facebook
Artificial Intelligence Research

Exploration-Exploitation in Reinforcement Learning

Finite-Horizon MDPs

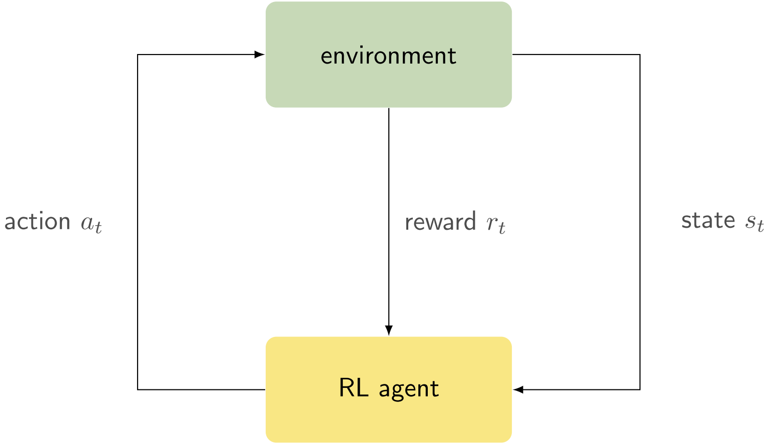
Matteo Pirotta
Facebook AI Research

Acknowledgements

These slides are part of a longer tutorial on exploration-exploitation in RL.

`https://rlgammazero.github.io/`

RL Agent-Environment Interaction



Website

https://rlgammazero.github.io

Markov Decision Process

[Puterman, 1994]

A **finite-horizon** Markov decision process (MDP) is a tuple $M = \langle \mathcal{S}, \mathcal{A}, r_h, p_h, H \rangle$

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Horizon H
- Transition distribution $p_h(\cdot|s,a) \in \Delta(\mathcal{S})$, $h = 1, \dots, H$
- Reward distribution with expectation $r_h(s,a) \in [0,1]$, $h = 1, \dots, H$

An agent acts according to a *time-variant policy*

$$\pi_h : \mathcal{S} \to \mathcal{A} \qquad h = 1, \dots, H$$

Markov Decision Process

[Puterman, 1994]

A **finite-horizon** Markov decision process (MDP) is a tuple $M = \langle \mathcal{S}, \mathcal{A}, r_h, p_h, H \rangle$

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Horizon H
- Transition distribution $p_h(\cdot|s,a) \in \Delta(\mathcal{S})$, $h = 1, \ldots, H$
- Reward distribution with expectation $r_h(s,a) \in [0,1]$, $h = 1, \ldots, H$

An agent acts according to a *time-variant policy*

$$\pi_h : \mathcal{S} \to \mathcal{A} \qquad h = 1, \ldots, H$$

In (contextual) bandit, actions do not influence the evolution of states

Value Functions and Optimality

Value functions

$$Q_h^{\pi}(s,a) = r_h(s,a) + \mathbb{E}\left[\sum_{l=h+1}^{H} r_l(s_l, \pi_l(s_l))\right]$$

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

Optimality

$$Q_h^{\star}(s,a) = \sup_{\pi} Q_h^{\pi}(s,a)$$

$$\pi_h^{\star}(s) = \arg\max_{a \in \mathcal{A}} Q_h^{\star}(s,a)$$

Value Functions and Optimality

Value functions

$$Q_h^{\pi}(s,a) = r_h(s,a) + \mathbb{E}\Bigg[\sum_{l=h+1}^{H} r_l(s_l, \pi_l(s_l))\Bigg]$$
$$V_h^{\pi}(s) = Q_h^{\pi}(s, \pi_h(s))$$

Optimality

$$Q_h^{\star}(s,a) = \sup_{\pi} Q_h^{\pi}(s,a)$$
$$\pi_h^{\star}(s) = \arg\max_{a \in \mathcal{A}} Q_h^{\star}(s,a)$$

Remark: given $r_h(s,a) \in [0,1]$, then $Q_h(s,a), V_h(s) \in [0, H-(h-1)]$

Bellman Equations

Policy Bellman equation

$$\begin{aligned} Q_h^\pi(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[Q_{h+1}^\pi(s', \pi_{h+1}(s'))\right] \\ &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[V_{h+1}^\pi(s')\right] \end{aligned}$$

Optimal Bellman equation

$$\begin{aligned} Q_h^\star(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q_{h+1}^\star(s', a')\right] \\ &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[V_{h+1}^\star(s')\right] \end{aligned}$$

Value Iteration (aka Backward Induction)

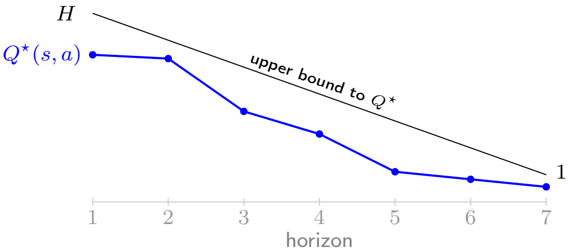
Input: $\mathcal{S}$, $\mathcal{A}$, r_h, p_h
Set $Q^\star_{H+1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

for $h = H, \dots, 1$ **do**
  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    Compute

$$
\begin{aligned}
Q^\star_h(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q^\star_{h+1}(s',a')\right] \\
&= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[V^\star_{h+1}(s')\right]
\end{aligned}
$$

  end
end
return $\pi^\star_h(s) = \arg\max_{a \in \mathcal{A}} Q^\star_h(s,a)$

Value Iteration (aka Backward Induction)



$$Q^\star_h(s,a) = \max_a \left\{ r_h(s,a) + \mathbb{E}_{s'|s,a}[V^\star_{h+1}(s')] \right\}$$

Online Learning Problem

Input: $\mathcal{S}, \mathcal{A}$ $\cancel{r_h, p_h}$

Initialize $Q_{h1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h = 1, \dots, H$, $\mathcal{D}_1 = \emptyset$

for $k = 1, \dots, K$ **do** `// episodes`

  Define π_k based on $(Q_{hk})_{h=1}^H$

  Observe initial state s_{1k} *(arbitrary)*

  for $h = 1, \dots, H$ **do**

    Execute $a_{hk} = \pi_{hk}(s_{hk})$

    Observe r_{hk} and $s_{h+1,k}$

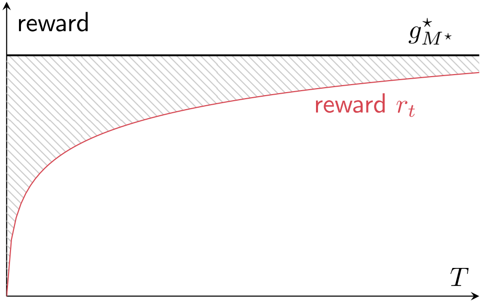
  end

  Add trajectory $(s_{hk}, a_{hk}, r_{hk})_{h=1}^H$ to $\mathcal{D}_{k+1}$

  Compute $(Q_{h,k+1})_{h=1}^H$ from $\mathcal{D}_{k+1}$

end

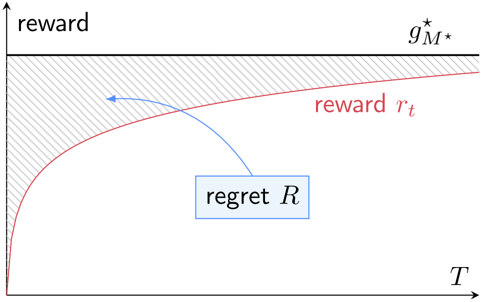
Frequentist Regret



$$R(K,\ M^\star,\ \mathfrak{A}) = \sum_{k=1}^{K} \Big(V^\star(s_{1k}) - V^{\pi_k}(s_{1k}) \Big)$$

Let $T = HK$ total number of steps executed in the environment

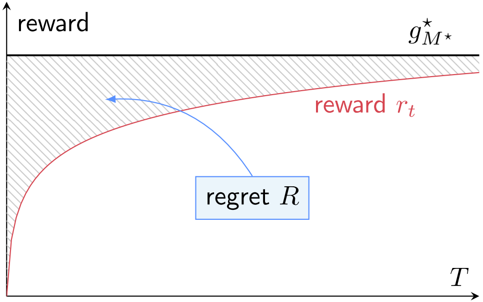
Frequentist Regret



$$R(K,\ M^\star,\ \mathfrak{A}) = \sum_{k=1}^{K} \left(V^\star(s_{1k}) - V^{\pi_k}(s_{1k}) \right)$$

Let $T = HK$ total number of steps executed in the environment

Frequentist Regret



unknown true MDP
$M^\star = \langle \mathcal{S}, \mathcal{A}, r, p, H \rangle$

algorithm $\mathfrak{A} = \{\pi_k\}_{k=1}^K$

$$R(K, M^\star, \mathfrak{A}) = \sum_{k=1}^{K} \left(V^\star(s_{1k}) - V^{\pi_k}(s_{1k}) \right)$$

policy selected by $\mathfrak{A}$

Let $T = HK$ total number of steps executed in the environment

Alternative Models

- Infinite-horizon undiscounted MDPs (average reward)
  $\Rightarrow$ regret minimization

- Infinite-horizon discounted MDPs
  $\Rightarrow$ PAC-MDPs

$$N(M^\star, \mathfrak{A}) = \sum_{t=0}^{\infty} \mathbb{I}\left\{ V^{\pi_t}(s_t) \leq V^\star(s_t) - \epsilon \right\}$$

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_{hk} = \begin{cases} \arg\max_{a \in \mathcal{A}} Q_{hk}(s_{hk}, a) & \text{w.p. } 1 - \epsilon_{hk}, \\ \mathcal{U}(\mathcal{A}) & \text{otherwise.} \end{cases}$$

- Q-learning update

$$Q_{h,k+1}(s_{hk}, a_{hk}) = (1 - \alpha_t) Q_{hk}(s_{hk}, a_{hk}) + \alpha_t \Big(r_{hk} + \max_{a' \in \mathcal{A}} Q_{h+1,k}(s_{h+1,k}, a') \Big)$$

