Convergence and Optimality of Model-Based Solutions

Infinite-Horizon Discounted MDP

Define the value function of a given policy μ

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

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, $V_\mu(i)$ is well-defined.

Infinite-Horizon Discounted MDPs

Then, it satifies the following Bellman equation

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

Infinite-Horizon Discounted MDPs

Contraction Mapping Theorem

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

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

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

1. There exists a unique x^* such that

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

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

We will prove this theorem later.

Infinite-Horizon Discounted MDPs

Uniqueness (for a given policy)

Given a stationary policy μ, there exists a unique $V = \begin{pmatrix} V(1) \\ V(2) \\ \vdots \end{pmatrix}$ which satisfies the following Bellman equation

$$V(i) = \bar{r}(i, \mu(i)) + \alpha \sum_j P_{ij}(\mu(i)) V(j), \quad \forall i \tag{1}$$

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

Infinite-Horizon Discounted MDPs

Proof of Uniqueness

Define $T_\mu(V) = \bar{r}_\mu + \alpha PV$. 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

Proof of Uniqueness (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 $\implies$ $V = T_\mu(V)$ has a unique solution. ■

Proof of contraction mapping theorem (Convergence)

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

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

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

Thus,

$$\begin{aligned} \|x_{n+l} - x_n\| &\leq (\alpha^{n+l-1} + \alpha^{n+l-2} + \cdots + \alpha^n)\|x_1 - x_0\| \\ &\leq \alpha^n(1 + \alpha + \ldots)\|x_1 - x_0\| \\ &\leq \frac{\alpha^n}{1-\alpha}\|x_1 - x_0\| \quad \text{(independent of } l\text{)} \end{aligned}$$

Proof of contraction mapping theorem (Convergence)

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

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

x_n is a Cauchy sequence (terms get arbitrarily close to each other as n increases). A Cauchy sequence in R^S converges to some limit, denoted by x^*.

Proof of contraction mapping theorem (Fixed point)

Now, from the convergence, we have

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

Since T is a contraction mapping,

$$\|T(x_n) - T(x^*)\| \leq \alpha \|x_n - x^*\|.$$

From the convergence, we have

$$0 \leq \lim_{n\to\infty} \|T(x_n) - T(x^*)\| \leq \alpha \lim_{n\to\infty} \|x_n - x^*\| = 0.$$

Therefore,

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

Proof of contraction mapping theorem (Uniqueness)

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

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

$$y^* = x^*$$

Value Iteration: Existence, Uniqueness and Convergence

Value Function

$$V^*(i) = \sup_{\mu_0,\mu_1,\ldots} E\left[\sum_{k=0}^{\infty} \alpha^k r(x_k, \mu_k(x_k)) \middle| x_0 = i\right].$$

- Theorem 1: Prove the Bellman equation for V^*
- Theorem 2: Prove that the value iteration is a contraction mapping
- Theorem 3: Show that an optimal policy can be obtained from V^*.

Value Iteration: Existence, Uniqueness and Convergence

Theorem 1: The Bellman Equation for V^*

$V^*(i)$ satisfies

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

or

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

Value Iteration: Existence, Uniqueness and Convergence

Proof of Theorem 1

Define

- $\mu = \{\mu_0, \mu_1, \dots\}$ where μ_k is a function of past history and current state

$$u_k = \mu_k(x_0, \dots, x_k, u_0, \dots, u_{k-1}, r_0, \dots, r_{k-1})$$

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

By definition, we have

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

Value Iteration: Existence, Uniqueness and Convergence

Proof of Theorem 1

Note that $V^*(j) \geq V_{\mu^1}(j) \quad \forall j$. So

$$V_\mu(i) \leq E[r(i, \mu_0(i))] + \alpha \sum_j P_{ij}(\mu_0(i)) V^*(j)$$

$$\leq \max_u \left\{ E[r(i,u)] + \alpha \sum_j P_{ij}(u) V^*(j) \right\}$$

Value Iteration: Existence, Uniqueness and Convergence

Proof of Theorem 1

Taking sup on both sides yields

$$
\begin{aligned}
V^*(i) &= \sup_{\mu} V_\mu(i) \leq \sup_{\mu} \max_{u} \left\{ E[r(i,u)] + \alpha \sum_j P_{ij}(u) V^*(j) \right\} \\
&= \max_{u} \left\{ E[r(i,u)] + \alpha \sum_j P_{ij}(u) V^*(j) \right\} \\
&= \max_{u_0} E[r(i,u_0) + \alpha V^*(x_1) | x_0 = i, u_0]
\end{aligned}
$$

Value Iteration: Existence, Uniqueness and Convergence

Proof of Theorem 1

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

$$V_{\mu^{(i)}}(i) \geq V^*(i) - \epsilon$$

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

At time 0, choose action u_0 and then follow policy $\mu^{(j)}$ if in state j

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

Value Iteration: Existence, Uniqueness and Convergence

Proof of Theorem 1

We have

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

Letting $\epsilon \to 0$, we obtain

$$V^*(i) \geq \max_{u_0} E[r(i, u_0) + \alpha V^*(x_1) | x_0 = i, u_0] \tag{2}$$

Value Iteration: Existence, Uniqueness and Convergence

Theorem 2: Contraction Mapping

Let

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

Then the value iteration algorithm can be written as

$$V_{k+1} = T(V_k)$$

and the Bellman Equation can be written as

$$V^* = T(V^*).$$

T is a contraction mapping given $\alpha \in [0, 1)$.

Value Iteration: Existence, Uniqueness and Convergence

Proof of Theorem 2

Two facts that are easy to prove

- Fact 1: If $V_1 \leq V_2$ (entry-wise), then

$$T(V_1) \leq T(V_2)$$

- Fact 2: Let $e = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix}$, then

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

Value Iteration: Existence, Uniqueness and Convergence

Proof of Theorem 2

Define $b = \max_i |V_1(i) - V_2(i)| = \|V_1 - V_2\|_\infty$, then

$$V_2 - be \leq V_1 \leq V_2 + be$$

$$T(V_2 - be) \leq T(V_1) \leq T(V_2 + be) \quad \text{(Fact 1)}$$

$$T(V_2) - \alpha be \leq T(V_1) \leq T(V_2) + \alpha be \quad \text{(Fact 2)}$$

$$\|T(V_1) - T(V_2)\|_\infty \leq \alpha b = \alpha \|V_1 - V_2\|_\infty$$

(Contraction mapping)

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

Value Iteration: Existence, Uniqueness and Convergence

Theorem 3: Obtain an Optimal Policy

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

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

Value Iteration: Existence, Uniqueness and Convergence

Proof of Theorem 3

Considering stationary policy μ^*,

$$\begin{aligned} T_{\mu^*}(V^*)(i) &= E[r(i,\mu^*(i))] + \alpha \sum_j P_{ij}(\mu^*(i)) V^*(j) \\ &= \max_u E[r(i,u)] + \alpha \sum_j P_{ij}(u) V^*(j) = V^*(i). \quad \text{(Theorem 1)} \end{aligned}$$

$$\implies T_{\mu^*}(V^*) = V^*$$

Therefore, V^* is a fixed point of T_{μ^*}. Since T_{μ^*} has a unique fixed point,

$$V_{\mu^*} = V^*.$$

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`