Computational Statistics

Homework 3

Romane PERSCH

December 1, 2016

5.15 Expectation-Maximization algorithm justification

Let f and g two density functions defined on $\mathcal{X}$.

Let $KL(f|g)$ the Kullback-Leibler information of g at f (or the entropy distance between f and g with respect to f). Note that this "entropy distance" does not define a distance, as $KL(f|g) \neq KL(g|f)$.

(a)-1 Show that $KL(f|g) \geq 0$

$$KL(f|g) = E_f[log(\frac{f(X)}{g(X)})]$$
$$KL(f|g) = E_f[-log(\frac{g(X)}{f(X)})]$$

As (-log) is convex, using Jensen's inequality[1] :

$$-log(E_f[\frac{g(X)}{f(X)}]) \leq E_f[-log(\frac{g(X)}{f(X)})]$$

Which is equivalent to :

$$-log(E_f[\frac{g(X)}{f(X)}]) \leq KL(f|g)$$
$$-log(\int \frac{g(x)}{f(x)} f(x)dx) \leq KL(f|g)$$
$$-log(\int g(x)dx) \leq KL(f|g)$$

As g is a density :

$$-log(1) \leq KL(f|g)$$
$$\boxed{0 \leq KL(f|g)}$$

(a)-2 Show that $KL(f|g) = 0 \quad \Leftrightarrow \quad f = g$

Reminder - Jensen's Inequality corollary : If ϕ is strictly convex, then :
$\phi(E[X]) = E[\phi(X)] \quad \Leftrightarrow \quad X$ is a constant variable equal almost surely to $E(X)$.

As (-log) is strictly convex, applying Jensen's Inequality corollary :
$KL(f|g) = 0 \quad \Leftrightarrow \quad \frac{g(X)}{f(X)}$ is a constant variable equal almost surely to $E_f[\frac{g(X)}{f(X)}]$
$KL(f|g) = 0 \quad \Leftrightarrow \quad \frac{g(X)}{f(X)} = 1$ almost surely

[1] Jensen's inequality : If ϕ is convex and X is an integrable random variable, then $\phi(E[X]) \leq E[\phi(X)]$

$KL(f|g) = 0 \quad \Leftrightarrow \quad g(X) = f(X)$ almost surely
As $X \sim f$:
$KL(f|g) = 0 \quad \Leftrightarrow \quad g(x) = f(x) \quad \forall x \in Supp(f)$[2] (except possibily on a finite set of points)
This implies $\int_{Supp(f)} f(x)dx = \int_{Supp(f)} g(x)dx$ and, as f and g are densities, we know that $\int_{\mathcal{X}} f(x)dx = \int_{Supp(f)} f(x)dx = 1$ and that $\int_{\mathcal{X}} g(x)dx = 1$.
So $\int_{Supp(f)} g(x)dx = 1 = \int_{\mathcal{X}} g(x)dx$. This implies, using Chasles relation, that $\int_{\mathcal{X} \setminus Supp(f)} g(x)dx = 0$.
As g is a density, it is positive. Hence : $\forall x \in \mathcal{X} \setminus Supp(f), \quad g(x) = 0$ (except possibly on a finite set of points).
Thus :

$$\boxed{KL(f|g) = 0 \quad \Leftrightarrow \quad f = g \quad \text{almost everywhere}}$$

(b) Show that this yields to Theorem 5.15

First, let us recall some notations. Assume we observe $X_1, ..., X_n$ iid from $g(x|\theta)$ and we want to compute the Maximimum Likelihood Estimator $\hat{\theta} = argmax \quad L(\theta|x) = argmax \quad \prod_{i=1}^n g(x_i|\theta)$. Assume there are latent variables Z (which we do not observe), such that $X, Z \sim f(x, z|\theta)$ and note the conditional distribution of the latent Z given the observed data x (using Bayes theorem) :

$$k(z|\theta, x) = \frac{f(x, z|\theta)}{g(x|\theta)}$$

.

Hence, the complete-data likelihood can be written as :

$$L^c(\theta|x, z) = \prod_{i=1}^n f(x_i, z_i|\theta) = \prod_{i=1}^n k(z_i|\theta, x_i)g(x_i|\theta)$$

.

Thus :

$$\log L^c(\theta|x, z) = \sum_{i=1}^n \log k(z_i|\theta, x_i) + \log L(\theta|x)$$

Hence, using Linearity of Expectation and integrating with respect to $Z = (Z_1, ..., Z_n)$ where θ_0 is assumed to be the true parameter (i.e : integrating with respect to the joint density $\prod_{i=1}^n k(z_i|\theta_0, x)$) :

$$\forall \theta_0 \quad E_{\theta_0}(\log L(\theta|x)) = E_{\theta_0}(\log L^c(\theta|x, Z)) - E_{\theta_0}(\sum_{i=1}^n \log k(Z_i|\theta, x_i))$$

As $\log L(\theta|x)$ does not depend on Z, we get :

$$\forall \theta_0 \quad \log L(\theta|x) = E_{\theta_0}(\log L^c(\theta|x, Z)) - \sum_{i=1}^n E_{\theta_0}(\log k(Z_i|\theta, x_i)) \qquad (*)$$

Let denote the expected complete-data log-likelihood by :

$$Q(\theta|\theta_0, x) = E_{\theta_0}(\log L^c(\theta|x, Z))$$

The idea behind the EM-algorithm is, at each step, to compute :

$$\hat{\theta}_{(j+1)} = \arg\max_{\theta} Q(\theta|\hat{\theta}_{(j)}, x)$$

This should converge, under certain conditions, to a Maximum Likelihood Estimator (MLE). The purpose of this question is thus to understand why.

