4 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a powerful method of calculating probabilities or expectations that are intractable by analytical methods or other numerical approaches [84, 105]. The aim of MCMC is to efficiently draw samples from a given probability distribution. In many applications the samples are used afterwards to estimate a variety of integrals wrt. the probability distribution (e.g., thermodynamical integrals in statistical physics [36]).

A MCMC method consists of two steps, *Modeling* and *Realization*:

1. *Modeling*: for a given probability density f a Markov chain $\mathcal{X}$ is constructed with f as its unique invariant density. The crucial point in this construction is the so-called *detailed balance condition*, which enables to explore global properties of f by a series of local moves.

2. *Realization*: The algorithmic part consists of realizing one or several paths of $\mathcal{X}$. What one obtains is a series of more or less correlated random samples from f as guaranteed by Markov chain theory.

Essentially, MCMC transforms "simple" random numbers (typically uniform or Gaussian random numbers) into f-distributed random vectors. Since f represents a complicated distribution in a high-dimensional state space, other approaches like the direct Monte Carlo method would fail to produce f-distributed random vectors. Also note, that the aim of MCMC is somehow opposite to other algorithms, where for a given Markov chain the focus lies on the investigation of its invariant distribution.

One of the most prominent representatives of MCMC methods is the Metropolis (or Metropolis-Hastings) algorithm [66, 89, 90]. We describe its transition kernel in Sect. 4.1, the resulting Metropolis algorithm in Sect. 4.2, and discuss convergence diagnostics in Sect. 4.3.

Confronted with the trapping problem, many sophisticated Metropolis-based MCMC schemes have been proposed; all of them aiming at drawing samples by means of a rapidly mixing Markov chain. Among these, one outstanding method is *Hybrid Monte Carlo* (HMC) [15, 33], which combines *molecular dynamics* with the Metropolis algorithm. As we see in Sect. 4.2.2, HMC enables large moves in the state space by using short molecular dynamics trajectories of the underlying Hamiltonian system as proposal steps. Not surprisingly, HMC is a popular approach to explore the state space of biomolecules [11]. As a further step to overcome the trapping problem we introduce *Adaptive Temperature HMC* (ATHMC) and the related *Potential Scaling HMC* (PSHMC) in Sect. 4.4, both designed to draw samples from simple bridge distributions. Finally, we discuss some state-of-the-art extensions of the basic MCMC scheme in Sect. 4.5.

4.1 Metropolis Transition Kernel

We already introduced the general form of a Markov kernel in Sect. 3.3.1. Let us now turn to the construction of the Metropolis transition kernel, which will shed light on the underlying mathematical structure of the well-known Metropolis algorithm.

Let $f > 0$ be a probability density on the state space Ω. Our goal is to construct a transition kernel K having f as its unique invariant density as defined in (30). To this end we define an arbitrary irreducible transition kernel

$$Q(x, dy) = q(x, y)\, dy$$

on Ω (the so-called proposition kernel), together with the *acceptance function*

$$\alpha(x, y) = \begin{cases} \min\left\{1, \frac{q(y,x)\, f(y)}{q(x,y)\, f(x)}\right\} & \text{for } q(x, y) > 0 \\ 1 & \text{otherwise} \end{cases} . \tag{41}$$

In order to evaluate α one only needs ratios of the form $f(y)/f(x)$. Since $f(x) = h(x)/Z_{h(x)}$ is given explicitly and the unknown normalizing constant Z_h cancels out, the computation of $f(y)/f(x)$ reduces to the simple computation of the ratio of unnormalized densities $h(y)/h(x)$.

Together with Q and α we define the transition kernel K by

$$K(x, dy) = k(x, y)\, dy + r(x)\delta_x(dy), \tag{42}$$

that splits into an absolutely continuous part

$$k(x, y) = \begin{cases} q(x, y)\alpha(x, y) & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

and a singular component

$$r(x) = 1 - \int k(x, y)\, dy.$$

The construction of K guarantees that $\mathcal{X}$ is irreducible, provided that Q is irreducible. Therefore, we can state that f is the unique invariant density of $\mathcal{X}$, because for all $x, y \in \Omega$ with $x \neq y$ the detailed balance condition

$$f(x)\, k(x, y) = f(y)\, k(y, x) \tag{43}$$

holds. Due to (43) the transition kernel K is reversible wrt. f, from which self-adjointness in L^2_π and hence a real spectrum $\sigma(P)$ follows for the associated Markov operator (see Sect. 3.3.2). In practice, we can further assume $\mathcal{X}$ to be aperiodic, which for example would be already guaranteed for an average acceptance probability below 1 (for details, see e.g. [123]).

Finite State Space. If the state space is finite, the transition kernel reduces to a stochastic matrix. Without having to take into account all the details of a continuous state space and for a better understanding of how the Metropolis algorithm really works, we proof by the following theorem the key properties of the resulting transition matrix.

Theorem 10 *Let $Q = (q_{ij})$ be an arbitrary irreducible and stochastic matrix, for which $(q_{ij} \neq 0 \iff q_{ji} \neq 0)$ holds.*

Let $\Upsilon : (0,\infty) \to (0,1]$ be a function for which

$$\frac{\Upsilon(x)}{\Upsilon(1/x)} = x \qquad \text{for all} \quad x \in (0,\infty) \tag{44}$$

holds. For a given a probability vector π, the matrices A and T are defined as follows: $A \in \mathrm{Mat}_n(\mathbb{R})$ by

$$a_{ij} = \begin{cases} \Upsilon\left(\dfrac{\pi_j\, q_{ji}}{\pi_i\, q_{ij}}\right) \in (0,1], & \text{for } q_{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

and $T = (t_{ij}) \in \mathrm{Mat}_n(\mathbb{R})$ by

$$t_{ij} = \begin{cases} q_{ij}\, a_{ij}, & \text{for } i \neq j \\ 1 - \displaystyle\sum_{\substack{l=1 \\ l\neq i}}^{n} q_{il} a_{il}, & \text{if } i = j \end{cases} . \tag{45}$$

Then for T holds:

1. *T is an irreducible stochastic matrix;*
2. *and $\pi' T = \pi'$.*

Proof.

1. Since $\pi \geq 0$ and $Q \geq 0$ per definition, we have $A \geq 0$ and hence $T \geq 0$. Moreover, if $q_{ij} > 0$ then $a_{ij} > 0$ and thus $t_{ij} > 0$. Therefore, irreducibility is inherited from Q to A. Also, for all $i \in \{1,\ldots,n\}$ the rows sum up to 1:

$$\sum_{k=1}^{n} t_{ik} = \sum_{\substack{k=1 \\ k\neq i}}^{n} q_{ik} a_{ik} + 1 - \sum_{\substack{k=1 \\ k\neq i}}^{n} q_{ik} a_{ik} = 1.$$

2. We show that

$$\pi_i \, t_{ij} = \pi_j \, t_{ji} \qquad \text{for all} \quad i, j \in \{1, \ldots, n\} \tag{46}$$

holds, from which immediately $\pi' \, T = \pi'$ follows:

$$(\pi' \, P_M)_j = \sum_{i=1}^{n} \pi_i \, t_{ij} = \pi_j \sum_{i=1}^{n} t_{ji} = \pi_j \qquad \text{for all} \quad j \in \{1, \ldots, n\}.$$

Equation (46) is obviously fulfilled for the case $i = j$; also for $t_{ij} = 0$ the definition of Q implies $t_{ji} = 0$. For the case $t_{ij} > 0$ all occurring components of π, Q, and A are positive and due to the definition of Υ we have

$$\frac{a_{ij}}{a_{ji}} = \frac{\Upsilon\left(\frac{\pi_j \, q_{ji}}{\pi_i \, q_{ij}}\right)}{\Upsilon\left(\frac{\pi_i \, q_{ij}}{\pi_j \, q_{ji}}\right)} = \frac{\pi_j \, q_{ji}}{\pi_i \, q_{ij}},$$

which can directly be transformed into Eq. (46).