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_{hk} = \begin{cases} \arg\max_{a \in \mathcal{A}} Q_{hk}(s_{hk}, a) & \text{w.p. } 1 - \epsilon_{hk}, \\ \mathcal{U}(\mathcal{A}) & \text{otherwise.} \end{cases}$$

- Q-learning update

$$Q_{h,k+1}(s_{hk}, a_{hk}) = (1 - \alpha_t) Q_{hk}(s_{hk}, a_{hk}) + \alpha_t \big(r_{hk} + \max_{a' \in \mathcal{A}} Q_{h+1,k}(s_{h+1,k}, a') \big)$$

The exploration strategy relies on **biased** estimates Q_{hk}

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_{hk} = \begin{cases} \arg\max_{a \in \mathcal{A}} Q_{hk}(s_{hk}, a) & \text{w.p. } 1 - \epsilon_{hk}, \\ \mathcal{U}(\mathcal{A}) & \text{otherwise.} \end{cases}$$

- Q-learning update

$$Q_{h,k+1}(s_{hk}, a_{hk}) = (1 - \alpha_t) Q_{hk}(s_{hk}, a_{hk}) + \alpha_t \big(r_{hk} + \max_{a' \in \mathcal{A}} Q_{h+1,k}(s_{h+1,k}, a') \big)$$

- The exploration strategy relies on **biased** estimates Q_{hk}
- Samples are used **once**

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_{hk} = \begin{cases} \arg\max_{a \in \mathcal{A}} Q_{hk}(s_{hk}, a) & \text{w.p. } 1 - \epsilon_{hk}, \\ \mathcal{U}(\mathcal{A}) & \text{otherwise.} \end{cases}$$

- Q-learning update

$$Q_{h,k+1}(s_{hk}, a_{hk}) = (1 - \alpha_t) Q_{hk}(s_{hk}, a_{hk}) + \alpha_t \Big(r_{hk} + \max_{a' \in \mathcal{A}} Q_{h+1,k}(s_{h+1,k}, a') \Big)$$

- The exploration strategy relies on **biased** estimates Q_{hk}
- Samples are used **once**
- **Dithering effect:** exploration is not effective in covering the state space
- **Policy shift:** the policy changes at each step

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_{hk} = \begin{cases} \arg\max_{a \in \mathcal{A}} Q_{hk}(s_{hk}, a) & \text{w.p. } 1 - \epsilon_{hk}, \\ \mathcal{U}(\mathcal{A}) & \text{otherwise.} \end{cases}$$

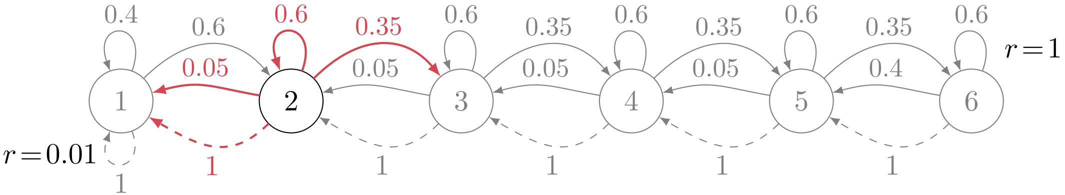
- Q-learning update

$$Q_{h,k+1}(s_{hk}, a_{hk}) = (1 - \alpha_t) Q_{hk}(s_{hk}, a_{hk}) + \alpha_t \big(r_{hk} + \max_{a' \in \mathcal{A}} Q_{h+1,k}(s_{h+1,k}, a') \big)$$

- The exploration strategy relies on **biased** estimates Q_{hk}
- Samples are used **once**
- **Dithering effect:** exploration is not effective in covering the state space
- **Policy shift:** the policy changes at each step
- **Regret:** $\Omega\Big(\min\{T, A^{H/2}\}\Big)$ [Jin et al., 2018]

River Swim: Markov Decision Processes

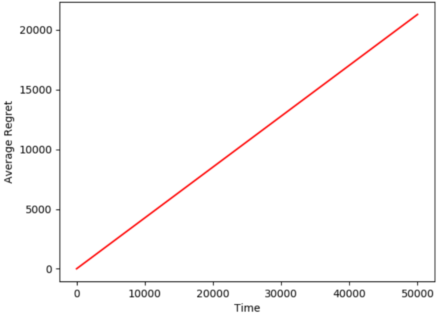
Strehl and Littman [2008]

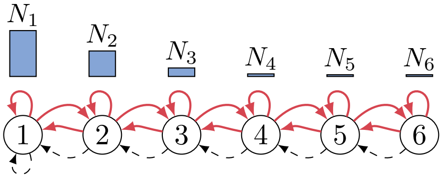


- $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A} = \{L, R\}$
- $\pi_L(s) = L$, $\pi_R(s) = R$

River Swim: Q-learning w\ ϵ-greedy Exploration

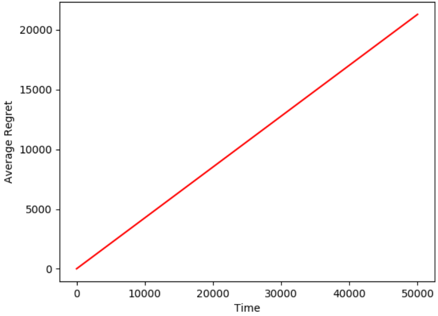
- $\epsilon_t = 1.0$

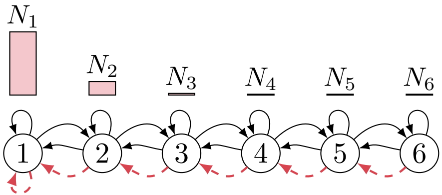




River Swim: Q-learning w\ ϵ-greedy Exploration

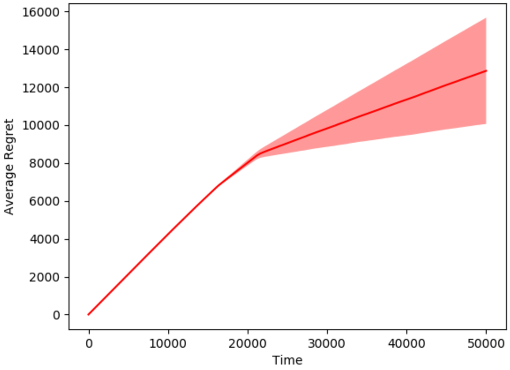
- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$





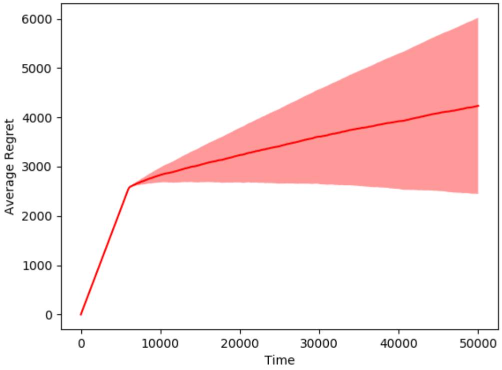
River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$
- $\epsilon_t = \dfrac{\epsilon_0}{(N(s_t) - 1000)^{2/3}}$



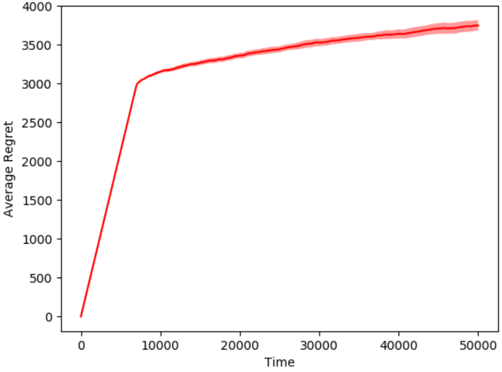
River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$
- $\epsilon_t = \dfrac{\epsilon_0}{(N(s_t) - 1000)^{2/3}}$
- $\epsilon_t = \begin{cases} 1.0 & t < 6000 \\ \dfrac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$



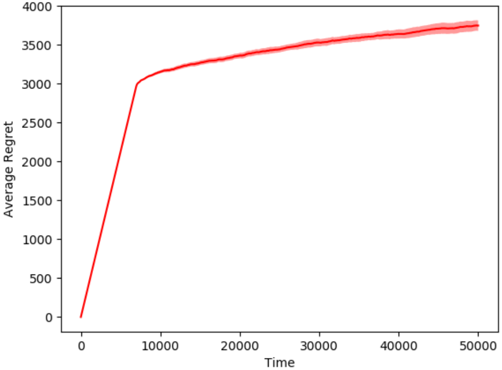
River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$
- $\epsilon_t = \dfrac{\epsilon_0}{(N(s_t) - 1000)^{2/3}}$
- $\epsilon_t = \begin{cases} 1.0 & t < 6000 \\ \dfrac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$
- $\epsilon_t = \begin{cases} 1.0 & t < 7000 \\ \dfrac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$



River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$
- $\epsilon_t = \frac{\epsilon_0}{(N(s_t) - 1000)^{2/3}}$
- $\epsilon_t = \begin{cases} 1.0 & t < 6000 \\ \frac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$
- $\epsilon_t = \begin{cases} 1.0 & t < 7000 \\ \frac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$



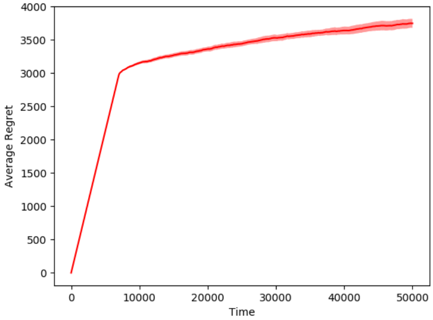
Tuning the ϵ schedule is difficult and problem dependent

