Algorithmic and Theoretical Foundations of RL

MDP and Bellman Optimality

Ke Wei
School of Data Science
Fudan University

Table of Contents

Markov Decision Process and Bellman Equations

Bellman Optimality Equations

Markov Chain

Definition 1 (Markov Chain)
Let $\mathcal{S} = \{s_1, \cdots, s_n\}$ be a finite state space. The discrete-time dynamic system $(s_t)_{t\in\mathbb{N}} \in \mathcal{S}$ is a Markov chain if it satisfies the Markov property:

$$P\left(s_{t+1} = s \mid s_t, s_{t-1}, \ldots, s_0\right) = P\left(s_{t+1} = s \mid s_t\right).$$

Transition Matrix:

$$P = \begin{bmatrix} p_{s_1 s_1} & p_{s_1 s_2} & \cdots & p_{s_1 s_n} \\ p_{s_2 s_1} & p_{s_2 s_2} & \cdots & p_{s_2 s_n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s_n s_1} & p_{s_n s_2} & \cdots & p_{s_n s_n} \end{bmatrix}, \quad \text{where } p_{s_i s_j} = P\left(s_{t+1} = s_j | s_t = s_i\right).$$

- It is easy to see that $P\mathbf{1} = \mathbf{1}$.

Stationary Distribution of Markov Chain

Given an initial visiting probability vector x_0 over the state space $\mathcal{S}$, it is not hard to see that

$$x_t = (P^T)^t x_0$$

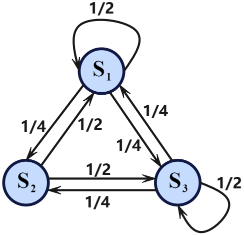
is visiting probability vector at time t. The limit (exists under mild conditions)

$$x = \lim_{t \to \infty} x_t$$

is known as the stationary distribution of Markov chain. It is easy to show that

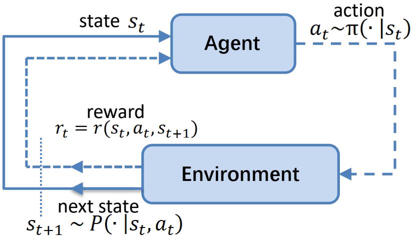
$$P^T x = x.$$

Illustrative Example



$$P = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}, \quad x = \begin{bmatrix} 2/5 \\ 1/5 \\ 2/5 \end{bmatrix}.$$

Markov Decision Process (MDP)

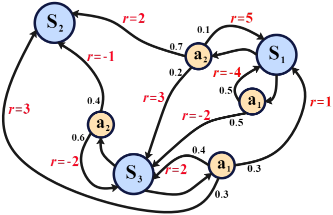

state s_t
Agent
action
$a_t \sim \pi(\cdot|s_t)$
reward
$r_t = r(s_t, a_t, s_{t+1})$
Environment
next state
$s_{t+1} \sim P(\cdot|s_t, a_t)$

Markov chain augmented with decision and reward: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$

- $\mathcal{S}$: state space (状态空间)
- $P(\cdot|s, a)$: state transition model (状态转移模型)
- $\gamma \in [0, 1]$: discount factor (折扣因子)
- $\mathcal{A}$: action space (动作空间)
- $r(s, a, s')$: immediate reward (即时奖励)
- $\pi(\cdot|s) : \mathcal{A} \to \Delta$ (策略)

Unless otherwise stated, we typically assume $|\mathcal{S}| < \infty$, $|\mathcal{A}| < \infty$, and r is bounded. For simplicity, we consider the case where r is fully determined by (s, a, s') (i.e, no randomness in r).

Illustrative Example



- three states: $\mathcal{S} = \{s_1, s_1, s_3\}$
- two actions: $\mathcal{A} = \{a_1, a_2\}$

Each edge is associated with a transition probability and a reward.

For instance, we can observe that:

$$P(s_3|s_3, a_2) = 0.6, \; P(s_2|s_3, a_2) = 0.4,$$
$$r(s_3, a_2, s_3) = -2, \; r(s_3, a_2, s_2) = -1.$$

Assume s_2 will always transfers to s_2 with reward 0 no matter what action is taken.

