# How to Quantitate a Markov Chain?

Stochostic project 1

Chin-Ning,Chou Wei-chang,Lee

*PROFESSOR RAOUL NORMAND*

April 17, 2015

**Abstract**

In this project, we want to quantitatively evaluate a Markov chain. In the beginning, we give a simple introduction to Information Theory ,and then focus on the inequality of mutual information of a Markov process. In the main part of this project, we try to prove the second law of thermodynamics in Markov chains, and derived some quantities to measure the inter-state behaviour of a stationary Markov chain. We will use Matlab to simulate Ehrenfest model to bridge realistic world and our quantitative approach.Finally, we will comment on the second law of thermodynamics under other entropy definition to.

*Keywords:* Entropy , Markov chain , Kullback-Leibler divergence , Ehrenfest model , Thermodynamics

## Introduction

In classical thermodynamics, people like to use entropy to evaluate the structural tendency in a system. While the well known Second Law of Thermodynamics claims that in a natural thermodynamic process, the entropy in the participated system will increase, we want to find out whether there also exists some properties in a Markov chain that quantitatively reveal the structural tendency in the system.

In the beginning, we applied the definition of entropy that Claude Shannon used in Information theory[**?**] to construct the entropy in a Markov chain. However, after deducing some corresponding results , we found that such definition does not elegantly exposure the tendency in the system. As a result, we surveyed and tried other models to describe the structure of a

Markov chain, rethinking how to interpret Shannon's entropy in a Markov chain.

In this report, we will first apply Shannon's entropy to examine the corresponding results in a Markov chain. Next, some other structural properties will be introduced. In the third part of the report, we will use a simple urn problem which is known as Ehrenfest model to demonstrate the properties we prove in this project. And we will give some comments on each structural properties and discuss the difference of Shannon's entropy and the entropy in thermodynamic at last.

## Basic Knowledge in Information Theory

In this section, we briefly introduce Shannon entropy, mutual information, Kullback-liebler divergence and some useful results such as data processing inequality.

**Shannon Entropy** The entropy of a random variable is a measure of the uncertainty of this random variable. It is a measure of the amount of information required on the average to describe the random variable.

The entropy H(x) of a discrete random variable X is defined by:

$$\begin{aligned} H(X) &= -\sum_{x \in X} p(x) \log p(x) \\ &= E_p[\frac{1}{\log p(x)}] \end{aligned}$$

And the conditional entropy is given by

$$\begin{aligned} H(Y|X) &= \sum_{x \in X} p(x) H(Y|X=x) \\ &= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x) \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \end{aligned}$$

For a pair of discrete random variable X and Y:

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x)p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}
$$

**Theorem.** (Uniform Distribution Maximize Shannon Entropy)
*Given a finite state Markov chain, it's Shannon entropy is maximized if and only if current distribution is uniform.*

The theorem gives us an intuition that the entropy of a Markov chain is maximized when the probability of every microscopic state to show up is the same! That is, if we start the stochastic process with uniform distribution, the entropy is maximized in the very beginning. So, it's clearly that the *Second Law of Thermodynamics* does not hold under Shannon entropy!

**Kullback-liebler divergence** We can use the definition above to derive relative entropy, which is a measure of distance between two distributions. The relative entropy or Kullback-liebler divergence between two pmf p(x) and q(x) is defined as:

$$
\begin{aligned}
D(p||q) &= -\sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \\
&= E_p[\log\frac{1}{q(x)}] - E_p[\log\frac{1}{p(x)}]
\end{aligned}
$$

The Kullback-Leibler divergence measures the expected number of extra bits required to code samples from P when using Q. It is non-negative and asymmetric :

$$
D(p||q) \neq D(q||p) \geq 0
$$

**Theorem.** (Kullback-liebler Divergence)
*For a Markov chain with unique stationary distribution, the Kullback-liebler divergence between the distribution of each step and the stationary distribution will monotonically decrease.*

To prove this theorem, we first need the following lemma.

**Lemma** If $u_n$ and $u'_n$ are two distribution on the same Markov chain at step n. Then, $D(u_n||u'_n)$ is monotonically decreasing.