River Swim: Q-learning w\ ϵ-greedy Exploration

Main drawbacks of Q-learning with ϵ-greedy

- ϵ-greedy performs *undirected* exploration
- *Inefficient use* of samples

Regret: $\Omega\Big(\min\{T, A^{H/2}\}\Big)$

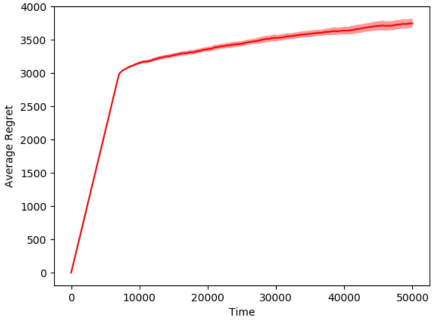


River Swim: Q-learning w\ ϵ-greedy Exploration

Main drawbacks of Q-learning with ϵ-greedy

- ϵ-greedy performs *undirected* exploration
- *Inefficient use* of samples

Regret: $\Omega\Big(\min\{T, A^{H/2}\}\Big)$



Uncertainty-driven exploration-exploitation

Minimax Lower Bound

Theorem (adapted from Jaksch et al. [2010])

For any MDP $M^\star = \langle \mathcal{S}, \mathcal{A}, p_h, r_h, H \rangle$ with stationary $(p_1 = p_2 = \ldots = p_H)$ transitions, any algorithm $\mathfrak{A}$ at any episode K suffers a regret of at least

$$\Omega\left(\sqrt{HSAT}\right)$$

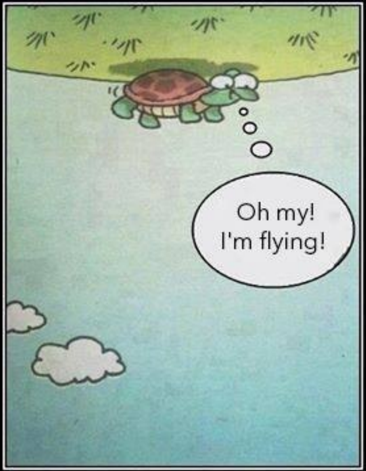
with $T = HK$.

- If *non-stationary* transitions
  - $p_1, \ldots, p_H$ can be arbitrary different
  - Effective number of states is $S' = HS$
  - Lower bound

$$\Omega\left(H\sqrt{SAT}\right)$$

Tabular MDPs: Outline

The Optimism Principle: Intuition



OPTIMISM
It's the best way to see life.

The Optimism Principle: Intuition

Exploration vs. Exploitation

The Optimism Principle: Intuition

Exploration vs. Exploitation

Optimism in Face of Uncertainty

When you are uncertain, consider the **best possible world (reward-wise)**

The Optimism Principle: Intuition

Exploration vs. Exploitation

Optimism in Face of Uncertainty

When you are uncertain, consider the **best possible world (reward-wise)**

If the best possible world is **correct**

$\Longrightarrow$ **no regret**

Exploitation

If the best possible world is **wrong**

$\Longrightarrow$ **learn useful information**

Exploration

The Optimism Principle: Intuition

Exploration vs. Exploitation

Optimism in **value function**

Optimism in Face of Uncertainty

When you are uncertain, consider the **best possible world (reward-wise)**

If the best possible world is **correct**

$\Longrightarrow$ **no regret**

Exploitation

If the best possible world is **wrong**

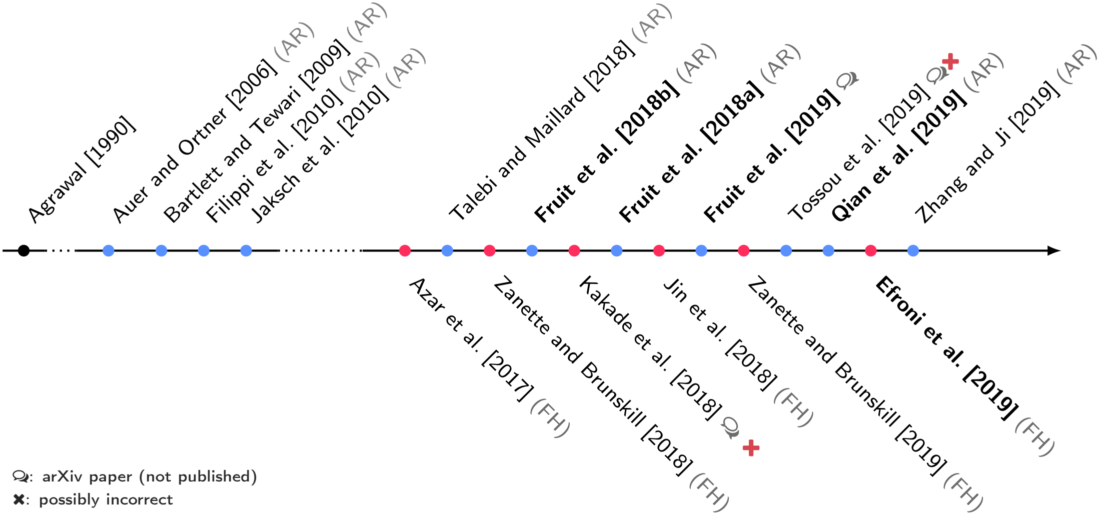
$\Longrightarrow$ **learn useful information**

Exploration

History: *OFU* for Regret Minimization

Tabular MDPs

FH: finite-horizon
AR: average reward



: arXiv paper (not published)
✖: possibly incorrect

Learning Problem

Input: $\mathcal{S}$, $\mathcal{A}$ $\cancel{r_h, p_h}$

Initialize $Q_{h1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h = 1, \dots, H$, $\mathcal{D}_1 = \emptyset$

for $k = 1, \dots, K$ **do** // `episodes`

Observe initial state s_{1k} *(arbitrary)*

Compute $(Q_{h,k})_{h=1}^{H}$ from $\mathcal{D}_k$

Define π_k based on $(Q_{hk})_{h=1}^{H}$

for $h = 1, \dots, H$ **do**

Execute $a_{hk} = \pi_{hk}(s_{hk})$

Observe r_{hk} and $s_{h+1,k}$

end

Add trajectory $(s_{hk}, a_{hk}, r_{hk})_{h=1}^{H}$ to $\mathcal{D}_{k+1}$

end

Learning Problem

Input: $\mathcal{S}$, $\mathcal{A}$ ~~r_h, p_h~~
Initialize $Q_{h1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h = 1, \dots, H$, $\mathcal{D}_1 = \emptyset$

for $k = 1, \dots, K$ **do** `// episodes`
    Observe initial state s_{1k} *(arbitrary)*
    Compute $(Q_{h,k})_{h=1}^{H}$ from $\mathcal{D}_k$
    Define π_k based on $(Q_{hk})_{h=1}^{H}$
    for $h = 1, \dots, H$ **do**
        Execute $a_{hk} = \pi_{hk}(s_{hk})$
        Observe r_{hk} and $s_{h+1,k}$
    end
    Add trajectory $(s_{hk}, a_{hk}, r_{hk})_{h=1}^{H}$ to $\mathcal{D}_{k+1}$
end

Defines the type of algorithm

Model-based Learning

Input: $\mathcal{S}$, $\mathcal{A}$ $\cancel{r_h, p_h}$

Initialize $Q_{h1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h = 1, \dots, H$, $\mathcal{D}_1 = \emptyset$

for $k = 1, \dots, K$ **do** `// episodes`

Observe initial state s_{1k} *(arbitrary)*

Estimate empirical MDP $\widehat{M}_k = (\mathcal{S}, \mathcal{A}, \widehat{p}_{hk}, \widehat{r}_{hk}, H)$ from $\mathcal{D}_k$

$$\widehat{p}_{hk}(s'|s,a) = \frac{\sum_{i=1}^{k-1} \mathbb{1}\left((s_{hi}, a_{hi}, s_{h+1,i}) = (s,a,s')\right)}{N_{hk}(s,a)}, \quad \widehat{r}_{hk}(s,a) = \frac{\sum_{i=1}^{k-1} r_{hi} \cdot \mathbb{1}\left((s_{hi}, a_{hi}) = (s,a)\right)}{N_{hk}(s,a)}$$

Planning (by backward induction) for π_{hk}

for $h = 1, \dots, H$ **do**

Execute $a_{hk} = \pi_{hk}(s_{hk})$

Observe r_{hk} and $s_{h+1,k}$

end

Add trajectory $(s_{hk}, a_{hk}, r_{hk})_{h=1}^{H}$ to $\mathcal{D}_{k+1}$

end

Measuring Uncertainty

Bounded parameter MDP [Strehl and Littman, 2008]

$$\mathcal{M}_k = \Big\{ \langle \mathcal{S}, \mathcal{A}, r_h, p_h, H \rangle : \forall h \in [H]$$

$$r_h(s,a) \in B^r_{hk}(s,a),\; p_h(\cdot|s,a) \in B^p_{hk}(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A} \Big\}$$

Compact *confidence sets*

$$B^r_{hk}(s,a) := \left[\widehat{r}_{hk}(s,a) - \beta^r_{hk}(s,a),\; \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) \right]$$

$$B^p_{hk}(s,a) := \left\{ p(\cdot|s,a) \in \Delta(\mathcal{S}) :\; \|p(\cdot|s,a) - \widehat{p}_{hk}(\cdot|s,a)\|_1 \leq \; \beta^p_{hk}(s,a) \right\}$$

Measuring Uncertainty

Bounded parameter MDP [Strehl and Littman, 2008]

$$\mathcal{M}_k = \Big\{ \langle \mathcal{S}, \mathcal{A}, r_h, p_h, H \rangle : \forall h \in [H]$$
$$r_h(s,a) \in B^r_{hk}(s,a),\ p_h(\cdot|s,a) \in B^p_{hk}(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A} \Big\}$$

