# Reinforcement Learning

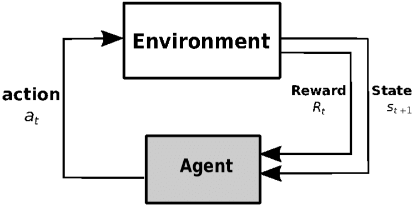
## Regret versus Sample Complexity

Emilie Kaufmann

M2 MVA, 2025/2026

# RL problems

A Markov Decision Process models the interaction of a learning agent with its environment :



But what do we care about ?

- Maximizing reward while learning, in a single interaction with the environment or a sequence of episodes (regret)
- Learning a good policy to apply later in the real world (sample complexity)

# 4 types of values

### Discounted MDP

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r_t \middle| s_1 = s\right]$$

### Episodic MDP

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{H} r_t \middle| s_1 = s\right]$$

### Average reward MDP

$$V^{\pi}(s) = \lim_{T\to\infty}\frac{1}{T}\mathbb{E}^{\pi}\left[\sum_{t=1}^{T} r_t \middle| s_1 = s\right]$$

### Goal-oriented MDP

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau_g^*} r_t \middle| s_1 = s\right]$$

$^*$ $g$ is an absorbing goal state, and $\tau_g$ is the time to reach the goal

# 4 types of values

## Discounted MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s\right]$$

## Episodic MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{H} r_t \middle| s_1 = s\right]$$

## Average reward MDP

$$V^\pi(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi\left[\sum_{t=1}^{T} r_t \middle| s_1 = s\right]$$

## Goal-oriented MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\tau_g^*} r_t \middle| s_1 = s\right]$$

$^*$ $g$ is an absorbing goal state, and $\tau_g$ is the time to reach the goal

**Question 1 :** How many interactions with the MDPs (or episodes) are needed to find a good policy ?

$$\mathbb{P}\left(V^\star - V^{\widehat{\pi}} \leq \varepsilon\right) \geq 1 - \delta$$

while minimizing the number of interactions with the environment

# 4 types of values

**Discounted MDP**

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s\right]$$

**Episodic MDP**

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{H} r_t \middle| s_1 = s\right]$$

**Average reward MDP**

$$V^{\pi}(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^{\pi}\left[\sum_{t=1}^{T} r_t \middle| s_1 = s\right]$$

**Goal-oriented MDP**

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau_g^*} r_t \middle| s_1 = s\right]$$

* $g$ is an absorbing goal state, and $\tau_g$ is the time to reach the goal

**Question 2 :** Can the algorithm learn to behave optimally ?
Can be measured by the number of sub-optimal plays or the regret

$$N = \sum_{t=1}^{\infty} \mathbb{1}\left(V^{\star}(s_t) - V^{\pi_t}(s_t) > \varepsilon\right)$$

(PAC-MDP framework, discounted MDP, see [Kakade, 2003])

# 4 types of values

### Discounted MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s\right]$$

### Episodic MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{H} r_t \middle| s_1 = s\right]$$

### Average reward MDP

$$V^\pi(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi\left[\sum_{t=1}^{T} r_t \middle| s_1 = s\right]$$

### Goal-oriented MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\tau_g^*} r_t \middle| s_1 = s\right]$$

[*] $g$ is an absorbing goal state, and $\tau_g$ is the time to reach the goal

**Question 2 :** Can the algorithm learn to behave optimally ?
Can be measured by the number of sub-optimal plays or the regret

$$R_T = TV^\star - \sum_{t=1}^{T} r_t \quad \text{in average-reward MDPs}$$

[Jaksch et al., 2010]

# 4 types of values

## Discounted MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s\right]$$

## Episodic MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{H} r_t \middle| s_1 = s\right]$$

## Average reward MDP

$$V^\pi(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi\left[\sum_{t=1}^{T} r_t \middle| s_1 = s\right]$$

## Goal-oriented MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\tau_g^*} r_t \middle| s_1 = s\right]$$

* $g$ is an absorbing goal state, and $\tau_g$ is the time to reach the goal

**Question 2 :** Can the algorithm learn to behave optimally ?
Can be measured by the number of sub-optimal plays or the regret

$$R_K = \sum_{k=1}^{K} \left(V^\star(s_1^k) - V^{\pi_k}(s_1^k)\right) \qquad \text{in episodic MDPs}$$

regret after $K$ episodes, $s_1^k$ : first state of episode $K$ [Azar et al., 2017a]

# 4 types of values

### Discounted MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s\right]$$

### Episodic MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{H} r_t \middle| s_1 = s\right]$$

### Average reward MDP

$$V^\pi(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi\left[\sum_{t=1}^{T} r_t \middle| s_1 = s\right]$$

### Goal-oriented MDP

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\tau_g^*} r_t \middle| s_1 = s\right]$$

* $g$ is an absorbing goal state, and $\tau_g$ is the time to reach the goal

**Question 2 :** Can the algorithm learn to behave optimally ?
Can be measured by the number of sub-optimal plays or the regret

$$R_K = \sum_{k=1}^{K} \left(V^\star(s_1^k) - V^{\pi_k}(s_1^k)\right) \qquad \text{in episodic MDPs}$$

regret after $K$ episodes, $s_1^k$ : first state of episode $K$ [Azar et al., 2017a]

# (Recap :) Solving an episodic MDP

$V^\pi(s) = V_1^\pi(s)$, introducing the value function from step $h$ :

$$V_h^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=h}^{H} r_t \middle| s_h = s\right]$$

(Non-stationary policy) $\pi = (\pi_1, \dots, \pi_H)$ :

$\pi_h(s)$ : action chosen if we are in state $s$ at step $h$ of the episode

## Bellman equations for a (deterministic) policy $\pi$

For all $h \in \{1, \dots, H\}$, for all $s \in \mathcal{S}$,

$$V_h^\pi(s) = r_h(s, \pi_h(s)) + \sum_{s' \in \mathcal{S}} p_h(s'|s, \pi_h(s)) V_{h+1}^\pi(s'),$$

with $V_{H+1}^\pi(s) = 0$ for all $s \in \mathcal{S}$.

