Markov Decision Processes

Chen-Yu Wei

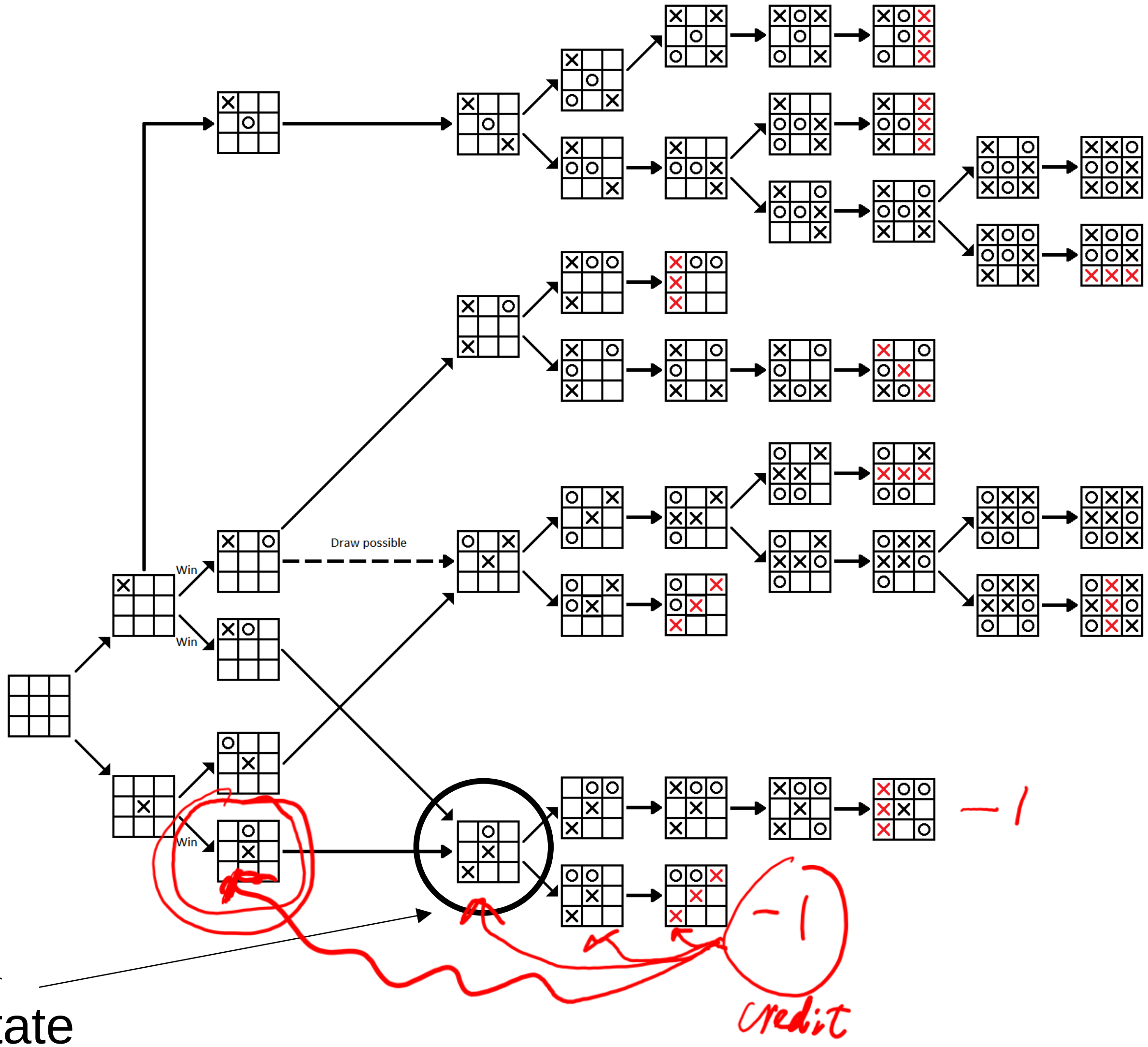
Sequence of Actions

To win the game, the learner has to take a sequence of actions $a_1 \to a_2 \to \cdots \to a_H$.

The effect of a particular action may not be revealed instantaneously.

- Some effect may be revealed instantaneously
- Some may be revealed later

Sequence of Actions



state

(a summary of the current status in a multi-stage game)

Interaction Protocol (Episodic Setting)

step

For **episode** $t = 1, 2, \ldots, T$:

$h \leftarrow 1$

Environment generates initial state $s_{t,1}$

While episode t has not ended:

Learner chooses an action $a_{t,h}$

Learner observes instantaneous reward $r_{t,h}$ with $\mathbb{E}[r_{t,h}] = R(s_{t,h}, a_{t,h})$

Environment generates next state $s_{t,h+1} \sim P(\cdot \mid s_{t,h}, a_{t,h})$

$h \leftarrow h + 1$

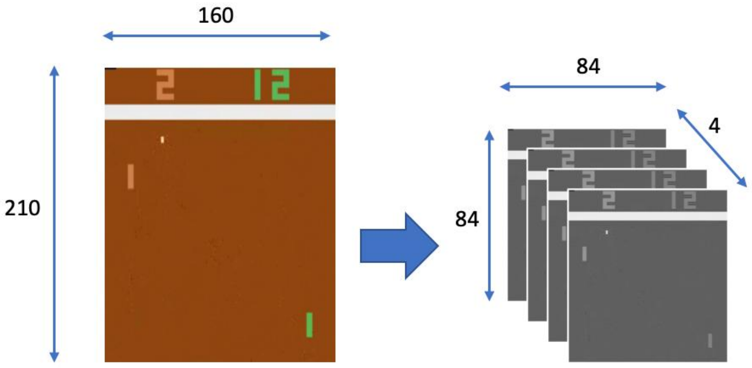
Markov assumption:

$r_{t,h}$ and $s_{t,h+1}$ are conditionally independent of $(s_{t,1}, a_{t,1}, \ldots, s_{t,h-1}, a_{t,h-1})$ given $s_{t,h}$

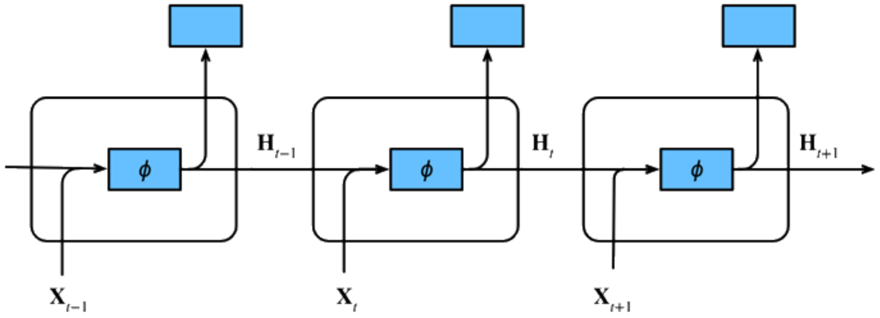
Goal: maximize $$\sum_{t=1}^{T} \sum_{h=1}^{\tau_t} R(s_{t,h}, a_{t,h})$$

τ_t : length of episode t

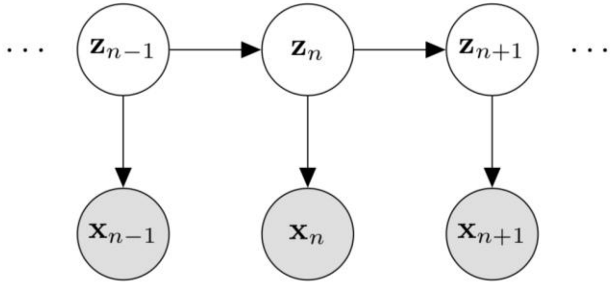
From Observations to States



Stacking recent observations



Recurrent neural network



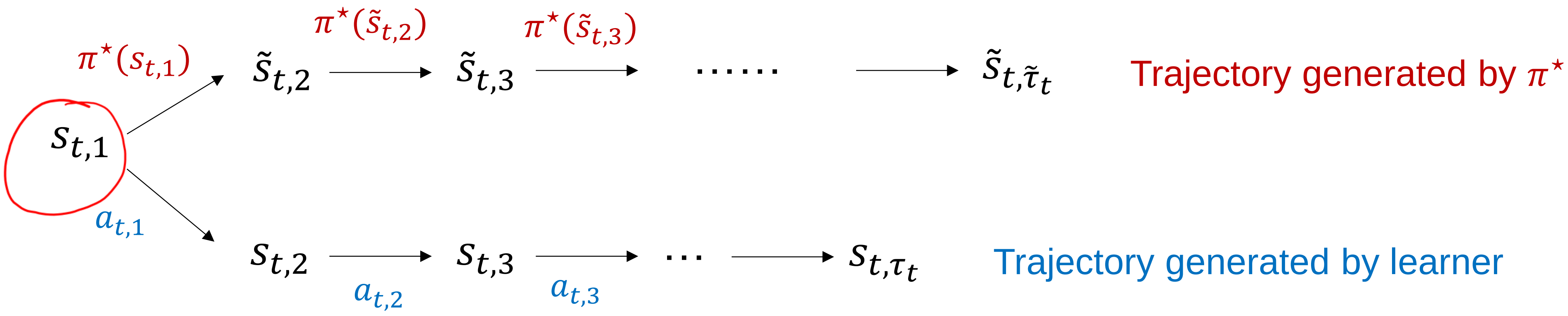
Hidden Markov model

Regret (Episodic Setting)