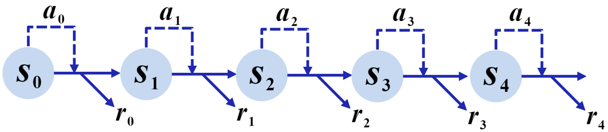
Tabular Setting

At state s_i take action a_j, transits to state $s_{ij} \in \{s_1, \cdots, s_{|\mathcal{S}|}\}$ and receive reward r_{ij}.

	s_1	s_2	$\cdots$	$s_{\vert\mathcal{S}\vert}$
a_1	(s_{11}, r_{11})	(s_{21}, r_{21})	$\cdots$	$(s_{\vert\mathcal{S}\vert 1}, r_{\vert\mathcal{S}\vert 1})$
a_2	(s_{12}, r_{12})	(s_{22}, r_{22})	$\cdots$	$(s_{\vert\mathcal{S}\vert 2}, r_{\vert\mathcal{S}\vert 2})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_{\vert\mathcal{A}\vert}$	$(s_{1\vert\mathcal{A}\vert}, r_{1\vert\mathcal{A}\vert})$	$(s_{2\vert\mathcal{A}\vert}, r_{2\vert\mathcal{A}\vert})$	$\cdots$	$(s_{\vert\mathcal{S}\vert\vert\mathcal{A}\vert}, r_{\vert\mathcal{S}\vert\vert\mathcal{A}\vert})$

Setting that should be kept in mind for theoretical discussion. A terminal state can exist.

State Value and Action Value



- Trajectory (轨迹):

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, s_3, \cdots), \quad r_t = r(s_t, a_t, s_{t+1}).$$

- Given s_0, the probability of trajectory τ is given by

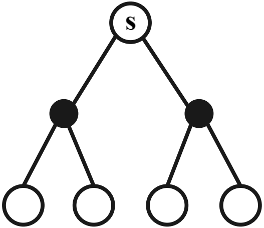
$$P_{s_0}^{\pi}(\tau) = \prod_{t=0}^{\infty} \pi(a_t|s_t)P(s_{t+1}|s_t, a_t).$$

- Infinite horizon discounted return (折扣回报):

$$r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots = \sum_{t=0}^{\infty} \gamma^t r_t.$$

Here we consider infinite horizon discounted return which enable us to focus on the stationary policy. In finite horizon problems, it may be beneficial to select a different action depending on the remaining time steps which has the form $\pi(s) = (\pi_0(s), \pi_1(s), \cdots)$.

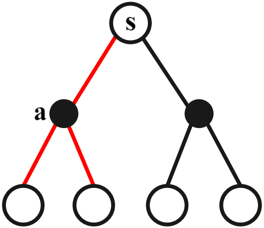
State Value and Action Value



- State value (状态价值函数):

$$V^{\pi}(s) = \mathbb{E}_{\tau \sim P_{s_0}^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} | s_0 = s \right], \quad \forall s \in \mathcal{S}.$$

State Value and Action Value



- Action value (*Q*-value, 动作价值函数):

$$Q^{\pi}(s,a) = \mathbb{E}_{\tau \sim P^{\pi}_{s_0,a_0}}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a\right], \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where the probability for $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, s_3, \cdots)$ is given by

$$P^{\pi}_{s_0,a_0}(\tau) = P(s_1|s_0,a_0)\prod_{t=1}^{\infty}\pi_t(a_t|s_t)P(s_{t+1}|s_t,a_t).$$

State Value and Action Value

- State and action values can be used to quantify goodness/badness of policies and actions.
- The relation between the state value and the action value is given by

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)} \left[Q^{\pi}(s, a) \right], \\
Q^{\pi}(s, a) &= \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s, a, s') + \gamma V^{\pi}(s') \right].
\end{aligned}
$$

- Computing the expectation seems not easy. However, the MDP structure enables us to compute the values by finding the solutions to linear systems (i.e., Bellman equations).

Remark

Action value or *Q*-value is the quantity that is primarily used in RL algorithms (but it is closely related to state value as already mentioned). Suppose we ask the question which action is the best at a state. The naive approach is to look one step forward following a fixed action, expand all the possible trajectories, compute the best reward (optimal action value) that can be obtained from that action, and then choose action that based on this reward. However, computing the best reward over all the possible trajectories starting from that action is difficult. Instead, we can look one step forward from the action and then use the reward obtained from a base policy to approximate the reward that can be obtained starting from the second step, which yields the action value. This intuition indeed underpins developments of value-based RL methods.