□

We call Q the proposal matrix, A the acceptance matrix, and T the transition matrix of the Metropolis Markov chain. The most commonly used function for Υ is the Metropolis function

$$\Upsilon_M(x) = \min\{1, x\}, \tag{47}$$

which we already applied directly in Eq. (41). Equation (47) was originally proposed by Metropolis et al. [89], and later also by Hastings [66] for the case of non-symmetric proposal steps. As alternatives to (47) several other acceptance functions that satisfy (44) have been considered [7, 66], but it was shown later that the Metropolis function (47) is in some sense an optimal choice [102].

Theorem 10 makes no assumption about T_V to be aperiodic, neither about a possible aperiodicity of T. This is no problem as long as we focus on the estimation of expectation values, since for that task aperiodicity is not required. Nevertheless, for a finite state space the existence of a state i with $t_{ii} > 0$ is sufficient to guarantee T to be aperiodic. Hence, in practice, we can assume aperiodicity, which furthermore guarantees the associated Markov chain to converge to its invariant distribution π.

4.2 Metropolis Algorithm

In the Metropolis algorithm the setup of the transition kernel K (or T for a finite state space) is directly used to generate a sample from its associated Markov chain $\mathcal{X}$. This can be done by iterating the following steps:

1. *Let $x^{(k)}$ be the present state.*
2. *Draw y from the proposal distribution $q(x, y)$ according to the proposal kernel Q.*
3. *Compute the acceptance probability $a(x, y)$ by*
$$a(x, y) = \Upsilon\left(\frac{f(y)\, q(y, x)}{f(x)\, q_(x, y)}\right).$$
4. *Draw $r \in [0, 1)$ uniformly.*
5. *Set*
$$x^{(k+1)} := \begin{cases} y, & \text{if } r \leq a(x, y) \\ x^{(k)}, & \text{otherwise} \end{cases}.$$

The quality of the samples will depend crucially on the relationship between Q and f. Although any proposal kernel will do the job, a proper choice of Q is essential to end up with a rapidly mixing $\mathcal{X}$. In practice, finding a suitable Q for a multimodal probability distribution is a challenging task. A good choice of Q should not only make the computations of steps (2) and (3) feasible, but also propose large moves in state space and lead to a high acceptance rate. For high-dimensional problems it is often necessary to perform some pre-simulations in order to tune parameters determining Q. Yet, there also exist some general schemes (like HMC, see Sect. 4.2.2) which can be applied out of the box and often work reasonable well for large problem classes.

Our initial aim was to compute ergodic averages $\mathbb{E}_f(g)$, which we now can estimate from the sample $\mathbf{x}$. According to the limit theorems from Sects. 3.2.1 and 3.3.2 we have

$$\mathbb{E}_f(g) \approx \frac{1}{n} \sum_{k=1}^{n} g(x^{(k)}) \tag{48}$$

for large n with an $\mathcal{O}(1/\sqrt{n})$-convergence towards $\mathbb{E}_f(g)$.

Due to the construction of K via the detailed balance condition, P is a reversible Markov operator. For that case, we have shown in Sects. 3.2.1 and 3.3.2 how convergence of such averages is connected to the eigenvalues of P, and that the spectral gap $\Lambda = 1 - \lambda_2$ plays an important role.

4.2.1 Examples for Proposal Steps

To illustrate the effect of different proposal steps on the Metropolis algorithm, we apply two simple Metropolis sampler to the canonical distribution associated to a one-dimensional asymmetric double-well potential. In Fig. 10 the potential $\mathcal{V}_{\mathrm{DW}}$ is shown together with its canonical density f_β for $\beta = 1$.

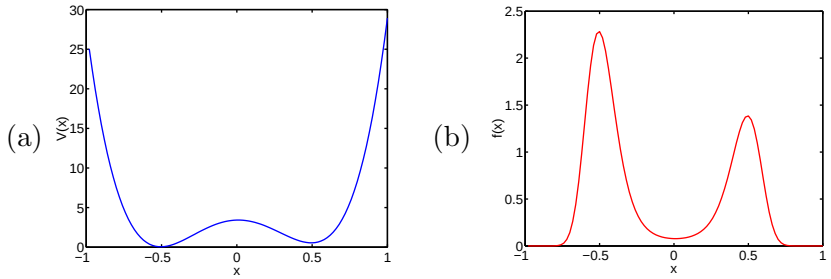

Figure 10: (a) One dimensional asymmetric double-well potential $\mathcal{V}_{\mathrm{DW}}$. (b) Canonical density of $\mathcal{V}_{\mathrm{DW}}$ for $\beta = 1$.

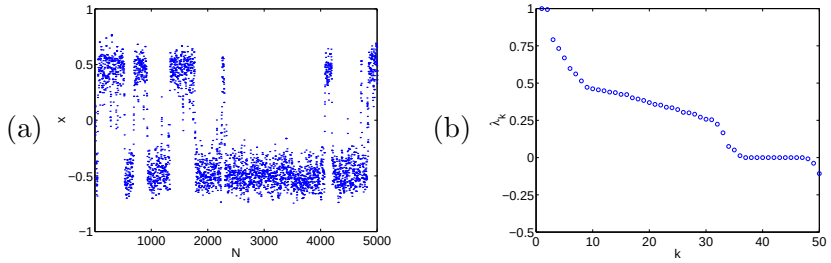

Figure 11: (a) A simulation with random walk proposals clearly shows a metastable behavior. (b) Eigenvalues of the discretized Markov operator T_{rand} for a discretization into 50 boxes. The 2nd eigenvalue $\lambda_2(T_{\mathrm{rand}}) = 0.9936$ is very close to 1.

