Markov Chains and Algorithmic Applications: Mini-Project

Hugo Aguettaz, Louis Amaudruz, Adrien Vandenbroucque

email: {hugo.aguettaz,louis.amaudruz,adrien.vandenbroucque}@epfl.ch

I. Introduction

In this project, we have the ability to produce samples coming from four unknown Markov chains. The first goal is to find various properties of these chains (time-homogeneity, existence of a limiting distribution, ...). Knowing those properties, we use some of the chains as *base chains* in the Metropolis-Hastings algorithm in order to sample from given distributions and analyze the convergence of the algorithm.

We first describe the methods used to compute the desired properties and design the proper Metropolis-Hastings chains. Then, we present the results obtained together with an analysis. Finally, a summary is given in the conclusion together with the general trends observed.

II. Methods

A. Section 1

1) Transition probability matrix and time-homogeneity: In order to approximate the transition matrices $P^{(t)}$ of a chain, we can only count the number of transition from one state to another in the generated sample. The idea is keep such counters for every possible transition in a matrix. After that, we just need to normalize every row of the count matrix.

In fact, in the Matlab implementation, we can increase the parameter `NChain`, giving us the ability to estimate $P^{(t)}$ for some t. For that, we count the transition from time t to $t+1$ as explained before. By increasing the parameter `Time`, we can specify for how many steps t we will compute $P^{(t)}$. In terms of data structure, we end up with a 3-dimensional array containing transition matrices for each time t.

Note that we also want to have an error (on average) less than 10^{-3} for every element of $P^{(t)}$ (for chains that are time-homogeneous) by choosing appropriate `Time` and `NChain`. In fact this has to do with choosing `Time` × `NChain`, which is related to the number of counts we will get. Moreover, since we use such a counting method, we are getting more and more precise, according to the Central Limit Theorem. In our case, the standard deviation of the sample means of the computed transitions are given by $\frac{\sigma_{ij}}{\sqrt{N_i}}$, the standard deviation of transition i to j, divided by the square root of the number of times we are at state i. Since we make a transition from i to j with an estimated probability $\hat{p}_{ij}$ (Bernoulli random variable), its standard deviation is $\sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})}$. So after setting `Time` and `NChain`, we can record the N_i and σ_{ij} as we estimate the transition probabilities, and finally compute the average error $\frac{1}{N}\sum_{i,j}\sqrt{\frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{N_i}}$, where N is the total number of elements in the matrix to check if it is smaller than 10^{-3}.

2) Limiting distributions: In order to compute the evolution of a uniform initial distribution, we successively multiply it by the $P^{(t)}$'s found from previous part. Note that since the chains can be time-inhomogeneous, we write the evolution from time 0 to t as

$$\pi^{(t)} = \pi^{(0)} \prod_{n=1}^{t} P^{(n)}. \tag{1}$$

We found that to get an error (on average) less than 10^{-3} for every element of the limiting π (when it exists), we simply need our approximations of $P^{(t)}$ to also have small errors, but this was ensured by the choice of `Time` and `NChain` we made in previous part.

In fact, if we want to compute the limiting distribution as was expected by the project, we would need to again sample from the chains, and compute for a large enough t what the distribution of the states is by counting. Using this method, named `expected_limiting` in our implementation, in order to have a small (average) error, we again need to choose `Time` and `NChain` following the same argument as for computing the transition matrices. Here, we want to choose `Time` so as to be close to the limiting, and `NChain` such that our frequency count is correct. So, for a high enough `Time, call it` t, we want to take `NChain` such that $\frac{1}{N}\sum_i \sqrt{\frac{\hat{\pi}_i^{(t)}(1-\hat{\pi}_i^{(t)})}{\texttt{NChain}}} < 10^{-3}$, where $\hat{\pi}_i^{(t)}$ is the estimate of $\pi_i^{(t)}$, and N is the number of states.

3) Total Variation distances and Eigendecompositions: For chains that have a limiting distribution, recall that from previous part, we have an easy access to their evolution for t such that $0 \leq t \leq t_{max}$, where t_{max} is chosen such that we are close to the limiting distribution. We can then compute the Total Variation distance between $\pi^{(t)}$ and $\pi^{(t_{max})}$, for $0 \leq t \leq t_{max}$ using the well-known formula seen in class.

We can of course repeat the process of computing the evolution of distributions and TV distances for the various asked initial distributions. In our code, this is done by a function called `compute_limiting_dist` that accepts multiple initial distributions and compute their evolutions at the same time using the corresponding $P^{(t)}$'s. Having these, we can compute the evolution of Total Variation distances for each of these distributions. From those plots, one can easily compute T_ϵ numerically.

