3 Markov Chains and Metastability

Stochastic processes change in a random way over time. A Markov chain describes a stochastic process where transitions between states are governed by probability distributions. More formally, a Markov chain is a sequence of random variables $\mathcal{X} = X^{(1)}, X^{(2)}, \ldots$ that are dependent on each other by the "Markov property". The Markov property implies a simple form of dependence which often is described as: "the future depends on the past only through the present."[1] Characterizing properties of Markov chains is done in two directions:

1. Probability theory provides limit theorems explaining the average behavior of a single realization in terms of sums of random variables (e.g., the limit behavior of $1/n \sum_{k=1}^{n} g(X^{(k)})$ of an observable g).

2. Linear algebra and functional analysis (for a finite and continuous state space, respectively) characterizes the global behavior of a Markov chain in terms of the distributions of $X^{(n)}$ (e.g., the limit distribution of $X^{(n)}$ for $n \to \infty$).

Both characterizations complement each other, and together they provide a coherent picture of the stochastic nature of Markov chains.

We first give a short outline of the classical limit theorems like the *law of large numbers* (LLN) and the *central limit theorem* (CLT) for independent random variables in Sect. 3.1, which are then applied to the static *Monte Carlo method* where all random variables are independent of one another. Next, we review the classical Frobenius-Perron-theory and state versions of LLN and CLT for the Markov chain case.

We then connect the phenomenon of metastability to dominant eigenvalues of the transition matrix and present a recently proposed idea of how to identify metastable sets which builds upon the structure of dominant eigenvectors (Sects. 3.2.2 and 3.2.3). In this context we shortly review the stochastic complementation technique due to Meyer, which is an approach for a fast computation of the stationary distribution of non-reversible finite Markov chains [91, 92].

A Markov chain on a continuous state space shows in many aspects the same behavior as on a finite state space. Yet, the associated Markov operator possesses a richer spectral structure (Sect. 3.3). We furthermore discuss in Sect. 3.3.3 ways to discretize a Markov operator on a high-dimensional continuous state spaces, which is a prerequisite for identifying its metastable sets.

[1] Unless stated otherwise, we use the term Markov chain to refer to a discrete time-homogeneous Markov chain on either a finite or continuous state space.

3.1 Classical Limit Theorems

To set the notation, let $(\Omega, \mathcal{F}, \nu)$ be a probability space where Ω is a nonempty set, $\mathcal{F}$ a σ-field of subsets of Ω, and ν a probability measure $\nu : \mathcal{B} \to \mathbb{R}$.

A random variable X is a measurable function $X : \Omega \to \mathbb{R}$. Its expectation value is given by

$$\mathbb{E}_\nu(X) = \int_\Omega X(\omega)\, d\nu.$$

In case ν is the Lebesgue measure on $\Omega \subseteq \mathbb{R}^d$, we simply write $\mathbb{E}(X)$. More generally, if the quantity $\mathbb{E}(X^k)$ exists, it is called the kth *moment* of X. The second moment of $X - \mathbb{E}(X)$ is called the *variance* of X and usually denoted by σ^2, i.e.,

$$\mathrm{Var}(X) = \sigma^2 = \mathbb{E}[(X - \mathbb{E}(X))^2].$$

The square root σ of the variance is the *standard deviation* of X.

Random variables are *independent* if for each n and for all measurable subsets $A_1, \ldots, A_n \in \mathbb{R}$

$$\nu\left(\bigcap_{k=1}^{n}\left(X^{(k)} \in A_k\right)\right) = \prod_{k=1}^{n} \nu\left(X^{(k)} \in A_k\right). \tag{13}$$

A sequence $(X^{(k)})_{k\in\mathbb{N}}$ of random variables is said to be independent when all finite subcollections satisfy condition (13).

Limit theorems are often stated in terms of sequences of independent random variables (see, e.g., the textbooks of Billingsley [13] or Feller [35]).

The general form of the *strong law of large numbers* (LLN) ([82], Chapt. 2, Sect. 9) states an almost sure convergence of the average sum $1/n \sum_{k=1}^{n}(X^{(k)} - \mu_k)$ of a series of independent random variables towards zero:

Theorem 1 (Strong Law of Large Numbers (LLN))
Let $X^{(1)}, X^{(2)}, \ldots$ be independent random variables with means μ_n and variances σ_n^2. Suppose that $\sum \sigma_n^2/n^2 < \infty$. Then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}(X^{(k)} - \mu_k) = 0\right) = 1.$$

In addition, for random variables with a common distribution the following version of the *central limit theorem* (CLT) ([82], Chapt. 3, Sect. 16) states that the limit distribution converges to a normal distribution, which is given by

$$\mathcal{N}_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp[-u^2/(2\sigma^2)]\, du.$$

Theorem 2 (Central Limit Theorem (CLT))
Suppose that $X^{(1)}, X^{(2)}, \ldots$ are independent random variables with a common distribution having mean $\mu = \mathbb{E}(X^{(k)})$ and (finite) variance σ^2, and let $S_n = X^{(1)} + \cdots + X^{(n)}$ denote their partial sum. Then, with

$$F_n(x) = \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right),$$

$$\lim_{n\to\infty} F_n(x) = \mathcal{N}_1(x).$$

With $\hat{\mu}_n = 1/n \sum_{k=1}^{n} X^{(k)}$, we can write

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma}(\hat{\mu}_n - \mu),$$

which lead us to an often used short form of the CLT, namely

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}_\sigma, \tag{14}$$

where $\mathcal{D}$ denotes distributional convergence. From (14) we obtain an $\mathcal{O}(1/\sqrt{n})$-convergence towards the expectation value μ, which, of course, has to be interpreted in terms of a distributional convergence.

Yet, the CLT does not provide a convergence rate towards the normal distribution. For this task, under the stronger assumption that all third moments of the random variables are finite, a stronger version of Theorem 2 can be stated ([82], Chapt. 3, Sect. 16):

Theorem 3 (Berry-Esséen Theorem)
Let $X^{(1)}, X^{(2)}, \ldots$ be a sequence of independent random variables with the same distribution, and suppose that distribution has mean 0, variance $\sigma^2 > 0$, and finite third moment $M = E(|X^{(k)}|^3)$. Then there exists a constant $C \leq 3$ such that for every $x \in (-\infty, \infty)$,

$$\left|\mathbb{P}\left(\frac{S_n}{\sqrt{n}} \leq x\right) - \mathcal{N}_\sigma(x)\right| \leq \frac{CM}{\sigma^3\sqrt{n}}. \tag{15}$$

Since for fixed $X^{(1)}$, the values of M and σ^3 are constant, the Berry-Esséen theorem also states a convergence rate of $\mathcal{O}(1/\sqrt{n})$, which denotes how fast the distributions F_n converge towards the normal distribution $\mathcal{N}_\sigma$.

