Lecture 5: Optimal Policies and Value Functions

Course: Reinforcement Learning Theory
Instructor: Lei Ying
Department of EECS
University of Michigan, Ann Arbor

Optimal policy

Define

$$\begin{aligned}
J_0^N(i) &= \max_{\mu_0,\ldots,\mu_N} E\left[\sum_{k=0}^{N} \alpha^k r(x_k,u_k)\middle| x_0 = i\right] \\
&= \max_{\mu_0,\ldots,\mu_N} E\left[r(x_0,u_0) + \sum_{k=1}^{N} \alpha^k r(x_k,u_k)\middle| x_0 = i\right] \\
&= \max_{\mu_0}\Bigg\{E[r(x_0,u_0)|x_0 = i] + \max_{\mu_1,\ldots,\mu_N} E\left[\sum_{k=1}^{N} \alpha^k r(x_k,u_k)\middle| x_1 = j\right] \\
&\qquad\qquad \times P_{ij}\left(\mu_0(x_0)|x_0 = i\right)\Bigg\} \\
&= \max_{\mu_0} E[r(x_0,u_0)|x_0 = i] + \alpha \sum_j P_{ij}(\mu_0(i)) J_1^N(j).
\end{aligned}$$

Optimal policy

The Bellman Equation for Finite-Horizon MDP

$$J_k^N(i) = \max_{\mu_k} E[r(x_k, u_k)|x_k = i] + \alpha \sum_j P_{ij}(\mu_k(i)) J_{k+1}^N(j)$$

- Extend this result to the infinite-horizon scenario $(N \to \infty)$

Remark:
We hope $J_k^\infty(i) = J_{k+1}^\infty(i)$. The optimal cost-to-go should not depend on time k because both $k \to \infty$ and $k+1 \to \infty$ include infinite time steps.

Optimal policy

Theorem 1

Let

$$J^*(i) = \sup_{\mu_0, \mu_1, \dots} \lim_{N \to \infty} E\left[\sum_{k=0}^{\infty} \alpha^k r(x_k, u_k) \middle| x_0 = i\right]$$

Then, $J^*(i)$ satisfies

$$J^*(i) = \max_u E[r(i,u) + \alpha J^*(x_1) | x_0 = i, u_0 = u]$$

or

$$J^*(i) = \max_u E[r(i,u)] + \alpha \sum_j P_{ij}(u) J^*(j)$$

Optimal policy

Proof: Define $\mu = \{\mu_0, \mu_1, \dots\}$ (possibly randomized policy).

Define $\mu^k = \{\mu_k, \mu_{k+1}, \dots\}$: policy starting from time k.

$$J_\mu(i) = E\left[r(i, \mu_0(i)) + \alpha \sum_j P_{ij}(\mu_0(i)) J_{\mu^1}(j) \middle| x_0 = i \right]$$

Recall that $J^*(j) \geq J_\mu(j) \quad \forall \mu$. So, for any μ,

$$\begin{aligned} J_\mu(i) &\leq E[r(i, \mu_0(i))] + \alpha \sum_j P_{ij}(\mu_0(i)) J^*(j) \\ &\leq \max_u \left\{ E[r(i,u)] + \alpha \sum_j P_{ij}(u) J^*(j) \right\} \end{aligned}$$

Optimal Policy

$$
\begin{aligned}
J^*(i) &= \sup_{\mu} J_{\mu}(i) \\
&\leq \sup_{\mu} \max_{u} \left\{ E[r(i,u)] + \alpha \sum_{j} P_{ij}(u) J^*(j) \right\} \\
&= \max_{u} \left\{ E[r(i,u)] + \alpha \sum_{j} P_{ij}(u) J^*(j) \right\}.
\end{aligned}
$$

Optimal policy

Next consider for each i. Let $\mu^{(i)}$ be a policy such that

$$J_{\mu^{(i)}}(i) \geq J^*(i) - \epsilon$$

Note that $\mu^{(i)}$ exists by the definition of J^*.

At time 0, choose action u_0, so we have

$$\begin{aligned} J^*(i) &\geq E[r(i,u_0)] + \alpha \sum_j P_{ij}(u_0) J_{\mu^{(j)}} \\ &\geq E[r(i,u_0) + \alpha J^*(x_1) | x_0 = i, u_0] - \alpha\epsilon \quad \forall u_0 \end{aligned}$$

Optimal policy

We have

$$J^*(i) \geq \max_{u_0} E[r(i, u_0) + \alpha J^*(x_1) | x_0 = i, u_0] - \alpha\epsilon$$

Letting $\epsilon \to 0$, we obtain

$$J^*(i) \geq \max_{u_0} E[r(i, u_0) + \alpha J^*(x_1) | x_0 = i, u_0] \qquad \blacksquare$$

Optimal policy

Theorem 2

Let

$$T(J)(i) = \max_u E[r(i,u)] + \alpha \sum_j P_{ij}(u) J(j)$$

Then the Bellman Equation can be written as $J^* = T(J^*)$, where T is a contraction mapping with $\alpha \in [0, 1)$.

Optimal policy

Proof:

Two facts:

- If $J_1 \leq J_2$ (entry-wise), then

$$T(J_1) \leq T(J_2)$$

- Let $e = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix}$, then

$$T(J + \gamma e) = T(J) + \alpha \gamma e \quad \forall \gamma \text{ (scalar)}$$

Define $r = \max_i |J_1(i) - J_2(i)| = \|J_1 - J_2\|_\infty$, then

$$J_2 - re \leq J_1 \leq J_2 + re$$

$$T(J_2) - \alpha re \leq T(J_1) \leq T(J_2) + \alpha re$$

Optimal policy

$$\|T(J_1) - T(J_2)\|_\infty \leq \alpha r = \alpha \|J_1 - J_2\|_\infty$$

(Contraction mapping)

Thus, $J^* = T(J^*)$ has a unique solution.

Theorem 3

$$\mu^*(i) \in \arg\max_u E[r(i,u)] + \alpha \sum_j P_{ij}(u) J^*(j)$$

where $J^* = T(J^*)$. Then μ^* is an optimal policy.

Proof for Theorem 3

Proof: $J_{\mu^*} = T_{\mu^*}(J_{\mu^*}) \Leftarrow$ mapping under policy μ^*.

$$\begin{aligned} T_{\mu^*}(J^*)(i) &= E[r(i, \mu^*(i))] + \alpha \sum_j P_{ij}(\mu^*(i)) J^*(j) \\ &= \max_u E[r(i,u)] + \alpha \sum_j P_{ij}(u) J^*(j) \\ &= J^*(i). \\ \implies T_{\mu^*}(J^*) &= J^* \end{aligned}$$

J^* is a fixed point of T_{μ^*}, which is unique, so we have

$$J_{\mu^*} = J^*. \qquad \blacksquare$$

Reference

- This lecture is based on R. Srikant's lecture notes on *MDPs with discounted cost* available at `https://sites.google.com/illinois.edu/mdps-and-rl/lectures?authuser=1`

Acknowledgements: I would like to thank Alex Zhao for helping prepare the slides, and Honghao Wei and Zixian Yang for correcting typos/mistakes.