1. Problem: Markov chains (10 points)

1.1. Definition

The most simple class of Markov processes are the so-called *Markov chains*, which are characterized by the following properties:

- The discrete time-variable t takes only only integer values $t = n \in \mathbb{Z}$.
- The random process $\hat{x}(t)$ only assumes a discrete set of values, which can be labeled by integers i.
- The process is time-homogeneous.

An even simpler special case are *finite* Markov chains, for which additionally the number of states is finite: $i = 1, \ldots, R$ for some integer R.

1.2. Matrix of transition probabilities and stationary solutions

The transition probability $p_{1|1}(x_2, t_2|x_1, t_1)$ of a Markov chain can be written as a (possibly infinitely dimensional) matrix $\mathsf{P}^{(n)}$ with components $P^{(n)}_{i_2,i_1}$, which depends only on the discrete time-difference $n = t_2 - t_1 \in \mathbb{Z}$. Furthermore, the one-time probabilities $p_1(x, t)$ can be written as a vector $\mathbf{p}^{(n)}$.

(a) Derive from the Chapman-Kolmogorov equation for Markov processes equation the relation

$$\mathsf{P}^{(n)} = \mathsf{P}^n \,, \tag{1}$$

where $\mathsf{P} := \mathsf{P}^{(1)}$ is the matrix of transition probabilities for a single time-step $n \to n + 1$.

(b) Write $\mathbf{p}^{(n)}$ in terms of the initial probabilities $\mathbf{p}^{(0)}$ at time $n = 0$?

(c) Obviously, the transition probabilities P_{ij} have to be non-negative. Additionally, they have to fulfill the condition

$$\sum_i P_{ij} = 1 \,. \tag{2}$$

Why?

(d) Any normalized vector $\mathbf{p}_\mathrm{s}$ with

$$\mathsf{P}\,\mathbf{p}_\mathrm{s} = \mathbf{p}_\mathrm{s} \tag{3}$$

is called a *stationary solution*. Show that this condition indeed guarantees that $\mathbf{p}^{(n)} = \mathbf{p}_\mathrm{s}$ is a solution of the Markov chain, i.e., conforms to the dynamics as described in (b).

Derive from Eq. (2) that such a solution must always exist. Hint: Write Eq. (2) in the form of an eigenvalue problem for the matrix P^T, i.e., the transpose of P. What does this imply for the eigenvalue problem of P?

1.3. Example: Cumulative maximum

Suppose a device measures a quantity that assumes integer values $1, \ldots, R$ at discrete times $n = 1, 2, \ldots$ and only records the largest value observed so far. We assume that the measured values $\hat{x}_n$ are independent and identically distributed with probabilities $q_i := \mathrm{Prob}(\hat{x}_n) = i$. The recorded result can be written as

$$\hat{m}_n = \max(\hat{x}_1, \ldots, \hat{x}_n)\,. \tag{4}$$

(a) Argue that $\hat{m}_n$ is a finite Markov chain and express the matrix P of the transition probabilities in terms of the probabilities q_i.

(b) Now, consider the special case of a uniform distribution $q_i = 1/R$ of the $\hat{x}_n$. Calculate the probability $p_m^{(n)} = \mathrm{Prob}(\hat{m}_n = m)$ either by directly using the definition (4) of the process (Hint: What are all possible ways to obtain the result m at time n?) or by making use of relation (1).

Discuss the long-time behavior of the two border cases $p_1^{(n)}$ and $p_R^{(n)}$.

Find a stationary solution for this problem.

(c) Now consider again the general case and derive an expression for the probabilities $p_m^{(n)}$.

2. Metropolis-Hastings algorithm (no points)

2.1. Motivation: Calculating weighted averages by sampling

An important application of Markov chains arises in the context of Monte-Carlo simulations for the evaluation of weighted averages of the form

$$\bar{f} := \sum_i f_i\, w_i\,. \tag{5}$$

Here, f_i is the quantity to be averaged and $w_i > 0$ is a normalized weighting function: $\sum_i w_i = 1$. A typical example is the *thermal average* with $w_i = (1/Z)\exp(-E_i/k_\mathrm{B}T)$ where E_i is the energy of a state i, k_B denotes Boltzmann's constant, T is the temperature and the normalization factor is given by the partition function $Z = \sum_i \exp(-E_i/k_\mathrm{B}T)$.

