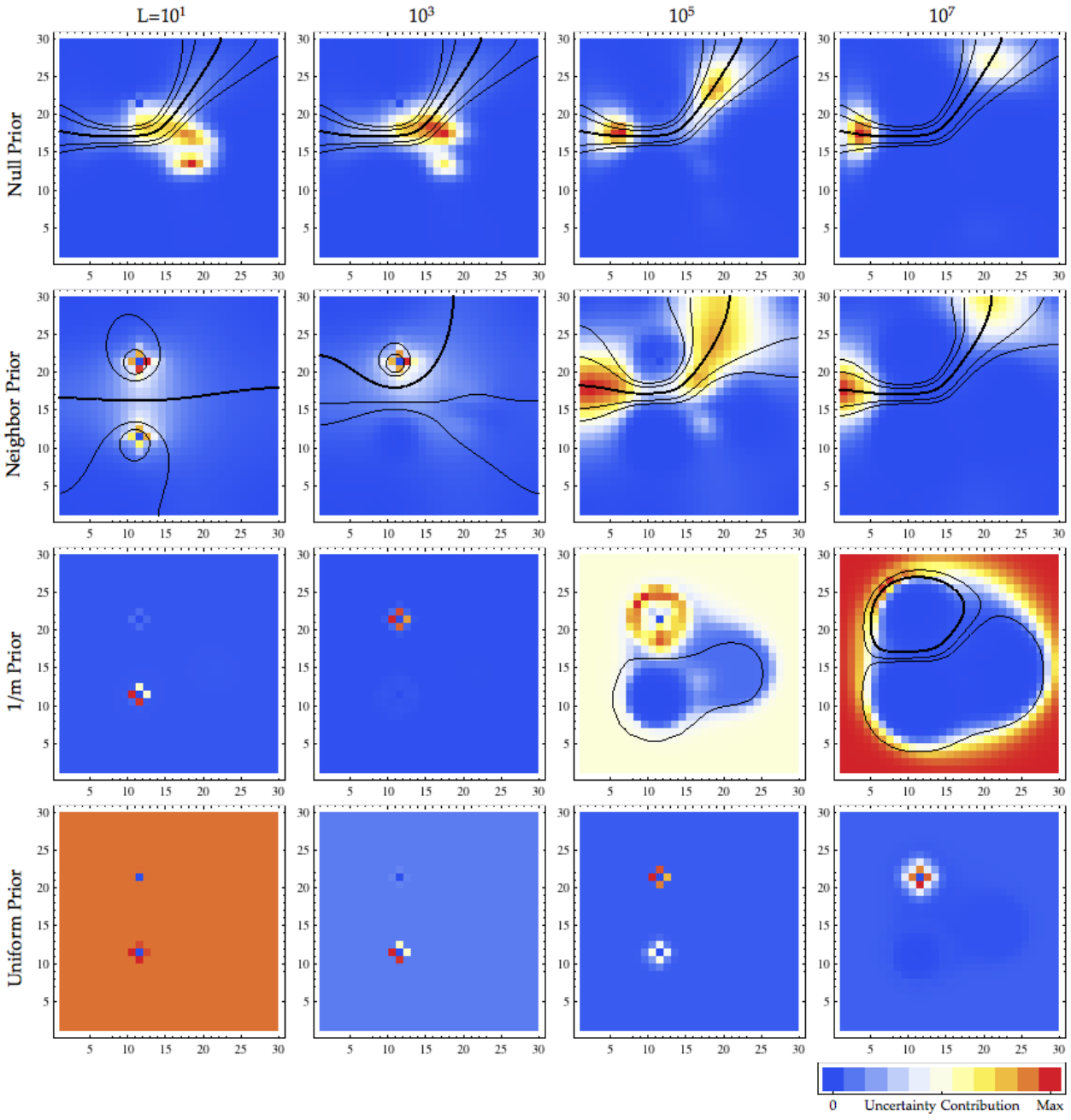

Figure 7. 2D three-well model: Sensitivity vector w_i in Eq. (38) for different prior choices prior choices (rows: null prior, neighbor prior, $1/m$ prior, full uniform) and simulation lengths (columns: $L = \{10^1, 10^3, 10^5, 10^7\}$). Isocommittor surfaces from Fig. (3) shown in black. Blue indicates vanishing sensitivity, red maximal sensitivity for each plot separately, thus absolute comparison is not possible between plots. This was chosen to better indicate the highest uncertainty contributions. The absolute sensitivity is given in Fig. (5). The figure shows that in the case of the uniform prior a length of $L = 10^7$ is insufficient for an accurate description of the sensitivity.

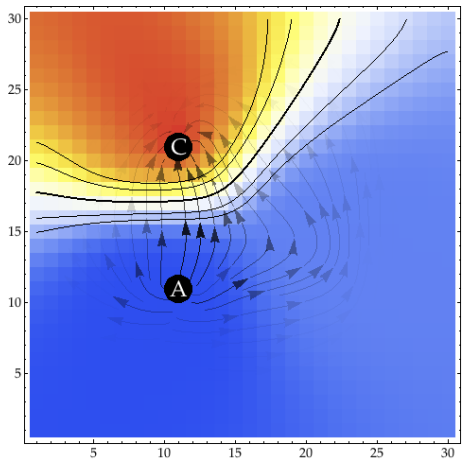