Random-walk Metropolis. A simple and widespread method to generate proposals is to perform a random walk in state space. Samples are drawn from the correct distribution in that proposals are rejected in the acceptance step. Since this method makes use of symmetric proposals, the acceptance step reduces for the Metropolis function to

$$a(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}.$$

Applied to the double-well potential, Fig. 11 shows how local updates make it difficult to move from one well into the other. That each well forms

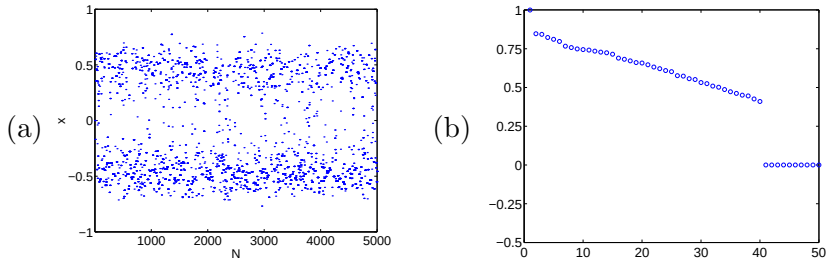

Figure 12: (a) A simulation with independent proposals is not affected by the energy barrier. (b) Eigenvalues of the discretized Markov operator T_{indep} for a discretization into 50 boxes. The 2nd eigenvalue $\lambda_2(T_{\mathrm{indep}}) = 0.8471$ is bounded far away from 1.

a metastable set is reflected by the spectral structure, with $\lambda_2 = 0.9936$ close to 1 followed by a spectral gap ($\lambda_3 = 0.7912$). The energy barrier in $\mathcal{V}_{\mathrm{DW}}$ is in fact a dynamical barrier for the Markov chain.

Independence Sampler. The name of this Metropolis sampler refers to the fact that the proposal distribution does not depend on x (i.e., $q(x, y) = q(y)$). For our example we apply the independence sampler in its simplest form with $q(y)$ being uniformly distributed on the state space $[-1, 1]$.

For independent proposals energy barriers of a potential have no impact on transitions between different parts of the state space, which is illustrated in Fig. 12 for $\mathcal{V}_{\mathrm{DW}}$. A second eigenvalue of $\lambda_2 = 0.8471$ confirms the absence of metastability in the Markov chain.

The independence sampler is a good choice, if it is possible to draw samples from a proposal distribution $q(y)$ that is close to $f(y)$ (for $q(y) \equiv f(y)$ the Markov chain would reduce to a sequence of independent random variables). Although finding good proposal distributions may be easy for many low-dimensional problems, it becomes more and more problematic in higher dimensions.

4.2.2 Hybrid Monte Carlo

Hybrid Monte Carlo (HMC) is a sophisticated sampling scheme that combines the theory of MCMC with the power of molecular dynamics in applications to molecular systems. It was invented in the late 80ies by Duane et al. for problems arising in quantum chromodynamics [33], and was soon thereafter applied to classical molecular systems [15, 87].

As outlined in Sect. 2.4 and Eq. (8), for a separable Hamiltonian $\mathcal{H}(x, p) = \mathcal{V}(x) + \mathcal{T}(p)$ the canonical distribution separates into the potential part $f_{\mathcal{V}}$ and momenta part $f_{\mathcal{T}}$.

For HMC, the momenta p serve as augmented variables in order to provide better mixing properties in position space. A Markov chain is realized in position space by propagating the system through state space by a series of trajectories, with new momenta drawn from the multidimensional Gaussian $f_{\mathcal{T}}$ before each step. For a given temperature the HMC Markov chain draws samples from $f_{\mathcal{V}}(\beta)$. More precisely, the steps required for one update from $x^{(k)}$ to $x^{(k+1)}$ are as follows:

1. *Generate new momenta p from the Gaussian distribution*

$$p \propto \exp\left[-\beta \sum_{i=1}^{d} \frac{p_i^2}{2m_i}\right].$$

2. *Run a time-reversible and volume preserving integration scheme with initial values $(x^{(k)}, p)$ and l iterations of time-step τ:*

$$(x', p') = (\Psi^{\tau})^l (x^{(k)}, p).$$

The vector x' is the new proposal for the Markov chain.

3. *Compute the energy difference ΔH in phase space, i.e.,*

$$\Delta H = H(x', p') - H(x^{(k)}, p).$$

4. *Accept the proposed state x' (i.e., set $x^{(k+1)} = x'$) with probability*

$$\min\{1, \exp[-\beta \Delta H]\},$$

otherwise set $x^{(k+1)} = x^{(k)}$.

Reversibility and volume preservation of Ψ is crucial for satisfying detailed balance, a rigorous proof is given in [84].

If we would propagate the system with the phase flow (i.e., $\Phi \equiv \Psi$), we have $\exp[-\beta \Delta H] = 0$ due to energy conservation (see Sect. 2.1), and the acceptance probability would be 1. In practice, however, we have to resort to an integration scheme which inherits reversibility and volume preservation. A well-known time-reversible integration scheme is the Leapfrog algorithm, which due to its symplecticity is also volume preserving (see Sect. 2.3). The only requirement for HMC is to keep the energy difference $\exp[-\beta \Delta H]$ small, which lead to a reasonable acceptance rate; there is no need for the integrator to be close to the analytical solution. Therefore, although much more sophisticated schemes that are time-reversible and volume preserving do exist, Leapfrog turned out to be very efficient in the context of HMC.

Extensions of HMC. Since its invention, HMC has been extended and combined with other advanced MCMC methods in a variety of ways. For example, it is possible to generate a new proposal by integrating wrt. an arbitrary Hamiltonian $\tilde{\mathcal{H}}$. The method is still valid if the acceptance probability is computed wrt. $\mathcal{H}$. Yet, in practice it is a difficult task to find a smoother or more suited $\tilde{\mathcal{H}}$ other than $\mathcal{H}$.

Another idea is to vary the trajectory length τl by drawing l from some distribution before each HMC step. More sophisticated is the "windowing" method proposed by Neal [95]. The trajectory is computed a fixed number of steps longer (the "window"), and the same number of additional steps backwards from the current state. The actual proposal state is then chosen from the window samples according to the Gibbs distribution; together with the additional "backward steps", which have an influence on the acceptance step, detailed balance is preserved. Furthermore, it is possible to incorporate HMC into other sampling approaches (e.g., multicanonical sampling [65] described in Sect. 4.5.1 or adaptive temperature HMC in Sect. 4.4.2).

4.3 Convergence Diagnostics

Deciding when to stop a Markov chain $\mathcal{X}$ is an important matter in practice. In general, one aims at generating a (finite) sample $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})$ being well distributed according to the invariant density f of $\mathcal{X}$.

For ergodic averages $1/n \sum_{k=1}^{n} g(x^{(k)})$ we outlined in Sect. 3 the limiting behavior for $n \to \infty$ in terms of the asymptotic variance σ_a. Knowledge of σ_a would be quite helpful in setting the sample size n, yet obtaining estimates for σ_a seems to be as difficult as for the ergodic average itself. In practice, the only source of information are the already drawn sample points $(x^{(1)}, \ldots, x^{(n)})$, and one has to decide from different kind of estimators whether or not to simulate further.

A big problem for all diagnostics tools is that they can falsely indicate convergence, which can happen when (maybe due to metastability) relevant parts of the state space were never visited. Therefore, it is surely advisable to run the chain at least a minimum number of steps (e.g., a certain amount of time or a fixed number of steps), which decreases the probability of false diagnostics. It is also often suggested to discard steps from the beginning (the so-called burn-in phase), in order to start the chain in equilibrium (i.e., one wants to have $x^{(1)} \sim f$).

