Tutorial on Sampling (MCMC and Simulated Annealing)*

(SLT 2021)

March 8, 2021

These are **informal** notes whose purpose is to help you to follow the tutorial class (March 15) and for some parts of the first coding exercise.

Basic Properties (Markov chains)

Just a few examples of very simple Markov chains:
A Markov chain is just a matrix P where $P(c, c')$ is the probability of moving from c to c' i **one step**.[1] Note that these two properties are always true:

1. Each row sums up to 1,

$$\sum_{c'} P(c, c') = 1 \tag{1}$$

   because this is just the probability that from c we are in one step in some state.

2. If we take **powers** of this matrix, $P^2, P^3, \ldots, P^t$, we get the probability that in t **steps** we go from some c to some c'.
   (**Exercise:** Convince yourself that this is true using (1); Try $t = 2$ to get the idea.)

What does a Markov chains? We start "somewhere" and "move" to the next state according to P, which corresponds to a sequence of random variables:

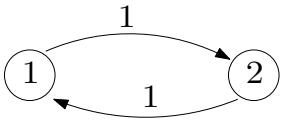
$$X_0, X_1, \ldots, X_t, \ldots$$

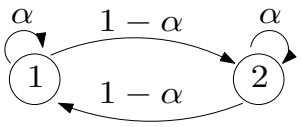
which satisfy the condition that the "next state depends only the current state":

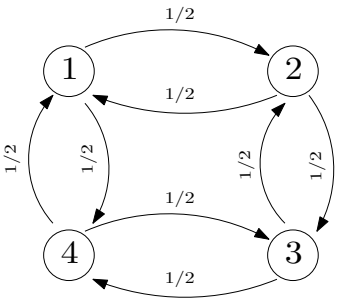
$$\Pr(X_{t+1} = c' \mid X_t = c,\ X_{t-1} = x_{t-1} \ldots, X_0 = x_0) = \Pr(X_{t+1} = c' \mid X_t = c) = P(c, c')$$

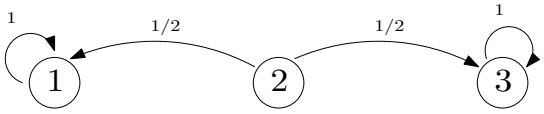
*These notes are based on earlier notes of this course. All mistakes are mine (Paolo Penna).

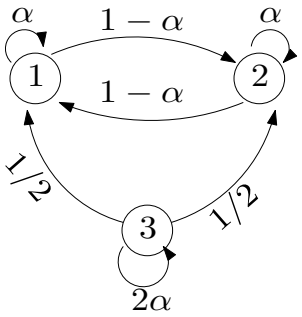
[1] Here we restrict to so called *time invariant* Markov chains over a *finite state space*.











What happens after **sufficiently many steps**? Does the **starting state** matter?

A remarkable property of "well behaving" Markov chains is that, after sufficiently many steps, we "**converge**" to a unique **invariant or stationary distribution** π:

$$P^t(c, c') \to \pi(c')$$

The next two conditions capture "well behaving" Markov chains.

Irreducible (informal): Can go from any state to any state (finite number of steps).
Aperiodic (informal): Chain does not go "back and forth" forever. Enough to set

$$P(c, c) > 1/2 \tag{2}$$

The reason we want these two properties is to guarantee that our chain **converges** always to the **same distribution** regardless of our initial starting state.

Invariant or Stationary Distribution:

$$\pi P = \pi \tag{3}$$

Some intuition first. Suppose that in some chain with two states, after 10000 steps the probability of being in state 1 is roughly 1/3 while that of being in state 2 is roughly 2/3, regardless of the initial state c,

$$P^t(c,1) \approx \pi(1) = 1/3 \qquad\qquad P^t(c,2) \approx \pi(2) = 2/3\ . \tag{4}$$

Then we expect that the same should hold for 10001 steps (for t is very large, not much should change if we do t steps or $t+1$ steps):

$$P^{t+1}(c,1) \approx \pi(1) = 1/3 \qquad\qquad P^{t+1}(c,2) \approx \pi(2) = 2/3\ . \tag{5}$$

But these probabilities P^{t+1} can be computed from P^t:

$$\underbrace{P^{t+1}(c,1)}_{\pi(1)} = \sum_{c''} P^t(c,c'')P(c'',1) \overset{(4)}{=} \underbrace{\sum_{c''} \pi(c'')P(c'',1)}_{\pi P(1)} \tag{6}$$

which is the condition in (3).