To compute the eigenvalue decomposition of the transition matrices of time-homogeneous chains, we use a Matlab function called `eig`, which computes the eigenvalues together with the corresponding left and right-eigenvectors. To get the stationary distribution, we retrieve the eigenvector corresponding to eigenvalue 1 and normalize it. In fact, in our function `eig_decomposition`, we return the stationary distribution from the eigendecomposition and also compute the spectral gap $\gamma = 1 - \lambda_*$. From there, we can find the theoretical value of T_ϵ and the upperbound on the Total Variation distance, as we did many times in class.

B. Section 2

Here, we use the Metropolis-Hastings algorithm to sample from three arbitrary distributions (we will refer to them later with their respective numbers) :

1) $\pi_a = [16, 8, 4, 2, 1]/31$
2) $\pi_a = [1, 1, 4, 1, 1]/8$
3) $\pi_a = [4, 2, 1, 2, 4]/13$

The sampling is performed using only the time-homogeneous Markov chains of the first part as base chains.

First we compute the transition matrix of the base chain with transition probabilities ψ_{ij} as explained in section 1 of the project. Using our desired distribution and our base chain, we compute acceptance probabilities to make a move in a direction according to the formula :

$$a_{ij} = \min\left(1, \frac{\pi_j \psi_{ji}}{\pi_i \psi_{ij}}\right)$$

Finally, we compute the transition matrix P as :

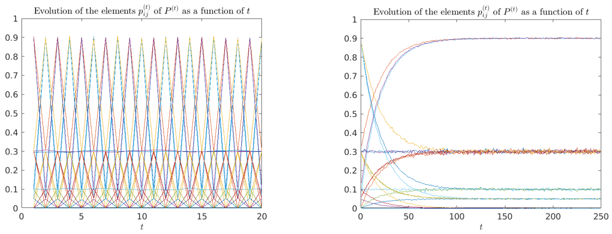
$$p_{ij} = \begin{cases} \psi_{ij} a_{ij} & \text{if } i \neq j \\ 1 - \sum_{i \neq k} \psi_{ik} a_{ik} & \text{otherwise} \end{cases}$$

In order to generate samples, we simulate the chain by using the `simulate` function from the `dtmc` object framework, a package to analyze Markov chains in *Matlab*. To analyze the convergence rates, we reuse ideas from section 1. That is, we first compute the evolution of an initial distribution until it converges to some limiting distribution (which should be π_a at some time t_{max} large enough). From these, we compute the Total Variation distances between $\pi^{(t)}$ and $\pi^{(t_{max})}$, for $0 \leq t \leq t_{max}$. One can easily determine numerically the mixing time T_ϵ for $\epsilon = 0.005$. Finally, we additionally include the theoretical upperbound on the mixing time, and for that we use the eigenvalue decomposition of the Metropolis-Hastings chain using the same techniques as in section 1, and compute the spectral gap. From there, we determine the theoretical value of T_ϵ for the desired ϵ.

III. Results

A. Section 1

1) Transition probability matrix and time-homogeneity: Figure 1 depicts the evolution of the elements of the transition matrices for chains 3 and 4, which both turn out to be time-inhomogeneous.



(a) Chain 3

(b) Chain 4

Figure 1: Transition probability matrix elements for each chain over time

In fact, chains 1 and 2 have a transition matrix that doesn't seem change over time, indicating that they are time-homogeneous. In order to compute the transition matrices with small error, we found that in practice choosing `NChain` $\times$ `Time` $\approx 500'000$ gives an average error just below 10^{-3}.

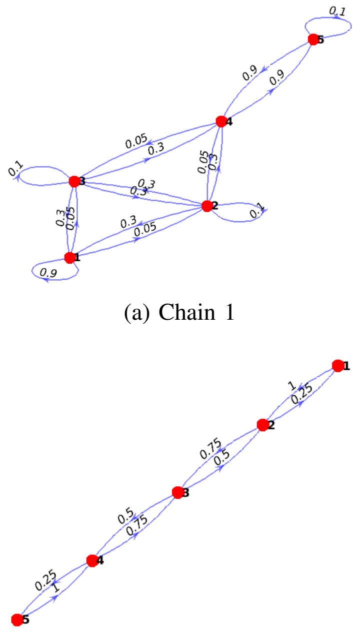
As mentioned earlier, the other two chains clearly have their transition matrices depend on time. For chain 3, we can see that the transition probability matrices alternate between two distinct transition kernels. Chain 4 is also very interesting as it converges to a fixed transition matrix. That is, this Markov process converges to a time-homogeneous Markov process. In practice, we notice that it takes around 100 steps to converge to this regime. As this chain behaves like a time-homogeneous one in the long run, we decide to analyze it as such and will sometimes present results about the transition matrix to which it converges.

