Lecture 4: Algorithms for MDP

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

Infinite-Horizon Discounted MDP

- Define the value function of a given policy μ

$$J_\mu(i) = \lim_{N\to\infty} E\left[\sum_{k=0}^{N} \alpha^k r(x_k, \mu(x_k)) \middle| x_0 = i\right]$$

- Note that $a_N = \sum_{k=0}^{N} \alpha^k r(x_k, \mu(x_k))$ is an increasing, upper-bounded sequence, so it has a finite limit. Therefore, $J_\mu(i)$ is well-defined.
- Assume the MC under policy μ is irreducible and aperiodic (thus it has a unique stationary distribution).

Infinite-Horizon Discounted MDPs

Then,

$$
\begin{aligned}
J_\mu(i) &= \lim_{N\to\infty} E\left[r(i,\mu(i)) + \sum_{k=1}^{N} \alpha^k r(x_k,\mu(x_k)) \middle| x_0 = i \right] \\
&= \bar{r}(i,\mu(i)) + \sum_j P_{ij}(\mu(i)) \lim_{N\to\infty} E\left[\sum_{k=1}^{N} \alpha^k r(x_k,\mu(x_k)) \middle| x_1 = j \right] \\
&= \bar{r}(i,\mu(i)) + \\
&\sum_j P_{ij}(\mu(i)) \alpha \lim_{N\to\infty} E\left[r(j,\mu(j)) + \sum_{k=2}^{N} \alpha^{k-1} r(x_k,\mu(x_k)) \middle| x_1 = j \right] \\
&= \bar{r}(i,\mu(i)) + \alpha \sum_j P_{ij}(\mu(i)) J_\mu(j).
\end{aligned}
$$

Infinite-Horizon Discounted MDPs

Lemma: $J_\mu(i)$ satisfies the following Bellman equation:

$$J_\mu(i) = \bar{r}(i, \mu(i)) + \alpha \sum_j P_{ij}(\mu(i)) J_\mu(j), \quad \forall i \qquad \text{(A)}$$

For convenience, let $P_{ij} = P_{ij}(\mu(i))$.

Theorem: There exists a unique $J_\mu = \begin{pmatrix} J_\mu(1) \\ J_\mu(2) \\ \vdots \end{pmatrix}$ which satisfies (A).

Proof 1: Based on norm.

$$J_\mu = r_\mu + \alpha P_\mu J_\mu$$

where $r_\mu = \begin{pmatrix} E[r(1,\mu(1))] \\ E[r(2,\mu(2))] \\ \vdots \end{pmatrix}$ and P_μ is the probability transition matrix.

Then,

$$r_\mu = (I - \alpha P_\mu) J_\mu$$

And $J_\mu = (I - \alpha P_\mu)^{-1} r_\mu$ if $(I - \alpha P_\mu)$ is invertible. So, it must be that $(I - \alpha P_\mu)$ has no eigenvalue of 0.

A sufficient condition for this is that the eigenvalues of P_μ have magnitude ≤ 1, i.e. $|\lambda(P_\mu)| \leq 1$

Infinite-Horizon Discounted MDPs

We have

$$\max_i |\lambda_i(P)| \leq \max_i \sum_j |P_{ij}|$$

Since P_μ is a probability transition matrix, $\sum_j P_{ij}(\mu) = 1$.