Two main approaches for convergence diagnostics are *single chain* [54] and *multiple chain* [51] diagnostics. A general overview of MCMC convergence diagnostics is given in [23]. In view of the central limit theorem it seems advisable to run a single chain as long as possible, since this is what MCMC really is about. Although questionable, it is often argued that in the presence of metastability it is more likely to observe a transition in a single long chain than in multiple shorter ones. On the contrary, multiple shorter runs can also be of diagnostic value. If the chains get stuck in different metastable sets, it is easy to diagnose lack of convergence. Especially, when overdispersed starting points are available, multiple chains seems to be favorable. Moreover, running several chains in parallel is much easier than implementing a parallel version for a single chain. The question of whether running one long chain or several smaller chains is discussed controversially (see [51, 54] and discussions therein), but it is generally acknowledged that no method works well in every situation. The choice depends on the particular problem as well as the computing facilities available. Since we later on use a multiple chain approach we shortly outline the steps needed to set up an estimator.

Multiple Chain Diagnostics. The general idea behind this approach is to run multiple Markov chains in parallel (ideally from overdispersed starting points), and analyze the samples at fixed time intervals until all estimates of interesting observables do agree adequately.

A popular estimator of that kind has been proposed by Gelman and

Rubin [49, 51]. Suppose that we started m (say, $m = 5$) Markov chains and run each of them for n steps. For each scalar observable g we label the m chains of length n as (g_{ij}), $j = 1, \ldots, m$; $i = 1, \ldots, n$, and compute the *between-sequence variance* B and the *within-sequence variance* W:

$$B = \frac{n}{m-1} \sum_{i=1}^{m} ((\bar{g}_{i\cdot}) - (\bar{g}_{\cdot\cdot}))^2, \text{ where } \quad \bar{g}_{i\cdot} = \frac{1}{n} \sum_{i=1}^{n} g_{ij}, \bar{g}_{\cdot\cdot} = \frac{1}{m} \sum_{i=1}^{m} \bar{g}_{i\cdot}$$

$$W = \frac{1}{m} \sum_{i=1}^{m} \sigma_i^2, \quad \text{where} \quad \sigma_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (g_{ij} - (\bar{g}_{i\cdot}))^2.$$

The within-sequence variance W is assumed to be an underestimate of $\mathrm{Var}(g)$. From B and W another estimator of $\mathrm{Var}(g)$ is constructed,

$$\widehat{\mathrm{Var}}(g) = \frac{n-1}{n} W + \frac{1}{n} B,$$

which is an overestimate under the assumption that the starting points are overdispersed. With these two estimates of $\mathrm{Var}(g)$, the so-called *estimated potential scale reduction* is given by

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{\mathrm{Var}}(g)}{W}}.$$

The value of $\hat{R}$ converges to 1 in the limit $n \to \infty$ and should start declining to 1 from above if all parallel chains are essentially overlapping. A criterion to stop the simulation automatically would be to wait until $\hat{R}$ decreased below a given threshold (say, 1.1) for all observables.

Remark. It is advisable to apply convergence diagnostics to a variety of suitable observables (e.g., the total energy and all important torsion angles for a biomolecular system). The computational cost for convergence diagnostics is typically negligible in comparison to obtaining the samples (this is especially true for methods like HMC). Therefore many practitioners recommend to monitor convergence by different methods in order to get the most out of these sometimes controversially discussed diagnostic methods (cf. discussion in [51]). Within UC, we use the multiple chain diagnostics proposed by Gelman and Rubin. This allows for simple parallel implementation, and we can benefit from available overdispersed starting points from previous bridge samples at higher temperatures, which makes estimation more robust and lowers the probability of overlooking important parts of the state space.

4.4 Bridge Distributions

The MCMC methods presented so far directly draw samples from a (canonical) distribution $f(\beta)$ for a given inverse temperature β, without embedding $f(\beta)$ into a series of tempered distributions (or any other form of auxiliary distributions). A practical way to extend $f(\beta)$ are the so-called *bridge distributions*, which can be thought of as elementary auxiliary distributions.

Suppose, we are interested in drawing samples at a low temperature $\mathrm{T}_{\mathrm{low}} = 1/(k_B\beta_{\mathrm{low}})$, and that the (to our knowledge) best available MCMC method for $f_{\mathrm{low}} = f(\beta_{\mathrm{low}})$ is still slowly mixing. Let us further assume that by increasing the temperature to $\mathrm{T}_{\mathrm{high}} = 1/(k_B\beta_{\mathrm{high}})$ the MCMC method becomes rapidly mixing for $f_{\mathrm{high}} = f(\beta_{\mathrm{high}})$. Unfortunately, a sample $\mathbf{x}_{\mathrm{high}}$ from f_{high} provides only very limited information about the situation at f_{low}, since it mainly draws samples from different parts of the state space.

At this point bridge distribution techniques come into play, which enable to combine drawing samples from f_{low} with the mixing properties of f_{high}. The most direct way to define a bridge distribution is in terms of the normalized densities (say, $f_{\mathrm{high,low}} = \sigma f_{\mathrm{high}} + (1-\sigma) f_{\mathrm{low}}$). Yet, for practical purposes it is more natural to define $f_{\mathrm{high,low}}$ via its unnormalized bridge density $h_{\mathrm{high,low}}$ in terms of the unnormalized densities h_{low} and h_{high}.

Simply using the sum $h_{\mathrm{high}} + h_{\mathrm{low}}$ would be unbalanced with the effect of putting most of the probability on one of its parts. To avoid this undesirable effect one has to introduce shift parameters that enables to balance the probability of the two canonical distributions. A generic form of a bridge density is

$$h_{\mathrm{high,low}} = \sigma_1 h_{\mathrm{high}} + \sigma_2 h_{\mathrm{low}}, \tag{49}$$

where suitable values for σ_1 and σ_2 enable to equalize the influence of the two distributions on $h_{\mathrm{high,low}}$. In particular, by defining $f_{\mathrm{high,low}}$ as in (49) we expect to satisfy:

1. The Markov chain $\mathcal{X}_{\mathrm{high,low}}$ associated with $f_{\mathrm{high,low}}$ is rapidly mixing.

2. A sample $\mathbf{x}$ of $f_{\mathrm{high,low}}$ allows a statistical reasonable reweighting to f_{high} and f_{low}; this presupposes that all important parts of f_{high} and f_{low} are covered by $f_{\mathrm{high,low}}$.

3. Estimation of ratio of normalizing constants $Z_{h_{\mathrm{high}}}/Z_{h_{\mathrm{low}}}$ by reweighting (this will become important in Sect. 5.3).

The algorithmic steps necessary to apply bridge simulation techniques is to (a) define a "suitable" bridge density $f_{\mathrm{high,low}}$, (b) draw a sample $\mathbf{x}$ from $f_{\mathrm{high,low}}$, and (c) reweight the sample to f_{low} afterwards. For step (a), making a good choice of the parameters becomes easier if the two distributions are closer to each other. For canonical distributions this means that the temperature difference $T_{\mathrm{high}} - T_{\mathrm{low}}$ cannot be chosen arbitrarily high. We

discuss the important task of parameter estimation for concrete situations in Sects. 4.4.2 and 4.4.3.

