Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

ALEXANDER FISCHER

An Uncoupling-Coupling Technique for Markov Chain Monte Carlo Methods

ZIB-Report 00-04 (September 2000)

An Uncoupling-Coupling Technique for Markov Chain Monte Carlo Methods

Alexander Fischer*

September 12, 2000

Abstract

Uncoupling-coupling Monte Carlo (UCMC) combines uncoupling techniques for finite Markov chains with Markov chain Monte Carlo methodology. By determining almost invariant sets of the associated Markov operator, the Monte Carlo sampling splits by a hierarchical annealing process into the essential regions of the state space; therefore UCMC aims at avoiding the typical metastable behavior of Monte Carlo techniques. From the viewpoint of Monte Carlo, a slowly converging long-time Markov chain is replaced by a limited number of rapidly mixing short-time ones. The correct weighting factors for the various Markov chains are obtained via a coupling matrix, that connects the samplings from the different almost invariant sets. The underlying mathematical structure of this approach is given by a general examination of the uncoupling-coupling procedure. Furthermore, the overall algorithmic scheme of UCMC is applied to the n-pentane molecule, a well-known example from molecular dynamics.

Keywords: almost invariant sets, bridge sampling, cluster analysis, hierarchical annealing, Markov chains, Monte Carlo, n-pentane molecule, ratio of normalizing constants, reweighting, uncoupling-coupling.

Mathematical Subject Classification: 60J22, 65C05, 65C40, 82B80

1 Introduction

A large class of problems can be stated as the computation of expectation values $\mathrm{E}_f(g) = \int g(x) f(x)\, dx$ of a function or observable g w. r. t. a density f. Most common are problems from statistical physics [8] and Bayesian inference arising from statistical modeling [15]; in both cases f has the form of a Gibbs–Boltzmann distribution. The widely used Markov chain Monte Carlo (MCMC) methodology provides a flexible and general framework for approximations of such expectation values by averaging over the realization of an appropriate Markov chain generated by the Monte Carlo algorithm.

The main reasons of the popularity of the MCMC approach is that the convergence rate of averages is independent of the high-dimensionality of the state

*Freie Universität Berlin, Institut für Mathematik I, Arnimallee 2–6, 14195 Berlin, Germany. Email: alexander.fischer@math.fu-berlin.de; Internet: http://www.math.fu-berlin.de/~afischer

space and that at no point knowledge about the normalizing constant of a given unnormalized density $\hat{f}$ is needed. However, usually one has to tackle the *trapping problem*, i.e., the Markov chain resides for a very long time in one part of the state space before it moves on to another part. Such undesirable behavior of the Markov chain is caused by *metastable regions*—also called *modes* or *conformations*—in the state space. Because rare transitions between modes result in a very poor convergence rate of averages to the analytic expectation values, there exists a huge literature addressing the trapping problem; either by some sophisticated update schemes for more or less special application classes [1, 7, 8] or by general extensions of the MCMC methodology as multigrid Monte Carlo [16], reweighting techniques [9], simulated tempering in statistical physics [21], reversible jump algorithms in Bayesian inference [17] or macrostate dissection [3].

Herein, by the uncoupling-coupling scheme, we propose another extension of the MCMC methodology: the key idea of uncoupling-coupling Monte Carlo (UCMC) is to regard metastable regions in the state space as *almost invariant sets* w. r. t. some Markov operator corresponding to the Markov chain. It was recently shown by DELLNITZ/JUNGE [5] and SCHÜTTE ET AL. [27] that these almost invariant sets are strongly connected to the spectral structure of the Markov operator, and that it is even possible for a wide range of problem classes to identify almost invariant sets by solving the eigenvalue problem [6]. Remarkably, by this approach we have a dynamical interpretation of metastability depending on the Markov chain and not as usual a geometric one based solely on the density. Therefore, with the information about the almost invariant sets we simultaneously possess the knowledge how to decompose the slowly mixing Markov chain into its rapidly mixing parts.

In most cases, we can embed f in a hierarchy of densities $f(\gamma)$, where $f(\gamma)$ results in a smoothed density by varying the parameter γ (e. g., increasing the temperature if we choose γ to be a temperature parameter). Starting with a MCMC scheme, we recursively decompose the state space and restart annealed Markov chains—which are restricted to their almost invariant set—until all almost invariant sets w. r. t. the Markov operator corresponding to f are resolved. Then, the coupling step consists of setting up a coupling matrix C between the almost invariant sets in such a way that the stationary distribution of C contains the correct weighting factors between these sets, which seemingly got lost in the uncoupling step. It will show up that we have to use bridge sampling densities in the uncoupling step for efficient approximations of ratio of normalizing constants [12], which arise as components of C.

The UCMC scheme combines aspects from simulated annealing approaches in optimization [19], aggregation-disaggregation techniques evolved from the SIMON-ANDO theory [28] and stochastic complementation techniques investigated by MEYER [23] on finite state space Markov chains. A hierarchical annealing structure is also used by CHURCH ET AL. in the macrostate dissection approach for thermodynamical integrals [3]; there the strategy is followed to dissect the state space in a geometric oriented manner in terms of Gaussian integration kernels.

A crucial part of the algorithm is to find a decomposition in a not a-priori known number of almost invariant sets. Typically the number of almost invariant sets is small for a wide class of problems even for high-dimensional problems, e. g., for biomolecules [1]. In that case, we can obtain a decomposition as described in [27] and use an appropriate dynamics-based cluster algorithm as proposed by DEUFLHARD ET AL. [6]. The decomposition is guided solely by the spectral structure of the corresponding Markov operators: This guarantees that the convergence rate increases substantially for subsequent simulation runs, because the second eigenvalue decreases substantially for subsequent simulation runs.

Another feature of UCMC is that due to the different Markov chains emerging in UCMC, an implementation of it is well suited for parallel computation. In contrast to simply multiplying the number of simulations for the Markov chain, we herein can distribute all the various rapidly converging chains emerging from the uncoupling steps during the annealing procedure.

This paper is organized as follows: Section 2 starts with a short overview of MCMC. In Section 3 we investigate the relation between almost invariant sets of a given MCMC method and the spectral structure of the corresponding Markov operator. It is also described, how a clustering of the state space into almost invariant sets can be performed. Then in Section 4 the structure of the coupling matrix is given. To apply the uncoupling-coupling step effectively in an algorithmic scheme, we embed in Section 5 the uncoupling step in a hierarchical structure, which will led us naturally to bridge sampling techniques. An application of UCMC to the n-pentane molecule is given in the last section.

2 Markov Chain Monte Carlo

The goal of MCMC methods is to sample from a probability measure μ and use the output of a Markov chain to compute expectation values w. r. t. that measure. To set the notation, let $(\Omega, \mathcal{B}, \lambda)$ be the underlying measure space and μ a probability measure on $(\Omega, \mathcal{B})$. We suppose in the following, that λ is the Lebesgue measure on $\Omega \subseteq \mathbf{R}^d$, and that μ possesses a density

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

with $f > 0$ where dx denotes integration w. r. t. the Lebesgue measure λ.

In most cases, f is defined in terms of an *unnormalized density* $\hat{f}$ via

$$f(x) = \frac{\hat{f}(x)}{\int_\Omega \hat{f}(x)\, dx},$$

where $Z_{\hat{f}} = \int_\Omega \hat{f}(x)\, dx$ denotes the *normalizing constant* of $\hat{f}$.

A *transition kernel*[1] $K : \Omega \times \mathcal{B} \to [0, 1]$ defines a (homogeneous) Markov chain $\mathcal{X} = (X_n)_{n \geq 0}$ through the relation

$$P\{X_{n+1} \in A | X_0, \ldots, X_n\} = K(X_n, A)$$

[1]For an exact definition of commonly used terms as *transition kernel*, *irreducibility* or *aperiodicity*, we refer to the monograph [24].

