# Stochastic Modeling

Philippe Chevalier
UCLouvain

Section 2: Finite State Markov Chains

# Introduction

**definition**

A Markov Chain is a stochastic process with fixed intervals $\{X_n, n \geq 0\}$ such that each random variable $X_n, n \geq 1$ depends on the past only through the most recent random variable $X_{n-1}$

$$P[X_n = j | X_{n-1} = i, X_{n-2} = k, \ldots, X_0 = m] = P[X_n = j | X_{n-1} = \\ = P_{ij}$$

The random variable $X_n$ is called the state of the Markov Chain

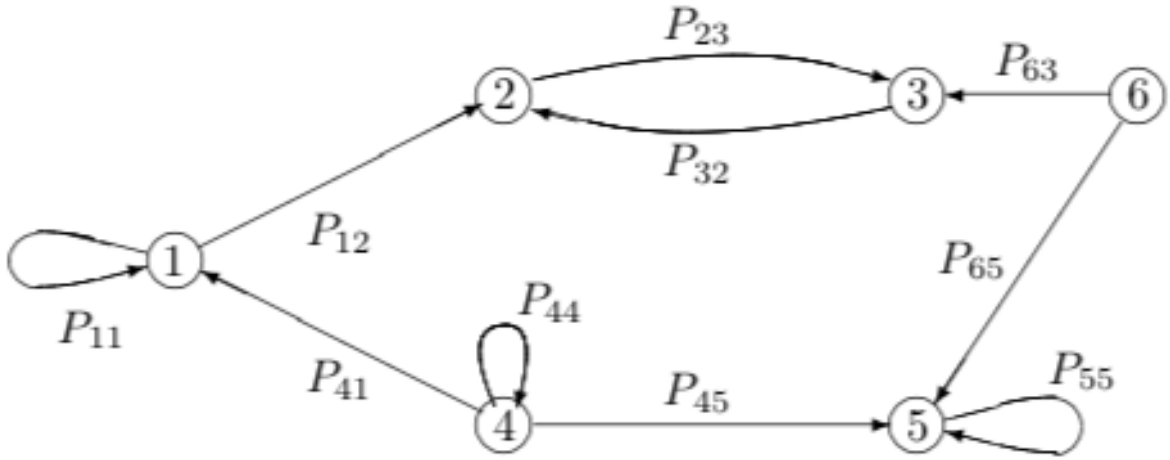
The set of possible sample values for the states lie in a countable set (of finite size in this section)

We will generally call the possible states $\{1, 2, \ldots, J\}$.

When the transition probabilities depend on the period $(p_{ij}(n))$, we have a non-homogeneous Markov Chain (not studied in this course)

# Two representations

Graphical




Matrix

$$P = [P_{ij}] = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1J} \\ P_{21} & P_{22} & \cdots & P_{2J} \\ \vdots & & \ddots & \vdots \\ P_{J1} & P_{J2} & \cdots & P_{JJ} \end{bmatrix}$$

# Accessibility

**Definition**

A state $j$ is accessible from $i$ (abbreviated as $i \to j$) if there is a walk in the graph from $i$ to $j$

.

- We denote $P_{ij}^n = P[X_n = j | X_0 = i]$
- $i \to j$ if and only if $P_{ij}^n > 0$ for some $n$.

# Communication

**Definition**

Two distinct states $i$ and $j$ communicate (abbreviated as $i \leftrightarrow j$) if $i$ is accessible form $j$ and $j$ is accessible form $i$

$i \leftrightarrow i$ by definition

Note that if $i \leftrightarrow j$ and $m \leftrightarrow j$ then $i \leftrightarrow m$

# Classification of states

**Definition**

A class $C$ of states is a non-empty set of states such that for each $i \in C$, each state $j \neq i$ satisfies $j \in C$ if $i \leftrightarrow j$ and $j \notin C$ if $j \not\leftrightarrow i$.

- A state $i$ is *recurrent* if it is accessible from all states that are accessible from $i$. ($i$ is recurrent if $i \to j$ implies that $j \to i$).
- A transient state is a state that is not recurrent.
- All states from a same class will be of the same type.
- A finite state Markov Chain has at least one recurrent class.

# Periodicity

**Definition**

The period of a state $i$, denoted $d(i)$, is the greatest common divisor (gcd) of those values of $n$ for which $P_{ii}^n > 0$.
If the period is 1, the state is *aperiodic*, and if the period is 2 or more, the state is *periodic*.

