IE 3186: Approximate Dynamic Programming **Fall 2018**

Lecture 12: Q-Learning for Optimal Stopping

Lecturer: Daniel Jiang *Scribes: Kamal Basulaiman, Jing Yang*

References:

J. N. Tsitsiklis, B. Van Roy, *Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to high-dimensional financial derivatives*, IEEE Transactions on Automatic Control, 1999.

D. P. Bertsekas, J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996. (§6.8)

12.1 Approximate Q-Learning for Optimal Stopping

- Consider a Markov chain $\{i_k\}$ taking values in $\{1, 2, \ldots, n\}$, where n is large. The original paper deals with stochastic process with $i_k \in \mathbb{R}^d$, but we will examine the simpler case.
- Transition probabilities (p_{ij}), where p_{ij} is the probability of transitioning from state i to state j in one period. Suppose there is a steady state distribution $\xi = (\xi_1, \xi_2, \ldots, \xi_n) > 0$.
- Decisions: {stop, go}. If "stop," pay cost $c(i)$. If "go," pay cost $g(i_k, i_{k+1})$. Crucial point here is that decisions don't affect i_k.
- Applications:
  1. Optimal replacement problems.
  2. When to start a treatment to maximize patient quality of life?
  3. When to exercise an option (finance)?

A compact formulation of the MDP is as follows:

$$Q^*(i_k) = \mathbf{E}\left[g(i_k, i_{k+1}) + \gamma \min\left(c(i_{k+1}), Q^*(i_{k+1})\right)\right],$$

where $Q^*(i_k)$ is interpreted as the cost of continuing starting from state i_k. The cost of stopping is always $\mathbf{E}(g(i_k, i_{k+1}))$, so we can write the MDP in a simple form. Goal:

design an approximate Q-learning algorithm that uses basis function approximations specifically for this problem. Define Bellman operator:

$$(FQ)(i) = \sum_j p_{ij}\left[g(i,j) + \gamma \min\left(c(j), Q(j)\right)\right] \iff FQ = g + \gamma P f(Q),$$

where $g(i) = \sum_j p_{ij}\, g(i,j)$, and $f(Q)(i) = \min\{c(i), Q(i)\}$.

Proposition 12.1. *F is contraction on $\|\cdot\|_\infty$ and Q^* is the fixed point.*

Proposition 12.2. *F is contraction in the weighted Euclidean norm $\|\cdot\|_\xi$.*

Proof. For all Q, Q', we have

$$\begin{aligned}\left|FQ(i) - FQ'(i)\right| &\leqslant \gamma \sum_j p_{ij}\left|f(Q)(i) - f(Q')(i)\right| \\ &\leqslant \gamma \sum_j p_{ij}\left|Q(i) - Q'(i)\right|\end{aligned}$$

since $|\min(a,x) - \min(a,y)| \leqslant |x-y|$. Therefore $\left|FQ - FQ'\right| \leqslant \gamma P\left|Q - Q'\right|$ componentwise, so $\|FQ - FQ'\|_\xi \leqslant \gamma \|P\left|Q - Q'\right|\|_\xi \leqslant \gamma\|Q - Q'\|_\xi$ by the non-expansiveness of P. □

Suppose that we take a basis function approach. Consider the algorithm $\Phi r_{k+1} = (\Pi F)(\Phi r_k)$, which has a fixed point since ΠF is a contraction. Equivalently,

$$r_{k+1} = \arg\min_r \|\Phi r - \left(g + \gamma P f(\Phi r_k)\right)\|_\xi^2.$$

12.1.1 SGD Approach

Sample $i_0 \sim \xi$ and $i_{k+1} \sim P_{i_k}$. Consider this update:

$$r_{k+1} = r_k + \alpha_k\, \Phi(i_k)\left[g(i_k) + \gamma f(\Phi r_k)(i_{k+1}) - (\Phi r_k)(i_k)\right]$$

Some intuition (thought experiment):

- Suppose $\Phi(i_k) > 0$, then increasing $r(1)$ will increase $Q \approx \Phi r$.
- If new estimate of Q − current estimate of $Q > 0$, then the current estimate is too small. We should increase $r(1)$ to increase the estimate. This is exactly what the update will do.

