Computer Engineering Department

Value-based Theoretical Guarantees

Mohammad Hossein Rohban, Ph.D.

Spring 2025

Courtesy: Most of slides are adopted from ML course EE3001 by Jie Wang.

Bellman's Optimality Equation

- Assume a stochastic reward function.

$$\Pr(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a), \forall\, s, s' \in \mathcal{S}, r \in \mathcal{R}, a \in \mathcal{A},$$

which is abbreviated by $p(s', r|s, a)$.

$$\begin{aligned} q_*(s, a) &= \max_\pi \mathbb{E}[G_t | S_t = s, A_t = a] \\ &= \max_\pi \mathbb{E}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= \mathbb{E}[R_{t+1} | S_t = s, A_t = a] + \gamma \max_\pi \mathbb{E}[G_{t+1} | S_t = s, A_t = a]. \end{aligned}$$

Bellman's Optimality Equation (cont.)

$$\mathbb{E}[R_{t+1}|S_t = s, A_t = a] = \sum_r r \sum_{s'} p(s', r|s, a).$$

$$\begin{aligned}
\mathbb{E}[G_{t+1}|S_t = s, A_t = a] &= \sum_{s',a'} p(s', a'|s, a)\mathbb{E}[G_{t+1}|S_{t+1} = s', A_{t+1} = a', S_t = s, A_t = a] \\
&= \sum_{s',a'} p(s'|s, a)p(a'|s', s, a)\mathbb{E}[G_{t+1}|S_{t+1} = s', A_{t+1} = a'] \\
&= \sum_{s',a'} p(s'|s, a)\pi(a'|s')q_\pi(s', a') \\
&= \sum_{s'} p(s'|s, a) \sum_{a'} \pi(a'|s')q_\pi(s', a').
\end{aligned}$$

Bellman's Optimality Equation (cont.)

$$q_*(s,a) = \sum_r r \sum_{s'} p(s', r|s,a) + \gamma \max_\pi \sum_{s'} p(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a').$$

$$q_*(s,a) = \sum_r r \sum_{s'} p(s', r|s,a) + \gamma \max_\pi \sum_{s'} p(s'|s,a) \max_{a'} q_\pi(s',a').$$

Bellman's Optimality Equation (cont.)

$$
\begin{aligned}
q_*(s,a) &= \sum_r r \sum_{s'} p(s', r|s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} q_*(s',a') \\
&= \sum_{r,s'} p(s', r|s,a)(r + \gamma \max_{a'} q_*(s',a')).
\end{aligned}
$$

Questions

- Does there exist q_* functions satisfying the Bellman's Eq.?
- Is this function unique?
- Can value iteration find this function?

Fixed Point

- For an operator T, we call x a fixed point if $Tx = x$.
- q_* is a fixed point of the Bellman's Eq.
- Why?

Fixed Point (cont.)

Theorem 1 (Banach Fixed Point Theorem). *Suppose that X is a nonempty complete metric space and $T : X \to X$ is a contraction mapping on X. Then T has a unique fixed point.*

Definition 1 (Contraction Mapping). [1] Let (X, d) be a metric space. A mapping $T : X \to X$ is called a *contraction mapping* on X if there is a positive real number $\alpha < 1$ such that for any $x, y \in X$

$$d(Tx, Ty) \leq \alpha d(x, y).$$

Existence Proof

- Pick an arbitrary point x_0.
- Construct a sequence: $x_k = Tx_{k-1},\ k = 1, 2, \ldots.$
- Let $C = d(x_1, x_0)$.
- Note that

$$d(x_{k+1}, x_k) \leq \alpha d(x_k, x_{k-1}) \leq \cdots \leq \alpha^k d(x_1, x_0) = \alpha^k C, \forall, k = 1, 2, \ldots.$$

$$d(x_m, x_n) \leq \sum_{m-n-1}^{m-n-1} d(x_{n+i+1}, x_{n+i}).$$

$$d(x_m, x_n) \leq \sum_{i=0}^{m-n-1} \alpha^{n+i} C = \alpha^n C \frac{1-\alpha^{m-n}}{1-\alpha} \leq \alpha^n \frac{C}{1-\alpha}.$$

Existence Proof

- Thus for any $\epsilon > 0$, if $N \geq \frac{\log \epsilon(1-\alpha) - \log C}{\log \alpha}$ then $d(x_m, x_n) \leq \epsilon.$
- Hence x_n is a Cauchy sequence.
- Therefore, it converges to a point, let's call x.
- Now, we show that x is a fixed point of T.
- Note that:

$$d(Tx, x) \leq d(Tx, x_k) + d(x_k, x) \leq \alpha d(x, x_{k-1}) + d(x_k, x), \forall\, k = 1, 2, \ldots.$$

$$d(Tx, x) = 0,$$

Uniqueness

- Proof by contradiction.
- Let *x′* be another such fixed point.
- Then, $d(x, x') = d(Tx, Tx') \leq \alpha d(x, x'),$
- Which is a contradiction.

Application to the Bellman's Eq.

- Define the operator *T* as:

$$Tq(s,a) = \sum_{r,s'} p(r,s'|s,a)(r + \gamma \max_{a'} q(s',a')),$$

T in Bellman is contraction

Lemma 1. *For a finite MDP, the mapping T in Eq. (10) is a contraction mapping.*

Proof. We consider the complete metric space $(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, d)$, where $d(q_1, q_2) = \|q_1 - q_2\|_\infty$ for any $p, q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. Then,

$$
\begin{aligned}
\|Tq_1 - Tq_2\|_\infty &= \max_{s,a} |Tq_1(s,a) - Tq_2(s,a)| \\
&= \gamma \max_{s,a} \sum_{r,s'} p(r, s'|s,a) |\max_{a'} q_1(s', a') - \max_{a'} q_2(s', a')| \\
&\leq \gamma \max_{s,a} \sum_{s'} p(s'|s,a) \max_{a'} |q_1(s', a') - q_2(s', a')| \\
&\leq \gamma \max_{s,a} \max_{s'} \max_{a'} |q_1(s', a') - q_2(s', a')| \\
&= \gamma \max_{s',a'} |q_1(s', a') - q_2(s', a')| \\
&= \gamma \|q_1 - q_2\|_\infty,
\end{aligned}
$$

Why value iteration converges to the fixed point?

- Let’s discuss!

Policy Improvement Improves!

- If we set the new policy to maximize q(s, a) over a, the new policy leads to higher v(s) values for all states s.
- Let’s discuss!

Policy Iteration Converges

Theorem. Policy iteration is guaranteed to converge and at convergence, the current policy and its value function are the optimal policy and the optimal value function!

Proof sketch:

(1) *Guarantee to converge*: In every step the policy improves. This means that a given policy can be encountered at most once. This means that after we have iterated as many times as there are different policies, i.e., (number actions)$^{\text{(number states)}}$, we must be done and hence have converged.

(2) *Optimal at convergence*: by definition of convergence, at convergence $\pi_{k+1}(s) = \pi_k(s)$ for all states s. This means $\forall s \;\; V^{\pi_k}(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_i^{\pi_k}(s') \right]$

Hence V^{π_k} satisfies the Bellman equation, which means V^{π_k} is equal to the optimal value function V*.