- All the states of a class will have the same periodicity
- If a class has a period $d > 1$, then there exists a partition $C_1, C_2, \ldots, C_d$ of the states of this class in $d$ subclasses such that all the transitions from a state of $C_n$ go to a state of class $C_{n+1}$ and all transitions from $C_d$ go to a state of $C_1$
- If $j \in C_n$ and $p_{jk} > 0$ then $k \in C_{n+1}$
- A class that is aperiodic and recurrent is called *ergodic*

# Computing transition probabilities

$$P[X_{n+2} = j | X_n = i] = P_{ij}^2 = \sum_{k=1}^{J} P_{ik} P_{kj}$$

$$\Rightarrow P^2 = P.P \quad \text{et} \quad P^n = \underbrace{P.P.P \ldots P}_{n}$$

More generally:

$$P_{ij}^{m+n} = \sum_{k=1}^{J} P_{ik}^m P_{kj}^n$$

$$P^{m+n} = P^m.P^n$$

# Convergence

- A matrix P will be called a stochastic matrix iff the matrix is square, non- negative and each row sums to 1.
- What would be the limit of $P^n$ when $n$ tends to infinity?
- If $P$ is ergodic, we would expect all rows to converge towards the same value, let us call this row vector $\pi$, then
  $P^\infty = P^\infty P \Rightarrow \pi = \pi P$
  We must also impose that the values of the vector $\pi$ sum to 1.
- When does this system have a solution?
  When is it unique?
  When does $P^\infty$ converge?

# A bit of Matrix theory

- Note that because $P$ is a stochastic matrix we have : $Pe = e$
- This implies that 1 is an eigenvalue of the matrix $P$
- The solution $\pi$ is a left eigenvector for the eigenvalue 1
- The number of linearly independent solutions corresponds to the multiplicity of the eigenvalue 1
  - There will be one independent solution for each recurrent class of $P$
- If $P$ is aperiodic then $\lim_{n\to\infty} P^n = e\pi$
- else $\pi$ will be the average over the different subclasses

# Tropical weather

- A little tired with Belgian weather, you decide to go for a vacation in the tropics ...
- After some research you find some reliable statistics about the weather there. It appears that the weather follows a clear pattern:
  - If today is sunny, there is an 80% chance of a sunny day tomorrow, else it will be cloudy.
  - If today is cloudy, there is a 25% chance that it might rain tomorrow, a 25% chance that it stays cloudy, else the weather will be sunny again.
  - If it rains, there is a 50% chance that the rain lasts for one day else the rain lasts for 2 days, when the rain finishes, the weather is always sunny.
- What is the probability of a sunny day on your arrival?
  How many days of sunshine would you expect during a week?

# Find the Stationary probabilities

$$P = \begin{pmatrix} 0,1 & 0,9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0,4 & 0 & 0,4 & 0 & 0 & 0,2 & 0 & 0 & 0 & 0 \\ 0,25 & 0 & 0,15 & 0,2 & 0 & 0 & 0.2 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0,25 & 0,75 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,75 & 0,25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0,5 & 0,5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0,25 & 0,25 & 0 & 0,5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0,5 & 0 & 0 & 0,5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0,25 & 0,75 & 0 & 0 \end{pmatrix}$$

# Markov chains with rewards

Let $r_i$ be a reward associated with state $i$
We could also define $r_{ij}$ as the reward associated to a transition between states $i$ and $j$, the expected gain for each passage through state $i$ is then:

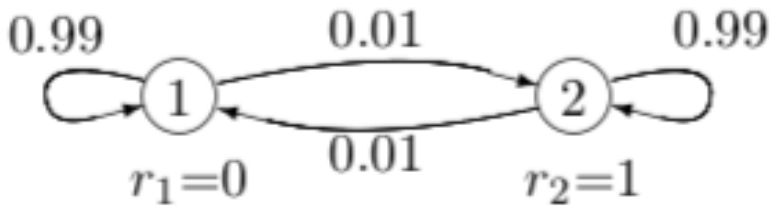
$$r_i = \sum_j P_{ij} r_{ij}$$

For an ergodic chain, we observe that $g$ the average reward per period will be

$$g = \sum_i r_i \pi_i$$

# Transient rewards

We will see that the transient rewards are very interesting to understand the behaviour of Markov chains.
To motivate this, let us use the following example:



# Expected reward over multiple transitions

Let $X_m$ be the state at time $m$ and let $R_m = R(X_m)$ be the reward at time $m$, i.e., if the sample value of $X_m$ is $i$, then $r_i$ is the sample value of $R_m$.
Conditional on $X_m = i$, the aggregate expected reward $v_i(n)$ over n periods from $X_m$ to $X_{m+n-1}$ is

$$\begin{aligned} v_i(n) &= E[R(X_m) + R(X_{m+1}) + \cdots + R(X_{m+n-1}) | X_m = i] \\ &= r_i + \sum_j P_{ij} r_j + \cdots + \sum_j P_{ij}^{n-1} r_j \end{aligned}$$

In vector notation:

$$\boldsymbol{v}(n) = \boldsymbol{r} + [P]\boldsymbol{r} + \cdots + [P^{n-1}]\boldsymbol{r} = \sum_{h=0}^{n-1} [P^h]\boldsymbol{r}$$

Assuming the Markov chain is an *ergodic unichain* (has a single ergodic class with possibly some transient classes),

$$\lim_{n\to\infty}[P^n] = \boldsymbol{e\pi}$$

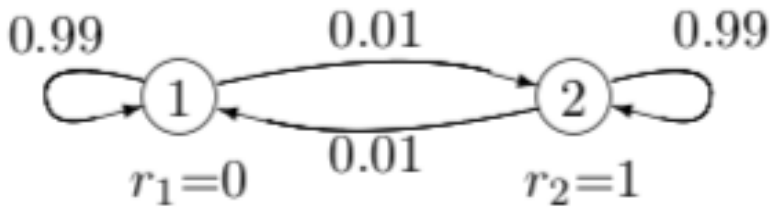
$$\Rightarrow \lim_{n\to\infty}[P^n]\boldsymbol{r} = \boldsymbol{e\pi r} = g\boldsymbol{e}$$

So when $n$ grows, the expected reward per period converges to $g$. The transient effect can thus be evaluated by:

$$\lim_{n\to\infty}(\boldsymbol{v}(n) - ng\boldsymbol{e}) = \lim_{n\to\infty}\sum_{h=0}^{n-1}[P^h - \boldsymbol{e\pi}]\boldsymbol{r}$$

From the properties of $P$ we can show that this limit exist. It is called the relative-gain vector and denoted by $\boldsymbol{w}$.

# Example illustration




| $n$ | $v_1(n)$ | $v_2(n)$ |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 0.01 | 1.99 |
| 4 | 0.0592 | 3.9408 |
| 10 | 0.4268 | 9.5732 |
| 40 | 6.1425 | 33.8575 |
| 100 | 28.3155 | 71.6845 |
| 400 | 175.007 | 224.9923 |

# The relative gain vector $w$

The relative gain vector $\boldsymbol{w}$ can be computed by solving the following equations

$$\boldsymbol{w} + g\boldsymbol{e} = [P]\boldsymbol{w} + \boldsymbol{r} \qquad \text{and} \qquad \boldsymbol{\pi w} = 0$$

Proof:

$$\begin{aligned}
[P]\boldsymbol{w} &= \lim_{n\to\infty} \sum_{h=0}^{n-1} [P^{h+1} - \boldsymbol{e\pi}]\boldsymbol{r} = \lim_{n\to\infty} \sum_{h=1}^{n} [P^{h} - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \lim_{n\to\infty} \left( \sum_{h=0}^{n} [P^{h} - \boldsymbol{e\pi}]\boldsymbol{r} \right) - [P^0 - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \boldsymbol{w} - [P^0 - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \boldsymbol{w} - \boldsymbol{r} + g\boldsymbol{e}
\end{aligned}$$

Note that this linear system always admits a solution, the normalization constraint on $\boldsymbol{w}$ is arbitrary.

# Markov Decision Processes

Suppose that in each state $i$, we can choose between $K^i$ different possibilities with rewards $r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(K_i)}$ with the corresponding transition probabilities
$\{P_{ij}^{(1)}, 1 \leq j \leq J\}, \{P_{ij}^{(2)}, 1 \leq j \leq J\}, \ldots, \{P_{ij}^{(K_i)}, 1 \leq j \leq J\}$

Example :


- Can we find an optimal stationary policy?
- Can we find an optimal dynamic policy?