➜ policy evaluation using backwards induction

# (Recap :) Solving an episodic MDP

$V^\star(s) = V_1^\star(s)$, introducing the optimal value function from step $h$ :

$$V_h^\star(s) = \max_\pi \mathbb{E}^\pi \left[ \sum_{t=h}^{H} r_t \middle| s_h = s \right]$$

## Bellman equations for the optimal policy

For all $h \in \{1, \dots, H\}$, for all $s \in \mathcal{S}$, letting $V_{H+1}^\star = 0$

$$\begin{aligned} V_h^\star(s) &= \max_{a\in\mathcal{A}} \left[ r_h(s,a) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a) V_{h+1}^\star(s') \right] \\ \pi_h^\star(s) &= \underset{a\in\mathcal{A}}{\text{argmax}} \left[ r_h(s,a) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a) V_{h+1}^\star(s') \right] \end{aligned}$$

➜ The optimal policy $\pi^\star$ is non-stationnary and can be computed using backwards induction (dynamic programming)

# (Recap :) Solving an episodic MDP

$V^\star(s) = V_1^\star(s)$, introducing the optimal value function from step $h$ :

$$V_h^\star(s) = \max_\pi \mathbb{E}^\pi \left[ \sum_{t=h}^{H} r_t \middle| s_h = s \right]$$

## Bellman equations for the optimal policy

For all $h \in \{1, \dots, H\}$, for all $s \in \mathcal{S}$, letting $V_{H+1}^\star = 0$

$$\begin{aligned} Q_h^\star(s,a) &= r_h(s,a) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a) \left[\max_b Q_{h+1}^\star(s',b)\right] \\ \pi_h^\star(s) &= \underset{a \in \mathcal{A}}{\text{argmax}}\ Q_h^\star(s,a) \end{aligned}$$

➜ The optimal policy $\pi^\star$ is non-stationnary and can be computed using backwards induction (dynamic programming)

# (Recap :) Solving an episodic MDP

$V^\star(s) = V_1^\star(s)$, introducing the optimal value function from step $h$ :

$$V_h^\star(s) = \max_\pi \mathbb{E}^\pi \left[ \sum_{t=h}^{H} r_t \middle| s_h = s \right]$$

## Bellman equations for the optimal policy

For all $h \in \{1, \dots, H\}$, for all $s \in \mathcal{S}$, letting $V_{H+1}^\star = 0$

$$\begin{aligned} Q_h^\star(s,a) &= r_h(s,a) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a) \left[\max_b Q_{h+1}^\star(s',b)\right] \\ \pi^\star &= \text{greedy}(Q^\star) \end{aligned}$$

➜ The optimal policy $\pi^\star$ is non-stationnary and can be computed using backwards induction (dynamic programming)

# Learning in episodic MDPs

For each episode $t$, an episodic RL algorithm

- starts in some initial state $s_1^t \sim \rho$ (e.g. $s_1^t = s_1$)
- selects a policy $\pi^t$ (based on observations from past episodes)
- uses this policy to generate an episode of length $H$ :

$$s_1^t, a_1^t, r_1^t, s_2^t, \dots, s_H^t, a_H^t, r_H^t$$

where $a_h^t = \pi_h^t(s_h^t)$ and $(r_h^t, s_{h+1}^t) = \text{step}(s_h^t, a_h^t)$

## Definition

The (pseudo)-regret of an algorithm $\pi = (\pi^t)_{t\in\mathbb{N}}$ in $T$ episodes is

$$R_T(\pi) = \sum_{t=1}^{T} \left[ V^\star(s_1^t) - V^{\pi^t}(s_1^t) \right].$$

# Learning in episodic MDPs

For each episode $t$, an episodic RL algorithm

- starts in some initial state $s_1^t \sim \rho \quad$ (e.g. $s_1^t = s_1$)
- selects a policy $\pi^t$ (based on observations from past episodes)
- uses this policy to generate an episode of length $H$ :

$$s_1^t, a_1^t, r_1^t, s_2^t, \dots, s_H^t, a_H^t, r_H^t$$

where $a_h^t = \pi_h^t(s_h^t)$ and $(r_h^t, s_{h+1}^t) = \text{step}(s_h^t, a_h^t)$

- the algorithm may decide to stop after $\tau = t$ episodes, and output $\widehat{\pi}_t$

**Definition**

The algorithm is $(\varepsilon, \delta)$-PAC for Best Policy Identification if

$$\mathbb{P}\left(V^\star(s_1) - V^{\widehat{\pi}_\tau}(s_1) \leq \varepsilon\right) \geq 1 - \delta.$$

The sample complexity is $\tau$, the number of episodes before stopping.

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Bandits without rewards ?

$\mathcal{B}(\mu_1)$

$\mathcal{B}(\mu_2)$

$\mathcal{B}(\mu_3)$

$\mathcal{B}(\mu_4)$

$\mathcal{B}(\mu_5)$

For the $t$-th patient in a clinical study,

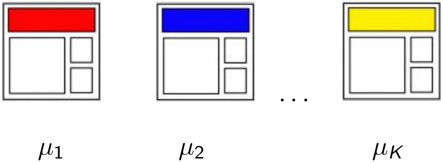
- choose a treatment $A_t$
- observe a response $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t}$

**Maximize rewards** $\leftrightarrow$ cure as many patients as possible

**Alternative goal :** identify as quickly as possible the best treatment (without trying to cure patients during the study)

# Bandits without rewards ?

Probability that some version of a website generates a conversion :



**Best version** : $a_\star = \underset{a=1,\dots,K}{\operatorname{argmax}} \mu_a$

**Sequential protocol** : for the $t$-th visitor :

- display version $A_t$
- observe conversion indicator $X_t \sim \mathcal{B}(\mu_{A_t})$.