We conclude that $J_\mu = (I - \alpha P_\mu)^{-1} c_\mu$. ■

Infinite-Horizon Discounted MDPs

Contraction Mapping Theorem

Let T be a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$. Assume T is a contraction mapping:

$$\|T(x) - T(y)\| \leq \alpha \|x - y\| \quad \forall x, y \in \mathbb{R}^n$$

where $\alpha \in [0, 1)$ and $\| \cdot \|$ is some norm. Then,

(a) There exists a unqiue x^* such that

$$x^* = T(x^*) \quad \text{(fixed point)}$$

(b) The iteration $X_{k+1} = T(X_k)$ converges to x^* from any $X_0 \in \mathbb{R}^n$

Infinite-Horizon Discounted MDPs

Proof 2: Contraction

Define $T_\mu(J_\mu) = r_\mu + \alpha P J_\mu$. Then

$$\begin{aligned}
T_\mu(x) - T_\mu(y) &= \alpha P(x-y) \\
\|T_\mu(x) - T_\mu(y)\|_\infty &= \alpha \max_i |P(x-y)|_i \\
&= \alpha \max_i |\sum_j P_{ij}(x_j - y_j)| \\
&\leq \alpha \max_i \sum_j P_{ij} \underbrace{\max_j |x_j - y_j|}_{\|x-y\|_\infty} \\
&= \alpha \max_i \sum_j P_{ij} \|x-y\|_\infty
\end{aligned}$$

Infinite-Horizon Discounted MDPs

(Cont'd)

$$\underbrace{\alpha \max_i \sum_j P_{ij}}_{=1} \|x - y\|_\infty = \alpha\|x - y\|_\infty$$

Thus T_μ is a contraction mapping $\implies J_\mu = T_\mu(J_\mu)$ has a unique solution. ■

Proof of contraction mapping theorem

Contraction Mapping Theorem

Let $T : D \in \mathbb{R}^n \to D$ such that

(i) D is closed

(ii) $\|T(x) - T(y)\| \leq \alpha \|x - y\|$ for some $\alpha \in [0, 1)$

Then, there exists a unique x^* such that $T(x^*) = x^*$ and

$$\lim_{k \to \infty} T^k(x_0) = x^* \quad \forall x_0.$$

Proof:

Fix x_0 and define $x_1 = T(x_0), x_2 = T(x_1) = T^2(x_0), \ldots$

$$\|x_n - x_{n+l}\| \leq \|x_n - x_{n+1}\| + \cdots + \|x_{n+l-1} - x_{n+l}\|$$

$$\begin{aligned} \|x_{n+1} - x_n\| &= \|T(x_n) - T(x_{n-1})\| \\ &\leq \alpha \|x_n - x_{n-1}\| \leq \cdots \leq \alpha^n \|x_1 - x_0\| \end{aligned}$$

Proof of contraction mapping theorem

Thus,

$$\begin{aligned}\|x_n - x_{n+l}\| &\leq (\alpha^n + \alpha^{n+1} + \cdots + \alpha^{n+l-1})\|x_1 - x_0\| \\ &\leq \alpha^n(1 + \alpha + \cdots + \alpha^{l-1})\|x_1 - x_0\| \\ &\leq \frac{\alpha^n}{1-\alpha}\|x_1 - x_0\| \quad \text{(independent of } l\text{)}\end{aligned}$$

Given $\epsilon > 0, \exists N_\epsilon$ such that

$$\begin{aligned} &\|x_n - x_{n+l}\| \leq \epsilon \quad \forall n \geq N_\epsilon \text{ and } l \geq 1 \\ \implies &\|x_n - x_m\| \leq \epsilon \quad \forall n, m \geq N_\epsilon\end{aligned}$$

Proof of contraction mapping theorem

Therefore, X_n is a Cauchy sequence, so it converges to x^*. Since D is closed, $x^* \in D$.

By continuity of T (since T is a contraction),

$$x^* = \lim_{n\to\infty} x_n = \lim_{n\to\infty} T(x_{n-1}) = T(\lim_{n\to\infty} x_{n-1})$$

$$x^* = T(x^*)$$

Suppose $y^* \neq x^*$ such that $T(y^*) = y^*$. Then,

$$\|y^* - x^*\| = \|T(y^*) - T(x^*)\| \leq \alpha \|y^* - x^*\|$$

But $\alpha < 1$, so it must be

$$\|y^* - x^*\| = 0$$

$$y^* = x^*$$

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.