Figure 8. 2D three-well model: Net Flux between states A and C computed from the reference transition matrix $\hat{T}_{ij}$. The underlying colors represent the reference committor. Arrows indicate the direction of the flux and the opacity the intensity. Most flux travels over the direct barrier from state A to state C.

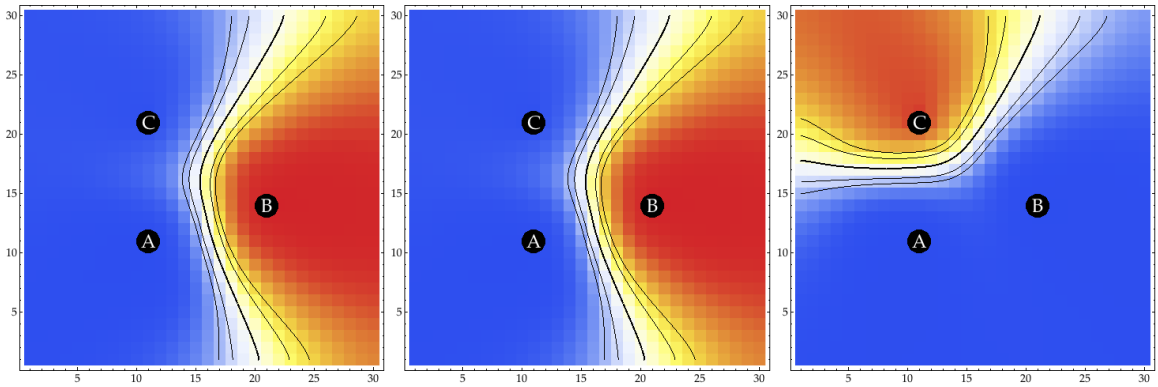

Figure 9. Committor Computed for 3 states from Eq. (29). The committor clearly shows a clear separation of the configurational space into 3 subsets divided by the potential barriers.

i	Sign	Mean $\bar{x}$	Std Dev σ
1	-	$\{0.0, 0.0, -0.2\}$	0.10
2	-	$\{-0.6, 0.2, -0.6\}$	0.08
3	-	$\{-0.6, 0.4, 0.4\}$	0.08
4	+	$\{0.4, -0.6, -0.6\}$	0.05
5	-	$\{-0.6, -0.6, -0.6\}$	0.05

Table II. 3D Ligand:protein model: Parameters for the manually defined potential U.