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

$$\int_A f(x)\,dx = \int_\Omega K(x, A) f(x)\,dx \tag{1}$$

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

In the Metropolis-Hastings algorithm a transition kernel K which satisfies (1) is realized by first defining an arbitrary but *irreducible* transition kernel

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

together with the *acceptance function*

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

In α only ratios of the form $f(y)/f(x)$ have to be computed, which is feasible even if the normalizing constant Z_f is unknown.

Together with Q and α we can define K by

$$K(x, dy) = k(x, y)\,dy + r(x)\delta_x(dy),$$

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

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

With this K one step in the realization of the Markov chain from the state $X_k = x$ consists of proposing y distributed according to $q(x, y)$, and accept this step—that is set $X_{k+1} = y$—with probability $\alpha(x, y)$; otherwise the proposal is rejected and X_{k+1} is set to x.

The construction of K guarantees that $\mathcal{X}$ is irreducible, provided that Q is irreducible. If we further assume that $\mathcal{X}$ is aperiodic—which is guaranteed whenever proposals get rejected during the sampling process—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

$$f(x)\,k(x, y) = f(y)\,k(y, x) \tag{2}$$

holds (for details, see e. g. [29]). Due to (2) K is called *reversible* w. r. t. f.

A realization $\{x_k\}$ of $\mathcal{X}$ now enables us to calculate expectation values

$$\mathrm{E}_f(g) = \int_\Omega g(x) f(x)\,dx$$

w. r. t. f by using the estimator

$$\mathrm{E}_f^N(g) = \frac{1}{N}\sum_{k=1}^{N} g(x_k), \tag{3}$$

which converges to $\mathrm{E}_f(g)$ for $N \to \infty$ by the strong law of large numbers for Markov chains. Altogether, we can say that MCMC is a method that allows to obtain samples from f without knowledge of the normalizing constant $Z_{\hat{f}}$.

In the following we want to understand the global behavior of a Markov chain via the spectral structure of its associated *propagator* (also called *transfer operator*). This propagator $P : L^s \to L^s$ for $s \in \{1, 2\}$ with $u \mapsto Pu$ is defined in terms of the transition kernel K by

$$Pu(y) = \int_\Omega k(x, y)u(x)\,dx + r(y)u(y), \tag{4}$$

where $L^s = \{u : \Omega \to \mathcal{C} \mid \int_\Omega |u(x)|^s\,dx < \infty\}$. Instead of generating a series of states as the Markov chain $\mathcal{X}$ does, P describes the propagation of densities.

Two important properties of the propagator are [26, 27]:

(i) P is a Markov operator: for all $u \in L^1$ 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 w. r. t. a weighted scalar product $< \cdot, \cdot >_\mu$ due to the reversibility of K.

From (i) and (ii) follows that in L^2 the spectrum $\sigma(P)$ of P is real and bounded by 1, so we have $\sigma(P) \subseteq [-1, 1]$. Similar to the well-known Frobenius–Perron theorem for countable state spaces, the irreducibility and aperiodicity of K implies that P has a simple eigenvalue $\lambda_1 = 1$, for which f is an eigenfunction, i. e. $Pf = f$.

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

If $r_{\mathrm{ess}} < 1$, the Markov chain $\mathcal{X}$ is called uniformly ergodic [24], which is a desirable property for the rate of convergence of the MCMC algorithm. More precisely, let the eigenvalues of $\sigma_{\mathrm{discr}}(P)$ be ordered due to their modulus, i. e., $\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq \ldots$ and assume, that $r_{\mathrm{ess}} < |\lambda_2|$. Then $\mathcal{X}$ converges with geometric rate $|\lambda_2|$; the more $|\lambda_2|$ is bounded away from $\lambda_1 = 1$, the faster $\mathcal{X}$ produces a good sampling of the density f. If the above assumption about r_{ess} holds, a Markov chain $\mathcal{X}$ is called *rapidly mixing*, if $\lambda_2 \ll 1$.

3 The Uncoupling Step

Our aim in this section is to replace a slowly mixing Markov chain $\mathcal{X}$ by n rapidly mixing Markov chains $\mathcal{X}_1, \ldots, \mathcal{X}_n$. In this sense, the uncoupling step refers to decomposing the state space Ω into almost invariant sets $A_1, \ldots, A_n$ and perform restricted samplings in these sets by restarted Markov chains.

3.1 Almost Invariant Sets

A typical but undesirable behavior of a realization $\{x_k\}$ of a Markov chain $\mathcal{X}$ is to remain for a long time in a *metastable* region—also *mode* or *conformation*—of the state space, before it moves on to another one. The convergence properties of expectation values are then mainly dependent on rare transitions between metastable regions of the state space, whereas in these regions $\mathcal{X}$ is rapidly mixing.

This behavior can be understood via the concept of *almost invariant sets* [5, 27]. For measurable sets $A, B \subseteq \Omega$ the transition probability between A and B is given by

$$w(A, B, P) = \frac{1}{\int_A f(x)\,dx} \int_A P(x, B) f(x)\,dx,$$

which can be interpreted as the probability to move from the set A under the propagator P to the set B. With this definition, clearly $w(\Omega, \Omega, P) = 1$, i. e., all probability remains in the state space Ω. We will denote a set A as almost invariant w. r. t. P, if $w(A, A, P) \approx 1$.

A computational approach to the identification of almost invariant sets by means of a discretized eigenvalue problem was proposed by DELLNITZ/JUNGE [5] in the context of low-dimensional dynamical systems; there the algorithm was based on multilevel techniques for a discrete approximation of the so-called *Frobenius–Perron* operator. The underlying concept was reformulated by SCHÜTTE ET AL. [27] for high-dimensional Hamiltonian systems, as they occur in molecular dynamics. In this approach, an approximation of almost invariant sets is obtained by the simulation of a *hybrid Monte Carlo* method and a discretization of the corresponding propagator in *essential degrees of freedom*. This is exactly the approach we follow in the UCMC algorithm to detect almost invariant sets of propagators associated to more general Markov chains.

Crucial to these investigations is that all propagators are Markov operators which allow to link the existence of almost invariant sets to the occurrence of eigenvalues near the unique eigenvalue $\lambda_1 = 1$ in the spectra of the Markov operators. Under this viewpoint we can say, that a Markov chain is slowly mixing due to the existence of at least one almost invariant set corresponding to the 2nd eigenvalue λ_2; at least this is true, as far as λ_2 is actually close to 1 and not located elsewhere near the unit circle. It is also shown, that eigenvalues anywhere near the unit circle correspond to an *almost cyclic behavior* [5]. We will not pursue the existence of almost cyclic behavior further, first because our Markov operators are symmetric and possess therefore real spectra and second the typical spectral structure of the Markov operators arising from applications do not have eigenvalues near -1 but rather a well-separated cluster of eigenvalues near 1.

3.2 Cluster Analysis

Suppose, that $\{x_k\}$ is the output of a Markov chain $\mathcal{X}$ corresponding to a propagator P and that we are interested in detecting almost invariant sets

$A_1, \ldots, A_n$ of P, where the number n of almost invariant sets is not specified in advance.

A discrete approximation of P can be obtained by discretizing the in general high-dimensional state space Ω in its essential coordinates; loosely speaking these are the coordinates which mainly governs the dynamics of P. By this coarse graining we arrive at a decomposition of Ω in $\mathcal{B} = (B_1, \ldots, B_N)$ boxes, from which we can set up a stochastic transition matrix $T \in \mathrm{Mat}_{N \times N}$ by simply counting the transitions from B_k to B_l for consecutive states $(x_k \in B_k, x_{k+1} \in B_l)$ of $\mathcal{X}$ (for details see [27]). Furthermore, the symmetry of the propagator P defined in (4) can be inherited to T, so that we end up with T as a reversible stochastic matrix.