*Proof.* Denote the corresponding distribution of $\{u_n\}$ and $\{u'_n\}$ by $p$ and $q$. This can be easily shown by observing,

$$\begin{aligned}
&D(p(x_n, x_{n+1})||q(x_n, x_{n+1})) \\
&\quad = D(p(x_n)||q(x_n)) + D(p(x_n|x_{n+1})||q(x_n|x_{n+1})) \\
&\quad = D(p(x_{n+1})||q(x_{n+1})) + D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n))
\end{aligned}$$

Since $p(x_{n+1}|x_n) = q(x_{n+1}|x_n)$, it's straight forward that

$$D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) = 0$$

Thus,

$$D(p(x_n||q(x_n)) \geq D(p(x_{n+1})||q(x_{n+1}))$$

□

Finally, to prove the monotonically decreasing of Kullback-liebler divergence between current distribution and stationary distribution, just simply set $q = \pi$ then the result is shown.

**Mutual Information** The mutual information I(X;Y) is the relative entropy between the joint distribution and the product distribution p(x).p(y):

$$\begin{aligned}
I(X;Y) = &= -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= -\sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= H(X) - H(X|Y) \geq 0
\end{aligned}$$

The mutual information I(X;Y) is the reduction in the uncertainty of X due to the knowledge of Y. It is easily to derive following property:

$$\begin{aligned}
I(X;Y) &= I(Y;X) \\
I(X;X) &= H(X;X) \\
H(X) &\geq H(X|Y)
\end{aligned}$$

We can use a Venn diagram to summarize the relationship between the entropy concept we discuss above:

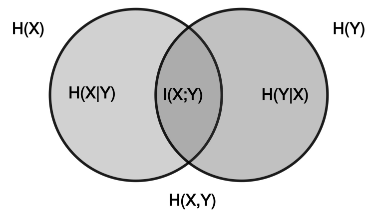


Figure 1: Venn's diagram for mutual information

## Data processing Inequlity

Under simple Markov assumption, We can show that no clever manipulation of the data can improve the inferences that can be made from the original data.

$$If\ X \rightarrow Y \rightarrow Z\ forms\ a\ Markov\ chain,\ then\ I(X;Y) \geq I(X;Z)$$

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) && (1) \\ &= I(X;Y) + I(X;Z|Y) && (2) \end{aligned}$$

First focus on $I(X;Z|Y)$, we notice that:

$$\begin{aligned} I(X;Z|Y) &= H(X|Y) - H(X|Z,Y) \\ &= E_P(x,y,z)\log\frac{p(x,z|y)}{p(x|y).p(z|y)} \\ &= E_P(x,y,z)\log\frac{\frac{p(x,y,z)}{p(y)}}{p(x|y).p(z|y)} \\ &= E_P(x,y,z)\log\frac{\frac{p(x,y)p(z|y)}{p(y)}}{p(x|y).p(z|y)} \\ &= E_p(x,y,x)\log 1 && = 0 \end{aligned}$$

From (1) and (2) we can easily get $I(X;Y) \geq I(X;Z)$ , Z can be written as a function of Y, thus we cannot increase the mutual information about X and Y by manipulating Y. Note that $I(X;Y|Z) \leq I(X;Y)$ can be derived from (1) and (2), which means that the dependency of X and Y is decreased by observing Z. It is because transition matrix is invertible , $Z \rightarrow Y \rightarrow X$ also forms a Markov chain. Z reveals some information about X and Y.

