Comparaison de méthodes de type UCB appliqués à la résolution de MDP

Lucas Gerretsen
Juliette Mitjans

Encadré par Odalric-Ambrym Maillard

18/03/2018

1 Introduction

L'apprentissage par renforcement consiste à observer les récompenses dues aux actions passées, afin de déterminer les actions futures. La difficulté réside dans le caractère inconnu des dynamiques du système. L'apprentissage par renforcement peut être appliqué aux problème de bandits armés qui répondent à des problématiques de minimisation du regret. Dans ce cadre, la notion de tradeoff entre exploration et exploitation est centrale. En comparaison, les techniques classiques de Q-learning, seules, ne permettent pas de minimiser efficacement le regret.

Les algorithmes classiques UCB (Uper Confidence Bound), UCB-V (Variance) et KL-UCB (Kullback Leibler) surestiment le regret. Une approche plus récente a été prouvée plus proche de la solution optimale : posterior sampling, ou Thompson Sampling. (G. AGRAWAL 2013b) Le principe de base est de construire une posterior bayésienne pour les récompenses dues à chaque action, de générer des échantillons indépendants, puis de choisir l'action qui correspond à la valeur de récompense attendue la plus élevée parmi les échantillons.

On se place ainsi dans le cadre de la résolution de MDP avec minimisation du regret. Notre objectif est de comparer les algorithmes UCRL2 (JAKSCH 2010) et PSRL (J. AGRAWAL 2017).

2 Définitions

Définition 1 *Un MDP (processus de Markov décisionnel) est un quadruplet $M = (S, A, T, R)$ où : - S est l'ensemble des états, ici fini - A est l'ensemble des actions, ici fini - dans un cas général, T est une application de $SxAxS \rightarrow [0; 1]$: $T(s, a, s')$ représente, pour un agent en s qui effectue l'action a, la probabilité de transiter vers le nouvel état s' - dans un cas détérministe, R est une application $SxAxS \rightarrow [0; 1]$: $R(s, a, s')$ représente, pour un agent qui transite de s à s' par l'action a, la récompense observée.*

Définition 2 *Une stratégie est une application π de S dans A qui associe une action a à chaque état s.*

Définition 3 *Le diamètre $D(M)$ de M est défini comme suit :*

$$D(M) = max_{s \neq s', (s,s') \in S^2} min_{\pi : S \rightarrow A} T^{\pi}_{s,s'}$$

où $T^{\pi}_{s,s\prime}$ est le nombre moyen d'étapes pour parvenir à $s\prime$ en partant de s et en suivant la stratégie π.

Définition 4 *Un MDP M est dit communiquant s'il a un diamètre $D(M)$ fini.*

Nous verrons que la condition que M est communiquant est essentielle (théorème 6). Cette condition réduit le champ des MDP solvables avec les algorithmes que nous présenterons. Par exemple, les MDP finis pour lesquels un état ne peut être visité qu'une fois sont courants mais ne remplissent pourtant pas cette condition.

Définition 5 *Le gain d'une stratégie π, partant de $s_1 = s$, est défini comme suit :*

$$\lambda^{\pi}(s) = \mathbb{E}[lim_{T\to\inf}\frac{1}{T}\Sigma_{t=1..T}r(s_t, \pi(s_t), s_{t+1})\|s_1 = s]$$

où $(s_t)_{t\geq 1}$ désigne une trajectoire aléatoire partant de s_1.

S'ensuit alors le théorème suivant, qui intervient dans les garanties des algorithmes :

Définition et théorème 6 (théorèmes 8.1.2 et 8.3.2 (PUTERMAN 2014)) *Si le MDP M est communiquant, on a :*

$$\exists \pi*, \forall s \in S, max_{s'\in S} max_{\pi}\lambda^{\pi}(s') = \lambda^{\pi*}(s)$$

On définit alors cette quantité $\lambda*$ comme le gain optimal, et la stratégie $\pi*$ comme la stratégie optimale.

Le théorème 6 est en particulier utile en pratique dans le sens où les garanties de convergence vers la stratégie optimale $\pi*$ ne dépendent pas de l'état initial.

L'objectif pour un agent est d'adopter à terme une stratégie afin de maximiser ses récompenses. La difficulté réside dans le fait que le MDP possède des propriétés de transition aléatoires dont la loi est inconnue. Bien que les définitions ci-dessus ne l'incluent pas, nous verrons que les algorithmes s'adaptent également lorsque les récompenses sont elles-aussi aléatoires. Pour résoudre un MDP, on pourrait imaginer un agent qui suit dans un premier temps une stratégie exploratoire pendant un temps indéfiniment long afin d'estimer les propriétés du MDP, puis dans un second temps qui en déduit et suit la stratégie optimale en terme de récompenses cumulées.

Dans le cadre de minimisation du regret, on ajoute la contrainte que l'on souhaite minimiser le regret de l'agent, ce qui impose un compromis entre exploration et exploitation propre au problème du bandit armé.

Définition 7 *On définit ainsi le regret à horizon de temps T induit par une trajectoire finie $(s_t, a_t)_{t=1..T}$:*

$$\Delta(M, T) = T\lambda* - \Sigma_{t=1..T}t_{s_t,a_t}$$

3 Bornes sur le regret

UCRL2 et PSRL ont une approche générale algorithmique très similaire. PSRL conçu quelques années après UCRL, bénéficie d'une meilleure borne sur le regret.

