5 Uncoupling-Coupling

In this section we present a detailed examination of the uncoupling-coupling procedure. We first work out the main ideas of the uncoupling and coupling for a decomposition into two metastable sets in Sect. 5.1. We next formalize the two steps of our UC method in a general setting: *uncoupling* in terms of restricted sampling (Sect. 5.2) together with a hierarchical decomposition of the state space by means of annealing (Sect. 5.2.3); and *coupling* via extracting information from simulation runs in bridge densities for regaining coupling factors between decoupled metastable sets (Sect. 5.3). Eventually, in Sect. 5.3.3, we discuss two ways to set up a weighted sample that is distributed according to our target distribution. The resulting algorithm is summarized in Sect. 5.4.

5.1 Bridging the Barrier

Let us examine a decomposition of a state space Ω into two metastable sets A and B. This is the simplest possible case for which UC can be employed and will provide us a better understanding of the leading ideas behind the uncoupling-coupling procedure. To make this more illustrative, we consider in Fig. 17 again the potential $\mathcal{V}$, which we already know from the introductory example (see Fig. 2 (a) on page 7). To set the notation for this example, let $\Omega = A \cup B$ be a decomposition of the state space $\Omega = [-2, 2]$ into the sets $A = [-2, 0]$ and $B = (0, 2]$. Denote by $f_{\text{high}}(\Omega)$ and $f_{\text{low}}(\Omega)$ two canonical densities on Ω corresponding to high and low temperatures T_{high} and T_{low}, respectively. In the same way, we write $f_{\text{high}}(A)$, $f_{\text{high}}(B)$, $f_{\text{low}}(A)$ and $f_{\text{low}}(B)$ for restricted canonical densities on A and B for T_{high} and T_{low}, respectively. Thereby, the term *restricted density* means that the respective unnormalized densities are identical on the restricted set. Since $f(\Omega) = h(\Omega)/Z_{h(\Omega)}$, where $h(\Omega)$ denotes the unnormalized density, we write $h(A) = \mathbf{1}_A h(\Omega)$. Although $\mathcal{V}$ is one-dimensional in our example, nothing prevents us in the following from thinking of the two sets A and B as a decomposition of a high dimensional state space.

Suppose, we are interested in expectation values with respect to the target distribution $f_{\text{low}}(\Omega)$ or simply in obtaining a (possibly weighted) data set distributed according to $f_{\text{low}}(\Omega)$. Let us suppose that all parameters of a MCMC sampler are fixed, so that the notion of a density f sufficiently characterizes the resulting Markov chain. This enables us to write $\mathcal{X}(f) := \mathcal{X}_f = X_f^{(1)}, X_f^{(2)}, \ldots$ for a Markov chain corresponding to this MCMC sampler. In our example, we use the HMC algorithm as MCMC sampler, with fixed trajectory length and internal step size of the Verlet integrator.

As we have seen in Fig. 2 (c), drawing samples directly from $f_{\text{low}}(\Omega)$ via a realization of $\mathcal{X}_{f_{\text{low}}(\Omega)}$ is much too slow to get a satisfiable result; this

was reflected by a 2nd eigenvalue of 0.9988 of $P(\mathcal{X}_{f_{\text{low}}(\Omega)})$. Yet, we have also seen in Fig. 2 (d) that drawing samples from $f_{\text{low}}(A)$ and $f_{\text{low}}(B)$ is quite efficient, since the 2nd eigenvalues 0.8341 and 0.8129 of $P(\mathcal{X}_{f_{\text{low}}(A)})$ and $P(\mathcal{X}_{f_{\text{low}}(B)})$, respectively, are bounded far away from 1.

To take advantage of these rapidly mixing properties, we write $f_{\text{low}}(\Omega)$ as a weighted sum of its restricted densities on A and B:

$$f_{\text{low}}(\Omega) = \pi_A f_{\text{low}}(A) + \pi_B f_{\text{low}}(B)\,, \tag{56}$$

with coupling factors

$$\pi_A = \frac{Z_{h_{\text{low}}(A)}}{Z_{h_{\text{low}}(A)} + Z_{h_{\text{low}}(B)}} \quad \text{and} \quad \pi_B = \frac{Z_{h_{\text{low}}(B)}}{Z_{h_{\text{low}}(A)} + Z_{h_{\text{low}}(B)}}\,. \tag{57}$$

Thus expectations wrt. $f_{\text{low}}(\Omega)$ of a random variable $g : \Omega \to \mathbb{R}$ are given by

$$\begin{aligned}
\mathbb{E}_{f_{\text{low}}(\Omega)}(g) &= \int_\Omega g(x) f_{\text{low}}(\Omega)(x)\, dx \\
&= \pi_A \int_A g(x) f_{\text{low}}(A)(x)\, dx + \pi_B \int_B g(x) f_{\text{low}}(B)(x)\, dx \\
&= \pi_A \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g\left(X^{(k)}_{f_{\text{low}}(A)}\right) + \pi_B \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g\left(X^{(k)}_{f_{\text{low}}(B)}\right)
\end{aligned} \tag{58}$$

In other words, we can compute $\mathbb{E}_{f_{\text{low}}(\Omega)}(g)$ via realizations of $\mathcal{X}(f_{\text{low}}(A))$ and $\mathcal{X}(f_{\text{low}}(B))$, without the need to compute a realization of the slowly mixing $\mathcal{X}(f_{\text{low}}(\Omega))$. Yet, if we really intend to replace the slowly mixing $\mathcal{X}(f_{\text{low}}(\Omega))$ by $\mathcal{X}(f_{\text{low}}(A))$ and $\mathcal{X}(f_{\text{low}}(B))$, or by similar rapidly mixing Markov chains, we have at least to solve two essential problems:

1. identification of metastable sets A and B, and
2. computation of coupling factors π_A and π_B

We will address the former problem in the uncoupling step, and the latter in the coupling step.

Uncoupling Step. We first draw samples $\mathbf{x}_\Omega = (x_\Omega^{(1)}, \ldots, x_\Omega^{(N_\Omega)})$ at the higher temperature T_{high} from the canonical density $f_{\text{high}}(\Omega)$ via the Markov chain $\mathcal{X}_\Omega := \mathcal{X}(f_{\text{high}}(\Omega))$. We assume a realization of $\mathcal{X}_\Omega$ to mix well between all relevant parts of $f_{\text{high}}(\Omega)$, which at least is the case in our example (see Fig. 17 (b)).

A discretization of the associated Markov operator $P(\mathcal{X}_\Omega)$ as outlined in Sect. 3.3.3 by means of our samples $\mathbf{x}_\Omega$ results in a spectrum

j	1	2	3	4
λ_j	1	0.9820	0.6968	0.6313

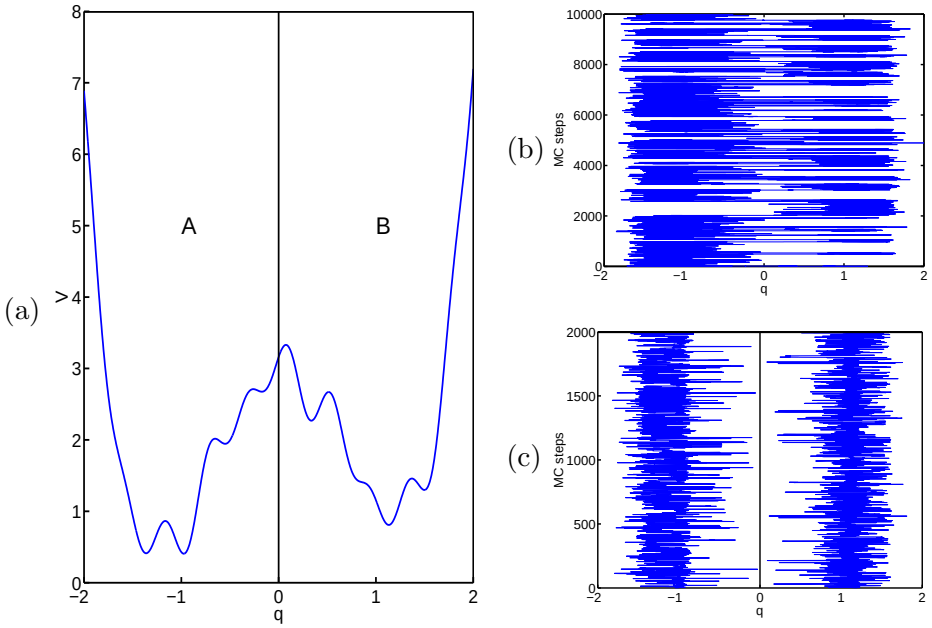

Figure 17: UC simulation of a one-dimensional multiminima potential $\mathcal{V}$ with two distinct wells. (a) The potential $\mathcal{V}$ with state space Ω is decomposed into two sets A and B, which typically turn out to be metastable wrt. the Markov operator under consideration. (b) Simulation on Ω at an increased temperature. We observe frequent transitions between the two metastable sets A and B (2nd eigenvalue of $P(\mathcal{X}_\Omega)$ is 0.9820; also cf. with Fig. 2 on page 7. (c) Two restricted bridge density simulations on A and B with 2nd eigenvalues 0.6745 and 0.7744, respectively; the low 2nd eigenvalues and thus fast convergence allows short simulation runs.

and therefore indicates the existence of two metastable sets. In fact, an identification of these two sets as outlined in Sect. 3.2.3 would result approximately in the sets $A = [-2, 0]$ and $B = (0, 2]$, which we have chosen beforehand in our example. The remarkable point is, that the metastable sets of $P(\mathcal{X}_\Omega)$ are in principal the same as for $P(\mathcal{X}_{f_{\text{low}}(\Omega)})$, only less metastable. Therefore, we can make use of the faster mixing $\mathcal{X}_\Omega$ to identify metastable sets for $\mathcal{X}_{f_{\text{low}}(\Omega)}$.

