1 Markov chains

A Markov chain, named after Andrey Markov, is a process that works on a countable discrete **state space** and changes between them in discrete steps of time, according to some random mechanism. In most cases relevant for this book, the state space is finite, having n states. For simplicity we will label the states by numbers 1 through n.

$$S = \{1, ..., n\}.$$

While in general the Markov chain steps are not associated to a physical time, when applying them to describe molecular kinetics we will associate each Markov chain step with an advance of the physical time by a fixed amount τ.

A **realization** or **trajectory** of the Markov chain is a time sequence,

$$(x_0, ..., x_N)$$

where $x_k \in S$ denotes the state at time step k. Each jump, such as $x_0 \to x_1$, is called a transition.

Markov property: Markov processes are defined by the fact that the propagation of the system is entirely determined by knowing its present state x_k, and is thus independent on its past. The process has no "memory" of its past. For Markov chains, this can be formalized as:

$$\mathbb{P}(x_k \mid x_{k-1},\, x_{k-2}, ..., x_0) = \mathbb{P}(x_k \mid x_{k-1}).$$

We shall be interested in **time-homogeneous** Markov chains. In general, the probability to make a transition from state i to state j at time k could depend on these states and on the time k. In time-homogeneous Markov chains, there is no dependence on time, and the transition probabilities depend on the states exclusively. In this case we can define a transition probability matrix, or short, **transition matrix**:

$$\mathbf{T} \in \mathbb{R}^{n \times n} \quad : \quad T_{ij} = \mathbb{P}(x_k = j \mid x_{k-1} = i)$$

whose element (i, j) yields the conditional transition probability that the Markov chain will be in state j at time k given that it has been in state i at time $k - 1$. As the transition matrix defines how the present state will propagate into the future state, it is also called the **propagator** of the system. When defined this way, the transition matrix is said to be **row-dominant** or **row-stochastic**. This is because each row i of $\mathbf{T}$ is a probability distribution of the state found at the next time-step:

$$\mathbf{T}_{i*} = (T_{i1}, ..., T_{in}).$$

There are different conventions for defining the transition matrix. In statistical texts, transition matrices are often denoted by the symbol $\mathbf{P}$. In many physical chemistry texts, transition matrices are often defined column-dominant rather than row-dominant. Column- and row-dominant transition matrices can be interconverted by transposition. Row-stochastic transition matrices have the properties:

$$T_{ij} \geq 0 \; \forall i, j$$
$$\sum_{j=1}^{n} T_{ij} = 1 \; \forall i$$

Example 1: Two-state protein folding Let us consider a two-state Markov model of a protein consisting of a folded state 1 and an unfolded state 2. A two-state Markov chain is characterized by the transition matrix

$$\mathbf{T} = \begin{bmatrix} 1 - T_{12} & T_{12} \\ T_{21} & 1 - T_{21} \end{bmatrix}$$

with the transition probabilities T_{12} and T_{21}. The self-transition probabilities are fixed by row-stochasticity to $T_{11} = 1 - T_{12}$ and $T_{22} = 1 - T_{21}$. Fig. 1a shows a graphical representation of three different two-state Markov chains. The self-transition probabilities are not drawn for clarity.

Example 2: A game in a tennis match

We model the time evolution of a game in a tennis match between players A and B. We suppose that for each ball played, player A scores with a probability p_A, and player B scores with a probability $p_B = 1 - p_A$. For each time scored, a player is advanced on the point list consisting of $\{0, 15, 30, 40\}$. This yields the 16 states 1-16 representing all possible combinations of points between the two players. When one of the players has 40 points and scores another time, he wins the game — the Markov chain reaches one of the terminal states 19 or 20. An exception is the score 40/40, at which any player needs to score twice before winning the game. Thus, two "advantage" states 17 and 18 are introduced, which either lead to winning the game, or going back to the 40/40 state. See Fig. 1b for an illustration.

Example 3: Discrete random walk

Consider the discrete random walk on the integers which starts at zero and transitions $+1$ or -1 with equal probability p at each step, while remaining at its current position with probability $1-2p$. This is an example for a Markov chain with an infinite, albeit countable, state space.

Generating realizations / trajectories Let us assume that in general the first state x_0 is drawn from an initial distribution $\mathbf{p}_0$. Then, a realization of length $N+1$ can be generated as follows:

1. Draw x_0 from the initial distribution $\mathbf{p}_0$
2. For $k = 0, ..., N-1$: draw x_{k+1} from the discrete distribution $[T_{x_k,1}, ..., T_{x_k,n}]$

Due to the stochasticity of the process, individual realizations may be very different. The probability of observing a specific realization $(x_0, ..., x_N)$ is given by:

$$\begin{aligned} \mathbb{P}(x_0, ..., x_N) &= p_{0,x_0} T_{x_0,x_1} \cdots T_{x_{N-1},x_N} \\ &= p_{0,x_0} \prod_{i=0}^{N-1} T_{x_{i-1},x_i} \end{aligned}$$

Fig. 1b shows sample trajectories of length 10,000 generated for the two-state folding models shown from Fig. 1a.

Examples Figs 1e-g show some statistics that can be straightforwardly computed by running many trajectories of the tennis game model from Fig. 1d and averaging over them: Fig 1e shows the probability that player A wins the set depending on the probability of scoring a ball, p_A. It is seen that the way a game is played corresponds to a nonlinear transformation of p_A to the game-winning probability, in such a way that only if the players are rather similar, both have a good chance of winning the game. If one player dominates (e.g. $p_A \geq 0.8$), the probability that he loses a game is near 0. Fig. 1f shows the mean number of balls played in one game, equivalent to the mean number of steps a Markov chain realization needs to reach states 19 or 20 when starting in 1. As expected, the games tend to take longer when the players have similar strength. Fig. 1g shows the statistics of the number of balls played between two players with equal ball scoring probability, $p_A = p_B = 0.5$. Although the mean of this distribution is between 6 and 7 balls, the distribution has a long tail and games with 15 or more balls are not rare. When a game takes 8 balls or more, it has to go through the "advantage A" or "advantage B" state and thus always takes an even number of balls to finish.