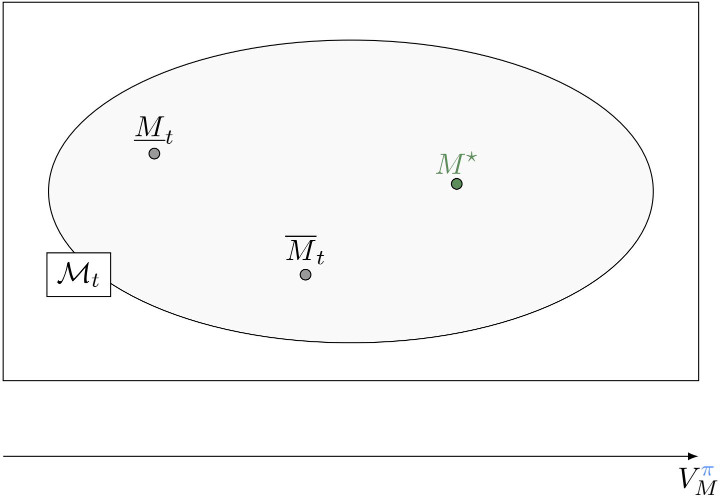
Compact *confidence sets*

$$B^r_{hk}(s,a) := \Big[\widehat{r}_{hk}(s,a) - \beta^r_{hk}(s,a),\ \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) \Big]$$
$$B^p_{hk}(s,a) := \Big\{ p(\cdot|s,a) \in \Delta(\mathcal{S}) :\ \|p(\cdot|s,a) - \widehat{p}_{hk}(\cdot|s,a)\|_1 \leq\ \beta^p_{hk}(s,a) \Big\}$$

Confidence bounds based on [Hoeffding, 1963] and [Weissman et al., 2003]

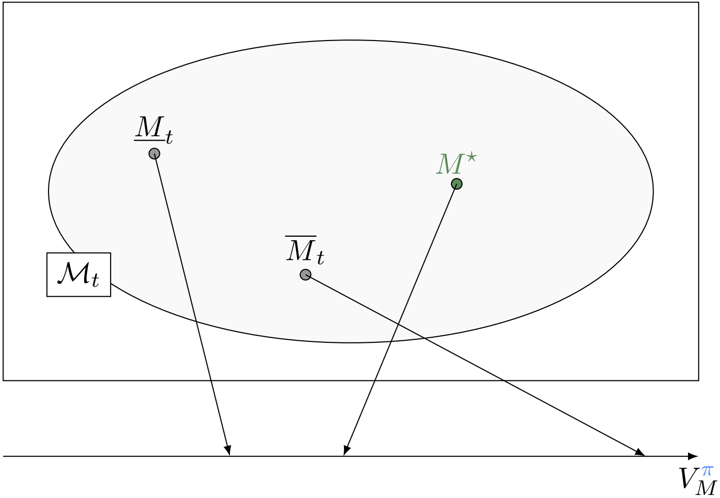
$$\beta^r_{hk}(s,a) \propto \sqrt{\frac{\log(N_{hk}(s,a)/\delta)}{N_{hk}(s,a)}}, \qquad \beta^p_{hk}(s,a) \propto \sqrt{\frac{S \log(N_{hk}(s,a)/\delta)}{N_{hk}(s,a)}}$$

Bounded Parameter MDP: Optimism



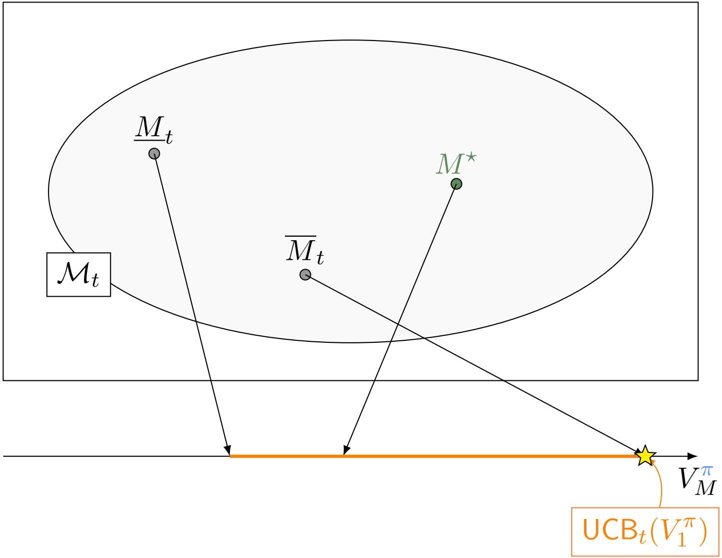
Fix a *policy* π

Bounded Parameter MDP: Optimism



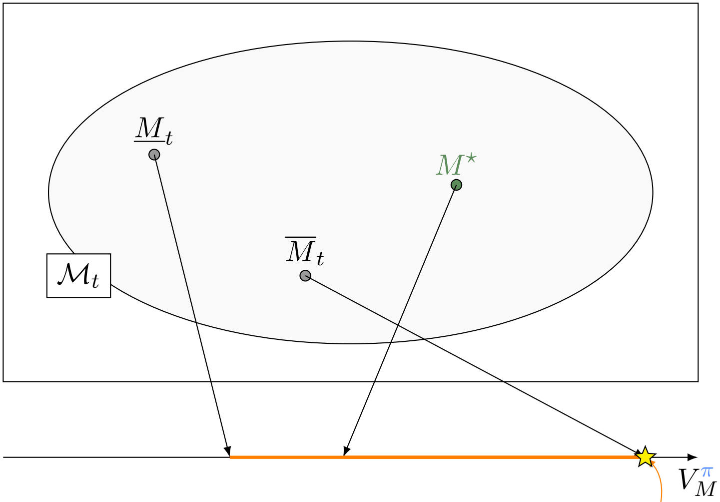
Fix a *policy* π

Bounded Parameter MDP: Optimism



Fix a *policy* π

Bounded Parameter MDP: Optimism



Fix a *policy* π

Optimism: $\mathsf{UCB}_t(V_1^\pi) = \max_{M \in \mathcal{M}_t} V_{1,M}^\pi \geq V_{1,M^\star}^\pi$ $\boxed{\mathsf{UCB}_t(V_1^\pi)}$

Extended Value Iteration

[Jaksch et al., 2010]

Input: $\mathcal{S}$, $\mathcal{A}$, B^r_{hk}, B^p_{hk}

Set $Q_{H+1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

for $h = H, \dots, 1$ **do**

  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

    Compute

$$Q_{hk}(s,a) = \max_{r_h \in B^r_{hk}(s,a)} r_h(s,a) + \max_{p_h \in B^p_{hk}(s,a)} \mathbb{E}_{s' \sim p_h(\cdot|s,a)} \Big[V_{h+1,k}(s') \Big]$$

$$= \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) + \max_{p_h \in B^p_{hk}(s,a)} \mathbb{E}_{s' \sim p_h(\cdot|s,a)} \Big[V_{h+1,k}(s') \Big]$$

$$V_{hk}(s) = \min\Big\{ H - (h-1), \max_{a \in \mathcal{A}} Q_{hk}(s,a) \Big\}$$

  end

end

return $\pi_{hk}(s) = \arg\max_{a \in \mathcal{A}} Q_{hk}(s,a)$

Extended Value Iteration

[Jaksch et al., 2010]

Input: $\mathcal{S}$, $\mathcal{A}$, B^r_{hk}, B^p_{hk}

Set $Q_{H+1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

for $h = H, \ldots, 1$ **do**

  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

    Compute

$$Q_{hk}(s,a) = \max_{r_h \in B^r_{hk}(s,a)} r_h(s,a) + \max_{p_h \in B^p_{hk}(s,a)} \mathbb{E}_{s' \sim p_h(\cdot|s,a)} \Big[V_{h+1,k}(s') \Big]$$

$$= \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) + \max_{p_h \in B^p_{hk}(s,a)} \mathbb{E}_{s' \sim p_h(\cdot|s,a)} \Big[V_{h+1,k}(s') \Big]$$

$$V_{hk}(s) = \min \Big\{ H - (h-1),\ \max_{a \in \mathcal{A}} Q_{hk}(s,a) \Big\}$$

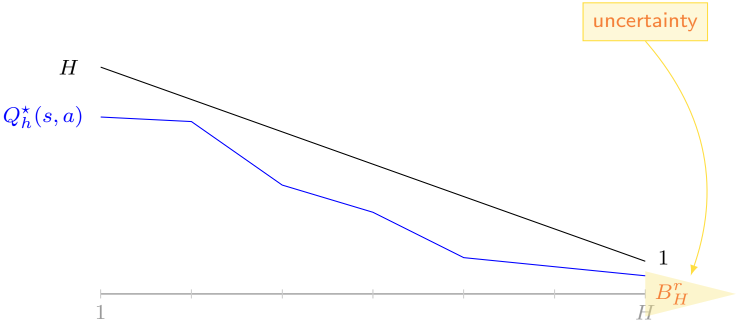
  end

end

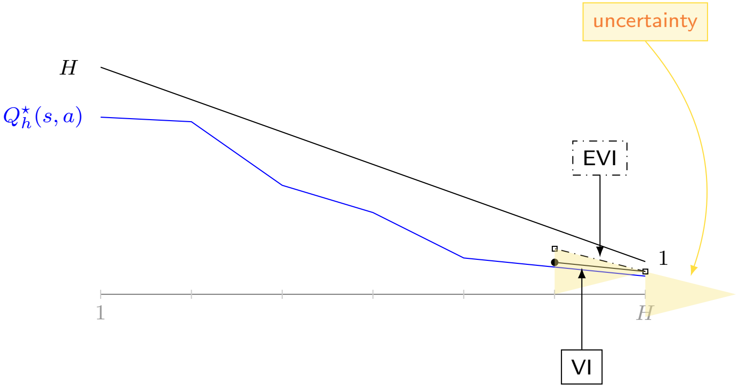
return $\pi_{hk}(s) = \arg\max_{a \in \mathcal{A}} Q_{hk}(s,a)$

