CS 6789: Foundations of Reinforcement Learning

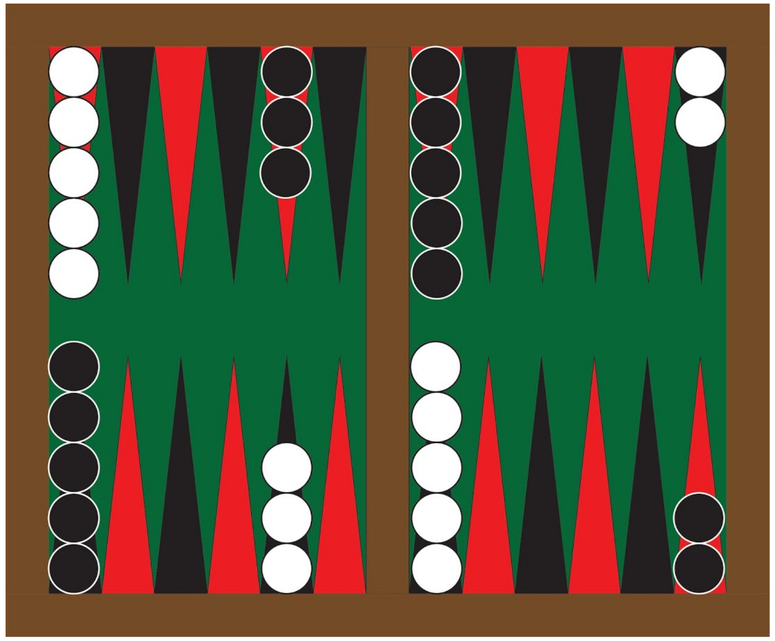
Sham Kakade (UW & MSR) & Wen Sun

TA: Jonathan Chang

https://wensun.github.io/CS6789.html

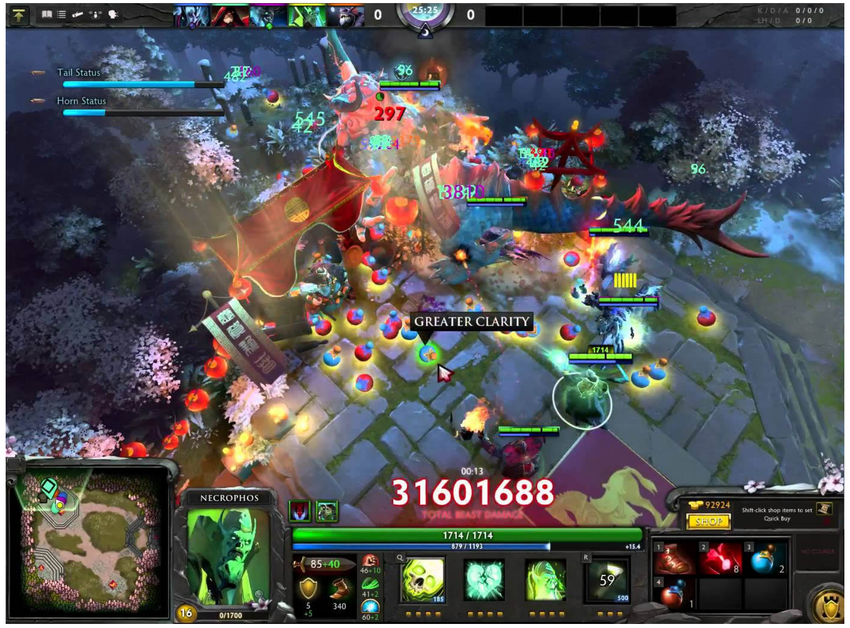
Fall 2020

Progress of RL in Practice



TD GAMMON [Tesauro 95]

[AlphaZero, Silver et.al, 17]



[OpenAI Five, 18]

This course focuses on RL Theory

We care about sample complexity:

Total number of environment interactions
needed to learn a high quality policy (i.e., achieve the task)

Four main themes we will cover in this course:

1. Exploration strategies (not just random)
2. Policy Optimization (gradient descent)
3. Control (LQR and nonlinear control)
4. Imitation Learning (i.e., learning from demonstrations)

