IEOR 8100-001: Learning and Optimization for Sequential Decision Making 03/23/16

Lecture 17: Optimality of Markovian Stationary Policy

Instructor: Shipra Agrawal Scribed by: Chun Ye

Recall the setting of Markovian multi-armed bandit from last lecture. We start from states $s_{1,1}, \ldots, s_{N,1}$ for arms $1, \ldots, N$. At each time t, we observe the state $s_{a,t} \in \mathcal{S}_a$, for each arm a. Our policy then takes an action $a_t \in \mathcal{A}$ (i.e. pulls an arm). As a consequence of our arm pull, we receive a reward $r_t = r_{a_t}(s_{a_t,t})$. Note that the reward only depends on the state of the pulled arm at time t. After the arm pull, the state of arm a_t transitions from $s = s_{a_t,t}$ to $s' = s_{a_t,t+1} \in \mathcal{S}_a$ with probability $p_{a_t}(s, s')$ while states of all arms that were not pulled in time t stay the same. Our objective is to minimize the expected total discounted reward, i.e.

$$\lim_{T\to\infty} V^{\pi}_{1,T}(s_1) = E[\sum_{t=1}^{\infty} \gamma^{t-1} r_{a_t}(s_{a_t,t})].$$

We will show that an optimal policy can be obtained by computing the so-call Gitten's index. An index for every potential state of every arm. Given the states of arms $s_{1,t}, \ldots, s_{N,t}$, this optimal pulls the arm with the highest Gitten's index, i.e.

$$a_t = \arg \max_{a=1,\ldots,N} G_a(s_{a,t}).$$

The Gitten's index is defined to be:

$$\forall a, s \in S_a \quad G_a(s) = \sup_{\tau \geq 1} \frac{E[\sum_{t=1}^{\tau} \gamma^{t-1} r_a(\tilde{s}_{a,t})]}{E[\sum_{t=1}^{\tau} \gamma^{t-1}]},$$

the numerator denotes the expected reward that we can obtain if we were to pull a single arm a for τ units of time from $t = 1$, where τ is some stopping time. The proof will be divided into two steps.

1. There exists an optimal Markovian stationary policy for this problem.
2. The Gitten's index policy is optimal amongst all Markovian stationary policies.

We will focus on proving the optimality of a Markovian stationary policy in this lecture. Actions taken by a Markovian stationary policy depend on the current state, and is independent of time. More precisely, let $\pi(.)$ be a Markovian stationary policy then

$$\pi(s_{1,t}, \ldots, s_{N,t}) = \pi(s_{1,t'}, \ldots, s_{N,t'}),$$

whenever $(s_{1,t}, \ldots, s_{N,t}) = (s_{1,t'}, \ldots, s_{N,t'})$.

Note that a priori, the optimal policy could be history dependent, meaning that the action at time t depends on the observed history up till time t. We denote this action by $a_t = \pi(\mathcal{H}_t)$. Given a policy π, we denote the expected reward from l till time T given $s_l = (s_{1,l}, \ldots, s_{N,l})$ of π by: $V^{\pi}_{l,T}(s) = E[\sum_{t=l}^{T} \gamma^{t-l} r_{a_t}(s_{a_t,t}) | s_l = s]$ and we define

$$V^{\pi}_{l,\infty}(s) = \lim_{T\to\infty} E[\sum_{t=l}^{T} \gamma^{t-l} r_{a_t}(s_{a_t,t}) | s_l = s] = E[\sum_{t=l}^{\infty} \gamma^{t-l} r_{a_t}(s_{a_t,t}) | s_l = s].$$

Note that we can change the limit and expectation by dominated convergence theorem as the rewards are bounded for all state-action pair and $\gamma < 1$.

Before we prove the main theorem, we will derive a recursion formula for $V^{\pi}_{l,T}(s)$.

$$
\begin{aligned}
V^{\pi}_{l,T}(s) &= E[\sum_{t=l}^{T} \gamma^{t-1} r_{a_t}(s_{a_t,t}) | s_l = s] \\
&= E[r_{a_l}(s) + \gamma \sum_{t=l+1}^{T} \gamma^{t-(l-1)} r_{a_t}(s_{a_t,t}) | s_l = s] \\
&= E[r_{a_l}(s) + \sum_{s'} \gamma E[\sum_{t=l+1}^{T} \gamma^{t-(l+1)} r_{a_t}(s_{a_t,t}) | s_{l+1} = s'] p_{a_l}(s, s') | s_l = s] \\
&= E[r_{a_l}(s) + \sum_{s'} \gamma E[\sum_{t=l+1}^{T} \gamma^{t-(l+1)} r_{a_t}(s_{a_t,t}) | s_l = s, s_{l+1} = s'] p_{a_l}(s, s') | s_l = s] \\
&= E[r_{a_l}(s) + \gamma \sum_{s'} V^{\pi}_{l+1,T}(s'; s_l = s) p_{a_l}(s, s') | s_l = s]
\end{aligned}
$$