Remark. Bridge distributions have been used in the statistics community [50, 55, 88] to combine samples from two different distributions in order to compute the ratio of normalizing constants by an *a-posteriori* estimator. In contrast, our use of bridge distributions aims at determining *a-priori* a bridge density $f_{\text{high,low}}$, and then directly draw samples from $f_{\text{high,low}}$ by some adopted MCMC method.

4.4.1 Reweighting

Bridge sampling and a variety of other extended MCMC methods are based on *reweighting techniques*. Suppose, that we are given a sample $\mathbf{x}$ distributed according to an auxiliary distribution f_{aux}. We are not directly interested in f_{aux}, but rather about some "nearby" target distribution f_* in order to extract from $\mathbf{x}$ valuable information about f_* (in the context of bridge distributions we would have $f_{\text{aux}} = f_{\text{high,low}}$ and $f_* = f_{\text{low}}$). The key idea is to estimate from a sample $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ weights $\mathbf{w}_* = (w_*^{(1)}, w_*^{(2)}, \ldots, w_*^{(n)})$ such that the weighted sample $(\mathbf{w}, \mathbf{x})$ is distributed according to f_*.

The estimator for $\mathbf{w}_*$ is given by

$$w_*^{(k)} = \frac{h_*(x^{(k)}) \,/\, h_{\text{aux}}(x^{(k)})}{\sum_{l=1}^{n} h_*(x^{(l)}) \,/\, h_{\text{aux}}(x^{(l)})} \qquad \text{for} \quad k = 1, \ldots, n. \tag{50}$$

With these weight factors we can estimate $\mathbb{E}_{f_*}(g)$ for some observable A. By using $A(x)\, w_*(x)$ as an observable wrt. f_{aux} the unnormalized density h_{aux} as well as the normalizing constant $Z_{h_{\text{aux}}}$ cancel out (though in a different way), and we end up estimating $A(x)$ wrt. f_*:

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} A(x^{(k)})\, w_*^{(k)} &= \frac{\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} A(x^{(k)})\, h_*(x^{(k)}) \,/\, h_{\text{aux}}(x^{(k)})}{\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} h_*(x^{(k)}) \,/\, h_{\text{aux}}(x^{(k)})} \\
&= \frac{\int A(x)\, [h_*(x) \,/\, h_{\text{aux}}(x)]\, f_{\text{aux}}(x)\, dx}{\int [h_*(x) \,/\, h_{\text{aux}}(x)]\, f_{\text{aux}}(x)\, dx} \qquad (51) \\
&= \frac{\int A(x)\, h_*(x)\, dx}{\int h_*(x)\, dx} \\
&= \mathbb{E}_{f_*}(g)
\end{aligned}$$

Equation (51) states an asymptotic result, but provides no information about the quality of estimates for $\mathbb{E}_{f_*}(g)$ given a finite sample of f_{aux}. If

f_* is not covered sufficiently by f_{aux}, the weighted sample $(\mathbf{w}, \mathbf{x})$ will be statistically unreliable.

Reweighting is a standard tool for many MCMC-based algorithms, and plays an important role in analyzing bridge distributions in statistics [50, 55, 88]. In statistical physics reweighting between nearby canonical distributions was proposed by Ferrenberg and Swendsen [37]. In the context of UC we make extensive use of reweighting in Sect. 5.3.2 (including estimation of quotients of normalizing constants). Ferrenberg and Swendsen also proposed a reweighting scheme for combining samples from different but partly overlapping distributions [38]. We consider this more complicated case of *reweighting mixtures* in Sect. 5.3.3.

4.4.2 Adaptive Temperature HMC

Adaptive Temperature HMC (ATHMC) [42] provides a HMC scheme for bridge distributions constructed out of two canonical distributions. Suppose we are interested in the (unnormalized) canonical distribution given by

$$h_{\text{low}}(x) = \exp[-\beta_{\text{low}}(\mathcal{V}(x)]$$

for a given potential $\mathcal{V}$ and temperature $T_{\text{low}} = 1/(k_B\beta_{\text{low}})$, and that we face the sampling problem described at the beginning of Sect. 4.4. Let us consider bridge densities $f_{\text{aux}} = h_{\text{aux}}/Z_{h_{\text{aux}}}$ of the form

$$h_{\text{aux}}(x) = \frac{1}{2}\left(\exp[-\beta_{\text{low}}(\mathcal{V}(x) - \nu)] + \exp[-\beta_{\text{high}}(\mathcal{V}(x) - \nu)]\right), \tag{52}$$

which we denote as *mixed-canonical distribution* in the following. The mixed-canonical distribution consists of the sum of two canonical distributions with temperatures β_{low} and β_{high}, respectively, whereby the constant ν determines which of the two distributions dominates f_{aux}.

Clearly, for $\beta_{\text{low}} = \beta_{\text{high}}$, the mixed-canonical distribution reduces to f_{low}. Furthermore, for $\beta_{\text{low}} > \beta_{\text{high}}$ the mixed-canonical distribution h_{aux} converges to f_{high} or f_{low} if ν tends to ∞ or $-\infty$, respectively.

By setting

$$\sigma_1 = 2\exp[\beta_{\text{low}}\,\nu] \qquad \text{and} \qquad \sigma_2 = 2\exp[\beta_{\text{high}}\,\nu]$$

we see that Eq. (52) is a special case of (49). However, the form of Eq. (52) is better suited to determine the parameter ν such that f_{low} and f_{high} are both well represented by f_{aux}. Drawing samples with potential energies above ν is similar to drawing samples from f_{high}, and the same holds for energies below ν and f_{low}. By setting

$$\nu \approx \frac{1}{2}\left(\mathbb{E}_{f_{\text{low}}}(\mathcal{V}) + \mathbb{E}_{f_{\text{high}}}(\mathcal{V})\right) \tag{53}$$

an approximately equal balance between f_{low} and f_{high} can be assumed. Equation (53) let us presume that we already need samples $\mathbf{x}_{\text{low}}$ and $\mathbf{x}_{\text{high}}$ in order to estimate ν. Fortunately, it is already sufficient to have some sample $\mathbf{x}_{\text{high}}$ (e.g., from some short preliminary simulation of f_{high} or as part of a more complex sampling strategy like UC); instead of (53) we can use the similar (though not equivalent) expectation value

$$\nu \approx \mathbb{E}_{f_{\text{avg}}}(\mathcal{V}) \quad \text{where} \quad \beta_{\text{avg}} = \frac{2\beta_{\text{low}}\beta_{\text{high}}}{\beta_{\text{low}} + \beta_{\text{high}}} \tag{54}$$

wrt. the averaged temperature β_{avg}. If the temperature difference is not too high, reweighting of $\mathbf{x}_{\text{high}}$ to f_{avg} is possible, from which one obtains an estimation of ν.