The existence of almost invariant sets and its relation to the spectral structure is already treated in the SIMON-ANDO theory for finite Markov chains and investigated there under the term *nearly completely decomposable systems* [23, 28]. However, therein only such systems are studied for which a suitable decomposition is known in advance. In contrast, an algorithmic approach has to determine the almost invariant sets for given eigenvalues and corresponding eigenvectors.

This is done in an identification algorithm proposed by DEUFLHARD ET AL. [6] for reversible stochastic matrices. Depending on a spectral gap, which separates a cluster of eigenvalues $\lambda_1 = 1, \lambda_2, \ldots, \lambda_n$ in the neighborhood of 1, the number n of almost invariant sets for the decomposition is determined. By exploiting the sign structure in the eigenvectors corresponding to these n eigenvalues, an appropriate assignment of boxes $B \in \mathcal{M}$ to almost invariant sets A_i for $i = 1, \ldots, n$ is realized. The algorithm presented therein is justified by an perturbation analysis, which makes use of the reversibility of T. All in all, the algorithm clusters the set $\mathcal{M}$ of boxes into almost invariant sets $A_i = \cup_{k \in I_i} B_k$ for $i = 1, \ldots, n$.

It would be desirable that the essential spectral radius $r_{\mathrm{ess}}(P)$ is sufficiently bounded away from 1; at least $r_{\mathrm{ess}}(P)$ should be less than λ_n, otherwise some of the eigenvalues from the discrete spectrum $\sigma_{\mathrm{discr}}(P)$ could get interfered by eigenvalues in T, which emerge as a discretization of $\sigma_{\mathrm{ess}}(P)$. Theoretical investigations have already shown, that $r_{\mathrm{ess}}(P)$ is bounded away from 1 for some special classes of problems [26]. However, whether there exists actually an inference problem with $\sigma_{\mathrm{ess}}(P)$, has yet to be investigated.

By this identification algorithm, a dynamical clustering of the state space is performed in a way that the resulting decomposition depends on the dynamics of the Markov chain and reflects therefore our goal to detect sets A_i with $w(A_i, A_i, P) \approx 1$. This is different from the vast majority of cluster algorithms, where a geometric clustering of the data—in our case the states $\{x_k\}$—is obtained without using any transitions between these states [18].

3.3 Restricted Sampling

Ending up with the sets $A_1, \ldots, A_n$ as the output of the cluster algorithm we now want to sample separately in each A_l, for $k = 1, \ldots, n$. Therefore, for each

l we define a restricted Markov kernel K_l from K on A_l by setting

$$K_l(x, dy) = k_l(x, y)\mu(dy) + r_l(x)\delta_x(dy) \tag{5}$$

with

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

and

$$r_l(x) = 1 - \int k_l(x, y)\, dy.$$

Clearly, the detailed balance condition still holds, so that K_l is again a reversible Markov kernel. Now, let $\hat{f}_l = \mathbf{1}_{A_l}\hat{f}$ be the restricted unnormalized density on A_l, with $\mathbf{1}_A$ denoting the indicator function on A, i.e., $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ otherwise. Then, under the assumption, that K_l is irreducible, $f_l = \hat{f}_l / Z_{\hat{f}_l}$ is the unique invariant density of K_l. Note, that the question, whether K_l inherits the irreducibility from K, depends on both, the set A_l and the Markov kernel K. Altogether, for K_l being irreducible, a sampling $\{x_i^l\}$ from the restricted Markov chain $\mathcal{X}_l$ is obtained by starting at a point $X_0^l = x \in A_l$ and performing the same update procedure as for $\mathcal{X}$ with the exception that proposals $y \neq A_l$ are rejected.

As defined in equation (4), we denote by P_l the corresponding propagator of K_l. If we assume that A_l is almost invariant and that it cannot be subdivided further into two or more almost invariant sets, then we can state the following: First, P_l is irreducible, otherwise there would exist a decomposition into two or more invariant subsets. Second, the 2nd eigenvalue λ_2 of P_l is substantially less than 1, otherwise there would exist a decomposition into two or more almost invariant subsets. Third, as a consequence, due to $\lambda_2 \ll 1$, the corresponding Markov chain $\mathcal{X}_l$ is rapidly mixing.

4 The Coupling Step

In the coupling step we will show that it is possible to regain information about a global density $f = \sum_{k=1}^N \pi_k f_k$ in terms of densities f_k, by setting up a coupling matrix C with π as its stationary distribution. This together with the decomposition from the uncoupling step allows us to formulate the algorithmic hierarchical annealing scheme in the next section.

Now suppose that arbitrary unnormalized densities $\hat{f}_1, \ldots, \hat{f}_N$ are given. We denote by $A_k = \mathrm{supp}(f_k)$ the support in the state space Ω and by $\phi_{ij} = \mathbf{1}_{A_i \cap A_j}$ the common support of the densities f_i and f_j.

To obtain information about the density f corresponding to the global unnormalized density $\hat{f} = \sum_{k=1}^N \hat{f}_k$, it is sufficient to know the ratios of normalizing constants $\pi_k = Z_{\hat{f}_k} / Z_{\hat{f}}$, because then we can reconstruct f from the f_k's by

$$\sum_{k=1}^N \pi_k\, f_k = \sum_{k=1}^N \frac{Z_{\hat{f}_k}}{Z_{\hat{f}}} \frac{\hat{f}_k}{Z_{\hat{f}_k}} = \frac{\hat{f}}{Z_{\hat{f}}} = f.$$

Having in mind an algorithmic approach, we have to compute the π_k's (or at least approximations of them) without directly referring to $Z_{\hat{f}}$. This resembles the standard MCMC method, where one avoids the normalizing constant by evaluating ratios depending only on the unnormalized density. In the same way we define the *coupling matrix* $C = (c_{ij}) \in \mathrm{Mat}_{N\times N}$ by

$$c_{ij} = \begin{cases} \frac{1}{N} \frac{Z_{\phi_{ij}\hat{f}_i}}{Z_{\hat{f}_i}} \min\left(1, \frac{Z_{\phi_{ji}\hat{f}_j}}{Z_{\phi_{ij}\hat{f}_i}}\right) & \text{for } i \neq j \text{ and } \mu(A_i \cap A_j) > 0 \\ 0 & \text{for } i \neq j \text{ and } \mu(A_i \cap A_j) = 0 \\ 1 - \sum_{k=1(k\neq i)}^{N} c_{ik} & \text{else} \end{cases} \tag{6}$$

Obviously, C is a stochastic matrix, because for $i \neq j$ we have $0 \leq c_{ij} \leq 1/N$, while due to the diagonal entries the sum of each row is 1. The Markov chain corresponding to C is also aperiodic, simply because $c_{ii} \geq 1/N$ for each diagonal entry.

Furthermore, let us assume in the following that each A_i is connected to any A_j in the sense that there exists a sequence of sets $A_i = A_{m_1}, A_{m_2}, \ldots, A_{m_{k-1}}, A_j = A_{m_k}$, such that $\mu(A_{m_l} \cap A_{m_{l+1}}) > 0$ for $l = 1, \ldots, k-1$. Then for all i and j there exists a path from the state i to the state j in C, which makes C irreducible (this condition will get clear in Section 5, see also Fig. 2).

The key point in the construction of C is that

$$\pi = \frac{1}{Z_{\hat{f}}}(Z_{\hat{f}_1}, \ldots, Z_{\hat{f}_n})$$

is the unique stationary distribution due to the aperiodicity and irreducibility of C. This follows immediately from the detailed balance condition

$$\pi_i \frac{Z_{\phi_{ij}\hat{f}_i}}{Z_{\hat{f}_i}} \min\left(1, \frac{Z_{\phi_{ji}\hat{f}_j}}{Z_{\phi_{ij}\hat{f}_i}}\right) = \pi_j \frac{Z_{\phi_{ji}\hat{f}_j}}{Z_{\hat{f}_j}} \min\left(1, \frac{Z_{\phi_{ij}\hat{f}_i}}{Z_{\phi_{ji}\hat{f}_j}}\right), \tag{7}$$

which moreover shows that C is reversible.