Now, with A and B identified, we set up bridge densities $f_{\text{high,low}}(A)$ and $f_{\text{high,low}}(B)$ which both encompass the corresponding canonical densities at T_{high} and T_{low}. For this task we use bridge densities as defined for ATHMC (or alternatively bridge densities for PSHMC), and then extract the missing parameter from $\mathbf{x}_\Omega$ as outlined in Sect. 4.4.2. Regarding our simple one-dimensional example, the set up of bridge densities may look oversized—which actually is true for this case. But since in a general high-dimensional state space $f_{\text{low}}(A)$ and $f_{\text{high}}(B)$ are almost entirely separated by the energy, the need for bridge densities becomes apparent.

The restricted samplings of $\mathcal{X}_A$ and $\mathcal{X}_B$ are shown in Figure 17 (c), with resulting 2nd eigenvalues 0.6745 and 0.7744 of $P(\mathcal{X}_A)$ and $P(\mathcal{X}_B)$,

respectively. Like $\mathcal{X}_{\text{low}(A)}$ and $\mathcal{X}_{\text{low}(B)}$, the Markov chains $\mathcal{X}_A$ and $\mathcal{X}_B$ are rapidly mixing on the metastable sets A and B, respectively.

In summary, identification of A and B via $\mathcal{X}_\Omega$ enable us via the bridge densities $\mathcal{X}_A$ and $\mathcal{X}_B$ to draw samples in Ω that encompass all relevant parts of our target distribution. In Sect. 5.2 we generalize and refine this approach to a hierarchical structure of bridge densities.

Coupling Step. Our aim is to set up an estimator for expectation values wrt. $f_{\text{low}}(\Omega)$ from the samples $\mathbf{x}_\Omega = (x_\Omega^{(1)}, \ldots, x_\Omega^{(N_\Omega)})$, $\mathbf{x}_A = (x_A^{(1)}, \ldots, x_A^{(N_A)})$, and $\mathbf{x}_B = (x_B^{(1)}, \ldots, x_B^{(N_B)})$ of $\mathcal{X}_\Omega$, $\mathcal{X}_A$, and $\mathcal{X}_B$, respectively.

By reweighting as outlined in Sect. 4.4.1 $\mathbf{x}_A$ and $\mathbf{x}_B$ from $f_{\text{high,low}}(A)$ and $f_{\text{high,low}}(B)$ to $f_{\text{low}}(A)$ and $f_{\text{low}}(B)$, respectively, we obtain weights

$$\mathbf{w}_{f_{\text{low}}(A)} = \left(w_{f_{\text{low}}(A)}^{(1)}, \ldots, w_{f_{\text{low}}(A)}^{(N_A)}\right)$$

and

$$\mathbf{w}_{f_{\text{low}}(B)} = \left(w_{f_{\text{low}}(B)}^{(1)}, \ldots, w_{f_{\text{low}}(B)}^{(N_B)}\right) .$$

With these weights we can rewrite (58), with the only difference that the approximations of $\mathbb{E}_{f_{\text{low}}(A)}(g)$ and $\mathbb{E}_{f_{\text{low}}(B)}(g)$ are given in terms of weighted random variables due to the use of bridge densities:

$$\begin{aligned}\mathbb{E}_{f_{\text{low}}(\Omega)}(g) &= \pi_A \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} w_{f_{\text{low}}(A)}^{(k)} g\left(X_A^{(k)}\right) \\ &\quad + \pi_B \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} w_{f_{\text{low}}(B)}^{(k)} g\left(X_B^{(k)}\right) \end{aligned} \tag{59}$$

What we still need are the coupling factors π_A and π_B between these two weighted samples. The quotient of normalizing constants

$$c = \frac{Z_{h_{\text{low}}(A)}}{Z_{h_{\text{low}}(B)}} \tag{60}$$

is the key to compute them, since then

$$\pi_A = \frac{c}{1+c} = \frac{Z_{h_{\text{low}}(A)}}{Z_{h_{\text{low}}(A)} + Z_{h_{\text{low}}(B)}} \quad \text{and} \quad \pi_B = 1 - \pi_A . \tag{61}$$

A direct computation of (60) is not possible, but we can build a "bridge" between $f_{\text{low}(A)}$ and $f_{\text{low}(B)}$ via $f_{\text{high}(\Omega)}$ by setting

$$\frac{Z_{h_{\text{low}}(A)}}{Z_{h_{\text{low}}(B)}} = \frac{Z_{h_{\text{high}}(B)}}{Z_{h_{\text{low}}(B)}} \frac{Z_{h_{\text{high}}(A)}}{Z_{h_{\text{high}}(B)}} \frac{Z_{h_{\text{low}}(A)}}{Z_{h_{\text{high}}(A)}} , \tag{62}$$

thus expanding (60) into three quotients of normalizing constants. The first and third part can be estimated from the bridge density samples of $\mathbf{x}_A$ and

$\mathbf{x}_B$, respectively; the middle part from $\mathbf{x}_\Omega$. However, we postpone the actual estimation of these quotients to Sect. 5.3.2, where we treat them within the general coupling step.

Let us for now assume that we already obtained an estimate $\hat{c}$ of (60) via (62). Then, we can compute approximations $\hat{\pi}_A$ and $\hat{\pi}_B$ from (61), which eventually lead to an estimator for (59):

$$\hat{\mathbb{E}}_{f_{\text{low}}(\Omega)}(g) = \hat{\pi}_A \frac{1}{N_A} \sum_{k=1}^{N_A} w^{(k)}_{f_{\text{low}}(A)} g\left(x_A^{(k)}\right) + \hat{\pi}_B \frac{1}{N_B} \sum_{k=1}^{N_B} w^{(k)}_{f_{\text{low}}(B)} g\left(x_B^{(k)}\right) \quad (63)$$

Although samples from x_Ω are not explicitely part of (63), they play an important role in the estimation of $\hat{\pi}_A$ and $\hat{\pi}_B$, thus a good sample from $f_{\text{high}}(\Omega)$ is nevertheless essential. We will see in Sect. 5.3.3 that it is also possible to include $\mathbf{x}_\Omega$ explicitely.

The coupling matrix of UC, which we introduce in Sect. 5.3, will turn out to be a generalized scheme of the coupling steps described here.

Computational Cost. Essentially, we replaced a slowly mixing Markov chain $\mathcal{X}_{\text{low}}(\Omega)$ with $\lambda_2(P_{\mathcal{X}_{\text{low}}(\Omega)}) = 0.9988$ by three Markov chains $\mathcal{X}_\Omega$, $\mathcal{X}_A$, and $\mathcal{X}_B$ with corresponding 2nd eigenvalues 0.9820, 0.6745, and 0.7744, respectively. There is no doubt for $\mathcal{X}_A$ and $\mathcal{X}_B$ to be rapidly mixing, and even for $\mathcal{X}_\Omega$ we have $\lambda_2(P_\Omega) < 0.9822 = \lambda_2(P^{15}_{\mathcal{X}_{\text{low}}(\Omega)})$, which loosely spoken indicates that for one step of $\mathcal{X}_\Omega$ one has to perform 15 steps with $\mathcal{X}_{\text{low}}(\Omega)$ to obtain the same mixing behavior. One should be careful, however, in interpreting this factor, since it is mainly dependent on the choice of temperatures T_{low} and T_{high}. We will see in Sect. 6, that in general UC can lead in more complex situations to much higher computational gains. Although important for UC to be successful, the remaining overhead of the method (like dynamical clustering or reweighting) is negligible in terms of computational cost.

5.2 Uncoupling of Markov Chains

By the uncoupling step we refer on the one hand to a hierarchical decomposition of the state space Ω into metastable sets and on the other hand to restricted bridge sampling in these sets by restarted Markov chains. Yet, we first investigate a much simpler situation: given a Markov chain $\mathcal{X}$ which is slowly mixing due to n metastable sets, we outline how to set up n rapidly mixing Markov chains $\mathcal{X}_1, \ldots, \mathcal{X}_n$ on these metastable sets, and analyze their properties and the effect this imposes on the spectra of the associated Markov operators $P_{\mathcal{X}_1}, \ldots, P_{\mathcal{X}_n}$. The algorithmic usefulness of these theoretical investigations will become apparent in a hierarchical context in Sect. 5.2.3.