$\pi^*: S \to A$

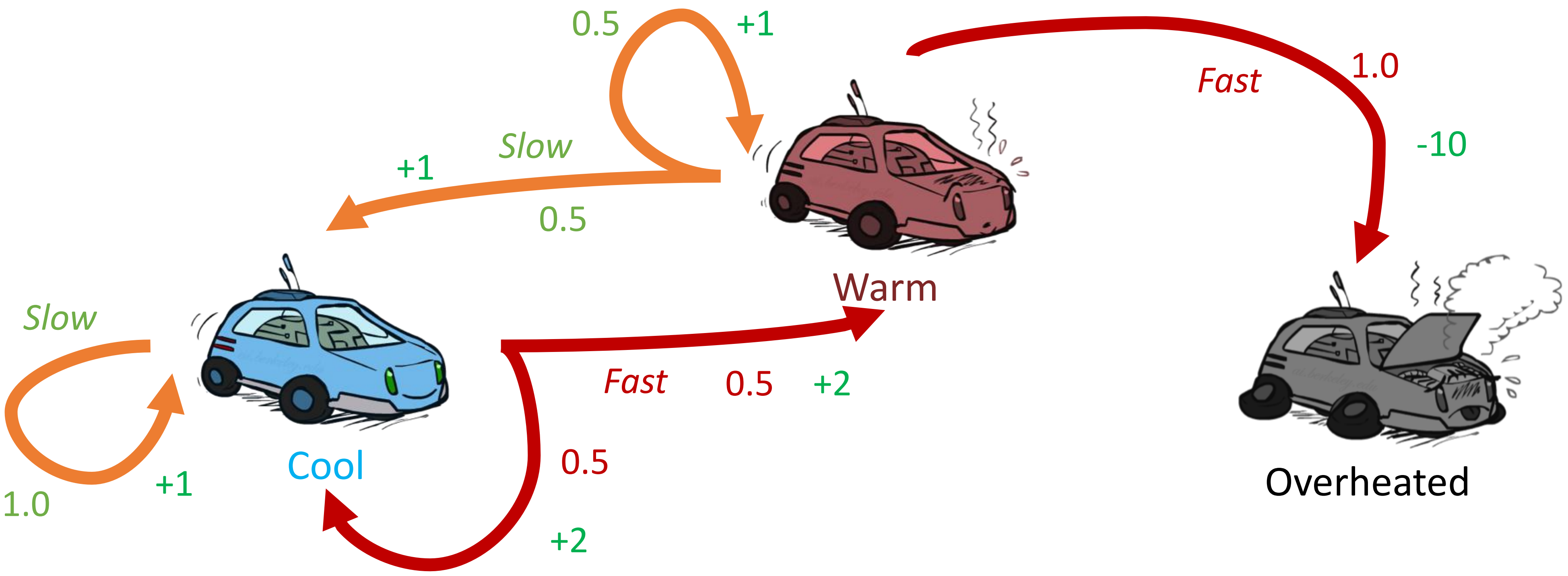
$$\text{Regret} = \underbrace{\max_{\pi^\star} \mathbb{E}^{\pi^\star}\left[\sum_{t=1}^{T}\sum_{h=1}^{\tilde{\tau}_t} R(\tilde{s}_{t,h}, \pi^\star(\tilde{s}_{t,h}))\right]}_{\text{Benchmark}} - \sum_{t=1}^{T}\sum_{h=1}^{\tau_t} R(s_{t,h}, a_{t,h})$$

CB $\max_{\pi^*} \sum_{t=1}^{T} R(x_t, \pi^*(x_t)) - \sum_{t=1}^{T} R(x_t, a_t)$



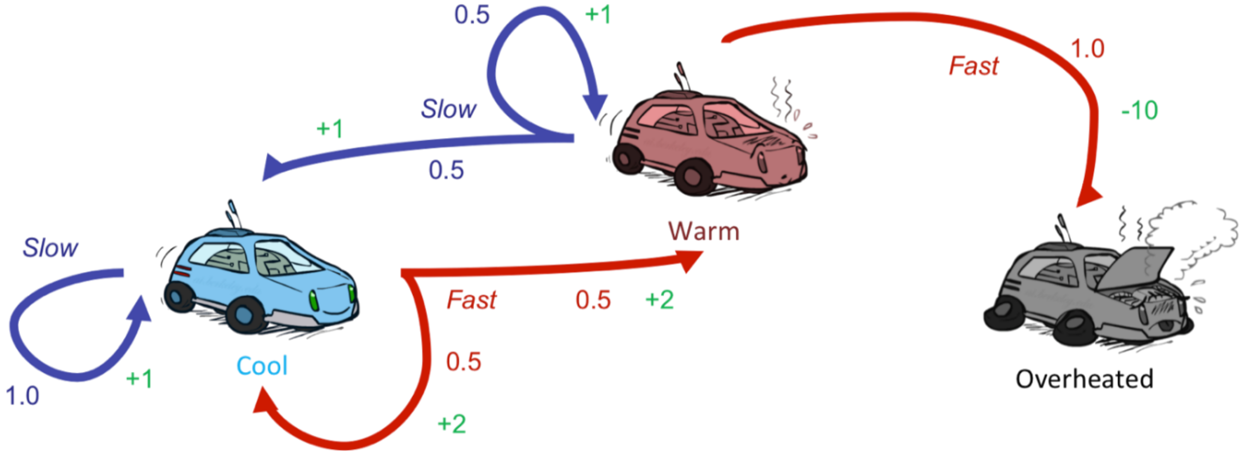
Example: Racing

- A robot car wants to travel far, quickly
- Three states: Cool, Warm, Overheated
- Two actions: *Slow*, *Fast*
- Going faster gets double reward



Example: Racing

s	a	s'	$P(s'\|s,a)$	$R(s,a)$
	Slow		1.0	+1
	Fast		0.5	+2
	Fast		0.5	+2
	Slow		0.5	+1
	Slow		0.5	+1
	Fast		1.0	–10
	(end)		1.0	0



Formulations

- Interaction Protocol
  - Fixed-Horizon
  - Variable-Horizon (Goal-Oriented)
  - Infinite-Horizon
- Performance Metric
  - Total Reward
  - Average Reward
  - Discounted Reward
- Policy
  - Markov policy
  - Stationary policy

Horizon = Length of an episode

Interaction Protocols (1/3): Fixed-Horizon

Horizon length is a fixed number H

$h \leftarrow 1$

Observe initial state $s_1 \sim \rho$

While $\boldsymbol{h \leq H}$:

Choose action a_h

Observe reward r_h with $\mathbb{E}[r_h] = R(s_h, a_h)$

Observe next state $s_{h+1} \sim P(\cdot\,|s_h, a_h)$

Examples: games with a fixed number of time

Interaction Protocols (2/3): Goal-Oriented

The learner interacts with the environment until reaching **terminal states** $\mathcal{T} \subset \mathcal{S}$

$h \leftarrow 1$

Observe initial state $s_1 \sim \rho$

While $s_h \notin \mathcal{T}$:

Choose action a_h

Observe reward r_h with $\mathbb{E}[r_h] = R(s_h, a_h)$

Observe next state $s_{h+1} \sim P(\cdot \,|s_h, a_h)$

$h \leftarrow h + 1$

Examples: video games, robotics tasks, personalized recommendations, etc.

Interaction Protocols (3/3): Infinite-Horizon

The learner continuously interacts with the environment

~~$h \leftarrow 1$~~

~~Observe initial state $s_1 \sim \rho$~~

Loop forever:

Choose action a_h

Observe reward r_h with $\mathbb{E}[r_h] = R(s_h, a_h)$

Observe next state $s_{h+1} \sim P(\cdot\,|s_h, a_h)$

$h \leftarrow h + 1$

Examples: network management, inventory management

Formulations

- Interaction Protocol
  - Fixed-Horizon
  - Variable-Horizon (Goal-Oriented)
  - Infinite-Horizon
- Performance Metric
  - Total Reward
  - Average Reward
  - Discounted Reward
- Policy
  - Markov policy
  - Stationary policy

Performance Metric

Total Reward (for episodic setting): $\sum_{h=1}^{\tau} r_h$ (τ: the step where the episode ends)

Average Reward (for infinite-horizon setting): $\lim_{H\to\infty} \frac{1}{H} \sum_{h=1}^{H} r_h$

Discounted Total Reward (for episodic or infinite-horizon): $\sum_{h=1}^{\tau} \gamma^{h-1} r_h$

τ: the step where the episode ends, or ∞ in the infinite-horizon case
$\gamma \in [0,1)$: discount factor

$\gamma = 0.99$

Interaction Protocols vs. Performance Metrics

	"natural" objective		
Fixed-Horizon	→	Total Reward	
Goal-Oriented	→	Total Reward	Could be unbounded
Infinite-horizon	→	Average Reward	Could have constant change for an infinitesimal change in policy

Discounted Total Reward?

