IEOR 8100-001: Learning and Optimization for Sequential Decision Making 04/04/16

Lecture 20: MDP/Reinforcement Learning

Instructor: Shipra Agrawal Scribed by: Amine Allouah, Yang Kang.

After having covered the theory of the Markov Decision Processes (MDP) in the previous classes, we will introduce in this lecture, the "reinforcement learning" problem, that is, MDP with unknown parameters.

For that, we consider the same formulation of MDP as in the previous class. Especially, we consider:

$$\mathcal{M} = (S, A, \mathbb{P}, r),$$

where S is the state space, A is the action space, $\mathbb{P}$ is the transition probability within the state space under the action space A and r is the reward function.

1 Problem formulation:

The dynamic happens as follow, at every time $t = 1, 2, \cdots$

- Observe state $s_t \in S$.
- Take an action a_t, observe reward $r_{a_t}(s_t)$.
- Observe transition from s_t to s_{t+1}, where $s_{t+1} = s'$ with probability $p_{a_t}(s_t, s')$.

In the reinforcement learning problem, the transition probabilities $p_a(s, s')$ are unknown. In fact, in a more general setting, for any given a, s rewards $r_a(s)$ are IID samples from some distribution $\nu_a(s)$, which is also unknown. But, for simplicity, in this class, we will consider rewards $r_a(s)$ to be known and observable whenever action a is taken in state s.

The objective or the **goal** is to maximize the reward over a certain period T, i.e:

$$\max \sum_{t=1}^{T} r_t,$$

where $r_t = r_{a_t}(s_t)$, $a_t \in A$ is the action taken by algorithm at time t.

Considering that there is one unknown distribution $(p_a(s, \cdot))$ for every s, a, we may want to think of this as a bandit problem with $|S||A|$ arms. So, can we achieve $O(\sqrt{SAT})$ regret bounds by using MAB algorithms? The challenge is that unlike the MAB problem, here we cannot arbitrarily choose (s_t, a_t) as any of these $|S||A|$ options at a given time step t – while we can choose any action a_t allowed in a given state, the current state s_t is determined by the previous actions and transition process.

1.1 Assumptions

Two assumptions are made in order to characterize the problem:

. The rewards are assumed to be bounded, i.e $r_a(s) \in [0,1]\ \forall (s,a) \in (S,A)$.

. The MDP is communicating.

Here after we define a communicating MDP.

Definition 1. *The diameter of a MDP is defined as follow:*

$$D = \max_{s_1 \neq s_2} \min_{\pi \in \Pi} \mathbb{E}[\tau(s_1, s_2, \pi)],$$

where $\tau(s_1, s_2, \pi)$ is a stopping time describing the number of steps needed to reach state s_2 from state s_1, and Π is the space of the feasible Markovian stationary policies.

Comment: $\min_{\pi \in \Pi} \mathbb{E}[\tau(s_1, s_2, \pi)]$ is the minimum average time needed to go from s_1 to s_2, so the diameter is capturing the maximum average time needed to transit from any state to any other state.

Definition 2. *a MDP is communicating if and only if its diameter is finite.*

Example 3. *The ergodic and the recurrent and the unichain MDPs are examples of communicating MDPs.*

1.2 Regret Definition

To define the regret, we need first to define a benchmark, for that, we use the expected average infinite horizon reward for MDP as benchmark:

$$\mathbb{E}[\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} r_t | s_1].$$

The following theorem characterizes the optimal policy that maximizes this benchmark.

Theorem 4 (Puterman 1994, Thm 8.1.2)**.** *Given any starting state s_1, there always exists a Markovian stationary (possibly randomized) policy that maximizes the expected average infinite horizon reward for MDP.*

In below, a policy will always assumed to be Markovian stationary unless otherwise stated. Before defining the regret, we still need one definition.

Definition 5. *A gain of policy π starting at state s is defined as follow:*

$$\rho^{\pi}(s) = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} r_{a_t}(s_t) | s_1 = s\right],$$

where $a_t \sim \pi(s_t)$.

The optimal gain at state s is defined as follow:

$$\rho^*(s) = \max_{\pi} \rho^{\pi}(s).$$

In Puterman 1994, Proposition 8.1.1, it is shown that the above limit exists for stationary policies.

For the communicating MDPs, we can find a relationship between all the optimal gains.

Theorem 6 (Puterman 1994, Theorem 8.3.2)**.** *If the MDP is communicating then:*

$$\forall s, \rho^*(s) = \rho^*,$$

it means that the optimal gain does not depend on the starting state.

Sketch of the proof: Suppose that that there exists $s_1 \neq s_2$ such that $\rho^*(s_1) > \rho^*(s_2)$.