Bellman Equation for State Value

Theorem 1
*Given policy π, state value satisfies the following **Bellman equation**:*

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V^\pi(s')\right].$$

*Alternatively, if for any $V \in \mathbb{R}^{|\mathcal{S}|}$, define the **Bellman operator**:*

$$[\mathcal{T}^\pi V](s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V(s')\right],$$

Bellman equation can be rewritten as

$$V^\pi = \mathcal{T}^\pi V^\pi.$$

That is, V^π is a fixed point of $\mathcal{T}^\pi$.

- $\mathcal{T}^\pi$ looks one step ahead using policy π.

Matrix Form for Bellman Operator

Lemma 1

The Bellman operator can be expressed as the following matrix form:

$$\mathcal{T}^{\pi} V = r^{\pi} + \gamma P^{\pi} V,$$

where

$$r^{\pi} = \begin{bmatrix} r^{\pi}(s_1) \\ \vdots \\ r^{\pi}(s_n) \end{bmatrix}, \quad P^{\pi} = \begin{bmatrix} p^{\pi}_{s_1 s_1} & \cdots & p^{\pi}_{s_1 s_n} \\ \vdots & \ddots & \vdots \\ p^{\pi}_{s_n s_1} & \cdots & p^{\pi}_{s_n s_n} \end{bmatrix},$$

and the entries of r_{π} and P^{π} are

$$r^{\pi}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') \right] \quad \text{and} \quad p^{\pi}_{ss'} = \sum_{a} \pi(a|s) P(s'|s,a).$$

- $p^{\pi}_{ss'}$ is the transition probability from s to s' under policy π.

Basic Properties

Properties about P^π

- $I - \gamma P^\pi$ is invertible
- $(I - \gamma P^\pi)^{-1} \geq I$
- if $r \geq 0$, then $(I - \gamma P^\pi)^{-1} r \geq r \geq 0$

$$\Rightarrow \quad V^\pi = (I - \gamma P^\pi)^{-1} r_\pi$$

Properties about $\mathcal{T}^\pi$

- $\mathcal{T}^\pi$ is monotone, i.e., $\mathcal{T}^\pi V_1 \leq \mathcal{T}^\pi V_2$ if $V_1 \leq V_2$.
- $\mathcal{T}^\pi$ is a contraction with respect to $\|\cdot\|_\infty$,

$$\|\mathcal{T}^\pi V_1 - \mathcal{T}^\pi V_2\|_\infty \leq \gamma \|V_1 - V_2\|.$$

$$\Rightarrow \quad V^{k+1} = \mathcal{T}^\pi V^k \to V^\pi$$

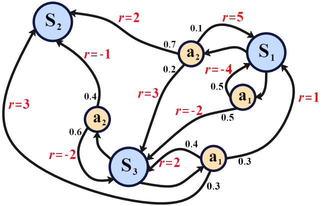
$V^k \to V^\pi$ can be established via fixed point theorem.

Fixed Point Theorem

Definition 2 (Contraction mapping)
Let (X, d) be a complete metric space. Then a map $\mathcal{T} : X \to X$ is called a contraction mapping on X if there exists $\rho \in [0, 1)$ such that $d(\mathcal{T}x, \mathcal{T}y) \leq \rho \cdot d(x, y)$ for all $x, y \in X$.

Theorem 2 (Fixed point theorem)
Let (X, d) be a non-empty complete metric space with a contraction mapping $\mathcal{T} : X \to X$. Then T admits a unique fixed point x^* in X (i.e. $\mathcal{T}x^* = x^*$). Furthermore, x^* can be obtained as follows: start with an arbitrary element $x_0 \in X$ and define a sequence $(x^k)_{k \in \mathbb{N}}$ by $x^k = \mathcal{T}x^{k-1}$ for $k \geq 1$. Then $\lim_{k\to\infty} x^k = x^*$.

