Planning in Markov Decision Processes

Announcements

HW0: due this Thursday 11:59pm
Gradescope (please self-enroll)

Announcements

HW0: due this Thursday 11:59pm
Gradescope (please self-enroll)

Waiting List

Recap: Infinite Horizon MDPs

$$\mathscr{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Recap: Infinite Horizon MDPs

$$\mathscr{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Stationary Policy $\pi : S \mapsto \Delta(A)$

Recap: Infinite Horizon MDPs

$$\mathscr{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Stationary Policy $\pi : S \mapsto \Delta(A)$ → π^* deterministic A^S

Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

Recap: Infinite Horizon MDPs

$$\mathscr{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Stationary Policy $\pi : S \mapsto \Delta(A)$

Value function $V^{\pi}(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

Q function $Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s, a), a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

Recap: Bellman Optimality

$$\mathscr{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Theorem 1: Bellman Optimality

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

Recap: Bellman Optimality

$$\mathscr{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Theorem 1: Bellman Optimality

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

Bell consistent

Theorem 2:

For any $Q : S \times A \to \mathbb{R}$, if $Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s', a') \right]$ for all s, a, then $Q(s,a) = Q^\star(s,a), \forall s, a$

Main Question for Today:

Given an MDP $\mathscr{M} = (S, A, P, \gamma)$, How to find $\pi^\star$ (stationary & deterministic)

P. r. are known

$\forall s, a, s'$, $P(s' \mid s a)$ ✓

$\forall s, a$, $r(s, a)$ ✓

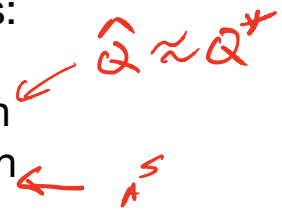
$\Rightarrow \pi^*$

Main Question for Today:

Given an MDP $\mathscr{M} = (S, A, P, \gamma)$, How to find $\pi^\star$ (stationary & deterministic)

Two Approaches:

1. Value Iteration
2. Policy Iteration



$f: S \times A \to \mathbb{R}$

Define Bellman Operator $\mathcal{T}$: $\mathcal{T} \cdot f: S \times A \to \mathbb{R}$

Given a function $f : S \times A \mapsto \mathbb{R}$,

$$\mathcal{T}f : S \times A \mapsto \mathbb{R},$$

$$(\mathcal{T}f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in A} f(s',a'), \forall s, a \in S \times A$$

Example: $f = Q^*$

$$\mathcal{T} Q^*(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \max_{a \in A} Q^*(s',a') = Q^*(s,a)$$

$$\mathcal{T} Q^* = Q^*$$

Value Iteration Algorithm:

1. Initialization: $\widehat{Q}^0 : \| \widehat{Q}^0 \|_\infty \in (0, \frac{1}{1-\gamma})$

2. Iterate until convergence: $\widehat{Q}^{t+1} = \mathcal{T} \widehat{Q}^t$

Intuition:

Via Bellman optimality theorem:

$$Q^\star = \mathcal{T} Q^\star$$ ⇐ Fixed point

Intuition:

Via Bellman optimality theorem:

$$Q^\star = \mathcal{T} Q^\star$$

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Intuition:

Via Bellman optimality theorem:

$$Q^\star = \mathcal{T} Q^\star$$

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

$\ell : \mathbb{R} \to \mathbb{R}$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

Intuition:

Via Bellman optimality theorem:

$$Q^\star = \mathcal{T} Q^\star$$

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

Intuition:

Via Bellman optimality theorem:

$$Q^\star = \mathcal{T} Q^\star$$

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

$|x_t - x^\star| =$

Intuition:

Via Bellman optimality theorem:

$$Q^\star = \mathcal{T} Q^\star$$

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

$$|x_t - x^\star| = |\ell(x_{t-1}) - \ell(x^\star)|$$

Intuition:

Via Bellman optimality theorem:

$$Q^\star = \mathcal{T} Q^\star$$

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

$f: \mathbb{R} \to \mathbb{R}$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

Lipchitz constant

$$|x_t - x^\star| = |\ell(x_{t-1}) - \ell(x^\star)| \leq L |x_{t-1} - x^\star|$$

Intuition:

