Apprentissage par renforcement

- Victor Boone (Université Grenoble Alpes)

  Title: Instance optimal regret in finite Markov Decision Processes

  Abstract: This talk will present a recent advance in reinforcement learning : instance optimal regret in finite Markov decision processes (MDPs). In this learning setting, a learner faces an unknown (finite) MDP that is controlled by playing actions. Overtime, the player gathers information on the environment and is expected to play better. The regret measures the learning performance by comparing the aggregate rewards of the learner to the score of an ominous player that knows the optimal actions in advance. We will present a tight lower bound of the regret and show its tightness by providing a algorithmic scheme whose asymptotic upper bound is equal to the lower bound. This work goes beyond the ergodic setting and covers all communicating MDPs.

- Aurelien Garivier (ENS Lyon)

  Title: Optimisation dans les processus de décision markoviens: au-delà de la moyenne

  Abstract: Les équations de Bellman permettent d'optimiser l'espérance de l'utilité dans les processus de décision markoviens. Mais comment faire si l'on souhaite optimiser d'autres fonctionnelles de l'utilité, par exemple pour des raisons de gestion des risques ? L'apprentissage distributionnel peut représenter un espoir intéressant, dans la mesure où il permet de garder une trace non seulement du comportement moyen, mais de l'ensemble de la distribution. On s'efforcera dans cet exposé de cerner quelles sont les fonctionnelles de l'utilité qui sont optimisables par programmation dynamique, et d'illustrer dans quelle mesure celles-ci répondent à la problématique de gestion des risques.

- Matthieu Jonckheere (LAAS CNRS)

  Title: Gradient estimation for structured reinforcement learning and convergence properties on infinite state space

  Abstract: Many Markov decision processes (MDPs) have large state and action spaces as well as non-convex objective functions, which hinders the convergence properties of many reinforcement learning (RL) algorithms. We show that these difficulties can be circumvented by exploiting some simple structure of the underlying MDP. We focus on a particular class of RL algorithms, called policy-gradient methods, which directly optimize the policy parameter via stochastic gradient ascent, without necessarily relying on value-function estimation. These methods are known to perform better on MDPs with large state and action spaces, but they sometimes experience slow convergence due to the high variance of the gradient estimator. We design a new family of gradient estimators, called score-aware gradient estimators (SAGEs), that apply to policy parameterizations under which the stationary distribution of the MDP forms an exponential family parameterized by the policy parameter. Our second contribution is a convergence result showing that, under appropriate assumptions, the policy under SAGE-based policy-gradient methods has a large probability of converging to an optimal policy, provided that it starts sufficiently close. This holds also when the state space is infinite, the objective function is nonconvex and has multiple optimal policies. Lastly, we compare numerically the performance of a SAGE-based policy gradient method with that of actor-critic. This is joint work with Céline Comte (LAAS CNRS, Toulouse), Jaron Sanders (Eindhoven University of Technology) and Albert Senen-Cerda (IRIT and LAAS, CNRS, Toulouse).