The conclusion of no clever manipulation of the data can improve the inferences may lead to misconception that the shuffle do not increase randomness of the cards. We can prove that it is not true if shuffle method T and card positions X are independent :

$$\begin{aligned}
H(TX) &\geq H(TX|T) && \textit{condition reduce entropy} \\
&= H(T^{-1}TX|T) && \textit{we know } T^{-1} \textit{ conditioning } T \\
&= H(X|T) && \textit{independence} \\
&= H(X)
\end{aligned}$$

By an further application of data processing inequality to the Markov chain $X_1 \rightarrow X_{n-1} \rightarrow X_n$ we have:

$$\begin{aligned}
I(X_1; X_{n-1}) &\geq I(X_1; X_n) \\
H(X_{n-1}) - H(X_{n-1}|X_1) &\geq H(X_n) - H(X_n|X_1) && \textit{expand}
\end{aligned}$$

**Theorem.** (Conditional Entropy given Initial Distribution)
*If $\{X_n\}$ is a stationary Markov chain, then*

$$H(X_{n+1}|X_1) \geq H(X_n|X_1) \quad \forall n$$

*Moreover, the conditional entropy given initial distribution will converge to the Shannon entropy of the stationary distribution.*

The conditional entropy $H(X_n|X_1)$ increases with n. Thus the conditional uncertainty of the future increases. Note that conditional reduces entropy, $H(X_n|X_1)$ will be bounded by $H(X_n)$ and since $H(X_n|X_1)$ is nondecreasing with n. It will finally reach a limit, which is the entropy of stational distribution $H(X_\pi)$

## Entropy and the Second Law of Thermodynamics

We can model the evolution of a isolated system as a Markov chain with transition matrix obeying the physical laws governing the system. And now we know some basic properties of entropy , what if we try to figure out the statement of the Second Law of Thermodynamics that the entropy of an isolated system is always nondecreasing?

Let $\mu_n$ and $\mu_n^{'}$ be two probability distribution at time n on the same space of Markov chain but different initial distribution and $\mu_{n+1}$ , $\mu_{n+1}^{'}$be the ones at time n+1. The corresponding joint mass function of time n and n+1 can be denoted as p and q. We have:

$$p(x_n, x_{n+1}) = p(x_n)r(x_{n+1}|x_n)$$
$$q(x_n, x_{n+1}) = q(x_n)r(x_{n+1}|x_n)$$

where $r(.|.)$ is the transition probability of this Markov chain. By the chain rule for relative entropy, we can rewrite the relative entropy of p and q as :

$$D(p(x_n, x_{n+1})||q(x_n, x_{n+1})) = D(p(x_n)||q(x_n)) + D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) \tag{3}$$
$$= D(p(x_{n+1})||q(x_{n+1})) + D(p(x_n|x_{n+1})||q(x_n|x_{n+1})) \tag{4}$$

Also,

$$D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) = D(r(x_{n+1}|x_n)||r(x_{n+1}|x_n)) = 0 \tag{5}$$
$$D(p(x_n|x_{n+1})||q(x_n|x_{n+1}) \geq 0 \tag{6}$$

Substitute (3)(4) with (5)(6), We have:

$$D(p(x_n)||q(x_n)) = D(\mu_n||\mu_n^{'}) \geq D(p(x_{n+1})||q(x_{n+1}) = D(\mu_{n+1}||\mu_{n+1}^{'})$$

Distance casued by different initial distribution is decreasing with time n. If we let $\mu_n^{'}$ be any stationary distribution $\pi$, $\mu_n^{'}$ will be the same as $\mu_{n+1}^{'}$ .

$$D(\mu_n||\pi) \geq D(\mu_{n+1}||\pi)$$

Which implies that any state distribution gets closer and closer to each stationary distribution as time goes by. $D(\mu_n||\pi)$ is a monotonically non-increasing non-negative sequence and thus have a limit.

The fact that the relative entropy always decreases does not imply that the entropy increases. Markov chain with a non-uniform stationary distribution starts from uniform distribution will decrease since we get maximum entropy when the distribution is uniform.

If the stationary distribution is uniform, we have:

$$D(\mu_n||\pi) = \log|\chi| - H(\mu_n) = \log|\chi| - H(X_n)$$

The monotonic decrease in relative entropy implies a monotonic increase in entropy.

## Simulation on Ehrenfest model

In this section, we use Erenfest model to demonstrate the concepts and the results in the previous part.

**Model Setting** First, let's construct the model settings.

- There are two urns: A, B in our system.
- Each urn contains $n$ balls in the beginning. And the ball in urn A is black while the ball in urn B is white.
- In each step, we take one ball with uniform probability from each urn and make an exchange.