5.2.1 Restricted Sampling

Suppose $\mathbf{x} = (x^{(1)}, \ldots, x^{(N)})$ to be generated from $\mathcal{X}$ and the sets $A_1, \ldots, A_n$ to be the output of our dynamical cluster algorithm. We further assume, that $\mathcal{X}$ is given by a Metropolis type transition kernel K as defined in Sect. 4.1. To sample separately in each A_l, for $k = 1, \ldots, n$ we define restricted Markov chains $\mathcal{X}_1, \ldots, \mathcal{X}_n$ with associated Markov kernels K_l, $l = 1, \ldots, n$ by restricting K on A_l:

$$K_l(x, dy) = k_l(x, y)\mu(dy) + r_l(x)\delta_x(dy) \tag{64}$$

with

$$k_l(x, y) = \begin{cases} q(x, y)\alpha(x, y) & \text{if } x \neq y \text{ and } y \in A_l \\ 0 & \text{otherwise} \end{cases} \tag{65}$$

and

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

That is, we can generate a realization of $\mathcal{X}_l$ by a simple alteration of the Metropolis algorithm for $\mathcal{X}$. One update step for the restricted chain $\mathcal{X}_l$ is then given by:

1. *suppose, we are in the state* $x_l^{(j)} \in A_l$
2. *compute proposal* y_l *for the Metropolis algorithm associated with* $\mathcal{X}$
3. *reject* y_l*, if* $y_l \notin A_l$ *(i.e., set* $x_l^{(j+1)} := x_l^{(j)}$*)*
4. *otherwise, accept (i.e., set* $x_l^{(j+1)} := y_l$*) with the same probability as for the Metropolis algorithm associated with* $\mathcal{X}$

Since A_l is metastable wrt. $\mathcal{X}$, mixing within A_l is fast, and proposal steps outside of A_l (which lead to rejections) will occur only rarely. Intuitively, the restricted Markov chain $\mathcal{X}_l$ should again be rapidly mixing. But before we analyze this in more detail, we investigate some general properties of $\mathcal{X}_l$.

Clearly, the detailed balance condition (43) wrt. f_l still holds for all $x, y \in A_l$, so K_l is again a reversible Markov kernel. Now, let $h_l = \mathbf{1}_{A_l} h$ be the restricted unnormalized density on A_l (as before, $\mathbf{1}_A$ denotes the indicator function on A). Then, under the assumption that K_l is irreducible, $f_l = h_l / Z_{h_l}$ is the unique invariant density of K_l. Therefore, the density f_l is a scalar multiple of the global density f. Thus, we can regain the global density via

$$f = \sum_{l=1}^{n} \pi_l f_l \tag{66}$$

in terms of the local densities f_l. Only the coupling factors $\pi_1, \ldots, \pi_n$ are unknowns which represent the neglected coupling between the sets A_l. As a generalization of (57), they are again given by

$$\pi_l = \frac{Z_{h_l}}{\sum_{k=1}^{n} Z_{h_k}} = \frac{Z_{h_l}}{Z_h} \,.$$

It is not our aim to estimate these coupling factors in this situation (which in fact would be intractable here), but postpone this problem after we introduced a hierarchical scheme in Sect. 5.2.3.

Illustration of Restricted Sampling. For ease of presentation we will now illustrate this procedure in a finite dimensional situation. To this end, let $\tilde{T}$ again denote the transition matrix of HMC for n-butane associated with the box discretization in 23 boxes given in Sec. 3.3.3.

The first four eigenvalues of $\sigma(\tilde{T})$, namely

j	1	2	3	4
λ_j	1	0.9779	0.9733	0.4850

clearly indicate the existence of three metastable sets.

Let us denote the associated chain by $\mathcal{X}$ in the following. Moreover, let A_1, A_2, A_3 be a decomposition of the state space $\Omega = \{1, \ldots, 23\}$ into three metastable sets. Applying (64) to $\tilde{T}$ results in a transition matrix $\tilde{T}_{\text{restr}}$ with entries

$$\tilde{T}_{\text{restr},kl} = \begin{cases} \tilde{T}_{kl}, & k,l \in A_i \text{ for } i \in \{1,2,3\} \text{ and } k \neq l \\ 0, & k \in A_i, l \in A_j \text{ for } i,j \in \{1,2,3\} \text{ and } i \neq j \\ \tilde{T}_{ll} + \sum\limits_{i \notin A_j} \tilde{T}_{li}, & k = l \in A_j \end{cases}$$

Consequently, if we assume the boxes for each subset A_l to be in a successive order, $\tilde{T}_{\text{restr}}$ has block-diagonal form. The three stochastic matrices $\tilde{T}_l$ on the block diagonal of $\tilde{T}_{\text{restr}}$ are then transition matrices associated with the uncoupled Markov chains $\mathcal{X}_1$, $\mathcal{X}_2$, and $\mathcal{X}_3$.

Figure 18 illustrates the situation when the sets $A_1 = \{1, \ldots, 7\}, A_2 = \{8, \ldots, 16\}$, and $A_3 = \{17, \ldots, 23\}$ are *good* approximations of the three metastable sets of $\mathcal{X}$. Figure 18 shows on the left the corresponding decomposition in the torsion angle of n-butane, in the middle $\tilde{T}_{\text{restr}}$, and on the right the ordered spectra of the three transition matrices T_l. The 2nd eigenvalues of the three T_l,

$\lambda_2(T_1)$	$\lambda_2(T_2)$	$\lambda_2(T_3)$
0.5353	0.6707	0.5010

are indeed substantially less than 1 (i.e., all three restricted Markov chains $\mathcal{X}_l$ are rapidly mixing).

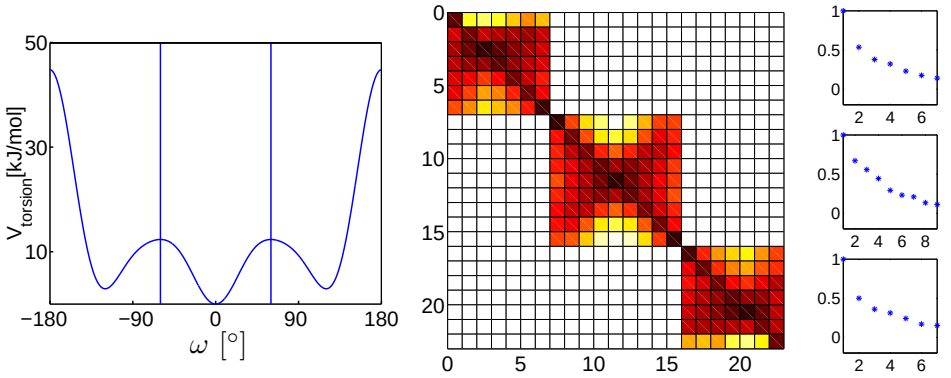

Figure 18: Left: Illustration of the entries of the transition matrix $\tilde{T}_{\text{restr}}$ (as defined in the text above) for a good choice of A_1, A_2, A_3. Intensity of entries due to logarithmic scale. Right: Ordered spectrum of $\tilde{T}$.

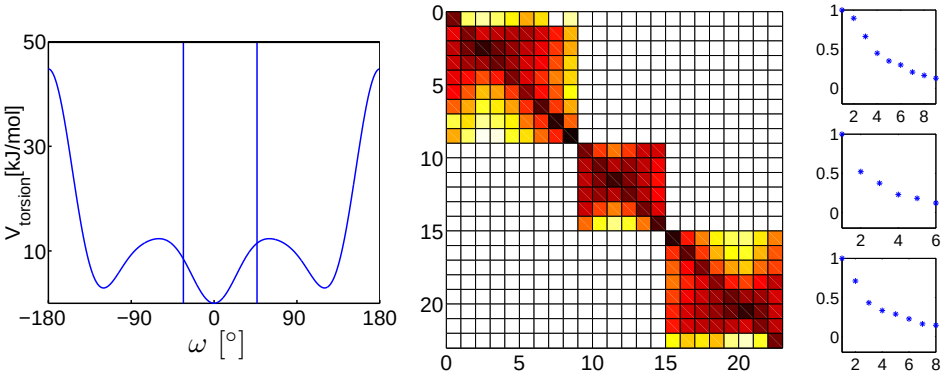

Figure 19: Left: Illustration of the entries of the transition matrix $\tilde{T}_{\text{restr}}$ for a deteriorated choice of A_1, A_2, A_3. Intensity of entries due to logarithmic scale. Right: Ordered spectrum of $\tilde{T}$.

Figure 19 now illustrates *bad* approximations of the three metastable sets of $\mathcal{X}$, namely $A_1 = \{1, \ldots, 9\}, A_2 = \{10, \ldots, 15\}$, and $A_3 = \{16, \ldots, 23\}$. Again, the right hand side of Fig. 19 shows the ordered spectra of the three transition matrices T_l. We observe that now the 2nd eigenvalues

$\lambda_2(T_1)$	$\lambda_2(T_2)$	$\lambda_2(T_3)$
0.8952	0.5215	0.7113

are much closer to 1 (nevertheless, compared to $\mathcal{X}$ the three restricted Markov chains $\mathcal{X}_l$ are still rapidly mixing).

A comparison of these two uncouplings of $\tilde{T}$ indicates that the transformation of $\tilde{T}$ into $\tilde{T}_{\text{restr}}$ due to (64) is in fact not very sensitive wrt. to the resulting spectra; at least, as long as the sets A_l are metastable wrt. $\mathcal{X}$.