Logistics

Four (HW0-HW3) assignments (total 55%) and Course Project (45%)

(HW0 10%, HW1-3 15% each)

HW0 is out today and due in one week

Prerequisites (HW0)

Deep understanding of Machine Learning, Optimization, Statistics

ML: sample complexity analysis for supervised learning (PAC)

Opt: Convex (linear) optimization, e.g., gradient decent for convex functions

Stats: basics of concentration (e.g., Hoeffding's), tricks such as union bound

Prerequisites (HW0)

Deep understanding of Machine Learning, Optimization, Statistics

ML: sample complexity analysis for supervised learning (PAC)

Opt: Convex (linear) optimization, e.g., gradient decent for convex functions

Stats: basics of concentration (e.g., Hoeffding's), tricks such as union bound

Check out HW0 asap!

Course projects (45%)

- Team work: size 3
- Midterm report (5%), Final presentation (15%), and Final report (25%)

- Basics: **survey** of a set of similar RL theory papers. Reproduce analysis and provide a coherent story
- Advanced: **identify** extensions of existing RL papers, **formulate** theory questions, and **provide** proofs

Course Notes

We will use the monograph: **Reinforcement Learning Theory and Algorithms**

https://rltheorybook.github.io/rl_monograph_AJK_V2.pdf

Basics of Markov Decision Processes

Supervised Learning

Supervised Learning

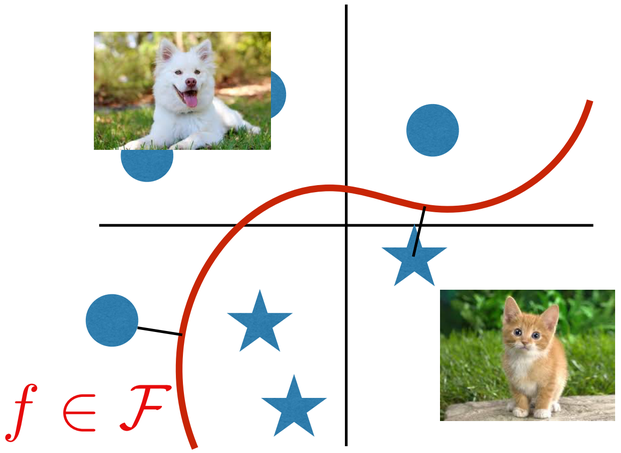
Given i.i.d examples at training:

( ,cat) ( ,cat) ( ,dog)

Supervised Learning

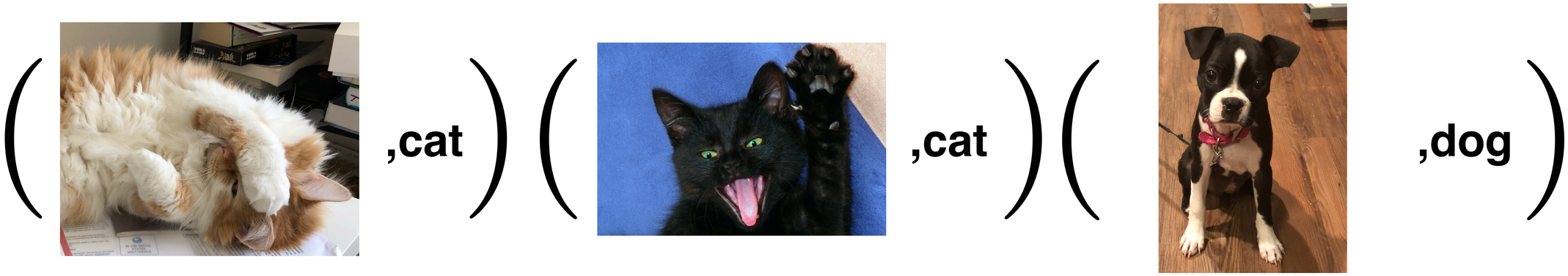
Given i.i.d examples at training:

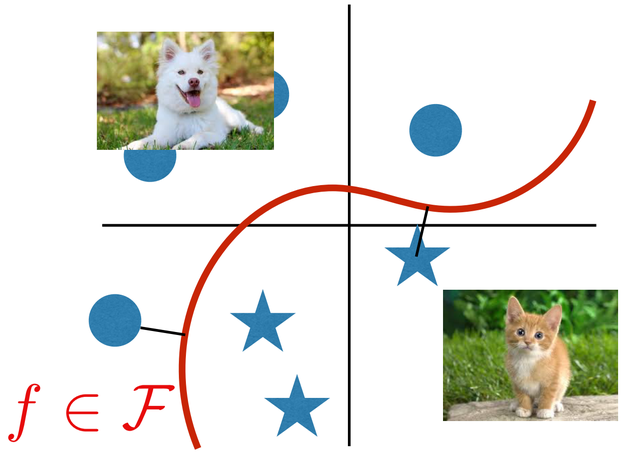
( ,cat) ( ,cat) ( ,dog)



Supervised Learning

Given i.i.d examples at training:





Passive:

Prediction

Data Distribution

Supervised Learning

Given i.i.d examples at training:

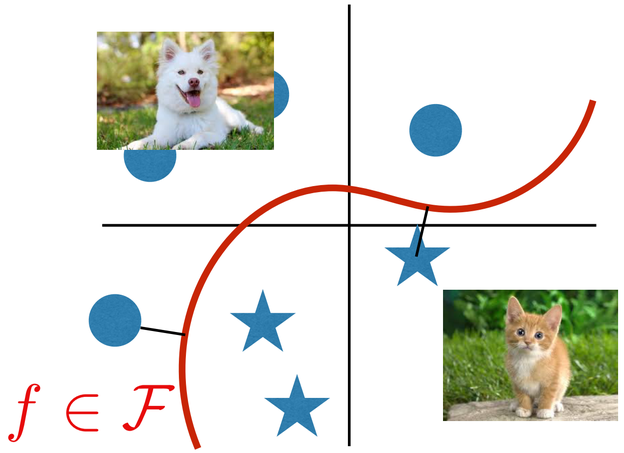
Cersei

,cat

,cat

Tasha

,dog



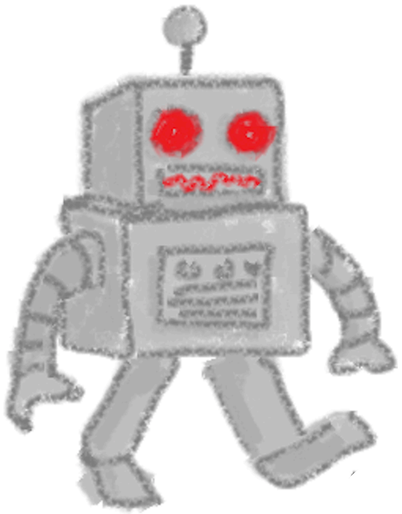
Passive:

Prediction

Data Distribution

Markov Decision Process

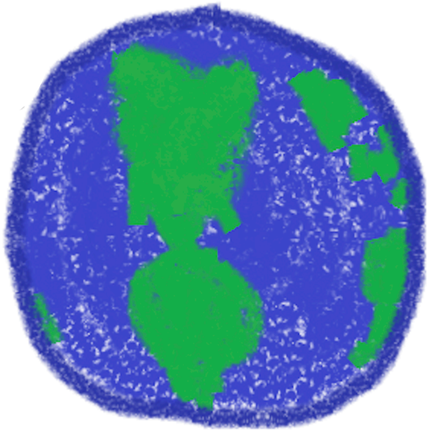
Learning Agent



$$\pi(s) \rightarrow a$$

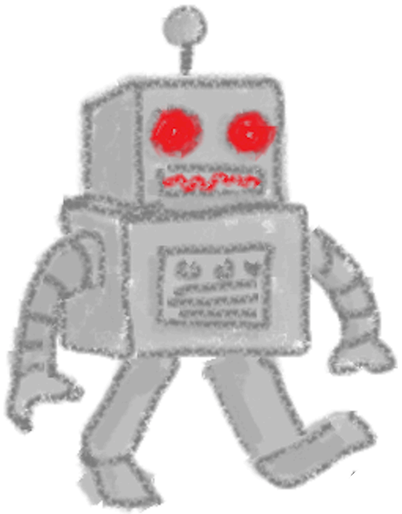
Policy: determine **action** based on **state**