Focusing more on the **recent** reward

There is a potential mismatch between our ultimate goal and what we optimized.

Formulations

- Interaction Protocol
  - Fixed-Horizon
  - Variable-Horizon (Goal-Oriented)
  - Infinite-Horizon
- Performance Metric
  - Total Reward
  - Average Reward
  - Discounted Reward
- Policy
  - Markov policy
  - Stationary policy

Policy for MDPs

$\pi = (\pi_1, \pi_2, \cdots, \pi_H, \cdots)$

Markov Policy

h : step index

$$a_h \sim \pi_h(\cdot \mid s_h) \in \Delta_A \quad \text{(space of dist)}$$
$$a_h = \pi_h(s_h) \in A$$

For **fixed-horizon** setting, there exists an optimal policy in this class ✓

Stationary Policy $\subseteq$ Markov Policy

$$a_h \sim \pi(\cdot \mid s_h)$$
$$a_h = \pi(s_h)$$

For **infinite-horizon/goal-oriented** settings, there exists an optimal policy in this class ✓

Fixed-horizon (Markov Policy) (total reward)

✓ Goal-oriented (Stationary Policy) (Discounted reward)

A **stationary policy** specifies

$\pi(\text{Slow} \mid \text{Cool})$

$\pi(\text{Fast} \mid \text{Cool})$

$\pi(\text{Slow} \mid \text{Warm})$

$\pi(\text{Fast} \mid \text{Warm})$

A **Markov policy** specifies

$\pi_h(\text{Slow} \mid \text{Cool})$

$\pi_h(\text{Fast} \mid \text{Cool})$

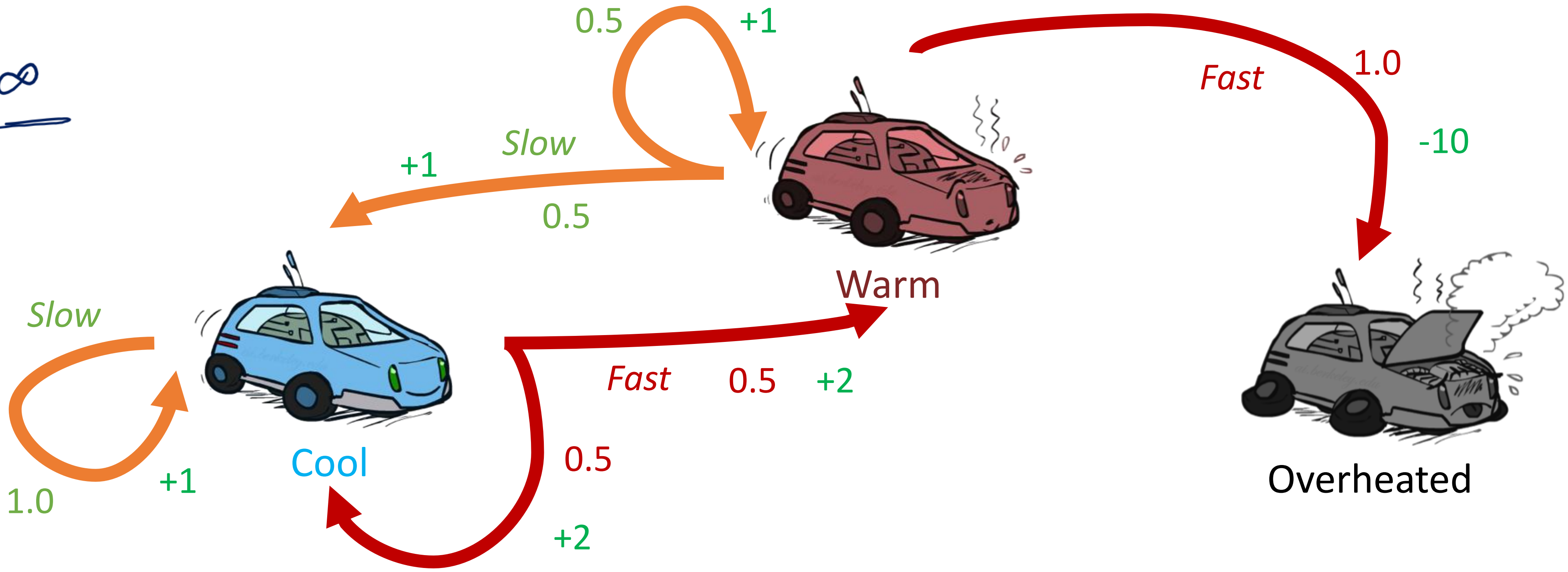
$\pi_h(\text{Slow} \mid \text{Warm})$

$\pi_h(\text{Fast} \mid \text{Warm})$

$\forall h$

$H = 5$

$H = \infty$



Value Iteration

(Fixed-Horizon)

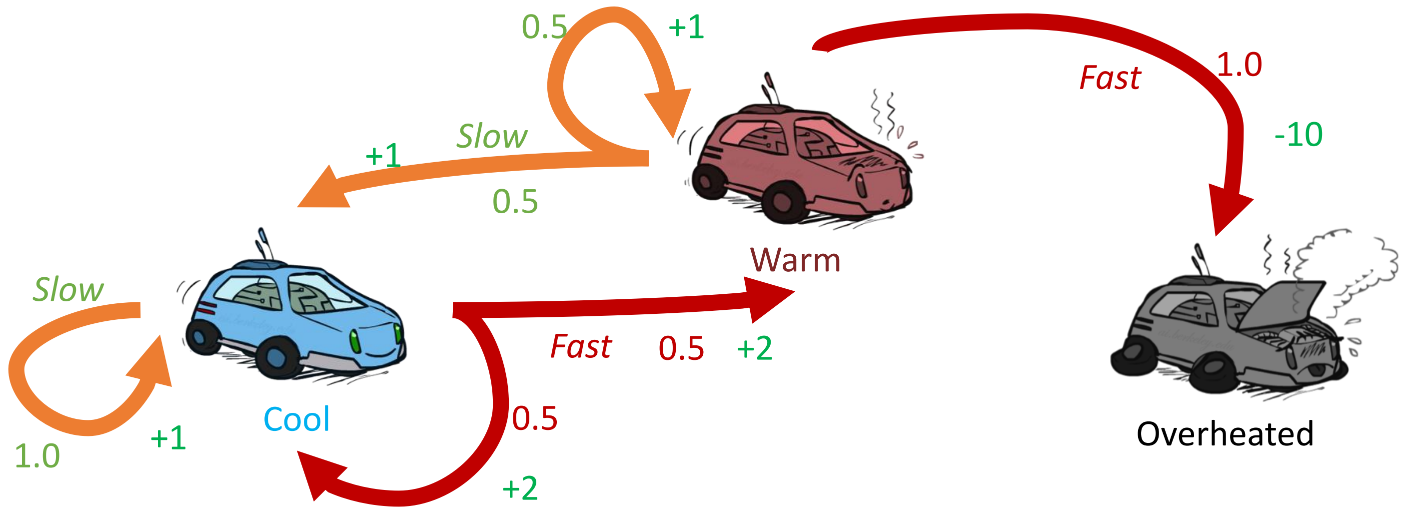
Two Tasks

- **Policy Evaluation:** Calculate the expected total reward of a given policy

  What is the expected total reward for the policy $\pi(\text{cool}) = \text{fast}, \pi(\text{warm}) = \text{slow}$?

  Policy Optimization: Find the best policy

  What is the policy that achieves the highest expected total reward?