Actually, this is exactly the way a transition kernel for the Metropolis-Hastings algorithm is constructed. Of course, from an algorithmic viewpoint we do not intend to perform a Monte Carlo sampling with C, but rather compute directly the stationary distribution from an appropriate approximation.

There are several remarks to make concerning the setup of C: First, for an approximation $\tilde{C}$ of C, the most problematic parts are the components c_{ij} corresponding to the ratio of normalizing constants $Z_{\phi_{ij}\hat{f}_i}/Z_{\phi_{ji}\hat{f}_j}$, which as opposed to the normalizing constant itself can at least in principle be approximated efficiently (see Sec. 5.2). In fact, like in the Metropolis algorithm we replace a direct computation of $Z_{\hat{f}}$ by the computation of ratios of normalizing constants between the $Z_{\hat{f}_k}$'s.

Second, we made a canonical choice of ϕ_{ij}, which is well suited for the hierarchical annealing approach, but other choices are possible as long as the detailed balance condition (7) holds. For example, one could incorporate P in

the components c_{ij} of C to reflect the dynamics of P in the sense of a coarse graining of Ω.

Third, if we suppose that we can compute expectation values for each f_k and we know the stationary distribution π of C, we are able to compute expectation values w. r. t. f, which are now given by

$$\mathrm{E}_f(g) = \mathrm{E}_{\sum_k \pi_k f_k}(g) = \sum_k \pi_k \int_{A_k} g(x) f_k(x)\, dx. \tag{8}$$

For finite state problems, MEYER [23] investigates the concept of *stochastic complementation* for a given—not necessarily reversible—irreducible stochastic matrix P'. By aggregation of states a coupling matrix C' is constructed in such a way that the stationary distribution π' of C' contains the correct weighting factors for the aggregates. The problematic part in the setup of a stochastic complement c_{ij}—which are the elements of C'—is, that information from P' is needed from all components of P'. The application in mind by stochastic complementation are the so-called *aggregation-disaggregation* techniques [4, 28], which try to perform a fast computation of the stationary distribution of P' for large finite state spaces, where a suitable decomposition of the state space is known in advance.

In contrast, in our approach it is the restriction to reversible chains and the use of the detailed balance condition for the setup of C, which enables us to make use of a coupling matrix in the context of MCMC, where the identification of the almost invariant sets are an important part of the algorithm. In particular, an entry c_{ij} is defined only in terms of $\hat{f}_i$ and $\hat{f}_j$, no matter about the remaining $\hat{f}_k$'s.

5 Hierarchical Annealing

It is not practical to start the simulation with a Markov chain for the density f of interest, because the Markov chain would immediately get trapped in one of its metastable regions. Actually, it is exactly this Markov chain, that we want to avoid by sampling only in its almost invariant sets. A way out of this problem is to embed f in a family of densities $f(\gamma)$ where $f = f(\gamma)$ is the original density for a fixed γ. By varying γ the densities $f(\gamma)$ transforms into smoother ones with less metastable regions. Often such a family of densities arises naturally, e. g., by varying the temperature parameter—physical or artificial—in the Gibbs-Boltzmann distribution.

Starting with a MCMC scheme for a density f_1 with a suitably chosen γ-parameter, we recursively decompose the state space and restart Markov chains with annealed γ-parameter—which are restricted to their almost invariant set—until we have sampled all almost invariant sets w. r. t. the Markov operator corresponding to f.

A straightforward approach to hierarchical annealing would be to decompose the initial sampling $\{x_k^1\}$ of f_1 on the state space $A_1 = \Omega$ into almost invariant sets $A_2, \ldots, A_l$ by clustering the data points $\{x_k^1\}$, setting up annealed densities

$f_2, \ldots, f_l$ on $A_2, \ldots, A_l$, generate data points $\{x_k^2\}, \ldots, \{x_k^l\}$, and recursively repeat the decomposition for $A_2, \ldots, A_l$ and annealing of the γ parameter until finally the original density f is sampled by the restricted densities $f_{l+m}, \ldots, f_n$ in the almost invariant sets $A_{l+m}, \ldots, A_n$ of f.

Though this approach allows to detect the almost invariant sets $A_{l+m}, \ldots, A_n$ and sample herein, it is not useful in regard to an approximation $\tilde{C}$ of the coupling matrix C, for which we need to extract the correct weighting factors $\pi_{l+m}, \ldots, \pi_n$ for the data points $\{x_k^{l+m}\}, \ldots, \{x_k^n\}$. To cope with this problem, we use *bridge densities*: Instead of sampling the annealed density f_j on the almost invariant set A_j, a bridge density f_{ij} between $\mathbf{1}_{A_j} f_i$ and f_j is constructed, which encompasses both densities. By using bridge densities f_{ij} (see Section 5.1) rather than f_j alone for all subsequent samplings, we will be able to get all the information we need to set up the coupling matrix (see Section 5.2).

By decreasing the parameter γ during the hierarchical annealing process we not only change the densities but simultaneously the corresponding Markov operators and with them the almost invariant sets. The idea of this method is, that the hierarchical decomposition comes along with a hierarchy of almost invariant sets in terms of the almost invariance measure $w(A_k, A_k, P)$ for the original propagator P. In regard to the spectral structure of $P(\gamma)$—the propagator corresponding to $f(\gamma)$—, this behavior is reflected by the increase of some eigenvalues towards 1 in dependence of γ. Actually, by defining restricted Markov operators as in Section 3.3 on the almost invariant sets, these eigenvalues become new eigenvalues $\lambda_1 = 1$ for the restricted operators; one can think of that as transforming almost invariant sets into invariant ones (see Fig. 1).

Another important aspect is the convergence rate of the Markov chains. During the hierarchical annealing various rapidly mixing Markov chains get computed, which have to be stopped automatically after a "reasonable" number N of iterations. Because N can vary drastically from one Markov chain to another depending on the almost invariant set A_k and the corresponding propagator, we use the "non-"convergence estimator from GELMAN and RUBIN [13, 11]. For this estimator, multiple realizations of a Markov chain $\mathcal{X}_k$ are generated to compute estimates depending on the variances between these realizations. There exist other estimators using only one Markov chain (see the discussion in [13]), but this one is especially suited for the UCMC approach, because we can start realizations for a subsequent Markov chain $\mathcal{X}_k$ with starting points already well distributed in A_k. However, in case all realizations get trapped in the same metastable region, like any other estimator this one wrongly indicates a convergence of the Markov chains.

5.1 Bridge sampling

To maintain the connectivity between the Markov chains which emerge during the annealing, samplings from a density f_j must be somehow concatenated to its parent density f_i in the hierarchy. If the densities f_i and f_j are directly sampled, the overlap between these densities is in general to small to extract a statistical reliable approximation of the ratio of normalizing constants from

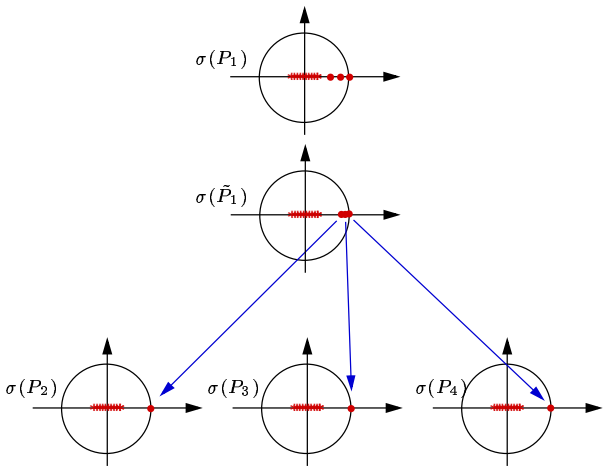

Figure 1: The splitting behavior of the spectral structure of the corresponding Markov operators (see also [23]). Due to the reversibility of the Markov operators all spectra $\sigma(P_k)$ are real. Here a situation for three almost invariant sets is shown. While the 2nd and 3rd eigenvalue of $\sigma(P_1)$ are significantly bounded away from $\lambda_1 = 1$, they move towards 1 when we anneal the γ-parameter without decomposing the state space, what is indicated by $\sigma(\tilde{P}_1)$. The decomposition into three almost invariant sets results in a splitting of these eigenvalues, each of them becomes $\lambda_1 = 1$ for the restricted Markov operators $P_2, \ldots, P_4$.

the sampled data.