Monte Carlo Integration. Let $g : \Omega \to \mathbb{R}$ be a Lebesgue integrable function on a compact subset $\Omega \subset \mathbb{R}^d$ of arbitrary dimension d. In a variety of settings one needs to approximate the Lebesgue integral

$$\mathbb{E}(g) = \int_\Omega g(x)\, dx, \tag{16}$$

which we can write as an expectation value $\mathbb{E}(g)$ by interpreting g as a random variable on $(\Omega, \mathcal{B}(\Omega), \mu)$, where μ denotes the normalized Lebesgue-measure. For low-dimensional $\Omega \subseteq \mathbb{R}^d$ many good deterministic quadrature methods are available [29], yet the curse of dimension prevents application for the case $d \gg 1$. For such high-dimensional problems Monte Carlo integration is the method of choice [34, 44]. By generating a series of random vectors $x^{(1)}, x^{(2)}, \ldots$ from independent distributed uniform random variables $X^{(1)}, X^{(2)}, \ldots$ on Ω, we estimate $\mathbb{E}(g)$ by averages $\hat{I}_n = \sum_{k=1}^{n} g(x^{(k)})$. A direct application of the LLN and CLT guarantees an $\mathcal{O}(1/\sqrt{n})$ convergence rate, which due to the probabilistic setting shows no explicit dependence on the dimension d.

Random Numbers. Monte Carlo integration relies on the availability of random numbers and hence we need some source of randomness. In practice, randomness is introduced in the form of pseudo-random number generators [1, 79, 131]. Pseudo-random numbers are not truly random, but rather computed from a deterministic sequence or simply taken from a precalculated list.

Good algorithms generate pseudo-random numbers that share important statistical properties of ideal random numbers, which makes them behave like ideal random numbers and thus suitable for Monte Carlo simulations. A pseudo-random number generator enables an easy and fast access to a source of randomness, which due to its deterministic nature can be replayed for use in several simulations.

Pseudo-random numbers are typically generated in two steps:

1. generation of random numbers uniformly distributed on $[0, 1)$,
2. transformation into random numbers distributed according to the target distribution.

One of the most popular algorithms for generating random numbers is the *linear congruential method*. Its recursion is given by

$$v_0 = \text{seed}, \quad v_{k+1} = (av_k + b) \bmod c$$

for natural numbers a, b, c, and a seed-value in $\{0, 1, \ldots, c-1\}$. The sequence (v_k) is then transformed into a sequence of pseudo-random numbers in $[0, 1)$ by

$$u_k = \frac{v_k}{c}.$$

A proper choice of a, b and c based on algebra and statistical tests are necessary to produce reliable pseudo-random numbers. Among others, recommended values are $a = 16807$, $b = 0$, and $c = 2^{31} - 1$ [101].

To accomplish the second step, transformation rules from uniform distributed random numbers into other simple distributions do exist. For example, Box and Muller proposed a transformation rule that produces random numbers distributed according to the normal distribution [14]. Yet, if the target distribution is more complicated and high-dimensional, a much more sophisticated approach is needed (e.g., *coupling from the past* proposed by Propp and Wilson [39, 130]).

Thermodynamical Integrals. In equilibrium statistical physics the computation of thermodynamical quantities of a system is computed via an integral with respect to a stationary distribution f (e.g., the canonical distribution $f_{\mathcal{V}}$ as in (8)). A thermodynamical integral takes the form

$$\mathbb{E}_f(g) = \int_\Omega g(x)f(x)\,dx = \frac{1}{Z_h}\int_\Omega g(x)h(x)\,dx, \quad \text{where} \quad Z_h = \int_\Omega h(x)\,dx \tag{17}$$

where the observable g and the unnormalized density h can be computed easily. Yet the integral representation of the normalizing constant Z_h prevents a direct application of the Monte Carlo method to the computation of (17). Much worse, even Z_h is not tractable by direct Monte Carlo due to a large variance of f. In fact, the computation of normalizing constants would solve many problems in statistical physics but remains up to now (except of some simple systems) a big challenge for physicists.

Markov chain Monte Carlo (MCMC) provides a way out of this dilemma. Instead of a series of independent uniformly distributed random variables, MCMC circumvents the problem of computing Z_h by producing a dependent series of random variables distributed according to f. The density f is evaluated only via quotients of the form $f(x)/f(y)$, where Z_h cancels out. MCMC can be seen as an extension of the direct Monte Carlo method which is based on Markov chain theory.

3.2 Markov Chains on Finite State Spaces

For Markov chains the dependence on the past is passed on only through the present state, which, as will be seen in the following, leads to a rich mathematical structure. We first restrict to the case of discrete time homogenous Markov chains [16, 19, 76, 99] on a finite state space $\Omega = \{1, \dots, m\}$. Let $\mathcal{X} = X^{(1)}, X^{(2)}, \dots$ be a sequence of random variables with values in Ω. The sequence $\mathcal{X}$ together with a probability vector $\pi^{(1)} \in \mathbb{R}^m$ is called a discrete time Markov chain on the state space Ω, if for every $k \in \mathbb{N}$ and $j, i_{k-1}, \dots, i_1 \in \Omega$ the condition

$$\mathbb{P}(X^{(k)} = j \,|\, X^{(k-1)} = i_{k-1}, \dots, X^{(1)} = i_1) = \mathbb{P}(X^{(k)} = j \,|\, X^{(k-1)} = i_{k-1}) \tag{18}$$

is satisfied, and $\mathbb{P}(X^{(1)} = i) = \pi_i^{(1)}$ for $i \in \{i_1, \ldots, i_m\}$. Equation (18) is called the *Markov property* and $\pi^{(1)}$ the *initial distribution*.

If in addition the transition probabilities are independent of k, i.e., for all $i, j \in \Omega$

$$\mathbb{P}(X^{(k)} = j \,|\, X^{(k-1)} = i) = t_{ij} \tag{19}$$

holds, the Markov chain is called *homogeneous*.

In the case of a homogenous Markov chain the matrix $T = (t_{ij})$ given by the transition probabilities t_{ij} forms a stochastic matrix, i.e.,

$$T \geq 0 \quad \text{and} \quad \sum_{j=1}^{m} t_{ij} = 1 \quad \text{for all } i,$$

and the initial distribution of $X^{(1)}$ is given by a probability distribution

$$\pi^{(1)} \quad \text{where} \quad \pi^{(1)} \geq 0 \quad \text{and} \quad \sum_{j=1}^{m} \pi_j^{(1)} = 1.$$

Each homogeneous Markov chain is associated to a transition matrix and vice versa. This relationship allows many properties of homogeneous Markov chains to be expressed in terms of linear algebra.

