Lecture Note 15

1 Optimal Stopping Problems

In the last lecture, we have analyzed the behavior of $TD(\lambda)$ for approximating the cost-to-go function in autonomous systems. Recall that much of the analysis was based on the idea of sampling states according to their stationary distribution. This was done either explicitly, as was assumed in approximate value iteration, or implicitly through the simulation or observation of system trajectories. It is unclear how this line of analysis can be extended to general controlled systems. In the presence of multiple policies, in general there are multiple stationary state distributions to be considered, and it is not clear which one should be used. Moreover, the dynamic programming operator T may not be a contraction with respect to any such distribution, which invalidates the argument used in the autonomous case. However, there is a special class of control problems for which analysis of $TD(\lambda)$ along the same lines used in the case of autonomous systems is successful. These problems are called *optimal stopping problems*, and are characterized by a tuple $(\mathcal{S}, P : \mathcal{S} \times \mathcal{S} \mapsto [0,1], g_0 : \mathcal{S} \mapsto \Re, g_1 : \mathcal{S} \mapsto \Re)$, with the following interpretation. The problem involves a Markov decision process with state space $\mathcal{S}$. In each state, there are two actions available: to stop (action 0) or to continue (action 1). Once action 0 (stop) is selected, the system is halted and a final cost $g_0(x)$ is incurred, based on the final state x. At each previous time stage, action 1 (continue) is selected and a cost $g_1(x)$ is incurred. In this case, the system transitions from state x to state y with probability $P(x, y)$. Each policy corresponds to a (random) stopping time τ_u, which is given by

$$\tau_u = \min\{k : u(x_k) = 0\}.$$

Example 1 (American Options) *An* American call option *is an option to buy stock at a price K, called the stock price, on or before an expiration date — the last time period the option can be exercised. The state of a such a problem is the stock price P_k. Exercising the option corresponds to the "stop" action and leads to a reward* $\max(0, P_k - K)$*; not exercising the option corresponds to the "continue" action and incurs no costs or rewards.*

Example 2 (The Secretary Problem) *In the secretary problem, a manager needs to hire a secretary. He interviews secretaries sequentially and must make a decision about hiring each one of them immediately after the interview. Each interview incurs a certain cost for the hours spent meeting with the candidate, and hiring a certain person incurs a reward that is a function of the person's abilities.*

In the infinite-horizon, discounted-cost case, each policy is associated with a discounted cost-to-go

$$J_u(x) = \mathrm{E}\left[\sum_{t=0}^{\tau_u} \alpha^t g_1(t) + \alpha^{\tau_u} g_0(x_{\tau_u}) | x_0 = x\right].$$

We are interested in finding the policy u with optimal cost-to-go

$$J^* = \min_u J_u.$$

Bellman's equation for the optimal stopping problem is given by

$$J = \min(g_0, g_1 + \alpha PJ) \triangleq TJ.$$

As usual, T is a maximum-norm contraction, and Bellman's equation has a unique solution corresponding to the optimal cost-to-go function J^*. Moreover, T is also a weighted-Euclidean-norm contraction.

Lemma 1 *Let π be a stationary distribution associated with P, i.e., $\pi^T P = \pi^T$, and let $D = \mathrm{diag}(\pi)$ be a diagonal matrix whose diagonal entries correspond to the elements in vector π. Then, for all J and $\bar{J}$, we have*

$$\|TJ - T\bar{J}\|_{2,D} \leq \alpha \|J - \bar{J}\|_{2,D}$$

Proof: Note that, for any scalars c_1, c_2 and c_3, we have

$$|\min(c_1, c_3) - \min(c_2, c_3)| \leq |c_1 - c_2|. \tag{1}$$

It is straightforward to show Eq.(1). Assume without loss of generality that $c_1 \geq c_2$. If $c_1 \geq c_2 \geq c_3$ or $c_3 \geq c_1 \geq c_2$, Eq(1) holds trivially. The only remaining case is $c_1 \geq c_3 \geq c_2$, and we have

$$|\min(c_1, c_3) - \min(c_2, c_3)| = |c_3 - c_2| \leq |c_1 - c_2|.$$

Applying (1), we have

$$\begin{aligned} |TJ - T\bar{J}| &= |\min(g_0, g_1 + \alpha PJ) - \min(g_0, g_1 + \alpha P\bar{J})| \\ &\leq |\alpha P(J - \bar{J})|. \end{aligned}$$