**Maximize rewards** $\leftrightarrow$ maximize the number of conversions

**Alternative goal :** identify the best version
(without trying to maximize conversions during the test)

# A Pure Exploration Problem

$$\mu_\star = \max_a \mu_a \qquad a_\star = \underset{a=1,\dots,K}{\mathrm{argmax}} \ \mu_a$$

**Goal :** identify an arm with mean close to $\mu_\star$ as quickly and accurately as possible (e.g. identify $a_\star$ if it is unique)

**Algorithm :** made of three components :

- ➜ sampling rule : $A_t$ (arm to explore)
- ➜ recommendation rule : $B_t$ (current guess for the best arm)
- ➜ stopping rule $\tau$ (when do we stop exploring ?)

**Probability of error**

The probability of error after $T$ rounds is

$$p_\nu(T) = \mathbb{P}_\nu \left( B_T \notin a_\star(\nu) \right).$$

# A Pure Exploration Problem

$$\mu_\star = \max_a \mu_a \qquad a_\star = \underset{a=1,\dots,K}{\operatorname{argmax}} \ \mu_a$$

**Goal :** identify an arm with mean close to $\mu_\star$ as quickly and accurately as possible (e.g. identify $a_\star$ if it is unique)

**Algorithm :** made of three components :

- ➜ sampling rule : $A_t$ (arm to explore)
- ➜ recommendation rule : $B_t$ (current guess for the best arm)
- ➜ stopping rule $\tau$ (when do we stop exploring ?)

**Simple regret** [Bubeck et al., 2011]

The simple regret after $n$ rounds is

$$r_\nu(n) = \mu_\star - \mu_{B_n}.$$

# A Pure Exploration Problem

$$\mu_\star = \max_a \mu_a \qquad a_\star = \underset{a=1,\dots,K}{\mathrm{argmax}} \ \mu_a$$

**Goal :** identify an arm with mean close to $\mu_\star$ as quickly and accurately as possible (e.g. identify $a_\star$ if it is unique)

**Algorithm :** made of three components :

- ➜ sampling rule : $A_t$ (arm to explore)
- ➜ recommendation rule : $B_t$ (current guess for the best arm)
- ➜ stopping rule $\tau$ (when do we stop exploring ?)

**Simple regret [Bubeck et al., 2011]**

The simple regret after $n$ rounds is

$$r_\nu(n) = \mu_\star - \mu_{B_n}.$$

$$\Delta_{\min} p_\nu(T) \leq \mathbb{E}_\nu[r_\nu(T)] \leq \Delta_{\max} p_\nu(T)$$

# Several objectives

**Algorithm :** made of three components :

➜ sampling rule : $A_t$ (arm to explore)

➜ recommendation rule : $B_t$ (current guess for the best arm)

➜ stopping rule $\tau$ (when do we stop exploring ?)

▶ **Objectives studied in the literature :**

| Fixed-budget setting | Fixed-confidence setting |
|---|---|
| input : budget $T$ | input : risk parameter $\delta$<br>(tolerance parameter $\epsilon$) |
| $\tau = T$<br>minimize $\mathbb{P}(B_T \neq a_\star)$<br>or $\mathbb{E}[r_T(\nu)]$ | minimize $\mathbb{E}[\tau]$<br>$\mathbb{P}(B_\tau \neq a_\star) \leq \delta$<br>or $\mathbb{P}(r_\nu(\tau) > \epsilon) \leq \delta$ |
| [Bubeck et al., 2011]<br>[Audibert et al., 2010] | [Even-Dar et al., 2006] |

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Can we use UCB ?

**Context :** bounded rewards ($\nu_a$ supported in $[0,1]$)

We know good algorithms to maximize rewards, for example UCB($\alpha$)

$$A_{t+1} = \underset{a=1,\dots,K}{\operatorname{argmax}} \ \hat{\mu}_a(t) + \sqrt{\frac{\alpha \ln(t)}{N_a(t)}}$$

- How good is it for best arm identification ?

# Can we use UCB ?

**Context :** bounded rewards ($\nu_a$ supported in $[0,1]$)

We know good algorithms to maximize rewards, for example UCB($\alpha$)

$$A_{t+1} = \underset{a=1,\dots,K}{\text{argmax}}\ \hat{\mu}_a(t) + \sqrt{\frac{\alpha \ln(t)}{N_a(t)}}$$

- How good is it for best arm identification ?

**Possible recommendation rules :**

| | |
|---|---|
| Empirical Best Arm<br>(EBA) | $B_t = \text{argmax}_a\ \hat{\mu}_a(t)$ |
| Most Played Arm<br>(MPA) | $B_t = \text{argmax}_a\ N_a(t)$ |
| Empirical Distribution of Plays<br>(EDP) | $B_t \sim p_t$, where<br>$p_t = \left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t}\right)$ |

[Bubeck et al., 2011]

# Can we use UCB ?

**Context :** bounded rewards ($\nu_a$ supported in $[0,1]$)

We know good algorithms to maximize rewards, for example UCB($\alpha$)

$$A_{t+1} = \underset{a=1,\dots,K}{\mathrm{argmax}} \ \hat{\mu}_a(t) + \sqrt{\frac{\alpha \ln(t)}{N_a(t)}}$$

- How good is it for best arm identification ?