Policy **executed** at episode k

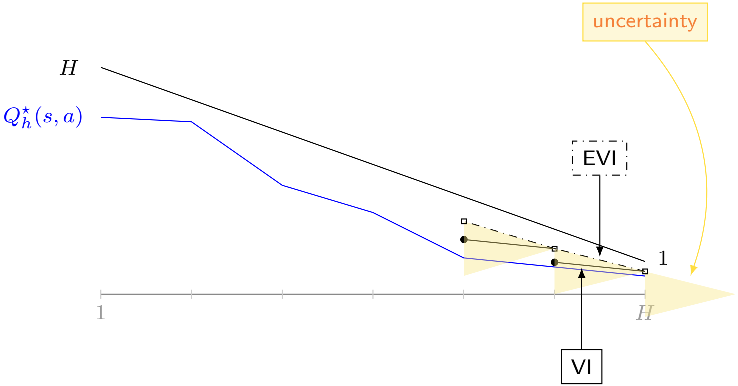
Optimism



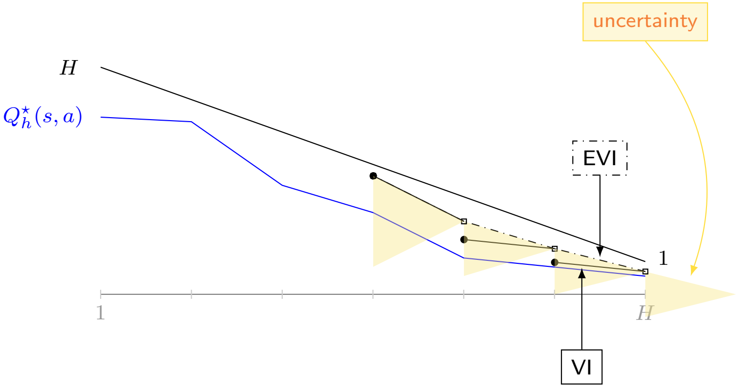
Optimism



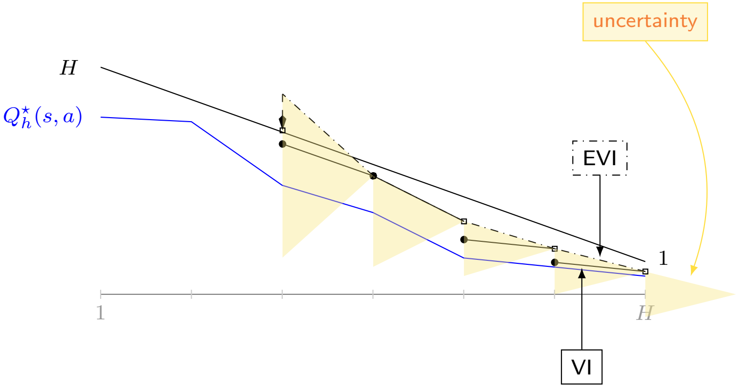
Optimism



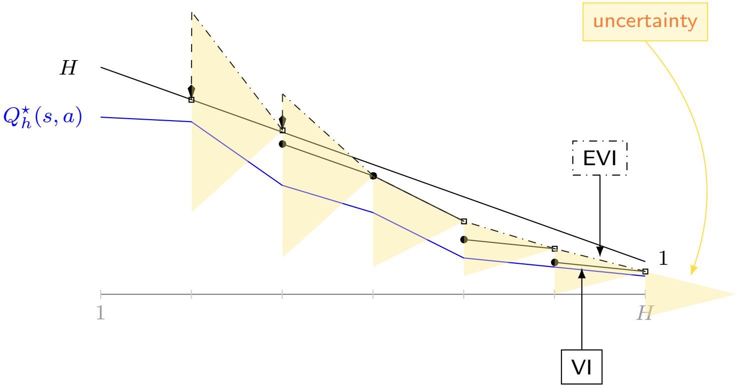
Optimism



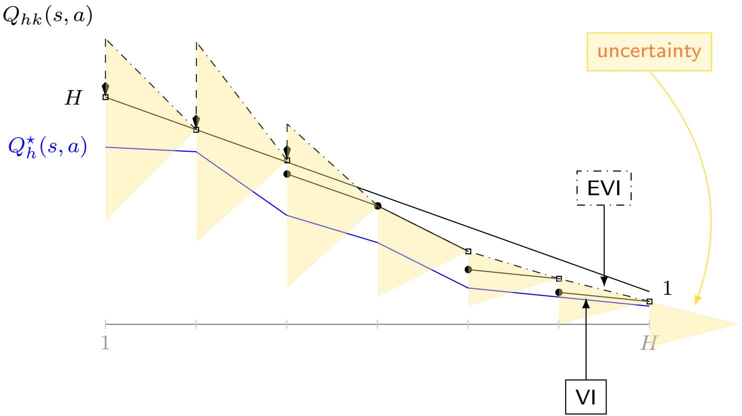
Optimism



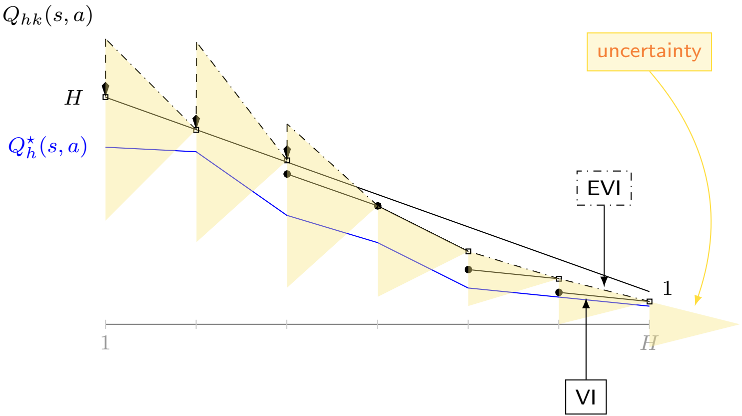
Optimism



Optimism



Optimism



$$\forall h \in [H], \forall (s,a), \qquad Q_{hk}(s,a) \geq Q^\star_h(s,a)$$

UCRL2-CH for Finite Horizon

Theorem (adapted from [Jaksch et al., 2010])

For any tabular MDP with stationary transitions, UCRL2 *with Chernoff-Hoeffding confidence intervals (UCRL2-CH), with high-probability, suffers a regret*

$$R(K, M^\star, \text{UCRL2-CH}) = \widetilde{\mathcal{O}}\left(HS\sqrt{AT} + H^2SA\right)$$

- Order optimal $\sqrt{AT}$
- $\sqrt{HS}$ factor worse than the lower-bound

Lower-bound: $\Omega(\sqrt{HSAT})$ ***(stationary transitions)***

Extended Value Iteration

$$Q_{hk}(s,a) = \max_{(r,p)\in B^r_{hk}(s,a)\times B^p_{hk}(s,a)} \left\{ r + p^\mathsf{T} V_{h+1,k} \right\}$$

$$= \max_{r\in B^r_{hk}(s,a)} r + \max_{p\in B^p_{hk}(s,a)} p^\mathsf{T} V_{h+1,k}$$

$$= \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) + \max_{p\in B^p_{hk}(s,a)} p^\mathsf{T} V_{h+1,k}$$

$$\leq \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) + \|p - \widehat{p}_{hk}(\cdot|s,a)\|_1 \|V_{h+1,k}\|_\infty + \widehat{p}_{hk}(\cdot|s,a)^\mathsf{T} V_{h+1,k}$$

$$\leq \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a) + \widehat{p}_{hk}(\cdot|s,a)^\mathsf{T} V_{h+1,k}$$

Extended Value Iteration

$$Q_{hk}(s,a) = \max_{(r,p)\in B^r_{hk}(s,a)\times B^p_{hk}(s,a)} \left\{ r + p^\mathsf{T} V_{h+1,k} \right\}$$

$$= \max_{r\in B^r_{hk}(s,a)} r + \max_{p\in B^p_{hk}(s,a)} p^\mathsf{T} V_{h+1,k}$$

$$= \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) + \max_{p\in B^p_{hk}(s,a)} p^\mathsf{T} V_{h+1,k}$$

$$\leq \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) + \|p - \widehat{p}_{hk}(\cdot|s,a)\|_1 \|V_{h+1,k}\|_\infty + \widehat{p}_{hk}(\cdot|s,a)^\mathsf{T} V_{h+1,k}$$

$$\leq \widehat{r}_{hk}(s,a) + \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a) + \widehat{p}_{hk}(\cdot|s,a)^\mathsf{T} V_{h+1,k}$$

Exploration bonus $(1 + H\sqrt{S})\beta^r_{hk}(s,a)$ for the reward

UCBVI

[Azar et al., 2017]

Replace EVI with *Exploration Bonus*

Input: $\mathcal{S}$, $\mathcal{A}$, $\cancel{B^r_{hk}, B^p_{hk}}$, $\widehat{r}_{hk}, \widehat{p}_{hk}, b_{hk}$

Set $Q_{H+1,k}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

for $h = H, \dots, 1$ **do**

  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

    Compute

$$Q_{hk}(s,a) = \widehat{r}_{hk}(s,a) + b_{hk}(s,a) + \mathbb{E}_{s' \sim \widehat{p}_{hk}(\cdot|s,a)}\Big[V_{h+1,k}(s')\Big]$$

$$V_{hk}(s) = \min\Big\{H - (h-1), \max_{a' \in \mathcal{A}} Q_{hk}(s',a')\Big\}$$

  end

end

return $\pi_{hk}(s) = \arg\max_{a \in \mathcal{A}} Q_{hk}(s,a)$

Equivalent to value iteration on $\overline{M}_k = (\mathcal{S}, \mathcal{A}, \widehat{r}_{hk} + b_{hk}, \widehat{p}_{hk}, H)$

UCBVI: Measuring Uncertainty

- *Combine uncertainties* in rewards and transitions
- In a smart way

$$b_{hk}(s,a) = (H+1)\sqrt{\frac{\log(N_{hk}(s,a)/\delta)}{N_{hk}(s,a)}} < \beta^r_{hk} + H\beta^p_{hk}$$