Theorem 5.15 : The sequence $(\hat{\theta}_{(j)})_j$ (in the Expectation-Maximization algorithm) satisfies :

$$L(\hat{\theta}_{(j+1)}|x) \geq L(\hat{\theta}_{(j)}|x)$$

In other words, Theorem 5.15 tells us that at each step of the EM-algorithm, the likelihood increases. Note that this does not always imply that the sequence $(\hat{\theta}_{(j)})_j$ *converges* to the MLE. However, it is obviously a first important justification to the EM-algorithm.

[2] $Supp(f)$ is the support of f, i.e $Supp(f) = \{x \in \mathcal{X}, f(x) \neq 0\}$

Proof - Step 1 : The inequality in part (a) implies :

$$KL(f|g) \geq 0$$
$$E_f[log(\frac{f(X)}{g(X)})] \geq 0$$
$$E_f[log(f(X)) - log(g(X))] \geq 0$$

Using Linearity of Expectation :

$$\boxed{E_f[log(f(X))] \geq E_f[log(g(X))]}$$

Proof - Step 2 : Let j be fixed. We can apply the inequality obtained in Step 1 to :

$$f(z) = k(z|\hat{\theta}_{(j)}, x)$$

$$g(z) = k(z|\hat{\theta}_{(j+1)}, x)$$

For each $i = 1, ..., n$ this leads to :

$$E_{\hat{\theta}_{(j)}}[\log k(Z_i|\hat{\theta}_{(j)}, x_i)] \geq E_{\hat{\theta}_{(j)}}[k(Z_i|\hat{\theta}_{(j+1)}, x_i)]$$

If we sum over all i and multiply by (-1), we get :

$$-\sum_{i=1}^{n} E_{\hat{\theta}_{(j)}}[\log k(Z_i|\hat{\theta}_{(j)}, x_i)] \leq -\sum_{i=1}^{n} E_{\hat{\theta}_{(j)}}[k(Z_i|\hat{\theta}_{(j+1)}, x_i)] \qquad (A)$$

By definition of $\hat{\theta}_{(j+1)}$, we also have :

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)}, x) \leq Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)}, x) \qquad (B)$$

Hence, (A) and (B) imply that :

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)}, x) - \sum_{i=1}^{n} E_{\hat{\theta}_{(j)}}[\log k(Z_i|\hat{\theta}_{(j)}, x_i)] \leq Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)}, x) - \sum_{i=1}^{n} E_{\hat{\theta}_{(j)}}[k(Z_i|\hat{\theta}_{(j+1)}, x_i)$$

Using (*), this is equivalent to : $\boxed{L(\hat{\theta}_{(j)}|x) \leq L(\hat{\theta}_{(j+1)}|x)}$

5.9 Mixture model and EM-algorithm

Suppose that the random variable X has a mixture distribution ; that is, the X_i are independently distributed as :

$$X_i \sim \theta g(x) + (1 - \theta)h(x) = p(x|\theta)$$

Where h and g are known. An EM algorithm can be used to find the ML estimator of θ. Introduce $Z_1, ..., Z_n$, where Z_i indicates from which distribution X_i has been drawn, so :

$$X_i|Z_i = 1 \sim g(x)$$

$$X_i|Z_i = 0 \sim h(x)$$

(a) Complete-data likelihood

Let $p(x_i, z_i|\theta)$ be the joint distribution of (X_i, Z_i) when the parameter is θ. Thus, the complete-data likelihood is :

$$L^c(\theta|x, z) = \prod_{i=1}^{n} p(x_i, z_i|\theta)$$

Using Bayes theorem :

$$p(x_i, z_i|\theta) = p(x_i|z_i, \theta)p(z_i|\theta)$$

Where $p(x_i|z_i,\theta)$ is the conditional distribution of X_i given Z_i and $p(z_i|\theta)$ is the distribution of Z_i.
As $X_i|Z_i = 1 \sim g(x)$ and $X_i|Z_i = 0 \sim h(x)$, we have :

$$p(x_i|z_i,\theta) = z_i g(x_i) + (1-z_i)h(x_i)$$

In other words, $p(x_i|z_i,\theta) = g(x_i)$ if $z_i = 1$ and $p(x_i|z_i,\theta) = h(x_i)$ if $z_i = 0$.
Moreover, by definition Z_i is a Bernoulli variable of parameter θ, so its distribution can be written as :

$$p(z_i|\theta) = \theta^{z_i}(1-\theta)^{1-z_i}$$

Hence : $\boxed{L^c(\theta|x,z) = \prod_{i=1}^n [z_i g(x_i) + (1-z_i)h(x_i)]\theta^{z_i}(1-\theta)^{1-z_i}}$

(b) EM sequence

$E(Z_i|\theta,x_i)$ **computation**

$$E(Z_i|\theta,x_i) = \int_{\{0,1\}} z p(z|\theta,x_i)dz$$

Where we integrate with respect to the counting measure. Thus, using Bayes theorem :

$$\begin{aligned}
&= \int_{\{0,1\}} z \frac{p(x_i,z|\theta)}{p(x_i|\theta)}dz \\
&= \int_{\{0,1\}} z \frac{[zg(x_i) + (1-z)h(x_i)]\theta^z(1-\theta)^{1-z}}{\theta g(x_i) + (1-\theta)h(x_i)}dz \\
&= \sum_{z=0,1} z \frac{[zg(x_i) + (1-z)h(x_i)]\theta^z(1-\theta)^{1-z}}{\theta g(x_i) + (1-\theta)h(x_i)} \\
&= 0 + 1 \times \frac{[g(x_i) + 0 \times h(x_i)]\theta^1(1-\theta)^0}{\theta g(x_i) + (1-\theta)h(x_i)}
\end{aligned}$$

$$\boxed{E(Z_i|\theta,x_i) = \frac{\theta g(x_i)}{\theta g(x_i) + (1-\theta)h(x_i)}}$$

EM Sequence Keeping the same notations as in Problem 5.15, the EM sequence is defined by :

$$\hat{\theta}_{(j+1)} = \arg\max_\theta Q(\theta|\hat{\theta}_{(j)},x)$$

Where :

$$Q(\theta|\hat{\theta}_{(j)},x) = E_{\hat{\theta}_{(j)}}(\log L^c(\theta|x,Z))$$