**Possible recommendation rules** :

| | |
|---|---|
| Empirical Best Arm<br>(EBA) | $B_t = \mathrm{argmax}_a \ \hat{\mu}_a(t)$ |
| Most Played Arm<br>(MPA) | $B_t = \mathrm{argmax}_a \ N_a(t)$ |
| Empirical Distribution of Plays<br>(EDP) | $B_t \sim p_t$, where<br>$p_t = \left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t}\right)$ |

[Bubeck et al., 2011]

# Can we use UCB ?

▶ **UCB + Empirical Distribution of Plays**

$$\begin{aligned}
\mathbb{E}\left[r_\nu(T)\right] &= \mathbb{E}\left[\mu_\star - \mu_{B_T}\right] = \mathbb{E}\left[\sum_{b=1}^{K}(\mu_\star - \mu_b)\mathbb{1}_{(B_T=b)}\right] \\
&= \mathbb{E}\left[\sum_{b=1}^{K}(\mu_\star - \mu_b)\mathbb{P}(B_T = b|\mathcal{F}_T)\right] \\
&= \mathbb{E}\left[\sum_{b=1}^{K}(\mu_\star - \mu_b)\frac{N_b(T)}{T}\right] \\
&= \frac{1}{T}\sum_{b=1}^{K}(\mu_\star - \mu_b)\mathbb{E}[N_b(T)] \\
&= \frac{\mathcal{R}_\nu(T)}{T}.
\end{aligned}$$

➜ a conversion from cumulative regret to simple regret !

# Can we use UCB ?

- **UCB + Empirical Distribution of Plays**

$$\mathbb{E}\left[r_\nu\left(\texttt{UCB}(\alpha), T\right)\right] \leq \frac{\mathcal{R}_\nu(\texttt{UCB}(\alpha), T)}{T} \leq \frac{C(\nu)\ln(T)}{T}$$

# Can we use UCB ?

- **UCB + Empirical Distribution of Plays**

$$\mathbb{E}\left[r_\nu\left(\texttt{UCB}(\alpha), T\right)\right] \leq \frac{\mathcal{R}_\nu(\texttt{UCB}(\alpha), T)}{T} \leq C\sqrt{\frac{K\ln(T)}{T}}$$

# Can we use UCB ?

- **UCB + Empirical Distribution of Plays**

$$\mathbb{E}\left[r_\nu\left(\texttt{UCB}(\alpha), T\right)\right] \leq \frac{\mathcal{R}_\nu(\texttt{UCB}(\alpha), T)}{T} \leq C\sqrt{\frac{K\ln(T)}{T}}$$

- Almost optimal in the **worse case**

### Lower bound [Bubeck et al., 2011]

For every algorithm $\mathcal{A}$, there exists a bandit instance $\nu$ in which

$$\mathbb{E}[r_\nu(\mathcal{A}, T)] \geq \frac{1}{20}\sqrt{\frac{K}{T}}$$

# Can we use UCB ?

- **UCB + Empirical Distribution of Plays**

$$\mathbb{E}\left[r_\nu\left(\mathtt{UCB}(\alpha), T\right)\right] \leq \frac{\mathcal{R}_\nu(\mathtt{UCB}(\alpha), T)}{T} \leq C\sqrt{\frac{K\ln(T)}{T}}$$

- ... but potentially bad in the **instance-dependent** regime

The simple regret or the uniform sampling strategy decays exponentially :

$$\mathbb{E}_\nu\left[r_\nu\left(\mathtt{Unif}, T\right)\right] \leq (K-1)\Delta_{\max}\exp\left(-\frac{1}{2}\frac{T}{K}\Delta_{\min}^2\right)$$

➜ UCB does not always provably outperform uniform sampling...

# (Instance-dependent) sample complexity

With Uniform Sampling, the number of sample needed to get an error probability smaller than $\delta$ is of order

$$T \simeq \frac{K}{\Delta_{\min}^2} \log\left(\frac{1}{\delta}\right)$$

(for small $\delta$), assuming, e.g. rewards in $[0, 1]$.

Can this be improved for *more adaptive* algorithm, i.e. algorithm using an

- adaptive sampling rule
- adaptive stopping rule

to

$$\mathbb{E}[\tau] \simeq \mathcal{O}\left(H(\nu) \log\left(\frac{1}{\delta}\right)\right)$$

where $H(\mu) < \frac{K}{\Delta_{\min}^2}$ ?

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Successive Elimination

**Input :** risk parameter $\delta \in (0,1)$.
**Idea :** sample all remaining arm uniformly and perform eliminations of arms which look sub-optimal

```
**Initialization :** $\mathcal{S} = \{1, \dots, K\}$
**While** $|\mathcal{S}| > 1$
    Draw all arms in $\mathcal{S}$. $t \leftarrow t + |\mathcal{S}|$.
    $\mathcal{S} \leftarrow \mathcal{S} \backslash \{a\}$ if $\max_{i \in \mathcal{S}} \hat{\mu}_i(t) - \hat{\mu}_a(t) \geq 2\sqrt{\frac{\ln(Kt^2/\delta)}{t}}$.
**Output :** the unique arm $B_\tau \in \mathcal{S}$.
```

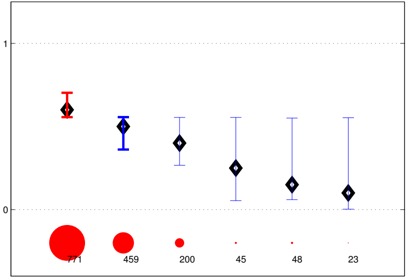
### Theorem [Even-Dar et al., 2006]

Successive Elimination satisfies $\mathbb{P}_\nu\left(B_\tau = a_\star\right) \geq 1 - \delta$. Moreover,

$$\mathbb{P}_\nu\left(\tau_\delta = O\left(\sum_{a=2}^{K} \frac{1}{\Delta_a^2} \ln\left(\frac{K}{\delta\Delta_a}\right)\right)\right) \geq 1 - \delta.$$