Environmen



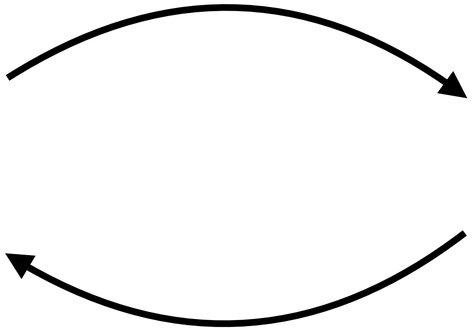
Markov Decision Process

Learning Agent



$$\pi(s) \rightarrow a$$

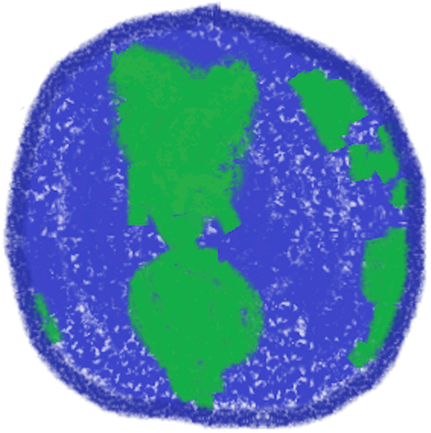
Policy: determine **action** based on **state**



Send **reward** and **next state** from a Markovian transition dynamics

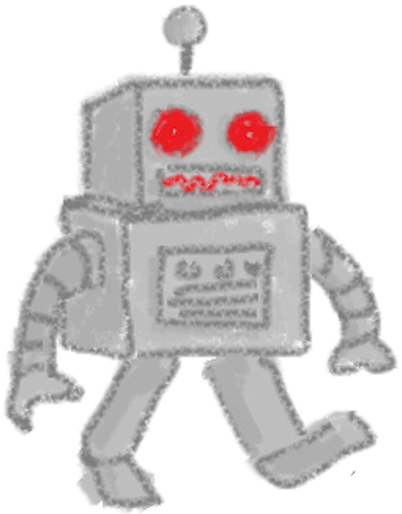
$$r(s,a), s' \sim P(\cdot \mid s,a)$$

Environmen



Markov Decision Process

Learning Agent



$$\pi(s) \rightarrow a$$

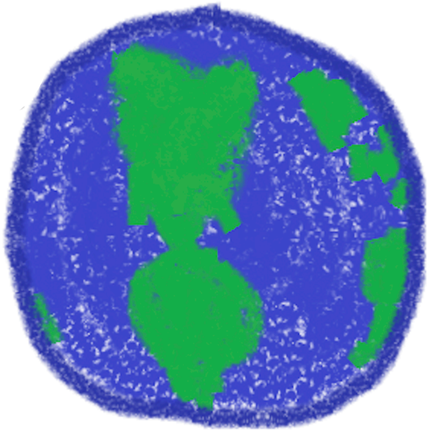
Policy: determine **action** based on **state**