# Dynamic Programming algorithm for the dynamic optimal policy

Let $\boldsymbol{v}(0)$ be the final reward, 1 period before the end we can compute the optimal action by:

$$v_i^*(1) = \max_k \{r_i^{(k)} + \sum_j P_{ij}^{(k)} v_j(0)\}$$

The expected global reward with 2 periods to go is: $r_i^{(k)} + \sum_j P_{ij}^{(k)} v_j(1)$, to maximize $\boldsymbol{v}(2)$ we will need $\boldsymbol{v}^*(1)$

$$v_i^*(2) = \max_k \{r_i^{(k)} + \sum_j P_{ij}^{(k)} v_j^*(1)\}$$

With $n$ periods to go:

$$v_i^*(n) = \max_k \{r_i^{(k)} + \sum_j P_{ij}^{(k)} v_j^*(n-1)\}$$

In vector format we have:

$$\boldsymbol{v}^*(n) = \max_A \{\boldsymbol{r}^A + [P^A]\boldsymbol{v}^*(n-1)\}$$

or equivalently:

$$\boldsymbol{v}^*(n) = \boldsymbol{r}^B + [P^B]\boldsymbol{v}^*(n-1)$$

With $B$ such that

$$\boldsymbol{r}^B + [P^B]\boldsymbol{v}^*(n-1) \geq \boldsymbol{r}^A + [P^A]\boldsymbol{v}^*(n-1) \qquad \forall A$$

Where $A$ and $B$ are policies (i.e. a decision $k_i$ for each state $i$)

# Optimal stationary policy

We assume for any policy $A$ the resulting Markov chain with matrix $[P^A]$ is an ergodic unichain.
We know that for a policy $A$

$$\begin{aligned}
\boldsymbol{w}^A &= \boldsymbol{r}^A - g^A\boldsymbol{e} + [P^A]\boldsymbol{w}^A \\
&= \boldsymbol{r}^A - g^A\boldsymbol{e} + [P^A](\boldsymbol{r}^A - g^A\boldsymbol{e} + [P^A]\boldsymbol{w}^A) \\
&= \boldsymbol{r}^A + [P^A]\boldsymbol{r}^A - 2g^A\boldsymbol{e} + [P^A]^2\boldsymbol{w}^A \\
&= \boldsymbol{r}^A + [P^A]\boldsymbol{r}^A + [P^A]^2\boldsymbol{r}^A - 3g^A\boldsymbol{e} + [P^A]^3\boldsymbol{w}^A \\
&= \boldsymbol{r}^A + [P^A]\boldsymbol{r}^A + [P^A]^2\boldsymbol{r}^A + \cdots + [P^A]^{n-1}\boldsymbol{r}^A - ng^A\boldsymbol{e} + [P^A]^n\boldsymbol{w} \\
\boldsymbol{v}^A(n) &= \boldsymbol{r}^A + [P^A]\boldsymbol{r}^A + [P^A]^2\boldsymbol{r}^A + \cdots + [P^A]^{n-1}\boldsymbol{r}^A + [P^A]^n\boldsymbol{v}(0)
\end{aligned}$$

Hence:

$$\boldsymbol{v}^A(n) = ng^A\boldsymbol{e} + \boldsymbol{w}^A + [P^A]^n(\boldsymbol{v}(0) - \boldsymbol{w}^A)$$

The policy $A$ that maximizes $g^A$ will be called the optimal stationary policy.

# A particular case

If $\boldsymbol{v}(0) = \boldsymbol{w}^B$ for a policy $B$ such that

$$\boldsymbol{r}^B + [P^B]\boldsymbol{w}^B \geq \boldsymbol{r}^A + [P^A]\boldsymbol{w}^B \qquad \forall A$$

then $B$ is the dynamic optimal policy for each time period and

$$\boldsymbol{v}^*(n) = \boldsymbol{w}^B + ng^B\boldsymbol{e}$$

Proof:

$$\boldsymbol{v}^*(1) = \boldsymbol{r}^B + [P^B]\boldsymbol{w}^B = \boldsymbol{w}^B + g^B\boldsymbol{e} \qquad \Rightarrow \boldsymbol{w}^B = \boldsymbol{v}^*(1) - g^B\boldsymbol{e}$$

Our hypothesis implies thus that:

$$\boldsymbol{r}^B + [P^B]\boldsymbol{v}^*(1) - g^B[P^B]\boldsymbol{e} \geq \boldsymbol{r}^A + [P^A]\boldsymbol{v}^*(1) - g^B[P^A]\boldsymbol{e} \quad \forall A$$

$$\Rightarrow \quad \boldsymbol{r}^B + [P^B]\boldsymbol{v}^*(1) \geq \boldsymbol{r}^A + [P^A]\boldsymbol{v}^*(1) \quad \forall A$$

We can then repeat the same argument for $\boldsymbol{v}^*(2), \boldsymbol{v}^*(3), \ldots$

# Theorem

If for any policy $A$ the Markov chain $[P^A]$ is an ergodic unichain, then the policy $B$ is an optimal stationary policy if and only if:

$$\boldsymbol{r}^B + [P^B]\boldsymbol{w}^B \geq \boldsymbol{r}^A + [P^A]\boldsymbol{w}^B \qquad \forall A \qquad (*)$$

Proof:

$\Leftarrow$

If $w^B$ were the final reward, then $B$ would be the optimal dynamic policy.