We display the underlying graph of the time-homogeneous chains (1 and 2) in figure 2.

From these graphs, the chain 1 is irreducible and aperiodic, so it is ergodic since the number of states is finite. This suggests that there is a limiting and stationary distribution, which is actually the case as explained later. Chain 2 is also irreducible, but has a period of 2 so a limiting distribution does not exist. However, since it is positive-recurrent, there should exist a stationary distribution.

The graph corresponding to the transition matrix that chain 4 converges to is in the supplementary material, and we can see that the chain is also ergodic. As such, there should exist a limiting distribution since when arriving at the point where the chain becomes time-homogeneous and ergodic, it will converge to the limiting distribution.

2) Limiting distributions: When computing the limiting distributions using the counting method (the one expected),



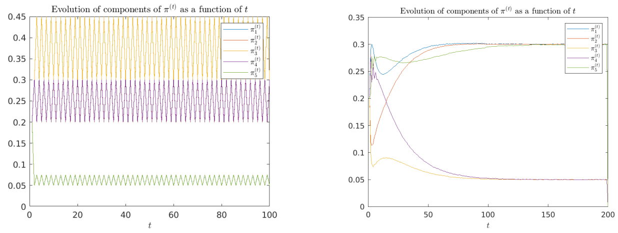
(a) Chain 1

(b) Chain 2

Figure 2: Transition graphs

we found that taking `Time` ≈ 50 and `NChain` $\approx 500'000$ gives an average error of less than 10^{-3} for each element of the limiting distributions.

The following results are those using our method of computing $\pi^{(t)}$, that is multiplying by the $P^{(t)}$'s. Figure 3 shows the evolution of the distribution starting from the uniform one as t increases for chains 2 and 4. We observe that (over the time range we chose), chains 1 and 4 seem to converge to some limiting distribution, while chains 2 and 3 show no convergence at all.



(a) Chain 2

(b) Chain 4

Figure 3: Evolution of probability distributions for chains 2 and 4 starting from a uniform distribution

This confirms our hypotheses that a limiting distributions exists for chain 1 and 4 but not for chain 2.

3) Total Variation distances and Eigendecompositions: For the Total Variation plots, we consider here results for chains 1 and 4 as they were the only ones having a limiting distribution. We show here the evolution of the TV distance starting from all states together with the theoretical upperbound.

The first comment we can make is that once again we have a confirmation that the limiting distributions we find are indeed limiting. That is, for the various initial $\pi^{(0)}$, we always converge to the same distribution. We also notice the exponentially decaying curve we saw in class for the TV distance. When looking at chain 1, one can notice that starting from states 2 or 3, convergence is faster than starting from 4 or 5, which is itself faster than starting from state 1.

The second comment we can make is that in practice we see that the evolution of the TV distance is indeed bounded by the theoretical upperbound derived in class. This can be seen in Figure 4 (for chain 1) but also in table I. For chain 1, the practical mixing time is smaller than the theoretical as expected. For chain 4, we can explain the fact that the theoretical mixing time is smaller because it is computed from the transition matrix the process converges to. However it does not take into account the transient steps before converging to the time-homogeneous process which is why it takes a few more steps before reaching mixing time in practice.

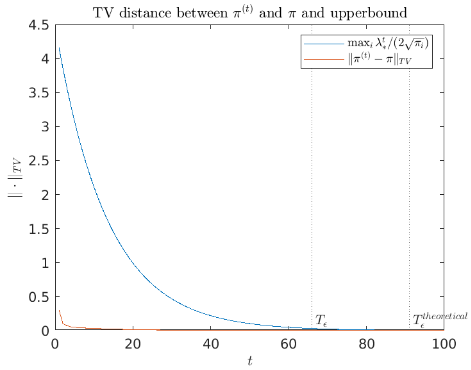


Figure 4: Total variation distance with upper bound between the limiting distribution and $\pi^{(t)}$ with $\pi^{(0)}$ uniform for chain 1

Chain	T_ϵ (practical)	T_ϵ (theoretical)
1	66	92
4	99	90

Table I: Mixing times T_ϵ of chains 1 and 4 for $\epsilon = 0.005$

Finally, the stationary distributions are shown in the various bar charts in the supplementary material. What we see is that as claimed before, chain 1 and chain 2 both have a stationary distribution. Once again here, we computed additionally what the stationary of chain 4 would be if it started as time-homogeneous directly.

B. Section 2

As explained, in this part, chain 1 and 2 will be used as base chains as they are time-homogeneous.

