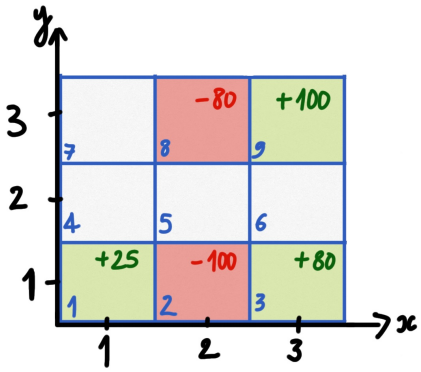
Introduction à l'apprentissage par renforcement

Master IASD

Année 2020–2021

1 Q-Learning

On considère le monde-grille ci-dessous et un agent essaie d'apprendre la politique optimale. Les états sont numérotés de 1 à 9 (dans le coin en bas à gauche de chaque case). L'état initial est la case 7. Toutes les cases grisées ou colorées correspondent à des états terminaux. Les autres états ont quatre actions possibles (aller au Nord, Sud, Est ou Ouest). Le résultat est *déterministe* (l'agent se déplace dans la case voisine suivant la direction donnée, ou il reste à sa place si l'action tentait de l'amener en dehors de la grille). Si le résultat de l'action est de tomber dans un état grisé/colorié, il obtient la récompense notée en haut à droite dans cet état pour la transition. On suppose que le taux d'escompte est $\gamma = 0.5$ et que le taux d'apprentissage est de $\alpha = 0.5$.



Question 1. *Rappeler la définition de la fonction valeur optimale ainsi que la forme de l'équation de Bellman.*

Question 2. *Quelle est la valeur de la fonction de valeur optimale pour les états suivants : $v^\star(6)$, $v^\star(5)$, $v^\star(7)$ (montrez bien comment vous obtenez le résultat) ?*

On donne trois épisodes complets. Chaque ligne dans un épisode est un tuple (s, a, s', r) où l'agent a exécuté l'action a dans l'état s, le conduisant à l'état s' et obtenant la récompense r.

Episode 1	Episode 2	Episode 3
7, S, 4, 0	7, S, 4, 0	7, S, 4, 0
4, E, 5, 0	4, E, 5, 0	4, E, 5, 0
5, E, 6, 0	5, E, 6, 0	5, S, 2, -100
6, S, 3, 80	6, N, 9, 100	

Question 3. *En utilisant les mises à jour de l'algorithme* Q*-learning, quelles sont les valeurs de $Q(6, N)$, $Q(4, S)$ et $Q(5, E)$ après les trois épisodes ci-dessus? On supposera que toutes les valeurs sont initialisées à 0, et que l'on fait les mises à jour dans l'ordre.*

Question 4. *Quelles valeurs aurait on obtenu si on avait utilisé l'algorithme* SARSA *?*

2 Bandit bayésien avec des valeurs discrètes

On s'intéresse à un modèle de bandit dans lequel les espérances inconnues des distributions correspondant aux K bras $\mu_1, \dots, \mu_k$ appartiennent à une grille fixée de d valeurs connues $\{v_1, \dots, v_d\}$, où $0 < v_1 < \cdots < v_d < 1$. On notera, comme dans le cours,

- $A_t \in \{1, \dots, K\}$ l'action effectuée au temps t;
- $X_t \in \{0, 1\}$ la récompense observée au temps t, dont on supposera qu'elle suit une loi de Bernoulli: $\mathbb{P}(X_t = 1 | A_t = k, \mathcal{H}_{t-1}) = \mu_k$ et $\mathbb{P}(X_t = 0 | A_t = k, \mathcal{H}_{t-1}) = 1 - \mu_k$, où $\mathcal{H}_{t-1}$ désigne les récompenses observées jusqu'au temps $t-1$;
- $N_k(t) = \sum_{s=1}^{t} \mathbb{1}\{A_s = k\}$ le nombre de tirages du bras k jusqu'au temps t;
- $\bar{X}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^{t} X_s \mathbb{1}\{A_s = k\}$ la moyenne empirique des récompenses obtenues en tirant le bras k jusqu'au temps t.

On notera T l'horizon temporel qui est ici supposé connu à l'avance.