(Reminder : $E_{\hat{\theta}_{(j)}}$ denotes the expectation with respect to $p(z|\hat{\theta}_{(j)},x)$)
In our context, using (a) :

$$Q(\theta|\hat{\theta}_{(j)},x) = E_{\hat{\theta}_{(j)}}(\log(\prod_{i=1}^n [Z_i g(x_i) + (1-Z_i)h(x_i)]\theta^{Z_i}(1-\theta)^{1-Z_i})) \quad (1)$$

$$= E_{\hat{\theta}_{(j)}}(\sum_{i=1}^n \log([Z_i g(x_i) + (1-Z_i)h(x_i)]\theta^{Z_i}(1-\theta)^{1-Z_i})) \quad (2)$$

$$= E_{\hat{\theta}_{(j)}}(\sum_{i=1}^n \log(Z_i g(x_i) + (1-Z_i)h(x_i)) + \sum_{i=1}^n \log(\theta^{Z_i}(1-\theta)^{1-Z_i})) \quad (3)$$

$$= E_{\hat{\theta}_{(j)}}(\sum_{i=1}^n \log(Z_i g(x_i) + (1-Z_i)h(x_i))) + E_{\hat{\theta}_{(j)}}(\sum_{i=1}^n \log(\theta^{Z_i}(1-\theta)^{1-Z_i})) \quad (4)$$

$$= E_{\hat{\theta}_{(j)}}(\sum_{i=1}^n \log(Z_i g(x_i) + (1-Z_i)h(x_i))) + E_{\hat{\theta}_{(j)}}(\sum_{i=1}^n Z_i \log(\theta) + (1-Z_i)\log(1-\theta)) \quad (5)$$

The first expectation does not depend on θ, hence :

$$\begin{aligned}
\max_\theta \quad Q(\theta|\hat{\theta}_{(j)}, x) \quad &\Leftrightarrow \quad \max_\theta \quad E_{\hat{\theta}_{(j)}}(\sum_{i=1}^{n} Z_i \log(\theta) + (1 - Z_i)\log(1-\theta)) \\
&\Leftrightarrow \quad \max_\theta \quad \sum_{i=1}^{n} E_{\hat{\theta}_{(j)}}(Z_i)\log(\theta) + (1 - E_{\hat{\theta}_{(j)}}(Z_i))\log(1-\theta) \\
&\Leftrightarrow \quad \max_\theta \quad \sum_{i=1}^{n} E_{\hat{\theta}_{(j)}}(Z_i)[\log(\theta) - \log(1-\theta)] + n\log(1-\theta)
\end{aligned}$$

As $E_{\hat{\theta}_{(j)}}$ denotes the expectation with respect to $p(z|\hat{\theta}_{(j)}, x)$:

$$\max_\theta \quad Q(\theta|\hat{\theta}_{(j)}, x) \quad \Leftrightarrow \quad \max_\theta \quad n\log(1-\theta) + [\log(\theta) - \log(1-\theta)] \sum_{i=1}^{n} E(Z_i|\hat{\theta}_{(j)}, x_i)$$

Let $H(\theta) = [\log(\theta) - \log(1-\theta)] \sum_{i=1}^{n} E(Z_i|\hat{\theta}_{(j)}, x_i)$. We are thus trying to maximize H wrt θ.

$$\begin{aligned}
H'(\theta) &= -n\frac{1}{1-\theta} + [\frac{1}{\theta} + \frac{1}{1-\theta}] \sum_{i=1}^{n} E(Z_i|\hat{\theta}_{(j)}, x_i) \\
&= -n\frac{1}{1-\theta} + \frac{1}{\theta(1-\theta)} \sum_{i=1}^{n} E(Z_i|\hat{\theta}_{(j)}, x_i)
\end{aligned}$$

Thus :

$$\begin{aligned}
H'(\theta) = 0 \quad &\Leftrightarrow \quad \frac{1}{\theta(1-\theta)} \sum_{i=1}^{n} E(Z_i|\hat{\theta}_{(j)}, x_i) = n\frac{1}{1-\theta} \\
&\Leftrightarrow \quad \sum_{i=1}^{n} E(Z_i|\hat{\theta}_{(j)}, x_i) = n\theta \\
&\Leftrightarrow \quad \frac{1}{n} \sum_{i=1}^{n} E(Z_i|\hat{\theta}_{(j)}, x_i) = \theta
\end{aligned}$$

Hence :

$$\hat{\theta}_{(j+1)} = \frac{1}{n} \sum_{i=1}^{n} E(Z_i|\hat{\theta}_{(j)}, x_i)$$

Using the previous result about $E(Z_i|\hat{\theta}_{(j)}, x_i)$: $\boxed{\hat{\theta}_{(j+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\theta}_{(j)} g(x_i)}{\hat{\theta}_{(j)} g(x_i) + (1-\hat{\theta}_{(j)}) h(x_i)}}$

(c) Show that $\hat{\theta}_{(j)}$ converges to $\hat{\theta}$, a Maximum Likelihood Estimator of θ

As explained in 5.15, the convergence of the sequence $(\hat{\theta}_{(j)})$ to a MLE is not always guaranteed. Theorem 5.15 only garantees that the likelihood is increasing at each step j and :

Theorem 5.16 If $Q(\theta|\theta_0, x)$ is continuous in both θ and θ_0 , then every limit point of an EM sequence $(\hat{\theta}_{(j)})$ is a stationary point of $L(\theta|x)$ and $L(\hat{\theta}_{(j)}|x)$ converges monotonically to $L(\hat{\theta}|x)$ for some stationary point $\hat{\theta}$.

So the convergence is only guaranteed to a stationary point (this can be a local maximizer or a saddle point).