Value Iteration for Policy Evaluation

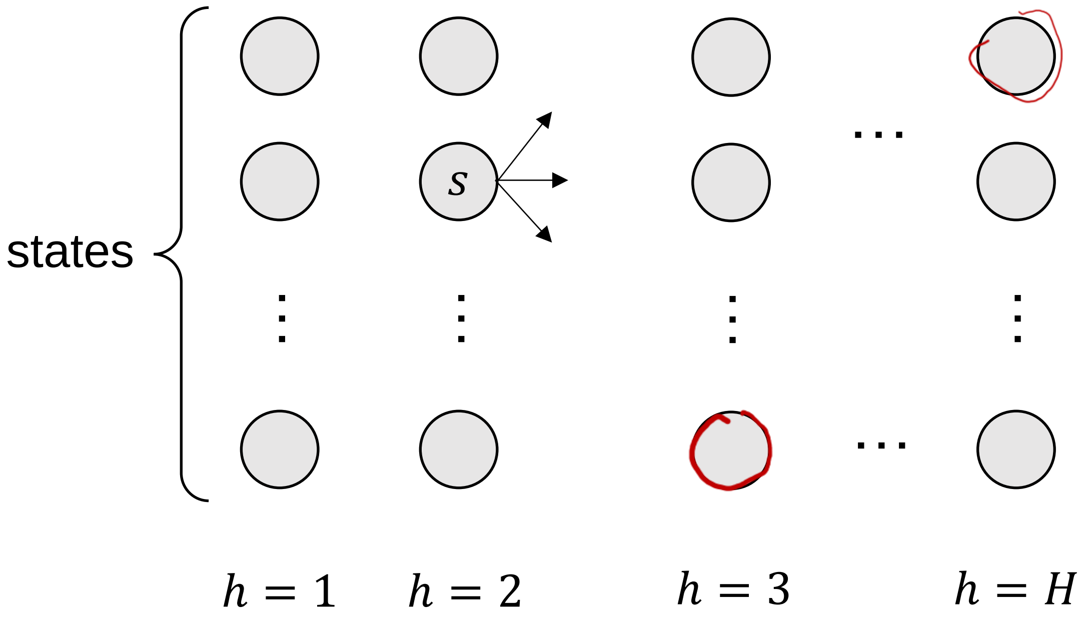
$\pi = (\pi_1, \cdots, \pi_H)$

$\mathbb{E}^{\pi}\left[\sum_{h=1}^{H} R(s_t, a_t)\right]$

$$Q_h^{\pi}(s,a) = \mathbb{E}^{\pi}\left[\sum_{k=h}^{H} R(s_k, a_k) \,\middle|\, (s_h, a_h) = (s,a)\right]$$

$$V_h^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{k=h}^{H} R(s_k, a_k) \,\middle|\, s_h = s\right]$$

$R(s,a)$



State transition: $P(s'|s,a)$

Reward: $R(s,a)$

$\boxed{V_1^{\pi}(s)}$

expected total

$= \sum_s \rho(s) V_1^{\pi}(s)$

Backward induction:

$Q_H^{\pi}(s,a) = R(s,a)$

$V_{H+1}^{\pi}(s) = 0 \qquad \forall s$

For $h = H, \ldots 1$: for all s, a

$$Q_h^{\pi}(s,a) = R(s,a) + \underbrace{\sum_{s'} P(s'|s,a)\, V_{h+1}^{\pi}(s')}_{\text{Expected total reward of } \pi \text{ from step } h+1}$$

$$V_h^{\pi}(s) = \sum_a \pi_h(a|s)\, Q_h^{\pi}(s,a)$$

Bellman Equation

Q_h^π is called "the state-action value functions of policy π"

V_h^π is called "the state value function of policy π"

Both can be just called "**value functions**"

$$Q_h^\pi(s,a) = R(s,a) + \sum_{s'} P(s'|s,a)\, V_{h+1}^\pi(s')$$

$$V_h^\pi(s) = \sum_a \pi_h(a|\, s) Q_h^\pi(s,a)$$

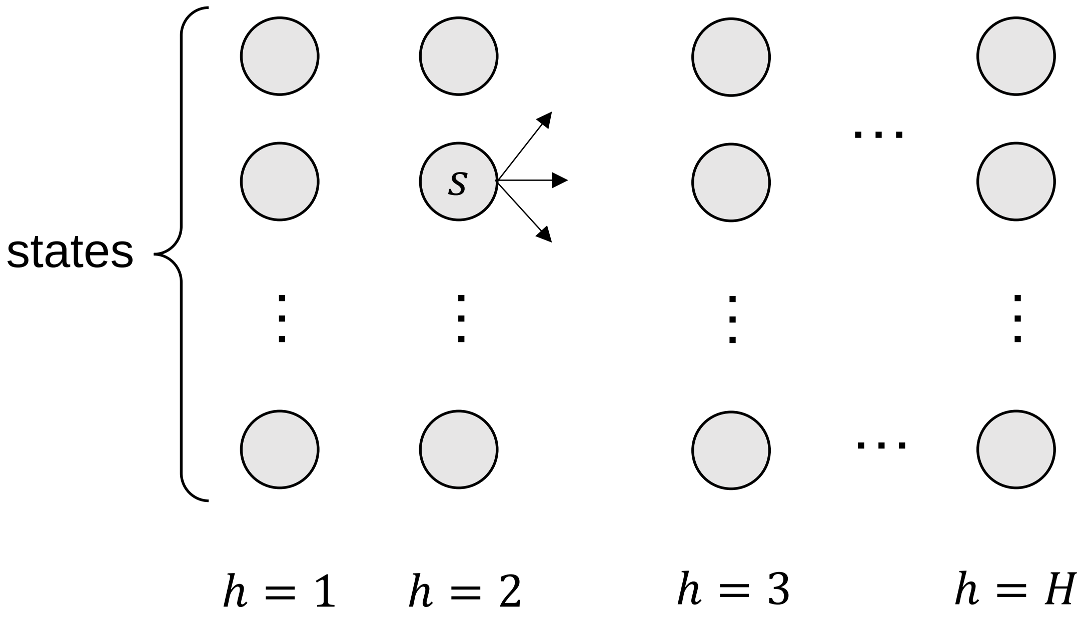
or

$$Q_h^\pi(s,a) = R(s,a) + \sum_{s',a'} P(s'|s,a)\, \pi_{h+1}(a'|s') Q_{h+1}^\pi(s',a')$$

or

$$V_h^\pi(s) = \sum_a \pi_h(a|\, s) \left(R(s,a) + \sum_{s'} P(s'|s,a)\, V_{h+1}^\pi(s') \right)$$

Value Iteration for Policy Optimization



State transition: $P(s'|s,a)$

Reward: $R(s,a)$

$$Q_h^\star(s,a) = \max_{\pi \in \Pi_M} \mathbb{E}^\pi \left[\sum_{k=h}^{H} R(s_k, a_k) \;\middle|\; (s_h, a_h) = (s,a) \right]$$

$$V_h^\star(s) = \max_{\pi \in \Pi_M} \mathbb{E}^\pi \left[\sum_{k=h}^{H} R(s_k, a_k) \;\middle|\; s_h = s \right]$$

Backward induction:

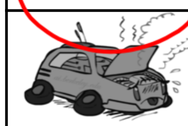
$V_{H+1}^\star(s) = 0 \qquad \forall s$

For $h = H, \dots 1$: for all s, a

$$Q_h^\star(s,a) = R(s,a) + \underbrace{\sum_{s'} P(s'|s,a)\, V_{h+1}^\star(s')}_{\text{Expected optimal total reward from step } h+1}$$

$$V_h^\star(s) = \max_a Q_h^\star(s,a) \qquad \pi_h^\star(s) = \operatorname*{argmax}_a Q_h^\star(s,a)$$

Exercise

s	a	s'	$P(s'\|s,a)$	$R(s,a)$
	Slow		1.0	+1
	Fast		0.5	+2
	Fast		0.5	+2
	Slow		0.5	+1
	Slow		0.5	+1
	Fast		1.0	−10
	(end)		1.0	0

Assume $H = 3$

$\boldsymbol{Q_3^\star(s,a)} = R(s,a)$

$Q_3^\star(\text{cool}, \text{slow}) = 1$

$Q_3^\star(\text{cool}, \text{fast}) = 2$

$Q_3^\star(\text{warm}, \text{slow}) = 1$

$Q_3^\star(\text{warm}, \text{fast}) = -10$

$\boldsymbol{V_3^\star(s)}$

$V_3^\star(\text{cool}) = 2$

$V_3^\star(\text{warm}) = 1$

$\boldsymbol{Q_2^\star(s,a)} = R(s,a) + \sum_{s'} P(s'|s,a)\, V_3^*(s')$

$Q_2^\star(\text{cool}, \text{slow}) = 1 + V_3^*(\text{cool}) = 3$

$Q_2^\star(\text{cool}, \text{fast}) = 2 + 0.5\, V_3^*(\text{cool}) + 0.5\, V_3^*(\text{warm}) = 3.5$

$Q_2^\star(\text{warm}, \text{slow}) = 1 + 0.5\, V_3^*(\text{cool}) + 0.5\, V_3^*(\text{warm}) = 2.5$