ATHMC Scheme. Suppose, that all parameters are set to reasonable values, and that we now want to draw samples from f_{aux}. For a general MCMC method we could directly start the simulation, but in the context of HMC we face the problem of choosing the "right" temperature in order to draw the momenta in dependence of a fixed β. We could use β_{avg} as defined in (54), yet it is more appropriate to adapt the temperature in dependence of the potential energy by the inverse temperature function (see Fig. 13 (a))

$$\beta(x) = -\frac{\ln h_{\text{aux}}(x)}{\mathcal{V}(x) - \nu}. \tag{55}$$

With this adaptive temperature choice one update step with ATHMC consists of (for details and a justification of the modification see [42]):

1. *Initialization of momenta p at the inverse temperature $\beta(x^{(k)})$, i.e.,*

$$p \propto \exp\left[-\beta(x^{(k)}) \sum_{j=1}^{d} \frac{p_j^2}{2m_j}\right],$$

   where d denotes the number of degrees of freedom.

2. *Calculation of new coordinates and momenta*

$$(x', p') = (\Psi^\tau)^l\,(x^{(k)}, p).$$

3. *Computation of $\beta(x')$.*

4. *Acceptance of new coordinates x' (i.e., set $x^{(k+1)} = x'$) with probability*

$$\min\left\{1, \frac{h_{\text{aux}}(x')\ \exp[-\beta(x')\,\mathcal{T}(p')]}{h_{\text{aux}}(x^{(k)})\ \exp[-\beta(x^{(k)})\,\mathcal{T}(p)]}\ \left(\frac{\beta(x')}{\beta(x^{(k)})}\right)^{d/2}\right\},$$

   otherwise set $x^{(k+1)} = x^{(k)}$.

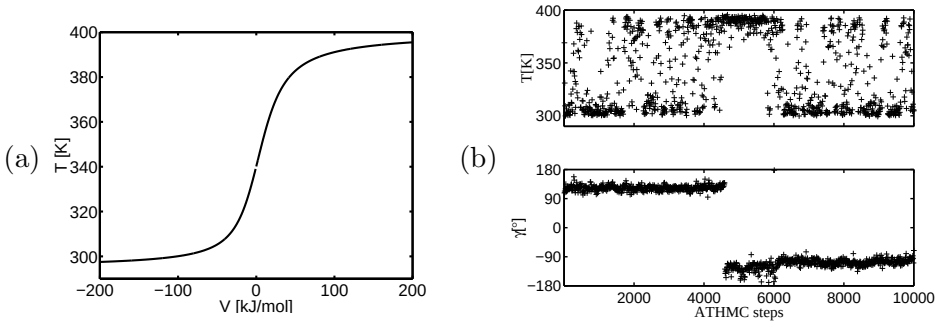

Figure 13: (a) Temperature function $T(x) = (k_B\beta(x))^{-1}$ in dependence of the potential energy with $T_{\text{low}} = 295\,\text{K}$, $T_{\text{high}} = 400\,\text{K}$ and $\nu = 0$. The temperature function reflects the locale structure of the mixed-canonical distribution. By changing ν the temperature function is shifted and f_{aux} is altered. (b) ATHMC for *r(ACC)* in a mixed-canonical distribution. The simulation was performed for $T = 295\,\text{K}$, $T = 400\,\text{K}$ and $\nu = -1121\,\text{kJ/mol}$. The temperature T and the torsion angle γ are displayed at every tenth step over the first 10000 steps. The transition in γ is induced from drawing samples at higher temperatures.

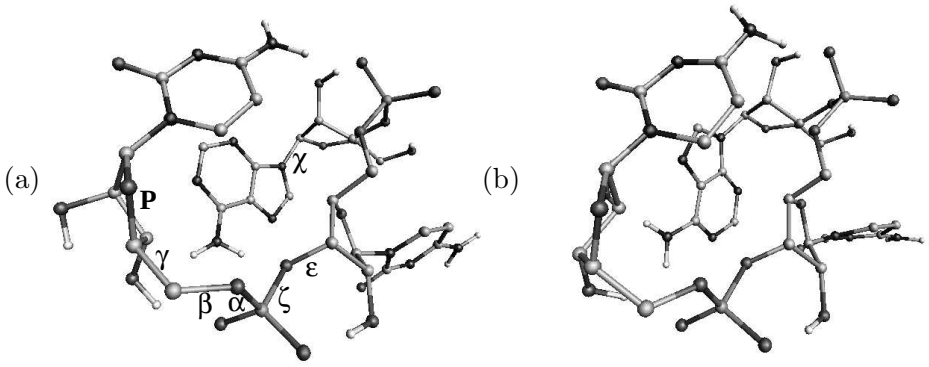

Figure 14: Representatives from two metastable sets of r(ACC), whereby differences show up in the torsion angle χ and ring structure **P**. (a) To state this more precisely in biochemical terms, the χ angle around the first glycosidic bond is in *anti* position ($-175°$) and the terminal ribose pucker **P** is in C(3')endo C(2')exo conformation. (b) The χ angle is in *syn* position (19° degrees) and the terminal ribose in C(2')endo C(3')exo conformation.

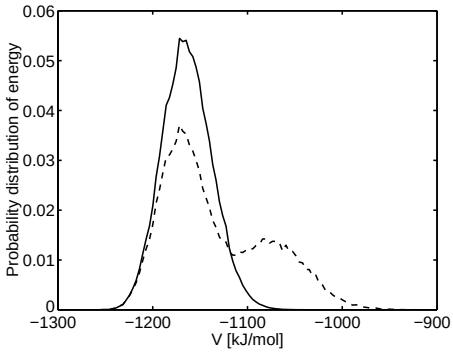

Figure 15: Reweighting of the ATHMC simulation from the sample of the mixed-canonical distribution (- - -) to the canonical distribution at β_{low} (—).

Example. We illustrate by a simulation of the triribonucleotide r(ACC) (see Fig. 5 on page 16) how ATHMC helps to induce transitions between metastable sets in a biomolecule.

The simulation was performed over 5×10^5 steps with parameter settings $T_{\text{low}} = 295\,\text{K}$, $T_{\text{high}} = 400\,\text{K}$ and $\nu = -1121\,\text{kJ/mol}$, whereby ν was obtained from short preliminary runs. In Fig. 13 (b), the first 10^4 steps of the simulation are shown for the temperature and the torsion angle γ. We observe frequent fluctuations between low and high temperature, and hence between low and high energy regions. A transition between metastable sets is indicated by γ, which happens while the simulation draws samples in high temperature regions. In Fig. 14 two metastable conformations are shown, which distinguish from one another by the orientation of the χ angle and the ribose pucker **P**.

The probability distribution of energy for f_{aux} and f_{low} (i.e., before and after reweighting) is shown in Fig. 13. Without reweighting we observe two maxima around the averaged potential energies at f_{low} and f_{high}. ATHMC stretches the energy range, but still puts high probability on f_{low}, as indicated by the large overlap of canonical and mixed-canonical distribution.

A detailed analysis of the ATHMC simulation for r(ACC) is given in [42], a conformational analysis based on the reweighted sample was done in [72].