If there existed a policy $A$ such that $g^A > g^B$ then in the long term the policy $A$ would generate a larger reward, hence a contradiction.

$\Rightarrow$

Let us assume $(*)$ is not satisfied.

Let $B = (k_1, k_2, \ldots, k_J)$, there must exist $i$ and $k$ such that

$$r_i^{(k_i)} + \sum_j P_{ij}^{(k_i)} w_j^B \leq r_i^{(k)} + \sum_j P_{ij}^{(k)} w_j^B$$

Let us build $A = (k'_1, k'_2, \ldots, k'_J)$ such that $k'_i$ maximizes $r_i^{(k)} + \sum_j P_{ij}^{(k)} w_j^B$

By our assumption there exists at least one index $i$ such that $k_i \neq k'_i$. As a result:

$$\boldsymbol{r}^B + [P^B]\boldsymbol{w}^B \overset{\neq}{\leq} \boldsymbol{r}^A + [P^A]\boldsymbol{w}^B$$

$$\boldsymbol{w}^B + g^B \boldsymbol{e} \overset{\neq}{\leq} \boldsymbol{r}^A + [P^A]\boldsymbol{w}^B$$

Hence:

$$\boldsymbol{\pi}^A \boldsymbol{w}^B + g^B \boldsymbol{\pi}^A \boldsymbol{e} < \boldsymbol{\pi}^A \boldsymbol{r}^A + \boldsymbol{\pi}^A [P^A] \boldsymbol{w}^B \quad \Rightarrow \quad g^B < g^A \qquad \square$$

# Policy Improvement algorithm

1. Choose an arbitrary policy $B$
2. Compute $w^B$
3. If $\boldsymbol{r}^B + [P^B]\boldsymbol{w}^B \geq \boldsymbol{r}^A + [P^A]\boldsymbol{w}^B \qquad \forall A$ then stop
   else continue
4. Find $A$ such that $\boldsymbol{r}^A + [P^A]\boldsymbol{w}^B \overset{\neq}{\geq} \boldsymbol{r}^B + [P^B]\boldsymbol{w}^B$
5. $B \leftarrow A$ go to step 2.

**Theorem**

Assuming for any policy $A$ the Markov chain $[P^A]$ is an ergodic unichain, if $B$ is an optimal stationary policy then

$$\lim_{n\to\infty} \boldsymbol{v}^*(n) - ng^B\boldsymbol{e} = \boldsymbol{w}^B + (\beta - \boldsymbol{\pi}^B\boldsymbol{w}^B)\boldsymbol{e}$$

- Average reward from optimal dynamic policy $= g^B$
- Relative gain vector for optimal dynamic policy $= \boldsymbol{w}^B$
- Relative gain from all optimal stationary policies are identical
- $\min_i[v_i^*(n) - v_i^*(n-1)] \leq g^B \leq \max_i[v_i^*(n) - v_i^*(n-1)]$

# Dynamic Programming algorithm for the stationary optimal policy

1. Fix an arbitrary vector $\boldsymbol{v}(0)$
2. Compute $\boldsymbol{v}^*(n) = \max_A\{\boldsymbol{r}^A + [P^A]\boldsymbol{v}^*(n-1)\}$
3. Compute $l = \min_i[v_i^*(n) - v_i^*(n-1)]$ and $u = \max_i[v_i^*(n) - v_i^*(n-1)]$
4. If $l < u$ go to step 2.
5. $A$ is the optimal stationary policy and the maximum stationary reward is $l = u = g^A$.

# Inventory Management example

You are managing inventories for a clinical trial, for drugs A the data are the following

- Weekly demand can be 0, 1 or 2 with respective probabilities 0.3 0.5 and 0.2
- Ordering cost is 10 per order + 20 per kit
- Holding cost is 4 per kit and per week
- The management says you cannot store more than 3 kits
- If you run out of units and a patient has to be supplied with an emergency shipment, it costs 50 per kit

Given that you can order at the end of a week for the beginning of next week (but you do not know the demand of next week), what would be the optimal ordering policy?