Apprentissage par renforcement

Aurelien Garivier

Journées MAS 2010, Bordeaux

Apprentissage par renforcement

Session organisée par **Aurélien Garivier**

Dans un problème d'*apprentissage par renforcement* [1, 2], un agent évoluant dans un environnement aléatoire doit cumuler un maximum de récompenses en choisisant au fil du temps la meilleure politique, c'est-à-dire la meilleure réaction possible à ses observations. Une telle situation est modélisée par un processus de décision markovien : on suppose que la suite des états que traverse l'agent est une chaîne de Markov dont les noyaux de transitions successifs sont déterminés par les actions choisies, et on admet que la récompense reçue à chaque instant est une fonction (aléatoires) de l'état courant. Quand les propriétés probabilistes de l'environnement sont connues, la détermination de la politique optimale, qui constitue le problème dit de *planification*, est typiquement un problème de programmation dynamique.

Mais quand l'environnement est inconnu, il n'existe pas de solution générale au problème, et le choix d'une politique doit s'appuyer sur des procédures d'estimation. Deux familles de méthodes sont envisageables. Dans les premières, le choix de l'agent se base sur l'estimation directe de la performance des différentes politiques qui s'offrent à lui : la question est donc d'abord de savoir évaluer une politique le plus efficacement possible. Différents algorithmes ont été proposés, dont il faut pouvoir contrôler le risque ; pour la méthode des différences temporelles par moindres carrés, il est possible de prouver des bornes de convergence non asymptotiques.

Les méthodes de la deuxième famille passent par l'estimation des paramètres du modèle, c'est-à-dire des lois de transistions et des distributions des récompenses. Un simple 'plug-in' de ces estimées dans le programme dynamique mène le plus souvent à des stratégies insuffisamment exploratoires mais, dans les cas les plus simples, il est possible de faire presque aussi bien qu'un agent connaissant la politique optimale dès l'origine. Une classe intéressante d'algorithmes se base sur le principe dit d'*optimisme face à l'incertitude* [3] : l'agent choisit d'agir comme s'il évoluait dans l'environnement le plus favorable parmi tous ceux qui rendent ses observations suffisamment vraisemblables. L'enjeu est alors de montrer que cette heuristique mène à des algorithmes présentant à la fois des garanties théoriques de performance et une faible complexité algorithmique. Parfois, l'agent s'autorise d'abord une phase exploratoire, pendant laquelle il ne tient pas compte des récompenses cumulées, avant que ne commence une phase d'exploitation où il doit immédiatement mettre à profit l'expérience accumulée : il est alors également possible d'exhiber des stratégies presque minimax.

Références :

[1] *Neuro-dynamic programming* (1996), D.P. Bertsekas and J.N. Tsitsiklis, Athena Scientific.
[2] *Reinforcement Learning and Dynamic Programming Using Function Approximators* (2010), L. Busoniu, R. Babuska, B. De Schutter et D. Ernst, CRC Press, Automation and Control Engineering Series.
[3] *Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem* (1995), R. Agrawal, Advances in Applied Probability 27, 1054-1078.

Adresse de l'organisateur :

Aurélien Garivier
LTCI, CNRS UMR 5141, Telecom ParisTech
Département TSI, site Dareau 46 rue Barrault 75634 Paris cedex 13 France
E-mail : aurelien.garivier@telecom-parisTech.fr
<http://www.telecom-paristech.fr/~garivier>

Journées MAS 2010, Bordeaux

Session : Apprentissage par renforcement

Finite sample analysis of Least Squares Temporal Differences

par **Rémi Munos**

L'exposé commencera par une brève introduction à l'apprentissage par renforcement, en insistant sur le compromis exploration-exploitation. Nous aborderons d'abord les solutions optimistes du problème des bandits multi-bras, avant d'en arriver aux problèmes de bandits sur les espaces métriques, qui peuvent être traités par un algorithme d'optimisation optimiste hiérarchique.

Ensuite, je m'intéresserai spécifiquement au problème de l'évaluation de politique, c'est-à-dire à l'apprentissage de la valeur d'une politique donnée, par la méthode des différences temporelles par moindres carrés (Least-Squares Temporal-Difference). Je présenterai une analyse non asymptotique de LSTD. La borne proposée est générale, dans le sens où aucune hypothèse n'est faite sur l'existence d'une distribution stationnaire de la chaîne de Markov résultante. Je terminerai par la présentation des extensions à différentes versions de LSTD.

Adresse :
Rémi Munos
INRIA Lille - Nord Europe, équipe SequeL
40 avenue Halley
59650 Villeneuve d'Ascq, FRANCE
E-mail : remi.munos@inria.fr
<http://sequel.futurs.inria.fr/munos>

Journées MAS 2010, Bordeaux

Session : Apprentissage par renforcement

Model-Free Monte Carlo-like Policy Evaluation

par Raphael Fonteneau, Susan A. Murphy, Louis Wehenkel et **Damien Ernst**

We propose an algorithm for estimating the finite-horizon expected return of a closed loop control policy from an a priori given (off-policy) sample of one-step transitions. It averages cumulated rewards along a set of "broken trajectories" made of one-step transitions selected from the sample on the basis of the control policy. Under some Lipschitz continuity assumptions on the system dynamics, reward function and control policy, we provide bounds on the bias and variance of the estimator that depend only on the Lipschitz constants, on the number of broken trajectories used in the estimator, and on the sparsity of the sample of one-step transitions.