UCBVI: Measuring Uncertainty

- *Combine uncertainties* in rewards and transitions
- In a smart way

$$b_{hk}(s,a) = (H+1)\sqrt{\frac{\log(N_{hk}(s,a)/\delta)}{N_{hk}(s,a)}} < \beta^r_{hk} + H\beta^p_{hk}$$

Save a $\sqrt{S}$ factor

$$\left| (p_h(\cdot|s,a) - \widehat{p}_{hk}(\cdot|s,a))^\mathsf{T} \underbrace{V^\star_h}_{\leq H} \right| \leq H \underbrace{\sqrt{\frac{\log(N_{hk}(s,a)/\delta)}{N_{hk}(s,a)}}}_{=\beta^p_{hk}/\sqrt{S}}$$

UCBVI-CH: Regret

Theorem (Thm. 1 of Azar et al. [2017])

For any tabular MDP with stationary transitions, UCBVI *with Chernoff-Hoeffding confidence intervals (UCBVI-CH), with high-probability, suffers a regret*

$$R(K, M^\star, \text{UCBVI-CH}) = \widetilde{\mathcal{O}}\left(H\sqrt{SAT} + H^2 S^2 A\right)$$

- Order optimal $\sqrt{SAT}$
- $\sqrt{H}$ factor worse than the lower-bound
- Long "warm up" phase
- If *non-stationary*, then $\widetilde{\mathcal{O}}\left(H^{3/2}\sqrt{SAT}\right)$

Lower-bound: $\Omega(\sqrt{HSAT})$ ***(stationary transitions)***

Refined Confidence Bounds

- UCRL2 with *Bernstein-Freedman bounds* (instead of Hoeffding/Weissman): *
  - see tutorial website

$$R(K, M^\star, \mathsf{UCRL2B}) = \widetilde{\mathcal{O}}\left(\sqrt{H\ \Gamma\ SAT} + H^2 S^2 A\right)$$

$$\Gamma = \max_{h,s,a} \|p_h(\cdot|s,a)\|_0 \leq S$$

Still not matching the lower-bound!

* **stationary model** $(p_1 = \ldots = p_H)$

Refined Confidence Bounds

- UCRL2 with *Bernstein-Freedman bounds* (instead of Hoeffding/Weissman): *
  - see tutorial website

$$R(K, M^\star, \text{UCRL2B}) = \widetilde{\mathcal{O}}\left(\sqrt{H\,\Gamma\,SAT} + H^2S^2A\right)$$

$$\Gamma = \max_{h,s,a} \|p_h(\cdot|s,a)\|_0 \leq S$$

  - Still not matching the lower-bound!

- UCBVI with *Bernstein-Freedman bounds*: *

$$R(K, M^\star, \text{UCBVI-BF}) = \widetilde{\mathcal{O}}\left(\sqrt{HSAT} + H^2S^2A + H\sqrt{T}\right)$$

  - Matching the Lower-Bound!
  - Long "warm up" phase

* **stationary model** $(p_1 = \ldots = p_H)$

Refined Confidence Bounds

- EULER [Zanette and Brunskill, 2019]
  keeps upper and lower bounds on $V_h^\star$

$$R(K, M^\star, \text{EULER}) = \mathcal{O}\left(\sqrt{\mathbb{Q}^\star SAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})\right)$$

Problem-dependent bound based on *environmental norm* [Maillard et al., 2014]

$$\mathbb{Q}^\star = \max_{s,a,h}\left(\mathbb{V}(r_h(s,a)) + \mathbb{V}_{x\sim p_h(\cdot|s,a)}(V_{h+1}^\star(x))\right)$$

$$\mathbb{V}_{x\sim p}(f(x)) = \mathbb{E}_{x\sim p}\left[\left(f(x) - \mathbb{E}_{y\sim p}[f(y)]\right)^2\right]$$

Can remove the dependence on H

Matching lower-bound in the worst case

UCRL2: RiverSwim

Hoeffding

$$b^r_{hk}(s,a) = r_{\max}\sqrt{\frac{L}{N}}$$

$$b^p_{hk}(s,a) = \sqrt{\frac{SL}{N}}$$

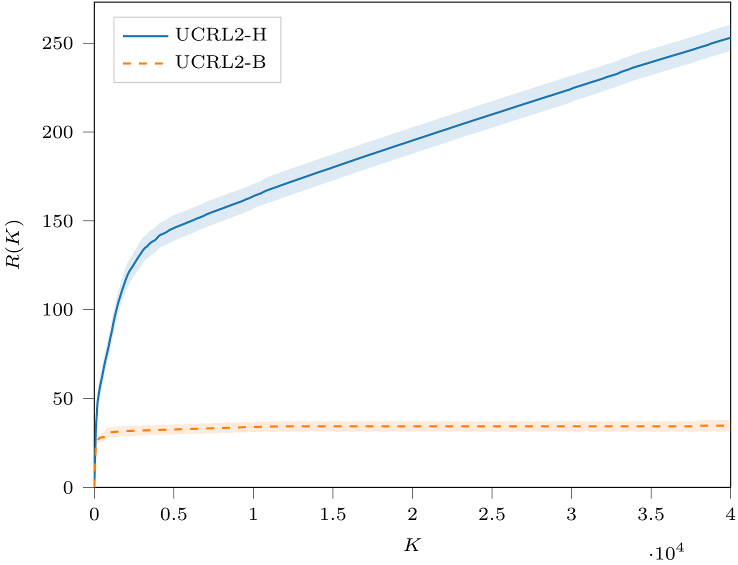
Bernstein

$$b^r_{hk}(s,a) = \sqrt{\frac{L\widehat{\mathbb{V}}(\widehat{r}_{hk})}{N}} + r_{\max}\frac{L}{N}$$

$$b^p_{hk}(s,a) = \sqrt{\frac{L\widehat{\mathbb{V}}(\widehat{p}_{hk})}{N}} + \frac{L}{N}$$

$$\widehat{\mathbb{V}}(\widehat{r}_{hk}) = \frac{1}{N}\sum_i (r_{h,i} - \widehat{r}_{hk})^2$$

is the population variance



$N = N_{hk}(s,a) \vee 1$
$L = \log(SAN/\delta)$

UCBVI: RiverSwim

Hoeffding

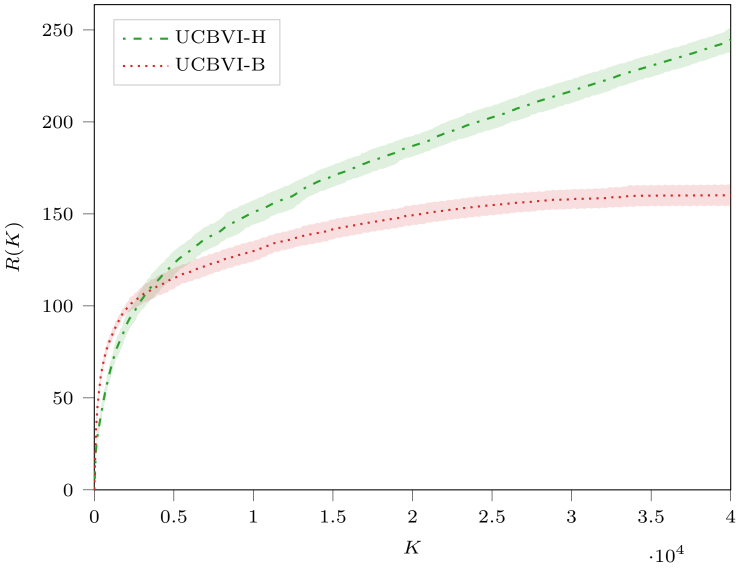
$$b_{hk}(s,a) = \frac{(H-h)L}{\sqrt{N}}$$

Bernstein

$$b_{hk}(s,a) = \sqrt{\frac{L\mathbb{V}_{\widehat{p}_{hk}}(V_{h+1,k})}{N}} + \frac{(H-h)L}{N} + \frac{(H-h)}{\sqrt{N}}$$

$\mathbb{V}_p(V) = \mathbb{E}_{x\sim p}[(V(x)-\mu)^2]$
with $\mu = \mathbb{E}_{x\sim p}[V(x)]$

$N = N_{hk}(s,a) \vee 1$
$L = \log(SAN/\delta)$



UCBVI: RiverSwim

Hoeffding

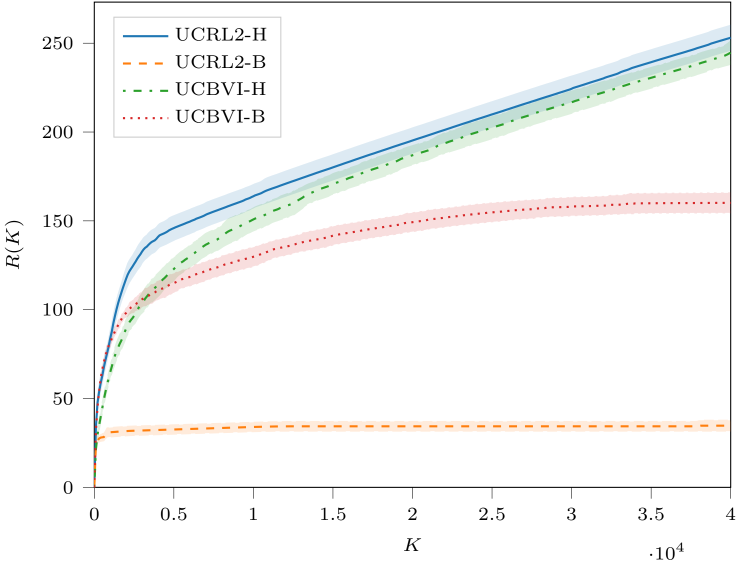
$$b_{hk}(s,a) = \frac{(H-h)L}{\sqrt{N}}$$

Bernstein

$$b_{hk}(s,a) = \sqrt{\frac{L\mathbb{V}_{\widehat{p}_{hk}}(V_{h+1,k})}{N}} + \frac{(H-h)L}{N} + \frac{(H-h)}{\sqrt{N}}$$