In the following we state some basic facts about Markov chains (for further background we refer to [16] and the introduction given in [10], Chapter 8). The distribution of a Markov chain is determined by its initial distribution and its transition matrix. The distribution of $X^{(k)}$ denoted by $\pi^{(k)}$ is given by

$$\pi_i^{(k)} = \mathbb{P}(X^{(k)} = i).$$

With the *n-step transition matrix* T^k we can compute $\pi^{(k)}$ via the relation

$$(\pi^{(k)})' = (\pi^{(1)})' \, T^{k-1}.$$

Often the Markov chain possesses some additional structure. Two important notions expressed in terms of its associated transition matrix are *irreducibility* and *aperiodicity*:

- A nonnegative matrix T is called *irreducible*, if for all $i, j \in \{1, \ldots, m\}$ there exists a $k \in \mathbb{N}$ with $(T^k)_{ij} > 0$.

- A nonnegative matrix T is called *aperiodic*, if for all $i \in \{1, \ldots, m\}$

  $$\gcd\left(k \in \mathbb{N} \;\mid\; (T^k)_{ii} > 0\right) = 1,$$

  where gcd denotes the greatest common divisor.

A matrix is called *primitive* if there exists a $k \in \mathbb{N}$ with $T^k > 0$. A matrix T is primitive if and only if it is irreducible and aperiodic.

We call π an invariant distribution of $\mathcal{X}$, if its associated transition matrix satisfies

$$\pi' = \pi' T. \tag{20}$$

Every Markov chain on a finite state space has at least one invariant distribution. If $\pi^{(1)}$ equals an invariant distribution, then $\pi^{(k)} \equiv \pi^{(1)}$ for all $k \in \mathbb{N}$, and we say that $\mathcal{X}$ is started in stationarity.

In general, the Perron-Frobenius theory investigates spectral properties of nonnegative matrices [76, 116], hence including stochastic matrices as a special case. The following theorem summarizes some essential facts of the Perron-Frobenius theory ([16], Chapt. 6.1).

Theorem 4 (Perron-Frobenius) *Let T be a nonnegative primitive $r \times r$ matrix. There exists a real eigenvalue λ_1 with algebraic as well as geometric multiplicity one such that $\lambda_1 > 0$ and $\lambda_1 > |\lambda_j|$ for any other eigenvalue. Moreover, the left eigenvector u_1 and the right eigenvector v_1 associated with λ_1 can be chosen positive and such that $u_1' v_1 = 1$.*

Let $\lambda_2, \lambda_3, \ldots, \lambda_r$ be the eigenvalues of T other than λ_1 ordered in such a way that

$$\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_r|$$

and if $|\lambda_2| = |\lambda_j|$ for some $j \geq 3$, then $m_2 \geq m_j$, where m_j is the algebraic multiplicity of λ_j. Then

$$T^n = \lambda_1^n v_1 u_1' + \mathcal{O}\left(n^{m_2-1}|\lambda_2|^n\right). \tag{21}$$

If $0 \leq B \leq T$ and β is an eigenvalue of B, then $|\beta| \leq \lambda_1$. Moreover, $|\beta| = \lambda_1$ implies $B = T$.

If in addition, T is stochastic, then $\lambda_1 = 1$.

In the case of a primitive stochastic matrix T we have $\lambda_1 = 1$ with a right eigenvector $v_1 \equiv \mathbf{1}$ where $\mathbf{1} = (1, \ldots, 1)$, and the stationary distribution π is identical to the left eigenvector u_1. Then, Eq. (21) simplifies to

$$T^k = e\,\pi' + \mathcal{O}\left(n^{m_2-1}\Lambda^n\right), \quad \Lambda = |\lambda_2|, \tag{22}$$

from which follows

$$\lim_{k\to\infty} (\pi^{(1)})' T^{k-1} = \lim_{k\to\infty} (\pi^{(k)})' = \pi'$$

for any initial distribution $\pi^{(1)}$ with convergence rate Λ.

For a sample path $\mathbf{x}$ of the associated irreducible Markov chain $\mathcal{X}$ this means that (independent of the initial state) for k being large enough, $x^{(k)}$ will be approximately drawn from π. In practice, this process is often denoted as convergence towards equilibrium or "burn-in" phase.

Reversible Markov Chains. Given a transition matrix T and a probability vector π the pair (T, π) is said to be *reversible*, if

$$\pi_i \, t_{ij} = \pi_j \, t_{ji} \quad \text{for all} \quad i, j. \tag{23}$$

This implies that π is an invariant distribution of T. If, in addition, T is irreducible, then π is the unique invariant distribution, and we simply say that T is reversible. In terms of some *weighting matrix* $D = \mathrm{diag}(\sqrt{\pi_i})$ we can write Eq. (23) as $D^2 T = T' D^2$, and we may introduce the π-weighted inner product

$$\langle x, y \rangle_\pi = x' \, D^2 \, y. \tag{24}$$

It is easy to see that a reversible matrix T is symmetric wrt. the inner product $\langle \cdot, \cdot \rangle_\pi$, since we immediately have $\langle x, Ty \rangle_\pi = x' D^2 T y = x' T' D^2 y = \langle Tx, y \rangle_\pi$.

As we have seen from Theorem 4, a wealth of information can be said about the spectrum of an irreducible stochastic matrix T. If, in addition, the Markov chain is reversible, even more structural properties hold (see [16, 30]):

1. There exists a basis of π-orthogonal right eigenvectors, which diagonalizes T.

2. For every right eigenvector u there is an associated left eigenvector $v = D^2 u$, which corresponds to the same eigenvalue.

3. All eigenvalues are real and contained in the interval $(-1, 1]$.

4. T is similar to the symmetric, in general non-stochastic matrix $T_{\mathrm{sym}} = DTD^{-1}$.

We will see in Sect. 4.1 that the detailed balance equation (23) is the key principle to construct transition matrices that have π as its invariant distribution.

3.2.1 Pathwise Limit Theorems

We now turn our focus to the limit behavior of sums of random variables, which is the typical objective in a MCMC setting.

