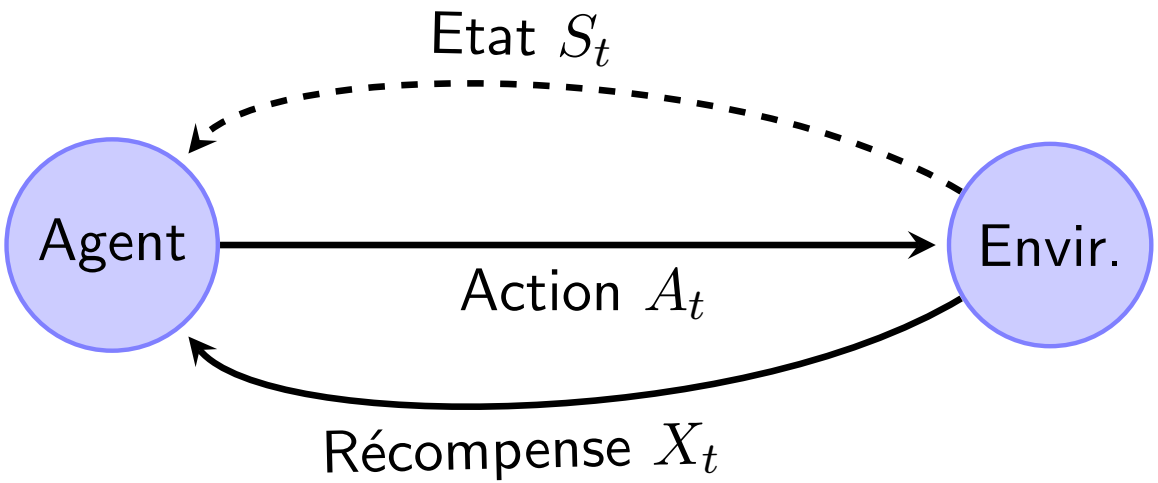
Exploration optimale à l'aide d'experts probabilistes [arXiv:1110.5447]

Sébastien Bubeck, Damien Ernst et Aurélien Garivier

Princeton, Université de Liège et CNRS-Telecom ParisTech

8 novembre 2011

Apprentissage par renforcement



dilemme
exploration
|
exploitation

RL $\neq$ apprentissage classique (notion de récompense)
RL $\neq$ théorie des jeux (environnement indifférent)

Problèmes de bandits

Essais cliniques séquentiels :

problème : des patients atteints d'une certaine maladies sont diagnostiqués au fil du temps

outils : on dispose de plusieurs traitements mal dont l'efficacité est a priori inconnue

déroulement : on traite chaque patient avec un traitement, et on observe le résultat (binaire)

objectif : soigner un maximum de patients (et pas connaître précisément l'efficacité de chaque traitement)

Principe d'optimisme

Algorithmes **optimistes** : [Lai&Robins '85 ; Agrawal '95]

Fais comme si tu te trouvais dans l'environnement qui t'est le plus favorable parmi tous ceux qui rendent les observations suffisamment vraisemblables

De façon plutôt inattendue, les méthodes optimistes se révèlent pertinentes dans des cadres très différents, efficaces, robustes et simples à mettre en oeuvre

Stratégies "Upper Confidence Bound"

UCB [Lai&Robins '85 ; Auer&al '02 ; Audibert&al '07]

- Construit une UCB pour chaque bras :

$$\underbrace{\frac{S_t(a)}{N_t(a)}}_{\text{récompense moyenne estimée}} + \underbrace{\sqrt{\frac{\log(t)}{2N_t(a)}}}_{\text{bonus d'exploration}}$$

- Choisis le bras qui la plus grande UCB

Avantage : comportement facilement interprétable et "acceptable"
Politique d'indice : on calcule un indice par bras et on choisit celui qui est le plus élevé, cf. [Gittins '79]

Good-UCB : convergence uniforme

Nombre d'objets intéressants trouvés par Good-UCB (trait plein), l'oracle (pointillés épais), et par échantillonnage uniforme (pointillé léger) en fonction du temps pour des tailles $N = 128, N = 500$, $N = 1000$ et $N = 10000$, dans un environnement à 7 experts.