

Cooperative Inverse Reinforcement Learning (CIRL)

Présentation - MAP578

Nathanaël Cuvelle-Magar, Jérémie Dentan, Philippe Nugnes

30 novembre 2021

Plan

- 1 Introduction
- 2 Illustration sur un exemple
- 3 Analyse théorique
- 4 Simulations
- 5 Conclusion
- 6 References et annexes

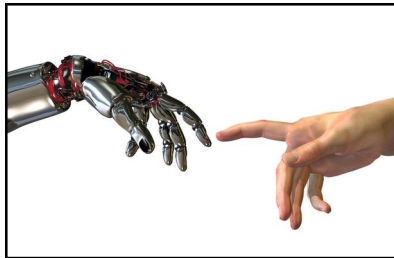
Introduction

Alignement

Comment s'assurer qu'un système autonome agit selon les intentions de ses concepteurs ?

Sur quelle mesure entraîner les algorithmes de façon à ce qu'ils répondent à nos besoins ?

Une solution proposée : l'IRL [Rus98]



Problématiques liées à l'IRL

Le comportement humain est imparfait

Le robot ne doit pas copier la fonction de gain de l'humain

Risque de l'apprentissage passif

Un robot simple observateur passif n'apprendra pas correctement.
Combiner erreur et expertise

Problématique : En quoi le CIRL répond à la problématique de l'alignement ? Quelles faiblesses et pistes d'amélioration ?

Plan

- 1 Introduction
- 2 Illustration sur un exemple**
- 3 Analyse théorique
- 4 Simulations
- 5 Conclusion
- 6 References et annexes

Énoncé du problème

Soient un humain H et un robot R . L'humain a une préférence politique $\theta \in [0, 1]$ et le robot dispose d'un ensemble de m publicités dont chacune est connotée politiquement, d'un facteur $y_i \in [0, 1]$.

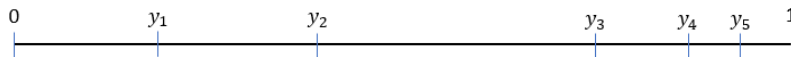
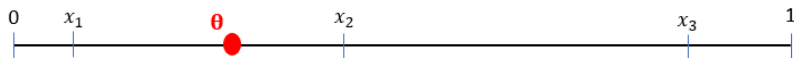
Objectif : Le robot choisit une publicité y_i , la plus proche possible de θ (au sens de la distance euclidienne)

Principe :

Phase 1 (*learning*) : l'humain présente au robot ses actions et lui permet de déterminer un encadrement de θ

Phase 2 (*deployment*) : Le robot estime θ comme le milieu du dernière intervalle obtenu et agit en conséquence

Description de la phase 1



Deux phases bien distinctes : *learning* et *deployment*.

Mécanisme de déploiement

Comportement du robot

Le robot agit comme si l'humain faisait une *demonstration by expert* (DBE) lors de la *learning phase*, et en déduit sa *belief* sur θ .

Théorème 2

Lors de la phase de déploiement, le robot maximise le *reward* associé à la moyenne de sa *belief* sur θ .

Plan

- 1 Introduction
- 2 Illustration sur un exemple
- 3 Analyse théorique**
- 4 Simulations
- 5 Conclusion
- 6 References et annexes

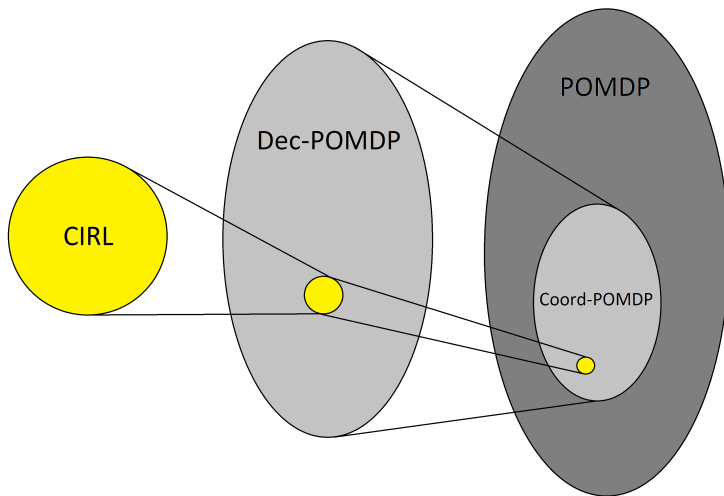


Figure – Vision globale

CIRL : une nouvelle formalisation du problème

Formalisme [DHM]

- des états du monde $s \in \mathcal{S}$;
- des actions possibles pour l'humain et le robot ($a^H \in \mathcal{A}^H$ et $a^R \in \mathcal{A}^R$) ;
- l'évolution du monde est contrôlée par une distribution $T(s'|s, a^H, a^R)$;
- la récompense est donnée à chaque étape par une fonction paramétrée $R : \mathcal{S} \times \mathcal{A}^H \times \mathcal{A}^R \times \Theta \longrightarrow \mathbb{R}$;
- l'état initial (s_0, θ) est tiré d'une distribution P_0

Objectif

Trouver un couple $\pi^H : (\mathcal{A}^H \times \mathcal{A}^R \times \mathcal{S})^* \times \Theta \longrightarrow \mathcal{A}^H$ et $\pi^R : (\mathcal{A}^H \times \mathcal{A}^R \times \mathcal{S})^* \longrightarrow \mathcal{A}^R$ de "politiques" optimal.

