# Devoir maison - Aspects probabilistes de l'informatique

## 1 Espérance de la récompense maximale

Soit $X_i$ l'état aléatoire à l'instant $i$ et $Y_i$ l'action aléatoire à l'instant $i$ d'un MDP. Etant donné une politique $\boldsymbol{\pi}$, l'*espérance de la récompense maximale* à horizon temporel $t$ de $\boldsymbol{\pi}$ est définie par :

$$M_t^{\boldsymbol{\pi}} \stackrel{\text{def}}{=} \mathbf{E}^{\boldsymbol{\pi}} \left(\max(r(X_i, Y_i) \mid 0 \leq i < t)\right)$$

Le vecteur de récompenses correspondant (qui dépend de l'état initial) est noté $\mathbf{M}_t^{\boldsymbol{\pi}}$. Comme vu en cours, le vecteur de récompense optimal $\mathbf{M}_t^*$ est défini par : pour tout $s \in S$, $\mathbf{M}_t^*[s] \stackrel{\text{def}}{=} \sup_{\boldsymbol{\pi}}(\mathbf{M}_t^{\boldsymbol{\pi}}[s])$.

**Question 1.** Exhiber un exemple de MDP tel qu'aucune politique markovienne ne soit optimale pour l'espérance de la récompense maximale à horizon 3.

**Question 2.** Soit $\mathcal{M}$ un MDP et $t$ un horizon. Proposer un algorithme qui renvoie la récompense optimale et la politique optimale pour l'espérance de la récompense maximale en temps polynomial par rapport à la taille de $\mathcal{M}$ et en temps pseudo-polynomial par rapport à $t$.
*Indication : L'algorithme construit un* MDP $\mathcal{M}'$ *tel qu'à partir de la récompense optimale et de la politique optimale pour l'espérance de la récompense totale pure dans* $\mathcal{M}'$*, on peut retrouver la récompense optimale et de la politique optimale pour l'espérance de la récompense maximale dans* $\mathcal{M}$*.*

## 2 Composants terminaux d'un MDP

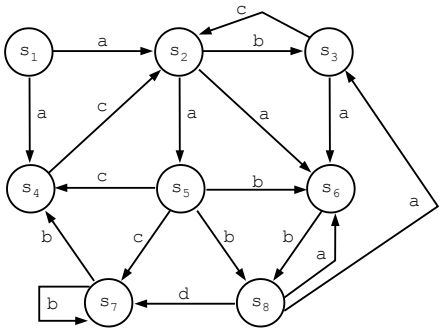
Soit $\mathcal{M}$ un MDP, nous introduisons la notion de sous-MDP. Un sous-MDP $\mathcal{M}'$ de $\mathcal{M}$ est un ensemble *non vide* de couples états-actions tel que $(s, a) \in \mathcal{M}'$ implique que $s \in S$ et $a \in A_s$. Le graphe sous-jacent de $\mathcal{M}'$, $G_{\mathcal{M}'} = (S', E')$ est défini par :

1. $S' \stackrel{\text{def}}{=} \{s \in S \mid \exists (s, a) \in \mathcal{M}'\}$ ;
2. $E' \stackrel{\text{def}}{=} \{(s, s') \in (S')^2 \mid \exists (s, a) \in \mathcal{M}'$ ave $p(s'|s, a) > 0\}$.

Un sous-MDP$\mathcal{M}'$ est un *composant terminal* de $\mathcal{M}$ si :

1. Pour tout $s, s' \in S$, $a \in A_s$, $(s, a) \in \mathcal{M}'$ et $p(s'|s, a) > 0$ implique $s' \in S'$ ;
2. $G_{\mathcal{M}'}$ est fortement connexe.

$\mathcal{M}'$, un composant terminal de $\mathcal{M}$, est *maximal* s'il n'y a pas de composant terminal de $\mathcal{M}''$ avec $S' \subseteq S''$, $E' \subseteq E''$ et $S' \cup E' \subsetneq S'' \cup E''$.