Multiple Steps

Send **reward** and **next state** from a Markovian transition dynamics

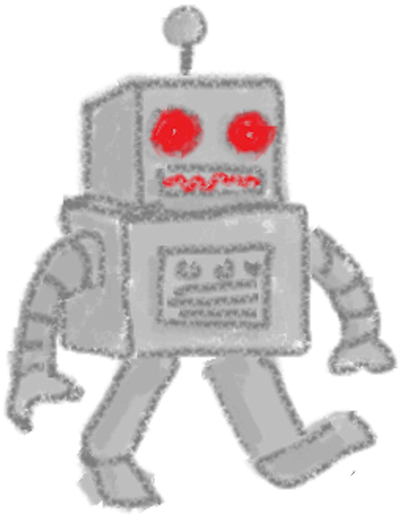
$$r(s, a), s' \sim P(\cdot \mid s, a)$$

Environmen



Markov Decision Process

Learning Agent



$$\pi(s) \rightarrow a$$

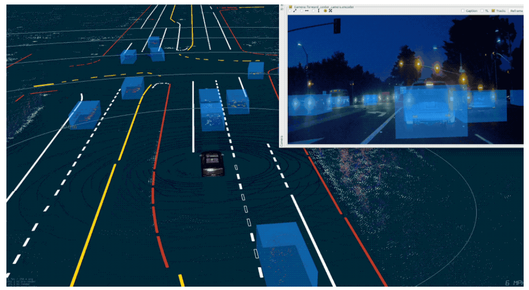
Policy: determine **action** based on **state**

Multiple Steps

Send **reward** and **next state** from a Markovian transition dynamics

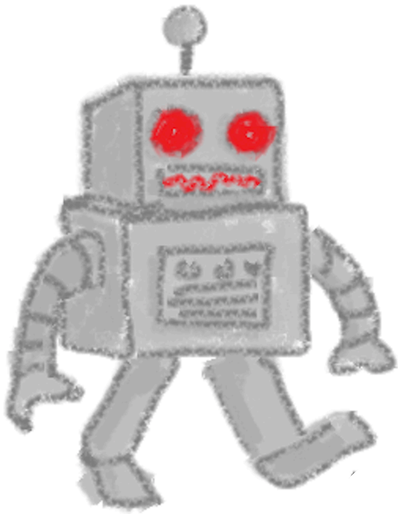
$$r(s, a), s' \sim P(\cdot \mid s, a)$$

Environmen



Markov Decision Process

Learning Agent



$$\pi(s) \rightarrow a$$

Policy: determine **action** based on **state**

Multiple Steps

Send **reward** and **next state** from a Markovian transition dynamics

$$r(s, a), s' \sim P(\cdot \mid s, a)$$

Environmen

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning					
Reinforcement Learning					

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓				
Reinforcement Learning	✓				

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓	✓			
Reinforcement Learning	✓	✓			

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓	✓			
Reinforcement Learning	✓	✓	✓		

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓	✓			
Reinforcement Learning	✓	✓	✓	✓	

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓	✓			
Reinforcement Learning	✓	✓	✓	✓	✓

Table content based on slides from Emma Brunskill

Infinite horizon Discounted Setting

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

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

Infinite horizon Discounted Setting

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

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

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

Infinite horizon Discounted Setting

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

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

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

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

Infinite horizon Discounted Setting

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

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

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

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

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

Bellman Equation:

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

Bellman Equation:

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

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

Bellman Equation:

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

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

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

Bellman Equation:

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

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

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

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

Optimal Policy

For infinite horizon discounted MDP, there exists a deterministic stationary policy

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^{\pi}(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.4 in the RL monograph]

Optimal Policy

For infinite horizon discounted MDP, there exists a deterministic stationary policy

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^{\pi}(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.4 in the RL monograph]

We denote $V^\star := V^{\pi^\star}, Q^\star := Q^{\pi^\star}$

Optimal Policy

For infinite horizon discounted MDP, there exists a deterministic stationary policy

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^{\pi}(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.4 in the RL monograph]

We denote $V^\star := V^{\pi^\star}, Q^\star := Q^{\pi^\star}$

Theorem 1: Bellman Optimality

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

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^\star(s))} V^\star(s')$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} V^\star(s')$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} V^\star(s')$$

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

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} V^\star(s')$$

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

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s',\widehat{\pi}(s'))} V^\star(s'') \right]$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} V^\star(s')$$

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

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s',\widehat{\pi}(s'))} V^\star(s'') \right]$$

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s',\widehat{\pi}(s'))} \left[r(s'', \widehat{\pi}(s'')) + \gamma \mathbb{E}_{s''' \sim P(s'',\widehat{\pi}(s''))} V^\star(s''') \right] \right]$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} V^\star(s')$$

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

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s',\widehat{\pi}(s'))} V^\star(s'') \right]$$

