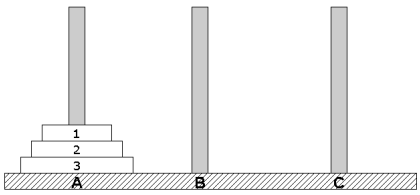
Examen RLD

Durée : 1 heure 30

(Documents autorisés)

Exercice 1 – Tours de Hanoï

On considère le jeu de la tour d'hanoï (il n'y a qu'un seul joueur) dont la disposition initiale est la suivante :



Le but du jeu est de déplacer une tour formée de disques de différentes tailles d'une position A à la position C en déplaçant les disques les uns après les autres sans que jamais un disque plus grand puisse être placé au-dessus d'un disque plus petit. Seulement trois positions sont disponibles : la position de départ A, une position intermédiaire B et la position d'arrivée C. Pour simplifier, on ne considère dans cet exercice que deux disques : un petit disque 1 et un grand disque 2.

Q 1.1 Dessiner le graphe correspondant au MDP de ce jeu (états, transitions, état terminal)

Q 1.2 Dans le cadre des algorithmes de plannification tels que Policy ou Value Iteration pour la découverte de politique optimale, donner la fonction objectif visée. Discuter de l'intérêt du facteur de discount γ. Que risque-t-il de se passer si $\gamma = 1$? **Q 1.3** Dans l'optique d'utiliser ces algorithmes de plannification, définir les récompenses associer aux transitions du

MDP malgré l'utilisation d'un facteur de discount $\gamma = 1$.

Q 1.4 Soit la fonction $V^\pi(s)$ l'espérance de récompenses discountées à partir de l'état s en suivant la politique π. Donner une expression récursive de $V^\pi(s)$ dans notre cas déterministe des tours de Hanoï.

Q 1.5 Soit π une politique aléatoire qui choisit uniformément les actions en chaque état. Appliquer deux itérations de l'algorithme d'évaluation de cette politique (suivant la relation précédente avec $\gamma = 1$ et $V_0 = 1$ pour tous les états).

Q 1.6 Considérons maintenant une version modifiée du jeu où les disques ne sont pas posés mais jetés : un jet d'un disque entre deux positions consécutives réussit à tous les coups, par contre un jet de la position A vers la position C (resp. de C vers A) échoue dans 50% des cas. En cas d'échec de l'envoi, la partie est perdue. Donner les modifications sur le MDP du jeu (états et transitions).

Q 1.7 Dans la nouvelle configuration du jeu, donner deux schémas de récompense menant vers des politiques optimales différentes, l'une prudente, l'autre plus risquée (mais ne menant pas vers un comportement suicidaire), avec $\gamma = 1$ (en précisant lesquelles). Même chose avec $\gamma = 0.1$.

Q 1.8 On considère maintenant le cas où l'on ne connaît pas le MDP. Proposer une méthode efficace pour l'estimation des valeurs V^π des différents états dans ce cas. Discuter des différentes versions de cet algorithme sur la base du compromis biais variance.

Q 1.9 Quel problème se pose pour la recherche de politique optimale lorsqu'on ne connaît pas le MDP ? Proposer une solution, basée sur des fonctions de valeurs, en détaillant l'algorithme.

Q 1.10 Pour mettre en évidence l'importance de l'exploration dans le cas des MDP inconnus, donner un scénario dans lequel l'algorithme précédent dans sa version gloutonne converge vers une politique sous-optimale (en précisant le schéma de récompense considéré).

Q 1.11 Dire ce qu'est une stratégie On-Policy dans ce cadre (avec exemple d'algo d'apprentissage) et quels en sont les avantages et inconvénients.

Exercice 2 – Policy Gradients

Les méthodes Policy Gradients s'intérèssent à la maximisation de la probabilité des trajectoires rapportant un fort reward cumulé selon la politique π_θ :

$$\theta^* = \arg\max_\theta J(\theta) = \arg\max_\theta \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\mathcal{R}(s_t, a_t, s_{t+1})\right]$$

Q 2.1 Pour réaliser cette maximisation, on s'intéresse au gradient :

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t=0}^{T} r_t\right]$$

D'où vient le log donné dans ce gradient ? Mentionnez trois grands avantages de cette écriture.