Suppose, that A_j is an almost invariant set from the decomposition of the set A_i and f_j is the corresponding annealed new density on A_j, e.g., by decreasing the γ-parameter. Let $f_i^j = \phi_{ij} f_i$ be the restriction to the set $A_i \cap A_j$ from the density f_i on A_i (cf. Fig. 2); as a consequence of the decomposition A_j is a subset of A_i, therefore we have $f_i^j = \mathbf{1}_{A_j} f_i$.

A generic choice for a bridge density on the set A_j is given by

$$\hat{f}_{ij} = \sigma \hat{f}_i + (1 - \sigma) \hat{f}_j^i \tag{9}$$

for some $\sigma \in [0, 1]$.

By mixing both densities into f_{ij}, we expect to satisfy in particular:

(i) The Markov chain $\mathcal{X}_{ij}$ corresponding to f_{ij} is rapidly mixing. This assumption is justified, because A_j is an almost invariant set w. r. t. P_j.

(ii) A simulation run $\{x_l^{ij}\}$ of f_{ij} allows a statistical reasonable reweighting to the densities f_i and f_j; this presupposes that all important parts of the densities f_i and f_j get sampled by $\mathcal{X}_{ij}$.

(iii) As we will see in 5.2, we can use the reweighted data from (ii) to approximate $Z_{\phi_{ij}\hat{f}_i}/Z_{\phi_{ij}\hat{f}_i}$.

However, a suitable choice for σ in (9) has to depend on a specific application as in the numerical example in Section 6.

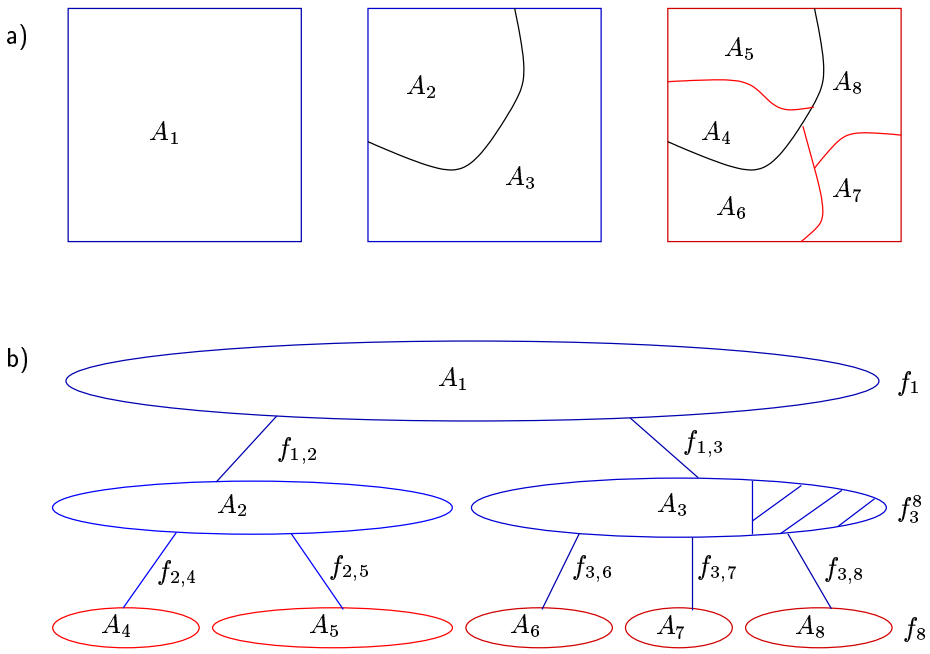

Figure 2: a) Schematic plot of a two level hierarchical decomposition. An initial sampling of f_1 decomposes the state space $\Omega = A_1$ into two subsets A_2 and A_3, which get further subdivided into $\{A_4, A_5\}$ and $\{A_6, A_7, A_8\}$. b) The same subdivision as in a), but now represented as a graph, where nodes correspond to sets A_k. We denote by $\rho(i)$ the index of the *parent set* of a node A_i; therefore edges $(\rho(i), i)$ represent bridge densities $f_{\rho(i)i}$. As an example, for the set A_8 we have $\rho(8) = 3$. The density $f_3^8 = \phi_{3,8} f_3$ corresponds to the hatched part of A_3; in UCMC neither f_3^8 nor f_8 is sampled, but rather the bridge density $f_{3,8}$, which sufficiently encompasses the important parts of f_3^8 and f_8 (cf. Fig. 6). Additionally, the tree structure of the graph guarantees that the coupling matrix C defined in equation (6) is irreducible, because an edge $(\rho(i), i)$ implies the entries $c_{\rho(i)i}$ and $c_{i\rho(i)}$ of C to be non-null. A shrinkage of the almost invariant sets is indicated by the ellipses getting smaller during annealing (cf. Fig. 4).

There exists also more theoretical investigations about bridge densities [12, 14, 22], yet with a slightly different point of view: bridge densities are primarily used to connect simulation runs from two densities and compute the ratio of normalizing constants by an *a-posteriori* data analysis. In contrast, for our purpose it is necessary to determine *a-priori* a bridge density f_{ij} to perform directly a Markov chain simulation in f_{ij}.

5.2 Approximation of the coupling matrix

In Section 4 the general form of the coupling matrix C was given. With the usage of bridge densities, we are now able to compute an approximation $\tilde{C} =$

$(\tilde{c}_{ij})$ of the non-zero entries

$$c_{ij} = \frac{1}{N} \frac{Z_{\phi_{ij}\hat{f}_i}}{Z_{\hat{f}_i}} \min\left(1, \frac{Z_{\phi_{ji}\hat{f}_j}}{Z_{\phi_{ij}\hat{f}_i}}\right) \qquad \text{for} \quad i \neq j \quad \text{and } \mu(A_i \cup A_j) > 0$$

of the hierarchical annealing process.

Note, that as a consequence of the hierarchical subdivision, only non-diagonal entries c_{ij} of C linking a density f_i to its "child" densities or "parent" density $f_{\rho(i)}$ are non-zero (see Fig. 2).

Let $\mathcal{S}_{ij} = \{x_k^{ij} | k = 1, \ldots, N_{ij}\}$ be the sampled data from the density f_{ij} resulting from the multiple realizations of the Markov chain $\mathcal{X}_{ij}$. We define with $\mathcal{W}_{ij}^l = \{\alpha_k^{l,ij} | k = 1, \ldots, N_{ij}\}$ for $l \in \{i, j\}$ data sets containing weighting factors for an unnormalized density $\phi_{ij}\hat{f}_l$ by reweighting the data points in $\mathcal{S}_{ij}$ with the reweighting formula [9]

$$\alpha_k^{l,ij} = \frac{\phi_{ij}(x_k^{ij})\hat{f}_l(x_k^{ij})/\hat{f}_{ij}(x_k^{ij})}{\sum_{m=1}^{N_{ij}} \phi_{ij}(x_m^{ij})\hat{f}_l(x_m^{ij})/\hat{f}_{ij}(x_m^{ij})}. \tag{10}$$

Because our sampling $\mathcal{S}_{ij}$ was already restricted to $A_i \cap A_j$, the indicator function ϕ_{ij} is always 1—i. e. $\phi_{ij}(x_k^{ij}) = 1$ for all k—and could therefore be omitted from formula (10). As a special case, for the initial sampling $\mathcal{S}_{11} = \{x_k^1 | k = 1, \ldots, N_{11}\}$ no reweighting is needed, so $\alpha_k^{11} = 1/N_{11}$ for all $k = 1, \ldots, N_{11}$.