Nous avons dessiné ci-dessus $G_{\mathcal{M}}$ le graphe sous-jacent d'un MDP $\mathcal{M}$ où une action $a$ étiquette un arc $(s, s')$ si $p(s'|s, a) > 0$.

**Question 3.** Soit $\mathcal{M}$ le MDP dont le graphe est dessiné ci-dessus. Trouver un composant terminal maximal de $\mathcal{M}$ et un composant terminal non maximal de $\mathcal{M}$.

Soit $\rho = s_0, a_0, s_1, a_1, \ldots$ une séquence infinie. Nous définissons $\omega(\rho) \stackrel{\text{def}}{=} \{(s,a) \mid \forall i \in \mathbb{N}\ \exists j \geq i\ (s_j, a_j) = (s,a)\}$, l'ensemble des couples états-actions apparaissant infiniment souvent dans $\rho$.

**Question 4.** Soit $\boldsymbol{\pi}$ une politique et $\rho = X_0, Y_0, X_1, Y_1, \ldots$ la séquence aléatoire d'un MDP. Prouver que :

$$\mathbf{Pr}^{\boldsymbol{\pi}}(\omega(\rho) \text{ est un composant terminal}) = 1$$

**Algorithme 1 :** Calcul des composants terminaux maximaux

```
ComposantsTermMax(M)
Input : M, un MDP
Output : SM, l'ensemble des composants terminaux maximaux de M
Data : i entier, s, s' états, a action, sub, sub' sous-MDP, stack, une pile de sous-MDP
sub ← {(s,a) | s ∈ S, a ∈ A_s} ; Empiler(stack, sub) ; SM ← ∅
while not Vide(stack) do
    sub ← Depiler(stack) ; S' ← {s | ∃(s,a) ∈ sub}
    for (s,a) ∈ sub do
        for s' ∈ S do
            if p(s'|s,a) > 0 and s' ∉ S' then sub ← sub \ {(s,a)}
        end
    end
    if sub ≠ ∅ then
        Calculer les composants fortement connexes de G_sub, S_1, ..., S_K
        if K > 1 then
            for i from 1 to K do sub' ← {(s,a) ∈ sub | s ∈ S_i} ; Empiler(stack, sub')
        else SM ← SM ∪ sub
    end
end
return SM
```

**Question 5.** Prouver que l'algorithme 1 renvoie l'ensemble des composants terminaux maximaux.

**Question 6.** Analyser la complexité (dans le pire des cas) de l'algorithme 1 par rapport à $|S|$ et $|A|$.

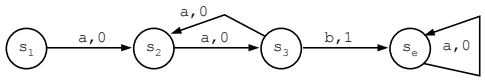
# 3 Minimisation du coût d'accessibilité

Soit $\mathcal{M}$ un MDP avec *des récompenses positives ou nulles* et $s_e$, un *état absorbant* : $A_{s_e}$ est un singleton dont la distribution de Dirac conduit à $s_e$ et dont la récompense est nulle. Nous faisons l'hypothèse qu'il existe des politiques qui garantissent d'atteindre $s_e$ avec probabilité 1 et ces politiques sont appelées des *politiques gagnantes*. Dans ce cas, il existe une politique gagnante stationnaire deterministe.
Le coût d'accessibilité d'une politique $\boldsymbol{\pi}$ (qui peut être infini) est défini par :

$$R^{\boldsymbol{\pi}} \stackrel{\text{def}}{=} \sum_{i \in \mathbb{N}} \mathbf{E}^{\boldsymbol{\pi}}(r(X_i, Y_i))$$

Le coût vectoriel correspondant (qui dépend de l'état initial) est noté $\mathbf{R}^{\boldsymbol{\pi}}$. Le coût vectoriel optimal $\mathbf{R}^*$ est défini par : pour tout $s \in S$, $\mathbf{R}^*[s] \stackrel{\text{def}}{=} \inf_{\boldsymbol{\pi}}(\mathbf{R}^{\boldsymbol{\pi}}[s] \mid \boldsymbol{\pi} \text{ est gagnante})$. Le problème du coût d'accesibilité consiste à trouver le coût d'accessibilité minimal $\mathbf{R}^*$ et une politique gagnante optimale.