Problème d'optimisation

On cherche à résoudre le problème d'optimisation suivant :

$$\max_{\pi^H, \pi^R} \mathbb{E}^{\pi^H, \pi^R} \left[\sum_{t \geq 0} \gamma^t R(s_t, a_t^H, a_t^R, \theta) \right]$$

Caractérisation des solutions

Afin de justifier l'existence de politiques optimales et de préciser leur structure, on réduit le CIRL à des problèmes déjà étudiés dans la littérature :

- le CIRL peut se reformuler comme un dec-POMDP, ce qui nous informe sur sa complexité [BZ100] *a priori*,
- [NMT12] proposent un formalisme plus général que le dec-POMDP, qu'ils réduisent* à un POMDP.

POMDP

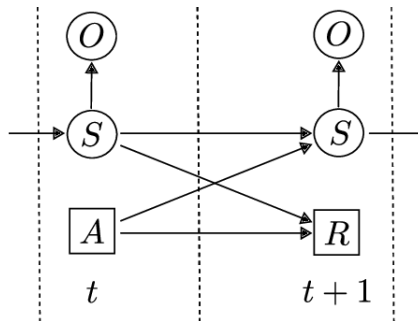


Figure – illustration d'un POMDP [BGD14]

Dec-POMDP

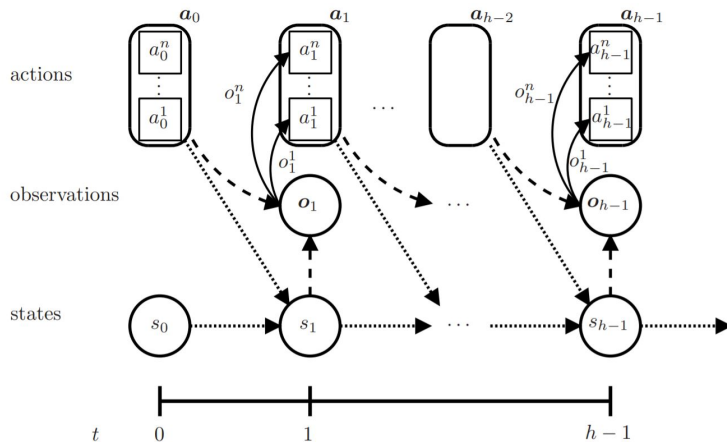


Figure – illustration d'un dec-POMDP [Oli]

Grandes lignes de la réduction : Dec-POMDP \rightarrow POMDP

Dec-POMDP

- états du monde : $x_t \in X$
- observations : $y_t^i \in Y^i, i = 1, \dots, n$
- actions : $u_t^i \in U^i, i = 1, \dots, n$
- (mémoire locale : $m_t^i \in (Y_{1:t-1}^i, U_{1:t-1}^i), i = 1, \dots, n$)
- (mémoire mise en commun à un instant donné :
 $z_t \in (Y^1, U^1) \times \dots \times (Y^n, U^n)$)

POMDP associé (Coord-POMDP)

- nouvel état du monde : $s = (x_t, (y_t^i)_{1 \leq i \leq n}, (m_t^i)_{1 \leq i \leq n})$
- observations : $o_t = z_{t-1}$
- nouvelles actions : $a_t = (\Gamma_t^i)_{1 \leq i \leq n}$, où les $\Gamma_t^i : Y^i \times M^i \rightarrow U^i$ sont des règles de décision.

Résultats

Dans le cas des CIRL :

- $X = S \times \Theta$, $Y^H = S \times \Theta$ et $Y^R = S$
- la seule information stockée en mémoire locale est le paramètre θ , le reste étant gardé dans la mémoire commune.

Théorème 1

Pour tout CIRL M d'espace d'états S et d'espace de paramètres de récompense Θ , il existe un POMDP M_C d'espace d'états de taille $|S|.|\Theta|$ tel qu'à toute paire de politiques de M soit associée une politique M_C de même espérance de gain.

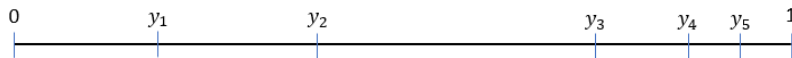
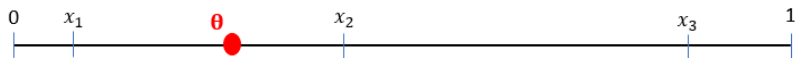
Corollaire 1

Pour tout CIRL, il existe une paire de politiques optimale ne dépendant que de l'état courant et de l'estimation de θ par R [SS73].