Question 5. *Définir la stratégie optimale à paramètre connu, ainsi que $\mathbb{E}[R_T]$ l'espérance du regret d'un algorithme séquentiel, exprimée en fonction de $N_1(T), \dots, N_K(T)$ et des paramètres du modèle.*

Question 6. *On rappelle le résultat vu en cours concernant le regret de l'algorithme ETC (Explore-Then-Commit)*

$$\mathbb{E}[R_T] \leq \sum_{\substack{k=1 \\ k\,:\,\Delta_k > 0}}^{K} \Delta_k \left(m + T e^{-m\Delta_k^2} \right).$$

où Δ_k représente l'écart $\max\{\mu_1, \dots, \mu_K\} - \mu_k$. Rappeler à quoi correspond le paramètre m. Pourquoi dans le modèle considéré ici est-il possible de régler m de façon à ce que $\mathbb{E}[R_T] \leq \alpha + \beta \log(T)$, avec α et β à préciser ?

On s'intéresse maintenant à la version bayésienne du problème dans laquelle on considère, l'ensemble des configurations possibles de $(\mu_1, \dots, \mu_k)$ munies de la loi a priori suivante :

- $\mu_1, \dots, \mu_k$ sont conjointement indépendants;
- $\mathbb{P}(\mu_k = v_i) = 1/d$, pour tous $1 \leq k \leq K$ et $1 \leq i \leq d$.

Question 7. *Montrer que le regret Bayésien de l'algorithme ETC, réglé comme à la question 6, est toujours bornée par $\alpha + \beta \log(T)$.*

On s'intéresse désormais à la façon de mettre en oeuvre l'algorithme de Thompson sampling dans ce modèle. On note comme dans le cours, pour $k \in \{1, \dots, K\}$, $(X_{k,m})_{m \geq 1}$ la séquence de tirages indépendants et identiquement distribués associée au bras k et $\bar{X}_{k,n} = 1/n \sum_{m=1}^{n} X_{k,m}$.

Question 8. *Montrer que*

$$\mathbb{P}(u_k = v_i | X_{k,1}, \dots X_{k,n}) = \frac{v_i^{n\bar{X}_{k,n}} (1 - v_i)^{n(1-\bar{X}_{k,n})}}{\sum_{j=1}^{d} v_j^{n\bar{X}_{k,n}} (1 - v_j)^{n(1-\bar{X}_{k,n})}}.$$

Question 9. *Donner l'expression de $\mathbb{E}(X_{k,n+1} | X_{k,1}, \dots X_{k,n})$.*

Question 10. *Utiliser ce qui précède pour spécifier l'algorithme de bandits qui cherche à maximiser, à chaque étape, $\mathbb{E}[X_t | \mathcal{H}_{t-1}]$.*

Question 11. *Décrire l'algorithme de Thompson sampling correspondant à ce modèle.*

3 Regret pour un bandit à un bras inconnu

On conserve les notations de l'exercice précédent en considérant cette fois un modèle à $K = 2$ bras où

- le bras 2 a une moyenne $\mu_2 = \alpha$, où $0 < \alpha < 1$ est une valeur connue;
- $\mu_1 = \mu$, où $\mu \in]0, 1[$ est un paramètre réel inconnu.

On note $\Delta = |\alpha - \mu|$.

Question 12. *Donner les expressions de l'espérance du regret d'un algorithme en fonction de $N_1(T)$, T et Δ dans les deux cas $\mu > \alpha$ et $\mu < \alpha$.*

On considère, l'algorithme suivant.

Algorithme 1.

Initialization: $M = 1$, $A_1 = 1$, observe X_1 and set $\bar{X}_1(1) = X_1$
while $\bar{X}_M(t) > \alpha - \sqrt{\frac{\log T}{2M}}$ **do**
  $M \leftarrow M + 1$
  Set $\mathcal{A}_M = 1$
  Observe X_M
  Compute $\bar{X}_1(M) = \frac{X_M + (M-1)\bar{X}_1(M-1)}{M}$
end
for $t = M + 1, \dots T$ **do**
  Play $\mathcal{A}_t = 2$
end

Question 13. *L'algorithme 1 peut être vu comme une variante adaptative de l'algorithme ETC. Pourquoi n'est-il pas nécessaire de tirer le bras 2 pendant la phase d'exploration ? Pourquoi, pour $t \leq M$, peut on écrire $\bar{X}_1(t) = \bar{X}_{1,t}$? De même, vérifier que $N_1(T) = M$.*

On rappelle que l'inégalité de Hoeffding indique que pour tout $s > 0$

$$\mathbb{P}(\bar{X}_{1,n} < \mu - s) \leq e^{-2ns^2} \quad \text{et} \quad \mathbb{P}(\bar{X}_{1,n} > \mu + s) \leq e^{-2ns^2}$$