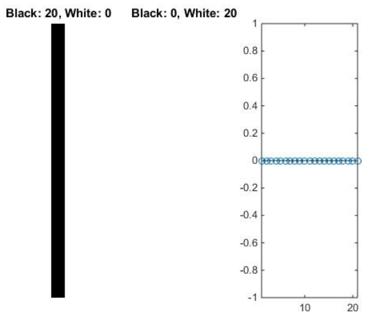


Figure 2: Initial setting

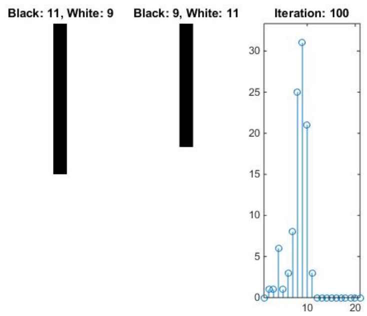


Figure 3: After 100 iterations

**Properties** With these initial settings, we can simply find out that this system is a Markov chain and easily derive the below properties of this process:

- The state space $S = \{S_i | i = 0, 1, ..., n\}$, where $S_i$ refers to the configuration that urn A contains $i$ black balls and $n - i$ white balls, while urn B contains $n - i$ black balls and $i$ white balls.
- The transition probability is

$$P(X_n = i, X_{n+1} = j) = \begin{cases} (\frac{i}{n})^2 & if \quad j = i - 1 \quad and \quad i \geq 1 \\ \frac{2i(n-i)}{n^2} & if \quad j = i \quad and \quad 1 \leq i \leq n-1 \\ (\frac{n-i}{n})^2 & if \quad j = i + 1 \quad and \quad i \leq n-1 \\ 1 & if \quad i = 0, j = 1 \quad or \quad i = n, j = n-1 \end{cases} \tag{7}$$

- The stationary distribution is

In the following simulation, we set the number of balls in each urn to 20 and observe the: Shannon entropy, Kullback-liebler divergence to the stationary distribution, conditioned entropy to initial distribution. Also, we consider two different initial distribution: uniform initial distribution and random initial distribution.

First, consider the uniform initial distribution. Note that, the Shannon entropy is maximized when the the distribution is uniform. The simulation results are in the following figure:

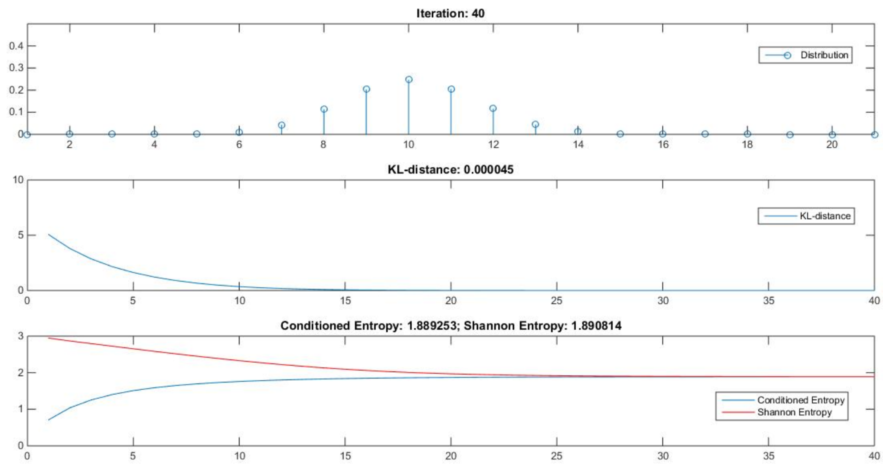


Figure 4: Start with uniform distribution

We can see the following results:

1. In the first graph, it's clearly that the distribution after 40 iterations is very close to the stationary distribution.
2. In the second graph, the Kullback-liebler divergence between current distribution and the stationary distribution is monotonically decreasing.
3. In the third graph, we can see that the Shannon entropy(red curve) is decreasing while the conditioned entropy on initial distribution is monotonically increasing.