Plan

- 1 Introduction
- 2 Illustration sur un exemple
- 3 Analyse théorique
- 4 Simulations**
- 5 Conclusion
- 6 References et annexes

Stratégie d'apprentissage CIRL / DBE



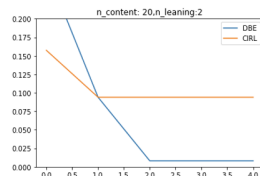
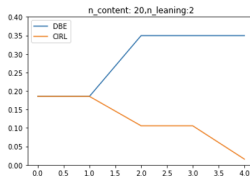
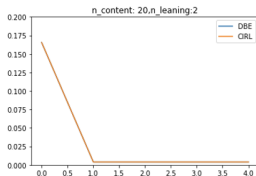
Implémentation

```
#Config
n_content, n_learning, epoch = (20, 2, 5)

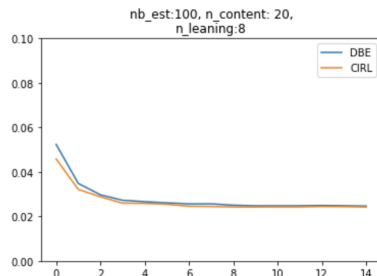
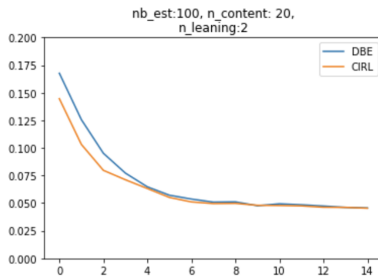
#Instances
hm = human()
rbt = robot(n_content = n_content)
cdt = [candidates(n_ = n_learning) for k in range(epoch)]

#Learning
for k in range(epoch):
    rbt.learn(hm, mode="CIRL", verbose = False, cdt = cdt[k])

#Deployment
print(hm.reward(rbt.action()))
```



Estimation de densité



Que retenir de cet exemple ?

Ces problématiques sont réelles et non abstraites

Situation concrète de notre vie numérique : ex : Netflix...

La stratégie CIRL peut être biaisée

Pour la dé-biaiser, il faut que le robot suppose un *noise* sur les actions de l'humain, et non pas une DBE parfaite.

Quantité d'information

Le gain d'efficacité de CIRL est surtout significatif en présence de peu d'information : `n_learning` ou `epoch` faibles.

Plan

- 1 Introduction
- 2 Illustration sur un exemple
- 3 Analyse théorique
- 4 Simulations
- 5 Conclusion**
- 6 References et annexes

Reviews

Domaine d'actions

Est-il possible que CIRL surperforme DBE lorsque le robot et l'humain ont le même domaine d'actions ?

Données expérimentales

Une stratégie d'apprentissage CIRL a-t-elle été effectivement implémentée, et si oui, avec quel succès ?

Manque de détails

De nombreux détails manquant ou des preuves évasives. Par exemple sur la notion de *decentralised*-POMDP.

Conclusion

Sur l'article...

- ➊ Des concepts nouveaux mais une théorie qui se base sur un domaine et des résultats connus ;
- ➋ Des pistes d'approfondissement : *optimal teaching*, spectres d'actions de H et R ...

Sur notre travail...

- ➊ Une formalisation et des détails qui manquent à l'article et qui sont réclamés dans les *reviews* ;
- ➋ Un exemple inédit, plus proche de problématiques modernes et quotidiennes.

Plan

- 1 Introduction
- 2 Illustration sur un exemple
- 3 Analyse théorique
- 4 Simulations
- 5 Conclusion
- 6 References et annexes

References I



Sebastian Brechtel, Tobias Gindele, and Rüdiger Dillmann, *Probabilistic decision-making under uncertainty for autonomous driving using continuous pomdps*, 17th International IEEE Conference on Intelligent Transportation Systems (ITSC) (2014), 392–399.







Daniel S. Bernstein, Shlomo Zilberstein, and Neil Immerman, *The complexity of decentralized control of markov decision processes*, UAI, 2000.



P. Abbeel S. Russell D. Hadfield-Menell, A. Dragan, *Cooperative inverse reinforcement learning*.

References II

-  Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis, *Decentralized stochastic control with partial history sharing : A common information approach*, 2012.
-  Frans A. Oliehoek, *Decentralized pomdps*.
-  Stuart Russell, *Learning agents from uncertain environments*, Proceedings of the Eleventh Annual Conference on Computational Learning Theory (1998).
-  Richard D. Smallwood and Edward J. Sondik, *The optimal control of partially observable markov processes over a finite horizon*, Operations Research **21** (1973), no. 5, 1071–88.

Annexe 1 : estimation de densité