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s',\widehat{\pi}(s'))} \left[r(s'', \widehat{\pi}(s'')) + \gamma \mathbb{E}_{s''' \sim P(s'',\widehat{\pi}(s''))} V^\star(s''') \right] \right]$$

$$\leq \mathbb{E} \left[r(s, \widehat{\pi}(s)) + \gamma r(s', \widehat{\pi}(s')) + \ldots \right] = V^{\widehat{\pi}}(s)$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we just proved $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we just proved $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

This implies that $\arg\max_a Q^\star(s,a)$ is an optimal policy

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$| V(s) - V^\star(s) | = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right|$$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$|\, V(s) - V^\star(s) \,| = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right|$$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$
\begin{aligned}
|\, V(s) - V^\star(s) \,| &= \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right| \\
&\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right| \\
&\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| V(s') - V^\star(s') \right|
\end{aligned}
$$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$
\begin{aligned}
|\, V(s) - V^\star(s) \,| &= \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right| \\
&\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right| \\
&\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| V(s') - V^\star(s') \right| \\
&\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left(\max_{a'} \gamma \mathbb{E}_{s'' \sim P(s',a')} \left| V(s'') - V^\star(s'') \right| \right)
\end{aligned}
$$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$
\begin{aligned}
|\, V(s) - V^\star(s) \,| &= \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right| \\
&\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} V^\star(s')) \right| \\
&\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| V(s') - V^\star(s') \right| \\
&\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left(\max_{a'} \gamma \mathbb{E}_{s'' \sim P(s',a')} \left| V(s'') - V^\star(s'') \right| \right) \\
&\leq \max_{a_1, a_2, \ldots a_{k-1}} \gamma^k \mathbb{E}_{s_k} |\, V(s_k) - V^\star(s_k) \,|
\end{aligned}
$$

Finite Horizon Setting

$$\mathscr{M} = \{S, A, P, r, \mu_0, H\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad H \in \mathbb{N}^+, \quad \mu_0 \in \Delta(S)$$

Finite Horizon Setting

$$\mathcal{M} = \{S, A, P, r, \mu_0, H\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad H \in \mathbb{N}^+, \quad \mu_0 \in \Delta(S)$$

Given $\pi : S \mapsto \Delta(A)$

$$s_0 \sim \mu_0, a_0 \sim \pi(s_0), r(s_0, a_0), s_1 \sim P(\cdot \mid s_0, s_0), \ldots, s_{H-1} \sim P(\cdot \mid s_{H-2}, a_{H-2}), a_{H-1} \sim \pi(s_{H-1}), r(s_{H-1}, a_{H-1})$$

Finite Horizon Setting

$$\mathscr{M} = \{S, A, P, r, \mu_0, H\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad H \in \mathbb{N}^+, \quad \mu_0 \in \Delta(S)$$

Given $\pi : S \mapsto \Delta(A)$

$$s_0 \sim \mu_0, a_0 \sim \pi(s_0), r(s_0, a_0), s_1 \sim P(\cdot \mid s_0, s_0), \ldots, s_{H-1} \sim P(\cdot \mid s_{H-2}, a_{H-2}), a_{H-1} \sim \pi(s_{H-1}), r(s_{H-1}, a_{H-1})$$

Objective function: $J(\pi) = \mathbb{E}\left[\sum_{h=0}^{H-1} r(s_h, a_h)\right]$

Finite Horizon Setting

$$\mathscr{M} = \{S, A, P, r, \mu_0, H\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad H \in \mathbb{N}^+, \quad \mu_0 \in \Delta(S)$$

Given $\pi : S \mapsto \Delta(A)$

$$s_0 \sim \mu_0, a_0 \sim \pi(s_0), r(s_0, a_0), s_1 \sim P(\cdot \,|\, s_0, s_0), \ldots, s_{H-1} \sim P(\cdot \,|\, s_{H-2}, a_{H-2}), a_{H-1} \sim \pi(s_{H-1}), r(s_{H-1}, a_{H-1})$$

Time-dependent value/Q function:

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \,|\, s_h = s, a_t \sim \pi(s_t)\right], \quad Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \,|\, s_h = s, a_h = a, a_t \sim \pi(s_t)\right]$$