Ensemble evolution The probability to find the chain at state i at time t, $p_{t,i}$, can be computed by considering all possible realizations from the previous step $t-1$:

$$\begin{aligned} p_{t,i} &= p_{t-1,1} T_{1i} + ... + p_{t-1,n} T_{1n} \\ &= \sum_{j=1}^{n} p_{t-1,j} T_{1j} \end{aligned}$$

Define the probability vector $\mathbf{p}_t = (p_{t,1}, ..., p_{t,n})^T$, this is compactly written as:

$$\mathbf{p}_t^T = \mathbf{p}_{t-1}^T \mathbf{T}.$$

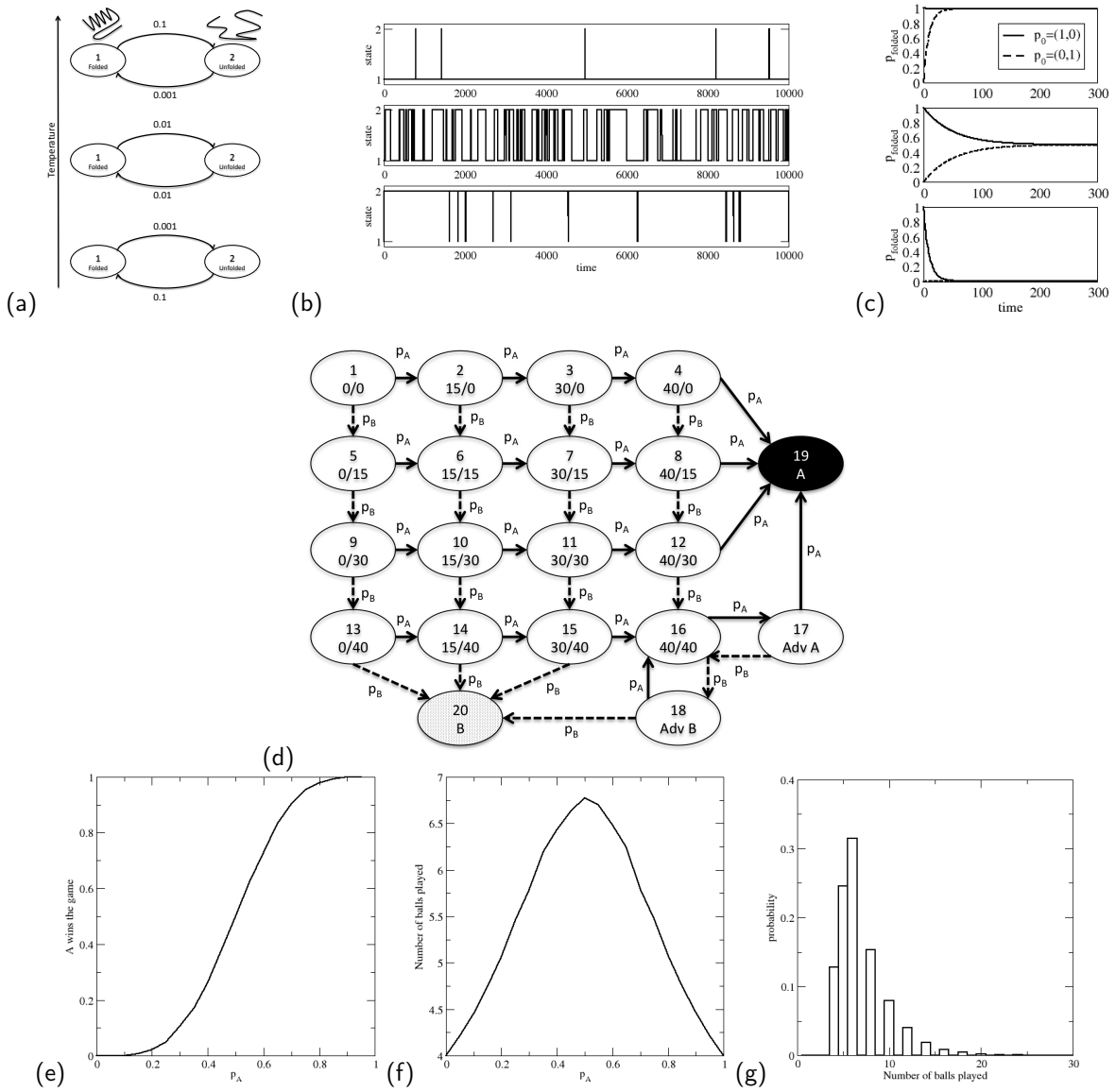


Figure 1: Examples for Markov chains and quantities that can be computed from their direct simulation. (a) A two-state Markov chain resembling a simplified folding/unfolding process of a two-state protein. (b) Protein folding model sample trajectories. (c) Ensemble probability $p_{t,1} = p_{t,\mathrm{folded}}$ when starting the ensemble from $\mathbf{p}_0^T = (1,0)$ or $\mathbf{p}_0^T = (0,1)$ (d) A Markov chain describing the evolution of a Tennis game between players A and B with respective scoring probabilities p_A and p_B. (e) The probability that A wins the game depending on his scoring probability p_A. (f) The mean number of balls played depending on p_A. (g) The probability distribution of balls plays for $p_A = 0.5$.

Applying this equation k times starting from $\mathbf{p}_0$ yields the **Chapman-Kolmogorow equation**:

$$\mathbf{p}_k^T = \mathbf{p}_0^T \mathbf{T}^k.$$

where $\mathbf{T}^k$ is the kth power of matrix $\mathbf{T}$. Thus, the powers of $\mathbf{T}$ are still stochastic matrices, and serve as the propagators for longer timesteps:

$$(\mathbf{T}^k)_{ij} = \mathbb{P}(x_k = j \mid x_0 = i)$$

Fig. 1c shows the evolution of $p_{t,1}$ when initializing the two-state folding models shown from Fig. 1a either from the folded state,$\mathbf{p}_0^T = (1, 0)$, or from the unfolded state, $\mathbf{p}_0^T = (0, 1)$.

Irreducibility and Connectivity Given a Markov chain transition matrix $\mathbf{T}$, it is essential for our purposes to verify that the state space is fully connected, that all the states can be reached from all other states under the dynamics of the Markov chain. Here we introduce the relevant terms from Markov chain theory and graph theory, and introduce a way to efficiently validate whether the state space is fully connected.

We start with the language of Markov chain theory. State j is **accessible** from state i (written $i \to j$), if and only if there exists a finite sequence of states

$$i = i_0, i_1, ..., i_{n-1}, i_n = j$$

such that $T_{i_k, i_{k+1}} > 0$ for all $k \in \{0, 1, ..., n-1\}$. Thus, $i \to j$ if there is a nonzero probability that the Markov chain reaches j after a finite number of steps when starting from i.

If both, $i \to j$ and $j \to i$, then we say that i and j **communicate** (written $i \leftrightarrow j$). Communication is a relation with following properties:

1. Reflexivity: each state communicates with itself, as for $n = 0$ we have the 1-sequence $i_0 = i = i_n$ with $(\mathbf{T}^0)_{ii} = 1$.
2. Symmetry: $i \leftrightarrow j$ is equivalent to $j \leftrightarrow i$.
3. Transitivity: If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.

A **communication class** $C \subseteq S$ is a set of states whose members communicate, i.e. $i \leftrightarrow j$ for all $i, j \in C$, , and no state in C communicates with any state not in C. Because of transitivity we can can equivalently say that there exists a $i \in C$ such that $i \leftrightarrow j$ for all $j \in C$. The communication class C is **closed** if additionally no state outside C is accessible from members inside C, i.e. when $i \in C$ and $i \to j$, then $j \in C$.