Question 14. *On suppose que $\mu > \alpha$, montrer que*

$$\mathbb{P}\left(\bar{X}_{1,n} < \alpha - \sqrt{\frac{\log T}{2n}}\right) \leq \frac{1}{T} e^{-2n\Delta^2}.$$

En déduire une majoration de $\mathbb{P}(M \leq T)$, puis que l'espérance du regret vérifie

$$\mathbb{E}[R_T] \leq \frac{\Delta}{1 - e^{-2\Delta^2}}.$$

Question 15. *On suppose maintenant que $\mu < \alpha$, montrer que*

$$\mathbb{P}\left(\bar{X}_{1,n} > \alpha - \sqrt{\frac{\log T}{2n}}\right) \leq e^{-n\Delta^2/2}$$

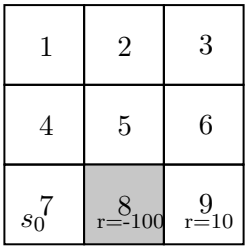
lorsque $n > \frac{2\log T}{\Delta^2}$. En déduire que l'espérance du regret est majorée par

$$\mathbb{E}[R_T] \leq \frac{2\log T}{\Delta} + \frac{\Delta}{1 - e^{-\Delta^2/2}}.$$

indication : On pourra utiliser la formule $\mathbb{E}[M] = \sum_{n \geq 0} \mathbb{P}(M > n)$.

4 Apprentissage de politique paramétrée

On se donne un processus de décision markovien défini sur l'espace d'états $\mathcal{S} = \{1, \ldots, 9\}$



- Où l'état numéro 7 représente l'état initial et les états 8 et 9 sont des états terminaux (sans action associée).
- Les actions $\mathcal{A}$ sont $\{N, E, S, O\}$ (alternativement notées $1, 2, 3, 4$) dans chaque état et correspondent à des déplacements d'une case (les diagonales sont interdites). Le résultat des actions est déterministe, sauf pour la case numéro 5 où, avec une probabilité 0.25, l'action choisie conduit dans la case 8 (imaginez qu'il s'agit du bord d'une falaise).
- Se déplacer vers la case 8 conduit à une récompense de -100, un déplacement vers la case 9 conduit à une récompense de +10. Tous les autres déplacements ont une récompense de -1 (essayer de passer au travers d'un bord comme aller au Nord depuis la case 2 a pour conséquence de rester sur la même case et de recevoir une récompense de -1).

On utilise une politique paramétrique $\pi_{\boldsymbol{\theta}}$ possédant un seul paramètre par couple état action (s,a). Le vecteur des paramètres est noté $\boldsymbol{\theta} \in \mathbb{R}^{28}$ et ses coordonnées θ_k. Pour tout k, $1 \leq k \leq 28$, on note μ_k la quantité :

$$\mu_k = \frac{1}{1 + \exp(-\theta_k)}. \tag{1}$$

Dans un état non terminal $i \in \{1, \ldots, 7\}$, la probabilité $\pi_{\boldsymbol{\theta}}(a = j | s = i)$ que la politique $\pi_{\boldsymbol{\theta}}$ choisisse l'action $j \in \{1, 2, 3, 4\}$ est égale à $\mu_{4(i-1)+j}$.

Question 16. *Donner intuitivement le comportement attendu de la politique optimale quand le facteur d'escompte vaut $\gamma = 1$.*

Question 17. *Quel problème peut apparaître dans le cas général si $\gamma = 1$? Peut-il se produire avec la politique $\pi_{\boldsymbol{\theta}}$ compte tenu de la paramétrisation choisie ?*

Question 18. *Montrer que*

$$\frac{\partial \ln \pi_{\boldsymbol{\theta}}(a = j | s = i)}{\partial \theta_k} = \frac{1}{1 + \exp(\theta_k)} \mathbb{1}\{k = 4(i-1) + j\}.$$

Question 19. *Un episode produit la sequence d'état, actions, recompenses suivante:*
$s_0 = 7, a = 1(N), r = -1$; $s_1 = 4, a = 2(E), r = -1$; $s_2 = 5, a = 2(E), r = -1$; $s_3 = 6, a = 2(E), r = -1$; $s_4 = 6, a = 3(S), r = 10$; $s_5 = 9$ *(FIN)*
Partant de $\boldsymbol{\theta} = \mathbf{0}$, donner les nouvelles valeurs de θ_k après application de l'algorithme REINFORCE *pour les valeurs de k correspondant aux états 3, 5 et 6 (pour $\gamma = 1$).*

Question 20. *Comment l'algorithme peut-il être modifié pour réduire la variance lors de l'apprentissage ?*