# Confidence-based algorithms : LUCB

$$\mathcal{I}_a(t) = [\mathrm{LCB}_a(t), \mathrm{UCB}_a(t)].$$



- At round $t$, draw

$$B_t = \underset{b}{\mathrm{argmax}}\ \hat{\mu}_b(t)$$
$$C_t = \underset{c \neq B_t}{\mathrm{argmax}}\ \mathrm{UCB}_c(t)$$

- Stop at round $t$ if

$$\mathrm{LCB}_{B_t}(t) > \mathrm{UCB}_{C_t}(t) - \epsilon$$

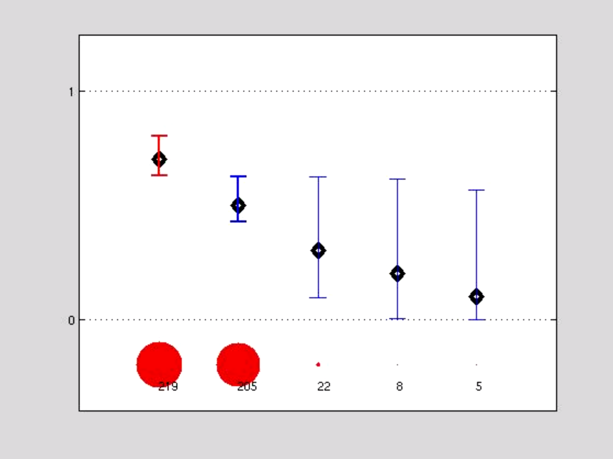
**Theorem** [Kalyanakrishnan et al., 2012]

For well-chosen confidence intervals, $\mathbb{P}_\nu(\mu_{B_\tau} > \mu_\star - \epsilon) \geq 1 - \delta$ and

$$\mathbb{E}[\tau_\delta] = \mathcal{O}\left(\left[\sum_{a=1}^{K} \frac{1}{\Delta_a^2 \vee \epsilon^2}\right] \ln\left(\frac{1}{\delta}\right)\right)$$

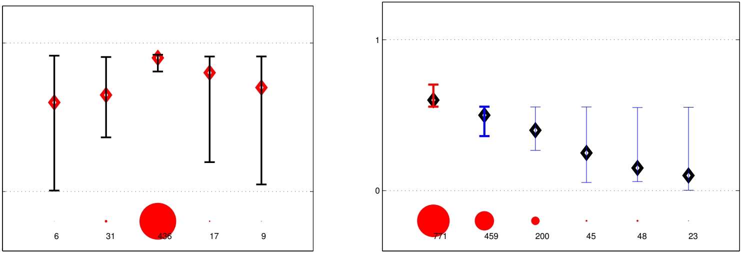
# (kl)-LUCB in action

$$\begin{aligned}\mathrm{UCB}_a(t) &= \max\left\{q \in [0,1] : N_a(t)\mathrm{kl}(\hat{\mu}_a(t), q) \leq \log(Ct^2/\delta)\right\} \\ \mathrm{LCB}_a(t) &= \min\left\{q \in [0,1] : N_a(t)\mathrm{kl}(\hat{\mu}_a(t), q) \leq \log(Ct^2/\delta)\right\}\end{aligned}$$



# A comparison with UCB

Regret minimizing algorithms and Best Arm Identification algorithms behave quite differently



Number of selections and confidence intervals for KL-UCB (left) and KL-LUCB (right)

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# A Sample Complexity Lower Bound

$\delta$-correct algorithm (exact BAI) : $\forall \boldsymbol{\mu}, \mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_\tau = a_\star(\boldsymbol{\mu})) \geq 1 - \delta$.

**Lower Bound** [Garivier and Kaufmann, 2016]

For $\delta$-correct algorithms for Gaussian bandits of variance $\sigma^2$,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T_\star(\boldsymbol{\mu}) \log\left(\frac{1}{3\delta}\right)$$

with

$$\left(T_\star(\boldsymbol{\mu})\right)^{-1} = \sup_{w \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(a_\star(\mu))} \sum_{a \in [K]} w_a \frac{(\mu_a - \lambda_a)^2}{2\sigma^2}$$

$\Delta_K = \{\boldsymbol{w} \in [0,1]^K : \sum_a w_a = 1\}$ and $\mathrm{Alt}(i) = \{\boldsymbol{\lambda} \in \mathbb{R}^K : a_\star(\boldsymbol{\lambda}) \neq i\}$.

**Proof.** Information theoretic argument
For all $\boldsymbol{\nu}'$ : $a_\star(\boldsymbol{\nu}') \neq a_\star(\boldsymbol{\nu})$, for any $\delta$-correct algorithm,

$$\sum_{a \in [K]} \mathbb{E}_{\boldsymbol{\nu}}[N_a(\tau)] \mathrm{KL}(\nu_a, \nu'_a) \geq \log\left(\frac{1}{3\delta}\right)$$

# A Sample Complexity Lower Bound

The "minimal distance" has a closed form :

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(a_\star(\mu))} \sum_{a \in [K]} w_a \frac{(\mu_a - \lambda_a)^2}{2\sigma^2} = \min_{a \neq a_\star} \frac{(\mu_a - \mu_{a_\star})^2}{2\sigma^2 \left( \frac{1}{w_a} + \frac{1}{w_{a_\star}} \right)}$$

but not the characteristic time

$$(T_\star(\boldsymbol{\mu}))^{-1} = \sup_{w \in \Delta_K} \min_{a \neq a_\star} \frac{(\mu_a - \mu_{a_\star})^2}{2\sigma^2 \left( \frac{1}{w_a} + \frac{1}{w_{a_\star}} \right)}$$

## Approximation of the characteristic time

$$\sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2} \leq T_\star(\boldsymbol{\mu}) \leq 2 \left( \sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2} \right)$$

➜ Can we still match this (non-explicit) lower bound ?