Q 2.2 De quelle partie de ce gradient pourrions-nous nous passer pour réduire la variance ?

Q 2.3 Dire ce qu'est une baseline dans ce cadre, à quoi celà sert, quelles en sont les propiétés requises, en donner un exemple utilisé par les approches Actor-Critic.

Q 2.4 Quelle est la principale différence entre les approaches Actor-Critic et la méthode Reinforce ? Quels en sont les avantages et inconvénients ?

Q 2.5 Dire ce qu'est l'Importance Sampling et quoi cela peut servir dans le cadre des approches Policy Gradients.

Q 2.6 Une manière de réduire la variance des techniques d'Importance Sampling est de considérer une normalisation $Z = \sum_i w_i$ avec w_i les poids d'importance Sampling, plutôt que le plus classique $Z = N$ (avec N le nombre d'exemples de l'estimateur). Montrer que dans ce cas le gradient $\nabla_\theta J_{is}(\theta)$ peut s'écrire :

$$\nabla_\theta J_{is}(\theta) = \frac{1}{Z}\sum_{\tau^{(i)}}\nabla_\theta w_i\left[R(\tau^{(i)}) - J_{is}(\theta)\right]$$

Q 2.7 Ok mais toujours pas mal de variance dans de nombreux problèmes. Une autre famille de méthodes considère :

$$J(\theta) = \mathbb{E}_{s\sim d^\beta}\big[\sum_{a\in\mathcal{A}} Q^\pi(s,a)\pi_\theta(a|s)\big]$$

avec $d^\beta(s) = \lim_{t\to\infty} P(S_t = s|\beta)$ la distribution stationnaire des états selon une politique de "behavior" β. Montrer qu'en prenant $\beta = \pi$, on peut obtenir un gradient non biaisé, à la manière des méthodes Actor-Critic classiques. S'appuyer sur ce résultat pour discuter de l'approximation réalisée par la plupart des méthodes Off-Policy Policy Gradient (type Off-PAC)

Q 2.8 Donner une déclinaison déterministe des méthodes Policy Gradients lorsque l'espace d'actions et l'espace d'états sont continus (en détaillant).

Exercice 3 – GANs Séquentiels Discrets

Q 3.1 Quelle est la difficulté de l'application des GANs pour des générateurs à sorties discrètes, telles que par exemple du texte ? À quoi doit-on avoir recours dans ce cas ?

Q 3.2 Quel intérêt des GANs dans ce cadre par rapport à une simple maximisation de vraisemblance classique ?

Q 3.3 Soit un ensemble de noeuds d'un réseau social, se transmettant des informations selon un graphe inconnu. Soit un ensemble d'épisodes de diffusion Γ d'entraînement, contenant pour chacun un vecteur de caractéristiques du contenu se diffusant dans l'épisode, ainsi que les identifiants des noeuds atteints par ce contenu et leur date de participation à l'épisode. On cherche à obtenir un générateur de séquences maximisant la vraisemblance des episodes réels (i.e., des séquences temporelles), conditionnés par le contenu diffusé. Formaliser le problème sous la forme d'un problème de maximisation de vraisemblance.

Q 3.4 Formaliser maintenant le problème sous la forme d'un GAN séquentiel, ne produisant des récompenses qu'en fin de génération, et donner l'algo d'optimisation correspondant.

Q 3.5 Adapter l'algorithme précédent en faisant en sorte d'éviter de donner uniquement des récompenses en fin de séquence, mais tout au long du processus.