It follows that

$$\begin{aligned} \|TJ - T\bar{J}\|_{2,D} &\leq \alpha \|P(J - \bar{J})\|_{2,D} \\ &\leq \alpha \|J - \bar{J}\|_{2,D}, \end{aligned}$$

where the last inequality follows from $\|PJ\|_{2,D} \leq \|J\|_{2,D}$ by Jensen's inequality and stationarity of π.

1.1 $TD(\lambda)$

Recall the Q function

$$Q(x, a) = g_a(x) + \alpha \sum_y P_a(x, y) J^*(y).$$

In general control problems, storing Q may require substantially more space than storing J^*, since Q is a function of state-action pairs. However, in the case of optimal stopping problems, storing Q requires essentially the same space as J^*, since the Q value of stopping is trivially equal to $g_0(x)$. Hence in the case of optimal stopping problems, we can set $TD(\lambda)$ to learn

$$Q^* = g_1 + \alpha PJ^*,$$

the cost of choosing to continue and behaving optimally afterwards. Note that, assuming that one-stage costs g_0 and g_1 are known, we can derive an optimal policy from Q^* by comparing the cost of continuing with the cost of stopping, which is simply g_0. We can express J^* in terms of Q^* as

$$J^* = \min(g_0, Q^*),$$

so that Q^* also satisfies the recursive relation

$$Q^* = g_1 + \alpha P \min(g_0, Q^*).$$

This leads to the following version of $TD(\lambda)$:

$$\begin{aligned} r_{k+1} &= r_k + \gamma_k z_k \left[g_1(x) + \alpha \min\left(g_0(x_{k+1}), \phi(x_{k+1}) r_k\right) - \phi(x_k) r_k\right] \\ z_k &= \alpha\gamma z_{k-1} + \phi(x_k) \end{aligned}$$

Let

$$HQ = g_1 + \alpha P \min(g_0, Q), \tag{2}$$

and

$$H_\lambda Q = (1-\lambda) \sum_{t=0}^{\infty} \lambda^t H^{t+1} Q.$$

We can rewrite $TD(\lambda)$ in terms of operator Q as

$$r_{k+1} = r_k + \gamma_k z_k (H_\lambda \Phi r_k - \Phi r_k + w_k). \tag{3}$$

The following property of H implies that an analysis of $TD(\lambda)$ for optimal stopping problems along the same lines of the analysis for autonomous systems suffices for establishing convergence of the algorithm.

Lemma 2

$$\|HQ - H\bar{Q}\|_{2,D} \leq \alpha \|Q - \bar{Q}\|_{2,D}.$$

Theorem 1 *[Analogous to autonomous systems] Let r_k be given by (3) and suppose that P is irreducible and aperiodic. Then $r_k \to r^*$ w.p.1, where r^* satisfies*

$$\Phi r^* = \Pi H_\lambda \Phi r^*$$

and

$$\|\Phi r^* - Q^*\|_{2,D} \leq \frac{1}{\sqrt{1-k^2}} \|\Pi Q^* - Q^*\|_{2,D}, \tag{4}$$

where $k = \frac{\alpha(1-\lambda)}{1-\alpha\lambda} \leq \alpha$.

