IIIS - AI (Fall 2018) Homework Set 3

Due: December 5, 2018

November 19, 2018

1 Policy iteration convergence (25 Points)

Recall the policy iteration algorithm and notation from Lecture 6 on Markov Decision Processes. We will alter notation slightly here. Let the initial policy be $\pi_0 = \hat{\pi}$ and let π_i be the policy after performing update i.

1.1 12 Points

Prove the following equation:

$$v^{\pi_i} \leq v^{\pi_{i+1}}$$

1.2 13 Points

Prove that policy iteration converges to the optimal solution.

2 Optional: Maximin Objective (30 Points)

In our lecture, we aims to find a policy for Markov Decision Processes to maximize the expected discounted reward. Now we assume that there are multiple agents, where each agent has its own reward function. We use a multi-agent Markov decision processes (MDP) to this problem, defined by a tuple $\langle I, S, A, T, \{R_i\}_{i \in I} \rangle$, where

I $= \{1, \ldots, n\}$ is a set of agent indices.

S is a finite set of states.

A $= \times_{i \in I} A_i$ is a finite set of joint actions, where A_i is a finite set of actions available for agent i.

T: $S \times A \times S \to [0, 1]$ is the transition function. $T(s'|s, a)$ is the probability of transiting to the next state s' after a joint action $a \in A$ is taken by agents in state s.

R_i: $S \times A \rightarrow \Re$ is a reward function of agent i and provides agent i with an individual reward $R_i(s, a)$ after a joint action a taken in state s.

Our goal for a given multi-agent MDP is to find a joint control policy π^* that maximizes the following objective value function

$$V(\pi) = \min_{i \in I} \mathbf{E}[\sum_{t=0}^{\infty} \lambda^t R_i(\mathbf{x}^t, \mathbf{a}^t) | \pi, b]. \quad (1)$$

where λ is the discount factor, the expectation operator $\mathbf{E}(\cdot)$ averages over stochastic action selection and state transition, b is the initial state distribution, and $\mathbf{x}^t$ and $\mathbf{a}^t$ are the state and the joint action taken at time t, respectively.

2.1 15 Points

Design an algorithm to compute an optimal policy for optimizing this maximin objective.

2.2 15 Points

If the transition function T is unknown, does there exist a model-free learning algorithm to learn the optimal joint policy? If so, describe it; otherwise, explain it.

3 Optional: Multi-Agent Learning (30 Points)

In Lecture 8, we discussed that the basic gradient ascent algorithm does not always converge and the gradient ascent with policy prediction only converge in two-play, two-action games.

3.1 15 Points

Can you design an algorithm converging in 2x3 games (one agent has two actions and the other agent has three actions?

3.2 15 Points

The convergence of gradient ascent with policy prediction assumes each agent can observe the payoff matrix and the other agent's policy. Can you design a new algorithm to relax these assumptions?