Another important aspect of restricted Markov chains according to (64) is illustrated in Fig. 20: the invariant density f of the unrestricted chain

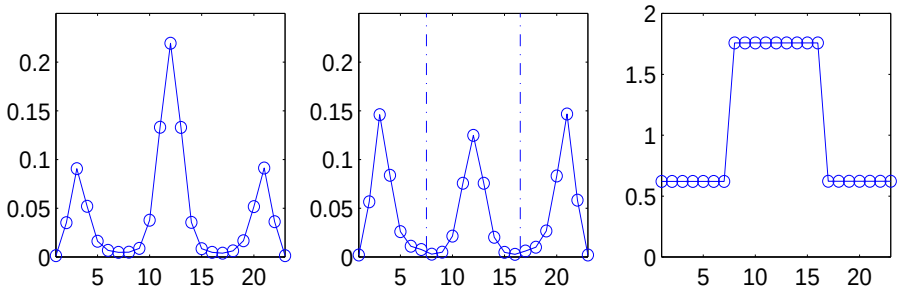

Figure 20: Left: Invariant distribution f. Center: Invariant distributions f_l, $l = 1, 2, 3$, in the three metastable sets A_l (as in Fig. 18). Right: Quotients $\pi_l = f/f_l$ in the three sets A_l.

is compared to the invariant densities f_l of the three restricted chains from Fig. 18. As a consequence of (66), we observe $f/f_l = \text{const} = \pi_l$ on each subset A_l, $l = 1, 2, 3$.

5.2.2 Eigenvalue Splitting

By defining restricted Markov operators on metastable sets as in Sect. 5.2.1, the eigenvalue cluster λ_l, $l = 1, \ldots, n$ in the vicinity of 1 of the original Markov operator P splits up into eigenvalues $\lambda_1(P_l) = 1$ for the restricted Markov operators P_l. Figure 21 gives an illustration of this behavior for n-butane.

We now give a mathematical justification of the assumption that lead us to the definition of (64), namely, that restricted Markov chains on metastable sets are indeed rapidly mixing (or at least possess increased mixing properties). Again, we restrict our investigations to a finite dimensional case.

To that end, let T be an arbitrary primitive and reversible $s \times s$ stochastic matrix with invariant distribution π, which we decompose into

$$T = D + E = \begin{pmatrix} D_{11} & E_{12} & \cdots & E_{1n} \\ E_{21} & D_{22} & \cdots & E_{2n} \\ \cdot & \cdot & \ddots & \cdot \\ E_{n1} & E_{n2} & \cdots & D_{nn} \end{pmatrix}. \tag{67}$$

At this point, we do not make any further assumptions about the size of the blocks $D_{11}, \ldots, D_{nn}$, nor do we assume the states of a block to be strongly coupled or the blocks among each other to be nearly uncoupled.

Via the row sums of $E = (e_{ij})$, we define the diagonal matrix E_{diag} by

$$E_{\text{diag}} = \text{diag}\left(\sum_{j=1}^{s} e_{1j}, \ldots, \sum_{j=1}^{s} e_{sj} \right) \tag{68}$$

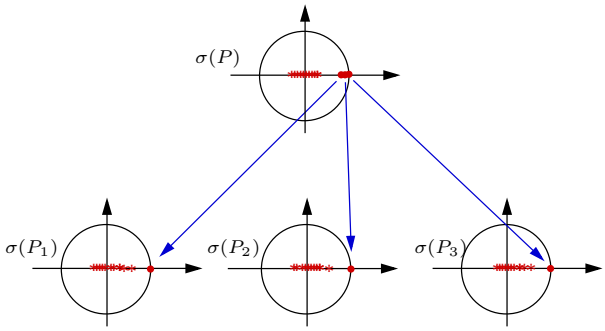

Figure 21: Typical splitting behavior of the spectral structure for restricted Markov operators (see also [91]). We here sketch the situation for n-butane, which was already illustrated in Figs. 8 and 18. Due to reversibility all spectra are real. By uncoupling into three metastable sets the eigenvalues $\lambda_1(P), \lambda_2(P)$, and $\lambda_3(P)$ split up into the spectra of the restricted Markov operators P_l, $l = 1, 2, 3$, each of these eigenvalues apparently transforms to $\lambda_1(P_l) = 1$.

and the restricted stochastic matrix

$$T_{\text{restr}} = D + E_{\text{diag}} = \begin{pmatrix} \bar{D}_{11} & 0 & \cdots & 0 \\ 0 & \bar{D}_{22} & \cdots & 0 \\ . & . & \ddots & . \\ 0 & 0 & \cdots & \bar{D}_{nn} \end{pmatrix}, \tag{69}$$

which is still reversible, since the pair(π, T_{restr}) still fulfills detailed balance.

Our goal is to make a statement about $\sigma(T_{\text{restr}})$ in terms of $\sigma(T)$. Since $\sigma(T_{\text{restr}})$ is uncoupled in n blocks $\bar{D}_{ll}$, $l = 1, \ldots, n$ (with each block forming a stochastic matrix), each $\lambda_1(\bar{D}_{ll}) = 1$ which directly implies $\lambda_k(T_{\text{restr}}) = 1$ for $k = 1, \ldots, n$.

In order to analyze $\lambda_{n+1}(T_{\text{restr}})$ in terms of $\lambda_{n+1}(T)$, we need the following theorem which is known as Weyl's inequalities ([12], III.2, pp. 62–63). It states relations between eigenvalues of symmetric matrices A, B, and $A+B$.

Theorem 11 (Weyl's Inequalities) *Let A, B be $s \times s$ symmetric matrices with eigenvalues $\lambda_1(A) \geq \cdots \geq \lambda_s(A)$ and $\lambda_1(B) \geq \cdots \geq \lambda_s(B)$. Then, for $j = 1, \ldots, s$,*

$$\begin{aligned} \lambda_j(A+B) &\leq \lambda_i(A) + \lambda_{j-i+1}(B) \quad \textit{for} \quad i \leq j\,, \\ \lambda_j(A+B) &\geq \lambda_i(A) + \lambda_{j-i+s}(B) \quad \textit{for} \quad i \geq j. \end{aligned}$$

If we put $i = j$ in the above inequalities, we immediately obtain

Corollary 1 *For each $j = 1, 2, \ldots, s$,*

$$\lambda_j(A) + \lambda_s(B) \leq \lambda_j(A+B) \leq \lambda_j(A) + \lambda_1(B).$$

With these preparations, we can state the following theorem, which provides a bound on $\lambda_{n+1}(T_{\text{restr}})$:

Theorem 12 *Let T, E, and T_{restr} be defined as in (67) and (69). Then,*

$$\lambda_k(T_{\text{restr}}) \leq \lambda_k(T) + 2\,\|E\|_\infty, \qquad \textit{for} \quad k = 1, \ldots, s. \tag{70}$$

Proof. With π being the invariant distribution of T, and $\mathcal{D}$ the diagonal matrix $\mathcal{D} = \text{diag}(\sqrt{\pi_1}, \ldots, \sqrt{\pi_s})$, we define similar symmetric matrices $T_{\text{sym}} = \mathcal{D}T\mathcal{D}^{-1}$, $E_{\text{sym}} = \mathcal{D}E\mathcal{D}^{-1}$, and $T_{\text{restr,sym}} = \mathcal{D}T_{\text{restr}}\mathcal{D}^{-1}$.

For the diagonal matrix E_{diag} as defined in (68) its spectral radius is obviously given by $r(E_{\text{diag}}) = \|E_{\text{diag}}\|_\infty$. Also, we can bound $r(E)$ by $\|E\|_\infty$, which can be seen as follows:

Consider the matrix

$$E_{\max} = E + \text{diag}\left(\|E\|_\infty - \sum_{j=1}^{s} e_{1j}, \ldots, \|E\|_\infty - \sum_{j=1}^{s} e_{sj} \right),$$

which is nonnegative with equal row-sum $\|E\|_\infty$ and satisfies $E \leq E_{\max}$.

Since ${\|E\|_\infty}^{-1} E_{\max}$ is stochastic, $\lambda_1({\|E\|_\infty}^{-1} E_{\max}) = 1$, and therefore $\lambda_1(E_{\max}) = \|E\|_\infty$. Applying Perron-Frobenius theory (see Theorem 4) with $E \leq E_{\max}$ then provides the inequality $\lambda_1(E) \leq \lambda_1(E_{\max})$, which directly translates into $r(E) \leq \|E\|_\infty$.

