Homework3: R.L. - Written Component

Feng Gao (2015011208)

December 2018

1 Policy iteration convergence (25')

Let the initial policy be $\pi_0 = \hat{\pi}$ and let π_i be the policy after performing update i.

1.1 Prove that $v^{\pi_i} \leq v^{\pi_{i+1}}$ (12')

Proof:

In policy iteration algorithm, we do policy evaluation first until convergence, then update policy using one-step look-ahead with resulting converged utilities as future values. Repeat these two steps until policy converges.

Policy evaluation:

$$v_{k+1}^{\pi_i}(s) = R(s) + \gamma \sum_{s'} P(s'|s,a) v_k^{\pi_i}(s')$$

Policy improvement:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} P(s'|s,a) v^{\pi_i}(s')$$

We can represent these two equations by means of matrices as following:

Policy evaluation:

$$V_n = R^{\pi_n} + \gamma T^{\pi_n} V_n$$

Policy improvement:

$$\pi_{n+1} = \arg\max_a R^a + \gamma T^a V_n$$

Let $H(V) = R + \gamma T V$, then we can know that $H^*(V_n) \geq H^{\pi_n}(V_n) = V_n$.

Then

$$H^*(V_n) = R^{\pi_{n+1}} + \gamma T^{\pi_{n+1}} V_n$$

Rearranging:

$$R^{\pi_{n+1}} \geq (I - \gamma T^{\pi_{n+1}}) V_n$$

Hence

$$V_{n+1} = (I - \gamma T^{\pi_{n+1}})^{-1} R^{\pi_{n+1}} \geq V_n$$

1.2 Prove that policy iteration converges to the optimal solution. (13')

Proof:

According to the question 1.1, we know that $V_{n+1} \geq V_n$.

We assume that action set A and state set S are finite, then there are finitely many policies and therefore the algorithm terminates in finitely many iterations.

At termination, $\pi_{n+1} = \pi_n$ and therefore

$$V_n = V_{n+1} = \max_a R_a + \gamma T^a V_n$$

2 Optional: Maximin Objective (30')

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 $< I, S, A, T, \{R_i\}_{i \in I} >$, where

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

S is a finite set of states.

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

$T : S \times A \times S \longrightarrow [0, 1]$ 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.

$R_i : S \times A \longrightarrow \mathcal{R}$ 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

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

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

2.1 **Design an algorithm to compute an optimal policy for optimizing this maximin objective. (15')**

2.2 **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. (15')**

3 Optional: Multi-Agent Learning (30')

3.1 **Can you design an algorithm converging in 2×3 games (one agent has two actions and the other agent has three actions? (15')**

3.2 **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? (15')**