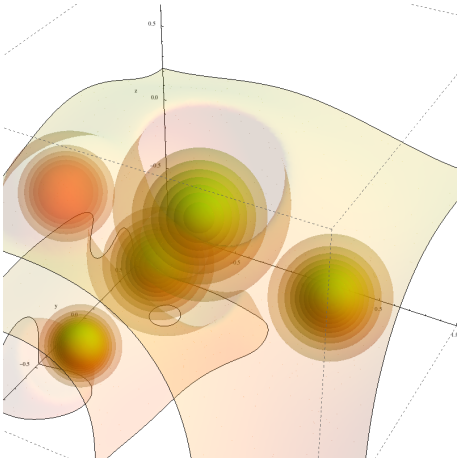

Figure 10. 3D Ligand:protein model: Equipotential surfaces.

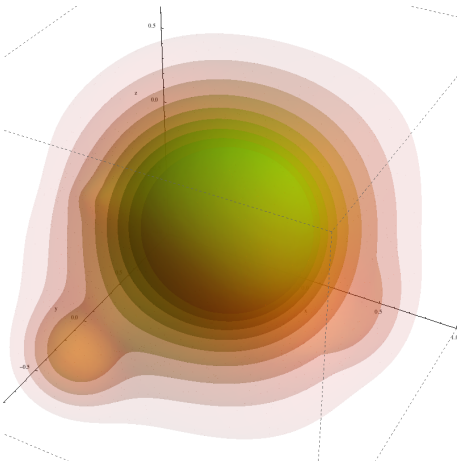

Figure 11. 3D Ligand:protein model: Isocommittor surfaces for the potential U.

of the potential U, effectively depicting surfaces of equal equilibrium probability.

The outer boundary of the grid is defined as the "unbound" state $\mathcal{A}$ while all states inside a sphere at the center of the grid with a radius of 0.2 define the "bound" state $\mathcal{B}$. The committor probability was computed using the procedure described in the theory section, employing the Power method to solve for the dominant eigenvector of the absorbing process [8]. The isocontours of the committor are shown in Fig. 11. It is seen that these contours are roughly spherical around the binding site B, but have protrusions due to the existence of local energy minima.

Fig. (12) shows some paths integrated along the normals to the isocommittor hypersurfaces. To compute these the committor function, given on each grid point, was inter-

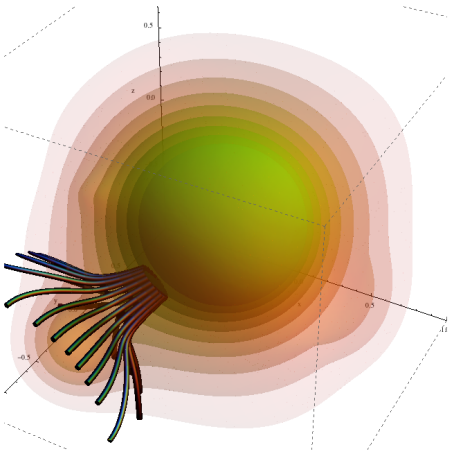

Figure 12. 3D Ligand:protein model: Bundle of path lines starting at the virtual binding site along the normals to the isocommittor surfaces.

polated by linear polynomials between each neighboring grid point and computed the normals from the continuous interpolation. As initial points a set of 20 circularly positioned points on the inner $\mathcal{B}$ state were chosen which were directed toward the potential minimum at Point 5 in Tab. II. The integrated paths define a bundle of field lines connecting the outer perimeter and the binding site, depicting the most probable paths towards the virtual binding site on the protein.

Using the committor also the reactivity g [15], *i.e.* the probability that a state contributes to a reactive trajectory, was computed using

$$g_i = q_i^+ \pi_i q_i^- \tag{40}$$

The results are shown in Fig. 13. Due to the higher equilibrium probability in Eq. (40), the density of reactive trajectories increases towards the binding site and especially in the local minima.