Next, consider the decomposition

$$T_{\text{restr}} = T + E_{\text{diag}} - E, \tag{71}$$

or in terms of the similar symmetric matrices

$$T_{\text{restr,sym}} = T_{\text{sym}} + E_{\text{diag}} - E_{\text{sym}}. \tag{72}$$

We first apply Corollary 1 with $j = 1$ to $E_{\text{diag}} - E_{\text{sym}}$:

$$\begin{aligned} \lambda_1(E_{\text{diag}} - E_{\text{sym}}) &\leq \lambda_1(E_{\text{diag}}) + \lambda_1(-E_{\text{sym}}) \\ &\leq \|E_{\text{diag}}\|_\infty + \lambda_1(-E) \\ &\leq 2\|E\|_\infty \end{aligned}$$

Applying Corollary 1 once more, we obtain for $k = 1, \ldots, s$

$$\begin{aligned} \lambda_k(T_{\text{restr,sym}}) &= \lambda_k(T_{\text{sym}} + E_{\text{diag}} - E_{\text{sym}}) \\ &\leq \lambda_k(T_{\text{sym}}) + \lambda_1(E_{\text{diag}} - E_{\text{sym}}) \\ &\leq \lambda_k(T_{\text{sym}}) + 2\|E\|_\infty. \end{aligned}$$

In terms of the matrices T_{restr} and E this inequality transforms into

$$\lambda_k(T_{\text{restr}}) \leq \lambda_k(T) + 2\|E\|_\infty.$$

□

For an arbitrary decomposition $T = D + E$ entries in E will be quite large (even for a small value of $\lambda_{n+1}(T)$) and a bound for $\lambda_{n+1}(T_{\text{restr}})$ given by Theorem 12 could even be greater than 1. This reflects the situation that we would not expect rapidly mixing in the blocks $\bar{D}_{ll}$, $l = 1, \ldots, n$; it would even be possible that one or all of the restricted Markov chains are no longer irreducible.

But now suppose that (e.g., as output of a dynamical cluster algorithm) we obtain a decomposition $T = D + E$ where $\|E_{\text{sym}}\|_\infty$ is small. Then, the bound on $\lambda_{n+1}(T_{\text{restr}})$ is still bounded away from 1, from which at least irreducibility for all D_{ll}, $l = 1, \ldots, n$, follows. Depending on the gap between 1 and $\lambda_{n+1}(T_{\text{restr}})$, Theorem 12 may help to assess the rapidly mixing property of restricted Markov chains.

Example. Inequality (70) can become strict. As we can see from the simple 2×2 stochastic matrix

$$T = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} \quad \text{with} \quad T_{\text{restr}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

we have $\lambda_2(T) = 1 - 2\epsilon$, and Eq. (70) yields

$$1 = \lambda_2(T_{\text{restr}}) \leq \lambda_2(T) + 2\epsilon = 1.$$

In general, since we made no assumptions on the internal structure of restricted blocks, inequality (70) has to take into account the worst case behavior that is compatible with $\sigma(T)$. This may lead to actual values of $\lambda_{n+1}(T_{\text{restr}})$ much below the bound provided by Theorem 12.

For example, if we return to the discretizations of n-butane illustrated in Fig. 18, we have to compare the bound provided by (70) with the (probable) increase of the eigenvalue $\lambda_4(T) = 0.4850$ on $\lambda_4(T_{\text{restr}})$. Yet, even for such a good metastable decomposition low weighted transition states keep $\|E\|_\infty$ quite large with $\|E\|_\infty = 0.3666$, so we have

$$\lambda_4(T_{\text{restr}}) = 0.6707 \leq \lambda_4(T) + 2 \times 0.3666 = 1.2182,$$

which is a trivial bound, since $\lambda_4(T_{\text{restr}})$ is already bounded by 1.

5.2.3 Hierarchical Annealing

Given a sample $\mathbf{x}$ of a Markov chain $\mathcal{X}$, dynamical clustering enables us to detect its metastable sets. Yet, this presupposes that we already generated a good sample $\mathbf{x}$ which to obtain was our initial algorithmic aim. We circumvented this problem in the introductory example (see Sect. 5.1) by constructing a patchwork of distributions via the canonical density $f_{\text{high}}(\Omega)$ and the bridge densities $f_{\text{high,low}}(A), f_{\text{high,low}}(B)$ of which we have extracted a

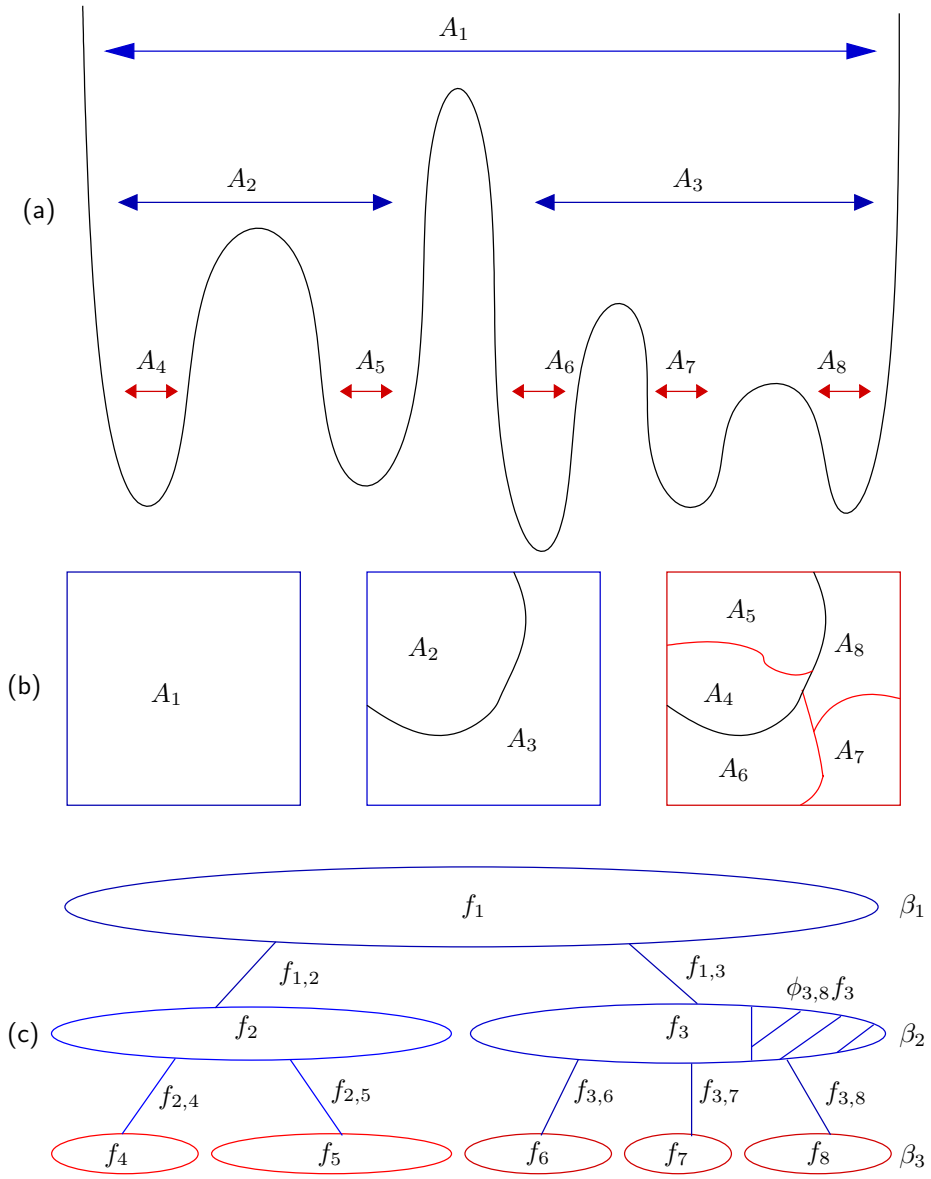

Figure 22: (a) Illustration of a one-dimensional potential $\mathcal{V}$ with hierarchical nested metastable sets. (b) Schematic plot of a three level hierarchical decomposition. An initial sampling of f_1 decomposes the state space $\Omega = A_1$ into two subsets A_2 and A_3, which get further subdivided into $\{A_4, A_5\}$ and $\{A_6, A_7, A_8\}$, respectively. The three levels are related to an annealing process; the top level is related to β_1, the intermediate level to $\beta_2 > \beta_1$, and the ground level to the inverse target temperature $\beta_3 = \beta_* > \beta_2$. (c) The same three level subdivision is now represented as a graph, where vertices correspond to canonical densities f_k with $\mathrm{supp}(f_k) = A_k$, and edges between two vertices i, j correspond to bridge densities $f_{i,j}$. By $\rho(k)$ we denote the index of the parent vertice of a density f_k. By drawing vertices as ellipses of different size, we also illustrate the extent of the respective distributions. As an example, the density $\phi_{3,8}f_3/Z_{\phi_{3,8}f_3}$ corresponds to the hatched part of f_3, and $\rho(8) = 3$; in UC neither $\phi_{3,8}f_3/Z_{\phi_{3,8}f_3}$ nor f_8 is sampled, but rather the bridge density $f_{3,8}$ (which sufficiently encompasses the important parts of $\phi_{3,8}f_3$ and f_8). The tree structure of the graph guarantees that the coupling matrix C (which we introduce in Sect. 5.3) is irreducible.

weighted sample with respect to the target distribution given by the density $f_* := f_{\text{low}}(\Omega)$. We now extend this approach into a hierarchical structure.

A natural way to represent a patchwork of distributions that is built up via bridge distributions is a graph $G = (V, E)$, whereby densities f_k, $k = 1, \ldots, M$, are associated with the set of vertices V, and all occurring bridge densities $f_{i,j}$ with the set of edges E. For a hierarchical annealing strategy (to which we restrict here) the graph G reduces to a rooted tree, with the initial sampling becoming its root.