$Q_2^\star(\text{warm}, \text{fast}) = -10$

$\boldsymbol{V_2^\star(s)}$

$V_2^\star(\text{cool}) = 3.5$ $\quad \pi_2^*(\text{cool}) = \text{fast}$

$V_2^\star(\text{warm}) = 2.5$ $\quad \pi_2^*(\text{warm}) = \text{slow}$

Bellman Optimality Equation

$Q_h^\star$: optimal state-action value functions

$V_h^\star$: optimal state value functions

or "**optimal value functions**"

$$Q_h^\star(s,a) = R(s,a) + \sum_{s'} P(s'|s,a)\, V_{h+1}^\star(s')$$

$$V_h^\star(s) = \max_a Q_h^\star(s,a)$$

or

$$Q_h^\star(s,a) = R(s,a) + \sum_{s'} P(s'|s,a) \left(\max_{a'} Q_{h+1}^\star(s',a')\right)$$

or

$$V_h^\star(s) = \max_a \left(R(s,a) + \sum_{s'} P(s'|s,a)\, V_{h+1}^\star(s') \right)$$

$$\pi_h^\star(s) = \operatorname*{argmax}_a \; Q_h^\star(s,a)$$

Recall: Regret

$$\text{Regret} = \max_{\pi^\star} \mathbb{E}^{\pi^\star}\left[\sum_{t=1}^{T}\sum_{h=1}^{\tilde{\tau}_t} R(\tilde{s}_{t,h}, \pi^\star(\tilde{s}_{t,h}))\right] - \sum_{t=1}^{T}\sum_{h=1}^{\tau_t} R(s_{t,h}, a_{t,h})$$

$$\mathbb{E}[\text{Regret}] = \mathbb{E}\left[\sum_{t=1}^{T}\left(\underline{V_1^\star(s_{t,1})} - \underline{V_1^{\pi_t}(s_{t,1})}\right)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\left(\underline{V_1^\star(\rho)} - \underline{V_1^{\pi_t}(\rho)}\right)\right]$$

$$\boxed{V_1^\pi(\rho) \triangleq \mathbb{E}_{s\sim\rho}[V_1^\pi(s)]}$$

$\underline{s_{t,1} \sim \rho}$

Value Iteration

(Discounted Variable-Horizon) (or variable-horizon)

Value Iteration for Policy Evaluation

$Q_i^\pi(s,a) = R(s,a) + \mathbb{E}^\pi\left[\gamma \sum_{h=2}^{i} \gamma^{h-2} R(s_h,a_h) \,\middle|\, s_2 \sim P(\cdot|s,a)\right]$

$i-1$

$\mathbb{E}\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h,a_h)\right] = \mathbb{E}\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h,a_h)\right]$

$V(s)$

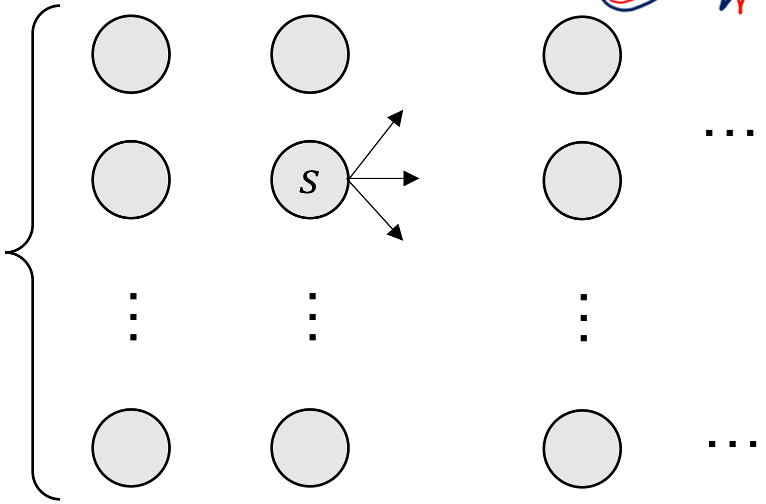
$$Q_i^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{h=1}^{i} \gamma^{h-1} R(s_h,a_h) \,\middle|\, (s_1,a_1) = (s,a)\right]$$

$$V_i^\pi(s) = \mathbb{E}^\pi\left[\sum_{h=1}^{i} \gamma^{h-1} R(s_h,a_h) \,\middle|\, s_1 = s\right]$$

For fixed horizon $i = (H+1) - h$

$$Q^\pi(s,a) = Q_\infty^\pi(s,a) \qquad V^\pi(s) = V_\infty^\pi(s)$$

states



$h=1$ $\quad h=2 \quad h=3$

weight $\quad 1 \quad \gamma \quad \gamma^2$

State transition: $P(s'|s,a)$

Reward: $R(s,a)$

$V_0^\pi(s) = 0 \quad \forall s$ $\qquad V_{H+1}^\pi(s) = 0$ $\quad$ fixed horizon

For $i = 1, 2, 3, \ldots$: $\quad$ for all s, a

$$Q_i^\pi(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}^\pi(s')$$

$$V_i^\pi(s) = \sum_a \pi(a|s)\, Q_i^\pi(s,a)$$

$Q^\pi(s,a) \approx Q_i^\pi(s,a)$

If $\left|Q_i^\pi(s,a) - Q_{i-1}^\pi(s,a)\right| \le \epsilon$ for all s, a: **terminate**

$$\begin{cases} Q^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h,a_h) \,\middle|\, (s_1,a_1)=(s,a)\right] \\ V^\pi(s) = \mathbb{E}^\pi\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h,a_h) \,\middle|\, \underline{s_1 = s}\right] = \underline{\sum_a \pi(a|s)\, \mathbb{E}^\pi\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h,a_h) \,\middle|\, s_1=s, a_1=a\right]} \end{cases}$$

$$= \sum_a \pi(a|s)\, Q^\pi(s,a)$$

$$Q^\pi(s,a) = R(s,a) + \underline{\mathbb{E}^\pi\left[\sum_{h=2}^{\infty} \gamma^{h-1} R(s_h,a_h) \,\middle|\, s_2 \sim p(\cdot|s,a)\right]}$$

$$= R(s,a) + \gamma \underline{\sum_{s'} p(s'|s,a)}\, \mathbb{E}^\pi\left[\sum_{h=2}^{\infty} \gamma^{h-2} R(s_h,a_h) \,\middle|\, s_2 = s'\right]$$

$$= R(s,a) + \gamma \sum_{s'} p(s'|s,a)\, \underbrace{\mathbb{E}^\pi\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h,a_h) \,\middle|\, s_1 = s'\right]}_{V^\pi(s')}$$

$$= R(s,a) + \gamma \sum_{s'} p(s'|s,a)\, V^\pi(s')$$

Bellman Equation

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

$\mathbb{E}_{s \sim p}\left(V^{\pi}(s)\right)$

$$Q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^{\pi}(s')$$

$$V^{\pi}(s) = \sum_{a} \pi(a|\, s) Q^{\pi}(s,a)$$

or

$$Q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s',a'} P(s'|s,a)\, \pi(a'|s') Q^{\pi}(s',a')$$

or

$$V^{\pi}(s) = \sum_{a} \pi(a|\, s) \left(R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^{\pi}(s') \right)$$

Convergence

$$|Q_i^{\pi}(s,a) - Q_{i-1}^{\pi}(s,a)| \leq \varepsilon \quad \forall s,a \qquad (*)$$

1. Value Iteration for policy evaluation will terminate.
2. When it terminates, it holds that

$$\left|Q_i^{\pi}(s,a) - Q^{\pi}(s,a)\right| \leq \frac{\epsilon}{1-\gamma} \quad \forall s,a$$

$$Q_i^{\pi}(s,a) = R(s,a) + \gamma \sum_{s',a'} P(s'|s,a)\,\pi(a'|s')\,Q_{i-1}^{\pi}(s',a')$$

$$= \boxed{R(s,a) + \gamma \sum_{s',a'} P(s'|s,a)\,\pi(a'|s')\,Q_i^{\pi}(s',a')} + \gamma \sum_{s',a'} P(s'|s,a)\,\pi(a'|s')\left(Q_{i-1}^{\pi}(s',a') - Q_i^{\pi}(s',a')\right)$$

$\in [-\varepsilon, \varepsilon]$ $\quad \leq \varepsilon$

If $(*)$ holds, then the last term can be upper bounded by $\gamma \cdot \varepsilon \leq \varepsilon$

$$\Rightarrow \left| Q_i^{\pi}(s,a) - \left(R(s,a) + \gamma \sum_{s',a'} P(s'|s,a)\,\pi(a'|s')\,Q_i^{\pi}(s',a')\right)\right| \leq \varepsilon$$

Convergence

1. Value Iteration for policy evaluation will terminate.
2. When it terminates, it holds that

$$\left|Q_i^\pi(s,a) - Q^\pi(s,a)\right| \le \frac{\epsilon}{1-\gamma} \quad \forall s,a$$

Proof strategy: (not the simplest proof)

1) Prove that VI will terminate (i.e., $\max_{s,a}\left|Q_i^\pi(s,a) - Q_{i-1}^\pi(s,a)\right| \le \epsilon$ will eventually holds)

2) At termination,

$$\text{BellmanError}(Q_i^\pi) = \max_{s,a}\left|Q_i^\pi(s,a) - \left(R(s,a) + \gamma \sum_{s',a'} P(s'|s,a)\pi(a'|s')Q_i^\pi(s',a')\right)\right| \le \epsilon$$

3) Use the **Value error** $\le (\mathbf{1-\gamma})^{-1}$ **Bellmen Error** lemma to claim

$$\left|Q_i^\pi(s,a) - Q^\pi(s,a)\right| \le \frac{\epsilon}{1-\gamma}.$$

Convergence (A More General Statement of 2.)