Furthermore, check that some of the Markov chains above do not have a unique stationary distribution (look at the last two chains).

$\underbrace{\text{Irreducible + Aperiodic}}_{\text{"well behaving" MC}}$ $\Rightarrow$ Converge to $\underbrace{\text{Unique Invariant Distribution}}_{\text{our target distribution}}$

This means that for $t \to \infty$ and for any c and c'

$$P^t(c,c') \to \pi(c')\ ,$$

so the probability of being in c' does not depend on the initial state c. If we **wait "long enough"** we can use a Markov chain to **sample** according to distribution π. This depends on the particular chain and it is usually denoted as the **mixing time** of the chain. This is the minimum $t_{mix}(\epsilon)$ such that for any $t \geq t_{mix}(\epsilon)$ we are close (by some factor ϵ) to the invariant distribution: for any initial states c

$$||P^t(c,\cdot) - \pi(\cdot)||_{TV} \leq \epsilon$$

where $||p-q||_{TV} = \sum_c |p(c)-q(c)|$ is the total variation distance.[2] Note that, if we want to sample **two** times with distribution π,

$$c_1, c_2, \sim \pi$$

we will have to wait t_{mix} for the first sample and then **another** t_{mix} for the second sample. (Convince yourself by looking at the two-states chain above, when the self-loop probability α is almost 1.)

[2] In this tutorial we do not consider techniques to prove bounds on the mixing time. Later we give some intuition of why some of our chains may be slowly mixing.

Detailed Balance (Our Tool)

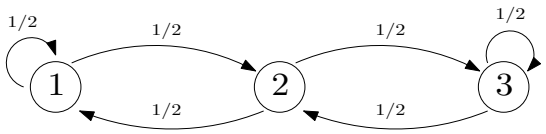
Detailed Balance: There is a π that, for any c and c', satisfies

$$\pi(c)P(c,c') = \pi(c')P(c',c) \tag{7}$$

Theorem 1. *If π satisfies detailed balance (7) then π is indeed the stationary distribution of P.*

Proof. **Exercise!** □

Example 2. *Consider this chain*



*The middle state is visited more often (every time we go from 1 to 3 we pass by 2). This means that the stationary distribution assigns higher probability to state 2 than the other two states (**true or false?**)*

Example 3 (Card Shuffling)**.** *We have a deck of n cards and we want to shuffle them (generate a **random** sequence). For this, we may use one of the following two methods:*

- ***M1*** *Pick a card in the deck uniformly at random and put it on top of the deck.*
- ***M2*** *Pick two cards at random and swap them.*

Try to answer the following two questions:

1. *Which of these methods yield a MC satisfying detailed balance?*
2. *Which of these methods generate a random shuffling (uniform distribution)?*

How to Use It (Metropolis-Hasting)

We want to **design** a chain having stationary distribution equal to some target distribution (typically a Gibbs distribution), so that we can use this MC to sample.

Example 4 (toy example)**.** *We have three possible solutions c_1, c_2, c_3 and some cost function $R(\cdot)$, with*

$$R(c_1) = 2 \qquad R((c_2) = 1 \qquad R(c_3) = 10 \tag{8}$$

We want to sample according to the Gibbs distribution

$$p_\beta(c_1) = \frac{e^{-2\beta}}{Z} \qquad p_\beta(c_2) = \frac{e^{-\beta}}{Z} \qquad p_\beta(c_3) = \frac{e^{-2\beta}}{Z} \tag{9}$$

where $Z = Z(\beta) = e^{-2\beta} + e^{-\beta} + e^{-100\beta}$ *. Our MC has transitions between any pair of these three states. Since we want stationary distribution* $\pi = p_\beta$ *we apply detailed balance* (7)*. For example, for the pair* (c_1, c_2) *this condition is*

$$p_\beta(c_1)P(c_1, c_2) = p_\beta(c_2)P(c_2, c_1)$$

that is

$$P(c_1, c_2) = \frac{p_\beta(c_2)}{p_\beta(c_1)}P(c_2, c_1) = \frac{e^{-\beta}/Z}{e^{-2\beta}/Z}P(c_2, c_1) = \frac{e^{-\beta}}{e^{-2\beta}}P(c_2, c_1)$$

and the main advantage is that ***we do need*** Z *(in general hard to compute).*

General Recipe (Metropolis-Hasting):

1. Define the "structure" of the MC (**proposal distribution**) $\Rightarrow$ Irreducible;
2. Set the transition probabilities (**acceptance probability)** $\Rightarrow$ Detailed Balance.
3. Make sure self loops have positive probability (**lazy chain**) $\Rightarrow$ Aperiodic.