On précise à nouveau les conditions dans lesquelles les théorèmes suivant sont valables : $M = (S, A, T, R)$ est supposé être un MDP communiquant, avec S et A finis, et connus de l'algorithme.

Théorème 7 (théorème 4 (JAKSCH 2010)) *Le regret de* UCRL2 *est, avec probabilité de plus de* $1 - 3\delta$ *:* $\forall s \in S, \forall T \geq 1, \forall \epsilon > 0$*,*

$$\Delta(M, \text{UCRL2}, s, T) \leq 34^2 \frac{D^2 S^2 A log(T/\delta)}{\epsilon} + \epsilon T$$

En prenant, dans le théorème 7, $\epsilon = DSA^{1/2}T^{-1/2}$ on obtient un regret $\tilde{O}(DS\sqrt{AT})$ où $\tilde{O}$ comprend des constantes et des facteurs en log des variables.

Par ailleurs, on a :

Théorème 8 (théorème 1 (J. AGRAWAL 2017)) *Le regret de* PSRL *est, avec probabilité de plus de* $1 - \delta$ *:* $\forall s \in S, \forall T \geq CDAlog^2(T/\delta)$*,*

$$\Delta(M, \text{PSRL}, s, T) \leq \tilde{O}(D\sqrt{SAT} + DS^{7/4}A^{3/4}T^{1/4} + DS^{5/2}A) \leq 34^2 \frac{D^2 S^2 A log(T/\delta)}{\epsilon}$$

qui se réécrit, pour $T \geq S^5 A$

$$\Delta(M, \text{PSRL}, s, T) \leq \tilde{O}(D\sqrt{SAT})$$

On voit avec ces deux derniers théorèmes que PSRL est en effet une amélioration de UCRL2 du point de vue de la borne du regret, d'un facteur $\sqrt{S}$.

Jaksch et al démontrent également une borne inférieure du regret $\Delta(M, \mathcal{A}, s, T)$ d'un algorithme $\mathcal{A}$ (théorème 5 (JAKSCH 2010)) en $\tilde{O}(\sqrt{DSAT})$. Ainsi, PSRL n'est au plus plus qu'à un facteur $\sqrt{D}$ du regret optimal.

4 Synthèse des 2 algorithmes

Les deux algorithmes UCRL2 et PSRL suivent la même démarche générale :

Entrées : Paramètre de confiance δ, S, A
For episodes $k = 1, 2, \ldots$ **do**
1. Déterminer une extension M_k de M (défini par des contraintes continues ou par un ensemble fini d'échantillons)
2. Estimer la politique optimale $\tilde{\pi}^k$ par Extended Value Iteration appliqué à M_k (fini ou infini)
3. Exécuter $\tilde{\pi}^k$ afin de générer un épisode partant de s_t, dont la longueur est telle qu'on a au maximum doublé le nombre d'observations des états

Essentiellement, UCRL2 travaille dans un sous-ensemble continu des extensions de M, délimité par des contraintes qui définissent des simplex. Ainsi, lors de l'étape de mise à jour de la fonction de valeur dans Value Iteration avec l'opérateur de Bellman optimal, le problème de maximisation se résout en parcourant le simplex correspondant au couple (s, a).

A contrario, PSRL travaille dans un sous-ensemble fini des extensions de M, échantillonnés suivant une loi de Dirichlet. Dans cette situation, la maximisation à l'étape de mise à jour s'effectue sur

un nombre fini de valeurs possibles. Le choix de la loi de Dirichlet est assez pratique : cette loi est conjuguée à elle-même ; aussi, entre deux itérations, les coefficients sont simplement mis à en fonction de nouvelles observations des états lors de la dernière génération de trajectoire.

5 Perspectives d'implémentation

Chacun des deux algorithmes est assez bien détaillé, et les deux semblent reproduisibles. Une fois implémentée, une comparaison empirique pourrait être envisagée pour une application à des MDP qui respectent les conditions vues dans les théorèmes précédents.

Dans les parties respectives de description des deux algorithmes, alors que le traitement des récompenses stochastiques est explicite dans l'algorithme UCRL2, il n'est que peu abordé dans l'algorithme PSRL. L'argument est donné que l'estimation exacte des transitions est davantage essentielle pour la détermination de la politique optimale que l'estimation des récompenses. Pour l'implémentation d'une extension de PSRL aux MDP avec récompenses stochastiques bornées, l'article renvoie aux techniques de Thompson Sampling de (G. AGRAWAL 2013b).

Références

AGRAWAL, Goyal (2013b). « Further Optimal Regret Bounds for Thompson Sampling. » In : *AISTATS*, p. 99–107.

AGRAWAL, Jia (2017). « Optimistic posterior sampling for reinforcement learning : worst-case regret bounds ». In : *31st Conference on Neural Information Processing System*.

JAKSCH Ortner, Auer (2010). « Near-optimal regret bounds for reinforcement learning. » In : *Journal of Machine Learning Research* 11(Apr), p. 1563–1600.

PUTERMAN (2014). « Markov decision processes : discrete stochastic dynamic programming . » In :