# Track-and-Stop

$$\left(T_\star(\boldsymbol{\mu})\right)^{-1} = \sup_{w \in \Delta_K} \min_{a \neq a_\star} \frac{(\mu_a - \mu_{a_\star})^2}{2\sigma^2 \left(\frac{1}{w_a} + \frac{1}{w_{a_\star}}\right)}$$

**Yes**, with an appropriate stopping rule

$$\tau = \inf \left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_a(t) - \hat{\mu}_{\hat{a}_t^\star}(t))^2}{2\sigma^2 \left(\frac{1}{N_a(t)} + \frac{1}{N_{\hat{a}_t^\star}(t)}\right)} > \beta(t, \delta) \right\}$$

where $\hat{a}_t^\star$ is the empirical best arm at time $t$

# Track-and-Stop

$$\left(T_\star(\boldsymbol{\mu})\right)^{-1} = \sup_{w \in \Delta_K} \min_{a \neq a_\star} \frac{(\mu_a - \mu_{a_\star})^2}{2\sigma^2\left(\frac{1}{w_a} + \frac{1}{w_{a_\star}}\right)}$$

**Yes**, with an appropriate GLR stopping rule

$$\tau = \inf\left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_a(t) - \hat{\mu}_{\hat{a}_t^\star}(t))^2}{2\sigma^2\left(\frac{1}{N_a(t)} + \frac{1}{N_{\hat{a}_t^\star}(t)}\right)} > \beta(t,\delta) \right\}$$

where $\hat{a}_t^\star$ is the empirical best arm at time $t$

➜ Generalized Likelihood Ratio Statistic for testing

$$\mathcal{H}_0 : (a_\star(\boldsymbol{\mu}) \neq \hat{a}_t) \quad \text{against} \quad \mathcal{H}_1 : (a_\star(\boldsymbol{\mu}) = \hat{a}_t)$$

# Track-and-Stop

$$(T_\star(\boldsymbol{\mu}))^{-1} = \sup_{w \in \Delta_K} \min_{a \neq a_\star} \frac{(\mu_a - \mu_{a_\star})^2}{2\sigma^2 \left( \frac{1}{w_a} + \frac{1}{w_{a_\star}} \right)}$$

**Yes**, with an appropriate GLR stopping rule

$$\tau = \inf \left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_a(t) - \hat{\mu}_{\hat{a}_t^\star}(t))^2}{2\sigma^2 \left( \frac{1}{N_a(t)} + \frac{1}{N_{\hat{a}_t^\star}(t)} \right)} > \beta(t, \delta) \right\}$$

where $\hat{a}_t^\star$ is the empirical best arm at time $t$
... and a sampling rule satisfying

$$\left( \frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t} \right) \to w^\star(\boldsymbol{\mu})$$

where $w^\star(\boldsymbol{\mu})$ is the maximizer in $w \in \Delta_K$

# Track-and-Stop

**Tracking sampling rule** : letting $U_t = \left\{a : N_a(t) < \sqrt{t}\right\}$,

$$A_{t+1} \in \begin{cases} \underset{a \in U_t}{\text{argmin}}\; N_a(t) \text{ if } U_t \neq \emptyset & (\textit{forced exploration}) \\ \underset{1 \leq a \leq K}{\text{argmax}} \left[ w_a^\star(\hat{\boldsymbol{\mu}}(t)) - \frac{N_a(t)}{t} \right] & (\textit{tracking}) \end{cases}$$

**Theorem** [Garivier and Kaufmann, 2016, Kaufmann and Koolen, 2021]

The Track-and-Stop strategy, that uses

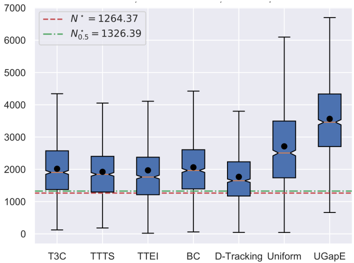
- the Tracking sampling rule
- the GLR stopping rule with $\beta(t, \delta) \simeq \log\left(\frac{K \log(t)}{\delta}\right)$
- and recommends $\hat{\imath}_t = a_\star(\hat{\boldsymbol{\mu}}(t))$

is $\delta$-correct for every $\delta \in ]0, 1[$ and satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} = T^\star(\boldsymbol{\mu}).$$

# In practise

Empirical distribution of $\tau_\delta$ for $\delta = 0.01$ for different algorithms on $\boldsymbol{\mu} = [1, 0.8, 0.75, 0.7]\,, \sigma^2 = 1$, estimated on 1000 runs



Using the right stopping rule makes a difference :

$$\text{LUCB} : \quad \forall a \neq \hat{a}_t^\star \quad , \; \hat{\mu}_{\hat{a}_t^\star}(t) - \hat{\mu}_a(t) > \sqrt{\frac{2\sigma^2\beta(t,\delta)}{N_{\hat{a}_t^\star}(t)}} + \sqrt{\frac{2\sigma^2\beta(t,\delta)}{N_a(t)}}$$

$$\text{GLR} : \quad \forall a \neq \hat{a}_t^\star \quad , \; \hat{\mu}_{\hat{a}_t^\star}(t) - \hat{\mu}_a(t) > \sqrt{2\sigma^2\beta(t,\delta)\left(\frac{1}{N_{\hat{a}_t^\star}(t)} + \frac{1}{N_a(t)}\right)}$$

Limitation : Higher computational cost for TaS ($w_\star$)

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

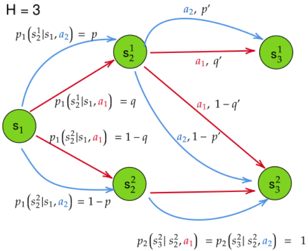
2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Finite Horizon Tabular MDPs