Q 3.6 Dans ce dernier cas faut-il mieux travailler en lambda-return ou bien en moyenne sur la séquence ? (justifier)

Q 3.7 On propose maintenant de modéliser un peu plus finement les dépendances des diffusions considérées : plutôt que de générer des séquences, on génère une arborescence, dans laquelle chaque noeud infecté possède un état dépendant de

la branche de laquelle il est issu (chemin menant du premier infecté jusqu'à lui pour le contenu en question). Cet état est ensuite utilisé pour conditionner les futures infections de l'arborescence enracinée en ce noeud. Adapter le modèle GAN à ce cadre, sachant que le corpus d'entraînement reste le même (uniquement des séquences). Ce genre de modélisation aurait-elle pu être réalisée sans utilisation de RL (si oui comment) ?

Exercice 4 – PPO

Soit l'algo PPO (version KL) suivant :

Algorithm 1: Algorithme PPO - version KL adaptatif

Input: Paramètres initiaux θ_0 et ϕ, KL cible δ, pas d'apprentissage α, nombre d'étapes d'optimisation K

$\beta_0 \leftarrow 1$
for $k = 0, 1, 2, ...$ **do**
  Collecte d'un ensemble de trajectoires $\mathcal{D}_k$ selon politique π_{θ_k}
  Calcul des avantages $\hat{A}^{\pi_{\theta_k}}$ pour toutes les transitions de $\mathcal{D}_k$ selon $TD(\lambda)$
  $\theta \leftarrow \theta_k$
  for s *de* 1 *à* K **do**
    $\theta \leftarrow \theta + \alpha\big(\nabla_\theta \mathcal{L}_{\theta_k}(\theta) - \beta_k \nabla_\theta \bar{D}_{KL}(\theta_k|\theta)\big)$
      avec $\nabla_\theta \mathcal{L}_{\theta_k}(\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{(s,a)\in\mathcal{D}_k} \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} \hat{A}^{\pi_{\theta_k}}(s,a)$
      et $\nabla_\theta \bar{D}_{KL}(\theta_k|\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{s\in\mathcal{D}_k} \nabla_\theta D_{KL}(\pi_{\theta_k}(.|s)|\pi_\theta(.|s))$.
  end
  $\theta_{k+1} \leftarrow \theta$
  if $\bar{D}_{KL}(\theta_k|\theta_{k+1}) \geq 1.5\delta$ **then**
    $\beta_{k+1} \leftarrow 2\beta_k$
  end
  if $\bar{D}_{KL}(\theta_k|\theta_{k+1}) \leq \delta/1.5$ **then**
    $\beta_{k+1} \leftarrow 0.5\beta_k$
  end
  Mise à jour de V_ϕ selon $TD(\lambda)$ sur $\mathcal{D}_k$
end

Q 4.1 Pourquoi considère-t-on un terme de KL dans cet algo ? quel est l'objectif ? quelle différence avec un réglage du pas d'apprentissage α ?

Q 4.2 Que vaut le gradient de la KL au premier passage ? À quoi correspond l'algo PPO si on prend $K = 1$?

Q 4.3 Au fait, où est passé l'habituel log-trick des Policy Gradients dans cet algo ?

Q 4.4 Montrer que l'on pourrait écrire la mise à jour de manière équivalente (bien qu'avec plus de variance) :

$$\theta \leftarrow \theta + \alpha\Big(\frac{1}{|\mathcal{D}_k|} \sum_{(s,a)\in\mathcal{D}_k} \nabla_\theta \pi_\theta(a|s)\Big[\frac{\hat{A}^{\pi_{\theta_k}}(s,a)}{\pi_{\theta_k}(a|s)} + \frac{\beta_k}{\pi_\theta(a|s)}\Big]\Big)$$

Q 4.5 Donner le pseudo-code du calcul de $\hat{A}^{\pi_{\theta_k}}(s,a)$ en fonction des transitions de la trajectoire τ et selon les paramètres habituels γ et λ

Q 4.6 Le schéma de retour de l'environnement CartPole est de déclarer terminal (done=True) l'état atteint à l'itération 500. Dire pourquoi cela peut entraîner une certaine instabilité et donner manière de contourner le problème (dont l'une en augmentant la représentation des états)