A finite Markov chain (or equivalently, its transition matrix $\mathbf{T}$) is **irreducible**, if it has a single communicating class $C = S$ (which is then automatically closed).

Example: the transition matrix

$$\mathbf{T} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

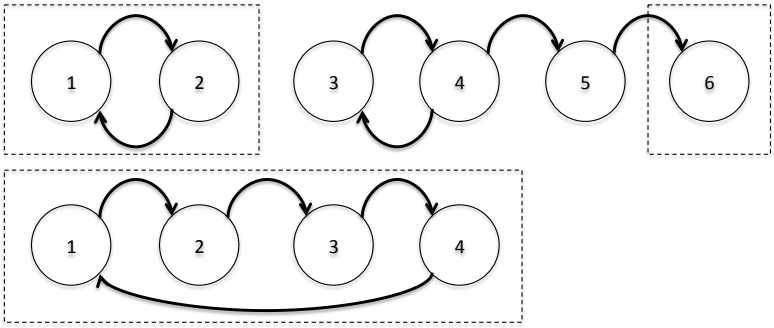
has the communication classes $\{1, 2\}$, $\{3, 4\}$, $\{5\}$, $\{6\}$. The communication class $\{3, 4\}$ is connected to $\{5\}$, and $\{5\}$ is connected to $\{6\}$, so $\{3, 4\}$ and $\{5\}$ are not closed. The only closed communication classes are $\{1, 2\}$ and $\{6\}$. $\mathbf{T}$ is reducible (not irreducible).

Example: the transition matrix

$$\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

has the single closed communication class $S = \{1, 2, 3, 4\}$. $\mathbf{T}$ is thus irreducible.

In order to design an efficient algorithm to compute the communication classes, it is useful to view the transition matrix as a graph. We define the **connectivity graph** $D = (S, A)$. D is a **directed graph**, consisting of a set of **nodes** and **arrows** connecting these nodes. Here, D has the node set S, i.e. each node represents a state of the Markov chain. The arrow set A consists of all arrows that connect state i to j if and only if $T_{ij} > 0$ and $i \neq j$. The connectivity graphs of the two transition matrices above are:



and the closed communication classes are marked by dashed boxes. In graph theory, two nodes $i, j \in C$ are **connected** if either a path from i to j or from j to exists. The are **strongly connected**, if both paths exist (i.e.: $i \leftrightarrow j$). A set C is called strongly connected if every pair of nodes $i, j \in C$ are strongly connected. A strongly connected set of nodes C is called **maximal** if none of its nodes are strongly connected with any node outside C. An equivalent expression for maximal strongly connected set is **strong component**. This is the graph-theoretic equivalent of communication class. When the entire set of nodes S is a strong component, we have a **strongly connected graph**. This is indicative of having an irreducible Markov chain. The corresponding Markov chain and Graph theoretical terms are summarized in the subsequent table:

Symbol	Markov chain term	Graph theory term
i	state	node / vertex
(i, j)	transition from i to j	arrow / edge from i to j
$i \to j$	j is accessible from i	$\{i, j\}$ are connected
$i \leftrightarrow j$	i and j communicate	$\{i, j\}$ are strongly connected
C	communication class	strong component
C	closed communication class	-
-	$\mathbf{T}$ is irreducible	D is strongly connected

Graph theory language is useful at this point because efficient algorithms exist for decomposing a graph into its strong components. Several efficient algorithms are available that solve the strong component decomposition in $O(n+|A|)$ CPU time, i.e. in time that is linear in the number of nodes and the number of nonzero elements of the transition matrix. Examples include the **path-based strong component algorithm** [1][2], **Tarjan's algorithm** [3] and **Kosaraju's algorithm** [4]. While Kosaraju's algorithm is slower than the other two algorithms, it is the easiest to understand, and we therefore introduce it here.

We first introduce the **depth-first search** algorithm as an approach to traverse the nodes of a graph by following its arrows:

[1] Dijkstra, Edsger (1976), A Discipline of Programming, NJ: Prentice Hall, Ch. 25.

[2] Gabow, Harold N. (2000), Path-based depth-first search for strong and biconnected components, Information Processing Letters 74 (3-4): 107-114, doi:10.1016/S0020-0190(00)00051-X, MR 1761551.

[3] Tarjan, R. E. (1972), Depth-first search and linear graph algorithms, SIAM Journal on Computing 1 (2): 146-160, doi:10.1137/0201010

[4] S. Rao Kosaraju, 1978 unpublished; Micha Sharir. A strong connectivity algorithm and its applications to data flow analysis. Computers and Mathematics with Applications 7(1):67-72, 1981.

Algorithm 1 DFS(D, v, E): Pseudocode for depth-first search in a digraph D, staring from node v

Input: Digraph D, starting node v, List of found nodes E.
Output: Updated E
Label node v as explored.
For all outgoing arrows $a = (v, w)$:
    If vertex w not explored then DFS(D, w, E)
Append v to list of found nodes E.

Depth-first search starts traversing the graph at some specified starting node v and then returns the set of nodes E that could be reached when starting from v. The set E is a connected set, but it may not be strongly connected, i.e. while all members of E are accessible from v, it is unclear if v is vice versa accessible from all members of E.

Kosaraju's algorithm then uses depth-first search and exploits the fact that the transpose graph of D (the same graph with the direction of every arrow reversed) has exactly the same strongly connected components as D:

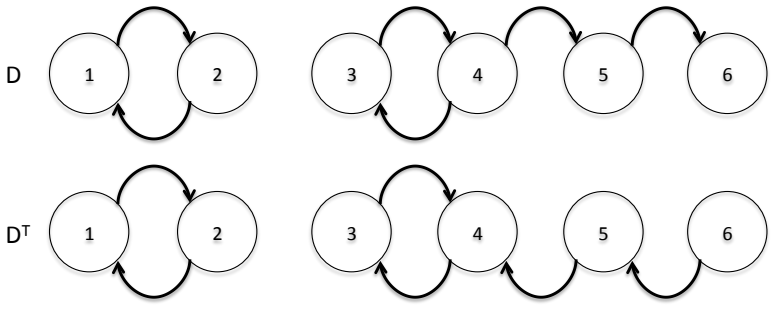
Algorithm 2 Kosaraju(D): Pseudocode of Kosaraju's strong component algorithm

Input: Digraph D
Output: Set of strong components, $\mathcal{C}$
Create empty list $V = ()$.
While V does not contain all vertices:
    Choose an arbitrary vertex v not in V.
    DFS (D, v, V)
Let D^T bet the transpose graph of D (directions of all arcs reversed)
While V is nonempty
    Let v be the last node in V
    Create empty list $C = ()$.
    DFS (D^T, v, C) with C initially empty.
    C is the strong component containing v. Add C to $\mathcal{C}$.
    Remove all nodes in C from the graph D and the list V.