In the following we assume the target density $f_* := f(\beta_*)$ to be defined as canonical density $h_* = \exp[-\beta_* \mathcal{V}]$ wrt. a given potential $\mathcal{V}$. As outlined in Sect. 4.4, by decreasing β in the canonical distribution we embed f_* into a hierarchy of distributions $f(\beta)$ which become easier to draw samples from.

Let T_1 be the temperature at which a given MCMC method can draw samples from $f(\beta_1)$ without suffering from trapping problems. In most practical situations $T_1 \gg T_*$, which prevents a direct bridge simulation as outlined in Sect. 4.4. Instead, we introduce a number of intermediate temperatures in such a way that bridge sampling between adjacent temperatures becomes feasible. This results in a hierarchy of temperature levels

$$T_1 > T_2 > \ldots > T_L = T_*,$$

or in terms of inverse temperatures, $\beta_1 < \beta_2 < \ldots < \beta_L = \beta_*$.

Our aim is now to build up a hierarchy via a series of bridge distributions between adjacent temperatures, each bridge distribution restricted to its higher temperature metastable set. Then, each Markov chain is rapidly mixing, and each sample will contain a vital piece of information for the coupling step.

How the hierarchy actually unfolds is illustrated in Fig. 22 for the situation of a three level decomposition with respect to three inverse temperatures $\beta_1 < \beta_2 < \beta_3 = \beta_*$. Essentially, what happens is a recursion of the following two steps:

1. Sampling of a bridge density $f_{i,j}$ on the set A_j between two adjacent temperature levels. The underlying assumption is that, since $\mathcal{X}_i$ is rapidly mixing on A_j, $\mathcal{X}_{i,j}$ is as well. The output is a sample $\mathbf{x}_{i,j}$.

2. Identification of metastable sets wrt. $\mathcal{X}_j$ by means of dynamical cluster analysis. For this purpose, $\mathbf{x}_{i,j}$ is reweighted towards the lower temperature.

This procedure is done until we reach the target temperature level T_*. We end up with samples from bridge distributions, each one covers a certain temperature range restricted to a metastable set in the state space. Extracting a weighted sample of $f(\beta_*)$ from these samples is part of the coupling step, which we describe in detail after the following remarks.

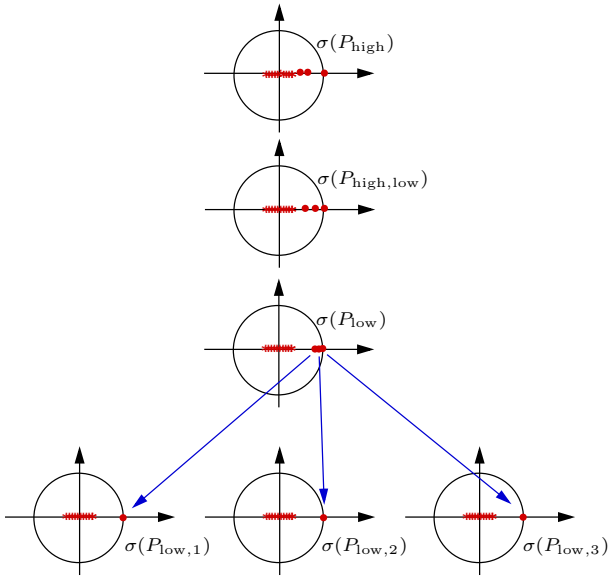

Figure 23: Extended splitting behavior of eigenvalues in the context of a hierarchical uncoupling. The scheme illustrates the spectra for three Markov operators all acting on the same metastable set A, and spectra resulting from a decomposition of A into three further metastable subsets. While the 2nd and 3rd eigenvalues of $\sigma(P_{\text{high}})$ are significantly bounded away from 1 they are closer to 1 in $\sigma(P_{\text{low}})$, which results from lowering the temperature from T_{high} to T_{low}. In UC we draw samples from $f_{\text{high,low}}$ via a still rapidly mixing Markov chain as indicated by $\sigma(P_{\text{high,low}})$. The key point is that reweighting the $f_{\text{high,low}}$-samples to f_{low} allows for detecting strong metastabilities in P_{low} without directly drawing samples from f_{low}. The decomposition into three metastable sets results in a splitting of three eigenvalues into $\lambda_1(P_{\text{low},l}) = 1$, $l = 1, 2, 3$, for the restricted Markov operators $P_{\text{low},l}$.

Spectral Structure. Irrespective of Markov chains being restricted on metastable sets or not, higher temperatures lead to better mixing rates. Figure 23 illustrates how this behavior is reflected by the spectra of the Markov operators associated to temperatures T_{high} and T_{low} as well as the bridge distribution $T_{\text{high,low}}$.

The situation is as follows: From the previous simulation in the hierarchy we are given a set A such that the Markov chain $\mathcal{X}_{\text{high}}$ at temperature T_{high} is rapidly mixing. What we want is to find out about metastable sets wrt. T_{low}. We solve this problem by setting up a bridge distribution, such that $\mathcal{X}_{\text{high,low}}$ is still rapidly mixing. Eigenvalues of $\sigma(P_{\text{high,low}})$ will, however, be somewhat closer to 1, since $\mathcal{X}_{\text{high,low}}$ also draw samples from the slowly mixing low temperature region T_{low}.

Reweighting the bridge distribution's sample to T_{low} directly provides a reliable statistical representation of its canonical distribution. Regarding our goal of identifying metastable sets wrt. P_{low}, a perfectly reweighted version of our dynamical cluster algorithm is not possible. The reason is

that dynamical transitions in higher temperature regions cannot be correctly reweighted to low temperature transitions. Yet, we can use an approximate scheme that practically identifies the same metastable sets. Suppose, that we have already computed weight factors $\mathbf{w}$ for a sample $\mathbf{x}$. Then, for a given discretization we use the same approximation via relative frequencies as in (38), (39), or (40), but additionally incorporate the weights $\mathbf{w} = (w^{(1)}, \ldots, w^{(n)})$. A reasonable choice is to weight a transition from $x^{(j)}$ to $x^{(j+1)}$ with the arithmetic mean of both weights, i.e., $(w^{(j)} + w^{(j+1)})/2$. With the notation from Sect. 3.3.3 Eq. (38) then turns into

$$\kappa(B_k, B_l) \approx \frac{\sum_{j=1}^{n-1} (w^{(j)} + w^{(j+1)})\, \mathbf{1}_{B_k}(x^{(j)})\, \mathbf{1}_{B_l}(x^{(j+1)})}{\sum_{j=1}^{n-1} (w^{(j)} + w^{(j+1)})\, \mathbf{1}_{B_k}(x^{(j)})},$$

and analogue equations can be derived from (39) and (40).

In summary, we are able to identify new metastable sets that emerge at T_{low} via the rapidly mixing Markov chain $\mathcal{X}_{\text{high,low}}$. This effect is exploited during the whole uncoupling procedure, thus avoiding the initial problem of analyzing the state space before a simulation even has been started.

Essential Hierarchy. What we described so far is an uncoupling strategy that produces a non-decreasing series of numbers M_k of metastable sets for each new hierarchical level β_k, $k = 1, \ldots, L$. This strategy can result in a strong increase of metastable sets due to identification of new emerging metastable subsets. Since all metastable sets are treated equal irrespective of their probability wrt. to $f(\beta_k)$, we would spend more and more computational effort on parts of the state space of low probability.

One way to avoid a possible computational explosion is to concentrate on the most probable metastable sets by discarding metastable sets of low probability. This can be achieved, e.g., by introducing a threshold on each temperature level β_k. The influence of discarding low weighted metastable sets on $f_* = f(\beta_L)$ is small, since these sets tend to become even less weighted during annealing anyway. We will see in Sect. 5.3.1 that discarding metastable sets causes no errors for the remaining sets in the coupling process.

5.3 Coupling Matrix

In the coupling step we will show that it is possible to regain information about a global density $f = \sum_{k=1}^{M} \pi_k f_k$ in terms of a patchwork of densities f_k by defining a coupling matrix C with π as its stationary distribution and estimating the entries of C from random samples of the f_k's.

5.3.1 Setup and Analysis

Now suppose that arbitrary unnormalized densities $h_1, \ldots, h_M$ are given. We denote by $A_k = \text{supp}(f_k)$ the support in the state space Ω and by

$\phi_{ij} = \mathbf{1}_{A_i \cap A_j}$ the indicator function of the common support of the densities f_i and f_j; μ denotes the underlying measure on Ω.