Define z and $\bar{z}$:

$$\begin{aligned} z(i_k, i_{k+1}, r_k) &= \Phi(i_k)\big(g(i_k) + \gamma f(\Phi r_k)(i_{k+1}) - (\Phi r_k)(i_k)\big), \\ \bar{z}(r_k) &= \mathbf{E}\big[z(i_k, i_{k+1}, r_k)\big] = \begin{pmatrix} \bar{z}(r_1) \\ \bar{z}(r_2) \\ \vdots \\ \bar{z}(r_n) \end{pmatrix}, \end{aligned}$$

so that $r_{k+1} = r_k + \alpha_k z(i_k, i_{k+1}, r_k)$. The j^{th}-component of $\bar{z}(r_k)$ is

$$\begin{aligned} \bar{z}_j(r_k) &= \mathbf{E}\Big[\Phi_j(i_k)\big[g(i_k) + \gamma f(\Phi r)(i_k) - (\Phi r)(i_k)\big]\Big] \\ &= \mathbf{E}\Big[\Phi_j(i_k)\big[g(i_k) + \gamma P\, f(\Phi r)(i_k) - (\Phi r)(i_k)\big]\Big] \\ &= \mathbf{E}\Big[\Phi_j(i_k)\big[(F\Phi r)(i_k) - (\Phi r)(i_k)\big]\Big] \\ &= \sum_i \xi_i \Big[\Phi_j(i)\big[(F\Phi r)(i) - (\Phi r)(i)\big]\Big] \\ &= \langle \Phi_j, F\Phi r - \Phi r\rangle_\xi \end{aligned}$$

Lemma 12.3. $(r - r^*)^T\, \bar{z}(r) < 0$ *for all* $r \neq r^*$*, i.e., the update direction is correct on average.*

Proof. Note that:

$$\begin{aligned} (r - r^*)^T \bar{z}(r) &= \sum_{j=1}^{m} (r(j) - r^*(j))\ \langle \Phi_j, F\Phi r - \Phi r\rangle_\xi \\ &= \sum_{j=1}^{m} (r(j) - r^*(j))\ \sum_{i=1}^{n} \xi_i \Phi_j(i)\big[(F\Phi r)(i) - (\Phi r)(i)\big] \\ &= \sum_i \xi_i \Big[\sum_j (r(j) - r^*(j))\ \Phi_j(i)\big[(F\Phi r)(i) - (\Phi r)(i)\big]\Big] \\ &= \sum_i \xi_i \big[(\Phi r)(i) - (\Phi r^*)(i)\big]\big[(F\Phi r)(i) - (\Phi r)(i)\big]\Big] \\ &= \langle \Phi r - \Phi r^*, F\Phi r - \Phi r\rangle_\xi \\ &= \langle \Phi r - \Phi r^*, F\Phi r - \Pi F\Phi r + \Pi F \Phi r - \Phi r\rangle_\xi \\ &= \langle \Phi r - \Phi r^*, \Pi F\Phi r - \Phi r\rangle_\xi \\ &= \langle \Phi r - \Phi r^*, \Pi F\Phi r - \Pi F\Phi r^* + \Phi r^* - \Phi r\rangle_\xi \\ &= \langle \Phi r - \Phi r^*, \Pi F\Phi r - \Pi F\Phi r^*\rangle_\xi - \|\Phi r - \Phi r^*\|_\xi^2 \\ &\leqslant \gamma\, \|\Phi r - \Phi r^*\|_\xi\ \|\Phi r - \Phi r^*\|_\xi - \|\Phi r - \Phi r^*\|_\xi^2 \end{aligned}$$

$$= \underbrace{(\gamma - 1)}_{<0} \underbrace{\|\Phi r - \Phi r^*\|_\xi^2}_{>0} < 0$$

where we used the Cauchy Schwarz inequality in the next to last line. □

Theorem 12.4. *Under some additional conditions (see Tsitsiklis, Van Roy* 1999*),*

- $r_k \;\to\; r^*$ *w.p. 1,*
- $\|\Phi r^* - Q^*\|_\xi \leqslant \dfrac{1}{\sqrt{1-\gamma^2}} \|\Pi Q^* - Q^*\|_\xi$*,*
- *Let* $J^*(i_0) = \min(c(i_0), Q^*(i_0))$ *and let* $J^{\Phi r^*}(i_0)$ *be the cost obtained by following policy induced by* Φr^**. Then,*

$$\mathbf{E}\big[J^{\Phi r^*}(i_0)\big] - \mathbf{E}\big[J^*(i_0)\big] \leqslant \frac{2}{(1-\gamma)\sqrt{1-\gamma^2}} \|\Pi Q^* - Q^*\|_\xi$$

where $i_0 \sim \xi$*.*

First part follows by SGD theorem and the above lemma. Second part is similar to the policy evaluation theorem from a previous lecture. We now prove the third part. Define a new operator:

$$(HQ)(i) = \begin{cases} c(i), & \text{if } c(i) \leqslant (\Phi r^*)(i) \\ Q(i), & \text{otherwise.} \end{cases}$$