Step 1 - $\hat{\theta}_{(j)}$ converges to a stationary point of $L(\theta|x)$ In our case, $Q(\theta|\theta_0, x)$ is obviously continuous in both θ and θ_0. This can be shown using equation (5) in (b), since :

- $\log(\theta)$ and $\log(1-\theta)$ are continuous in θ as $0 < \theta < 1$
- $E(Z_i|\theta_0, x_i)$ is continuous (in θ_0) using the first result of (b) (indeed, the denominator never equals 0)

- we can easily show that $E_{\theta_0}(\sum_{i=1}^{n} \log(Z_i g(x_i) + (1 - Z_i)h(x_i)))$ is continuous in θ_0 by writing that it is equal to $\sum_{i=1}^{n} \int \log(zg(x_i) + (1-z)h(x_i))p(z|\theta_0, x_i)dz$ and by doing a similar calculation as in the first part of (b).

Therefore, our EM-algorithm is guaranteed to converge to a stationary point of $L(\theta|x)$.

However, in our case, we are also able to show that all stationary points of $L(\theta|x)$ are maximum likelihood estimators.

Step 2 - $\hat{\theta}_{(j)}$ converges to a maximum likelihood estimator We will show that $\log L(\theta|x)$ is concave. This will therefore imply that any stationary point will necessarily be a global maximizer of $\log L(\theta|x)$, and therefore be a maximum likelihood estimator.

$$\begin{aligned}\log L(\theta|x) &= \sum_{i=1}^{n} \log p(x_i|\theta) \\ &= \sum_{i=1}^{n} \log(\theta g(x_i) + (1-\theta)h(x_i)) \\ &= \sum_{i=1}^{n} \log[(g(x_i) - h(x_i))\theta + h(x_i)]\end{aligned}$$

Lemma 1 - The sum of 2 concave functions remains concave Let f, g be two concave functions. Let $\lambda \in [0,1]$. Let x, y.

$$(f+g)(\lambda x + (1-\lambda)y) = f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y)$$

As f is concave and g is concave :

$$\begin{aligned}&\leq \lambda f(x) + (1-\lambda)f(y) + \lambda g(x) + (1-\lambda)g(y) \\ &\leq \lambda(f+g)(x) + (1-\lambda)(f+g)(y)\end{aligned}$$

Thus $f + g$ is concave.

Lemma 2 - $\forall (a,b) \neq (0,0)$ such that $b \geq 0$, $x \mapsto \log(ax+b)$ is concave Let $(a,b) \neq (0,0)$ such that $b \geq 0$ be fixed. Let $f(x) = \log(ax+b)$.
If $a = 0$, then $b > 0$ and $f(x) = \log(b)$, so f is a constant function and is therefore concave.
If $a > 0$, f is twice differentiable on $\left]\frac{-b}{a}; +\infty\right[$
If $a < 0$, f is twice differentiable on $\left]-\infty; \frac{-b}{a}\right[$
In both cases, we get :

$$f'(x) = \frac{a}{ax+b}$$

$$f''(x) = \frac{-a^2}{(ax+b)^2} \leq 0$$

Hence, f is concave.

As $\forall i, \quad h(x_i) \geq 0$ and as $(h(x_i) = 0 \Rightarrow x_i$ has been necessarily drawn from $g \Rightarrow g(x_i) > 0)$, the condition "$(g(x_i) - h(x_i), h(x_i)) \neq (0,0)$ and $h(x_i) \geq 0$" holds in our case. We can therefore apply Lemma 2 and we can conclude that $x \mapsto \log[(g(x_i) - h(x_i))\theta + h(x_i)$ is concave for all i.

Thus, using Lemma 1, $\log L(\theta|x)$ is concave. Hence, any stationary point is necessarily a global maximizer of $\log L(\theta|x)$. In other words, any stationary point is necessarily a maximum likelihood estimator.

Conclusion : $\hat{\theta}_{(j)}$is guaranteed to converge to $\hat{\theta}$, a Maximum Likelihood Estimator of θ

6.9 Show that an aperiodic Markov chain on a finite state-space is irreducible if and only if its transition matrix is regular

Let P be the transition matrix of the studied Markov Chain (X_n) on a finite state-space. Let E be this finite state-space and M its cardinality. In the following statements, we will do as if $E = 1, 2, ..., M$, since we can always map all states to $1, 2, ..., M$ with a bijective function.

Definition P is *regular* if and only if there exists $N \in \mathbb{N}$ such that P^N has no zero entries.

Definition (X_n) is *irreducible* if and only if for all states x, y there exists $n_{x,y}$ such that $P^{n_{x,y}}(x, y) > 0$.

Definition (X_n) is *aperiodic* if and only if all states are of period 1. In other words, (X_n) is *aperiodic* if and only if for all states x GCD $J(x) = 1$ where $J(x) = \{n \in \mathbb{N}^*, P^n(x, x) > 0\}$.

(X_n) **is assumed aperiodic.**

$\Leftarrow$

Assume P is regular. Therefore, there exists $N \in \mathbb{N}$ such that P^N has no zero entries.
Thus, for all states x, y, $P^N(x, y) > 0$. The Markov chain is therefore irreducible.

$\boxed{P \text{ is regular} \Rightarrow (X_n) \quad \text{irreducible}}$

$\Rightarrow$

Assume (X_n) is irreducible.

Lemma 1 Let $k \in \mathbb{N}^*$ an arbitrary but fixed natural number. Then there is a natural number $n_0 \geq 1$ such that $\{n_0, n_0 + 1, n_0 + 2...\} \subset \{n_1 k + n_2(k + 1); (n_1, n_2) \in \mathbb{N}^2 \setminus \{(0, 0)\}\}$.