Illustrative Example



Consider policy $\pi(a|s) = 0.5$ for all s, a and let $\gamma = 0.9$:

$$P^\pi = \begin{bmatrix} 0.3 & 0.35 & 0.35 \\ 0 & 1 & 0 \\ 0.15 & 0.35 & 0.5 \end{bmatrix},$$

$$r^\pi = [-0.25, 0, 0.2]^T,$$

$$V^\pi = [-0.21, 0, 0.31]^T.$$

We can also verify the correctness of V^π. Taking the state s_0 as an example, it is not hard to show that

$$\begin{aligned} V^\pi(s_3) =& \sum_a \pi(a|s_3) \sum_{s'} p(s'|s_3, a) \left(r(s_3, a, s') + \gamma V^\pi(s')\right) \\ =& 0.5\,(-1.6 + 0.9 \times 0.6 \times 0.31) + 0.5\,(2 + 0.9(0.4 \times 0.31 - 0.3 \times 0.21)) \\ =& 0.31. \end{aligned}$$

Bellman Equation for Action Value

Theorem 3
*Given policy π, action value satisfies the following **Bellman equation**:*

$$Q^\pi(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma \mathbb{E}_{a'\sim\pi(\cdot|s')}\left[Q^\pi(s',a')\right]\right].$$

*Alternatively, if for any $Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, define the **Bellman operator**:*

$$[\mathcal{F}^\pi Q](s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma \mathbb{E}_{a'\sim\pi(\cdot|s')}\left[Q(s',a')\right]\right],$$

Bellman equation can be rewritten as

$$Q^\pi = \mathcal{F}^\pi Q^\pi.$$

That is, Q^π is a fixed point of $\mathcal{F}^\pi$.

- $\mathcal{F}^\pi$ also admits a matrix form and it is also a contraction with infinity norm.

Table of Contents

Markov Decision Process and Bellman Equations

Bellman Optimality Equations

Optimal State Value and Optimal Policy

Definition:

- Optimal state value: $V^*(s) = \max_\pi V^\pi(s), \forall\, s \in \mathcal{S}$

For an MDP, optimal state value exists since we can restrict our focus on the set of deterministic policies (finite number) as implied by the following theorem.

Lemma 2 (Policy improvement)

For any policy π, if we define a new policy π' as follows:

$$\pi'(a|s) = \begin{cases} 1 & \textit{if } a = \arg\max_a \underbrace{\mathbb{E}_{s' \sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V^\pi(s')\right]}_{Q^\pi(s,a)}, \\ 0 & \textit{otherwise}, \end{cases}$$

then there holds $V^{\pi'} \geq V^\pi$.

For conciseness, we only consider stationary policies.

Proof of Lemma 2

By the Bellman equation, we have

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma V^{\pi}(s') \right] \\ &\leq \max_{a} \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma V^{\pi}(s') \right]. \end{aligned}$$

In terms of the Bellman operator, this implies that

$$V^{\pi} = \mathcal{T}^{\pi} V^{\pi} \leq \mathcal{T}^{\pi'} V^{\pi}.$$

Iterating this procedure yields that

$$V^{\pi} \leq (\mathcal{T}^{\pi'})^{k} V^{\pi} \to V^{\pi'},$$

which completes the proof.

Remark

Note that the key in the proof of Lemma 2 is

$$V^\pi \leq \mathcal{T}^{\pi'} V^\pi,$$

which relates two policies together in one inequality. This basically means π' is superior to π in one-step lookahead while $V^\pi \leq (\mathcal{T}^{\pi'})^k V^\pi$ means π' is superior to π in multi-step lookahead. Indeed, there is another way to show $V^{\pi'} \geq V^\pi$ based on this inequality:

$$V^{\pi'} - V^\pi = (I - \gamma P^{\pi'})^{-1}(r^{\pi'} - (I - \gamma P^{\pi'})V^\pi) \geq 0,$$

where we have used the identity

$$\mathcal{T}^{\pi'} V^\pi - V^\pi = r^{\pi'} + \gamma P^{\pi'} V^\pi - V^\pi = r^{\pi'} - (I - \gamma P^{\pi'})V^\pi.$$

As can be seen later, the weighted performance difference between $V^{\pi'}$ and V^π is an expectation of $\mathcal{T}^{\pi'} V^\pi(s) - V^\pi(s)$ over states under certain distribution.

Construction of Optimal Policy

Theorem 4 (Existence of optimal policy)

For an MDP, there exists a single deterministic optimal policy π^* *defined as follows*

$$\pi^*(a|s) = \begin{cases} 1 & \textit{if } a = \arg\max_a \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma V^*(s') \right], \\ 0 & \textit{otherwise}. \end{cases}$$

such that

$$V^*(s) = V^{\pi*}(s), \quad \forall\, s \in \mathcal{S}.$$

Proof of Theorem 4

By the definition of π^*, we have for any π,

$$
\begin{aligned}
V^\pi(s) &= \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V^\pi(s')\right] \\
&\le \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V^*(s')\right] \\
&\le \max_a \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V^*(s')\right].
\end{aligned}
$$

It follows that $\forall\, \pi$, $V^\pi \le \mathcal{T}^{\pi^*} V^*$. Thus, $V^* \le \mathcal{T}^{\pi^*} V^*$. Iterating this procedure yields as in the proof of Theorem 2 yields that $V^* \le V^{\pi^*}$. Since $V^* \ge V^{\pi^*}$ holds trivially, we have $V^* = V^{\pi^*}$.