Consider the first example shown above. If we start with node 1, the DFS algorithm would first identify nodes $\{1, 2\}$, if we continue with node 3, then it would subsequently find $\{3, 4, 5, 6\}$. Since the nodes are appended to the S-list once they are completely expanded, S will be:

$$S = (2, 1, 6, 5, 4, 3)$$

We now transpose the graph:



And call DFS starting from the last node in S, which is $v = 3$. We find the set $\{3, 4\}$. These nodes are removed from V and D^T (in fact they are just marked such that we don't work on them again). S is now:

$$V = (2, 1, 6, 5)$$

In the subsequent iterations we find the sets $\{5\}$, $\{6\}$, and finally $\{1,2\}$. The algorithm ends with the strong components:

$$\mathcal{C} = \{\{1,2\},\{3,4\},\{5\},\{6\}\}.$$

Note that strong components are not necessarily closed (e.g. the strong components $\{3,4\}$ and $\{5\}$ are not closed). However, it is easy to check if they are closed. Graph connectivity algorithms may be used to check if the entire state space S is connected under the action of $\mathbf{T}$, or to identify the largest closed communications class, and thereby the largest possible set of states supporting a unique stationary distribution (see below).

Stationary distribution An important property of a Markov chain with transition matrix $\mathbf{T}$ is its **stationary distribution**, or equivalently **equilibrium distribution**. A probability distribution $\boldsymbol{\pi} \in \mathbb{R}^n$ is a stationary distribution of $\mathbf{T}$ when

$$\boldsymbol{\pi}^T\mathbf{T} = \boldsymbol{\pi}^T \tag{1}$$

holds. Applying Eq. (1) n times leads to $\boldsymbol{\pi}^T\mathbf{T}^n = \boldsymbol{\pi}^T$, i.e. when applying the dynamics of the chain to the probability distribution $\boldsymbol{\pi}$, this distribution will never change, i.e. it is stationary, or in other words the distribution is in equilibrium with respect to $\mathbf{T}$. A related concept to stationary distribution is the **invariant measure** $\boldsymbol{\mu} \in \mathbb{R}^n$. For $\boldsymbol{\mu}$, Eq. (1) also holds, but $\boldsymbol{\mu}$ is not necessarily normalized. Thus, $\boldsymbol{\mu}$ can be scaled by an arbitrary constant, and when the elements of $\boldsymbol{\mu}$ sum up to a finite number, the stationary distribution can be obtained from any invariant measure as

$$\boldsymbol{\pi} = \frac{\boldsymbol{\mu}}{\sum_i \mu_i} \tag{2}$$

Depending on the structure of $\mathbf{T}$, a Markov chain may have no stationary distribution, one unique stationary distribution, or infinitely many stationary distributions.

Writing Eq. (1) as

$$\begin{aligned}\boldsymbol{\pi}^T\mathbf{T} &= 1\boldsymbol{\pi}^T \\ \mathbf{T}^T\boldsymbol{\pi} &= 1\boldsymbol{\pi}\end{aligned}$$

we see that this is an eigenvalue equation. Thus, $\boldsymbol{\pi}$ (and any multiple $\boldsymbol{\mu}$) is a left eigenvector of $\mathbf{T}$ with eigenvalue 1. We can also ask for the corresponding right eigenvector and find that

$$\mathbf{T}\mathbf{1} = \mathbf{1}$$

An irreducible Markov chain with a finite state space always has a unique stationary distribution. To validate that a given transition matrix is irreducible, its connectivity should be checked as described in the previous section. If the transition matrix is connected, $\boldsymbol{\pi}$ is unique and can be computed by computing the eigenvector of $\mathbf{T}$ to the eigenvalue 1, and subsequently normalizing that eigenvector according to Eq. (2).

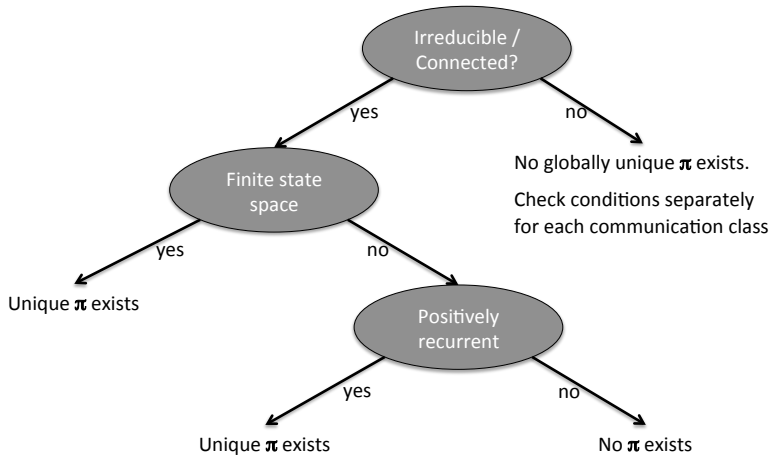
A reducible Markov chain, i.e. a chain with multiple communication classes may support separate independent stationary distributions on each of its communication classes. The stationary distribution should be analyzed for each communication class separately. The global stationary distribution is a linear combination of the individual stationary distributions, subject to the condition that the overall vector sums to 1. If more than one communication class possesses a stationary distribution, this means that there are infinitely many global stationary distributions.

For Markov chains with infinitely many states we additionally need the concept of recurrence. A Markov chain that has an invariant measure $\boldsymbol{\mu}$ whose elements have a finite sum is called **positively recurrent** and has a unique stationary distribution by virtue of normalization (2). However, when the sum of elements of $\boldsymbol{\mu}$ diverges, the Markov chain is called **null recurrent** and has no stationary distribution:

$$\begin{aligned}\sum_i \mu_i < \infty \qquad & \text{positively recurrent} \\ \sum_i \mu_i = \infty \qquad & \text{null recurrent.}\end{aligned}$$

Mathematically, the existence of a stationary distribution on such infinity chains simply depends on whether $\boldsymbol{\mu}$ is normalizable or not. Intuitively, a positively recurrent Markov chain has an "attractive" transition

probability structure such that finite-sized sets of states can hold the probability, while a null recurrent Markov chain has a "dissipative" transition probability structure, such that any probability will vanish into the infinite state space. Example c shows such a case.



