

CORRECTION DES TRAVAUX DIRIGÉS D'
APPRENTISSAGE PAR RENFORCEMENT
UNIVERSITÉ PARIS–SACLAY

Joon Kwon

mercredi 5 novembre 2024



EXERCICE 1 (*Labyrinthe*). —

- 1) On modélise le problème par :
- un ensemble d'états $\mathcal{S} = \{1, \dots, n\} \times \{1, \dots, n\}$ qui correspond aux cellules,
 - un ensemble d'actions

$$\mathcal{A} = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right\},$$

qui correspondent respectivement à "bas", "droite", "haut", "gauche",

- un ensemble de paiement $\mathcal{R} = \{0, 1\}$.

Soit également l'ensemble d'états appelés "murs intérieurs" $\mathcal{W} \subset \mathcal{A}$.

Les transitions sont toutes déterministes, et sont telles qu'on obtient un gain de 1 lorsqu'on se déplace vers la cellule d'arrivée depuis une cellule voisine, et un gain de 0 sinon. Si l'action choisie correspond à un déplacement non-autorisé (soit qu'il mène à l'extérieur du labyrinthe, soit qu'il mène vers un mur intérieur, soit que l'état actuel est un mur intérieur), l'état reste inchangé.

$$p(\cdot | s, a) = \begin{cases} \delta_{0, s+a} & \text{si } s \notin \mathcal{W} \text{ et } s+a \in \mathcal{S} \setminus (\{(n, n)\} \cup \mathcal{W}) \\ \delta_{0, s} & \text{si } s \in \mathcal{W} \text{ ou } s+a \notin \mathcal{S} \setminus \mathcal{W} \\ \delta_{(1, s+a)} & \text{si } s \notin \mathcal{W} \text{ et } s+a = (n, n). \end{cases}$$

- 2) Une politique optimale est la fonction valeur associée peuvent être représentés comme suit, où les murs intérieurs sont représentés en gris.

↓	↓	→	→	↓
↓	←	←	↓	↓
↓	←	↓	↓	←
↓	↓	→	→	↓
→	→	→	→	→

$\frac{1}{128}$	0	$\frac{1}{128}$	$\frac{1}{64}$	$\frac{1}{32}$
$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{256}$	0	$\frac{1}{16}$
$\frac{1}{32}$	$\frac{1}{64}$	0	$\frac{1}{4}$	$\frac{1}{8}$
$\frac{1}{16}$	0	$\frac{1}{4}$	$\frac{1}{2}$	0
$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1	0

EXERCICE 2. —

- 1) a) La distribution de l'état initial est $\mu = \delta_{(*,*,*,*,*)} \otimes \mathcal{U}(\{0, \dots, 9\})$.
b) $\alpha(s)$ correspond au nombre d'emplacements disponibles, $v(s)$ à l'entier formé (lorsqu'il n'y a plus d'emplacements disponibles), et $\sigma(s, a)$ correspond aux cinq premières composantes de l'état suivant lorsque dans l'état s on a choisi l'action a .
c) Une définition possible pour la transition est

$$p(\cdot | s, a) = \begin{cases} \delta_0 \otimes \delta_{\sigma(s,a)} \otimes \mathcal{U}(\{1, \dots, 9\}) & \text{si } s^{(a)} = * \text{ et } \alpha(s) > 1, \\ \delta_{v(\sigma(s,a))} \otimes \delta_{\sigma(s,a)} \otimes \delta_0 & \text{si } s^{(a)} = * \text{ et } \alpha(s) = 1, \\ \delta_0 \otimes \delta_s & \text{si } s^{(a)} \neq *. \end{cases}$$

- 2) a) La distribution de l'état initial est $\mu = \delta_{(1,1,1,1,1)} \otimes \mathcal{U}(\{1, \dots, 9\})$.
b) Une définition possible pour la transition est

$$p(\cdot | s, a) = \begin{cases} \delta_0 \otimes \delta_s & \text{si } s^{(a)} = 0 \\ \delta_{10^{a-1}s^{(6)}} \otimes \delta_{\sigma(s,a)} \otimes \mathcal{U}(\{1, \dots, 9\}) & \text{si } s^{(a)} = 1. \end{cases}$$

où $\sigma(s, a) \in \{0, 1\}^5$ est défini par

$$\sigma(s, a)_i = \begin{cases} 0 & \text{si } i = a \\ s_i & \text{sinon.} \end{cases}$$

- 3) a) Pour tout $s \in \mathcal{S}$,

$$\pi(\cdot | s) = \mathcal{U}(\{1, \dots, 5\}).$$

b) On peut montrer que

$$\mathbb{E}_{\mu, \pi} \left[\sum_{t=1}^{+\infty} R_t \right] = 49999.5.$$

c) Voir correction du TP.