Lemma 1 - Proof Let $n \geq k^2$. Following the Euclidean division theorem, $\exists m, d \in \mathbb{N} / n - k^2 = mk + d$ and $d < k$.
Hence :

$$\begin{aligned} n &= k^2 + mk + d \\ n &= (k + m - d)k + d(k + 1) \end{aligned}$$

Note that $k + m - d > 0$ since $d < k$.
Therefore, $\forall n \geq k^2, n \in \{n_1 k + n_2(k + 1); (n_1, n_2) \in \mathbb{N}^2 \setminus \{(0, 0)\}\}$. In other words, $n_0 = k^2$ is the desired number.

Lemma 2 Let $x \in E$. If $a, b \in J(x)$, then $\forall (n_1, n_2) \in \mathbb{N}^2 \setminus \{(0, 0)\}, n_1 a + n_2 b \in J(x)$. In other words, $\{n_1 a + n_2 b; (n_1, n_2) \in \mathbb{N}^2 \setminus \{(0, 0)\}\} \subset J(x)$.

Lemma 2 - Proof Let $a, b \in J(x)$. Let us first show that $a + b \in J(x)$.

$$P^{a+b}(x, x) = \sum_{z=1}^{M} P^a(x, z) P^b(z, x)$$

Since P is a transition matrix, all coefficients are positive and therefore all coefficients of powers of P are positive. Thus for all states z : $P^a(x, z)P^b(z, x) \geq 0$
Hence :

$$\begin{aligned} P^{a+b}(x, x) &= \sum_{z=1}^{M} P^a(x, z) P^b(z, x) \\ &\geq P^a(x, x) P^b(x, x) > 0 \end{aligned}$$

Because $a \in J(x) \Leftrightarrow P^a(x, x) > 0$ and $b \in J(x) \Leftrightarrow P^b(x, x) > 0$. So $a + b \in J(x)$.

Therefore, by recurrence, $\forall (n_1, n_2) \in \mathbb{N}^2 \setminus \{(0, 0)\}, n_1 a + n_2 b \in J(x)$.

Lemma 3 Let $x \in E$. Then, $J(x)$ contains two successive numbers. In other words, $\exists k^{(x)} \in J(x) / (k^{(x)} + 1) \in J(x)$.

Lemma 3 - Proof Let $x \in E$. As (X_n) is **aperiodic**, GCD $J(x) = 1$. Therefore, we can find a finite subset $\{l_1, ..., l_n\} \subset J(x)$ such that GCD $\{l_1, ..., l_n\} = 1$ (see Appendix for a proper proof of this statement). Following the generalization of Bézout's identity : $\exists \alpha_1, ..., \alpha_n \in \mathbb{Z}$ such that $1 = \alpha_1 l_1 + ... + \alpha_n l_n \quad (*)$.
Let us take $k = \sum_{i=1}^{n} |\alpha_i| l_i$. Then, $k \in \mathbb{N}$ and, using Lemma 2, $k \in J(x)$ (since $l_1, ..., l_n \in J(x)$, since $|\alpha_i| \in \mathbb{N}$ and since all $|\alpha_i|$ cannot be simultaneously equal to 0 otherwise $(*)$ would not hold).
Hence : $k + 1 = \sum_{i=1}^{n} (|\alpha_i| + \alpha_i) l_i$. As $(|\alpha_i| + \alpha_i) \geq 0$ and as they cannot be all equal to 0, using Lemma 2, this implies that $k + 1 \in J(x)$.
Therefore, we found two consecutive numbers in $J(x)$.

Main Proof - Step 1 : Show that there exists a certain threshold depending on the considered state x such that all successive integers superior to this threshold belong to $J(x)$ Let $x \in E$.
Using Lemma 3, $\exists k^{(x)} \in J(x) / (k^{(x)} + 1) \in J(x)$.
Hence, using Lemma 1 : $\exists n_0^{(x)} \geq 1$ such that $\{n_0^{(x)}, n_0^{(x)} + 1, n_0^{(x)} + 2...\} \subset \{n_1 k^{(x)} + n_2 (k^{(x)} + 1); n_1, n_2 \in \mathbb{N}\}$.
Hence, since $k^{(x)} \in J(x)$ and $k^{(x)} + 1 \in J(x)$, using Lemma 2 :

$$\exists n_0^{(x)} \geq 1 \text{ such that } \{n_0^{(x)}, n_0^{(x)} + 1, n_0^{(x)} + 2...\} \subset J(x)$$

Main Proof - Step 2 : Show that there exists a threshold depending on the states x and y such that for all successive integers superior to this threshold, the corresponding coefficient of P at this power is strictly positive Let $x, y \in E$. As (X_n) has been assumed **irreducible**, there exists $n_{xy} \in \mathbb{N}^*$ such that $P^{n_{xy}}(x, y) > 0$.
Following Step 1, there exists $n_0^{(y)}$ such that $\forall n \geq n_0^{(y)}, P^n(y, y) > 0$.
Let $n \geq n_0^{(y)}$. Therefore, as P is a transition matrix of a Markov chain and hence all its coefficients are positive (implying that all coefficients of powers of P are positive) :

$$\begin{aligned} P^{n_{xy}+n}(x, y) &= \sum_{z=1}^{M} P^{n_{xy}}(x, z) P^n(z, y) \\ &\geq P^{n_{xy}}(x, y) P^n(y, y) > 0 \end{aligned}$$

Main Proof - Step 3 : Show that we can find a common threshold N to all states x and y such that all corresponding coefficient of P at this power N are strictly positive Let $N = \max\{n_{xy} + n_0^{(y)}; x, y \in E\}$. (Note that we only choose one of the potential n_{xy} for each pair x, y and one of the potential $n_0^{(y)}$ for each y. Hence, $\{n_{xy} + n_0^{(y)}; x, y \in E\}$ is necessarily a finite set, since E is finite.)

$$\forall x, y \in E \quad P^N(x, y) = P^{n_{xy}+(N-n_{xy})}(x, y)$$

Since $N \geq n_{xy} + n_0^{(y)} : N - n_{xy} \geq n_0^{(y)}$. Therefore, using Step 2 with $n = N - n_{xy}$:

$$\forall x, y \in E \quad P^N(x, y) = P^{n_{xy}+(N-n_{xy})}(x, y) > 0$$

We thus showed that there exists N such that P^N has no zero entries.