Since the MDP is communicating there exists a policy π' using which we can go from s_2 to s_1 in finite expected time. Then we can construct a (possibly non-stationary) policy, which first goes from s_2 to s_1 using π' in finite expected time, and then uses the given policy π^* in state s_2. Such a policy will have gain $\rho^*(s_1) > \rho^*(s_2)$ in state s_2 which is a contradiction to assumption that $\rho^*(s_2)$ is the optimal gain in state s_2. (note that although the new policy with better gain could be non-stationary, by Theorem 4, there must exist a stationary policy with the same or better gain). □

Now we have all the ingredients to define the regret.

Definition 7. *The regret is defined as follow:*

$$R(T) = T\rho^* - \sum_{t=1}^{T} r_t.$$

2 Bias of a policy

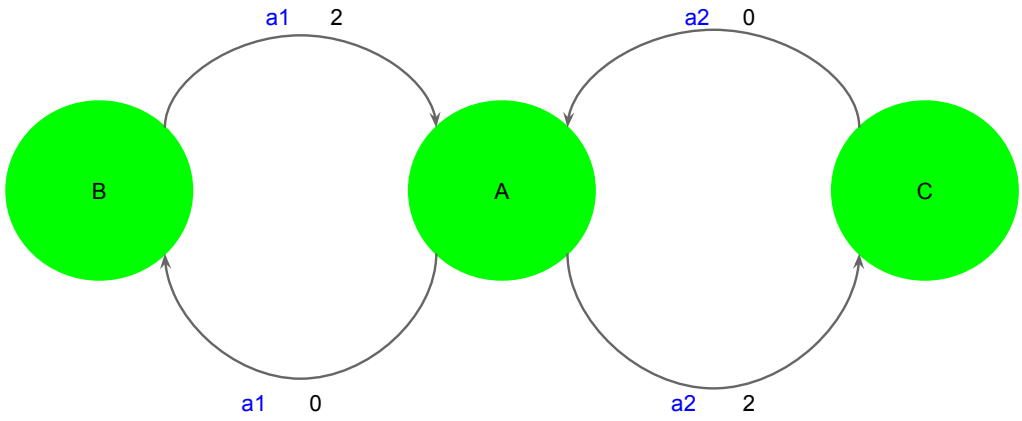
Here we define and explain "bias of a policy", this quantity will be useful in deriving optimality equations and in regret analysis.

Definition 8. *For a policy π and a state s, the bias is defined as follow:*

$$V^\pi(s) = \lim_{T\to\infty} E[\sum_{t=1}^{T}(r_t - \rho^\pi(s_t))|s_1 = s]$$

The limit in above is Cesaro limit, i.e. above is interpreted as:

$$V^\pi(s) = \lim_{n\to\infty} \frac{1}{n}\sum_{T=1}^{n} E[\sum_{t=1}^{T}(r_t - \rho^\pi(s_t))|s_1 = s]$$



Example 9. *Consider two policies: the one that selects a_1 in state A, and the one that selects a_2 in state A. The gain (average reward) for both policies is 1. For policy that selects a_1 in state A, the bias of state B alternates (for increasing T) between 1 and -1, the cesaro limit is 0. For policy that selects a_2 in state A, the bias of state B alternates between 2+1 and 2-1, cesaro limit is 2.*

We have:

$$\begin{aligned} V^*(s) &= \lim_{T\to\infty} E[\sum_{t=1}^{T}(r_t - \rho^*(s_t))|s_1 = s] \\ &= r_{a_1^*}(s) - \rho^* + \sum_{s'} \lim_{T\to\infty} E[\sum_{t=2}^{T}(r_t - \rho^*(s_t))|s_2 = s']p_{a_1^*}(s,s') \\ &= r_{a_1^*}(s) - \rho^* + \sum_{s'} V^*(s')p_{a_1^*}(s,s'). \end{aligned}$$

So

$$\rho^* = r_{a_1^*}(s) - V^*(s) + p_{a_1^*}(s,\cdot)^T V^*$$

In fact (proof in Puterman 1994, Theorem 8.4.1 for unichain MDPs, extended to weakly communicating MDPs in Chapter 9), we have following optimality equations,

$$\rho^* = \max_{a\in A} r_a(s) - V^*(s) + p_a(s,\cdot)^T V^*, \forall s$$

Therefore, we can find ρ^* by solving the following linear program:

$$\begin{array}{ll} \min_{\rho,h} & \rho \\ \text{s.t} & \rho \geq r_a(s) - h + p_a(s)^T h, \qquad \forall s, a \end{array}$$

Here, h is a $|S|$ dimensional vector, and $p_a(s)$ is an $|S|$-dimensional vector with components $\{p_a(s,s')\}_{s'}$. Given, optimal solution h^* to above, an optimal policy is

$$\pi^*(s) = \arg\max_a r_a(s) - h^* + p_a(s)^T h^*$$

Also for all s, h^*_s is equal to bias $V^{\pi^*}(s)$ of this policy within some additive constant factor, i.e., there exists some constant c such that,

$$V^{\pi^*}(s) = h^*_s + c, \forall s$$