Example a: The transition matrix

$$\mathbf{T} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

has a unique stationary distribution for $\alpha + \beta > 0$. It is easy to check that an invariant measure is

$$\boldsymbol{\mu}^T = (\beta, \alpha)$$

and the unique stationary distribution thus is

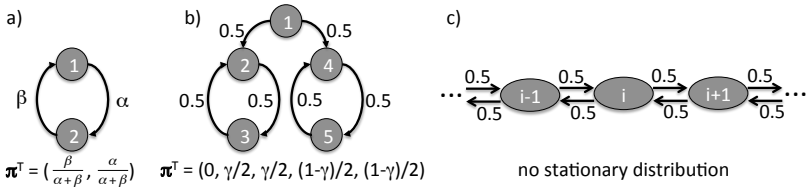
$$\boldsymbol{\pi}^T = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta} \right).$$

Example b: shows a reducible Markov chain, consisting of the communication classes {{1},{2,3},{4,5}}. Therefore, we can analyze the stationary distribution for each communication class separately. Set {1} is not closed, and does therefore not support a stationary distribution - any probability will flow out of {1} into either {2,3} or {4,5}, when the Markov chain is propagated. Sets {2,3} and {4,5} each have a stationary distribution of $(\frac{1}{2}, \frac{1}{2})$, but since these two sets are not connected, their two stationary distributions can be arbitrarily scaled with respect to one another. Thus, the Markov chain shown in b) has infinitely many stationary distributions, which can be parametrized by:

$$\boldsymbol{\pi}^T = \left(0, \frac{\gamma}{2}, \frac{\gamma}{2}, \frac{1-\gamma}{2}, \frac{1-\gamma}{2} \right)$$

where $\gamma \in [0, 1]$ is a free parameter. Any vector $\boldsymbol{\pi}$ of the above form will obey Eq. (1) and thus is a stationary distribution.

Example c: Consider the simple random walk on an infinite series of sites in a line (Example c below). An invariant measure is the infinite 1-vector, $(1, 1, ...)$. However, this invariant measure is not normalizable as $\sum_i \mu_i = \infty$, and therefore does not give rise to a stationary distribution.



Ergodicity Under appropriate conditions it is possible to sample from the stationary distribution $\boldsymbol{\pi}$ by simulating the Markov chain. In order to characterize the ability to compute $\boldsymbol{\pi}$ in this way, the concept of ergodicity is important.

Formally, a Markov chain that has a unique stationary distribution $\boldsymbol{\pi}$ is **ergodic** when

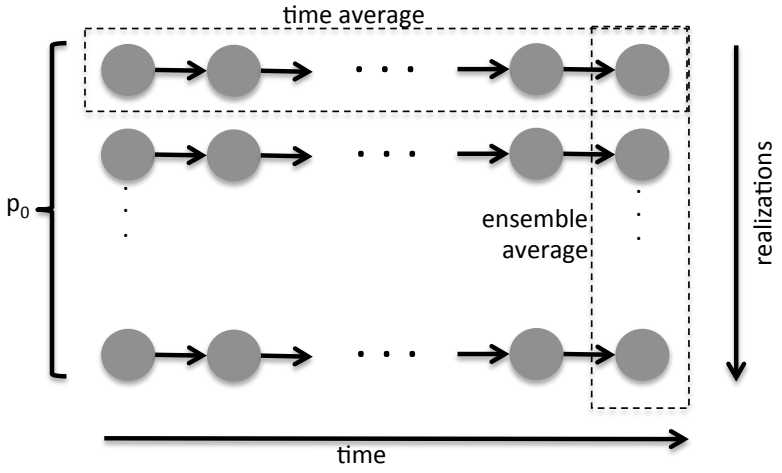
$$\lim_{n\to\infty} \mathbf{T}^n = \begin{bmatrix} \boldsymbol{\pi}^T \\ \vdots \\ \boldsymbol{\pi}^T \end{bmatrix}, \tag{3}$$

i.e. if any arbitrary starting distribution will converge to $\boldsymbol{\pi}$ under the dynamics of the Markov chain.

In an ergodic Markov chain, equilibrium quantities can be sampled either using time or ensemble averages. Consider an arbitrary function of state, $f(i)$, which assigns a value to each Markov state i. The expectation value of f is then $\mathbb{E}[f] = \sum_i \pi_i f(i)$. Here we consider two simulation setups: (1) a single Markov chain realization x_t, and (2) an ensemble of simulations started from initial probability distribution $\mathbf{p}_0$. In an ergodic chain the following limits are equal:

$$\underbrace{\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n} f(x_t)}_{\text{time average}} \quad = \quad \underbrace{\lim_{n\to\infty} \sum_i (\mathbf{p}_0 \mathbf{T}^n)_i f(i) = \sum_i \pi_i f(i) = \mathbb{E}[f]}_{\text{ensemble average}} \tag{4}$$

Note that a special case of $\mathbb{E}[f]$ is a vectorial function where $f(i)$ is a vector which is 1 at element i and 0 elsewhere, such that $\mathbb{E}[f]$ is the stationary distribution $\boldsymbol{\pi}$ itself. The equivalence of time average and ensemble average means that $\boldsymbol{\pi}$ and any equilibrium expectation $\mathbb{E}[f]$ can be sampled either by (1) simulating a single Markov chain for a long time, and taking the time average of states sampled; or (2) by simulating a large number of realizations starting from arbitrary starting conditions in parallel, for a sufficiently long time such that they decorrelate from their starting conditions, and then taking the ensemble average over the different realizations. This is illustrated in the subsequent scheme:



When is a Markov chain ergodic? It is not sufficient that a unique stationary distribution exists. For example the transition matrix:

$$\mathbf{T} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

has powers

$$\mathbf{T}^n = \begin{cases} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & n \text{ even} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & n \text{ odd} \end{cases}$$

and therefore does not converge. This is an example of a **periodic** chain: When simulating the chain it will visit each of its two states with a period of two steps. In general, a state is called periodic when the number of time steps at which it can be revisited has a greatest common divisor (gcd) of 2 or greater, and it is aperiodic when they have a gcd of 1. For example, a state which can be revisited after 4,6,8,... timesteps is periodic (gcd = 2) while a state which can be revisited after 4,5,6,... timesteps is not (gcd = 1). States within the same communication class have the same period, justifying the term **period of a communication class**. When all communication classes are aperiodic, the Markov chain is aperiodic.

In general, a sufficient condition for ergodicity, and thus for Eq. (3) to hold, is that the Markov chain is both irreducible and aperiodic. Since irreducible Markov chains have only one communication class, the entire chain is aperiodic if a single state is aperiodic. In summary, ergodicity can be shown by showing that the transition matrix is connected and has at least one state i with $T_{ii} > 0$.

Spectral decomposition The eigenvalues and eigenvectors of a transition matrix are of special interest, as they bear a lot of information about the stationary and kinetic properties of the ensemble dynamics of the Markov chain. First, notice that we can find both a left and right eigenvectors to the same eigenvalue λ_i:

$$\begin{aligned} \mathbf{T}\mathbf{r}_i &= \mathbf{r}_i\lambda_i && (5)\\ \mathbf{l}_i^T\mathbf{T} &= \lambda_i\mathbf{l}_i^T. && (6) \end{aligned}$$

Most eigenvector solvers, such as those in standard computer algebra packages, provide the right eigenvectors $\mathbf{r}_i$ by default. However, the second equation can be equivalently written as $\mathbf{T}^T\mathbf{l}_i = \lambda_i\mathbf{l}_i$, therefore the left eigenvectors can be simply obtained by computing the right eigenvectors of the transposed matrix $\mathbf{T}^T$.