Value error $\leq (\mathbf{1}-\boldsymbol{\gamma})^{-\mathbf{1}}$ **Bellmen Error**

Let $f\colon\ \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ be **any** function (not necessarily generated by Value Iteration)

If

$$\left|f(s,a)-\left(R(s,a)+\gamma\sum_{s',a'}P(s'|s,a)\pi(a'|s')f(s',a')\right)\right|\leq\epsilon\quad\forall s,a$$

then

$$|f(s,a)-Q^{\pi}(s,a)|\leq\frac{\epsilon}{1-\gamma}\quad\forall s,a$$

Given π, Assume we have

$$\underline{f(s,a)} \overset{\geq}{\leq} \underline{\cancel{R(s,a)} + \gamma \sum_{s',a'} p(s'|s,a)\,\pi(a'|s')\,f(s',a')} + \varepsilon \overset{-\varepsilon}{} \qquad \forall s,a$$

$$-\big)\quad Q^{\pi}(s,a) = \cancel{R(s,a)} + \gamma \sum_{s',a'} p(s'|s,a)\,\pi(a'|s')\,Q^{\pi}(s',a') \qquad \forall s,a$$

$$\underline{f(s,a) - Q^{\pi}(s,a)} \leq \gamma \sum_{s',a'} \underbrace{p(s'|s,a)\,\pi(a'|s')}\Big(f(s',a') - Q^{\pi}(s',a')\Big) + \varepsilon \qquad \forall s,a$$

$$\leq \gamma \max_{s',a'} \Big(f(s',a') - Q^{\pi}(s',a')\Big) + \varepsilon$$

$$\Rightarrow \underbrace{\max_{s,a}\Big(f(s,a) - Q^{\pi}(s,a)\Big)}\leq \gamma \underbrace{\max_{s',a'}\Big(f(s',a') - Q^{\pi}(s',a')\Big)} + \varepsilon$$

$$\Rightarrow (\cancel{1-\gamma}) \max_{s,a}\Big(f(s,a) - Q^{\pi}(s,a)\Big) \leq \frac{\varepsilon}{1-\gamma}$$

Similarly:

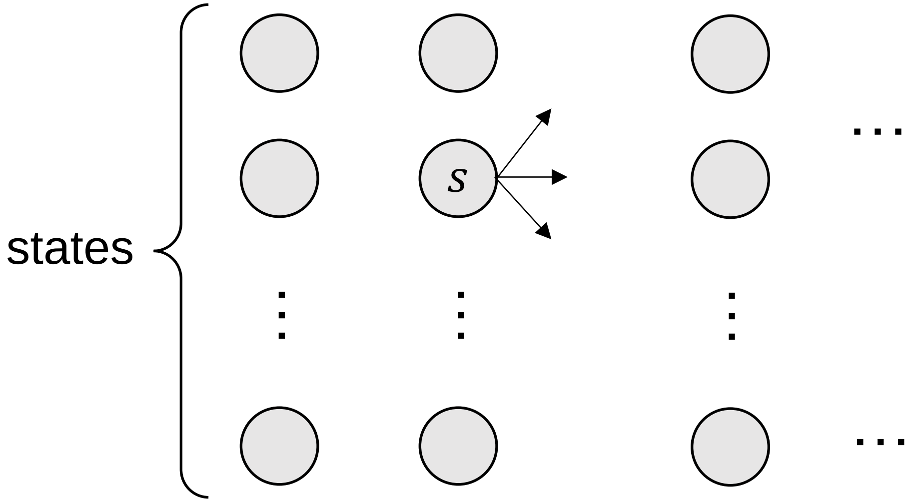
$$\left(\min_{s,a}\Big(f(s,a) - Q^{\pi}(s,a)\Big) \geq -\frac{\varepsilon}{1-\gamma}\right)$$

Value Iteration for Policy Optimization

$$Q_i^\star(s,a) = \max_\pi \mathbb{E}^\pi\left[\sum_{h=1}^{i} \gamma^{h-1} R(s_h,a_h) \,\middle|\, (s_0,a_0)=(s,a)\right]$$

$$V_i^\star(s) = \max_\pi \mathbb{E}^\pi\left[\sum_{h=1}^{i} \gamma^{h-1} R(s_h,a_h) \,\middle|\, s_0 = s\right]$$

$$Q^\star(s,a) = Q_\infty^\star(s,a) \qquad V^\star(s) = V_\infty^\star(s)$$



$h=1$ $h=2$ $h=3$

weight 1 γ γ^2

State transition: $P(s'|s,a)$

Reward: $R(s,a)$

$V_0^\star(s) = 0 \quad \forall s$

For $i = 1,2,3,\ldots$: for all s,a

$\neq Q^\pi(s,a)$

$$Q_i^\star(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}^\star(s')$$

$$V_i^\star(s) = \max_a Q_i^\star(s,a)$$

If $\left|Q_i^\star(s,a) - Q_{i-1}^\star(s,a)\right| \le \epsilon$ for all s,a: **terminate**

Bellman Optimality Equation

$$\pi^\star(s) = \underset{a}{\operatorname{argmax}}\ Q^\star(s,a)$$

$$Q^\star(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^\star(s')$$

$$V^\star(s) = \max_a Q^\star(s,a)$$

or

$$Q^\star(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q^\star(s',a')$$

or

$$V^\star(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^\star(s') \right)$$

Convergence

1. Value Iteration for policy optimization will terminate.
2. When it terminates, it holds that

$$\left|Q_i^\star(s,a) - Q^\star(s,a)\right| \leq \frac{\epsilon}{1-\gamma} \quad \forall s,a$$

3. When it terminates, it holds that

$$V^\star(s) - V^{\hat{\pi}}(s) \leq \frac{2\epsilon}{(1-\gamma)^2} \quad \forall s$$

where $\hat{\pi}(s) = \operatorname{argmax}_a Q_i^\star(s,a)$

$\pi^\star(s) = \operatorname{argmax}_a Q^\star(s,a)$

Convergence (A More General Statement of 2.)

Value error $\leq (\mathbf{1}-\boldsymbol{\gamma})^{-\mathbf{1}}$ **Bellmen Error**

Let $f\colon\ \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ be **any** function (not necessarily generated by Value Iteration)

If

$$\left|f(s,a)-\left(R(s,a)+\gamma\sum_{s'}P(s'|s,a)\max_{a'}f(s',a')\right)\right|\leq\epsilon\qquad\forall s,a$$

then

$$|f(s,a)-Q^{\star}(s,a)|\leq\frac{\epsilon}{1-\gamma}\qquad\forall s,a$$

Convergence (A More General Statement of 3.)

Suboptimality $\leq (\mathbf{1}-\boldsymbol{\gamma})^{-\mathbf{1}}$ Value Error

Let $f\colon\ \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be **any** function (not necessarily generated by Value Iteration)

If

$$|f(s,a) - Q^{\star}(s,a)| \leq \epsilon \qquad \forall s,a$$

then

$$V^{\star}(s) - V^{\pi_f}(s) \leq \frac{2\epsilon}{1-\gamma} \qquad \forall s$$

where $\pi_f(s) = \operatorname*{argmax}_a f(s,a)$

Review:

pure exploration $\qquad$ pure exploitation

$\underline{\hat{R}(a)}$ $\qquad$ $a_t = \arg\max_a \hat{R}(a)$

$\downarrow$

estimated value function

$$\boxed{\hat{a} = \arg\max_a \hat{R}(a)}$$

$$\forall a \quad |R(a) - \hat{R}(a)| \le \varepsilon$$

$$\boxed{a^* = \arg\max_a R(a)}$$

$$\underline{R(a^*) - R(\hat{a})} = \underbrace{\boxed{\hat{R}(a^*) - \hat{R}(\hat{a})}}_{\le 0} + \underbrace{R(a^*) - \hat{R}(a^*)}_{\le \varepsilon} + \underbrace{\hat{R}(\hat{a}) - R(\hat{a})}_{\le \varepsilon}$$

$$\le 2\varepsilon$$

Summary (Fixed Horizon)

Definitions

$$Q_h^\pi(s,a) \triangleq \mathbb{E}^\pi\left[\sum_{k=h}^{H} R(s_k,a_k) \;\middle|\; (s_h,a_h)=(s,a)\right]$$

$$V_h^\pi(s) \triangleq \mathbb{E}^\pi\left[\sum_{k=h}^{H} R(s_k,a_k) \;\middle|\; s_h = s\right]$$

$$Q_h^\star(s,a) \triangleq \max_\pi \mathbb{E}^\pi\left[\sum_{k=h}^{H} R(s_k,a_k) \;\middle|\; (s_h,a_h)=(s,a)\right]$$

$$V_h^\star(s) \triangleq \max_\pi \mathbb{E}^\pi\left[\sum_{k=h}^{H} R(s_k,a_k) \;\middle|\; s_h = s\right]$$