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, s_1)$$


H = 3
$p_1(s_2^1|s_1, a_2) = p$
$a_2,\ p'$
$a_1,\ q'$
$p_1(s_2^1|s_1, a_1) = q$
$a_1,\ 1-q'$
$p_1(s_2^2|s_1, a_1) = 1-q$
$a_2, 1-p'$
$p_1(s_2^2|s_1, a_2) = 1-p$
$p_2(s_3^2|\, s_2^2, a_1) = p_2(s_3^2|\, s_2^2, a_2) = 1$
$s_1$
$s_2^1$
$s_2^2$
$s_3^1$
$s_3^2$

## Value function

For a policy $\pi = \{\pi_h\}_{h\in[H]}$ for a reward function $r : [H] \times \mathcal{S} \times \mathcal{A} \to [0, 1]$

$$V_h^\pi(s; r) = \mathbb{E}^\pi\left[\sum_{\ell=h}^{H} r_\ell(S_\ell, A_\ell) \middle| S_h = s\right] \qquad \begin{array}{rcl} A_\ell & \sim & \pi_\ell(S_\ell) \\ S_{\ell+1} & \sim & p_\ell(\cdot|S_\ell, A_\ell) \end{array}$$

# Online episodic algorithm

In each episode $t = 1, 2, \ldots,$ the agent

- selects an exploration policy $\pi^t$ based on past data $\mathcal{D}_{t-1}$
- collects an episode under this policy

$$\mathcal{D}_t = \mathcal{D}_{t-1} \cup \left\{ \left( s_1^t, a_1^t, s_2^t, a_2^t, \ldots, s_H^t, a_H^t \right) \right\}$$

  where $s_1^t = s_1$, $a_h^t \sim \pi_h^t(s_h^t)$ and $s_{h+1}^t \sim p_h(\cdot | s_h^t, a_h^t)$
- can decide to stop exploration → adaptive stopping time $\tau$
- if so, can output a prediction, e.g. a good policy $\widehat{\pi}$

**Goal :** make a Probaby Approximately Correct (PAC) prediction

**Performance metric :** Sample Complexity $\tau$
$=$ number of episodes needed before stopping

# Best Policy Identification (BPI)

➜ Learn the optimal policy for a known reward function $r$

[Fiechter, 1994]

Algorithm :

- exploration policy $\pi^t$
- stopping rule $\tau$
- $\widehat{\pi}$ : guess for a good policy

**$(\varepsilon, \delta)$-PAC algorithm for Best Policy Identification**

$$\mathbb{P}\left(V_1^\star(s_1; r) - V_1^{\widehat{\pi}}(s_1; r) \leq \varepsilon\right) \geq 1 - \delta$$

# Reward Free Exploration (RFE)

➜ Learn the optimal policy for **any** reward function $r$ given afterwards

[Jin et al., 2020]

Algorithm :

- exploration policy $\pi^t$
- stopping rule $\tau$
- for any $r = (r_h(s,a)) \in [0,1]^{HSA}$, guess $\widehat{\pi}_r$ for a good policy

$(\varepsilon, \delta)$-PAC algorithm for Reward-Free Exploration

$$\mathbb{P}\left(\text{for any } r \in \mathcal{B}, V_1^\star(s_1; r) - V_1^{\widehat{\pi}_r}(s_1; r) \leq \varepsilon\right) \geq 1 - \delta$$

# Reward Free Exploration (RFE)

➜ Learn the optimal policy for **any** reward function $r$ given afterwards

[Jin et al., 2020]

Algorithm :

- exploration policy $\pi^t$
- stopping rule $\tau$
- for any $r \in \mathcal{B}$, guess $\widehat{\pi}_r$ for a good policy

$(\varepsilon, \delta)$-PAC algorithm for Reward-Free Exploration

$$\mathbb{P}\left(\text{for any } r \in \mathcal{B}, V_1^\star(s_1; r) - V_1^{\widehat{\pi}_r}(s_1; r) \leq \varepsilon\right) \geq 1 - \delta$$

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Optimistic RL algorithm

## Bellman equation

$$\pi_h^\star = \text{greedy}(Q_h^\star) \;\text{ with }\; Q_h^\star(s,a) = r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_b Q_{h+1}^\star(s',b)$$

**Optimism** : $\pi_h^{t+1} = \text{greedy}(\overline{Q}_h^t)$ where

$$\overline{Q}_h^t(s,a) = \max_{p \in \mathcal{M}_t} \left[ r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_b \overline{Q}_{h+1}^t(s',b) \right]$$

where $\mathcal{M}_t$ is a set of plausible MDPs.

NB : for simplicity, we assume that the reward function is known

# Plausible MDPs

$$\begin{aligned}
n_h^t(s,a) &= \sum_{\ell=1}^{t} \mathbb{1}(s_h^\ell = s, a_h^\ell = a) \\
n_h^t(s,a,s') &= \sum_{\ell=1}^{t} \mathbb{1}(s_h^\ell = s, a_h^\ell = a, s_{h+1}^\ell = s') \\
\hat{p}_h^t(s'|s,a) &= \frac{n_h^t(s,a,s')}{n_h^t(s,a)}
\end{aligned}$$

Set of plausible transition probabilities from $(h,s,a)$ :

$$\mathcal{C}_h^t(s,a) = \left\{ p \in \Delta_S : \mathrm{KL}(\hat{p}_h^t(\cdot|s,a), p) \leq \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\}$$

$\mathcal{M}_t(\delta) = \{\boldsymbol{p} = (p_h(\cdot|s,a))_{h,s,a} : \forall t \in \mathbb{N}, \forall (h,s,a), p_h(\cdot|s,a) \in \mathcal{C}_h^t(s,a)\}.$