Now let f_j be the child density of f_i and f_l the parent density of f_i, i. e. $\rho(j) = i$ and $\rho(i) = l$, respectively. Then the expectation value $\mathrm{E}_{f_i}(\phi_{ij}\hat{f}_i) = Z_{\phi_{ij}\hat{f}_i}/Z_{\hat{f}_i}$ can be approximated due to

$$\frac{Z_{\phi_{ij}\hat{f}_i}}{Z_{\hat{f}_i}} = \lim_{N_{il}\to\infty} \sum_{k=1}^{N_{il}} \alpha_k^{i,il} \phi_{ij}(x_k^{il}) \tag{11}$$

by means of the weighted data set of $\mathcal{S}_{il}$ and $\mathcal{W}_{il}$; this ratio can be interpreted as the probability to be in the set A_i w. r. t. f_i.

Similarly, we can approximate the ratio of normalizing constants solely by the data points $\mathcal{S}_{ij}$ from the bridge density f_{ij}, and the reweighting factors $\mathcal{W}_{ij}^i$ and $\mathcal{W}_{ij}^j$ corresponding to the densities f_i and f_j, respectively:

$$\frac{Z_{\phi_{ji}\hat{f}_j}}{Z_{\phi_{ij}\hat{f}_i}} = \lim_{N_{ij}\to\infty} \frac{\sum_{k=1}^{N_{ij}} \alpha_k^{j,ij} \hat{f}_j(x_k^{ij})/\hat{f}_{ij}(x_k^{ij})}{\sum_{k=1}^{N_{ij}} \alpha_k^{i,ij} \hat{f}_i(x_k^{ij})/\hat{f}_{ij}(x_k^{ij})}. \tag{12}$$

For finite N_{ij} one needs a reliable sampling as described in Section 5.1 to obtain a satisfactory approximation.

With (11) and (12) an irreducible approximation $\tilde{C}$ of C is given, of which we can directly compute its unique invariant distribution $\tilde{\pi}$. In general the matrix $\tilde{C}$ will not be exactly reversible, but it converges to the reversible matrix C for $N_{ij} \to \infty$ for all i, j which come into question. In case $\tilde{\pi}$ shows up to be complex—a case that never occurred until now while testing—one could use the real part of $\tilde{\pi}$ and additionally verify the quality of the approximation.

From $\tilde{\pi}$ we compute the reweighted components

$$\eta_l = \frac{\tilde{\pi}_l}{\sum_{k=m}^{N} \tilde{\pi}_k} \quad \text{for } l = m, \ldots, N \tag{13}$$

for the densities $\hat{f}_m, \ldots, \hat{f}_N$ which are a decomposition of the original density $\hat{f}$.

With (13) and the corresponding data sets $\mathcal{S}_{ij}$ and weighting factors $\mathcal{W}_{ij}^i$ we finally can replace the estimator given in (3) for computing $\mathrm{E}_f(g)$ by

$$\mathrm{E}_f^{N\Delta}(g) = \sum_{i=m}^{N} \eta_i \left(\sum_{k=1}^{N_{i\rho(i)}} g\left(x_k^{i\rho(i)}\right) \alpha_k^{i\rho(i)} \right), \tag{14}$$

where Δ is the index set over all pairs $(i, \rho(i))$ of bridge densities in question. Actually, this is a discrete counterpart of equation (8).

6 Example: The n-pentane molecule

Biomolecules are an important application class of MCMC methods in statistical physics. There exists a wide range of MCMC algorithms, which are trying to tackle the problems and challenges of biomolecular systems [1]. Biomolecules are well suited for the UCMC approach, because they possess in general only a small number of *conformations*, i. e., the most metastable 3-dimensional structures of a molecule, which can be thought of as almost invariant sets for a suitable Markov operator. Herein, we will test UCMC with the *n-pentane* molecule, which is a well-known test system for various MCMC algorithms [20]. It is represented by a *separated Hamiltonian* $\mathcal{H}(p,q) = \mathcal{T}(p) + \mathcal{V}(q)$, where the kinetic energy $\mathcal{T}(p)$ depends only on the momenta p and the potential energy $\mathcal{V}(q)$ only on the coordinates q. The United-Atoms representation of Ryckaert and Bellemans [25] is used to set up the Hamiltonian. Although the potential $\mathcal{V}$ is 15-dimensional, the overall structure and dynamic of n-pentane is mainly determined by its two torsion angles (see Fig. 3).

Our goal is to sample from the canonical density restricted to the potential $\mathcal{V}$, therefore the coordinate space Ω is the state space in the Monte Carlo setting. We have

$$f(q) = \frac{1}{Z_{\hat{f}}} \exp[-\beta\, \mathcal{V}(x)] \qquad \text{with} \quad \beta = \frac{1}{k_B T} \quad \text{and} \quad Z_{\hat{f}} = \int \exp[-\beta\, \mathcal{V}(x)]\, dx,$$

where k_B is Boltzmann's constant and T the temperature of the system. An embedding of f in a hierarchy of densities is given directly by the inverse temperature parameter β (cf. Fig. 4).

For the MCMC sampling we use hybrid Monte Carlo (HMC) [2, 7], which became a widely used method over the last decade for computing expectation values (mainly thermodynamic observables) in molecular systems [1]. In HMC, a proposal x' from the actual state x is obtained by generating random momenta p proportional to the canonical distribution of $\mathcal{T}$ and integrating the system

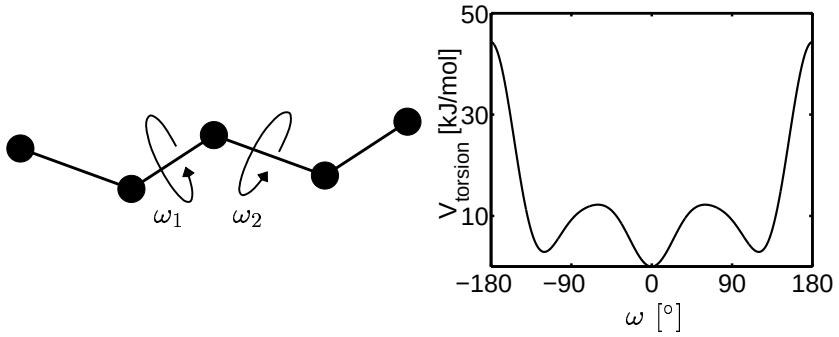

Figure 3: Left: United-Atom model due to [25] of the n-pentane molecule $CH_3 - CH_2 - CH_2 - CH_2 - CH_3$ with its two essential degrees of freedom, the torsion angles ω_1 and ω_2. Right: The potential part $\mathcal{V}_{\text{tor}}$ of a torsion angle. Each $\mathcal{V}_{\text{tor}}$ possesses three clearly distinct energy minima.

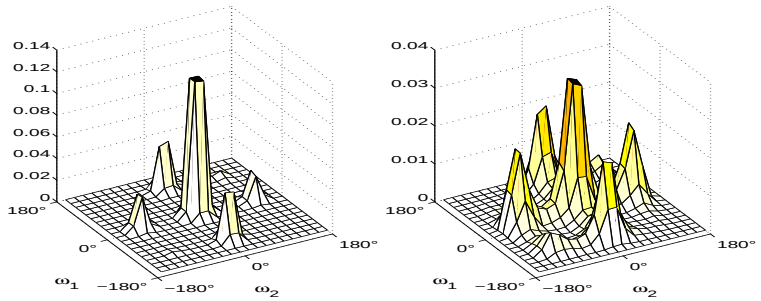

Figure 4: Invariant densities of n-pentane from long simulation runs at $T = 200\,K$ (left) and $T = 600\,K$ (right). Note that due to the projection on the torsion angles, it still can be the case that the overlap between the two densities is small. The form of the invariant densities are in good accordance with the torsion angle potential (cf. Fig. 3); the asymmetry in the four corners of the plot results from the influence of a Lennard-Jones term in $\mathcal{V}$.