CONCLUSIONS

In this paper we introduced an alternative way to compute the committor for space discrete system with dynamics given by both transition or rate matrices.

The method presented allows to retrieve efficiently, fast and easy to implement committor information for dynamical systems with a very large discrete configurational state

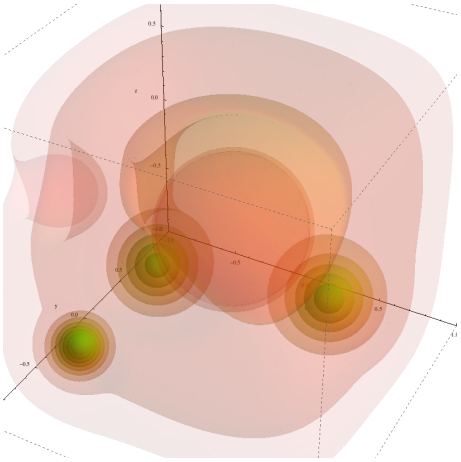

Figure 13. 3D Ligand:protein model: Density of reactive trajectories g_i as given in Eq. (40).

spaces. If the considered transition matrix is sparse enough even very large systems can be investigated with computational effort roughly proportional to the number of states and is thus limited only by memory constraints. In addition, the method in principle allows to compute the committor for the environment of a large protein with sufficient resolution to display folding bundles as we have demonstrated with the 3D example.

The sensitivity analysis provides a detailed error measure of the computed committor and also allows to an adaptive algorithm to be defined for fast computation of the committor by collecting information from different parts of the configurational space separately and combining this to produce more accurate estimations than possible from one single long simulation. Computation of the sensitivity requires the inversion of a matrix of the size of the number of states which is in general of cubic order, but can be made quadratic if the matrix is sufficiently sparse. The other computations are also maximally of quadratic order, which in principle also allows a sensitivity analysis for medium system sizes.

ACKNOWLEDGMENTS

We acknowledge funding from the DFG Research Foundation through the International Graduate College 710 and the Research Center Matheon. JCS acknowledges funding from the United States Department of Energy Biological and Environmental Research and Advanced Scientific Research sections "Multiscale Modeling" initiative.

* jan-hendrik.prinz@fu-berlin.de

† martin.held@fu-berlin.de

‡ smithjc@ornl.gov

§ frank.noe@fu-berlin.de

[1] Alexander M Berezhkovskii and Attila Szabo. Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices. *J. Chem. Phys*, 121(18):9186, Jan 2004.

[2] Robert B Best and Gerhard Hummer. Reaction coordinates and rates from transition paths. *Proc. Nat. Acad. Sci. USA*, 102(19):6732–6737, Jan 2005.

[3] Peter G Bolhuis, David Chandler, Christoph Dellago, and Phillip L Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291, Jan 2002.

[4] Christoph Dellago, Peter G Bolhuis, and Phillip L Geissler. Transition path sampling. *Advances in Chemical Physics*, 123:1, Jan 2002.

[5] Rose Du, Vijay S Pande, Alexander Yu Grosberg, Toyoichi Tanaka, and Eugene I Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys*, 108:334, 1998.

[6] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chem. Phys. Lett.*, 413(1-3):242–247, 2005.

[7] Bernd Ensing, Alessandro Laio, Francesco Luigi Gervasio, Michele Parrinello, and Michael L Klein. A minimum free energy reaction path for the e2 reaction between fluoro ethane and a fluoride ion. *J. Am. Chem. Soc.*, 126(31):9492–3, Aug 2004.

[8] G Golub and Van Loan. *Matrix computation*.

[9] I Horenko, R Klein, S Dolaptchiev, and C Schütte. Automated generation of reduced stochastic weather models i: simultaneous dimension and .... *Multiscale Modeling & Simulation*, Jan 2008.