$\boxed{(X_n) \text{irreducible} \Rightarrow P \text{ is regular}}$

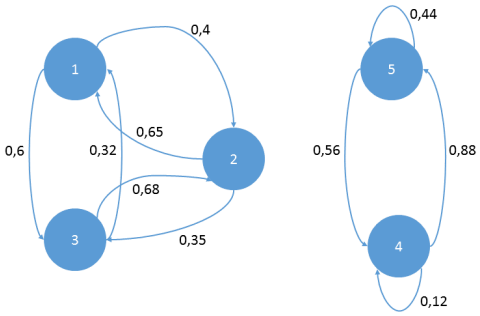
6.7 Given a transition matrix P, examine whether the corresponding Markov chain is irreducible and aperiodic

Let

$$P = \begin{pmatrix} 0.0 & 0.4 & 0.6 & 0.0 & 0.0 \\ 0.65 & 0.0 & 0.35 & 0.0 & 0.0 \\ 0.32 & 0.68 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.12 & 0.88 \\ 0.0 & 0.0 & 0.0 & 0.56 & 0.44 \end{pmatrix}$$

We can therefore draw the graph corresponding to this Markov Chain (see Figure 1).

Figure 1 – Markov Chain graph in Problem 6.7



Irreducibility

Definition (X_n) is *irreducible* if and only if for all states x, y $P_x(\tau_y < +\infty) > 0$ with τ_y being the first positive time y is visited.

Therefore, using Figure 1, we directly see that the corresponding Markov Chain is **not irreducible**, as for example it is impossible to go from state 1 to state 5. In other words, $P_1(\tau_5 < +\infty) = 0$.

More generally, Figure 1 shows us that states 4 and 5 cannot be reached from states 1, 2 and 3 and conversely.

The corresponding Markov Chain is not irreducible.

Aperiodicity

Definition (X_n) is *aperiodic* if and only if all states are of period 1. In other words, (X_n) is *aperiodic* if and only if for all states x GCD $J(x) = 1$ where $J(x) = \{n \in \mathbb{N}^*, P^n(x, x) > 0\}$.

Let us compute the period of each state in our case.

Period of state 5

$$P^1(5, 5) = 0.44 > 0 \quad \text{therefore} \quad 1 \in J(4)$$

Hence, GCD $J(5)$ is necessarily equal to 1, meaning that the period of state 5 is 1.

Period of state 4

$$P^1(4, 4) = 0.12 > 0 \quad \text{therefore} \quad 1 \in J(4)$$

Hence, GCD $J(4)$ is necessarily equal to 1, meaning that the period of state 4 is 1.

Period of state 3 Let us first show that $2 \in J(3)$.
As $(X_2 = 3, X_1 = 1 | X_0 = 3) \subset (X_2 = 3 | X_0 = 3)$:

$$P^2(3, 3) \geq P(X_2 = 3, X_1 = 1 | X_0 = 3)$$

Thus, using Bayes' formula and the Markov property :

$$P^2(3, 3) \geq P(X_1 = 1 | X_0 = 3) P(X_2 = 3 | X_1 = 1)$$

$$P^2(3, 3) \geq 0.32 \times 0.6 > 0 \quad \text{therefore} \quad 2 \in J(3)$$

Let us show that $3 \in J(3)$.
As $(X_3 = 3, X_2 = 2, X_1 = 1 | X_0 = 3) \subset (X_3 = 3 | X_0 = 3)$:

$$P^3(3, 3) \geq P(X_3 = 3, X_2 = 2, X_1 = 1 | X_0 = 3)$$

Thus, using Bayes' formula and the Markov property :

$$P^3(3, 3) \geq P(X_1 = 1 | X_0 = 3) P(X_2 = 2 | X_1 = 1) P(X_3 = 3 | X_2 = 2)$$

$$P^3(3,3) \geq 0.32 \times 0.4 \times 0.35 > 0 \quad \text{therefore} \quad 3 \in J(3)$$

Hence, $2 \in J(3)$ and $3 \in J(3)$. GCD $J(3)$ is thus necessarily equal to 1, as GCD {2,3} = 1. This means that the period of state 3 is 1.

Period of state 2 Using the same method, let us first show that $2 \in J(2)$.
As $(X_2 = 2, X_1 = 1 | X_0 = 2) \subset (X_2 = 2 | X_0 = 2)$:

$$P^2(2,2) \geq P(X_2 = 2, X_1 = 1 | X_0 = 2)$$

Thus, using Bayes' formula and the Markov property :

$$P^2(2,2) \geq P(X_1 = 1 | X_0 = 2) P(X_2 = 2 | X_1 = 1)$$

$$P^2(2,2) \geq 0.65 \times 0.4 > 0 \quad \text{therefore} \quad 2 \in J(2)$$

Let us show that $3 \in J(2)$.
As $(X_3 = 2, X_2 = 3, X_1 = 1 | X_0 = 2) \subset (X_3 = 2 | X_0 = 2)$:

$$P^3(2,2) \geq P(X_3 = 2, X_2 = 3, X_1 = 1 | X_0 = 2)$$

Thus, using Bayes' formula and the Markov property :

$$P^3(2,2) \geq P(X_1 = 1 | X_0 = 2) P(X_2 = 3 | X_1 = 1) P(X_3 = 2 | X_2 = 3)$$

$$P^3(2,2) \geq 0.65 \times 0.6 \times 0.68 > 0 \quad \text{therefore} \quad 3 \in J(2)$$

Hence, $2 \in J(2)$ and $3 \in J(2)$. GCD $J(2)$ is thus necessarily equal to 1, as GCD {2,3} = 1. This means that the period of state 2 is 1.