	$T_1 = 600\,K$	$T_2 = 400\,K$	$T_3 = 200\,K$	$T_4 = 100\,K$
λ_1	1.000	1.000	1.000	1.000
λ_2	0.924	0.972	0.999	0.331
λ_3	0.887	0.960	0.998	0.124
λ_4	0.878	0.957	0.997	-0.011
λ_5	0.854	0.938	0.996	0
λ_6	0.782	0.882	0.991	0
λ_7	0.760	0.876	0.985	0
λ_8	0.494	0.545	0.307	0
λ_9	0.310	0.502	0.151	0
HMC steps	10^4	3×10^4	8×10^4	2×10^5

Table 1: Eigenvalues of the HMC operator in dependence of the temperature. The lengths of the simulation runs result in roughly the same convergence estimate; except at $100\,K$, where the simulation run suffers from critical slowing down.

through the enhanced space—state space plus momenta space—by computing a short term trajectory $(x', p') = (\Psi^\tau)^n(x, p)$ with a discrete flow Ψ^τ with time step τ from the canonical equations $\dot{q} = \partial\mathcal{H}/\partial p$, $\dot{p} = -\partial\mathcal{H}/\partial q$ (for details, see [2]). Because HMC makes use of the canonical equations, i. e. the "physical" dynamics corresponding to the Hamiltonian $\mathcal{H}$, almost invariant sets of the HMC transfer operator are reflecting conformations of the molecule. For further connections between these two concepts we refer the reader to the work of SCHÜTTE ET AL. [27].

For the bridge sampling method between two adjacent temperatures we use *adaptive temperature* HMC (ATHMC) [10], an enhancement of HMC. In ATHMC first a bridge density is generated and then sampled in the same manner as in HMC by adapting the temperature for the proposal step according to the actual potential energy. ATHMC allows to sample from an energy range, which is just broad enough to encompass the two adjacent state space densities.

For comparison with the UCMC algorithm, let us first perform some standard HMC runs at different temperatures and compute the eigenvalues of the corresponding HMC Markov transition operators by a discretization of the state space. We choose a 6×6 Galerkin discretization in the two torsion angles (see Table 1). One can clearly see, how a well-separated eigenvalue cluster is moving towards $\lambda_1 = 1$ by decreasing the temperature from $600\,K$ to $200\,K$; simultaneously the sampling length increases, because the mixing rates for the corresponding Markov chains decreases drastically. At $T = 100\,K$ we are no longer able to sample from the canonical density due to a critical slowing down; the sampling even did not leave the initial almost invariant set during the simulation run.

While at $T = 600\,K$ the second eigenvalue $\lambda_2 = 0.924$ indicates that the corresponding Markov chain is rapidly mixing, at $T = 400\,K$ the subsequent eigenvalues of $\lambda_1 = 1$ are approaching 1, so we choose the HMC simulation run

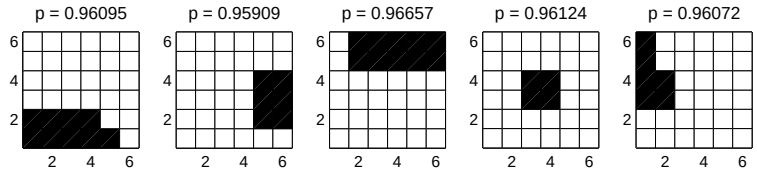

Figure 5: Initial clustering after the first simulation run in $A_1 = \Omega$ at $400\,K$ into 5 almost invariant sets $A_2, \ldots, A_6$. Each torsion angle is uniformly discretized into 6 boxes. The number above each almost invariant set A_k indicates the probability $p = w(A_k, A_k, P_1)$ to stay within these sets. Remember, that for n-pentane the densities are defined on the 15-dimensional state space, and we only visualize the almost invariant sets via projections into the two essential degrees of freedom. Exemplarily, we take a closer look at the set A_4 in Fig. 6.

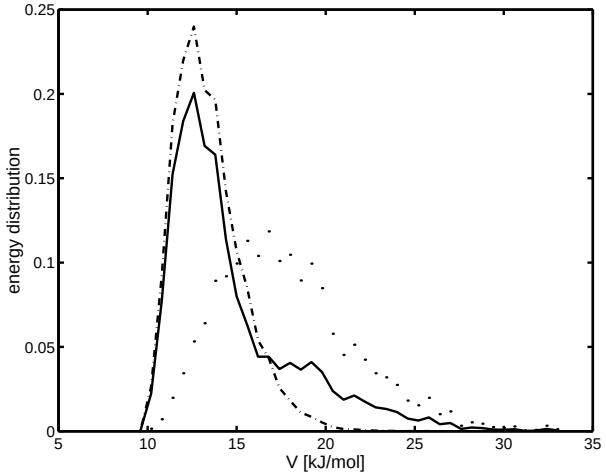

Figure 6: Energy distributions from the simulation run in the set A_4 (the third set from Fig. 5). The energy distribution from the bridge density f_{14} (—) encompasses satisfactorily the two reweighted energy distributions f_1^4 ($\cdots$) and f_4 ($\cdot - \cdot$), respectively. We can see from this simulation run, that the reweighted densities have a small overlap, in spite of f_4 is hierarchically embedded in f_1^4 by the annealing procedure. As to be expected, f_4 is more concentrated in lower energy regions, which can result in a shrinkage of the almost invariant sets during annealing on these low energy regions.

set	A_2	A_3	A_4	A_5	A_6
λ_2	0.968	0.539	0.972	0.342	0.849
λ_3	0.234	0.159	0.749	0.134	0.196
# steps	10^4	5000	10^4	5000	5000

Table 2: Eigenvalues λ_2 and λ_3 of the 5 bridge sampling densities $f_{1,2}, \ldots, f_{1,6}$. While A_3, A_5 and A_6 are rapidly mixing, the 2nd eigenvalue of A_2 and A_4 indicate that the corresponding Markov chains are slowly mixing; a further decomposition of these two sets is recommendable.

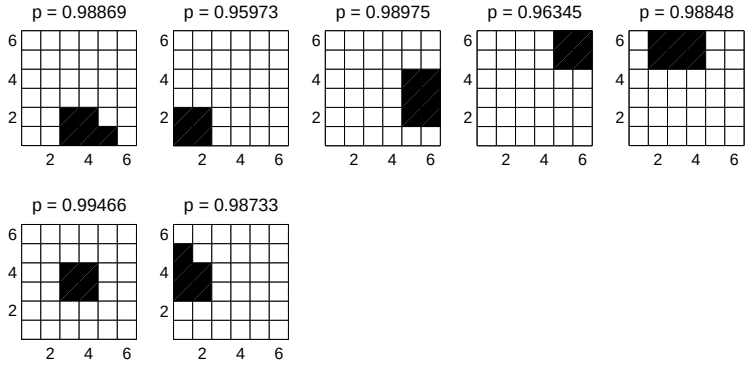

Figure 7: The bridge sampling runs of the 2nd level between $400\,K$ and $200\,K$ leads to a further subdivision of the sets A_2 and A_4 to a total of 7 almost invariant sets, namely $A_7, \ldots, A_{13}$. To take into account the rejections from proposals outside of the set A_k, p denotes here the probability $w(A_k, A_k, P_{\rho(k)k})$ minus the percentage of proposals outside of A_k.

at $T = 400\,K$ from Table 1 as the initial run for UCMC. Our aim is to compute expectation values at $T = 100\,K$. From an initial clustering into 5 almost invariant sets (see Fig. 5)—by interpreting the distance between $\lambda_5 = 0.938$ and $\lambda_6 = 0.882$ as some spectral gap—we generate bridge densities between $400\,K$ and $200\,K$ in these 5 sets. The eigenvalues of these bridge sampling runs are given in Table 2; we may interpret in the sets A_3 and A_4 the neglected eigenvalues $\lambda_6 = 0.882$ and $\lambda_7 = 0.876$ from $T = 400\,K$, which increased to $\lambda = 0.968$ and $\lambda = 0.972$. Due to these eigenvalues samplings within these two sets need the double length of 10^4 ATHMC steps to fulfill the convergence criterium. In addition, the energy distributions for the bridge density and its reweighted densities are given in Fig. 6.