We will now assume that the matrix $\mathbf{T}$ is diagonalizable (there are pathological cases where this is not possible), and define the following matrices:

$$\begin{aligned} \mathbf{\Lambda} &= \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = \mathrm{diag}(\lambda_1, ..., \lambda_n) \\ \mathbf{R} &= \begin{bmatrix} \mathbf{r}_1 & \cdots & \mathbf{r}_n \end{bmatrix} \\ \mathbf{L} &= \begin{bmatrix} \mathbf{l}_1^T \\ \vdots \\ \mathbf{l}_n^T \end{bmatrix}. \end{aligned}$$

These are the diagonal matrix of eigenvalues, $\mathbf{\Lambda}$, the matrix with right eigenvectors in the columns, $\mathbf{R}$, and the matrix with left eigenvectors in the rows, $\mathbf{L}$. We can then write all the eigenvalue problems $i = 1, ..., n$ in on set of matrix-vector equations as:

$$\begin{aligned} \mathbf{T}\mathbf{R} &= \mathbf{R}\mathbf{\Lambda} \\ \mathbf{L}\mathbf{T} &= \mathbf{\Lambda}\mathbf{L} && (7) \end{aligned}$$

allowing us to rewrite the transition matrix $\mathbf{T}$ in terms of its eigenvalue decomposition:

$$\begin{aligned} \mathbf{T} &= \mathbf{R}\mathbf{\Lambda}\mathbf{R}^{-1} \\ &= \mathbf{L}^{-1}\mathbf{\Lambda}\mathbf{L} \\ &= \mathbf{R}\mathbf{\Lambda}\mathbf{L} = \sum_{i=1}^{n} \lambda_i \mathbf{r}_i \mathbf{l}_i^T && (8) \end{aligned}$$

Using this decomposition we can understand the action of the eigenvectors and eigenvalues in the Markov chain dynamics. Let us consider the Chapman-Kolmogorow equation, describing the evolution of an ensemble probability $\mathbf{p}(0)$ under the dynamics of the Markov chain with transition matrix $\mathbf{T}$. We use the

eigenvalue decomposition of $\mathbf{T}$, allowing us to rewrite:

$$\begin{aligned}\mathbf{p}_k^T &= \mathbf{p}_0^T \mathbf{T}^k \\ &= \mathbf{p}_0^T \sum_{i=1}^{n} \lambda_i^k \mathbf{r}_i \mathbf{l}_i^T \\ &= \sum_{i=1}^{n} \lambda_i^k \langle \mathbf{p}_0, \mathbf{r}_i \rangle \mathbf{l}_i^T \end{aligned} \tag{9}$$

where $\langle , \rangle$ denotes the scalar product. We can therefore understand the probability vector at any time, $\mathbf{p}_k$, to be formed as the superposition of left eigenvectors $\mathbf{l}_i$. The left eigenvectors form a basis in which we can express probability vectors.

Using the Perron-Frobenius theorem[5], it can be shown that a finite-state transition matrix that is irreducible (connected) has one and only one Eigenvalue $\lambda_1 = 1$. We have seen above that this 1-Eigenvalue (also called Perron-Frobenius eigenvalue) is associated with the left eigenvector $\boldsymbol{\pi}$, i.e. the stationary distribution, and the right eigenvector $\mathbf{1}$. Inserting this as a first term into Eq. (9) yields (using $\langle \mathbf{p}_0, \mathbf{1} \rangle = \sum_i p_{0,i} = 1$):

$$\mathbf{p}_k^T = \boldsymbol{\pi}^T + \sum_{i=2}^{n} \lambda_i^k \langle \mathbf{p}_0, \mathbf{r}_i \rangle \mathbf{l}_i^T \tag{10}$$

This expansion has the following properties:

- The ensemble probability distribution at any time, $\mathbf{p}_k$, can be expressed as a superposition of the stationary distribution $\boldsymbol{\pi}$ and some "distortion" which is a superposition of the left eigenvectors. Each left eigenvector can thus be understood as a probability distortion.

- Following the Perron-Frobenius theorem, $\mathbf{T}$ has p eigenvalues of norm 1 where p is the period of the Markov chain, while all other $n - p$ eigenvalues have a norm strictly smaller than 1. Thus, if the Markov chain is aperiodic, and thereby ergodic, and we order eigenvalues by decreasing norm, we have:

  $$\begin{aligned}\lambda_1 &= 1 \\ |\lambda_2| &< 1 \\ |\lambda_k| &< |\lambda_{k+1}| \quad k = 2, ..., n\end{aligned}$$

  and we see that, consistently with Eq. (4), for an ergodic Markov chain any initial distribution $\mathbf{p}_0$ will converge towards the stationary distribution, as the terms $\lambda_i^k \to 0$ with $k \to \infty$:

  $$\lim_{k\to\infty} \mathbf{p}_k^T = \boldsymbol{\pi}^T + \lim_{k\to\infty} \sum_{i=2}^{n} \lambda_i^k \langle \mathbf{p}_0, \mathbf{r}_i \rangle \mathbf{l}_i^T = \boldsymbol{\pi}^T$$

  only if the Markov chain is periodic, it will have other (complex) eigenvalues with norm $|\lambda_i| = 1$ that do provide terms that oscillate in time and do not decay with $k \to \infty$.

- For an ergodic Markov chain, the amplitudes $\lambda_i^k \langle \mathbf{p}_0, \mathbf{r}_i \rangle$ of the distortions from equilibrium decay exponentially with discrete time k. By comparing this decay with an exponential function, $|\lambda_i^k| = \exp(-\kappa_i k)$, one may associate a decay/relaxation rate, κ_i, or a decay/relaxation timescale, t_i:

  $$\begin{aligned}\kappa_i &= -\ln |\lambda_i| \\ t_i &= -\frac{1}{\ln |\lambda_i|}\end{aligned}$$

  when the time step of the Markov chain is associated to a real time, the relaxation rate or timescale is measured in units of that time.