Finite Horizon Setting

$$\mathscr{M} = \{S, A, P, r, \mu_0, H\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad H \in \mathbb{N}^+, \quad \mu_0 \in \Delta(S)$$

Given $\pi : S \mapsto \Delta(A)$

$$s_0 \sim \mu_0, a_0 \sim \pi(s_0), r(s_0, a_0), s_1 \sim P(\cdot \,|\, s_0, s_0), \ldots, s_{H-1} \sim P(\cdot \,|\, s_{H-2}, a_{H-2}), a_{H-1} \sim \pi(s_{H-1}), r(s_{H-1}, a_{H-1})$$

Time-dependent value/Q function:

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \,|\, s_h = s, a_t \sim \pi(s_t)\right], \quad Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \,|\, s_h = s, a_h = a, a_t \sim \pi(s_t)\right]$$

$$V_h^\pi(s_h) = \mathbb{E}_{a_h \sim \pi(s_h)}\left[r(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim P(\cdot|s_h, a_h)} V_{h+1}^\pi(s_{h+1})\right]$$

Finite Horizon Setting

$$\mathscr{M} = \{S, A, P, r, \mu_0, H\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad H \in \mathbb{N}^+, \quad \mu_0 \in \Delta(S)$$

Given $\pi : S \mapsto \Delta(A)$

$$s_0 \sim \mu_0, a_0 \sim \pi(s_0), r(s_0, a_0), s_1 \sim P(\cdot \,|\, s_0, s_0), \ldots, s_{H-1} \sim P(\cdot \,|\, s_{H-2}, a_{H-2}), a_{H-1} \sim \pi(s_{H-1}), r(s_{H-1}, a_{H-1})$$

Time-dependent optimal policy $\pi^\star = \{\pi_0^\star, \ldots, \pi_{H-1}^\star\}$:

Finite Horizon Setting

$$\mathscr{M} = \{S, A, P, r, \mu_0, H\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad H \in \mathbb{N}^+, \quad \mu_0 \in \Delta(S)$$

Given $\pi : S \mapsto \Delta(A)$

$$s_0 \sim \mu_0, a_0 \sim \pi(s_0), r(s_0, a_0), s_1 \sim P(\cdot \,|\, s_0, s_0), \ldots, s_{H-1} \sim P(\cdot \,|\, s_{H-2}, a_{H-2}), a_{H-1} \sim \pi(s_{H-1}), r(s_{H-1}, a_{H-1})$$

Time-dependent optimal policy $\pi^\star = \{\pi_0^\star, \ldots, \pi_{H-1}^\star\}$:

$$Q_{H-1}^\star(s,a) = r(s,a), \pi_{H-1}^\star(s) = \arg\max_a Q_{H-1}^\star(s,a), V_{H-1}^\star(s) = \max_a Q_{H-1}^\star(s,a)$$

Finite Horizon Setting

$$\mathscr{M} = \{S, A, P, r, \mu_0, H\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad H \in \mathbb{N}^+, \quad \mu_0 \in \Delta(S)$$

Given $\pi : S \mapsto \Delta(A)$

$$s_0 \sim \mu_0, a_0 \sim \pi(s_0), r(s_0, a_0), s_1 \sim P(\cdot \mid s_0, s_0), \ldots, s_{H-1} \sim P(\cdot \mid s_{H-2}, a_{H-2}), a_{H-1} \sim \pi(s_{H-1}), r(s_{H-1}, a_{H-1})$$

Time-dependent optimal policy $\pi^\star = \{\pi_0^\star, \ldots, \pi_{H-1}^\star\}$:

$$Q_{H-1}^\star(s,a) = r(s,a), \pi_{H-1}^\star(s) = \arg\max_a Q_{H-1}^\star(s,a), V_{H-1}^\star(s) = \max_a Q_{H-1}^\star(s,a)$$

$$Q_h^\star(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} V_{h+1}^\star(s'), \pi_h^\star(s) = \arg\max_a Q_h^\star(s,a)$$

State (action) Occupancy

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

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

State (action) Occupancy

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

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