Contribution ID: **81** Type: **not specified**

The sliding regret

Thursday, June 20, 2024 11:00 AM (1 hour)

Optimistic reinforcement learning algorithms in Markov decision processes essentially rely on two ingredients to guarantee regret efficiency. The first one is the choice of well-tuned confidence bounds and the second is the design of a pertinent rule to end episodes. While many efforts have been dedicated to improve the tightness of confidence bounds, the management of episodes has remained essentially unaltered since the introduction of the doubling trick (DT) in UCRL2 (Auer et Al.2009). In this talk, I will present two solutions to move beyond (DT). The first one is the performance test (PT) that ends an episode as soon as the performance of the current policy becomes obviously sub-optimal. The second one is the vanishing multiplicative (VM) rule that is as simple as DT to implement and replace the doubling criterion by a weaker one. Both solutions keep the regret of the algorithm unaltered and induce a drastic reduction of the local regret taken from the start of exploration episodes (start of episodes where a sub-optimal policy is used). More specifically, classical algorithms such as UCRL2, KL-UCRL or UCRL2B, patched with our new rules get an immediate benefit. Their regret upper bound remain the same (up to a small negligible additive term) while their asymptotic local regret at exploration times decreases from Omega(T) to O(log T). I will also comment on numerical experiments confirming our asymptotic findings. The regret of algorithms under (VM) or (PT) becomes slightly better and significantly smoother, while the local regret at exploration times becomes sub-linear, even over finite times.

Session Classification: Keynote: Bruno Gaujal (INRIA)