Via Bellman optimality theorem:

$$Q^\star = \mathcal{T} Q^\star$$

i.e., $Q^\star$ is the fixed point solution of $f = \mathcal{T} f$

Consider the simple problem: finding fixed point solution $x^\star = \ell(x^\star)$

$$x_0, x_{t+1} = \ell(x_t), t = 0, \ldots,$$

$$|x_t - x^\star| = |\ell(x_{t-1}) - \ell(x^\star)| \leq L |x_{t-1} - x^\star|$$

$\cdots L^t |x_0 - x^*|$

$t \to \infty$

$L < 1$

$L^t \to 0$

If $L < 1$ (i.e., contraction), then it converges exponentially fast

Convergence of Value Iteration:

Bellman-operator

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

$\mathcal{T}$ $\mathcal{T}$

$\|f\|_\infty = \max_{s,a} |f(s,a)|$

Proof:

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$$|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right|$$

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$$|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \gamma \sum_{s'} P(s' \mid s,a) \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$$|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \gamma \sum_{s'} P(s'|s,a) \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right|$$

$$\leq \gamma \sum_{s'} P(s'|s,a) \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

$\mathbb{E}_{s' \sim P(\cdot|s,a)} f(s') \leq \max_{s'} f(s')$

Assume $f \geq 0$

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$$\begin{aligned}
|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| &= \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right| \\
&\leq \gamma \sum_{s'} P(s' \mid s,a) \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right| \\
&\leq \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right| \\
&\leq \gamma \max_{s'} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|
\end{aligned}$$

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Proof:

$\forall s,a \in S \times A$

$$\begin{aligned}
|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| &= \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right| \\
&\leq \gamma \sum_{s'} P(s' \mid s,a) \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right| \\
&\leq \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right| \\
&\leq \gamma \max_{s'} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right| \quad = \gamma \|Q - Q'\|_\infty
\end{aligned}$$

$\Rightarrow \|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$

$\gamma < 1$

Convergence of Value Iteration:

Lemma [Convergence]: Given $\widehat{Q}^0$, we have:

$$\| \widehat{Q}^t - Q^\star \|_\infty \leq \gamma^t \| \widehat{Q}^0 - Q^\star \|_\infty$$

Proof:

Convergence of Value Iteration:

Lemma [Convergence]: Given $\widehat{Q}^0$, we have:

$$\| \widehat{Q}^t - Q^\star \|_\infty \leq \gamma^t \| \widehat{Q}^0 - Q^\star \|_\infty$$

Proof: $Q^* = \mathcal{T} Q^*$

$$\| \widehat{Q}^{t+1} - Q^\star \|_\infty = \| \mathcal{T} \widehat{Q}^t - \mathcal{T} Q^\star \|_\infty \leq \gamma \| \widehat{Q}^t - Q^\star \|_\infty$$

← Recursion

Convergence of Value Iteration:

Lemma [Convergence]: Given $\widehat{Q}^0$, we have:

$$\| \widehat{Q}^t - Q^\star \|_\infty \leq \gamma^t \| \widehat{Q}^0 - Q^\star \|_\infty$$

Proof:

$$\| \widehat{Q}^{t+1} - Q^\star \|_\infty = \| \mathcal{T} \widehat{Q}^t - \mathcal{T} Q^\star \|_\infty \leq \gamma \| \widehat{Q}^t - Q^\star \|_\infty$$

$$\cdots \leq \gamma^{t+1} \| \widehat{Q}^0 - Q^\star \|_\infty$$

① Bell-operator γ-Contraction

② Recursion

$\gamma < 1$, $t \to \infty$ $\gamma^t \to 0$

$\widehat{Q}^t_{(s,a)} \to Q^\star_{(s,a)}$ $\forall s,a$

Final Quality of the Policy:

$$\hat{\pi}^t : \hat{\pi}^t(s) = \arg\max_a \widehat{Q}^t(s,a)$$

Theorem: $V^{\hat{\pi}^t}(s) \geq V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|\widehat{Q}^0 - Q^\star\|_\infty \forall s \in S$

Proof:

Final Quality of the Policy:

$$\hat{\pi}^t : \hat{\pi}^t(s) = \arg\max_a \widehat{Q}^t(s,a)$$

Theorem: $V^{\hat{\pi}^t}(s) \geq V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|\hat{Q}^0 - Q^\star\|_\infty \forall s \in S$

$s \in S$ Bell-Eqn

Proof:

$$V^{\hat{\pi}^t}(s) - V^\star(s) = Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