[10] Illia Horenko. Finite element approach to clustering of multidimensional time series. *to appear in SIAM J. Sci. Comp.*

[11] Gerhard Hummer. From transition paths to transition states and rate coefficients. *J. Chem. Phys*, 120(2):516–523, Jan 2004.

[12] Peter Lenz, Bojan Zagrovic, Jessica Shapiro, and Vijay S Pande. Folding probabilities: A novel approach to folding transitions and the two-dimensional ising-model. *J. Chem. Phys*, 120(14):6769, Jan 2004.

[13] A Ma and Aaron R Dinner. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B*, 109:6769–6779, Jan 2005.

[14] Luca Maragliano, Alexander Fischer, Eric Vanden-Eijnden, and Giovanni Ciccotti. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys*, 125(2):24106, Jul 2006.

[15] Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden. Transition path theory for markov jump processes. *Multiscale Model. Sim.*, 7(3):1192–1219, 2009.

[16] Hazime Mori. Transport, collective motion, and brownian motion. *Prog. Theo. Phys.*, 33(3):423–455, 1965.

[17] F Noé and S Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.*, Jan 2008.

[18] Frank Noé. Probability distributions of molecular observables computed from markov models. *J. Chem. Phys*, 128(24):244103, Jun 2008.

[19] Frank Noé, Christof Schütte, Lothar Reich, and Thomas R Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium simulations. *Proc. Nat. Acad. Sci. USA*, 2009.

[20] AC Pan and David Chandler. Dynamics of nucleation in the ising model. *J. Phys. Chem. B*, 108(51):19681–19686, Jan 2004.

[21] Vijay S Pande, Alexander Yu Grosberg, Toyoichi Tanaka, and Daniel S Rokhsar. Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.*, 8(1):68–79, Feb 1998.

[22] Baron Peters and Bernhardt L Trout. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys*, 125(5):054108, Jan 2006.

[23] Marco Sarich, Frank Noé, and Christof Schütte. Error in discretization of markov models. *Multiscale Model. Sim.*, 2009.

[24] Nina Singhal and Vijay S Pande. Error analysis and efficient sampling in markovian state models for molecular dynamics. *J. Chem. Phys*, 123(20):204909, Nov 2005.

[25] Nina Singhal, Christopher D Snow, and Vijay S Pande. Using path sampling to build better markovian state models: predicting the folding rate and mechanism of a tryptophan zipper

beta hairpin. *J. Chem. Phys*, 121(1):415, Jul 2004.

[26] A Spaar, Christian Dammer, Razif R Gabdoulline, Rebecca C Wade, and Volkhard Helms. Diffusional encounter of barnase and barstar. *Biophys. J.*, 90(6):1913–1924, 2006.

[27] Paul J Steinhardt, D R Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Phys. Rev. B*, 28(2):784–805, 1983.

[28] Floris Takens. Detecting strange attractors in turbulence. pages 366–381. 1981.

[29] H TAKETOMI, Y UEDA, and N GO. Studies on protein folding, unfolding and fluctuations by computer-simulation .1. effect of specific amino-acid sequence represented by specific inter-unit interactions. *Int J Pept Prot Res*, 7(6):445–459, Jan 1975.

[30] Eric Vanden-Eijnden. Transition path theory. volume 703, pages 453–493. 2006.

[31] Robert W Zwanzig. Memory effects in irreversible thermodynamics. *Phys. Rev.*, 124(4):983–992, 1961.

[32] Robert W Zwanzig. *Nonequilibrium statistical mechanics*. 2001.