When the weighting factor w_i is rather "peaked", i.e., depends strongly on the state i, it becomes rather inefficient to directly perform the sum in Eq. (5): One will most often hit states with $w_i \approx 0$, which do not contribute noticeably to the result. In such a situation, it is advantageous to reinterpret the sum as the expectation value of the random variable $f_{\hat{i}}$ where the $\hat{i}$ are drawn from the distribution w_i, i.e., $\mathrm{Prob}(\hat{i} = i) = w_i$:

$$\bar{f} = \langle f_{\hat{i}} \rangle \,. \tag{6}$$

For the evaluation of the right-hand side of this relation, one then samples a large number N of states $i_1, \ldots, i_N$ according to this distribution and approximates

$$\bar{f} \approx \frac{1}{N} \sum_{n=1}^{N} f_{i_n} \,. \tag{7}$$

This average will then typically yield a much improved approximation for $\bar{f}$ already for a rather small number of samples N (compared to the total number of states i).

This leaves us with the task of how to efficiently sample from a given distribution w_i, This problem often can be solved by the so-called Metropolis algorithm, or its generalization, the so-called Metropolis-Hastings algorithm.

2.2. Markov chain for Metropolis-Hastings algorithm

The basic idea of the Metropolis-Hastings algorithm is to generate samples from a given distribution as the stationary solution of a suitably constructed Markov chain. The stochastic dynamics described by the Markov chain is then simulated by a Monte-Carlo algorithm, which yields (after stationarity has been achieved) for every time-step a new sample. Doing so, implicitly replaces the ensemble average (7) by a time-average. We will not be able to discuss in detail here the underlying assumption of "ergodicity", which is based on Eq. (9) below, but just consider the concept of the algorithm.

We first choose a *proposal distribution* Q_{ij} for the transition probabilities of the Markov chain. As the name says, it generates transitions between the different states—not all of which will be "accepted", see below. In particular, it has to allow the process to reach (possible after several steps) from every starting state any state in the region we are interested in.[1] Technically, we require that $Q_{ij} = 0$ if and only if $Q_{ji} = 0$.

[1] The proposal distribution has to be chosen in such a way that sampling from it is easily possible. Often one thus employs a Gaussian "centered" around j or assumes a uniform distribution on a certain subset of states directly reachable from j. The appropriate choice can often be an art!

We then define the transition matrix of the Markov chain as[2]

$$P_{ij} := Q_{ij} \min\left(1, \frac{w_i\, Q_{ji}}{w_j\, Q_{ij}}\right) \quad \text{for } i \neq j \tag{8}$$

Why is it enough to give only the expression for the off-diagonal matrix elements?

Show that $\mathbf{w}$ is a stationary solution of this Markov chain which additionally fulfills the so-called *detailed balance* condition

$$P_{ij}\, w_j = P_{ji}\, w_i\,. \tag{9}$$

Hint: First verify this condition and then derive from it the stationarity (3).

How can Eq. (9) be interpreted in general and in the particular case of a thermal average?

Finally, we come to the question on how to simulate a Markov chain described by transition probabilities of the form (8). Such a simulation starts from an arbitrary initial state j and then jumps to a new state $i \neq j$, randomly according to the probability P_{ij} for fixed j; or just remains—with probability P_{jj}—in its original state.

In order to sample from these possible events with the correct probability, we first randomly draw a state i according to the proposal distribution Q_{ij}. Now, if $i = j$, the state remains the same. If $i \neq j$ we only jump to the new state i ("accept" the result) when

- either $\dfrac{w_i\, Q_{ji}}{w_j\, Q_{ij}} \geq 1$
- or a random number drawn uniformly from the interval $[0, 1]$ is smaller than $\dfrac{w_i\, Q_{ji}}{w_j\, Q_{ij}} < 1$.

Otherwise, we stay in the original state j. Show that this method results in the desired jump probability P_{ij}.

Interpret the acceptance criterion in the case of a thermal average with symmetric $Q_{ij} = Q_{ji}$.

This procedure is then repeated starting from the new (old) state i (j). After some time, the dynamics will "equilibrate" and the samples will be drawn from the stationary distribution $\mathbf{w}$. Then, every subsequent sample can be used, e.g., in the average (7). Convince yourself that correlations in the samples are not a problem.

Finally, notice that the expression (8) only contains ratios of the weighting factors w_i. Why can this be crucial?

[2]Note that in many cases, the proposal distribution is chosen to be symmetric: $Q_{ij} = Q_{ji}$. Then one finds $P_{ij} = Q_{ij} \min(1, w_i/w_j)$ for $i \neq j$. This was the case for the original Metropolis algorithm.