Preuves Cours 7 RLD

1 Preuves du modèle RLaR

Soit $\tilde{Q}_i^*(s, h_i, \vec{a}) = \hat{R}(s, \vec{a}) + \sum_{\omega_i \in \Omega_i} \hat{P}_i(\omega_i|s, \vec{a}) \max_{a'_i \in \mathcal{A}_i} Q_i^*(h'_i, a'_i)$ avec $h'_i = (h_i, a_i, \omega_i)$

On souhaite montrer qu'on a alors : $Q_i^*(h_i, a_i) = \sum_{s,a_{-i}} \hat{P}_i(a_i|s, h_i)\hat{P}_i(s|h_i)\tilde{Q}_i^*(s, h_i, \vec{a})$, avec $\hat{P}_i(a_{-i}|s, h_i)$ une estimée pour i de la probabilité des actions de tous les agents sauf i et $\hat{P}_i(s|h_i)$ un modèle de transition selon l'historique de l'agent i, ce qui donne lieu à la règle de mise à jour de la phase 2 de RLaR.

Si on connaît les dynamiques du monde, il est possible de définir dans le cadre des Dec-POMDPs :

$$Q_i^*(h_i, a_i) = \sum_{s,a_{-i},h_{-i}} \pi(a_{-i}|h_{-i}) P_i(s, h_{-i}|h_i) \left[R(s, \vec{a}) + \sum_{\omega_i \in \Omega_i} P_i(\omega_i|s, \vec{a}) \max_{b \in \mathcal{A}_i} Q_i^*((h_i, a_i, \omega_i), b) \right]$$

avec ω_i une observation pour l'agent i, $h_i = (\omega_i^0, a_i^1, \omega_i^1, ..., a_i^t, \omega_i^t)$ l'historique des actions-observations de l'agent i au temps t, a_{-i} l'ensemble des observations sauf celle de l'agent i et :

- $\pi(a_{-i}|h_{-i})$ retourne 1 seulement si toutes les actions sauf celle de i correspondent à celles choisies par les agents selon leurs historiques respectifs, i.e. $\pi(a_{-i}|h_{-i}) = 1$ si $\forall j \neq i, a_j = argmax_{a' \in \mathcal{A}_j} Q_j^*(h_j, a')$.
- $P_i(\omega_i|s, \vec{a}) = \sum_{s'} P(s'|s, \vec{a}) O_i(\omega_i|s', \vec{a})$, avec $O_i(\omega_i|s', \vec{a})$ la probabilité d'observer ω_i dans l'état s' si les actions précédentes étaient $\vec{a}$.

Selon cette définition, on a alors :

$$\begin{aligned}
Q_i^*(h_i, a_i) &= \sum_{s,a_{-i},h_{-i}} \pi(a_{-i}|h_{-i}) P_i(s, h_{-i}|h_i) \tilde{Q}_i^*(s, h_i, \vec{a}) \\
&= \sum_{s,a_{-i},h_{-i}} P_i(s, a_{-i}, h_{-i}|h_i) \tilde{Q}_i^*(s, h_i, \vec{a}) \\
&\qquad \text{(car } a_{-i} \text{ conditionnellement indépendant de } s \text{ et } h_i \text{ connaissant } h_{-i}) \\
&= \sum_{s,a_{-i}} \tilde{Q}_i^*(s, h_i, \vec{a}) \sum_{h_{-i}} P_i(s, a_{-i}, h_{-i}|h_i) \\
&= \sum_{s,a_{-i}} P_i(s, a_{-i}|h_i) \tilde{Q}_i^*(s, h_i, \vec{a}) \sum_{h_{-i}} P_i(h_{-i}|h_i, s, a_{-i}) \\
&= \sum_{s,a_{-i}} P_i(s, a_{-i}|h_i) \tilde{Q}_i^*(s, h_i, \vec{a}) \\
&= \sum_{s,a_{-i}} P_i(a_{-i}|s, h_i) P_i(s|h_i) \tilde{Q}_i^*(s, h_i, \vec{a})
\end{aligned}$$

Ce qui permet d'avoir une règle de mise à jour sans marginalisation sur tous les historiques possibles.

Algorithme néanmoins trop complexe dans la plupart des cas : marginalisation sur toutes les combinaisons d'actions + apprentissage du monde et modèle des autres agents pour chaque agent i.

2 Preuves du modèle COMA

Soit la fonction d'avantage pour l'agent a : $A^a(s,u) = Q(s,u) - b(s,u^{-a})$, avec $b(s,u^{-a}) = \sum_{u'^a} \pi_{\theta_a}(u'^a|\tau_t^a)Q(s,(u^{-a},u'^a))$.
On souhaite montrer que considérer la baseline $b(s,u^{-a})$ dans la fonction d'avantage (plutôt que le classique $V(s)$) ne biaise pas le gradient $\mathbb{E}_\pi[\sum_a \nabla_\theta \log \pi^a(u^a|\tau^a)A^a(s,u)]$.

$$\begin{aligned}
g_b &= \mathbb{E}_\pi[\sum_a \nabla_\theta \log \pi^a(u^a|\tau^a)b(s,u^{-a})] \\
&= \sum_s d^\pi(s) \sum_a \sum_{u^{-a}} \pi(u^{-a}|\tau^{-a}) \sum_{u^a} \pi^a(u^a|\tau^a)\nabla_\theta \log \pi^a(u^a|\tau^a)b(s,u^{-a}) \\
&= \sum_s d^\pi(s) \sum_a \sum_{u^{-a}} \pi(u^{-a}|\tau^{-a}) \sum_{u^a} \nabla_\theta \pi^a(u^a|\tau^a)b(s,u^{-a}) \\
&= \sum_s d^\pi(s) \sum_a \sum_{u^{-a}} \pi(u^{-a}|\tau^{-a})b(s,u^{-a})\nabla_\theta 1 \\
&= 0
\end{aligned}$$

avec $d^\pi(s)$ la distribution ergodique discountée sur les états.

La baseline b ne biaise donc pas le gradient de COMA.

À noter que ce gradient, de la même manière que pour les Actor-Critic classiques, est non biaisé uniquement pour des valeurs Q non biaisées, soit en version tabulaire, soit en utilisant une fonction compatible (dans l'article les auteurs utilisent une fonction compatible).