4.4.3 Potential Scaling HMC

Our goal is to reformulate the idea of ATHMC as a *potential scaling method*. This is motivated by (a) making the adjustment of bridge distribution parameters more flexible, and (b) providing an approach that has the potential to incorporate other parameters than the temperature for the construction of f_{aux}. In combination with HMC we call this method *Potential Scaling* HMC (PSHMC).

Suppose, we are interested in β_{low}, and our high (inverse) temperature is β_{high}. Then we can express f_{high} in terms of β_{low} via potential scaling, i.e.,

$$\begin{aligned} f_{\text{high}}(x) &= \exp[-\beta_{\text{high}}\mathcal{V}(x)] \\ &= \exp[-\beta_{\text{low}}\, s(\mathcal{V}(x))] \end{aligned}$$

with $0 < \alpha = \beta_{\text{high}}/\beta_{\text{low}} \leq 1$ and $s(y) = \alpha\, y$.

Our aim is to extend this relationship to a bridge distribution that encompasses given canonical distributions f_{low} and f_{high}. To that end, we set up a bridge distribution via a nonlinear scaling function $s : \mathbb{R} \to \mathbb{R}$ for some fixed temperature β_* (not necessarily β_{low}).

The nonlinear scaling function is subdivided into three parts:

$$s(x) = \begin{cases} \alpha_{\text{low}}\, x & \text{for } x < a \\ g(x) & \text{for } x \in [a,b] \\ \alpha_{\text{high}}\, x + (\alpha_{\text{low}} - \alpha_{\text{high}})\frac{a+b}{2} & \text{for } x > b \end{cases},$$

where g denotes the spline function

$$g(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3$$

with parameters

$$\begin{aligned} c_0 &= \alpha_{\text{low}}\, a \\ c_1 &= \alpha_{\text{low}} \\ c_2 &= \frac{3}{(b-a)^2}\left(\alpha_{\text{high}}\, b + (\alpha_{\text{low}} - \alpha_{\text{high}})\frac{a+b}{2} - \alpha_{\text{low}}\, a\right) \\ &\quad -\frac{1}{b-a}(\alpha_{\text{high}} + 2\,\alpha_{\text{low}}) \\ c_3 &= \frac{2}{(b-a)^3}\left(\alpha_{\text{low}}\, a - \alpha_{\text{high}}\, b + (\alpha_{\text{high}} - \alpha_{\text{low}})\frac{a+b}{2}\right) \\ &\quad +\frac{1}{(b-a)^2}(\alpha_{\text{low}} + \alpha_{\text{high}}) \end{aligned}$$

With these settings, g fulfills the boundary conditions $g(a) = \alpha_{\text{low}}\, a$, $g(b) = \alpha_{\text{high}}\, b + (\alpha_{\text{low}} - \alpha_{\text{high}})(a+b)/2$, $g'(a) = \alpha_{\text{low}}$, and $g'(b) = \alpha_{\text{high}}$; i.e., s is differentiable.

The bridge distribution is now defined as

$$f_{\text{aux}}(x) = \exp[-\beta_*\, \mathcal{V}_{\text{aux}}(x)] \quad \text{with} \quad \mathcal{V}_{\text{aux}}(x) = s(\mathcal{V}(x)).$$

In Fig. 16 the scaling function $s(x)$ and its derivative is plotted for $a = 50$, $b = 100$, $\alpha_{\text{low}} = 0.8$ and $\alpha_{\text{high}} = 0.4$. Drawing samples from f_{aux} shows

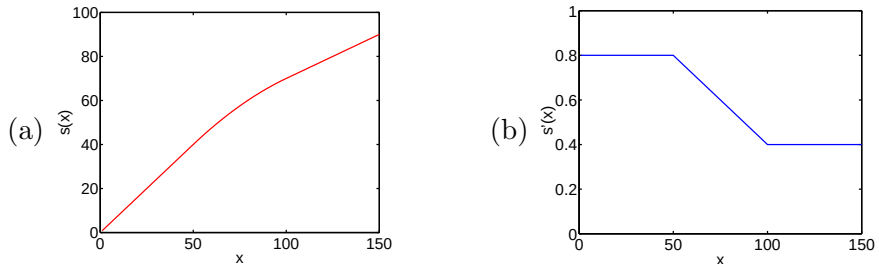

Figure 16: (a) Scaling function s for $a = 50$, $b = 100$, $\alpha_{\text{low}} = 0.8$ and $\alpha_{\text{high}} = 0.4$. (b) First derivative s'.

the following characteristics: below a the unnormalized density of f_{aux} is identical to f_{low} (up to a normalizing constant), and above b it is identical to f_{high}; between a and b the system moves from low to high temparature. Since s is strictly monotonic decreasing, no artificial local minima are introduced by this scaling procedure.

For fixed low and high temperature (i.e., scaling factors α_{low} and α_{high}) we need to determine suitable parameters a and b. It is more appropriate to describe a and b by the arithmetic mean $\nu = (a + b)/2$ and $\sigma = b - \nu$. Similar to ATHMC, a short pre-simulation at f_{high} is sufficient to estimate the parameters ν and σ. The parameter ν is evaluated via

$$\nu \approx \mathbb{E}_{f_*}(\alpha_{\text{avg}}\mathcal{V}) \quad \text{where} \quad \alpha_{\text{avg}} = \frac{\alpha_{\text{low}} - \alpha_{\text{high}}}{2},$$

which is analogue to Eq. (54) for ATHMC. A suitable choice for σ is

$$\sigma \approx \mathbb{E}_{f_*}(\alpha_{\text{high}}\,\mathcal{V}) - \mathbb{E}_{f_*}(\alpha_{\text{avg}}\,\mathcal{V}) - 2\,[\text{Var}_{f_*}(\alpha_{\text{high}}\,\mathcal{V})]^{1/2},$$

which is a compromise between a broad energy range for transitions between β_{low} and β_{high}, and a big overlap of f_{aux} with f_{low} and f_{high}.

With all parameters set, a sample $\mathbf{x}$ from the bridge distribution is obtained by applying standard HMC to $\mathcal{V}_{\text{aux}}$. The only extension is that in order to evaluate $\mathcal{V}'_{\text{aux}}(x) = s'(\mathcal{V}(x))\mathcal{V}'(x)$, each MD step of the integration scheme now also requires the evaluation of the original potential function $\mathcal{V}$. Then, by applying (51) to the sample $\mathbf{x}$, reweighting from $f_\beta(\mathcal{V}_{\text{aux}}(x))$ to $f_\beta(\mathcal{V}(x))$ is straightforward.

Remark. The additional parameter σ makes PSHMC more flexible than ATHMC. Moreover, in ATHMC detailed balance is obtained via a modified acceptance step, which can result in a decrease of the acceptance ratio. In contrast, the trajectories of a PSHMC simulation are directly computed in $\mathcal{V}_{\text{aux}}$, thus the acceptance ratio is comparable to HMC. On the other hand, an advantage for ATHMC is that all trajectories are computed wrt. $\mathcal{V}$, whereas computation wrt. $\mathcal{V}_{\text{aux}}$ is more expensive. We incorporate PSHMC as bridge sampling method into UC, and therefore postpone applications of PSHMC to biomolecular systems to Sect. 6.