We can also place a bound on the loss in the performance incurred by using a policy that is greedy with respect to Q^*, rather than the optimal policy. Specifically, consider the following stopping role based on Φr^*.

$$\tilde{u}(x) = \begin{cases} \text{stop}, & \text{if } g_0(x) \leq \Phi r^*(x) \\ \text{continue}, & \text{otherwise} \end{cases}$$

$\tilde{u}$ specifies a (random) *stopping time* $\tilde{\tau}$, which is given by

$$\tilde{\tau} = \min\{k : \phi(x_k) r^* \geq g_0(x_k)\},$$

and the cost-to-go associated with $\tilde{u}$ is given by

$$\tilde{J}(x) = \mathrm{E}\left[\sum_{t=0}^{\tilde{\tau}-1} \alpha^t g_1(x_t) + \alpha^{\tilde{\tau}} g_0(x_{\tilde{\tau}})\right].$$

The following theorem establishes a bound on the expected difference between $\tilde{J}$ and the optimal cost-to-go J^*.

Theorem 2

$$\mathrm{E}[\tilde{J}(x)|x \sim \pi] - \mathrm{E}[J^*(x)|x \sim \pi] \leq \frac{2}{(1-\alpha)\sqrt{1-K^2}}\|\Pi Q^* - Q^*\|_{2,D}$$

Proof: Let $\tilde{Q}$ be the cost of choosing to continue in the current time stage, followed by using policy $\tilde{u}$:

$$\tilde{Q} = g_1 + \alpha P \tilde{J}.$$

Then we have

$$\begin{aligned}
\mathrm{E}[\tilde{J}(x_0) - J(x_0)|x_0 \sim \pi] &= \mathrm{E}[\tilde{J}(x_1) - J(x_1)|x_0 \sim \pi] \\
&= \pi^T P(\underbrace{\tilde{J} - J}_{\geq 0}) \\
&= \sum_{x \in \mathcal{S}} \pi(x)|P(\tilde{J} - J^*)(x)| \\
&= \|P\tilde{J} - PJ^*\|_{1,\pi} \\
&\leq \|P\tilde{J} - PJ^*\|_{2,\pi} && (5) \\
&= \frac{1}{\alpha}\|g_1 + \alpha P\tilde{J} - g_1 + \alpha PJ^*\|_{2,\pi} \\
&= \frac{1}{\alpha}\|\tilde{Q} - Q^*\|_{2,\pi} && (6)
\end{aligned}$$

The inequality (5) follows from the fact that, for all J, we have

$$\begin{aligned}
\|J\|_{1,\pi}^2 &= \mathrm{E}[|J(x)| : x \sim \pi]^2 \\
&\leq \mathrm{E}[J(x)^2 : x \sim \pi] \\
&= \|J\|_{2,\pi}^2,
\end{aligned}$$

where the inequality is due to Jensen's inequality. Now define the operators $\tilde{H}$ and K, given by

$$\begin{aligned}
\tilde{H}Q &= g_1 + \alpha PKQ, \quad \text{where} \\
KQ &= \begin{cases} g_0(x), & \text{if } \Phi r^*(x) \geq g_0(x) \\ Q, & \text{otherwise.} \end{cases}
\end{aligned}$$

Recall the definition of operator H. Then it is easy to verify the following identities:

$$H\Phi r^* = \tilde{H}\Phi r^*$$
$$HQ^* = Q^*$$
$$\tilde{H}\tilde{Q} = \tilde{Q}.$$

Moreover, it is also easy to show that $\tilde{H}$ is a contraction with respect to $\|\cdot\|_{2,\pi}$. Now we have

$$\begin{aligned}
\|\tilde{Q} - Q^*\|_{2,\pi} &\leq \|Q^* - H\Phi r^*\|_{2,\pi} + \|\tilde{Q} - \tilde{H}\Phi r^*\|_{2,\pi} \\
&= \|HQ^* - H\Phi r^*\|_{2,\pi} + \|\tilde{H}\tilde{Q} - \tilde{H}\Phi r^*\|_{2,\pi} \\
&\leq \alpha\|Q^* - \Phi r^*\|_{2,\pi} + \alpha\|\tilde{Q} - \Phi r^*\|_{2,\pi} \\
&\leq \alpha\|Q^* - \Phi r^*\|_{2,\pi} + \alpha\big(\|\tilde{Q} - Q^*\|_{2,\pi} + \|Q^* - \Phi r^*\|_{2,\pi}\big) \\
&\leq 2\alpha\|Q^* - \Phi r^*\|_{2,\pi} + \alpha\|\tilde{Q} - \Phi r^*\|_{2,\pi}
\end{aligned}$$