Final Quality of the Policy:

$$\hat{\pi}^t : \hat{\pi}^t(s) = \arg\max_a \widehat{Q}^t(s,a)$$

Theorem: $V^{\hat{\pi}^t}(s) \geq V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|\widehat{Q}^0 - Q^\star\|_\infty \forall s \in S$

Proof:

$$V^{\hat{\pi}^t}(s) - V^\star(s) = Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$\stackrel{a}{=} Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \hat{\pi}^t(s)) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

+/−

Final Quality of the Policy:

$$\hat{\pi}^t : \hat{\pi}^t(s) = \arg\max_a \widehat{Q}^t(s,a)$$

Theorem: $V^{\hat{\pi}^t}(s) \geq V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|\widehat{Q}^0 - Q^\star\|_\infty \forall s \in S$

Proof:

$$V^{\hat{\pi}^t}(s) - V^\star(s) = Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$= Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \hat{\pi}^t(s)) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

Bell-Eqn

$$= \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

Final Quality of the Policy:

$$\hat{\pi}^t : \hat{\pi}^t(s) = \arg\max_a \widehat{Q}^t(s,a)$$

Theorem: $V^{\hat{\pi}^t}(s) \geq V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|\widehat{Q}^0 - Q^\star\|_\infty \forall s \in S$

Proof:

$$V^{\hat{\pi}^t}(s) - V^\star(s) = Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$= Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \hat{\pi}^t(s)) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) + Q^\star(s, \hat{\pi}^t(s)) - \widehat{Q}^t(s, \hat{\pi}^t(s)) + \widehat{Q}^t(s, \pi^\star(s)) - Q^\star(s, \pi^\star(s))$$

$\widehat{Q}^t(s, \hat{\pi}^t(s)) = \max_a \widehat{Q}^t(s,a) \geq \widehat{Q}^t(s, \pi^\star(s))$

$\widehat{Q}^t(s, \pi^\star(s)) - \widehat{Q}^t(s, \hat{\pi}^t(s)) \leq 0$

Final Quality of the Policy:

output of VI

$$\hat{\pi}^t : \hat{\pi}^t(s) = \arg\max_a \widehat{Q}^t(s,a)$$

Theorem: $V^{\hat{\pi}^t}(s) \geq V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|\widehat{Q}^0 - Q^\star\|_\infty \forall s \in S$

$Q^{\hat{\pi}^t}$

Proof:

$$V^{\hat{\pi}^t}(s) - V^\star(s) = Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$= Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \hat{\pi}^t(s)) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) + Q^\star(s, \hat{\pi}^t(s)) - \widehat{Q}^t(s, \hat{\pi}^t(s)) + \widehat{Q}^t(s, \pi^\star(s)) - Q^\star(s, \pi^\star(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) - 2\gamma^t \|\widehat{Q}^0 - Q^\star\|_\infty$$

Recursion

Final Quality of the Policy:

$$\hat{\pi}^t : \hat{\pi}^t(s) = \arg\max_a \widehat{Q}^t(s,a)$$

Theorem: $V^{\hat{\pi}^t}(s) \geq V^\star(s) - \frac{2\gamma^t}{1-\gamma}\|\widehat{Q}^0 - Q^\star\|_\infty \forall s \in S$

Proof:

$$V^{\hat{\pi}^t}(s) - V^\star(s) = Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$= Q^{\hat{\pi}^t}(s, \hat{\pi}^t(s)) - Q^\star(s, \hat{\pi}^t(s)) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$= \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) + Q^\star(s, \hat{\pi}^t(s)) - Q^\star(s, \pi^\star(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) + Q^\star(s, \hat{\pi}^t(s)) - \widehat{Q}^t(s, \hat{\pi}^t(s)) + \widehat{Q}^t(s, \pi^\star(s)) - Q^\star(s, \pi^\star(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}^t(s))}\left(V^{\hat{\pi}^t}(s') - V^\star(s')\right) - 2\gamma^t\|\widehat{Q}^0 - Q^\star\|_\infty \ldots \text{Recursion}$$

Policy Iteration Algorithm:

(focus on Deterministic policies)

1. Initialization: $\pi^0 : S \mapsto \mathcal{P}(A)$

Policy Iteration Algorithm:

1. Initialization: $\pi^0 : S \mapsto \Delta(A)$

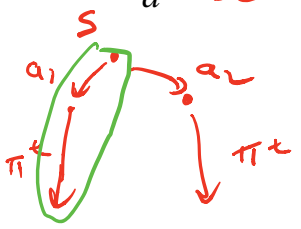
2. Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