Then we randomly choose the initial distribution and get the following results:

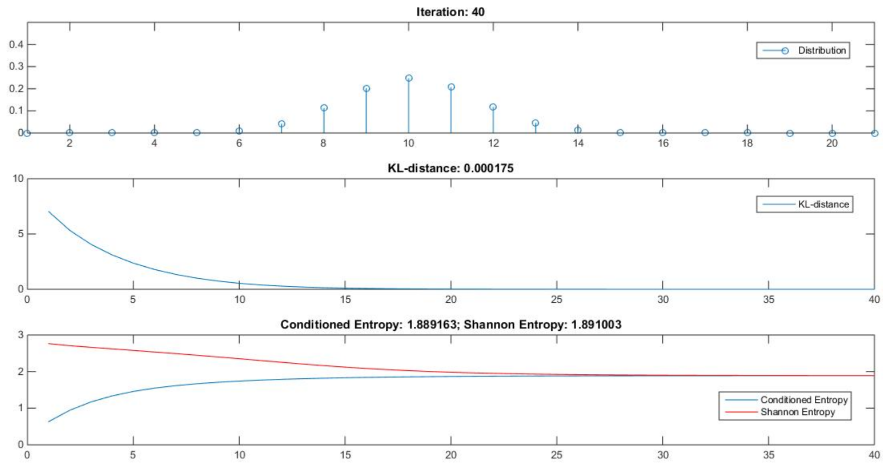


Figure 5: Start with random initial distribution

## Conclusion

Based on the ideas in information theory, we tried to apply various quantities to evaluate a Markov chain in light of demonstrating *The Second Law of Thermodynamics*. First, we find that although Shannon entropy sometimes perfectly describe a stochastic process, it has totally different physical meanings from the classical Boltzmann entropy, which emphasizes on single microscopic state.

As a result, we turn to Kullback-liebler divergence to quantitatively measure an ongoing Markov chain. And it turns out that the Kullback-liebler divergence between current distribution and stationary distribution will monotonically decreasing to 0 while the chain has an unique stationary distribution.

Finally, we showed that the conditional entropy given initial distribution of current distribution will monotonically increasing to the Shannon entropy of stationary distribution. That is, the future will be more and more difficult to predict from the beginning. However, the uncertainty cannot exceed an upper, the entropy of stationary distribution.

After surveying how to apply the concept in Information theory to describe the similar thermodynamic-like behaviour of Markov chain. Now we try to consider the intuition behind this Information theoretical approach.

First, the physical meaning behinds the classical thermodynamics view and information theoretic view have some fundamental difference. For classical thermodynamics, they believe a system will tend to a **single** equilibrium microscopic state. And they admit that they might exist some **fluctuations** near the equilibrium state. This point of view is *spatial*, which focus on the deterministic phenomenon.

However, for information theorist, the entropy they are interested is a description of **probabilistic**.