Thus,

$$\|\tilde{Q} - Q^*\|_{2,\pi} \leq \frac{2\alpha}{1-\alpha}\|Q^* - \Phi r^*\|_{2,\pi}. \tag{7}$$

The theorem follows from Theorem 1 and equations (6) and (7). □

1.2 Discounted Cost Problems with Control

Our analysis of $TD(\lambda)$ is based on comparisons with approximate value iteration:

$$\Phi r_{k+1} = \Pi T_\lambda \Phi r_k,$$

where the projection represented by Π is defined with respect to the weighted Euclidean norm $\|\cdot\|_{2,\pi}$ (or, equivalently, $\|\cdot\|_{2,D}$) and π is the stationary distribution associated with the transition matrix P. In controlled problems, there is not a single stationary distribution to be taken into account; rather, for every policy u, there is a stationary distribution π_u associated with transition probabilities T_u. A natural way of extending AVI to the controlled case is to consider a scheme of the form

$$\Phi r_{k+1} = \Pi_{u_k} T_{u_k} \Phi r_k, \tag{8}$$

where Π_u represents the projection based on $\|\cdot\|_{2,\pi_u}$ and u_k is the greedy policy with respect to Φr_k. Such a scheme is a plausible approximation, for instance, for an approximate policy iteration based on $TD(\lambda)$ trained with system trajectories:

1. Select a policy u_0. Let $k = 0$.

2. Fit $J_{u_k} \approx \Phi r_k$ (e.g., via $TD(\lambda)$ for autonomous systems);

3. Choose u_{k+1} to be greedy with respect to Φr_k. Let $k = k+1$. Go back to step 2.

Note that step 2 in the approximate policy iteration presented above involves training over an infinitely long trajectory, with a single policy, in order to perform policy evaluation. Drawing inspiration from asynchronous policy iteration, one may consider performing policy updates before the policy evaluation step is considered. As it turns out, none of these algorithms is guaranteed to converge. In particular, approximate value iteration (8) is not even guaranteed to have a fixed point. For an analysis of approximate value iteration, see [1].

A special situation where AVI and $TD(\lambda)$ are guaranteed to converge occurs when the basis functions are constant over partitions of the state space.

Theorem 3 *Suppose that* $\phi_i(x) = 1\{x \in A_i\}$, *where* $A_i \cap A_j = \emptyset, \forall i, j, i \neq j$. *Then*

$$\Phi r_{k+1} = \Pi T \Phi r_k$$

converges for any Euclidean projections Π.

Proof: We will show that, if Π is a Euclidean projection and ϕ_i satisfy the assumption of the theorem, then Π is a maximum-norm nonexpansion:

$$\|\Pi J - \Pi \bar{J}\|_\infty \leq \|J - \bar{J}\|_\infty.$$

Let

$$\begin{aligned}(\Pi J)(x) &= K_i, \quad \text{if } x \in A_i, \text{ where} \\ K_i &= \arg\min_r \sum_{x \in A_i} w(x)\Big(J(x) - r\Big)^2\end{aligned}$$

Thus

$$K_i = \mathrm{E}[J(x)|x \sim w_i], \text{ where } w_i = \frac{w(x)}{\sum_{x \in A_i} w(x)}.$$

Therefore

$$\begin{aligned}(\Pi J)(x) - (\Pi \bar{J})(x) &= \mathrm{E}\big[J(x) - \bar{J}(x)|x \sim w_i\big] \\ &\leq \|J - \bar{J}\|_\infty\end{aligned}$$

It follows that ΠT is a maximum-norm contraction, which ensures convergence of approximate value iteration. □

References

[1] D.P. de Farias and B. Van Roy. On the existence of fixed points for appproximate value iteration and temporal-difference learning. *Journal of Optimization Theory and Applications*, 105(3), 2000.