CSCE 475/875 Multiagent Systems

Handout 11: Reinforcement Learning and Other Learning

February 4, 2020

(Based on Shoham and Leyton-Brown 2011)

Reinforcement Learning in Unknown MDPs

Reinforcement learning does *not* explicitly model the opponent's strategy.

The specific family of techniques we look at are derived from the Q-learning algorithm for learning in unknown (single-agent) MDPs.

Consider (single-agent) MDPs. Value iteration assumes that the MDP is known. **What if we do not know the rewards or transition probabilities of the MDP?** It turns out that, if we always know what state we are in and the reward received in each iteration, we can still converge to the correct Q-values. (***Note***: The intuition is that we approximate the unknown transition probability by using the actual distribution of states reached in the game itself.)

Definition 7.4.1 (Q**-learning)** Q-*learning is the following procedure:*

Initialize the Q-function and V values (arbitrarily, for example)
Repeat *until convergence*
  Observe the current state s_t.
  Select action a_t and take it.
  Observe the reward $r(s_t, a_t)$ and next state s_{t+1}
  Perform the following updates (and do *not* update any other Q-values):

$$Q_{t+1}(s_t, a_t) \leftarrow (1 - \alpha_t) Q_t(s_t, a_t) + \alpha_t (r(s_t, a_t) + \beta V_t(s_{t+1}))$$
$$V_{t+1}(s) \leftarrow \max_a Q_t(s, a)$$

End Repeat

Theorem 7.4.2 Q-*learning guarantees that the* Q *and* V *values converge to those of the optimal policy, provided that each state-action pair is sampled an infinite number of times, and that the time-dependent learning rate* α_t *obeys* $0 \leq \alpha_t < 1, \sum_0^\infty \alpha_t = \infty$ *and* $\sum_0^\infty \alpha_t^2 < \infty$.

Issues:

(1) How to design the order in which the algorithm selects actions?

(2) What is the rate of convergence?

(3) It gives *no* assurance regarding the accumulation of optimal future discounted rewards by the agent—it could well be, depending on the discount factor, that by the time the agent converges to the optimal policy it has paid too high a cost, which cannot be recouped by exploiting the policy going forward. This is *not* a concern if the learning takes place during training sessions, and *only* when learning has converged sufficiently is the agent unleashed on the world (e.g., think of a fighter pilot being trained on a simulator before going into combat).

(4) But in general Q-learning should be thought of as guaranteeing good learning, but neither quick learning nor high future discounted rewards.

Belief-Based Reinforcement Learning

There is also a version of reinforcement learning that includes **explicit modeling of the other agent(s)**, given by the following equations.

$$Q_{t+1}(s_t, a_t) \leftarrow (1 - \alpha_t) Q_t(s_t, a_t) + \alpha_t \big(r(s_t, a_t) + \beta V_t(s_{t+1})\big)$$

$$V_t(s) \leftarrow \max_{a_i} \sum_{a_{-i} \subset A_{-i}} Q_t(s, (a_i, a_{-i}))\, Pr_i(a_{-i})$$

In this version, the agent updates the value of the game using the ***probability*** it assigns to the opponent(s) playing each action profile. Of course, the belief function must be updated after each play. How it is updated depends on what the belief function is.

Other Learning Behaviors

No-Regret Learning. As discussed above, a learning rule is universally consistent or (equivalently) exhibits no regret if, loosely speaking, against any set of opponents it yields a payoff that is no less than the payoff the agent could have obtained by playing any one of his pure strategies throughout. ***A learning rule is said to exhibit no regret if it guarantees that with high probability the agent will experience no positive regret.***

Targeted Learning. Here we discuss an alternative sense of "good," which retains the requirement of best response, but ***limits it to a particular class of opponents***. The intuition guiding this approach is that in any strategic setting, one has *some* sense of the agents in the environment. A chess player has studied previous plays of his/her opponent, and so on. And so it ***makes sense to try to optimize against this set of opponents, rather than against completely unknown opponents***.

Difference between No-Regret Learning and Targeted Learning: Consider learning in a repeated Prisoner's Dilemma game. Suppose that the target class consists of all opponents whose strategies rely on the past iteration (e.g., TfT). In this case successful targeted learning will result in constant cooperation, while no-regret learning prescribes constant defection.