4.5 Extended MCMC Methods

Most of the methods that try to tackle the trapping problem introduce considerable overhead, e.g., in form of presimulations for paramter estimation, reweighting, additional data analysis, or simply costly update steps as in HMC. Yet, in terms of efficiency they often outperform simple strategies up to some orders of magnitude [18, 36, 56, 83, 84].

Improved Updating. A direct way is to build intriguing update procedures that accelerate the mixing behavior of the Markov chain while still having the target distribution as its invariant distribution. There exist numerous methods of that kind, some popular ones are HMC [33], Swendsen-Wang algorithm [121], multigrid Monte Carlo [57], reversible jump MC [58], or configurational bias MC [117], to name a few. The efficiency and applicability often depend heavily on the actual problem at hand, and some methods are tailored to special application fields (e.g., discrete or continuous models).

Auxiliary Distributions. A Markov chain need not necessarily have the target distribution as its invariant distribution. A more general strategy, which further extends the concept of bridge distributions, is to make use of auxiliary distributions. An auxiliary distribution f_{aux} provide a very powerful framework that not only allows to obtain estimates wrt. the target distribution f_*, but also enables free energy computations and estimation of observables in dependence of some paramter range.

In general, extended MCMC methods based on an auxiliary distribution consist of the following steps:

1. *Construction of an auxiliary ensemble f_{aux}, where a Markov chain can move around freely; f_{aux} should contain all regions in f_* of high probability.*

2. *Determination of parameters for f_{aux} via initial sampling or some kind of iterative procedure. This is a crucial part for all extended ensemble methods, since a reasonable specification of f_{aux} needs information from the actual system under consideration.*

3. *A long simulation run by a (rapidly mixing) Markov chain that draw samples from the then fully specified f_{aux}.*

4. *Reweighting the sample from f_{aux} to f.*

Historically, Torrie and Valleau [124, 125] were one of the first who investigated auxiliary distributions by their *umbrella sampling* method, mainly to determine free energy differences. Auxiliary distributions which are used

in *umbrella sampling* are understood to span up a substantial range of different physical situations. Under this general description the bridge sampling methods ATHMC and PSHMC from Sect. 4.4 belong also to the category of umbrella sampling.

If f_* is a canonical distribution, the most natural way to construct f_{aux} is via the temperature. Methods based on such tempered distributions are *simulated tempering* [53, 86], where the Markov chain jumps between different canonical distributions, and simulations in a *multicanonical* [9] or $1/k$*-ensemble* [68]. All of these approaches can be used in connection with HMC [64].

We shortly outline two popular extended MCMC methods in the following, namely *multicanonical sampling* and *parallel tempering*.

4.5.1 Multicanonical Sampling

Multicanonical sampling proposed by Berg and Neuhaus [8, 9] aims at sampling the state space over a wide energy range by one long simulation run.

For a canonical distribution $f(x) = \exp[-\beta\,\mathcal{H}(x)]/Z$ with Z being the normalizing constant and $\mathcal{H} \equiv E$, the density in terms of the energy is

$$f(E) = n(E)\exp[-\beta\,E]/Z,$$

where $n(E)$ is the density of states (the canonical energy distribution with typically one single peak is illustrated in Fig. 6 (b)).

In contrast, the multicanonical method seeks to sample from a flat energy distribution. In order to define a multicanonical distribution a finite energy range is divided into L small segments, and we have

$$f_{\text{mult}}(E) = \frac{\exp[-S(\mathcal{H}(x))/k_B]}{L},$$

where k_B is Boltzmanns' constant and $S(\mathcal{H}(x)) = k_B \ln n(\mathcal{H}(x))$ is the entropy of the microcanonical ensemble at energy level $\mathcal{H}(x)$. In other words, the energy is uniformly distributed over the L energy segments. In practice, one has to determine a weight for each energy segment. Two step are required for the multicanonical method:

1. *parameter estimation*: Estimation of the weights is done in an iterative procedure of short preliminary simulations. The thereby defined density f_{mult} need not to produce a perfectly flat energy distribution, since a rough approximation does not affect the quality of reweighting from f_{mult}.

2. *multicanonical simulation*: A single Markov chain (or multiple independent Markov chains) draws samples from f_{mult} by some MCMC

method. A statistically correct and reliable reweighting of the resulting sample $\mathbf{x}_{\text{mult}}$ is possible to any canonical distribution that lies within the energy range covered by f_{mult}.

The flat energy distribution helps to overcome energy barriers. Yet, it still could happen that globally determined weights produce locally distorted energy distribution; the Markov chain would then be hindered to move freely over the energy range. Multicanonical sampling has been applied to continuum peptide models [133], and also combined with molecular dynamics, Langevin dynamics, and HMC [65].

4.5.2 Parallel Tempering

Parallel Markov chains that exchange information during simulation are another promising framework to attack the trapping problem. An established method of this kind is *parallel tempering* [53]. The method is also sometimes referred to under the name *exchange Monte Carlo*, when it was reinvented in [74].

In this method N Markov chains $\mathcal{X}_1, \ldots, \mathcal{X}_N$ run in parallel on the state space Ω, each one associated to a canonical density f_k at temperature T_k for $k = 1, \ldots, N$, such that

$$T_1 < T_2 < \ldots < T_N$$

span up a broad temperature range with T_1 associated to the target distribution. The resulting algorithm consists of a mixture of parallel and swapping update steps, which are chosen according to some iterative or random procedure:

- *parallel step:* Each Markov chain $\mathcal{X}_k$ performs an ordinary update step from $x_k^{(l)}$ to $x_k^{(l+1)}$ via its respective MCMC scheme.
- *swapping step:* Two neighboring temperature levels, say k and $k+1$, are chosen by random, and the respective states $x_k^{(l)}$ and $x_{k+1}^{(l)}$ swap between the Markov chains $\mathcal{X}_k$ and $\mathcal{X}_{k+1}$ with probability

$$\min\left\{1, \frac{f_k(x_{k+1}^{(l)})\, f_{k+1}(x_k^{(l)})}{f_k(x_k^{(l)})\, f_{k+1}(x_{k+1}^{(l)})}\right\}$$

Swapping allows to transfer states from slow mixing low temperature levels to rapidly mixing high temperature levels and vice versa, hence improving mixing at low temperatures. Without such swapping steps, one would simply run N independent Markov chains in parallel.

An analysis of this approach is done by directly dealing with the product space $\Omega_1 \times \cdots \times \Omega_N$ together with the joint probability distribution of the

canonical distributions on the product space. That way, one can show that both, parallel and swapping steps, fulfill detailed balance in this extended setting, and that each Markov chain $\mathcal{X}_k$ draw samples from the canonical distribution at T_k. Therefore, parallel tempering provides statistical information over the whole temperature range between T_1 and T_N .

An advantage of parallel tempering over other extended methods is that once the temperature values are chosen, a simulation can directly be started without further adjustments of parameters. Decisive for a fast mixing at T_1 is to choose a proper number of Markov chains; too few prevent neighboring Markov chains to exchange their state in the swapping step, too many prevent a fast mixing between low and high temperatures.