Interpretation: take Φr^*'s recommended decision, but evaluate cost-to-go using Q. Define:

$$\tilde{F}Q = g + \gamma P\, HQ.$$

Lemma 12.5. $\|\tilde{F}Q - \tilde{F}Q'\|_\xi$ *is* γ*-contraction in* $\|.\|_\xi$*.*

Proof. There are two cases:

$$(HQ)(i) = \begin{cases} HQ - HQ' = 0, & \text{if } c(i) \leqslant (\Phi r^*)(i), \\ HQ - HQ' = Q - Q', & \text{if } c(i) > (\Phi r^*)(i). \end{cases}$$

Therefore, we have $\|\tilde{F}Q - \tilde{F}Q'\|_\xi \leqslant \gamma \|HQ - HQ'\|_\xi \leqslant \|Q - Q'\|_\xi$. □

Lemma 12.6. $\tilde{Q} = g + \gamma P\, J^{\Phi r^*}$ *is a fixed point of* $\tilde{F}$*.*

Proof. We have:

$$(H\tilde{Q})(i) = \begin{cases} c(i), & \text{if } c(i) \leqslant (\Phi r^*)(i), \\ (g + \gamma P\, J^{\Phi r^*})(i), & \text{otherwise.} \end{cases}$$

This means $H\tilde{Q} = J^{\Phi r^*}$ and the result $\tilde{F}\tilde{Q} = \tilde{Q}$ follows. □

Remark 12.7. $\tilde{F}(\Phi r^*) = F(\Phi r^*)$.

Proof. We have:

$$(H\Phi r^*)(i) = \begin{cases} c(i), & \text{if } c(i) \leqslant (\Phi r^*)(i), \\ (\Phi r^*)(i), & \text{otherwise.} \end{cases}$$

Therefore, $(H\Phi r^*)(i) = \min(c(i), (\Phi r^*)(i))$ and $(\tilde{F}Qr^*)(i) = g(i) + \gamma(P\, f(\Phi r^*))(i) = (F\Phi r^*)(i)$. □

(iii). Note $i_0 \sim \xi$ and $\xi = \xi P$.

$$\begin{aligned}
\mathbf{E}\big[J^{\Phi r^*}(i_0) - J^*(i_0)\big] &\leqslant \big|\, \mathbf{E}\big[(PJ^{\Phi r^*})(i_0) - (PJ^*)(i_0)\big]\, \big| \\
&= \sqrt{\left|\sum_{i=1}^{n} \xi_i \big[(PJ^{\Phi r^*})(i) - (PJ^*)(i)\big]\right|^2} \\
&\leqslant \sqrt{\sum_{i=1}^{n} \xi_i \big(PJ^{\Phi r^*} - PJ^*\big)^2} \\
&= \|PJ^{\Phi r^*} - PJ^*\|_\xi = \frac{1}{\gamma}\, \|g + \gamma PJ^{\Phi r^*} - (g + \gamma PJ^*)\|_\xi \\
&= \frac{1}{\gamma}\, \|\tilde{Q} - Q^*\|_\xi,
\end{aligned}$$

where we used convexity and Jensen's inequality in the third line. Now,

$$\begin{aligned}
\|\tilde{Q} - Q^*\|_\xi &\leqslant \|Q^* - F\Phi r^* + \tilde{F}\Phi r^* - \tilde{Q}\|_\xi \\
&\leqslant \|Q^* - F\Phi r^*\|_\xi + \|\tilde{F}\Phi r^* - \tilde{F}\tilde{Q}\|_\xi \\
&\leqslant \gamma\, \|Q^* - \Phi r^*\|_\xi + \gamma\, \|\tilde{Q} - \Phi r^*\|_\xi \\
&\leqslant \gamma\, \|Q^* - \Phi r^*\|_\xi + \gamma\, \|\tilde{Q} - Q^*\|_\xi + \gamma\, \|Q^* - \Phi r^*\|_\xi.
\end{aligned}$$

Therefore:

$$\|Q^* - \tilde{Q}\|_\xi \leqslant \frac{2\gamma}{1-\gamma}\, \|Q^* - \Phi r^*\|_\xi \leqslant \frac{2\gamma}{1-\gamma}\, \frac{1}{\sqrt{1-\gamma^2}}\, \|\Pi Q^* - Q^*\|_\xi.$$

Combining with the previous step, we have

$$\mathbf{E}\Big[J^{\Phi r^*}(i_0) - J^*(i_0)\Big] \leqslant \frac{2}{(1-\gamma)\sqrt{1-\gamma^2}}\, \|\Pi Q^* - Q^*\|_\xi.$$

□