For an observable $g : \Omega \to \mathbb{R}$ with $\Omega = \{1, \ldots, m\}$ the expectation value of g is given by

$$\mathbb{E}_\pi(g) := \sum_{i=1}^{m} g(i)\pi(i).$$

Essential for the whole concept is that the *law of large numbers* and the *central limit theorem* also hold for Markov chains. The strong version of LLN (also denoted as ergodic theorem for Markov chains) now states [93, 99]:

Theorem 5 (Law of Large Numbers for Markov Chains)
Let $X^{(1)}, X^{(2)}, \ldots$ denote an irreducible and aperiodic Markov chain with finite state space Ω and unique invariant distribution π. Furthermore let $g: \Omega \to \mathbb{R}$ be a real-valued random variable on Ω. Then the convergence

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g(X^{(k)}) = \mathbb{E}_\pi(g)$$

happens almost surely.

Almost sure convergence guarantees that it is sufficient to generate one sufficiently long sample of the Markov chain to compute expectation values. Again, for convergence properties we have to resort to the CLT [77, 93]:

Theorem 6 (Central Limit Theorem for Markov Chains)
Let $X^{(1)}, X^{(2)}, \ldots$ denote an irreducible and aperiodic Markov chain with finite state space Ω and unique invariant distribution π. Moreover, let an observable $g: \Omega \to \mathbb{R}$ be given and denote the associated expectation value by $\mu = \mathbb{E}_\pi(g)$, and the associated partial sums by $S_n = g(X^{(1)}) + \cdots + g(X^{(n)})$. Then, the so-called asymptotic variance *σ_a^2 satifies*

$$\sigma_a^2 := \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{k=1}^{n} g(X^{(k)}) \right) < \infty, \tag{25}$$

and the distribution

$$F_n(x) = \mathbb{P}\left(\frac{S_n - n\mu}{\sigma_a \sqrt{n}} \leq x \right)$$

converges to a standard normal distribution with mean 0 for $n \to \infty$:

$$\lim_{n\to\infty} F_n(x) = \mathcal{N}_1(x).$$

With $\hat{\mu}_n = 1/n \sum_{k=1}^{n} g(X^{(k)})$, we thus again get the short form of the CLT, namely

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}_{\sigma_a}. \tag{26}$$

There also exist Berry-Esséen estimates for Markov chains in the same spirit as in Theorem 3 for the independent case [85]. But in contrast to the independent case, the dependence between the random variables of the Markov chain leads to a dramatic increase of the constant C, which makes the bound achieved so far still not applicable for practical purposes.

The CLT characterizes the long run behavior by the asymptotic variance σ_a. For a reversible primitive Markov chain with $m \times m$ transition matrix T, m eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_m$ with associated *left* eigenvectors

v_k, $k = 1, \ldots, m$, and stationary distribution π, the asymptotic variance for an observable g can be expressed in terms of its spectral decomposition [16]:

$$\sigma_a = \sigma_a(g, T, \pi) = \sum_{i=2}^{n} \frac{1 + \lambda_i}{1 - \lambda_i} |\langle g, v_i \rangle_\pi|^2 \tag{27}$$

We can see from this equation that measuring the quality of an estimate by means of the asymptotic variance includes all eigenvalues. However, the asymptotic variance can become extremely large whenever there is at least one eigenvalue close to $\lambda_1 = 1$. Negative eigenvalues, even those close to -1, are helpful in that they reduce the asymptotic variance. In Sect. 3.3.2, a more general version of (27) is given.

The CLT again seems to indicate that the rate of convergence along single pathwise realizations is of the form $C/\sqrt{n}$ where the constant C may become large with large asymptotic variance; since $C \propto \sigma_a$ it may become large if there is at least one eigenvalue close to $\lambda_1 = 1$.

The CLT is closely related to the following *large deviation result*:

$$\mathbb{P}\left(|\frac{1}{n} S_n \, - \, \mathbb{E}_\pi(g)| > \epsilon \right) \, \leq \, C \exp\left(-\frac{\epsilon}{\sigma_a^2} \, n \right), \tag{28}$$

that holds asymptotically for $n \to \infty$ for sufficiently small ϵ and a constant C that does not depend on n (for a detailed discussion of the conditions under which this statement may be valid see [80]). This shows that the probability to observe an error larger than ϵ is decreasing exponentially fast with n with decay rate ϵ/σ_a^2.

Summary. Distributional convergence describes the evolution of the initial distribution of a primitive Markov chain to the stationary distribution π as described by the Perron-Frobenius Theorem. Large positive as well as negative eigenvalues prevent the initial distribution to become almost stationary after a few iterations. This behavior is characterized in Eq. (22) by Λ. If, however, we are interested in pathwise convergence of expectation values, we have to resort to the CLT or large deviation results, (26) and (28), respectively. Therein, convergence is mainly influenced by the asymptotic variance σ_a, a constant which might be extremely large if the spectral gap $1 - \Lambda$ is significantly small.

3.2.2 Metastable Sets

If λ_2 is close to $\lambda_1 = 1$, we often find that the reason for the undesirably slow convergence is that the Markov chain remains for a long time in a *metastable* region (also called *mode* or *conformation*) of the phase space, before it moves on to another one. Such metastable behavior can be analyzed via the concept of *almost invariant sets* [24, 113].

We herein will exploit the following observation concerning metastability: If there are n eigenvalues close to $\lambda_1 = 1$ (including λ_1 itself) and a significant spectral gap to all remaining eigenvalues, then there also are n disjoint metastable sets and vice versa [91, 115]. If this is the case, the chain is rapidly mixing *within* the corresponding metastable subsets and the undesirably slow overall convergence results from the rareness of transitions between these metastable sets. This behavior is illustrated in Fig. 8, where entries in the 23×23 matrix T reflect spatial transitions probabilities in the torsion angle of n-butane (see Fig. 4).

The close connection between a separated cluster of dominant eigenvalues and the existence of metastable subsets has another very important algorithmic consequence: it has been shown for reversible Markov chains that one can identify the n metastable subsets only on basis of the *eigenvectors* associated with the n dominant eigenvalues [113, 115]. This insight leads to a significantly general identification algorithm [30] used for the detection of biomolecular conformations.

For sets $A, B \subseteq \Omega$ the transition probability between A and B is given by

$$\kappa(A, B) = \frac{\sum_{i \in A} \left(\pi_i \sum_{j \in B} t_{ij} \right)}{\sum_{i \in A} \pi_i}, \tag{29}$$

which can be interpreted as the probability to move from the set A to the set B wrt. T. With this definition, clearly $\kappa(\Omega, \Omega) = 1$ (i.e., all probability remains in the state space Ω). We will denote a set A as metastable wrt. T, if the transition probability from A to itself is close to one, i.e., if $\kappa(A, A) \approx 1$. Furthermore, we denote by $\mathcal{A} = \{A_1, \ldots, A_d\}$ a metastable decomposition of Ω wrt. T, if $\kappa(A_k, A_k) \approx 1$ for all $k = 1, \ldots, d$.