Relations (Bellman Equations)

$$Q_h^\pi(s,a) = R(s,a) + \sum_{s'} P(s'|s,a)\, V_{h+1}^\pi(s')$$

$$V_h^\pi(s) = \sum_a \pi_h(a|\, s) Q_h^\pi(s,a)$$

$$Q_h^\star(s,a) = R(s,a) + \sum_{s'} P(s'|s,a)\, V_{h+1}^\star(s')$$

$$V_h^\star(s) = \max_a Q_h^\star(s,a)$$

Calculation (VI)

Calculate $Q_h^\pi(s,a), V_h^\pi(s)\ \forall s,a$ from $h = H$ to $h = 1$

Calculate $Q_h^\star(s,a), V_h^\star(s)\ \forall s,a$ from $h = H$ to $h = 1$

Summary (Discounted Variable Horizon)

Definitions

$$Q^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h, a_h) \,\middle|\, (s_1, a_1) = (s,a)\right]$$

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h, a_h) \,\middle|\, s_1 = s\right]$$

$$Q^\star(s,a) = \max_\pi \mathbb{E}^\pi\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h, a_h) \,\middle|\, (s_1, a_1) = (s,a)\right]$$

$$V^\star(s) = \max_\pi \mathbb{E}^\pi\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h, a_h) \,\middle|\, s_1 = s\right]$$

Relations (Bellman Equations)

$$Q^\pi(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^\pi(s')$$

$$V^\pi(s) = \sum_a \pi(a|\, s) Q^\pi(s,a)$$

$$Q^\star(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^\star(s')$$

$$V^\star(s) = \max_a Q^\star(s,a)$$

Calculation (VI)

Calculate
$Q_i^\pi(s,a), V_i^\pi(s)\ \forall s, a$
for $i = 1, 2, \ldots$
until convergence

Calculate
$Q_i^\star(s,a), V_i^\star(s)\ \forall s, a$
for $i = 1, 2, \ldots$
until convergence

Policy Iteration

Policy Optimization

Policy Iteration

$\pi_i : S \to A$

Policy Iteration

$\pi_i(s) \neq \arg\max_a Q^{\pi_i}(s,a)$

For $i = 1, 2, \ldots$

$$\forall s, \qquad \pi_i(s) \leftarrow \arg\max_a \boxed{Q^{\pi_{i-1}}(s,a)}$$

Requires an inner VI for policy evaluation algo

Theorem (monotonic improvement). Policy Iteration ensures

$$\forall s, a, \qquad Q^{\pi_{i+1}}(s,a) \geq Q^{\pi_i}(s,a)$$

(We will prove this later.)

Generalized Policy Iteration

$N = \infty \Rightarrow$ Policy Iteration

(sub-routine: VI for policy evaluation)

$N = 1 \Rightarrow$ Value Iteration for policy optimization

For $i = 1, 2, \dots$

$\underbrace{Q_i(s,a)}= \underbrace{Q^{\pi_{i-1}}(s,a)}$ (inductive prove this)

$$\pi_i(s) = \max_a Q_i(s,a)$$ **Policy update**

$Q \leftarrow Q_i$

perform value iteration to evaluate π_i

Repeat for N times:

$$Q(s,a) \leftarrow R(s,a) + \gamma \sum_{s',a'} P(s'|s,a)\,\pi_i(a'|s')Q(s',a')$$ **Value update**

$Q_{i+1} \leftarrow Q$

$(s,a) \rightarrow \mathbb{R}$

Notice: in value iteration for PO, there may not exist a policy π such that $Q_i = Q^{\pi}$

$\boxed{Q_i = Q^{\pi_{i-1}}}$

In contrast, in policy iteration we have $Q_i = Q^{\pi_{i-1}}$

VI for PO can be viewed as PI **with incomplete policy evaluation**

Summary

- Value Iteration for Policy Optimization (VI for PO)
  - Is essentially a **dynamic programming** algorithm
  - Finds the value functions of the optimal policy

- Value Iteration for Policy Evaluation (VI for PE)
  - Also a **dynamic programming** algorithm
  - Finds the value functions of the given policy

- Policy Iteration (PI)
  - An iterative policy improvement algorithm (for PO)
  - Each iteration involves a policy evaluation subtask

- VI for PO and PI can be viewed as special cases of Generalized PI

Performance Difference Lemma

Unanswered Questions

- For an estimation $\hat{Q}(s,a) \approx Q^\star(s,a)$ with error, how can we bound

$$V^\star(\rho) - V^{\hat{\pi}}(\rho) \qquad \text{where } \hat{\pi}(s) = \max_a \hat{Q}(s,a)?$$

- How to show that Policy Iteration leads to monotonic policy improvement?

- Also, how are these methods related to the third challenge of online RL: credit assignment?

Performance Difference Lemma

For any two stationary policies π' and π in the discounted setting,

$$\mathbb{E}_{s\sim\rho}\left[V^{\pi'}(s)\right] - \mathbb{E}_{s\sim\rho}[V^{\pi}(s)] = \sum_{s,a} d_\rho^{\pi'}(s)\left(\pi'(a|s) - \pi(a|s)\right)Q^{\pi}(s,a)$$

$$= \sum_{s} d_\rho^{\pi'}(s,a)\left(Q^{\pi}(s,a) - V^{\pi}(s)\right)$$

$$d_\rho^{\pi}(s) \triangleq \mathbb{E}^{\pi}\left[\sum_{h=1}^{\infty}\gamma^{h-1}\mathbb{I}\{s_h = s\} \,\middle|\, s_1 \sim \rho\right]$$

Discounted occupancy measure on state s

$$d_\rho^{\pi}(s,a) \triangleq \mathbb{E}^{\pi}\left[\sum_{h=1}^{\infty}\gamma^{h-1}\mathbb{I}\{(s_h,a_h) = (s,a)\} \,\middle|\, s_1 \sim \rho\right]$$

Performance Difference Lemma

We can also swap the roles of π' and π and apply the same lemma

$$\mathbb{E}_{s\sim\rho}[V^{\pi}(s)] - \mathbb{E}_{s\sim\rho}\left[V^{\pi'}(s)\right] = \sum_{s,a} d_\rho^{\pi}(s)\left(\pi(a|s) - \pi'(a|s)\right)Q^{\pi'}(s,a)$$

$\times(-1)$

$$\Rightarrow \mathbb{E}_{s\sim\rho}\left[V^{\pi'}(s)\right] - \mathbb{E}_{s\sim\rho}[V^{\pi}(s)] = \sum_{s,a} d_\rho^{\pi}(s)\left(\pi'(a|s) - \pi(a|s)\right)Q^{\pi'}(s,a)$$

$=$

Original version:

$$\mathbb{E}_{s\sim\rho}\left[V^{\pi'}(s)\right] - \mathbb{E}_{s\sim\rho}[V^{\pi}(s)] = \sum_{s,a} d_\rho^{\pi'}(s)\left(\pi'(a|s) - \pi(a|s)\right)Q^{\pi}(s,a)$$

Performance Difference Lemma (Fixed-Horizon)

For any two Markov policies π' and π in the fixed-horizon setting,

$$\mathbb{E}_{s_1 \sim \rho}\left[V_1^{\pi'}(s_1)\right] - \mathbb{E}_{s_1 \sim \rho}[V_1^{\pi}(s_1)] = \sum_{h=1}^{H} \sum_{s,a} d_{\rho,h}^{\pi'}(s)\left(\pi'_h(a|s) - \pi_h(a|s)\right) Q_h^{\pi}(s,a)$$

$$= \sum_{h=1}^{H} \sum_{s,a} d_{\rho,h}^{\pi'}(s,a)\left(Q_h^{\pi}(s,a) - V_h^{\pi}(s)\right)$$

$$d_{\rho,h}^{\pi}(s) \triangleq \mathbb{E}^{\pi}[\mathbb{I}\{s_h = s\} \mid s_1 \sim \rho] = \mathbb{P}^{\pi}(s_h = s \mid s_1 \sim \rho)$$

$$d_{\rho,h}^{\pi}(s,a) \triangleq \mathbb{E}^{\pi}[\mathbb{I}\{(s_h, a_h) = (s,a)\} \mid s_1 \sim \rho] = \mathbb{P}^{\pi}((s_h, a_h) = (s,a) \mid s_1 \sim \rho)$$

The Meaning of Performance Difference Lemma

It tells us how **credit** are **assigned** to each state/step

The sub-optimality of a policy π:

$$\mathbb{E}_{s\sim\rho}[V^\star(s)] - \mathbb{E}_{s\sim\rho}[V^\pi(s)]$$

$$= \sum_{s,a} d_\rho^\pi(s)\left(\pi^\star(a|s) - \pi(a|s)\right)Q^{\pi^\star}(s,a)$$

$$= \sum_{s,a} d_\rho^\pi(s,a)\left(V^\star(s) - Q^\star(s,a)\right)$$

$$= \sum_{s,a} d_\rho^{\pi^\star}(s)\left(\pi^\star(a|s) - \pi(a|s)\right)Q^\pi(s,a)$$