The third equality is obtained by conditioning on the state s' that we transition to from s. Here s' and s agrees on all coordinates except for the a_l-th one. Moreover, $V^{\pi}_{l+1,T}(s'; s_l = s) = E[\sum_{t=l+1}^{T} \gamma^{t-(l+1)} r_{a_t}(s_{a_t,t}) | s_l = s, s_{l+1} = s']$ in the last equality. Given any partial history $\mathcal{P}_t$ by time t, let

$$
V^{\pi}_{l,T}(s; \mathcal{P}_l) = E[\sum_{t=l+1}^{T} \gamma^{t-(l+1)} r_{a_t}(s_{a_t,t}) | s_l = s, \mathcal{P}_l].
$$

we can show via similar arguments that

$$
V^{\pi}_{l,T}(s; \mathcal{P}_l) = E[r_{a_l}(s) + \gamma \sum_{s'} V^{\pi}_{l+1,T}(s'; \mathcal{P}_{l+1}) p_{a_l}(s, s') | s_l = s],
$$

where $\mathcal{P}_{l+1} = \mathcal{P}_l \cup \{s_l = s\}$.
The main theorem hinges on the following key lemma.

Lemma 1. *Let π^* denote an optimal policy, then*

$$
V^{\pi^*}_{l,T}(s) = \max_{a=1,\ldots,N} \{r_a(s) + \gamma \sum_{s'} V^{\pi^*}_{l+1,T}(s') p_a(s, s')\}.
$$

Moreover, $V^{\pi^*}_{l,T}(s) \geq V^{\pi}_{l,T}(s; \mathcal{P}_l)$ for all partial history $\mathcal{P}_l$ and policy π.

The lemma above implies that the optimal policy is Markovian. We are ready to prove the lemma via backward induction.

Proof. Base case: $l = T$, given any partial history $\mathcal{P}_T$ and any policy π,

$$
V^{\pi}_{T,T}(s; \mathcal{P}_T) = r_{a_T}(s) \leq \max_a r_a(s).
$$

Hence, it must be the case that $V^{\pi^*}_{T,T}(s) = \max_a r_a(s)$.

Inductive step: given any partial history $\mathcal{P}_l$ and any policy π,

$$\begin{aligned} V^{\pi}_{T,T}(s;\mathcal{P}_l) &= E[r_{a_l}(s) + \gamma \sum_{s'} V^{\pi}_{l+1,T}(s';\mathcal{P}_{l+1}) p_{a_l}(s,s') | s_l = s] \\ &\leq \max_{a=1,\ldots,N} \{r_a(s) + \gamma \sum_{s'} V^{\pi}_{l+1,T}(s';\mathcal{P}_{l+1}) p_a(s,s')\} \\ &\leq \max_{a=1,\ldots,N} \{r_a(s) + \gamma \sum_{s'} V^{\pi^*}_{l+1,T}(s') p_a(s,s')\} \end{aligned}$$

where the first inequality follows from expectation over actions is at most the action deterministic action that yields the most reward. The second inequality follows from the inductive hypothesis. $\max_{a=1,\ldots,N}\{r_a(s) + \gamma \sum_{s'} V^{\pi^*}_{l+1,T}(s') p_a(s,s')\}$ is attainable by a feasible policy starting at time l. Moreover, it is independent of history other than the starting state at time l. Hence, it must be the case that

$$V^{\pi^*}_{l,T}(s) = \max_{a=1,\ldots,N} \{r_a(s) + \gamma \sum_{s'} V^{\pi^*}_{l+1,T}(s') p_a(s,s')\},$$

which completes the proof. □

Note that the recursion given in the key lemma implies that $V^{\pi^*}_{l,T}(s)$ only depends on l and T through the value of $T - l$. Hence,

$$V^{\pi^*}_{l,T}(s) = V^{\pi^*}_{1,T-l+1}(s) \quad \forall s$$

Let $V^*(s) = V^{\pi^*}_{1,\infty}(s)$. Then we get that

$$\lim_{T\to\infty} V^{\pi^*}_{l,T}(s) = \lim_{T\to\infty} V^{\pi^*}_{1,T-l+1}(s) = V^*(s).$$

Hence we can rewrite the recursion formula

$$V^{\pi^*}_{l,\infty}(s) = \max_{a=1,\ldots,N} \{r_a(s) + \gamma \sum_{s'} V^{\pi^*}_{l+1,\infty}(s') p_a(s,s')\}$$

as the fixed point equation

$$V^*(s) = \max_{a=1,\ldots,N} \{r_a(s) + \gamma \sum_{s'} V^*(s') p_a(s,s')\}.$$

The above equation is known as the *Bellman equation*. It implies that the optimal policy is both Markovian and stationary.

Note that the space of Markovian stationary policy can be difficult to specify, as the optimal action to take may depend on the joint state of all arms. In the next lecture, we will show that the Gitten's index policy is optimal, which will drastically reduce the complexity of the optimal policy.