How to measure the quality of a metastable decomposition will depend on the specific application. For example, one could aim at maximizing the average sum

$$\max_{\mathcal{A}} \frac{1}{d} \sum_{k=1}^{d} \kappa(A_k, A_k).$$

Remark. For general non-reversible Markov chains, eigenvalues anywhere near the unit circle correspond to an *almost cyclic behavior* [24]. We will not pursue the existence of almost cyclic behavior further; we are only concerned about reversible Markov chains, and in that case the typical spectrum we will have to deal with in the following consists of a well-separated cluster of eigenvalues near 1 and no eigenvalues near -1.

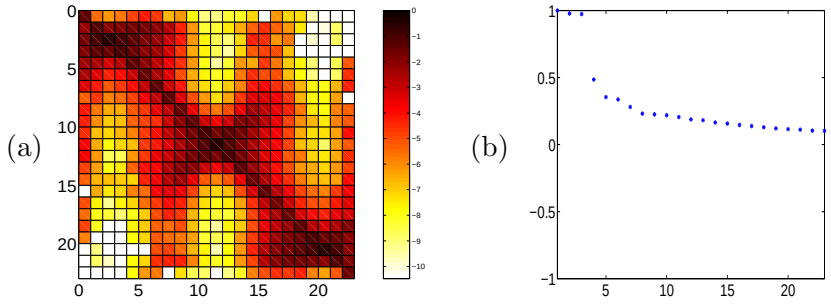

Figure 8: Discretized Markov operator of n-butane at $T = 300\,\mathrm{K}$ (details of discretization are discussed in Sect. 3.3.3). (a) Illustration of the entries of the transition matrix T. Intensity of entries due to logarithmic scale. (b) Ordered spectrum of T with a cluster of three eigenvalues close to $\lambda_1 = 1$ and a significant gap to all remaining ones.

3.2.3 Identification of Metastable Sets

Although the block dominant structure in T illustrates the corresponding metastable sets, the states belonging to a metastable set are in general not ordered (one could think of analyzing a stochastic matrix with randomly permuted states). We will present here only the main characteristics of the algorithm; a detailed description together with its motivation emerging from a perturbation analysis for the block dominant structure can be found in [30].

In the *unperturbed* case one would have a k-fold Perron-eigenvalue 1, and corresponding right eigenvectors are *constant* on their blocks. By interpreting a reversible irreducible matrix as the *perturbed* case it follows from perturbation theory that the k-fold eigenvalue 1 transforms into a simple eigenvalue 1 and $k-1$ eigenvalues close to 1 (which is also referred to as the Perron cluster); right eigenvectors now have to be *almost constant* on metastable sets, and are pairwise orthogonal.

The key algorithmic idea is to identify metastable sets via the *sign structure* of the right eigenvectors $u_1, \ldots, u_k$ corresponding to the dominant k eigenvalues. The sign structure is defined as

$$s_j = (\mathrm{sign}((u_1)_j), \ldots, \mathrm{sign}((u_k)_j)), \quad \text{for } j = 1, \ldots, m$$

Our aim is to find a map $a : \Omega \to \{1, \ldots, k\}$, where $\Omega = \{1, \ldots, m\}$ and $k \leq m$ (usually $k \ll m$), which assigns each state to its metastable set.

If the number of sign structures equals k, all states having the same sign structure form a metastable set and we are done. However, in practice we typically face the situation of having more sign structures than k, and our task is to merge sign structures which only differ slightly from each other. Based on the perturbation theory this is achieved by introducing equivalence classes of sign structures wrt. some threshold parameter ε that allows to interpret small entries in the eigenvectors to change their signs; we

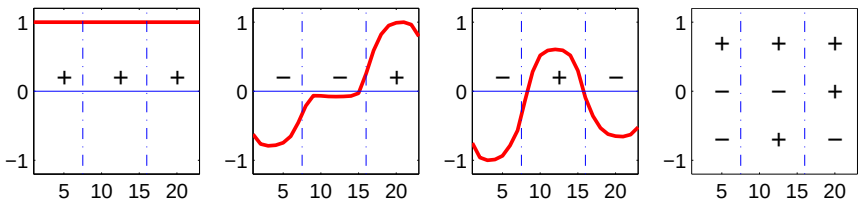

Figure 9: Subfigure 1–3: Right eigenvectors of the three dominant eigenvalues of the transition matrix T shown in Fig. 8. Subfigure 4: The three metastable subsets are characterized by three different sign combinations of these eigenvectors.

identify the smallest ε for which exactly k equivalence classes exist, which then again uniquely define the states for each of the k metastable sets.

Figure 9 illustrates the key idea of the algorithm for identifying metastable sets via these eigenvectors: For each state $j = 1, \ldots, 23$, we denote by $s_j \in \{+, -\}^3$ the 3-tuple of signs of the jth components in each of the three eigenvectors, the so-called sign combinations. The forth subfigure in Fig. 9 shows that there are only three different sign combinations, and that all states j with the same s_j belong to the same metastable set (in this simple example we do not face the more complicated case where we have more sign structures than metastable sets). Thus, metastable sets can be identified as sets of states with identical sign combination.

Alternative Approaches. We briefly want to mention some alternative methods for identifying metastable sets. Instead of using the sign-structure of dominant eigenvectors it is also possible to perform a fuzzy decomposition of the set $\mathcal{S}$ [129]. In terms of graph theory, our problem can be stated as an edge-weighted graph $\mathcal{G}(V, K)$, where a vertice v_i corresponds to a state s_i, and an edge k_{ij} is assigned the weight $\pi_i t_{ij}$ with T being the transition matrix. Done that, algorithms for graph partitioning can be applied [103]. These algorithms can be roughly divided into two groups: The first group, the so-called *greedy* algorithms, try to find a good decomposition into a given number of k sets of vertices by letting grow k initial empty sets and afterwards rearranging these sets by a given optimization criterion or cost function. The second group, often denoted as *spectral graph partitioners*, computes eigenvectors of the corresponding Laplacian matrix and tries to extract information from the "important" eigenvectors by a geometrical clustering of *eigenvector data* [3, 119]. In fact these kind of algorithms possess a structural similarity to our identification algorithm sketched above. More recently, it was also suggested to use the *congestion* of a graph, a notion which refers to a quantity specifying bottlenecks in a graph [25].