Policy Iteration Algorithm:

1. Initialization: $\pi^0 : S \mapsto \Delta(A)$

2. Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

3. Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$



Policy Iteration Algorithm:

1. Initialization: $\pi^0 : S \mapsto \Delta(A)$

2. Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

3. Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Policy Iteration Algorithm:

Closed-form for PE
(see 1.1.3 in Monograph)

1. Initialization: $\pi^0 : S \mapsto \Delta(A)$

2. Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

3. Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s, a$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s,a$

$\forall s,a \in S \times A$

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

↑ Bell-Eqn ↑ Bell-Eqn

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s,a$

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$= \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) + Q^{\pi^t}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s,a$

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$= \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) + Q^{\pi^t}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) \right]$$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s,a$

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$= \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) + Q^{\pi^t}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) \right] \geq \ldots, \geq \gamma^\infty = 0$$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s,a$

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$= \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) + Q^{\pi^t}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^t(s')) \right]$$

$$\geq \gamma \mathbb{E}_{s' \sim P(s,a)} \left[Q^{\pi^{t+1}}(s', \pi^{t+1}(s')) - Q^{\pi^t}(s', \pi^{t+1}(s')) \right] \geq \ldots, \geq \gamma^\infty = 0$$

$V^{\pi^{t+1}}(s) = Q^{\pi^{t+1}}(s, \pi^{t+1}(s))$

$\geq Q^{\pi^t}(s, \pi^{t+1}(s))$

$= \max_a Q^{\pi^t}(s,a) \geq Q^{\pi^t}(s, \pi^t(s)) = V^{\pi^t}(s)$

$V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s), \forall s$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$ $\cdots \gamma^t \|V^{\pi^0} - V^\star\|_\infty$

$t \to \infty$

$\gamma^t \to 0$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

$\forall s \in S$

Bell-opt

$$V^\star(s) - V^{\pi^{t+1}}(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^{t+1}}(s') \right]$$

Bell-Eqn

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

$V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s) \quad \forall s$

$$V^\star(s) - V^{\pi^{t+1}}(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^{t+1}}(s') \right]$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s') \right]$$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

$$
\begin{aligned}
V^\star(s) - V^{\pi^{t+1}}(s) &= \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^{t+1}}(s') \right] \\
&\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s') \right] \\
&= \max_a (r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \gamma V^\star(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s'))
\end{aligned}
$$

$= Q^{\pi^t}(s,a)$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

$$
\begin{aligned}
V^\star(s) - V^{\pi^{t+1}}(s) &= \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^{t+1}}(s') \right] \\
&\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s') \right] \\
&= \max_a (r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \gamma V^\star(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s')) \\
&\leq \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s') \right) \right)
\end{aligned}
$$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

$$V^\star(s) - V^{\pi^{t+1}}(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')\right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^{t+1}}(s')\right]$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')\right] - \left[r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s')\right]$$

$$= \max_a (r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \gamma V^\star(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s'))$$

$$\leq \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s') \right) \right)$$

$$\leq \gamma \|V^\star - V^{\pi^t}\|_\infty \quad \cdots \gamma^t \|V^\star - V^{\pi^0}\|_\infty$$

Value Iteration vs Policy Iteration

Which one is faster?

How many iterations (computation complexity) need to find the EXACT optimal policy?

A key Lemma: Performance Difference Lemma

Q: What's the performance difference of two policies π, π', i.e., $V^{\pi}(s) - V^{\pi'}(s)$

A key Lemma: Performance Difference Lemma

PDL Statement:

$$V^{\pi}(s_0) - V^{\pi'}(s_0) \overset{\Delta}{=} \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim d^{\pi}_{s_0}} \left[Q^{\pi'}(s,a) - V^{\pi'}(s) \right]$$

$s_0 \in S$

$A^{\pi'}(s,a)$

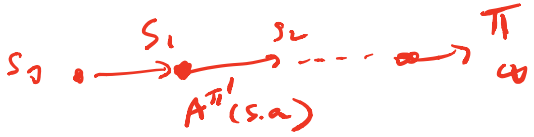
from π

$$d^{\pi}_{s_0}(s,a) = (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h(s,a;s_0,\pi)$$

$\mathbb{P}_h(s,a;s_0,\pi)$: probability of π visiting (s,a) at time step $h \in \mathbb{N}$, starting at s_0