$\mathbb{V}_p(V) = \mathbb{E}_{x\sim p}[(V(x)-\mu)^2]$
with $\mu = \mathbb{E}_{x\sim p}[V(x)]$

$N = N_{hk}(s,a) \vee 1$
$L = \log(SAN/\delta)$



Model-Based *Advantages*

Learning efficiency

- First order optimal
- Matching lower-bound

Counterfactual reasoning

- Optimistic/Pessimistic value estimate for any π
- Usefull for inference (e.g., safety)

Model-Based *Issues*

Complexity

- Space $O(HS^2A)$

$$\textit{non-stationary model} \implies H(\underbrace{S^2A}_{\textit{transitions}} + \underbrace{SA}_{\textit{rewards}})$$

- Time $O(K \underbrace{HS^2A}_{\textit{planning by VI}})$

Model-Based *Issues*

Complexity

- Space $O(HS^2A)$

$$\text{non-stationary model} \implies H(\underbrace{S^2A}_{\text{transitions}} + \underbrace{SA}_{\text{rewards}})$$

- Time $O(K \underbrace{HS^2A}_{\text{planning by VI}})$

incremental updates

Tabular MDPs: Outline

Posterior Sampling (PS)

a.k.a. Thompson Sampling [Thompson, 1933]

Keep Bayesian posterior for the *unknown* MDP

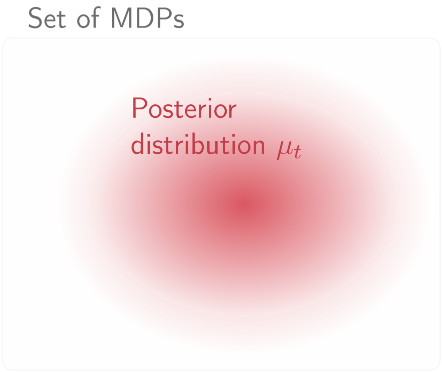
👉 A sample from the posterior is used as an estimate of the unknown MDP

Exploration

Few samples $\Longrightarrow$ uncertainty in the estimate

More samples $\Longrightarrow$ posterior concentrates on the true MDP

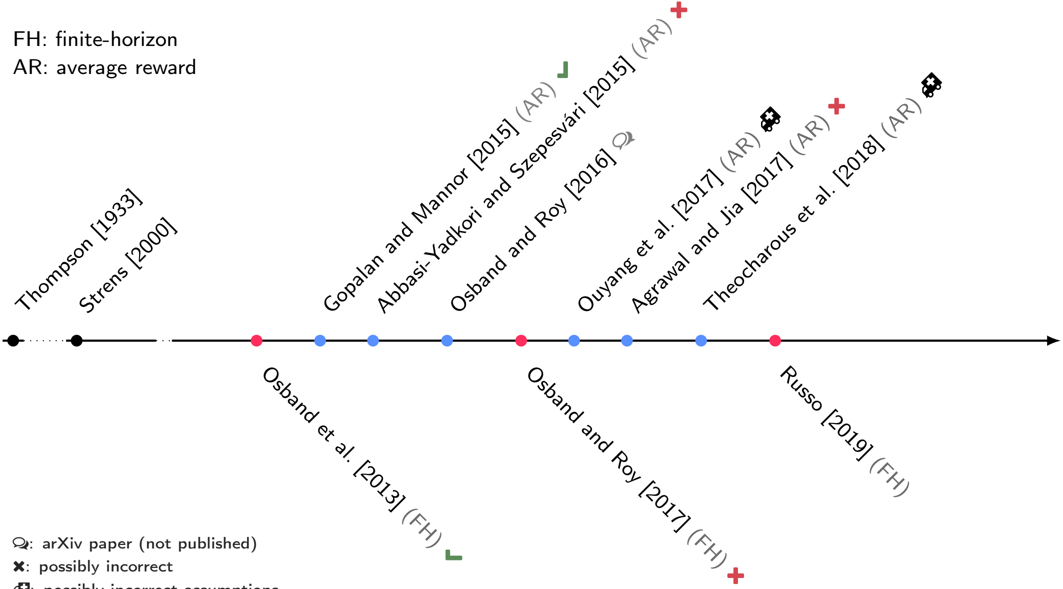
Exploitation



History: PS for Regret Minimization

Tabular MDPs

FH: finite-horizon
AR: average reward



: arXiv paper (not published)
✖: possibly incorrect
: possibly incorrect assumptions

Bayesian Regret

$$R^B(K, \mu_1, \mathfrak{A}) = \mathbb{E}_{M^\star \sim \mu_1}\left[\underbrace{\overline{R}(K, M^\star, \mathfrak{A})}_{:=\mathbb{E}\left[R(K, M^\star, \mathfrak{A})\right]} \right] = \mathbb{E}_{M^\star}\left[\sum_{k=1}^{K} V^\star_{1,M^\star}(s_{1k}) - V^{\pi_k}_{1,M^\star}(s_{1k})\right]$$

Posterior Sampling

[Osband and Roy, 2017]

Input: $\mathcal{S}$, $\mathcal{A}$, $\cancel{r_h, p_h}$, *prior* μ_1
Initialize $\mathcal{D}_1 = \emptyset$

for $k = 1, \ldots, K$ **do** `// episodes`

Observe initial state s_{1k} *(arbitrary)*

Sample $M_k \sim \mu_k(\cdot|\mathcal{D}_k)$
Compute

$$\pi_k \in \arg\max_{\pi}\{V^{\pi}_{1,M_k}\}$$

for $h = 1, \ldots, H$ **do**

Execute $a_{hk} = \pi_{hk}(s_{hk})$
Observe r_{hk} and $s_{h+1,k}$

end

Add trajectory $(s_{hk}, a_{hk}, r_{hk})_{h=1}^{H}$ to $\mathcal{D}_{k+1}$

end

Posterior Sampling

[Osband and Roy, 2017]

Input: $\mathcal{S}$, $\mathcal{A}$, $\cancel{r_h, p_h}$, *prior* μ_1
Initialize $\mathcal{D}_1 = \emptyset$

for $k = 1, \dots, K$ **do** `// episodes`
  Observe initial state s_{1k} *(arbitrary)*

  Sample $M_k \sim \mu_k(\cdot|\mathcal{D}_k)$
  Compute

$$\pi_k \in \arg\max_{\pi}\{V^{\pi}_{1,M_k}\}$$

  for $h = 1, \dots, H$ **do**
    Execute $a_{hk} = \pi_{hk}(s_{hk})$
    Observe r_{hk} and $s_{h+1,k}$
  end
  Add trajectory $(s_{hk}, a_{hk}, r_{hk})_{h=1}^{H}$ to $\mathcal{D}_{k+1}$
end

Prior distribution:

$$\forall \Theta, \quad \mathbb{P}(M^* \in \Theta) = \mu_1(\Theta)$$

Posterior distribution:

$$\forall \Theta, \quad \mathbb{P}(M^* \in \Theta | \mathcal{D}_k, \mu_1) = \mu_k(\Theta)$$

Priors

- Dirichlet (transitions)
- Beta, Normal-Gamma, etc. (rewards)

Model Update with Dirichlet Priors

⚠ *assume r is known*

$$\underbrace{\left\{\mu_t,\ (s_t, a_t, s_{t+1})\right\}}_{\sim H_t} \mapsto \mu_{t+1}$$

Model Update with Dirichlet Priors

⚠ *assume r is known*

$$\underbrace{\{\mu_t,\ (s_t, a_t, s_{t+1})\}}_{\sim H_t} \mapsto \mu_{t+1}$$

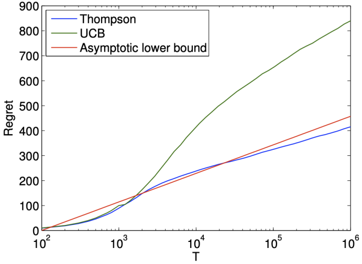
- $\mu_t(s,a) = \text{Dirichlet}(\alpha_1, \dots, \alpha_S)$ on $p(\cdot|s,a)$
- Observe $s_{t+1} \sim p(\cdot|s_t, a_t)$ (outcome of a multivariate Bernoulli) such that $s_{t+1} = i$. The Bayesian posterior is

$$\mu_{t+1}(s,a) = \text{Dirichlet}(\ \alpha_1, \dots, \alpha_i + 1, \dots, \alpha_S\ )$$

- Posterior mean vector $\widehat{p}_{t+1}(s_i|s,a) = \frac{\alpha_i}{n}$
- Variance bounded by $\frac{1}{n}$

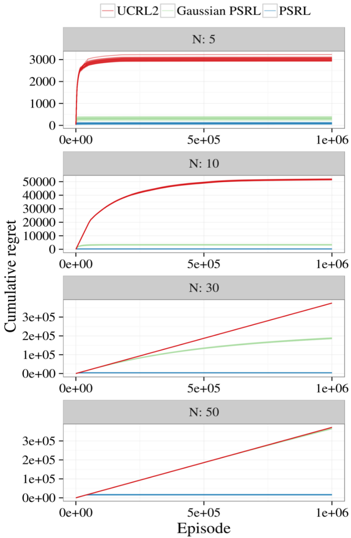
$$n = \sum_{i=1}^{S} \alpha_i$$

Posterior Sampling is Usually Better



Bandit

[Chapelle and Li, 2011]



Finite horizon RL

[Osband and Roy, 2017]

PSRL: Regret

Theorem (Osband and Roy [2017] revisited)

For any prior μ_1 with any independent Dirichlet prior over stationary transitions, the Bayesian regret of PSRL is bounded as