# Appendix A. Simulation Code

1. main program

```
%% Init
clc;
clear;

n = 20; % number og balls

% transition matrix
A = zeros(n+1,n+1);
A(1,2) = 1;
A(n+1,n) = 1;
for i = 2:n
    A(i,i-1) = (i*i)/(n*n);
    A(i,i+1) = ((n-i)*(n-i))/(n*n);
    A(i,i) = 1 - A(i,i-1) - A(i,i+1);
end
stationary = null(A'-eye(n+1));
stationary = stationary/sum(stationary);

%% Real Simulation
% In this section, we simulate a single trail in the
% urn problem. That is, we start from an initial
% state and then using the transition matrix to
% decide the next state of the system.
```

```
% Initial state
%   Using the number of balls to record which is the
%   current state
a = n; % urn A
niter = 0; % number of iterations

% image setting
imgA = zeros(n,1);
subplot(1,3,1); imshow(kron(imgA,ones(20)));
imgB = ones(n,1);
title(sprintf('Black: %d, White: %d', a, n-a));
subplot(1,3,2); imshow(kron(imgB,ones(20)));
title(sprintf('Black: %d, White: %d', n-a, a));
% state record
s = zeros(n+1,1);
s(1) = 0;
subplot(1,3,3);
stem(s);
axis([1,n+1,0,niter/3]);

for i = 1:niter
    if rand < a/n % urn A choose black
        if rand <= (n-a)/n % urn B choose black
        else % urn B choose white
            imgA(a) = 1;
            imgB(n-a+1) = 0;
            a = a-1;
        end
    else % urn A choose white
        if rand <= (n-a)/n % urn B choose black
            imgA(a+1) = 0;
            imgB(n-a) = 1;
            a = a+1;
        else % urn B choose white
        end
    end

    % update s
    s(n-a+1) = s(n-a+1)+1;

    % show result
```

```
    if mod(i,5) == 0
        subplot(1,3,1);
        imshow(kron(imgA,ones(20)));
        title(sprintf('Black: %d, White: %d', a, n-a));
        subplot(1,3,2);
        imshow(kron(imgB,ones(20)));
        title(sprintf('Black: %d, White: %d', n-a, a));
        subplot(1,3,3);
        stem(s); hold on;
        axis([1,n+1,0,niter/3]);
        title(sprintf('Iteration: %d', i));
        hold off;
        drawnow;
    end
end

%% Distribution Simulation
% We will focus on three properties:
%   1) Shannon Entropy
%   2) Kullback-liebler divergence

% Initial state
start_type = 1;
if start_type == 1 % random
    s = rand(1,n+1);
    s = s/sum(s);
elseif start_type == 2 % uniform
    s = ones(1,n+1);
    s = s/sum(s);
end
ss = s; % record initial state distribution

% Transition matrix from first state
B = A;

niter = n*2;
e = zeros(niter);
d = zeros(niter);
r = zeros(niter);

figure();
```

```
subplot(3,1,1);
stem(s); hold on;
legend('Distribution'); hold off;
axis([1,n+1,0,0.5]);
subplot(3,1,2);
plot(e); hold on;
legend('KL-distance'); hold off;
subplot(3,1,3);
plot(d); hold on;
legend('Conditioned Entropy','Shannon Entropy');
hold off;

for i = 1:niter
    s = s*A;
    d(i) = KL_distance(s',stationary);
    e(i) = entropy(s);
    % relative entropy
    r(i) = relativeentropy(ss,B);
    B = B*A;
    % plot
    subplot(3,1,1); stem(s);
    axis([1,n+1,0,0.5]);
    title(sprintf('Iteration: %d', i));
    legend('Distribution');
    subplot(3,1,2); plot(real(d(1:i)));
    axis([0,niter,0,10]);
    title(sprintf('KL-distance: %f', d(i)));
    legend('KL-distance');
    subplot(3,1,3); plot(real(r(1:i))); hold on;
    plot(e(1:i),'r-'); axis([0,niter,0,3]);
    title(sprintf('Conditioned Entropy: %f; ...
    Shannon Entropy: %f', r(i), e(i)));
    legend('Conditioned Entropy', 'Shannon Entropy');
    hold off;
    drawnow; pause(0.1);
end
```

2. entropy.m

```
function [H] = entropy(p)
H = 0;
for i = 1:length(p)
```

```
    tmp = -p(i)*log(p(i));
    if ~isnan(tmp) && ~isinf(tmp)
        H = H + tmp;
    end
end
end
```

3. KL_distance.m

```
function [d] = KL_distance(s, mu)

if length(s) ~= length(mu)
    disp('KL-dis␣error,␣different␣length');
end

d = 0;
for i = 1:length(s)
    tmp = s(i) .* log(s(i)/mu(i));
    if ~isnan(tmp) && ~isinf(tmp)
        d = d+tmp;
    end
end;
end
```

4. relativeentropy.m

```
function [H] = relativeentropy(ss,B)
H = 0;
A = -B.*log(B);
for i = 1:size(A,1)
    for j = 1:size(A,2)
        if isnan(A(i,j))
            A(i,j) = 0;
        end
    end
end
tmp = ss*(A);
for i = 1:length(tmp)
    if ~isnan(tmp(i))
        H = H + tmp(i);
    end
end
end
```