Proof of PDL

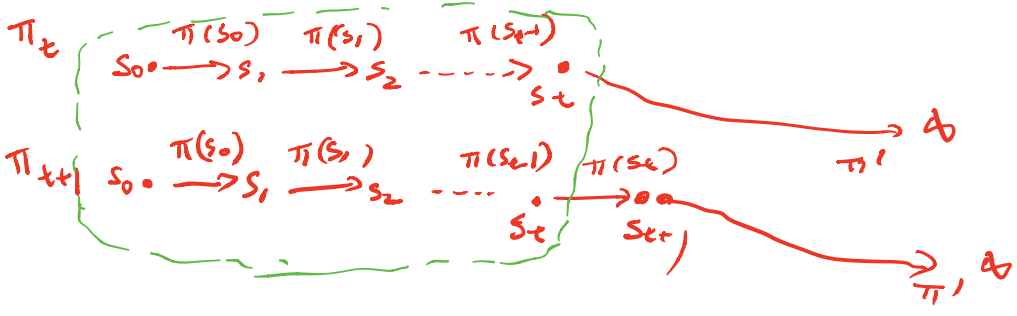
$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$$



Proof of PDL

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$



Proof of PDL

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

Proof of PDL

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0)$$

Proof of PDL

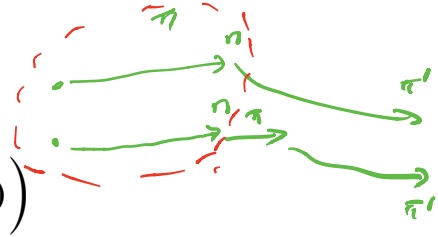
$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0)$$

$$= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(r(s_n,\pi(s_n)) + \gamma\mathbb{E}_{s_{n+1}\sim P_{s_n,\pi(s_n)}} V^{\pi'}(s_{n+1}) - V^{\pi'}(s_n)\right)$$



Proof of PDL

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$\begin{aligned}
&V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(r(s_n,\pi(s_n)) + \gamma\mathbb{E}_{s_{n+1}\sim P_{s_n,\pi(s_n)}} V^{\pi'}(s_{n+1}) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(Q^{\pi'}(s_n,\pi(s_n)) - V^{\pi'}(s_n)\right)
\end{aligned}$$

Proof of PDL $V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$\begin{aligned}
&V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(r(s_n,\pi(s_n)) + \gamma\mathbb{E}_{s_{n+1}\sim P_{s_n,\pi(s_n)}} V^{\pi'}(s_{n+1}) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(Q^{\pi'}(s_n,\pi(s_n)) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(A^{\pi'}(s_n,\pi(s_n))\right)
\end{aligned}$$

Proof of PDL $V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi}_{s_0}}\left[A^{\pi'}(s,a)\right]$

$$\pi_0 = \{\pi', \ldots, \pi'\}, \pi_1 = \{\pi, \pi', \ldots, \pi'\}, \pi_2 = \{\pi, \pi, \pi', \ldots, \pi'\}, \ldots \pi_\infty = \{\pi, \ldots, \pi\}$$

$$V^{\pi}(s_0) = V^{\pi_\infty}(s_0), \quad V^{\pi'}(s_0) = V^{\pi_0}(s_0)$$

$$\begin{aligned}
&V^{\pi_{n+1}}(s_0) - V^{\pi_n}(s_0) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(r(s_n,\pi(s_n)) + \gamma\mathbb{E}_{s_{n+1}\sim P_{s_n,\pi(s_n)}} V^{\pi'}(s_{n+1}) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(Q^{\pi'}(s_n,\pi(s_n)) - V^{\pi'}(s_n)\right) \\
&= \gamma^n \mathbb{E}_{s_n\sim\mathbb{P}_n(\cdot;s_0,\pi)}\left(A^{\pi'}(s_n,\pi(s_n))\right)
\end{aligned}$$

Telescoping: $V^{\pi'} - V^{\pi} = \sum_{n=0}^{\infty} V^{\pi_n} - V^{\pi_{n+1}}$

$V^{\pi_0}(s_0) \to V^{\pi_\infty}(s_0)$

$V^{\pi_0} - V^{\pi_1} + V^{\pi_1} - V^{\pi_2} + V^{\pi_2} - V^{\pi_3} = V^{\pi_0} - V^{\pi_\infty}$