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

$$\sum_{k=1}^{M} \pi_k \, f_k = \sum_{k=1}^{M} \frac{Z_{h_k}}{Z_h} \, \frac{h_k}{Z_{h_k}} = \frac{h}{Z_h} = f.$$

Having in mind an algorithmic realization, we have to compute the π_k's (or at least approximations of them) without directly referring to Z_h. 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* $C = (c_{ij}) \in \mathbb{R}^{M \times M}$ by

$$c_{ij} = \begin{cases} \frac{1}{M} \, \frac{Z_{\phi_{ij} h_i}}{Z_{h_i}} \, \min\left\{1, \frac{Z_{\phi_{ji} h_j}}{Z_{\phi_{ij} h_i}}\right\} & \text{for } i \neq j \text{ and } \mu(A_i \cap A_j) > 0 \\ 0 & \text{for } i \neq j \text{ and } \mu(A_i \cap A_j) = 0 \\ 1 - \sum_{k=1 (k \neq i)}^{M} c_{ik} & \text{otherwise} \end{cases} \tag{73}$$

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

The key point in the construction of C is that

$$\pi = \frac{1}{Z_h}(Z_{h_1}, \ldots, Z_{h_M})$$

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

$$\pi_i \, \frac{Z_{\phi_{ij} h_i}}{Z_{h_i}} \, \min\left\{1, \frac{Z_{\phi_{ji} h_j}}{Z_{\phi_{ij} h_i}}\right\} = \pi_j \, \frac{Z_{\phi_{ji} h_j}}{Z_{h_j}} \, \min\left\{1, \frac{Z_{\phi_{ij} h_i}}{Z_{\phi_{ji} h_j}}\right\}, \tag{74}$$

which in addition shows the reversibility of C.

Expectation Values. If we suppose that we can estimate expectation values for each f_k and we know the stationary distribution π of C, we are able to estimate expectation values wrt. f, which are then given by

$$\mathbb{E}_f(g) = \sum_{k=1}^{M} \pi_k \int_{A_k} g(x) f_k(x)\, dx. \tag{75}$$

In general, we are more interested in expectation values wrt. f_*. Now suppose, that $h_* = \sum_{k=m}^{M} h_k$ (i.e., the sum of the unnormalized densities belonging to the hierarchical level β_*). Then, we can restrict (75) to f_* by

$$\mathbb{E}_{f_*}(g) = \sum_{k=m}^{M} \pi_k^* \int_{A_k} g(x) f_k(x)\, dx. \tag{76}$$

where

$$\pi_l^* = \frac{\pi_l}{\sum_{k=m}^{M} \pi_k} \quad \text{for } l = m, \ldots, M.$$

Example. Equation (73) defines the coupling matrix in a general setting for a patchwork of overlapping densities. In our setting, the hierarchically nested f_k's are reflected in the structure of C. For example, the decomposition of Fig. 22 results in a coupling matrix $C \in \mathbb{R}^{8\times 8}$, where non-zero entries show the pattern

$$C = \begin{pmatrix} \bullet & \bullet & \bullet & & & & & \\ \bullet & \bullet & & \bullet & \bullet & & & \\ \bullet & & \bullet & & & \bullet & \bullet & \bullet \\ & \bullet & & \bullet & & & & \\ & \bullet & & & \bullet & & & \\ & & \bullet & & & \bullet & & \\ & & \bullet & & & & \bullet & \\ & & \bullet & & & & & \bullet \end{pmatrix}.$$

The hierarchy of the decomposition clearly shows up. By setting $m = 4$ and $M = 8$, we can estimate expectation values wrt. f_* by means of (76).

Essential Hierarchy. At the end of Sect. 5.2.3 we discussed the possibility to discard low weighted metastable sets during hierarchical annealing. We now investigate the influence of this procedure on the remaining metastable sets of high probability.

In a complete hierarchy, we have

$$h_* = \sum_{k=m}^{M} h_k \quad \text{and} \quad f_* = \sum_{k=m}^{M} \pi_k^* f_k,$$

where π^* is derived from C.

In contrast, in an essential hierarchy we have only an incomplete representation of f_* with d missing densities, which we denote as $f_{*,\text{ess}}$. For ease of notation, we suppose these d densities to have indices $m, \ldots, m+d-1$. Then, $h_{*,\text{ess}} = \sum_{k=m+d}^{M} h_k$, and we would like to set up

$$f_{*,\text{ess}} = \sum_{k=m+d}^{M} \pi_k^{*,\text{ess}} f_k,$$

with accordingly renormalized coupling factors $\pi_k^{*,\text{ess}}$ from π. In an essential hierarchy we cannot derive $\pi_k^{*,\text{ess}}$ from π, since we would not set up C but an "essential" coupling matrix C_{ess}. Yet, the detailed balance condition (74) guarantees that essential coupling factors computed via C_{ess} are identical to the ones derived from C. This means that by discarding some metastable sets during annealing we do not introduce any other errors in the estimation of (76) than neglecting the contribution of the low weighted discarded parts of f_*.

5.3.2 Approximation of the Coupling Matrix

Equation (73) defines the analytical form of the coupling matrix C. Clearly, C cannot be evaluated analytically. Instead, we have to estimate its entries

$$c_{ij} = \frac{1}{M} \frac{Z_{\phi_{ij} h_i}}{Z_{h_i}} \min\left\{1, \frac{Z_{\phi_{ji} h_j}}{Z_{\phi_{ij} h_i}}\right\} \qquad \text{for} \quad i \neq j \quad \text{and } \mu(A_i \cap A_j) > 0. \tag{77}$$

from the given random samples. We will show in the following how to obtain an estimation $\hat{C} = (\hat{c}_{ij})$ of C. 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 $f_{\rho(i)}$ as well as their respective symmetric entries c_{ji} are non-zero (see Fig. 22).

Reweighting. In order to estimate the first quotient of normalizing constants in (77), we have to compute weight factors for bridge samples at first. For all bridge densities f_{ij} let

$$\mathbf{x}_{ij} = \left(x_{ij}^{(1)}, \ldots, x_{ij}^{(N_{ij})}\right) \tag{78}$$

be the samples from the Markov chains $\mathcal{X}_{ij}$. We define with

$$\mathbf{w}_{ij} = \left(w_{ij}^{(1)}, \ldots, w_{ij}^{(N_{ij})}\right) \tag{79}$$

weight factors for an unnormalized density $\phi_{ij} h_i$ by reweighting each sample $\mathbf{x}_{ij}$ to the lower temperature via

$$w_{ij}^{(k)} = \frac{\phi_{ij}(x_{ij}^{(k)}) h_i(x_{ij}^{(k)}) / h_{ij}(x_{ij}^{(k)})}{\sum_{m=1}^{N_{ij}} \phi_{ij}(x_{ij}^{(m)}) h_i(x_{ij}^{(m)}) / h_{ij}(x_{ij}^{(m)})}. \tag{80}$$

Since our sampling $\mathbf{x}_{ij}$ was already restricted to $A_i \cap A_j$, the indicator function ϕ_{ij} is always 1, i.e., $\phi_{ij}(x_{ij}^{(k)}) = 1$ for all k, and could therefore be omitted from (80). No reweighting is necessary for the initial sampling $\mathbf{x}_{11} = \{x_{11}^{(1)}, \ldots, x_{11}^{(N_{11})}\}$, therefore $w_{11}^{(k)} = 1/N_{11}$ for $k = 1, \ldots, N_{11}$.

Statistical weight of a Metastable Set. For f_j being a child density of f_i, and f_i being a child density of f_l (i.e., $A_j \subseteq A_i \subseteq A_l$), the expectation value $\mathbb{E}_{f_i}(\phi_{ij} h_i) = Z_{\phi_{ij} h_i}/Z_{h_i}$ can be approximated by the bridge density sample $\mathbf{x}_{il}$ due to

$$\frac{Z_{\phi_{ij} h_i}}{Z_{h_i}} = \lim_{N_{il} \to \infty} \sum_{k=1}^{N_{il}} w_{il}^{(k)} \phi_{ij}(x_{il}^{(k)}) \tag{81}$$

by means of $\mathbf{x}$ and $\mathbf{w}$. This ratio can be interpreted as the probability to be in the set A_j wrt. f_i.

Ratio of Normalizing Constants. We can estimate the ratio of normalizing constants solely by the samples $\mathbf{x}_{ij}$ from the bridge density f_{ij}:

$$\frac{Z_{\phi_{ji} h_j}}{Z_{\phi_{ij} h_i}} = \lim_{N_{ij} \to \infty} \frac{\sum_{k=1}^{N_{ij}} h_j(x_{ij}^{(k)})/h_{ij}(x_{ij}^{(k)})}{\sum_{k=1}^{N_{ij}} h_i(x_{ij}^{(k)})/h_{ij}(x_{ij}^{(k)})}. \tag{82}$$

For finite N_{ij} one needs a reliable sampling as described in Sect. 4.4 in order to obtain a satisfactory estimation.

Remark. With (81) and (82) an estimation $\hat{C}$ of C is given, of which we can directly compute its unique invariant distribution $\hat{\pi}$. In general $\hat{C}$ will not be exactly reversible as it is the case for C, but since each entry $\hat{c}_{ij}$ is an non-negative estimate of c_{ij} it converges to the reversible matrix C for $N_{ij} \to \infty$ for all i, j which come into question.

5.3.3 Reweighting

We describe two reweighting methods, which both allow us to set up estimators for expectation values. The first method restricts to bridge distribution samples connected to the lowest hierarchical level β_L, the second method allows to incorporate all samples from all levels $\beta_1, \ldots, \beta_L$. As introduced in Sect. 5.2.3, $\rho(i)$ denotes the index of the parent density in the hierarchy.

