Course mini-project: Chain identification

What is the aim?

The aim of this mini-project is to 1. Find properties (e.g. transition probabilities, stationary distribution, convergence rate) of Markov Chains by looking at their samples, and 2. Use these chains to sample from an arbitrary distribution - by using Metropolis Hasting algorithm.

What are the files you received?

You have access to 4 MATLAB functions which are saved as `.p` files - e.g. `chain_1.p`. The essential property of `.p` files is that you can execute them on your computer, but you do not have access to their source code. In this project, you are supposed to find the source code of these functions by executing them!

Each function is corresponding to a Markov Chain (not necessarily time-homogeneous) on the state space $\{1, 2, 3, 4, 5\}$. You can sample from each of these chains by executing its corresponding function. For example, to sample from the 1st chain, you can use the code `X = chain_1(N_chain, Time, pi0)`. It produces `N_chain` $\in \mathbb{N}$ realisations of chain 1 with length `Time` $\in \mathbb{N}$ and initial distribution $\pi^{(0)} \in \mathbb{R}^5$ equal to `pi0`. Each column of the matrix `X` is corresponding to one of the realisations. For example, consider the code

```
N_chain = 5;
Time = 100;
pi0 = ones(1,5)/5;
X = chain_1(N_chain, Time, pi0);
```

As the result of this code, `X` is a 100 by 5 matrix where `X(t,n)` corresponds to X_{t-1} of realisation n of chain 1. As it is defined, X_0 is distributed uniformly over $\{1, 2, 3, 4, 5\}$ for all realisations.

Note that, if `pi0` is an integer (in the set $\{1, 2, 3, 4, 5\}$) rather than a probability distribution, then the function uses it as the first state of the chain, i.e. $X_0 =$ `pi0` which is equivalent to consider $\pi_i^{(0)} = \delta(i, \texttt{pi0})$.

What should you do? Section 1: Identify chains!

Since the chains are not necessarily time homogeneous, let us use a notation different from the course: We show the transition probability matrix at time t by $P^{(t)}$, i.e. the probabilities of transitions from `X(t+1,:)` to `X(t+2,:)`. Note that this is not equivalent to what you had seen in the course for $P^{(t)}$, which were the transition probabilities in t steps for time homogeneous chains, i.e. the probabilities of transitions from `X(1,:)` to `X(t,:)`.

For each of the 4 chains (i.e. `chain_1.p`, `chain_2.p`, `chain_3.p`, and `chain_4.p`),

a) Estimate $P^{(t)}$, and plot the values of its elements over time - for a reasonable range of time.

b) Is this chain time-homogeneous?

c) If it is time-homogeneous:

c.1) Find its time-homogeneous transition probability and save it as the variable `P_hat` in the file

`P_hat_chain_i.mat` where `i` is the number of the chain. Choose `N_chain` and `Time` in a way to have an error (on average) less than 10^{-3} for each elements of P - explain your approach.

c.2) Draw the underlying graph of the chain.

d) If it is not time-homogeneous, explain the dynamics of $P^{(t)}$ over time, e.g. how it changes, whether it converges, etc.

e) When `pi0` is an uniform distribution, estimate $\pi^{(t)}$ (i.e. the distribution of states at time t, i.e. `X(t+1,:)`), and plot the values of its elements over time.

f) Does there exist any limiting distribution for this chain? If yes, save it as the variable `pi_hat` in the file `pi_hat_chain_i.mat` where `i` is the number of the chain. Choose `N_chain` and `Time` in a way to have an error (on average) less than 10^{-3} for each element of π - explain your approach. Furthermore, plot the values of the limiting distribution as a bar-plot.

g) For the case where the limiting distribution π exists, plot the total-variation distance over time. Plot the total-variation distance over time also for the cases where $X^{(0)} = i, \forall i \in \{1, 2, 3, 4, 5\}$. Which initial state has the worst convergence rate? Can you estimate (numercically) an upper-bound for T_ϵ when $\epsilon = 0.005$?

h) If the chain is time-homogeneous, find its stationary distribution using the eigenvalue decomposition of your estimation of P. Check whether your findings are consistent with what you found in parts **f)** and **g)**. Furthermore, plot the stationary distribution with bar-plots.

What should you do? Section 2: Sample from an arbitrary distribution!

Assume that chain `i` is time-homogeneous. The aim of this part is to use the Metropolis-Hasting algorithm to change this chain in a way to have the arbitrary distribution π_a as its limiting distribution.

Given X_t as the state of the chain at time t, the next state X_{t+1} can be sampled as `[`X_t`;` X_{t+1}`] = chain_i(1, 1,` X_t`)`. With the Metropolis-Hasting algorithm, you can find the probability a_{t+1} with which you should accept the new sample X_{t+1} (otherwise stay at the same state as X_t) to finally have a limiting distribution π_a.

For each time-homogeneous Markov Chain among the 4 chains (i.e. `chain_1.p`, `chain_2.p`, `chain_3.p`, and `chain_4.p`),

a) Write a function as `X = MP_chain_i(N_chain, Time, pi_a, x0)` which produces `N_chain` $\in \mathbb{N}$ realisations of the modified version (in a way to have the limiting distribution `pi_a`) of chain `i` with length `Time` $\in \mathbb{N}$ and initial state $X_0 =$ `x0`. Note that you are not allowed to directly sample from `pi_a`. The function should be saved as `MP_chain_i.m`.

b) For each of your chains, and for each of the cases
1. `pi_a = [16,8,4,2,1]/31;`
2. `pi_a = [1,1,4,1,1]/8;`
3. `pi_a = [4,2,1,2,4]/13;`
do the following steps:

b.1) Evaluate your code, i.e. show that your modified chain has a limiting distribution equal to `pi_a` for all choices of the initial state `x0`.

b.2) Plot the total variation distance over time, and analyze the effect of the initial state `x0` on the convergence rate of the algorithm. Does it depend on the desired distribution `pi_a`? If yes, explain how.

b.3) For each distribution `pi_a`, can you estimate (numerically) an upper-bound for T_ϵ when $\epsilon = 0.005$?

c) For each of the three choices of `pi_a` in part **b)**, which chain has a better convergence rate? Can you explain your observations intuitively?

What should you submit?

1. Report: It should be at most 4 pages. It should include the results of your analyses, answers to the questions, interesting observations, and employed approaches. You do not need to put all of the figures we want in the main report, you can put them in Supplementary Materials (see below) and refer to them in the main report. The main part of your grade is based on your report.

2. `.mat` Files: `P_hat_chain_i.mat` and `pi_hat_chain_i.mat` files.

3. `.m` Files for Functions: `MP_chain_i.m` files. *Codes should be well-written and commented with details.*

4. `.m` Files for Analysis: `Analysis_chain_i.m` files for identifying chain `i`. `Analysis_MP_chain_i.m` files for evaluation of your sampling algorithm for chain `i`. *Codes should be well-written and commented with details.*

5. Supplementary Materials: It should be at most 4 pages. It should include extra figures, the details of any non-trivial approaches that you used, etc. It is obvious that there is no need to explain the details of well-known materials of the course, e.g., the Metropolis-Hasting algorithm.

When should you submit?

Deadline: Sunday, December 15, at 23:55.