Period of state 1 Using the same method, let us first show that $2 \in J(1)$.
As $(X_2 = 1, X_1 = 3 | X_0 = 1) \subset (X_2 = 1 | X_0 = 1)$:

$$P^2(1,1) \geq P(X_2 = 1, X_1 = 3 | X_0 = 1)$$

Thus, using Bayes' formula and the Markov property :

$$P^2(1,1) \geq P(X_1 = 3 | X_0 = 1) P(X_2 = 1 | X_1 = 3)$$

$$P^2(1,1) \geq 0.6 \times 0.32 > 0 \quad \text{therefore} \quad 2 \in J(1)$$

Let us show that $3 \in J(1)$.
As $(X_3 = 1, X_2 = 2, X_1 = 3 | X_0 = 1) \subset (X_3 = 1 | X_0 = 1)$:

$$P^3(1,1) \geq P(X_3 = 1, X_2 = 2, X_1 = 3 | X_0 = 1)$$

Thus, using Bayes' formula and the Markov property :

$$P^3(1,1) \geq P(X_1 = 3 | X_0 = 1) P(X_2 = 2 | X_1 = 3) P(X_3 = 1 | X_2 = 2)$$

$$P^3(1,1) \geq 0.6 \times 0.68 \times 0.65 > 0 \quad \text{therefore} \quad 3 \in J(1)$$

Hence, $2 \in J(1)$ and $3 \in J(1)$. GCD $J(1)$ is thus necessarily equal to 1, as GCD {2,3} = 1. This means that the period of state 1 is 1.

Hence, all states are of period 1.
The corresponding Markov Chain is thus aperiodic.

6.54 Show that several examples of Markov Chains are reversible

Before starting the exercise, see Problem 6.53 in Appendix to understand what a reverse Markov Chain is.

A Markov Chain is *reversible* if $P = \tilde{P}$. In other words, it is reversible if $P(X_n = i | X_{n-1} = j) = P(X_{n-1} = j | X_n = i) \quad \forall n$: there is the same probability to go from one state to a new one and to come back from the new one to the previous one if we go back in time.

Hence, a Markov Chain is *reversible* if it satisfies :

$$p_{ij} = \frac{\pi_j p_{ji}}{\pi_i}$$

$$\Leftrightarrow$$

$$\pi_i p_{ij} = \pi_j p_{ji}$$

This last equation is called the the detailed balance condition.

Show that every two-state ergodic chain is reversible Let (X_n) an ergodic Markov Chain in a finite state space with cardinality equal to 2. Thus, its transition matrix P can be written :

$$P = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

As (X_n) is ergodic, there exists an invariant probability distribution π. Thefore, π verifies $\pi P = \pi$ (1) and $\sum_{i=1}^{2} \pi_i = 1$ (2).

$$(1) \quad \Leftrightarrow \quad \begin{cases} a\pi_1 + c\pi_2 = \pi_1 \\ b\pi_1 + d\pi_2 = \pi_2 \end{cases} \quad \Leftrightarrow \quad \begin{cases} c\pi_2 = (1-a)\pi_1 \\ d\pi_2 = (1-b)\pi_2 \end{cases}$$

Moreover, as P is a transition matrix, we have :

$$(*) \begin{cases} a + b = 1 \\ c + d = 1 \end{cases} \quad \Leftrightarrow \quad \begin{cases} b = 1 - a \\ d = 1 - c \end{cases}$$

Let us now check if (X_n) verifies the detailed condition balance.

$$\pi_1 p_{12} = b\pi_1 = (1-a)\pi_1 \quad \text{using } (*)$$

$$\pi_1 p_{12} = c\pi_2 \quad \text{using (1)}$$

Hence : $\boxed{\pi_1 p_{12} = \pi_2 p_{21}}$ Moreover, $\pi_1 p_{11} = \pi_1 p_{11}$ and $\pi_2 p_{22} = \pi_2 p_{22}$ are obviously verified. The detailed condition balance is therefore verified, which implies that (X_n) is reversible.

Show that an ergodic Markov Chain with symmetric transition matrix is reversible Let (X_n) an ergodic Markov Chain with symmetric transition matrix. Let P be the transition matrix and let m be the cardinality of the finite state space on which (X_n) is defined.
As (X_n) is ergodic, there exists a unique invariant probability distribution π.

Let us first show that π is the uniform distribution.
Let u be the uniform distribution on {1, ..., m} : $u = (\frac{1}{m}, ..., \frac{1}{m})$.

$$(uP)_i = \sum_{k=1}^{m} u_k p_{ki}$$

$$(uP)_i = \frac{1}{m} \sum_{k=1}^{m} p_{ki}$$

Since P is a transition matrix, the sum of each row is equal to 1. Therefore $\sum_{k=1}^{m} p_{ik} = 1$. However, as P is symmetric, $p_{ik} = p_{ki}$. Hence : $\sum_{k=1}^{m} p_{ki} = 1$ (the sum of each column is also equal to 1).
Therefore :

$$(uP)_i = \frac{1}{m} = u_i$$

Hence :

$$uP = u \text{ and } \sum_{i=1}^{m} u_i = 1$$

u is thus an invariant distribution. As the invariant distribution is unique, we necessarily have : $\pi = u$.