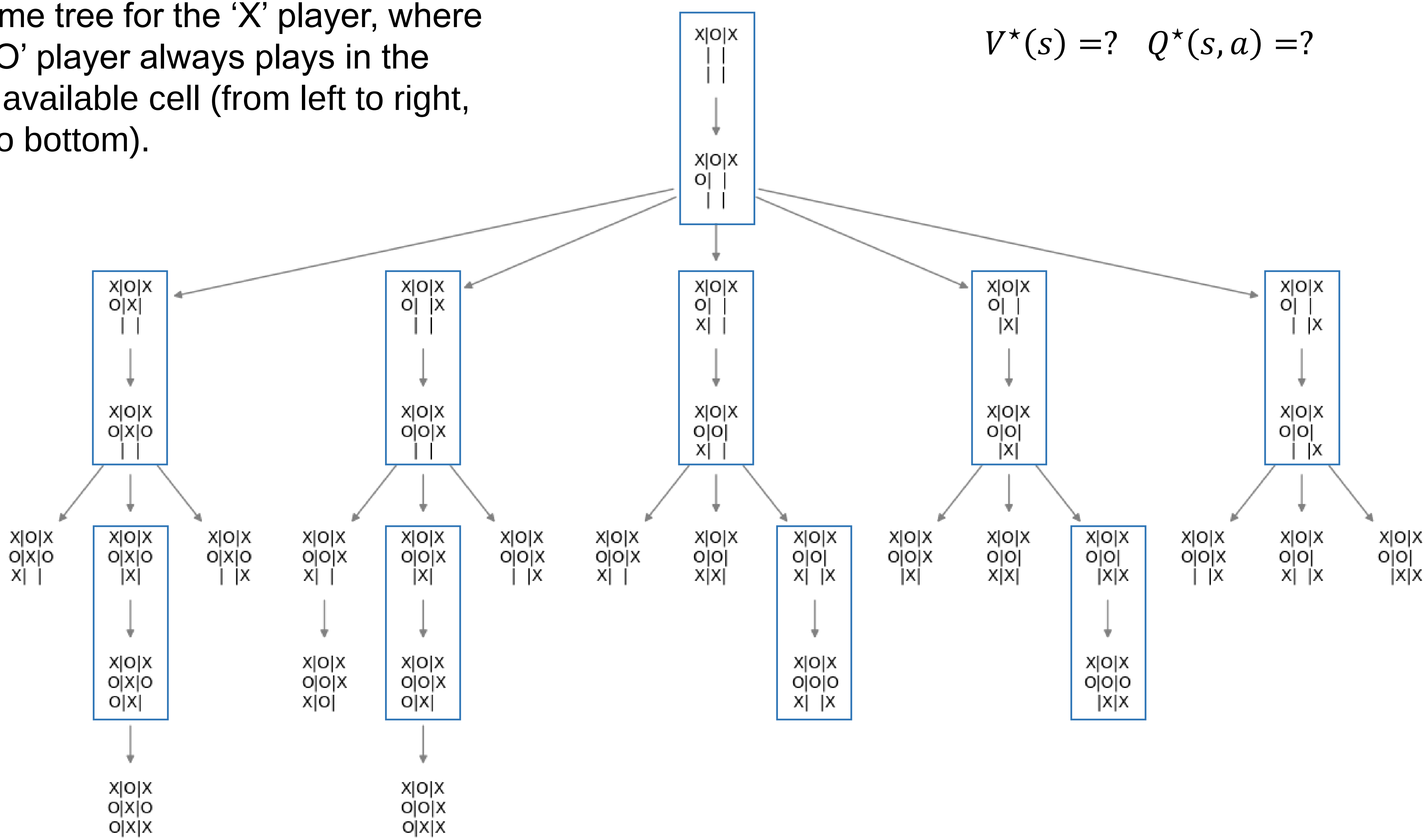
$$= \sum_{s,a} d_\rho^{\pi^\star}(s,a)\left(Q^\pi(s,a) - V^\pi(s)\right)$$

If π is highly sub-optimal, then we can always find

1) An (s,a)-pair on the path of π such that $V^\star(s) - Q^\star(s,a)$ is positive and large
2) An (s,a)-pair on the path of $\pi^\star$ such that $Q^\pi(s,a) - V^\pi(s)$ is positive and large

A game tree for the 'X' player, where the 'O' player always plays in the **first** available cell (from left to right, top to bottom).

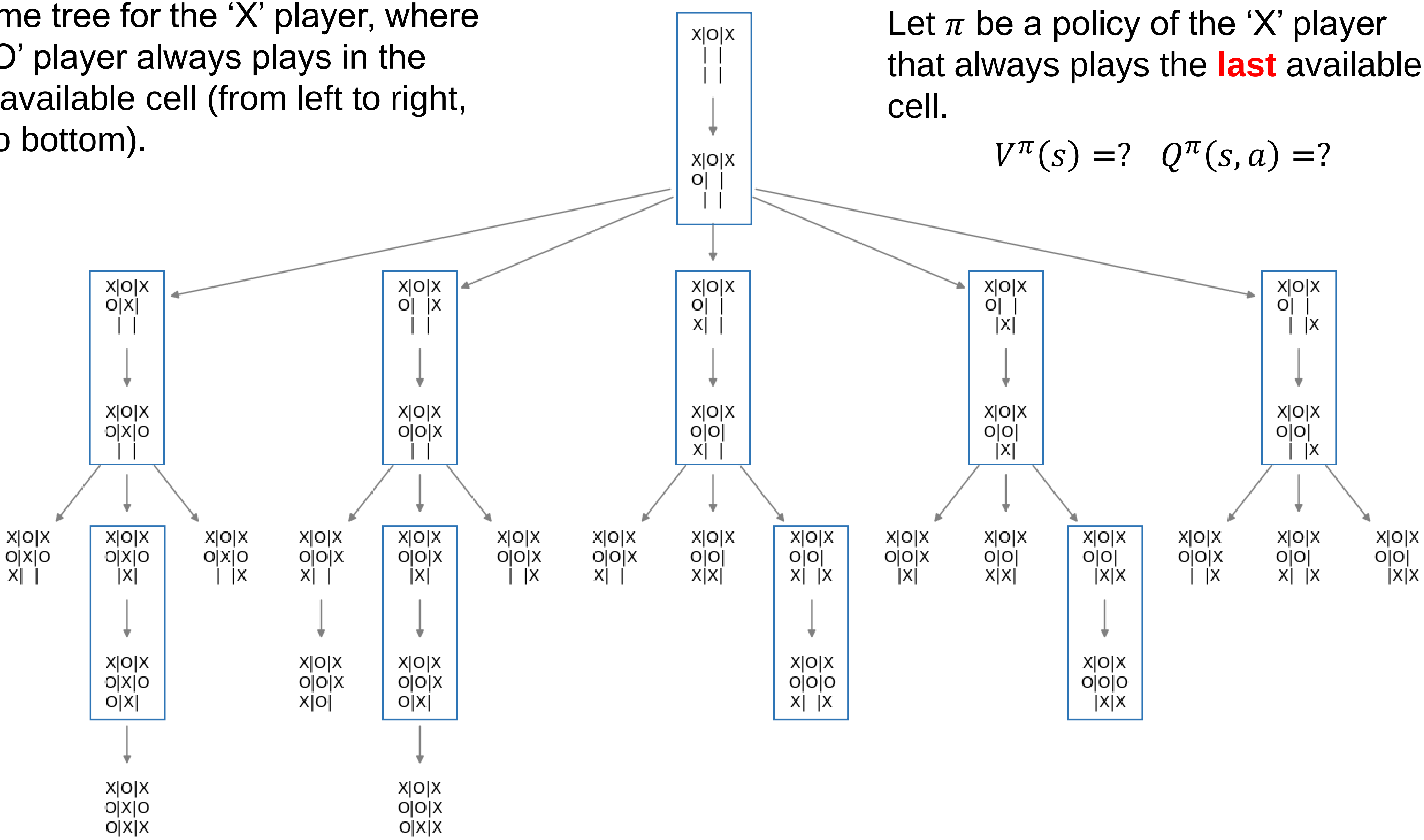
$V^\star(s) = ?\quad Q^\star(s,a) = ?$



A game tree for the ‘X’ player, where the ‘O’ player always plays in the **first** available cell (from left to right, top to bottom).

Let π be a policy of the ‘X’ player that always plays the **last** available cell.

$$V^{\pi}(s) = ? \quad Q^{\pi}(s, a) = ?$$



Proof (Sketch) of Performance Difference Lemma

Unanswered Question 1

- For an estimation $\hat{Q}(s,a) \approx Q^\star(s,a)$ with error, how can we bound

$$V^\star(\rho) - V^{\hat{\pi}}(\rho) \qquad \text{where } \hat{\pi}(s) = \max_a \hat{Q}(s,a)?$$

Unanswered Question 2

- Why Policy Iteration leads to monotonic policy improvement?