Analysis of Classification-based Policy Iteration Algorithms

Mohammad Ghavamzadeh (INRIA Lille)

April 29, 2011

We present a variant of the classification-based approach to policy iteration which uses a cost-sensitive loss function weighing each classification mistake by its actual regret, i.e., the difference between the action-value of the greedy action and of the action chosen by the classifier. For this algorithm, we provide a full finite-sample analysis. Our results state a performance bound in terms of the number of policy improvement steps, the number of rollouts used in each iteration, the capacity of the considered policy space (classifier), and a capacity measure which indicates how well the policy space can approximate policies that are greedy w.r.t. any of its members. The analysis reveals a tradeoff between the estimation and approximation errors in this classification-based policy iteration setting. Furthermore, it confirms the intuition that classification-based policy iteration algorithms can be favorably compared to value function based approaches when the good policies are easier to be represented and learned than their corresponding value functions. We also study the consistency of the algorithm when there exists a sequence of policy spaces with increasing capacity.