Simple Reweighting. The samples that are closest to f_* are obviously the samples drawn from the bridge distributions connected to the lowest hierarchical level β_L. We denote by $\hat{\pi}^*$ estimated normalized coupling factors

$$\hat{\pi}_l^* = \frac{\hat{\pi}_l}{\sum_{k=m}^{N} \hat{\pi}_k} \quad \text{for } l = m, \ldots, M, \tag{83}$$

for the densities $f_m, \ldots, f_M$ belonging to β_L, i.e., $f_* = f(\beta_L) = \sum_{k=m}^{M} \hat{\pi}_k^* f_k$.

The corresponding samples $\mathbf{x}_{\rho(i)i}$ and weight factors $\mathbf{w}_{\rho(i)i}$ then allows us to set up the estimator

$$\hat{\mathbb{E}}_{f_*}(g) = \sum_{i=m}^{M} \hat{\pi}_i^* \left(\sum_{k=1}^{N_{\rho(i)i}} w_{\rho(i)i}^{(k)} \right) g\left(x_{\rho(i)i}^{(k)} \right), \tag{84}$$

which is a discrete counterpart of (76).

Reweighting Mixtures. Equation (84) provides an estimator for expectation values wrt. the target density f_* by using sample points from the samples of the UC hierarchy that best represent the high probability parts of f_*. Yet, expectation values in dependence of the temperature or free energy differences are sensitive to a reliable sampling in low weighted parts of f_* (i.e., higher energy regions). If we want to estimate such quantities we should include the sample points of all bridge samples in order to improve the statistics in low weighted parts of f_*.

A direct and simple way to include these samples would be to reweight each sample separately to the respective restricted part of f_*, and then stick them together in a similar way as in (84). Unfortunately, this approach is not feasible in practice, since reweighting between distributions that do not overlap well is prone to error.

A much more promising approach is to construct a mixture distribution in which all samples contribute to a more or less equal part. That way, our sample points are spread equally all over the sampled parts of the state space; reweighting from that distribution results in reliable weight factors, which are the basis for a good estimator.

A suitable mixture distribution is given by the density

$$f_{\text{mix}} = \sum_{k=1}^{M} \frac{N_{\rho(k)k}}{N} f_k = \sum_{k=1}^{M} \frac{N_{\rho(k)k}}{N} \frac{h_k}{Z_{h_k}} \tag{85}$$

where $N = \sum_{k=1}^{M} N_{\rho(k)k}$ is the total number of sample points from all samples.

If we replace the Z_{h_k}'s by the coupling factors π_k we obtain an unnormalized mixture density

$$h_{\text{mix}} = \sum_{k=1}^{M} \frac{N_{\rho(k)k}}{N} \frac{h_k}{\pi_k}. \tag{86}$$

Using the estimate $\hat{\pi}$ instead of π we are actually able to evaluate (86). Expectations wrt. f_* are then computed by

$$\hat{\mathbb{E}}_{f_*}(g) = \sum_{k=1}^{M} \sum_{l=1}^{N_{\rho(k)k}} w_{\text{mix},\rho(k)k}^{(l)} \, g\left(x_{\rho(k)k}^{(l)} \right), \tag{87}$$

where weight factors are given by

$$w^{(j)}_{\mathrm{mix},\rho(i)i} = \frac{w^{(j)}_{\rho(i)i} h_*(x^{(j)}_{\rho(i)i}) / h_{\mathrm{mix}}(x^{(j)}_{\rho(i)i})}{\sum_{k=1}^{M} \sum_{l=1}^{N_{\rho(k)k}} w^{(l)}_{\rho(k)k} h_*(x^{(l)}_{\rho(k)k}) / h_{\mathrm{mix}}(x^{(l)}_{\rho(k)k})}. \tag{88}$$

Equation (88) is the usual reweighting formula, only this time applied to a mixture density for already weighted sample points.

Remarks. Reweighting from multiple Markov chains was introduced in the field of statistical physics by Ferrenberg and Swendsen [38], where they proposed an iterative procedure to combine the data. In the setting of mathematical statistics Geyer [55] proposed to determine normalizing constants by using *reverse logistic regression* in order to reweight from a mixture distribution. The difficult part in both approaches is to compute quantities that allow to combine the data. In our UC framework it is fairly easy to reweight from a mixture distribution, since the coupling factors π_k are in fact the unknown quantities needed for an overall reweighting.

5.4 Uncoupling-Coupling Scheme

The problem that we address by UC is to draw samples from a target distribution given by its density f_*. We still assume f_* to be a canonical density restricted to the potential part of a separable Hamiltonian $\mathcal{H} = \mathcal{T} + \mathcal{V}$ with state space Ω at some inverse temperature β_*. We further assume that the chosen Metropolis Markov chain is slowly mixing due to a hierarchy of metastable sets. Putting together the different steps described in this section, we arrive at the following hierarchical scheme of the UC algorithm (see also Fig. 3 on page 8):

1. *choose an initial inverse temperature β_1 and an annealing scheme $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_L = \beta_*$*
2. *draw initial sampling $\mathbf{x}_{11}$ from $f_1 = f(\beta_1)$ on $A_1 = \Omega$ with Hybrid Monte Carlo (HMC); set $l := 1$*
3. *perform a cluster analysis of each sample $\mathbf{x}_{\rho(k)k}$ in question and identify metastable sets wrt. f_k via a dynamical cluster algorithm*
4. *derive suitable parameters from previous samples to set up bridge densities between β_l and β_{l+1}*
5. *draw samples from restricted bridge distributions (e.g., by means of ATHMC or PSHMC) between β_l and β_{l+1} in all identified metastable sets; set $l := l + 1$*

6. *iterate 3. to 5. until the target temperature β_L is reached*
7. *set up the coupling matrix $\hat{C}$, thereby*
   - *estimate normalizing constants*
   - *compute its stationary distribution $\hat{\pi}$*
8. *estimate expectation values by means of simple reweighting or mixture reweighting*

This algorithmic scheme is not dependent on a specific realization of any of its constituents (MCMC method, cluster algorithm, annealing scheme, ...), but rather provides a general framework for uncoupling and coupling of Markov chains. We want to conclude this section with discussing some implementation issues as well as obvious generalizations. Issues concerning dynamical clustering and identification of metastable sets have already been discussed in Sect. 3.

MCMC Scheme. All our numerical experiments for biomolecules are based on Hybrid Monte Carlo (HMC) [15]. HMC is used for initial sampling at fixed temperature, and ATHMC or PSHMC discussed in Sect. 4.4 as bridge sampling methods. The symplectic and reversible Leapfrog integrator is used to calcute the trajectories as part of HMC and its variants. Yet, UC is by no means dependent on HMC; we can replace it by any suitable method that draw samples via a reversible Markov chain. The same applies to the choice of bridge distributions; one could think of replacing ATHMC or PSHMC by more sophisticated methods like multicanonical sampling or parallel tempering.

For convergence assessment we use an estimator proposed by Gelman and Rubin based on parallel chains [51, 49], which we described in Sect. 4.3. This estimator is especially suitable in the context of UC, since (except from the initial sampling) we can choose well distributed initial values from the parent sample in the hierarchy.

Annealing Strategy. The temperature parameter β in f_* is used to embed the target density in a family $f(\beta)$ of smoothed densities. A compromise has to be made between the number of levels in the annealing scheme $\beta_1 \leq \ldots \leq \beta_L$ and the overall computational cost. More temperature levels lead to more bridge samplings, but allow for a better set up and easier sampling of bridge distributions. The initial temperature should be high enough to overcome all metastabilities. Yet, no a-priori rule exists for determining β_1, which therefore has to be adjusted on the basis of some preliminary runs.

In our hierarchical annealing scheme we use a sequence of fixed temperatures. A possible extension would be to use a more flexible annealing scheme in which the temperature is estimated together with other bridge sampling

parameters. This would allow for a better adaptation to the local structure of the potential.

Essential Hierarchy. The use of an essential hierarchy instead of a complete hierarchy has been discussed in Sects. 5.2.3 and 5.3.1. Metastable sets of low probability are discarded by introducing a threshold wrt. the total probability of the current hierarchical level. To check such a criterium, we need to set up a coupling matrix on each hierarchical level after the corresponding bridge samplings. Extraction of normalizing constants from a bridge sampling need only be computed once, since they do not change their value in the sequel of coupling matrices.

Parallelization. Generating samples is by far the most time consuming part of UC, therefore parallelization should focus on this task. Since we make use of multiple Markov chains for convergence assessment, parallelization is straightforward. In our simulations, five chains of the same length run in parallel for each sample. Information exchange between these parallel processes is limited at simulation points where the convergence criterion is checked. In addition, Markov chains restricted to different metastable sets on a temperature level are mutually independent. They can be run in parallel as well, though they are not necessarily of the same length.

Analyzing Samples. Ending up with samples $\mathbf{x}_{\rho(k)k}$, weights $\mathbf{w}_{\rho(k)k}$, and estimated coupling factors $\hat{\pi}_k$ for each f_k the samples can be analyzed in different ways: For simply computing averages at the target temperature it is sufficient to construct a weighted sample from the lowest level in the hierarchy. To compute quantities that need information from a broader energy range (e.g., free energy differences, or quantities in dependence of the temperature) we have to set up a mixture distribution and employ the method of mixture reweighting. Moreover, reweighting is the first step to determine physical relevant metastable conformations together with its transition rates (for details see [72, 112, 113]).