**Question 7.** En vous appuyant sur le MDP représenté ci-dessous (avec uniquement des distributions de Dirac) montrer qu'une politique non gagnante peut avoir un plus petit coût d'accessibilité que toute politique gagnante.




Dans la suite, nous supposons que pour toute politique non gagnante $\boldsymbol{\pi}$ il existe $s \in S$ tel que : $\mathbf{R}^{\boldsymbol{\pi}}[s] = \infty$.

L'opérateur $L$ sur $Rew \stackrel{\text{def}}{=} \{\mathbf{v} \in \mathbb{R}^S \mid \mathbf{v}[s_e] = 0 \wedge \forall s \in S\ \mathbf{v}[s] \geq 0\}$ est défini par :

$$\forall s \in S\ L(\mathbf{v})[s] = \min_{a \in A_s} \left( r(s,a) + \sum_{s' \in S} p(s'|s,a)\mathbf{v}[s'] \right)$$

**Question 8.** Soit $\mathbf{v} \in Rew$ un point fixe de $L$. Prouver que $\mathbf{v} \leq \mathbf{R}^*$.

**Question 9.** Soit $d^\infty$ une politique stationnaire. Montrer que $\mathbf{R}^{d^\infty} = \sum_{i \in \mathbb{N}} (\mathbf{P}_d)^i \mathbf{r}_d$ (avec les notations des notes de cours).

Soit $d^\infty$ une politique gagnante. Puisque $d^\infty$ est stationnaire, on peut ordonner $S = \{s_1, \ldots, s_n\}$ de telle façon que $s_1 = s_e$ et pour tout $s_i$ avec $i > 1$ il existe $\alpha_i < i$ tel que $\mathbf{P}_d[i, \alpha_i] > 0$. Soit $p = \min(\min(\mathbf{P}_d[i, \alpha_i] \mid i > 1), \frac{1}{2})$. Définissons $\mathbf{v} \in Rew$ par $\mathbf{v}[s_i] = 1 - p^{2i}$ pour $i > 1$.

**Question 10.** Montrer que $\mathbf{P}_d \mathbf{v} \leq \gamma \mathbf{v}$ avec $\gamma \stackrel{\text{def}}{=} \frac{1-p^{2n-1}}{1-p^{2n}}$. En déduire que $\mathbf{R}^{d^\infty}$ est fini.

Soit $d$ une règle de décision, l'opérateur $L_d$ sur $Rew$ est défini par :

$$L_d(\mathbf{v}) = \mathbf{r}_d + \mathbf{P}_d \mathbf{v}$$

**Question 11.** Soit $d^\infty$ une politique gagnante. Montrer que $\mathbf{R}^{d^\infty}$ est un point fixe de $L_d$.

**Question 12.** Soit $d$ une règle de décision telle qu'il existe $\mathbf{v} \in Rew$ avec $L_d(\mathbf{v}) \leq \mathbf{v}$.
Montrer que $d^\infty$ est une politique gagnante.
*Indication : utiliser l'hypothèse sur les politiques non gagnantes.*

**Question 13.** Soit $d^\infty$ une politique gagnante déterministe stationnaire telle que $L(\mathbf{R}^{d^\infty}) \lneq \mathbf{R}^{d^\infty}$. Soit $d'$ une règle de décision déterministe telle que $L(\mathbf{R}^{d^\infty}) = L_{d'}(\mathbf{R}^{d^\infty})$.
Montrer que $\mathbf{R}^{d'^\infty} \lneq \mathbf{R}^{d^\infty}$.

**Question 14.** Deduire des questions précédentes qu'il existe une politique gagnante déterministe stationnaire $d^\infty$ telle que $L(\mathbf{R}^{d^\infty}) = \mathbf{R}^{d^\infty}$ et que $d^\infty$ est une politique optimale pour le problème du coût d'accessibilité.

**Question 15.** Proposer un programme linéaire telle que sa solution soit $\mathbf{R}^*$.