We discuss this method with another simple example. We have $n = 3$ points and we want to construct $k = 2$ clusters. Each cluster is some c with

$$c = (c_1, c_2, c_3) \qquad c_i \in \{1, 2\}.$$

For some cost function $R(c|X)$ and we want to sample according to the Gibbs distribution

$$p_\beta(c) = \frac{e^{-\beta R(c|X)}}{Z(\beta)} \qquad Z(\beta) = \sum_{c \in C} e^{-\beta R(c|X)} \tag{10}$$

Our goal: a MC with stationary distribution $\pi = p_\beta$ (Gibbs distribution)

1. Proposal distribution ('change one coordinate'). Given the current state c, pick a random index i and change this coordinate choosing c_i' uniformly at random in $\{1, 2\}$. For any two states c and c' that differ in at most one coordinate

$$q(c, c') = 1/n$$

is the proposal distribution (where in our example $n = 3$).

2. Acceptance probability (satisfy detailed balance). Accept the previous "move" with some probability $A(c, c')$ to be defined. This means that our MC has transition probabilities

$$P(c, c') = q(c, c')A(c, c') \tag{11}$$

for all solutions that differ in at most one coordinate. We want to satisfy detailed balance for $\pi = p_\beta$, that is,

$$p_\beta(c)P(c, c') = p_\beta(c')P(c', c) \tag{12}$$

We do not need to write/compute $Z(\beta)$ anymore

$$\frac{e^{-\beta R(c)}}{Z(\beta)}P(c, c') = \frac{e^{-\beta R(c')}}{Z(\beta)}P(c', c) \Leftrightarrow \tag{13}$$

$$e^{-\beta R(c)}P(c, c') = e^{-\beta R(c')}P(c', c) \tag{14}$$

and from this we just use (11)

$$e^{-\beta R(c)}A(c, c') = e^{-\beta R(c')}A(c', c) \Leftrightarrow \tag{15}$$

$$A(c, c') = e^{-\beta[R(c')-R(c)]}A(c', c) \tag{16}$$

Since $A(\cdot)$ must be a probability distribution, we set

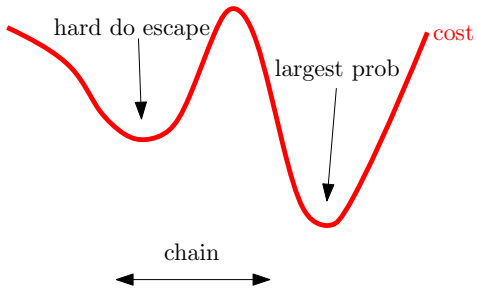
$$A(c, c') = \min\{1, e^{-\beta[R(c')-R(c)]}\}$$

3. Self loops (lazy chain work). To make the chain aperiodic, it is enough that $P(c, c) > 0$ for all c. A standard trick is to make your chain **lazy** meaning that, at every step, with probability $1/2$ we do nothing, and with probability $1/2$ we actually do one step of the (original) chain. The modified chain is only a factor 2 slower and has the same stationary distribution of the original one (this trick often used to prove bounds on the convergence time to the stationary distribution which may be affected by "too small' self loop probabilities).

Time to Converge?

How many steps do we need to make?

For some β our chain may need **many steps** before approaching to the desired stationary distribution $\pi = p_\beta$. Pictorially:



A numerical example is a one dimensional chain which moves left or right over four states with cost

$$3, 1, 10, 0, 3$$

As $\beta \to \infty$ all probability p_β concentrates on the state of cost 0. From the state of cost 1, the chain must first "accept" to move to a worst state of cost 10,

$$P(2,3) = \frac{1}{2} e^{-\beta(10-1)}$$

where the '1/2' comes from a 'left or right' proposal distribution.

Large $\beta \Rightarrow$ Too long to converge

Simulated Annealing

This observation suggested a powerful method called **simulated annealing** where we gradually adjust β (or the temperature $T = 1/\beta$). We gradually reduce the temperature and, for the current temperature, we run our MC for a certain number of steps. The scheme is fully specified by parameters

$$(\beta_0, n_0), (\beta_1, n_1), \ldots, (\beta_L, n_L)$$

which corresponds to run the MC (Metropolis-Hasting) for the Gibbs distribution p_{β_i} for $t = n_i$ steps, and then continuing with the MC for the Gibbs $p_{\beta_{i+1}}$ (this chain starts from the last state visited by the previous chain). This schemes yields a so-called *time dependent* Markov chain.