Stochastic Complementation. The connection between eigenvalues close to 1 and metastable sets is also described by Meyer [91] in the context of

a given (not necessarily reversible) irreducible stochastic matrix T. Meyer introduces the concept of *stochastic complementation* in order to address the problem of determining the stationary distribution of an irreducible Markov chain T with a large number of states by uncoupling it into several smaller independent Markov chains. This can be thought of as the inverse problem that we described so far, namely identifying metastable sets of a given stationary distribution via spectral properties of a reversible transition matrix T.

In [91], a coupling matrix $C = (c_{ij})$ is constructed by aggregation of states in such a way that the stationary distribution π of C contains the correct weighting factors for the aggregates. This is achieved by using global information to set up the *stochastic complements* c_{ij}. The coupling matrix C can be interpreted as a coarsed grained stochastic version of T. The application in mind by stochastic complementation are *aggregation-disaggregation* techniques [22, 118], which perform a fast computation of the stationary distribution of T for large finite state spaces, where a suitable decomposition of the state space is known in advance. Like UC, stochastic complementation makes use of coupling factors, although in a different context.

3.3 Markov Chains on Continuous State Spaces

Markov chains are also considered on *general state spaces* where Ω is an arbitrary topological space [93]. For our purpose we restrict to the case of a continuous state space $\Omega \subseteq \mathbb{R}^d$. All Markov chains are still considered as a discrete time stochastic process.

We will see that versions of the limit theorems from Sect. 3.2.1 also exist for continuous state spaces [32, 93, 107]. Especially, under some further assumptions resulting from the richer structure of continuous state spaces, we can draw analogous conclusions between the spectral gap and convergence rate as we have done in the finite case.

3.3.1 Transition Kernel

To set the notation, let $(\Omega, \mathcal{B}, \lambda)$ be the underlying measure space and π a probability measure on $(\Omega, \mathcal{B})$. We suppose in the following, that λ is the Lebesgue measure on $\Omega \subseteq \mathbb{R}^d$, and that π possesses a density

$$f(x)\, dx = \pi(dx)$$

with $f > 0$ where dx denotes integration wrt. the Lebesgue measure λ.

A *transition kernel*[2] $K : \Omega \times \mathcal{B} \to [0, 1]$ defines a chain $\mathcal{X} = (X^{(k)})_{k \in \mathbb{N}}$ through the relation

$$\mathbb{P}\{X^{(k+1)} \in A | X^{(k)}, \ldots, X^{(1)}\} = K(X^{(k)}, A)$$

[2]For an exact definition of commonly used terms as *transition kernel*, *irreducibility* or *aperiodicity*, we refer to the monograph [93].

where $K(x, A)$ denotes the probability to move in one step from the point x into the set A.

We call f an *invariant density* of $\mathcal{X}$, if

$$\int_A f(x)\,dx = \int_\Omega K(x, A) f(x)\,dx \tag{30}$$

holds for all $A \in \mathcal{B}$.

The Markov chain is reversible, if the transition kernel satisfies the detailed balance condition:

$$\int_A K(x, B) f(x)\,dx \;=\; \int_B K(y, A) f(y)\,dy, \quad \text{for all } A, B \in \mathcal{B}.$$

If the kernel K possesses a density $k(x, y)$ wrt. the invariant measure π, i.e.,

$$K(x, A) = \int_A k(x, y) f(y)\,dy \quad \text{for all} \quad A \in \mathcal{B}, \tag{31}$$

then this simplifies to

$$k(x, y) \;=\; k(y, x)$$

for every $x, y \in \Omega$.

3.3.2 The Markov Operator

In the following we want to understand the global behavior of a Markov chain with unique invariant density f given by $f(x)\,dx = \pi(dx)$ via the spectral structure of its associated Markov operator on the space L^2_π; the spaces L^s_π, $1 \le s < \infty$, are herein defined via

$$L^s_\pi = \{u : \Omega \to \mathcal{C} \mid \int_\Omega |u(x)|^s f(x)\,dx < \infty\}.$$

L^2_π then is a Hilbert space with the scalar product

$$\langle u, v\rangle_\pi \;=\; \int_\Omega u(x)\overline{v(x)} f(x) dx, \quad \forall u, v \in L^2_\pi.$$

The transition kernel K induces a Markov operator $P : L^2_\pi \to L^2_\pi$, $u \mapsto Pu$, also called *propagator*, that is defined by

$$Pu(y)\, f(y)\,dy = \int_\Omega K(x, dy) u(x)\, f(x)\,dx, \tag{32}$$

where the notation means that

$$\int_A Pu(y)\, f(y)\,dy = \int_\Omega K(x, A) u(x)\, f(x)\,dx, \quad \forall A \in \mathcal{B},$$

and simplifies to

$$Pu(y) = \int_\Omega k(x,y)u(x)\,f(x)\,dx,$$

if the kernel K possess a density $k(x,y)$ wrt. the invariant measure π.

Instead of generating a series of states as one does when computing a realization of the Markov chain $\mathcal{X}$, P describes the propagation of densities.

Two important properties of the so-defined operator are [112, 113]:

(i) P is a Markov operator: for all $u \in L^1_\pi$ we have $\int |Pu(x)|\,dx = \int |u(x)|\,dx$ and from $u \geq 0$ follows $Pu \geq 0$.

(ii) P is a symmetric operator in L^2_π wrt. the scalar product $\langle\cdot,\cdot\rangle_\pi$ due to the reversibility of K.

From (i) and (ii) follows that in L^2_π the spectrum $\sigma(P)$ of P is real and bounded in modulus by 1, so we have $\sigma(P) \subseteq [-1,1]$. Similar to the well-known Frobenius-Perron theorem for finite state spaces, irreducibility and aperiodicity of K implies that P has a simple eigenvalue $\lambda_1 = 1$, for which the constant function $\mathbf{1}$, $\mathbf{1}(x) = 1$, for all $x \in \Omega$ is an eigenfunction, i.e. $P\mathbf{1} = \mathbf{1}$. Thus, by introducing the orthogonal space of the eigenspace of the eigenvalue $\lambda_1 = 1$, i.e.,

$$L^{2,0}_\pi = \{u \in L^2_\pi : \langle u, \mathbf{1}\rangle_\pi = 0\},$$

we can decompose P into two parts via

$$Pu = \langle \mathbf{1}, u\rangle_\pi + P_0 u, \tag{33}$$

where P_0 acts and is self-adjoint on $L^{2,0}_\pi$.