[5] Perron, Oskar (1907), "Zur Theorie der Matrices", Mathematische Annalen 64 (2): 248–263, doi:10.1007/BF01449896
Frobenius, Georg (1912), "Ueber Matrizen aus nicht negativen Elementen", Sitzungsber. Königl. Preuss. Akad. Wiss.: 456–477
Frobenius, Georg (1908), "Über Matrizen aus positiven Elementen, 1", Sitzungsber. Königl. Preuss. Akad. Wiss.: 471–476
Frobenius, Georg (1909), "Über Matrizen aus positiven Elementen, 2", Sitzungsber. Königl. Preuss. Akad. Wiss.: 514–518

- Each deviation from equilibrium enters with a constant contribution $\langle \mathbf{p}_0, \mathbf{r}_i \rangle$, which measures how much the initial distribution overlaps with the ith eigenvector. For the special case that the initial distribution deviates from the equilibrium distribution only by a multiple c_i of one of the Markov chain eigenvectors, we get the behavior:

$$\mathbf{p}_k^T = \boldsymbol{\pi}^T + \lambda_i^k c_i \mathbf{l}_i^T = \boldsymbol{\pi}^T + c_i \exp(-\kappa_i k)\mathbf{l}_i,$$

  i.e. a single exponential relaxation to the equilibrium distribution with rate κ_i

- From Eq. (10) we can compute the speed of convergence to the stationary distribution of an ergodic Markov chain. When starting an ensemble from initial distribution $\mathbf{p}_0$, the deviation (in some suitable metric $\|\cdot\|$, e.g. the two-norm) of the ensemble distribution from the stationary distribution will be:

$$\begin{aligned} E(k) &= \left\|\boldsymbol{\pi}^T - \mathbf{p}_k^T\right\| = \left\|\sum_{i=2}^{n} \lambda_i^k \langle \mathbf{p}_0, \mathbf{r}_i \rangle \mathbf{l}_i^T\right\| \\ &= \lambda_2^k \left\|\langle \mathbf{p}_0, \mathbf{r}_2 \rangle \mathbf{l}_2^T + \sum_{i=3}^{n} \left(\frac{\lambda_i}{\lambda_2}\right)^k \langle \mathbf{p}_0, \mathbf{r}_i \rangle \mathbf{l}_i^T\right\| \\ &\leq \lambda_2^k \left\|\sum_{i=2}^{n} \langle \mathbf{p}_0, \mathbf{r}_i \rangle \mathbf{l}_i^T\right\| = \exp(-\kappa_2 k) \left\|\sum_{i=2}^{n} \langle \mathbf{p}_0, \mathbf{r}_i \rangle \mathbf{l}_i^T\right\| \end{aligned}$$

  and can thus be bounded by an exponential decay with the slowest timescale t_2 (rate κ_2) in the system. For times much greater than $k \gg t_3 = \kappa_3^{-1}$, all but the first term in line two vanish and we obtain the approximation:

$$E(k) = \lambda_2^k \left\|\langle \mathbf{p}_0, \mathbf{r}_2 \rangle \mathbf{l}_2^T\right\|$$

Example A:

The 2×2 matrix

$$\mathbf{T} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

depicted in Fig. (2) has the eigenvalues $(1, 1-a-b)$. Associated to eigenvalue 1 we find a left eigenvalue that can be normalized to the stationary distribution:

$$\boldsymbol{\pi}^T = \left(\frac{b}{a+b}, \frac{a}{a+b}\right)$$

which is immediately intuitive: when $b = a$, states 1 and 2 will have the same stationary probability. When $b > a$, more stationary probability will be in state 1, while $a > b$ means that more stationary probability will be in state 2. The second eigenvector

$$\mathbf{l}_2^T = (-1, 1)$$

is associated with shifting probability between states 1 and 2, i.e. the transition between the two states. Its eigenvalue $1 - a - b$ is associated to the rate

$$\kappa_2 = -\ln(1-a-b).$$

If a and b are very small parameters, i.e. the transitions between states 1 and 2 are rare events, we can use the first-order Taylor approximation to the logarithm around 1: $\ln(x) \approx x - 1$, and obtain the approximate relaxation rate:

$$\kappa_2 \approx a + b$$

or relaxation timescale $t_2 = (a+b)^{-1}$.

Example B:

The system

$$\mathbf{T} = \begin{bmatrix} 0.9 & 0.1 & & \\ 0.1 & 0.89 & 0.01 & \\ & 0.01 & 0.79 & 0.2 \\ & & 0.2 & 0.8 \end{bmatrix}$$

has a uniform stationary distribution, resulting from the symmetry of $\mathbf{T}$. The more interesting aspect is that this Markov chain has a hierarchy of timescales: 3 and 4 communicate most rapidly, 1 and 2 communicate less rapidly, and the sets $\{1,2\}$ and $\{3,4\}$ only communicate slowly. Such a system is called metastable: the sets $\{1,2\}$ and $\{3,4\}$ are metastable sets, i.e. the Markov chain stays in one of these sets relatively long until it switches to the other set. Metastability is apparrent from the relaxation timescales that can be calculated from the eigenvalues:

$$\begin{aligned} t_2 &= 99.50 \\ t_3 &= 4.24 \\ t_4 &= 1.90 \end{aligned}$$

showing clearly a timescale gap $t_2 \gg t_3$. Clearly, the number of metastable sets generally depends on the timescale of interest. In this system, only one clear timescale gap exists, while other systems may have multiple (or no clear) timescale gaps, and thus different numbers of metastable sets, depending what timescales one defines to be the threshold between "fast" and "slow".

The eigenvectors depicted in Fig. 2B exhibit the dynamic processes associated with each timescale. For example, the second eigenpair, associated with timescale $t_2 = 99.5$ is associated with an eigenvector that changes sign between sets $\{1,2\}$ and $\{3,4\}$. Therefore, the slowest relaxation process is indeed clearly associated with the transition between these two metastable sets. Also the other two timescales can be clearly associated to structural processes: t_3 is assigned to the transition $1 \leftrightarrow 2$, because the corresponding eigenvector has a sign change between these two states. Moreover, the elements 3 and 4 of $\mathbf{l}_3$ are nearly 0, indicating that the process mainly takes place between 1 and 2. Similar, the fastest timescale t_4 can be associated to $3 \leftrightarrow 4$.

Example C:

The system

$$\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

is periodic with a periodicity of 3 steps. Indeed it has three eigenvalues with norm one: the eigenvalues 1, $-0.5 - i0.87$, $-0.5 + i0.87$, which lie on the unit circle in the complex plane. The eigenvalue 1 is unique because the system is irreducible. The other two eigenvalues form a complex conjugate pair. The system is not ergodic, and does in fact not have a single finite relaxation timescale. Thus, although it has a unique stationary distribution $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, this stationary distribution can not be reached from an arbitrary starting condition $\mathbf{p}_0$ because the limit $\lim_{k\to\infty} \mathbf{T}^k$ does not exist.

Example D:

The system

$$\mathbf{T} = \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0 & 0.9 & 0.1 \\ 0.1 & 0 & 0.9 \end{bmatrix}$$

also has a circular structure that breaks reversibility (the detailed balance equations are not fulfilled), however the system is no periodic anymore. This is because due to the transition probabilities smaller than 1, each state can be revisited at any time of at least 3 steps, giving rise to a gcd of return times of 1. As a result, the system is ergodic, and has only one eigenvalue of norm 1. Again the eigenvalues 2 and 3 are complex conjugate. Since the system is ergodic, both are associated to a timescale, t_2 and t_3, which are identical 6.37.

