Travaux dirigés d'
Apprentissage par renforcement

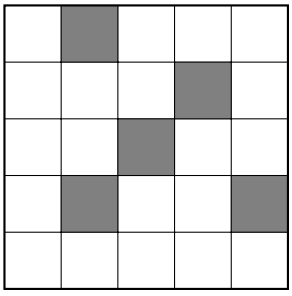
Université Paris–Saclay

Joon Kwon

mercredi 5 novembre 2024

Exercice 1 (*Labyrinthe*). — Soit $n \geqslant 1$. On considère un labyrinthe carré de $n \times n$ cases. Chaque case est désignée par un couple $(i, j) \in \{1, \dots, n\}^2$ où i correspond à la ligne et j à la colonne. Les cases de départ et d'arrivée sont respectivement $(1, 1)$ et (n, n). Certaines cases sont murées et ne peuvent pas être accédées : on note $\mathscr{W}$ l'ensemble des cases murées. Le but du problème est de déterminer le chemin le plus court du départ à l'arrivée.

1) Modéliser le problème par un MDP.



2) Pour l'exemple donné en figure, et avec un taux d'escompte $\gamma = 1/2$, donner sans preuve une politique optimale π_* et les valeurs de la fonction état-valeur associée v_{π_*}.

Exercice 2. — On considère un problème où le but de l'agent est de former un nombre à 5 chiffres, le plus grand possible. On considère 5 emplacements, initialement vides, correspondant aux chiffre des unités, des dizaines, des centaines, des milliers, et des dizaines de milliers. À chaque étape, un chiffre (entre 0 et 9) est tiré uniformément et présenté à l'agent. Celui-ci doit le placer dans l'un des emplacements disponibles. Le nombre est donc formé au bout de 5 étapes.

1) On cherche à modéliser le problème par un MDP dont
   — l'ensemble d'états est $\mathcal{S} = \{*, 0, \dots, 9\}^5 \times \{0, \dots, 9\}$, où les cinq premières composantes de l'état correspondent au nombre partiellement formé ($*$ représentant un emplacement libre), et la sixième composante le chiffre qui vient d'être tiré et qui est à placer,
   — l'ensemble de paiements est $\mathcal{R} = \{0, 1, \dots, 99999\}$,
   — l'ensemble d'actions est $\mathcal{A} = \{1, \dots, 5\}$, où 1 correspond au choix de l'emplacement des unités, 2 à celui des dizaines, etc.

   Pour un état $s \in \mathcal{S}$, on note $s = (s^{(1)}, s^{(2)}, \dots, s^{(6)})$ ses composantes.
   a) Quelle est la distribution de l'état initial ?
   b) Donner une interprétation pour chacune des fonctions suivantes.

$$\alpha(s) = \operatorname{Card}\left\{i \in \{1, \dots, 5\},\ s^{(i)} = *\right\}, \quad s \in \mathcal{S},$$

$$\nu(s) = \sum_{i=1}^{5} 10^{i-1} s^{(i)}, \quad \text{pour } s \in \mathcal{S} \text{ tel que } \alpha(s) = 0,$$

$$\sigma(s, a) = (\mathbb{1}_{\{i \neq a\}} s^{(i)} + \mathbb{1}_{\{i=a\}} s^{(6)})_{1 \leqslant i \leqslant 5}, \quad (s, a) \in \mathcal{S} \times \mathcal{A}.$$

   c) Pour tous $s \in \mathcal{S}$ et $a \in \mathcal{A}$ donnés, exprimer la distribution de transition $p(\cdot\,|s, a)$, qui est un élément de $\Delta(\mathcal{R} \times \mathcal{S})$. On pourra faire intervenir les fonctions α, ν et σ.

2) La modélisation de la question précédente a l'inconvénient de comporter un très grand nombre d'états. On souhaite dorénavant modéliser le problème par un MDP dont
   — l'espace d'états est $\mathcal{S} = \{0, 1\}^5 \times \{0, \dots, 9\}$ où pour les cinq premières composantes, un 1 correspond à un emplacement libre,
   — l'ensemble de paiements est $\mathcal{R} = \{0, 1, \dots, 99999\}$,
   — l'ensemble d'actions est $\mathcal{A} = \{1, \dots, 5\}$,

   et dont le paiement n'est pas escompté ($\gamma = 1$).
   a) Quelle est la distribution de l'état initial (qu'on notera μ) ?

b) Exprimer la fonction de transition p.

3) On considère la politique π qui à chaque étape choisit uniformément un des emplacements disponibles.
   a) Écrire formellement π.
   b) Calculer théoriquement

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

   c) Estimer empiriquement cette quantité (*voir notebook*).

4) Proposer une meilleure politique π' que celle de la question précédente, l'implémenter, et l'évaluer empiriquement (*voir notebook*).

Exercice 3 (*Différence de performance*). — Soit $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ un MDP fini et $0 < \gamma < 1$ le taux d'escompte. Soit π, π' deux politiques stationnaires. Pour $s \in \mathcal{S}$, on définit $d_{s,\pi} \in \mathbb{R}^{\mathcal{S}}$ comme suit :

$$d_{s,\pi}(s') = (1-\gamma)\sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{s,\pi}\left[\mathrm{S}_t = s'\right], \quad s' \in \mathcal{S}.$$

1) Montrer que $d_{s,\pi} \in \Delta(\mathcal{S})$.

2) Soit une variable aléatoire $\mathrm{S} \sim d_{s,\pi}$ et on définit $\alpha_\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ par

$$\alpha_\pi(s', a) = q_\pi(s', a) - v_\pi(s'), \quad (s', a) \in \mathcal{S} \times \mathcal{A}.$$

   Montrer que :

$$v_\pi(s) - v_{\pi'}(s) = \frac{1}{1-\gamma}\mathbb{E}\left[\mathbb{E}_{\mathrm{A}\sim\pi(\mathrm{S})}\left[\alpha_{\pi'}(\mathrm{S}, \mathrm{A})\right]\right].$$