To further investigate the spectral structure we define the discrete spectrum $\sigma_{\text{discr}}(P)$ to consist of all isolated eigenvalues of P with finite multiplicity and the essential spectrum by $\sigma_{\text{ess}}(P) = \{\lambda \in \sigma(P) | \lambda \notin \sigma_{\text{discr}}(P)\}$; the essential spectral radius is given by $r_{\text{ess}}(P) = \sup_{\lambda\in\sigma_{\text{ess}}(P)} |\lambda|$.

If $r_{\text{ess}} < 1$, the Markov chain $\mathcal{X}$ is called *geometrically ergodic* [93], which is a desirable property for the rate of convergence of the MCMC algorithm as we will see next: According to [77, 123] the central limit theorem in its form (26) also holds for Markov chains on continuous state spaces under the additional assumptions that we consider observables $g \in L^2_\pi$. The asymptotic variance σ_a is given by [54, 106]

$$\sigma_a^2 = \lim_{n\to\infty} \frac{1}{n} \text{Var}\left(\sum_{k=1}^n g(X^{(k)})\right) \tag{34}$$

$$= \langle \overline{b}, \overline{b}\rangle_\pi + 2\sum_{k=1} \langle \overline{b}, P^k \overline{b}\rangle_\pi, \tag{35}$$

where $\overline{b} = g - \pi(g) \in L^{2,0}_\pi$ with $\pi(g) = \langle \mathbf{1}, g\rangle_\pi = \int g(x) f(x)\,dx$.

The following theorems given in [77] show how (like in the case of a finite state space) the asymptotic variance can be computed by means of the spectral decomposition of the self-adjoint Markov operator P in L^2_π:

Theorem 7 *Let P be the operator associated with a reversible, irreducible Markov chain $\mathcal{X}$, and let P_0 be defined according to (33). Moreover, let $E_{g,P_0}(\cdot)$ be the spectral measure associated with $g \in L^2_\pi$ and P_0. Then*

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}\left(\sum_{k=1}^{n} g(X^{(k)})\right) = \int_{\sigma(P_0)} \frac{1+\lambda}{1-\lambda} E_{g,P_0}(d\lambda).$$

Theorem 8 *Let P be the operator associated with a reversible, irreducible Markov chain $\mathcal{X}$, and let $\Lambda = \Lambda(P_0) = \sup_{\lambda\in\sigma(P_0)} \lambda$. Then*

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}\left(\sum_{k=1}^{n} g(X^{(k)})\right) \leq \frac{1+\Lambda}{1-\Lambda}\pi(g^2) < \frac{2}{1-\Lambda}\pi(g^2).$$

If we assume that the chain is geometrically ergodic, then the *spectral gap* $1-\Lambda(P_0) > 0$ is closely related to the asymptotic variance. If we furthermore assume that there is a second eigenvalue $\lambda_2 \geq r_{\mathrm{ess}}(P)$ then $1-\Lambda(P_0) = 1-\lambda_2$. Then (due to our discussion of the central limit theorem) $\mathcal{X}$ converges with geometric rate due to λ_2, i.e., with increasing λ_2, we need exponentially increasing sampling length for $\mathcal{X}$ to produce good samplings of the density f. If the above assumption wrt. r_{ess} holds, a Markov Chain $\mathcal{X}$ is therefore called *slowly mixing* if λ_2 is close to 1, and *rapidly mixing* if $\lambda_2 \ll 1$.

The upper bound given in terms of $1-\Lambda(P_0)$ is a universal upper bound for any function $g \in L^2_\pi$. For a particular function, however, it may also happen that $1-\Lambda(P_0)$ has only a slight impact on the asymptotic variance (also see [54]).

3.3.3 Metastable Sets of a Markov Operator

Our aim is to extend the concept of transition probabilities from Eq. (29) and the identification strategy for metastable sets as presented in Sects. 3.2.2 and 3.2.3 to the continuous case.

Metastable Sets. Suppose, that we consider a transition kernel K of the form (31). Then, for two sets $A, B \subseteq \Omega$ the *transition probability* between A and B within an ensemble distributed wrt. the density f and after one step of the Markov chain is given by

$$\kappa(A,B) = \frac{1}{\int_A f(x)\,dx}\int_A\int_B k(x,y)f(x)\,dy\,dx = \frac{\langle \mathbf{1}_B, P\mathbf{1}_A\rangle_\pi}{\langle \mathbf{1}_A, \mathbf{1}_A\rangle_\pi}. \qquad (36)$$

where $\mathbf{1}_A$ denotes the indicator function of some set A, i.e., $\mathbf{1}_A(x) = 1$ if $x \in A$, and $\mathbf{1}_A(x) = 0$ otherwise.

The last formula allows to give a mathematical statement relating dominant eigenvalues, the corresponding eigenfunctions and a decomposition of the state space into metastable subsets. For later reference we define the *metastability of a decomposition* $\mathcal{D} = \{D_1, \ldots, D_m\}$ as the sum of the metastabilities $\kappa(D_i, D_i)$ of its subsets D_i. The next result can be found in [73]; a version for two subsets was published in [71].

Theorem 9 *Let $P : L^2_\pi \to L^2_\pi$ denote the Markov operator of a reversible and geometrically ergodic Markov chain. Then the spectrum of P has the form*

$$\sigma(P) \subset [a, b] \cup \{\lambda_n\} \cup \ldots \cup \{\lambda_2\} \cup \{1\}$$

with $-1 < a \leq b < \lambda_n \leq \ldots \leq \lambda_1 = 1$ and isolated, not necessarily simple eigenvalues of finite multiplicity that are counted according to multiplicity. Assume that $n > 1$ in this representation, i.e., there are at least two isolated dominant eigenvalues with modulus larger than the essential spectral radius. Denote by $v_n, \ldots, v_1$ the corresponding eigenfunctions, normalized to $\|v_k\|_2 = 1$. Let Q be the orthogonal projection of L^2_π onto $\operatorname{span}\{\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}\}$. The metastability of an arbitrary decomposition $\mathcal{D} = \{A_1, \ldots, A_n\}$ of the state space Ω can be bounded from above by

$$\kappa(A_1, A_1) + \ldots + \kappa(A_n, A_n) \leq 1 + \lambda_2 + \ldots + \lambda_n,$$

while it is bounded from below according to

$$1 + \eta_2\lambda_2 + \ldots + \eta_n\lambda_n + c \leq \kappa(A_1, A_1) + \ldots + \kappa(A_n, A_n)$$

where $\eta_j = \|Qv_j\|^2_\pi$ and $c = a\,(1 - \eta_2) \ldots (1 - \eta_n)$.