Let us now show that (X_n) is reversible by checking that the detailed balance condition is verified. Since P is symmetric :

$$\pi_i p_{ij} = \pi_i p_{ji}$$

Since $\pi = u$, $\forall i, j \quad \pi_i = \frac{1}{m} = \pi_j$. Therefore :

$$\pi_i p_{ij} = \pi_j p_{ji}$$

The detailed balance condition is thus verified. $\boxed{(X_n) \text{ is reversible.}}$

Examine whether the matrix P is reversible

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Let us find first if there is an invariant distribution π, i.e if $\exists \pi$ such that $\pi P = \pi$ and $\sum_{i=1}^{5} \pi_i = 1$.

$$\begin{cases} \pi P = \pi \\ \sum_{i=1}^{5} \pi_i = 1 \end{cases} \Leftrightarrow \begin{cases} 0.5\pi_2 = \pi_1 \\ 0.5\pi_3 = \pi_2 \\ \pi_1 + 0.5\pi_2 + 0.5\pi_4 + \pi_5 = \pi_3 \\ 0.5\pi_3 = \pi_4 \\ 0.5\pi_4 = \pi_5 \\ \sum_{i=1}^{5} \pi_i = 1 \end{cases}$$

$$\begin{cases} \pi P = \pi \\ \sum_{i=1}^{5} \pi_i = 1 \end{cases} \Leftrightarrow \begin{cases} \pi_2 = 2\pi_1 \\ \pi_3 = 4\pi_1 \\ \pi_1 + \pi_1 + \pi_1 + \pi_1 = 4\pi_1 \\ \pi_4 = 2\pi_1 \\ \pi_5 = \pi_1 \\ \sum_{i=1}^{5} \pi_i = 1 \end{cases}$$

$$\begin{cases} \pi P = \pi \\ \sum_{i=1}^{5} \pi_i = 1 \end{cases} \Leftrightarrow \begin{cases} \pi_2 = 2\pi_1 \\ \pi_3 = 4\pi_1 \\ \pi_4 = 2\pi_1 \\ \pi_5 = \pi_1 \\ \pi_1 + 2\pi_1 + 4\pi_1 + 2\pi_1 + \pi_1 = 1 \end{cases}$$

$$\begin{cases} \pi P = \pi \\ \sum_{i=1}^{5} \pi_i = 1 \end{cases} \Leftrightarrow \begin{cases} \pi_2 = \frac{2}{10} \\ \pi_3 = \frac{4}{10} \\ \pi_4 = \frac{2}{10} \\ \pi_5 = \frac{1}{10} \\ \pi_1 = \frac{1}{10} \end{cases}$$

$$\begin{cases} \pi P = \pi \\ \sum_{i=1}^{5} \pi_i = 1 \end{cases} \Leftrightarrow \pi = (0.1, 0.2, 0.4, 0.2, 0.1)$$

Let us now check if the detailed balance condition is verified.

$$\pi_2 p_{21} = 0.2 \times 0.5 = 0.1$$

$$\pi_1 p_{12} = 0.1 \times 0 = 0$$

Hence :

$$\pi_2 p_{21} \neq \pi_1 p_{12}$$

The detailed balance condition is thus not verified. The Markov Chain defined by P is not reversible

Appendix

Appendix to 6.9

Lemma Let S a set (finite or infinite) such as GCD $S = 1$. Then, there exists a finite set S' such that GCD $S' = 1$.

Proof : In the case where S is finite, the statement is obvious. In the case where S is infinite :
Let $n \in S / n > 1$ (as S is infinite, such n necessarily exists).
Let d a divisor of n such that $d > 1$. Hence, $\exists n_d \in S$ such that d does not divide n_d (otherwise, d would

be a common divisor of all elements in S, which is contradictory to GCD $S = 1$). Let then D be the set of such divisors of n (all divisors of n that are > 1).
Let $S' = \{n\} \cup \{n_d : d \in D\}$.
Hence : GCD $S' = 1$ and S' is finite.

Appendix to 6.54 : Problem 6.53 - Understanding Reverse Markov Chains

Given a finite state-space Markov chain (X_n), with transition matrix P, define a second transition matrix $\tilde{P}$ by :

$$p_{ij}(n) = \frac{P_\mu(X_{n-1} = j)P(X_n = i|X_{n-1} = j)}{P_\mu(X_n = i)}$$

Where μ is the initial distribution of the Markov Chain with transition matrix P.
N.B: There is a typo in the exercise : the denominator is $P_\mu(X_n = i)$ and not $P_\mu(X_n = j)$.

(a) Show that $\tilde{p}_{ij}(n)$ does not depend on n if the chain is stationary (i.e if $\mu = \pi$) By definition and since $\mu = \pi$:

$$\tilde{p}_{ij}(n) = \frac{P_\pi(X_{n-1} = j)P(X_n = i|X_{n-1} = j)}{P_\pi(X_n = j)}$$

If μ is equal to the invariant probability distribution π, then the chain is stationary. Indeed, $X_n \sim \mu P^n = \pi P^n = \pi$ (by iteration, since $\pi P = \pi$). Thus, all X_n have the same distribution π. Hence :

$$P_\pi(X_{n-1} = j) = P(X_0 = j) = \pi_j$$

$$P_\pi(X_n = i) = P(X_0 = i) = \pi_i$$

And by definition of a transition matrix :

$$P(X_n = i|X_{n-1} = j) = p_{ji}(n)$$

Since the Markov Chain is stationary :

$$P(X_n = i|X_{n-1} = j) = P(X_1 = i|X_0 = j) = p_{ji}(1) = p_{ji}$$

Where p_{ji} is the coefficient located at row j and column i of the matrix P. Hence :

$$\tilde{p}_{ij}(n) = \frac{\pi_j p_{ji}}{\pi_i}$$

Conclusion : If the chain is stationary, $\tilde{p}_{ij}(n)$ does not depend on n

(b) Explain why in this case the chain with transition matrix $\tilde{P}$ is called the reverse Markov Chain Using (a) :

$$\tilde{p}_{ij} = \frac{\pi_j p_{ji}}{\pi_i} = \tilde{p}_{ij}(n) \quad \forall n$$

This is equivalent to :

$$\tilde{p}_{ij} = \frac{P_\pi(X_{n-1} = j)P(X_n = i|X_{n-1} = j)}{P_\pi(X_n = j)} \quad \forall n$$

Using Bayes' formula, this is also equivalent to :

$$\tilde{p}_{ij} = P(X_{n-1} = j|X_n = i) \quad \forall n$$

Therefore, the transition matrix $\tilde{P}$ defines the "same" Markov chain as P but with time running backwards.