At this stage, we already can make a comparison between the direct HMC sampling at $200\,K$ from Table 1 and an analysis of the bridge sampling densities by setting up an coupling matrix $\tilde{C} \in \mathrm{Mat}_{6\times 6}$ on the sets $A_1, \ldots, A_6$ and compute its stationary distribution π.

set	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
λ_2	0.118	0.111	0.113	0.157	0.133	0.306	0.133
# steps	5000	5000	5000	5000	5000	5000	5000
$\hat{\pi}$	0.0207	3.18×10^{-5}	0.0193	2.83×10^{-5}	0.0199	0.9187	0.0213

Table 3: The bridge sampling densities $f_{2,7}$, $f_{2,8}$, $f_{3,9}$, $f_{4,10}$, $f_{4,11}$, $f_{5,12}$ and $f_{6,13}$ between $200\,K$ and $100\,K$ all yield rapidly mixing Markov chains in the sets $A_7, \ldots, A_{13}$, respectively.

Due to the high-dimensionality of Ω, there are no analytical results available to verify our results, but we can compare the probabilities to stay in symmetric conformations. From the Hamiltonian we know that the analytic values of the probabilities to be within each part of a 3×3 discretization of the torsion angles have the following symmetric entries:

$$S = \begin{pmatrix} d & b & c \\ b & a & b \\ c & b & d \end{pmatrix}.$$

Reweighting the data from the UCMC output and computing approximations of S according to the estimator (14) yields

$$S^{200}_{\mathrm{UCMC}} = \begin{pmatrix} 9.29 \times 10^{-6} & 0.0856 & 0.0017 \\ 0.0872 & 0.6712 & 0.0732 \\ 0.0016 & 0.0794 & 1.02 \times 10^{-5} \end{pmatrix},$$

which is similar to a direct HMC sampling with 5×10^5 steps that results in

$$S^{200}_{\mathrm{HMC}} = \begin{pmatrix} 6.00 \times 10^{-6} & 0.0846 & 0.0024 \\ 0.1064 & 0.6005 & 0.0941 \\ 0.0034 & 0.1105 & 0 \end{pmatrix}.$$

Now, let us proceed the UCMC sampling with a decomposition of the sets A_2 and A_4, which yields a total of 7 sets (see Fig. 7). A bridge sampling between $200\,K$ and $100\,K$ results in a fast convergence for each of the sets $A_7, \ldots, A_{13}$ (see Table 3).

Again, we can set up a coupling matrix $\tilde{C} \in \mathrm{Mat}_{13\times 13}$ and compute the stationary distribution $\tilde{\pi}$ for extracting η as given in (13). At $100\,K$ a direct sampling suffers from critical slowing down (cf. Table 1), in contrast the symmetry of S is still reflected from the reweighted data sets of the UCMC simulation at $100\,K$:

$$S^{100}_{\mathrm{UCMC}} = \begin{pmatrix} 0 & 0.0199 & 2.83 \times 10^{-5} \\ 0.0213 & 0.9187 & 0.0193 \\ 3.18 \times 10^{-5} & 0.0207 & 0 \end{pmatrix}.$$

Moreover, our insight into the simple Hamiltonian of n-pentane allows us to sample on the entire state space from a fine tuned bridge density between the canonical density at $T = 100\,K$ and a modified density—where the crucial torsion angle potential $\mathcal{V}_{\text{tor}}$ is faded out—which yields a good sampling in the $T = 100\,K$ region. From such a Scaled Potential Monte Carlo (SPMC) simulation with 2.5×10^5 HMC steps we get the corresponding reweighted probability matrix

$$S^{100}_{\text{SPMC}} = \begin{pmatrix} 7.45 \times 10^{-11} & 0.0241 & 1.93 \times 10^{-5} \\ 0.0253 & 0.9053 & 0.0239 \\ 2.19 \times 10^{-5} & 0.0214 & 1.07 \times 10^{-10} \end{pmatrix},$$

which can also be used as an estimation for the analytic values. Therefore, S^{100}_{UCMC} seems to reproduce the correct probabilities. Note, that in S^{100}_{UCMC} the entries s_{13} and s_{31} are exactly 0, because these parts of the state space vanished during the hierarchical annealing due to their almost zero weights.

References

[1] B. J. Berne and J. E. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opinion in Struct. Biol.*, 7:181–189, 1997.

[2] A. Brass, B. J. Pendleton, Y. Chen, and B. Robson. Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. *Biopolymers*, 33:1307–1315, 1993.

[3] B. W. Church, A. Ulitsky, and D. Shalloway. Macrostate dissection of thermodynamic Monte Carlo integrals. In D. M. Ferguson, J. I. Siepmann, and D. G. Truhlar, editors, *Monte Carlo Methods in Chemical Physics*, volume 105 of *Advances in Chemical Physics*. J. Wiley & Sons, New York, 1999.

[4] P. J. Courtois. *Decomposability: Queueing and Computer System Applications.* Academic Press, New York, 1977.

[5] M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *SIAM J. Num. Anal.*, 36(2):491–515, 1999.

[6] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.

[7] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–222, 1987.

[8] D. M. Ferguson, J. I. Siepmann, and D. G. Truhlar, editors. *Monte Carlo Methods in Chemical Physics*, volume 105 of *Advances in Chemical Physics*. Wiley, New York, 1999.

[9] A. M. Ferrenberg and R. H. Swendsen. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.*, 63(12):1195–1197, 1989.

[10] A. Fischer, F. Cordes, and C. Schütte. Hybrid Monte Carlo with adaptive temperature in mixed-canonical ensemble: Efficient conformational analysis of RNA. *J. Comput. Chem.*, 19(15):1689–1697, 1998.

[11] A. Gelman. Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 131–143. Chapman & Hall, 1996.

[12] A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185, 1998.

[13] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, 7(4):457–511, 1992.

[14] C. J. Geyer. Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical report, School of Statistics, Univ. Minnesota, 1994. Technical Report 568.

[15] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

[16] J. Goodman and A. D. Sokal. Multigrid Monte Carlo method. Conceptual foundations. *Phys. Rev. D*, 40(6):2035–2071, 1989.

[17] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[18] W. Huisinga, C. Best, R. Roitzsch, C. Schütte, and F. Cordes. From simulation data to conformational ensembles: Structure and dynamic based methods. *J. Comp. Chem.*, 20(16):1760–1774, 1999.

[19] P. J. M. v. Laarhoven and E. H. L. Aarts. *Simulated Annealing: Theory and Applications*. Reidel, Dordrecht, 1987.

[20] Z. Liu and B. J. Berne. Method for accelerating chain folding and mixing. *J. Chem. Phys.*, 99:6071–6077, 1993.

[21] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.*, 19(6):451–458, 1992.

[22] X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica*, 6:831–860, 1996.

[23] C. D. Meyer. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Rev.*, 31:240–272, 1989.

[24] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, Berlin, 1993.

[25] J.-P. Ryckaert and A. Bellemans. Molecular dynamics of liquid alkanes. *Faraday Discuss.*, 66:95–106, 1978.

[26] C. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules*. Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 1998. Available as Report SC-99-18 via http://www.zib.de/bib/pub/pw/.

[27] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.

[28] H. A. Simon and A. Ando. Aggregation of variables in dynamic systems. *Econometrica*, 29(2):111–138, 1961.

[29] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, 22:1701–1762, 1994.