Emilie Kaufmann

From regret to PAC RL

In online reinforcement learning, regret is perhaps the most studied performance metric in the literature on theoretical RL. In this talk we will consider episodic MDP and study the dual PAC RL framework, in which the goal is to identify near-optimal policies with high confidence, relaxing the need to maximize rewards while learning. In particular, we will be interested in the reward-free exploration problem, in which the goal is to learn a good policy with respect to *any* reward function that is given after the exploration phase. We will introduce different algorithms that can be viewed as variant of the UCB-VI algorithm (whose regret has been well studied) incorporating some appropriate (intrinsic) rewards to foster exploration. In particular, we will present the first algorithm with a sample complexity bound that does not depend only on the size of the MDP, going beyond minimax guarantees.