1) Limiting distributions: With both chain 1 and 2 as base chains, we are able to sample from all desired distributions π_a. Figure 5 shows the limiting distributions obtained when using chain 1 as base chain and distribution $\pi_a = [16, 8, 4, 2, 1]/31$ for computing acceptance probabilities, starting from all possible states. The plots presented in the supplementary material show the same trend for other choices of π_a, which is that we are always able to converge to the desired limiting distribution.

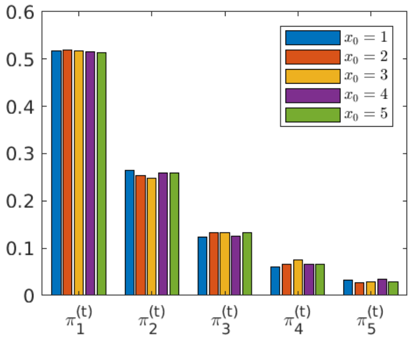
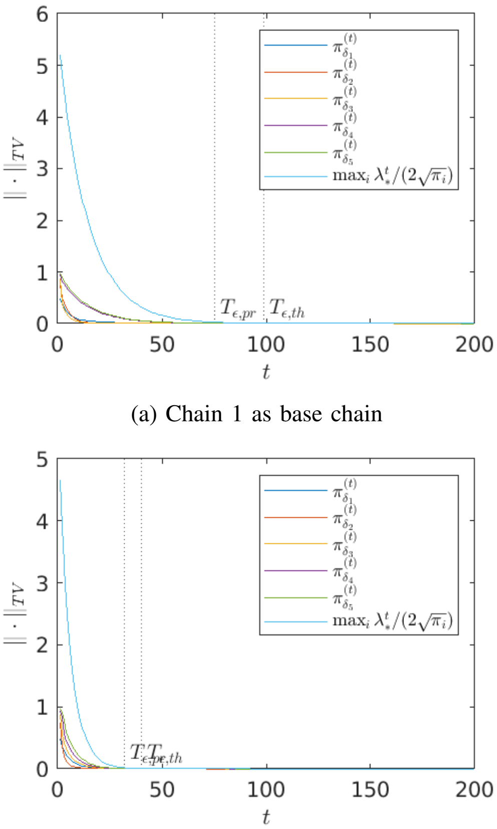


Figure 5: Limiting distribution for every initial state $\delta^{(0)}$ with chain 1 as base chain and $\pi_a = [16, 8, 4, 2, 1]/31$

2) Total Variation distances: For convergence rates, observe in figure 6 how the initial distribution affects the convergence for both chains. In these plots, we also include the theoretical upper bound, computed from the eigendecomposition of the transition matrix built from the Metropolis-Hastings procedure. Note that chain 2 seems to give a faster convergence to the distribution π_a compared to chain 1, which is discussed now.

It is interesting to compare what are the differences between choosing chain 1 or 2 as base chain in order to sample from π_a. First of all, it appears the second chain has faster convergence in all cases, which can be deduced when looking at the mixing times in table II. This can be explained by computing the theoretical convergence rates as we explained above, which basically consists in finding the spectral gaps. Our results show that the Metropolis chain induced by chain 2 has a larger spectral gap, which means a faster convergence rate. It also appears that the mixing time is dependent on the distribution we want to sample from, as the third distribution is always the slowest to reach. Once again, computing the spectral gaps explains this behavior.



(a) Chain 1 as base chain

(b) Chain 2 as base chain

Figure 6: TV distances with upper bound with $\pi_a = [16, 8, 4, 2, 1]/31$ and experimental and theoretical mixing times for $\epsilon = 0.005$

	Chain 1		Chain 2	
π_a	T_ϵ	γ	T_ϵ	γ
$[16, 8, 4, 2, 1]/31$	49 / 68	0.0695	25 / 28	0.1643
$[1, 1, 4, 1, 1]/8$	72 / 73	0.0601	27 / 27	0.1766
$[4, 2, 1, 2, 4]/13$	95 / 95	0.0481	84 / 84	0.0544

Table II: Mixing times T_ϵ (practical/theoretical) and spectral gap with chains 1 and 2 as base chains for $\epsilon = 0.005$

IV. Conclusion

In this project, we used much of the concepts seen in class. By counting, the frequencies of transitions, we were able to discover basic properties such as the time-homogeneity, or the existence of limiting and stationary distributions. Moreover, using some of these chains, we were able to sample from arbitrary distributions by implementing the Metropolis-Hastings algorithm.

On the practical side, we identified the challenge of choosing the right parameters `NChain` and `Time` in order to get good estimates of transition matrices and limiting distributions. We were able to plot and easily see the convergence speeds to limiting distributions, and check that the theoretical upper bound on the mixing time holds.

Using a different approach for computing the evolution of distributions gave us very good convergence results and insights for both sections of the project.