APPENDIX

Derivation of the committor from rate matrices

From the rate matrix $\mathbf{K}$ we can compute transition matrices for arbitrary timescales by

$$\mathbf{T}(\tau) = \exp(\tau \mathbf{K})$$

and from the Taylor expansion of the matrix exponential

$$\mathbf{T}(\tau) = \sum_{n=0}^{\infty} \frac{1}{n!} (\tau \mathbf{K})^n$$

it is easy to see that $\mathbf{T}(\tau)$ and $\mathbf{K}$ have the same eigenvectors. That means the committor is independent of the lagtime chosen and is a static property of the dynamic of our system. This is consistent with the fact that the hitting times h scales linearly with the lagtime τ but not the relative probability $\mathbb{P}_i(h^{\mathcal{B}} < h^{\mathcal{A}})$ that defines the committor. Thus, the lagtime can be chosen arbitrarily and in the limit of vanishing lagtimes only the linear term in the expansion of $\mathbf{T}(\tau)$ survives

$$\mathbf{T}(\tau) = \mathbb{1} + \tau \mathbf{K} + \mathcal{O}\left(\tau^2\right)$$

which we use in Eq. (10) yielding

$$
\begin{aligned}
q_i^+ &= \sum_j \left(\delta_{ij} + \tau K_{ij} + \mathcal{O}\left(\tau^2\right)\right) q_j^+ \\
q_i^+ &= \sum_j \delta_{ij} q_j^+ + \tau \sum_j K_{ij} q_j^+ + \mathcal{O}\left(\tau^2\right) \\
q_i^+ &= q_i^+ + \tau \sum_j K_{ij} q_j^+ + \mathcal{O}\left(\tau^2\right) \\
0 &= \sum_j K_{ij} q_j^+ + \mathcal{O}\left(\tau\right)
\end{aligned}
$$

for all states $i \notin \mathcal{A} \cup \mathcal{B}$.

Derivation of the Committor Covariance

To derive the committor covariance we start with the linear error propagation for the committor and use the sensitivity $\mathbf{S}$, given in Eq. (33), to extend the error in the transition matrix Σ as follows:

$$
\begin{aligned}
\mathrm{Cov}\left(q_a, q_d\right) &= \sum_{i,b,c=1}^{m} S_{ab}^i \Sigma_{bc}^i \left(S^T\right)_{cd}^i \\
&= \sum_{i,b,c=1}^{m} \frac{\partial \tilde{q}_a}{\partial T_{ib}} \Sigma_{bc}^i \frac{\partial \tilde{q}_d}{\partial T_{ic}} \\
&= \sum_{i,b,c=1}^{m} \left(\tilde{\mathbf{A}}^{-1}\right)_{ai} q_b \Sigma_{bc}^i \left(\tilde{\mathbf{A}}^{-1}\right)_{di} q_c \\
&= \sum_{i=1}^{m} \left(\tilde{\mathbf{A}}^{-1}\right)_{ai} \left(\tilde{\mathbf{A}}^{-1}\right)_{di} \sum_{b,c=1}^{m} q_b \Sigma_{bc}^i q_c.
\end{aligned}
$$

We then insert the analytical expression for the uncertainty in the transition matrix in Eq. (35) to obtain

$$
\mathrm{Cov}\left(q_a, q_a\right) = \sum_{i=1}^{m} \left(\tilde{\mathbf{A}}^{-1}\right)_{ai} \left(\tilde{\mathbf{A}}^{-1}\right)_{di} \sum_{b,c=1}^{m} q_b \frac{\alpha_{ib}\left(\alpha_i \delta_{bc} - \alpha_{ic}\right)}{\alpha_i^2(\alpha_i + 1)} q_c.
$$

This can be rewritten in a form that is quadratic in the number of states

$$
\mathrm{Cov}\left(q_a, q_a\right) = \sum_{i=1}^{m} \frac{1}{\alpha_i^2(\alpha_i + 1)} \left(\tilde{\mathbf{A}}^{-1}\right)_{ai}^2 \left(\alpha_i \sum_{b=1}^{m} q_b \alpha_{ib} q_b - \left(\sum_{b=1}^{m} q_b \alpha_{ib}\right)\left(\sum_{c=1}^{m} \alpha_{ic} q_c\right)\right)
$$

leaving us with the inversion of $\mathbf{A}$ as the most expensive operation of cubic order.