Reversible Markov chain / Detailed balance A Markov chain with a unique stationary distribution that is positive everywhere is additionally **reversible** if the absolute probability to see a "forward" transition $i \to j$ is the same as the "backward" transition $j \to i$, for all pairs i, j. Using the definition of conditional probability, this can be formulated as:

$$\begin{aligned} \mathbb{P}(x_k = i,\, x_{k+1} = j) &= \mathbb{P}(x_k = j,\, x_{k+1} = i) \\ \mathbb{P}(x_k = i)\,\mathbb{P}(x_{k+1} = j \mid x_k = i) &= \mathbb{P}(x_k = j)\,\mathbb{P}(x_{k+1} = i \mid x_k = j) \\ \pi_i T_{ij} &= \pi_j T_{ji} \qquad \forall i, j \in S \end{aligned} \tag{11}$$

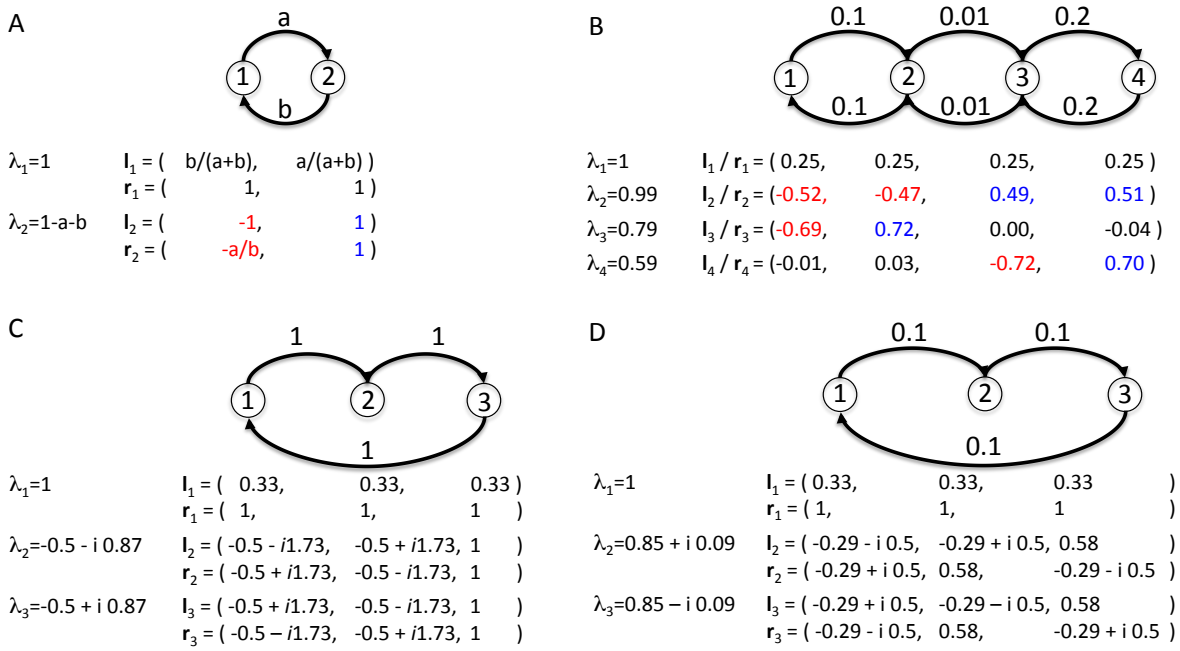


Figure 2: Spectral decomposition

where the conditions in the last row are known as the **detailed balance** conditions. Row 1 can be interpreted as time being reversed, hence the term reversible Markov chain. We can consequently define a backward propagator, which will be used later, as:

$$\tilde{T}_{ij} = \frac{\pi_j}{\pi_i} T_{ji}$$

in the reversible case, the forward and the backward propagator are identical:

$$\tilde{\mathbf{T}} = \mathbf{T}.$$

Using the matrix

$$\mathbf{\Pi} = \begin{bmatrix} \pi_1 & & 0 \\ & \ddots & \\ 0 & & \pi_n \end{bmatrix} = \mathrm{diag}(\pi_1, ..., \pi_n)$$

we can write all detailed balance equations (11) in one matrix equation:

$$\mathbf{C} = \mathbf{\Pi T} = (\mathbf{\Pi T})^T = \mathbf{C}^T. \tag{12}$$

where we have defined the matrix $\mathbf{C} = \mathbf{\Pi T}$ containing the absolute transition probabilities $\mathbb{P}(x_k = i, x_{k+1} = j)$ as elements. $\mathbf{C}$ is called the **correlation matrix**. For a reversible Markov chain, the correlation matrix is symmetric. Inserting the Eigenvalue decomposition of $\mathbf{T}$ (Eq. 7) into Eq. (12), we find that:

$$\mathbf{\Pi R \Lambda L} = \mathbf{L}^T \mathbf{\Lambda \Pi R}$$

and therefore:

$$\begin{aligned} \mathbf{L}^T &= \mathbf{\Pi R} \\ \mathbf{l}_i^T &= \mathbf{\Pi r}_i. \end{aligned}$$

That is, for a reversible Markov chain / a transition matrix fulfilling detailed balance, the right and left eigenvectors can be interconverted using the stationary probability distribution. We can thus express all

quantities using either the set of left or right eigenvectors. In particular, the spectral decomposition of the ensemble dynamics can be written as:

$$\begin{aligned}\mathbf{p}_k &= \mathbf{\Pi}\sum_{i=1}^{n}\lambda_i^k\langle\mathbf{p}_0,\mathbf{r}_i\rangle\mathbf{r}_i^T \\ &= \sum_{i=1}^{n}\lambda_i^k\langle\mathbf{p}_0,\mathbf{\Pi}^{-1}\mathbf{l}_i\rangle\mathbf{l}_i^T\end{aligned}$$

Furthermore, summing the detailed balance equations (11) over i gives us:

$$\begin{aligned}\sum_i \pi_i T_{ij} &= \pi_j \quad \forall j \in S \\ \boldsymbol{\pi}^T\mathbf{T} &= \boldsymbol{\pi}^T\end{aligned}$$

This means: if a distribution $\boldsymbol{\pi}$ which symmetrizes the transition matrix $\mathbf{T}$, so that the detailed balance equations (11) are obtained, $\boldsymbol{\pi}$ is a stationary distribution. If $\mathbf{T}$ is also finite and irreducible, $\boldsymbol{\pi}$ is unique (see stationary distribution, above).

Examples:

Fig. 2A,B represent reversible Markov chains, whereas Fig. 2C,D do not exhibit detailed balance. Every tridiagonal transition matrix with positive off-diagonal elements can be shown to exhibit detailed balance. One consequence of this is that every 2×2 matrix with positive off-diagonal elements exhibits detailed balance. This can be easily seen by using the stationary distribution $\boldsymbol{\pi}^T = \left(\frac{T_{12}}{T_{12}+T_{21}}, \frac{T_{21}}{T_{12}+T_{21}}\right)$ in the detailed balance equation $\pi_1 T_{12} = \pi_2 T_{21}$.