Bellman Optimality Equation for Optimal State Value

Theorem 5 (Bellman optimality equation)

The optimal state value satisfies the following ***Bellman optimality equation****:*

$$V^*(s) = \max_a \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma V^* \left(s' \right) \right] .$$

Alternatively, if for any $V \in \mathbb{R}^{|\mathcal{S}|}$, define the ***Bellman optimality operator****:*

$$[\mathcal{T}V](s) = \max_a \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma V \left(s' \right) \right] ,$$

Bellman optimality equation can be rewritten as

$$V^* = \mathcal{T}V^* .$$

That is, V^* is a fixed point of $\mathcal{T}$.

Remark

- Bellman optimality equation provides a more tractable characterization of optimal values and value-based RL methods are essentially about solving Bellman optimality equation under various settings.
- Bellman optimality operator can be viewed as one-step improvement operator. It is easy to see that $\mathcal{T}$ is monotone, $\mathcal{T}V_1 \leq \mathcal{T}V_2$ if $V_1 \leq V_2$. In addition, $\mathcal{T}$ has the following matrix form

$$\mathcal{T}V = \max_{\pi} \mathcal{T}^{\pi} V = \max_{\pi} \{r^{\pi} + \gamma P^{\pi} V\}.$$

Proof of Theorem 5

Since $V^*(s) = V^{\pi^*}(s)$, by Bellman equation for $V^{\pi^*}(s)$, we have

$$\begin{aligned} V^*(s) = V^{\pi^*}(s) &= \mathbb{E}_{a\sim\pi^*(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V^{\pi^*}(s')\right] \\ &= \mathbb{E}_{a\sim\pi^*(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V^*(s')\right] \\ &= \max_a \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V^*(s')\right], \end{aligned}$$

where the last line follows the definition of π^*.

Contraction Property of Bellman Optimality Operator

Lemma 3

The Bellman optimality operator of state value is a contraction with respect to infinity norm,

$$\|\mathcal{T}V_1 - \mathcal{T}V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty.$$

It follows that there exists a unique solution for Bellman optimality equation of state value.

Proof: The proof is based directly on the following observation:

$$|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|.$$

- Together with Theorem 5 and fixed point theorem, Lemma 3 implies that V^* is the unique solution to the Bellman optimality equation. This provides a tractable and dynamical programming approach to find V^* in contrast to the naive search over $|\mathcal{A}|^{|\mathcal{S}|}$ policies.

Optimal Action Value

Definition:

- Optimal action value: $Q^*(s,a) = \max_\pi Q^\pi(s,a), \forall\, s \in \mathcal{S}, a \in \mathcal{A}$

Lemma 4

Recalling the definition of optimal state value $V^*(s)$, there hold

$$Q^*(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma V^* \left(s'\right) \right],$$
$$V^*(s) = \max_a Q^*(s,a).$$

Proof: The equalities follow directly from Theorems 4 and 5.

- The optimal policy can be defined as

$$\pi^*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_a Q^*(s,a), \\ 0 & \text{otherwise.} \end{cases}$$

  This fact forms the foundation of Q-learning, which is essentially about learning optimal action values.

Bellman Optimality Equation for Optimal Action Value

Theorem 6

*The optimal action value satisfies the following **Bellman optimality equation**:*

$$Q^*(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma \max_{a' \in \mathcal{A}} Q^*(s',a') \right].$$

*Alternatively, if for any $Q \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, define the **Bellman optimality operator**:*

$$[\mathcal{F}Q](s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma \max_{a' \in \mathcal{A}} Q(s',a') \right],$$

Bellman optimality equation can be rewritten as

$$Q^* = \mathcal{F}Q^*.$$

That is, Q^* is a fixed point of $\mathcal{F}$.

- $\mathcal{F}$ is also a contraction with infinity norm.

Questions?