4) Voir correction du TP.

EXERCICE 3 (Différence de performance). —

1) $d_{s, \pi}$ a claieant des composantes positives. De plus,

$$\begin{aligned} \sum_{s' \in \mathcal{S}} d_{s, \pi}(s') &= \sum_{s' \in \mathcal{S}} (1 - \gamma) \sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{s, \pi} [S_t = s'] \\ &= (1 - \gamma) \sum_{t=0}^{+\infty} \gamma^t \sum_{s' \in \mathcal{S}} \mathbb{P}_{s, \pi} [S_t = s'] \\ &= (1 - \gamma) \sum_{s' \in \mathcal{S}} \gamma^t = 1. \end{aligned}$$

Donc $d_{s, \pi} \in \Delta(\mathcal{S})$.

2) Soit $S'_0, A'_0, R'_1, \dots \sim \mathbb{P}_{s, \pi}$. On peut d'abord prouver¹ que pour $t \geq 1$,

$$\mathbb{E} [q_{\pi'}(S'_t, A'_t)] = \mathbb{E} [R'_{t+1} + \gamma v_{\pi'}(S'_{t+1})].$$

Donc,

$$\begin{aligned} v_{\pi}(s') - v_{\pi'}(s) &= \mathbb{E} \left[\sum_{t=0}^{+\infty} \gamma^t R'_{t+1} \right] - v_{\pi'}(s) \\ &= \mathbb{E} \left[\sum_{t=0}^{+\infty} \gamma^t (R'_{t+1} + v_{\pi'}(S'_t) - v_{\pi'}(S'_t)) - v_{\pi'}(S'_0) \right] \\ &= \mathbb{E} \left[\sum_{t=0}^{+\infty} \gamma^t (R'_{t+1} + \gamma v_{\pi'}(S'_{t+1}) - v_{\pi'}(S'_t)) \right] \\ &= \mathbb{E} \left[\sum_{t=0}^{+\infty} \gamma^t (q_{\pi'}(S'_t, A'_t) - v_{\pi'}(S'_t)) \right] \\ &= \mathbb{E} \left[\sum_{t=0}^{+\infty} \gamma^t \alpha_{\pi'}(S'_t, A'_t) \right]. \end{aligned}$$

1. à détailler

Par ailleurs, on peut facilement vérifier que π étant stationnaire, $A'_t | S'_t \sim \pi(S'_t)$ et donc pour $t \geq 0$,

$$\mathbb{E} [\alpha_{\pi'}(S'_t, A'_t)] = \mathbb{E} [\mathbb{E} [\alpha_{\pi'}(S'_t, A'_t) | S'_t]] = \mathbb{E} [\mathbb{E}_{A \sim \pi(S'_t)} [\alpha_{\pi'}(S'_t, A'_t)]] .$$

Par conséquent,

$$\begin{aligned} v_{\pi}(s') - v_{\pi'}(s) &= \sum_{t=0}^{+\infty} \gamma^t \mathbb{E} [\mathbb{E}_{A \sim \pi(S'_t)} [\alpha_{\pi'}(S'_t, A)]] \\ &= \sum_{t=0}^{+\infty} \gamma^t \mathbb{E} \left[\sum_{s' \in \mathcal{S}} \mathbb{P}[S'_t = s'] \cdot \mathbb{E}_{A \sim \pi(s')} [\alpha_{\pi'}(s', A)] \right] \\ &= \sum_{s' \in \mathcal{S}} \sum_{t=0}^{+\infty} \gamma^t \mathbb{P}[S'_t = s'] \cdot \mathbb{E}_{A \sim \pi(s')} [\alpha_{\pi'}(s', A)] \\ &= \sum_{s' \in \mathcal{S}} \frac{1}{1-\gamma} d_{s,\pi}(s') \cdot \mathbb{E}_{A \sim \pi(s')} [\alpha_{\pi'}(s', A)] \\ &= \frac{1}{1-\gamma} \mathbb{E} [\mathbb{E}_{A \sim \pi(S)} [\alpha_{\pi'}(S, A)]] , \end{aligned}$$

où la dernière égalité découle du fait que $S \sim d_{s,\pi}$ par hypothèse.