Adresses :
Raphael Fonteneau
Université de Liège
Institut Montefiore, Sart-Tilman B28
B-4000 Liège BELGIQUE
E-mail : raphael.fonteneau@ulg.ac.be
<http://sites.google.com/site/raphaelfonteneau/home>

Susan A. Murphy
University of Michigan
439 West Hall, 1085 S. Univ. University of Michigan Ann Arbor, MI 48109-1107
E-mail : samurphy@umich.edu
<http://www.stat.lsa.umich.edu/~samurphy/>

Louis Wehenkel
Université de Liège
Institut Montefiore, Sart-Tilman B28
B-4000 Liège BELGIQUE
E-mail : L.Wehenkel@ulg.ac.be
<http://www.montefiore.ulg.ac.be/~lwh/>

Damien Ernst
Université de Liège
Institut Montefiore, Sart-Tilman B28
B-4000 Liège BELGIQUE
E-mail : dernst@ulg.ac.be
<http://www.montefiore.ulg.ac.be/~ernst/>

Session : Apprentissage par renforcement

Journées MAS 2010, Bordeaux

Session : Apprentissage par renforcement

Optimisme en apprentissage par renforcement et divergence de Kullback-Leibler

par **Sarah Filippi**, Olivier Cappé et Aurélien Garivier

We consider model-based reinforcement learning in finite Markov Decision Processes (MDPs), focussing on so-called optimistic strategies. Optimism is usually implemented by carrying out extended value iterations, under a constraint of consistency with the estimated model transition probabilities. In this paper, we strongly argue in favor of using the Kullback-Leibler (KL) divergence for this purpose. By studying the linear maximization problem under KL constraints, we provide an efficient algorithm for solving KL-optimistic extended value iteration. When implemented within the structure of UCRL2, the near-optimal method introduced by [Auer&al, 2009], this algorithm also achieves bounded regrets in the undiscounted case. We however provide some geometric arguments as well as a concrete illustration on a simulated example to explain the observed improved practical behavior, particularly when the MDP has reduced connectivity. To analyze this new algorithm, termed KL-UCRL, we also rely on recent deviation bounds for the KL divergence which compare favorably with the L_1 deviation bounds used in previous works.

Adresses :

Sarah Filippi
LTCI, CNRS UMR 5141, Telecom ParisTech
Département TSI, site Dareau 46 rue Barrault 75634 Paris cedex 13 France
E-mail : sarah.filippi@telecom-paristech.fr
<http://www.telecom-paristech.fr/>

Olivier Cappé
LTCI, CNRS UMR 5141, Telecom ParisTech
Département TSI, site Dareau 46 rue Barrault 75634 Paris cedex 13 France
E-mail : olivier.cappe@telecom-parisTech.fr
<http://www.telecom-paristech.fr/~cappe>

Aurélien Garivier
LTCI, CNRS UMR 5141, Telecom ParisTech
Département TSI, site Dareau 46 rue Barrault 75634 Paris cedex 13 France
E-mail : aurelien.garivier@telecom-parisTech.fr
<http://www.telecom-paristech.fr/~garivier>

Journées MAS 2010, Bordeaux

Session : Apprentissage par renforcement

Open Loop Optimistic Planning

par **Sébastien Bubeck**

We consider the problem of planning in a stochastic and discounted environment with a limited numerical budget. More precisely, we investigate strategies exploring the set of possible sequences of actions, so that, once all available numerical resources (e.g. CPU time, number of calls to a generative model) have been used, one returns a recommendation on the best possible immediate action (or sequence of actions) to follow based on this exploration. The performance of a strategy is assessed in terms of its simple regret, that is the loss in performance resulting from choosing the recommended action instead of an optimal one. We first provide a minimax lower bound for this problem, and show that a uniform planning strategy matches this minimax rate (up to a logarithmic factor). Then we propose a UCB (Upper Confidence Bounds)-based planning algorithm, called OLOP (Open-Loop Optimistic Planning), which is also minimax optimal, and prove that it enjoys much faster rates when there is a small proportion of near- optimal sequences of actions. Finally, we compare our results with the regret bounds one can derive for our setting with bandits algorithms designed for an infinite number of arms.

Adresse :
Sébastien Bubeck
INRIA Lille - Nord Europe, équipe SequeL
Parc scientifique de la Haute-Borne, 40 avenue Halley
59650 Villeneuve d'Ascq, FRANCE
E-mail : sebastien.bubeck@inria.fr
<http://sequel.futurs.inria.fr/bubeck>