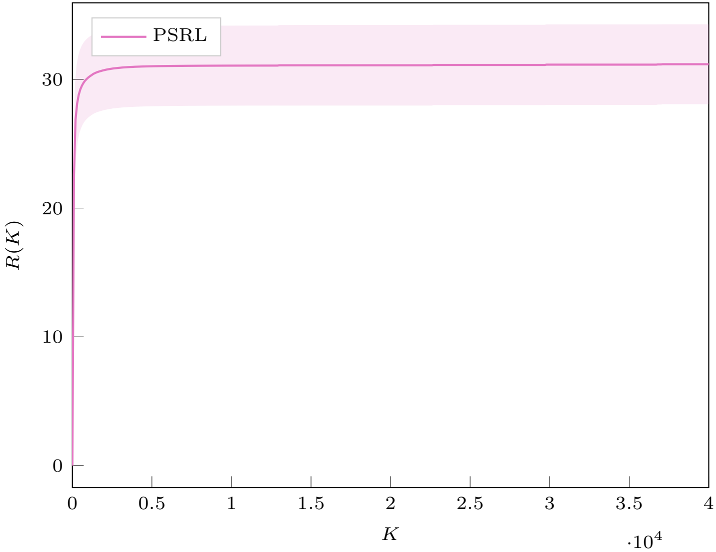
$$R^B(K, \mu_1, PSRL) = \widetilde{O}(HS\sqrt{AT})$$

- Order optimal $\sqrt{AT}$
- $\sqrt{HS}$ factor suboptimal

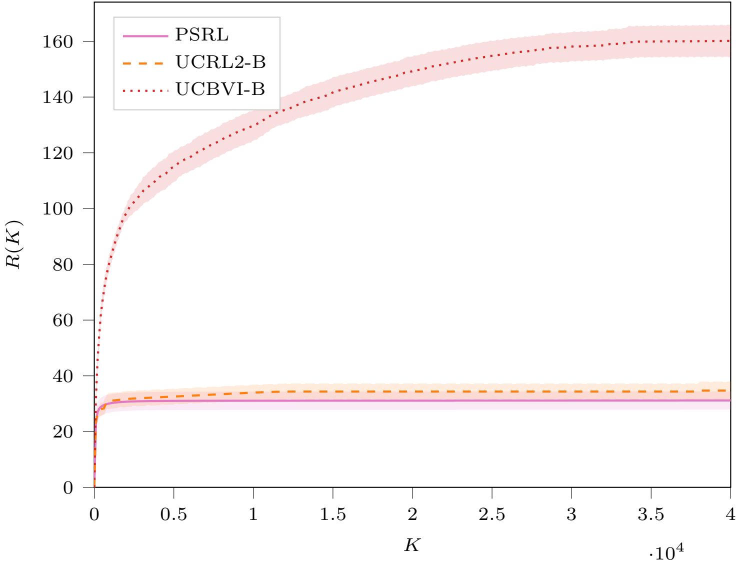
Lower-bound: $\Omega(\sqrt{HSAT})$ ***(stationary transitions)***

* in [Osband and Roy, 2017] is $\widetilde{O}(H\sqrt{SAT})$ for stationary MDPs
but there is a mistake in Lem. 3 (see [Qian et al., 2020])

PSRL: RiverSwim



PSRL: RiverSwim



Tabular MDPs: Outline

Model-Based *Issues*

Complexity

- Space $O(HS^2A)$

nonstationary model $\Longrightarrow H(\underbrace{S^2A}_{\textit{transitions}} + \underbrace{SA}_{\textit{rewards}})$

- Time $O(K \underbrace{HS^2A}_{\textit{planning by VI}})$

Solutions

1. Time complexity: incremental planning (e.g.,Opt-RTDP)

Model-Based *Issues*

Complexity

- Space $O(HS^2A)$

$$\textit{nonstationary model} \implies H(\underbrace{S^2A}_{\textit{transitions}} + \underbrace{SA}_{\textit{rewards}})$$

- Time $O(K \underbrace{HS^2A}_{\textit{planning by VI}})$

Solutions

1. Time complexity: incremental planning (e.g.,Opt-RTDP)
2. Space complexity: avoid to estimate rewards and transitions

Model-Based *Issues*

Complexity

- Space $O(HS^2A)$

$$\text{\textit{nonstationary model}} \implies H(\underbrace{S^2A}_{\text{transitions}} + \underbrace{SA}_{\text{rewards}})$$

- Time $O(K \underbrace{HS^2A}_{\text{planning by VI}})$

Solutions

1. Time complexity: incremental planning (e.g.,Opt-RTDP)
2. Space complexity: avoid to estimate rewards and transitions

Optimistic Q-learning (Opt-QL)

Space: $\mathcal{O}(HSA)$ Time: $\mathcal{O}(HAK)$

Optimistic Q-learning

Input: $\mathcal{S}$, $\mathcal{A}$, $\cancel{r_h, p_h}$

Initialize $Q_h(s,a) = H - (h-1)$ and $N_h(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h = [H]$

for $k = 1, \dots, K$ **do** `// episodes`

  Observe initial state s_{1k} *(arbitrary)*

  for $h = 1, \dots, H$ **do**

    Execute $a_{hk} = \pi_{hk}(s_{hk}) = \arg\max_a \widehat{Q}_h(s_{hk}, a)$

    Observe r_{hk} and $s_{h+1,k}$

    Set $N_h(s_{hk}, a_{hk}) = N_h(s_{hk}, a_{hk}) + 1$

    Update

$$Q_h(s_{hk}, a_{hk}) = (1-\alpha_t) Q_h(s_{hk}, a_{hk}) + \alpha_t \left(r_{hk} + \widehat{V}_{h+1}(s_{h+1,k}) + b_t \right)$$

Upper-Confidence Bound

    Set $\widehat{V}_h(s_{hk}) = \min \left\{ H - (h-1), \max_{a \in \mathcal{A}} Q_h(s_{hk}, a) \right\}$

  end

end

Step size and bonus for Opt-Qlearning

Let $t = N_{hk}(s,a)$

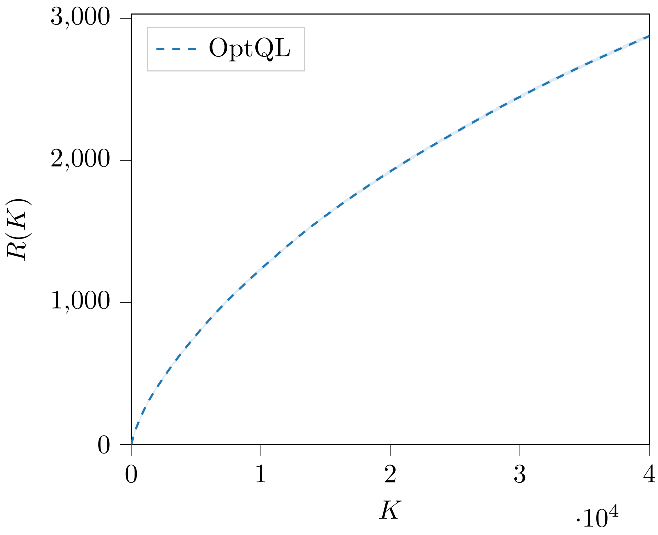
$$\alpha_t = \frac{H+1}{H+t}$$

Bonus

$$\left| \sum_{i=1}^{t} \alpha_t^i \Big(V_{h+1}^{\star}(s_{h+1,k_i}) - \mathbb{E}_{s'|s,a}[V_{h+1}^{\star}(s')] \Big) \right| \le \underbrace{c\sqrt{\frac{H^3 \log(SAT/\delta)}{t}}}_{:= b_t}$$

Opt-Qlearning: Example

Not so good!



Thank you!

facebook
Artificial Intelligence Research

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *UAI*, pages 1–11. AUAI Press, 2015.

Rajeev Agrawal. Adaptive control of markov chains under the weak accessibility. In *29th IEEE Conference on Decision and Control*, pages 1426–1431. IEEE, 1990.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *NIPS*, pages 1184–1194, 2017.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *NIPS*, pages 49–56. MIT Press, 2006.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 2017.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI*, pages 35–42. AUAI Press, 2009.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011. URL `http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf`.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *NeurIPS*, pages 12203–12213, 2019.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and kullback-leibler divergence. *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122, 2010.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. In *NeurIPS*, pages 2998–3008, 2018a.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *ICML*, Proceedings of Machine Learning Research. PMLR, 2018b.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved analysis of UCRL2B, 2019. URL `https://rlgammazero.github.io/docs/ucrl2b_improved.pdf`.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 861–898. JMLR.org, 2015.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. URL `http://www.jstor.org/stable/2282952`.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *NeurIPS*, pages 4868–4878, 2018.

Sham Kakade, Mengdi Wang, and Lin F. Yang. Variance reduction methods for sublinear reinforcement learning. *CoRR*, abs/1802.09184, 2018.

Odalric-Ambrym Maillard, Timothy A. Mann, and Shie Mannor. How hard is my mdp?" the distribution-norm to the rescue". In *NIPS*, pages 1835–1843, 2014.

Ian Osband and Benjamin Van Roy. Posterior sampling for reinforcement learning without episodes. *CoRR*, abs/1608.02731, 2016.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710. PMLR, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011, 2013.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *NIPS*, pages 1333–1342, 2017.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

Jian Qian, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Exploration bonus for regret minimization in discrete and continuous average reward mdps. In *NeurIPS*, pages 4891–4900, 2019.

Jian Qian, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Concentration inequalities for multinoulli random variables. *CoRR*, abs/2001.11595, 2020.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In *NeurIPS*, pages 14410–14420, 2019.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Malcolm Strens. A bayesian framework for reinforcement learning. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pages 943–950. ICML, 2000.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *ALT*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR, 2018.

Georgios Theocharous, Zheng Wen, Yasin Abbasi, and Nikos Vlassis. Scalar posterior sampling with applications. In *NeurIPS*, pages 7696–7704, 2018.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Aristide C. Y. Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *CoRR*, abs/1905.12425, 2019.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. 2003.

Andrea Zanette and Emma Brunskill. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 5732–5740. JMLR.org, 2018.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR, 2019.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *NeurIPS*, pages 2823–2832, 2019.