Theorem 9 highlights the strong relation between a decomposition of the state space into metastable subsets and dominant eigenvalue close to 1 (in almost the same manner as in the case of a finite state space). It states that the metastability of an arbitrary decomposition $\mathcal{D}$ cannot be larger than $1+\lambda_2+\ldots+\lambda_n$, while it is at least $1+\eta_2\lambda_2+\ldots+\eta_n\lambda_n+c$, which is "large" whenever the dominant eigenfunctions $v_2, \ldots, v_n$ are almost constant on the metastable subsets $A_1, \ldots, A_n$ implying $\eta_j \approx 1$ and $c \approx 0$. The term c can be interpreted as a correction that is small, whenever $a \approx 0$ or $\eta_j \approx 1$. It is demonstrated in [73] that the lower and upper bounds are sharp and asymptotically exact.

Discretization. In order to eventually compute the dominant eigenmodes of the Markov operator P given by K we have to discretize the corresponding eigenvalue problem [113, 114]. This can be done by *coarse graining* the Markov chain given by K with an arbitrary box decomposition of the state

space Ω into m disjoint sets $B_1, \ldots, B_m \subset \Omega$ with $\cup B_j = \Omega$. Based on this box decomposition, we introduce the new finite phase space $\{B_1, \ldots, B_m\}$ and define the transition function $\tilde{K}$ on $\{B_1, \ldots, B_m\}$ via

$$\tilde{K}(B_k, B_l) \;=\; \kappa(B_k, B_l). \tag{37}$$

The finite dimensional Markov chain defined by $\tilde{K}$ again is reversible wrt. its stationary density $\tilde{\pi}$ given by $\tilde{\pi}_k = \pi(B_k) = \int_{B_k} f(x)dx$. Whenever f is unique for K, $\tilde{\pi}$ is also unique for $\tilde{K}$.

Since after discretization the state space is finite, P becomes an $m \times m$ transition matrix T which simply is the column stochastic matrix with entries $T_{lk} = \tilde{K}(B_k, B_l) = \kappa(B_k, B_l)$.

The eigenmodes of P (if associated with isolated eigenvalues) can be approximated by eigenmodes of T; self-adjointness and boundedness of P allow to prove convergence in the limit of arbitrary fine box coverings, i.e., $m \to \infty$.

In order to set up T we have to estimate $\kappa(B_k, B_l)$ for arbitrary box numbers k and l from a realization $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})$ of $\mathcal{X}$. Then, the relative frequencies approximate $\kappa(B_k, B_l)$ in the sense that

$$\kappa(B_k, B_l) = \lim_{n\to\infty} \frac{\sum_{j=1}^{n-1} \mathbf{1}_{B_k}(x^{(j)})\; \mathbf{1}_{B_l}(x^{(j+1)})}{\sum_{j=1}^{n-1} \mathbf{1}_{B_k}(x^{(j)})}. \tag{38}$$

For a reversible Markov chain we can also take the reversed sample path $(x^{(n)}, \ldots, x^{(1)})$ for approximating relative frequencies:

$$\kappa(B_k, B_l) = \lim_{n\to\infty} \frac{\sum_{j=2}^{n} \mathbf{1}_{B_k}(x^{(j)})\; \mathbf{1}_{B_l}(x^{(j-1)})}{\sum_{j=2}^{n} \mathbf{1}_{B_k}(x^{(j)})}. \tag{39}$$

Putting together all transitions (forward and reversed) results in

$$\kappa(B_k, B_l) = \lim_{n\to\infty} \frac{\sum_{j=1}^{n-1} \mathbf{1}_{B_k}(x^{(j)})\; \mathbf{1}_{B_l}(x^{(j+1)}) + \sum_{j=2}^{n} \mathbf{1}_{B_k}(x^{(j)})\; \mathbf{1}_{B_l}(x^{(j-1)})}{\sum_{j=1}^{n-1} \mathbf{1}_{B_k}(x^{(j)}) + \sum_{j=2}^{n} \mathbf{1}_{B_k}(x^{(j)})}, \tag{40}$$

which best reflects reversibility of $\mathcal{X}$. Estimation due to (40) has the nice property that the resulting discretization T is reversible for any finite sampling length n, which in (38) and (39) is only true for the limit $n \to \infty$.

3.3.4 Dynamical Clustering

In practice, we have to determine a suitable box discretization from a given sample $\mathbf{x}$. For this task, two problems has to be taken into account: (a) discretization of a high-dimensional state space and (b) a given finite sample size. Concerning problem (a) a discretization of P should be as fine as possible. In view of (b), an evaluation of an entry t_{ij} of T between two

boxes B_i and B_j are based on the finite sample size; hence good statistics requires a small number of boxes.

A straightforward identification strategy consists of two separate steps:

1. Geometric discretization of the Markov operator P by pre-clustering the sample $\mathbf{x}$ into boxes; this enables to set up a transition matrix T.

2. Identification of metastable sets as outlined in Sect. 3.2.3.

Further enhancement can be obtained by intertwining steps (1) and (2).

If the state space Ω would be low dimensional, a direct discretization of Ω would cause no problems. In order to discretize a high-dimensional state space, a variety of approaches has been proposed. Among them, some were especially designed to be applied to a geometric discretization of the state space of biomolecules:

Essential Degrees of Freedom. In many situations the number of degrees of freedom can be reduced by using a-priori knowledge about the system under consideration. Biomolecules are typically described in terms of their torsion angles (e.g., the single torsion angle of n-butane, see Fig. 4), and even from these only a subset is sufficient to describe the large-scale dynamics (e.g., the two essential torsion angles of n-pentane, see Fig. 25). Further reduction of the number of coordinates is possible via *Principal Component Analysis* (PCA) [5, 72].

Self Organizing Neural Networks. Neural networks are often used for cluster analysis. The task of clustering a large amount of sample points distributed in a high-dimensional continuous state space into clusters with geometrical similar sample points is often done by means of self organizing maps. The so obtained clusters could then be used to set up a transition matrix T, whereby each cluster represents one state of T. An extension of this approach are the so-called self-organizing box maps [47]. In addition to the clusters, boxes are assigned to each cluster (which are needed, e.g., for restricted sampling in identified metastable sets). We make use of this approach as part of the UC algorithm in Sect. 6.2 for simulations of n-pentane.

Combined Geometric and Dynamic Clustering. The above two approaches follow a strict separation of pre-clustering and identification. In order to obtain a discretization that already reflects metastability, it is also possible to intertwine steps (1) and (2). For the simulations of the biomolecule presented in Sect. 6.3 we employ such an approach, where a dynamical clustering on torsion angles identifies the most metastable coordinates in a hierarchical manner, which then gives rise to a Galerkin discretization on the whole state space [21].