**Time-uniform concentration [Jonsson et al., 2020]**

For $\beta(n,\delta) = \log(SAH/\delta) + (S-1)\log(e(1+n/(S-1)))$, for any $\boldsymbol{p}$,

$$\mathbb{P}_{\boldsymbol{p}}\left(\forall t \in \mathbb{N}, \forall (h,s,a), p_h(\cdot|s,a) \in \mathcal{C}_h^t(s,a)\right) \geq 1-\delta.$$

# An upper bound on $\overline{Q}_h$

$$\overline{Q}_h^t(s,a) = \max_{p \in \mathcal{C}_h^t(s,a)} \left[ r_h(s,a) + \sum_{s'} p(s') \max_b \overline{Q}_{h+1}^t(s',b) \right]$$

with $\mathcal{C}_h^t(s,a) = \left\{ p : \mathrm{KL}(\hat{p}_h^t(\cdot|s,a), p) \leq \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\}$

# An upper bound on $\overline{Q}_h$

$$\overline{Q}_h^t(s,a) = \max_{p \in \mathcal{C}_h^t(s,a)} \left[ r_h(s,a) + \sum_{s'} p(s') \max_b \overline{Q}_{h+1}^t(s',b) \right]$$

with $\mathcal{C}_h^t(s,a) = \left\{ p : \mathrm{KL}(\hat{p}_h^t(\cdot|s,a), p) \leq \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\}$

For any $p$ in $\mathcal{C}_h^t(s,a)$,

$$\begin{aligned}
&\sum_{s'} p(s')\overline{V}_{h+1}^t(s') \\
&= \sum_{s'} \hat{p}_h^t(s'|s,a)\overline{V}_{h+1}^t(s') + \sum_{s'} (p(s') - \hat{p}_h^t(s'|s,a))\overline{V}_{h+1}^t(s') \\
&\leq \sum_{s'} \hat{p}_h^t(s'|s,a)\overline{V}_{h+1}^t(s') + ||\hat{p}_h^t(\cdot|s,a) - p||_1 (H-h) \\
&\leq \sum_{s'} \hat{p}_h^t(s'|s,a)\overline{V}_{h+1}^t(s') + (H-h)\sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}
\end{aligned}$$

where the last step follows from Pinsker's inequality.

# UCB-VI

## UCB-VI style algorithm

$\pi_h^{t+1} = \text{greedy}\left(\overline{Q}_h^t\right)$ for the optimistic Q-function

$$\begin{aligned}\overline{Q}_h^t(s,a) &= \left[r_h(s,a) + \mathrm{B}_h^t(s,a) + \sum_{s'\in\mathcal{S}} \hat{p}_h^t(s'|s,a)\max_b \overline{V}_{h+1}^t(s')\right] \wedge (H-h)\\ \overline{V}_h^t(s) &= \max_b \overline{Q}_h^t(s,b).\end{aligned}$$

Different exploration bonuses $\mathrm{B}_h^t(s,a)$ yield different guarantees

- Hoeffding bonuses $\mathrm{B}_h^t(s,a) \simeq \sqrt{\frac{\log(SAH/\delta)+S\log(n_h^t(s,a))}{n_h^t(s,a)}}$ (*"UCRL"*)
- Bernstein bonuses (more complex) (*UCB-VI* [Azar et al., 2017b])

$n_h^t(s,a)$ : number of visits of $(s,a)$ in step $h$ in the first $t$ episodes

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Regret guarantees

The (pseudo)-regret of an episodic RL algorithm $\pi = (\pi^t)_{t\in\mathbb{N}}$ is

$$\mathcal{R}_T(\pi) = \sum_{t=1}^{T} \left[ V_1^{\star}(s_1^t) - V_1^{\pi^t}(s_1^t) \right].$$

### Regret of UCB-VI [Azar et al., 2017b]

For appropriately chosen bonuses (depending on $\delta$) UCB-VI satisfies

$$\mathbb{P}\left(\mathcal{R}_T(\pi) = \mathcal{O}\left(\sqrt{H^3SAT}\right)\right) \geq 1 - \delta$$

which is minimax optimal in time-inhomogeneous MDPs. [Domingues et al., 2021]

# From regret to PAC guarantees

The (pseudo)-regret of an episodic RL algorithm $\pi = (\pi^t)_{t\in\mathbb{N}}$ is

$$\mathcal{R}_T(\pi) = \sum_{t=1}^{T} \left[ V_1^\star(s_1^t) - V_1^{\pi^t}(s_1^t) \right].$$

## Regret to PAC conversion [Jin et al., 2018]

Running UCB-VI for $T = \mathcal{O}\left(\frac{SAH^3}{\varepsilon^2\delta^2}\right)$ and outputting

$$\widehat{\pi} = \pi^N \quad \text{where} \quad N \sim \mathcal{U}(\{1,\dots,T\})$$

yields an $(\varepsilon, \delta)$-PAC algorithm for Best Policy Identification.

Proof. On the board.

# Is it optimal ?

The regret-to-PAC conversion yields a deterministic (fixed in advance) sample complexity

$$T = \mathcal{O}\left(\frac{SAH^3}{\varepsilon^2 \delta^2}\right)$$

## Minimax lower bound [Domingues et al., 2021]

For any $(\varepsilon, \delta)$-PAC BPI algorithm, there exists an MDP for which $\mathbb{E}[\tau] \geq c \frac{SAH^3}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)$

➜ to get the optimal dependency in $\delta$ as well, we need adaptive stopping rules

# An adaptive stopping rule

**BPI-UCRL** [Kaufmann et al., 2021] : UCB-VI with Hoeffding bonuses together with the stopping rule

$$\tau = \inf\left\{t \in \mathbb{N} : \overline{V}_1^t(s_1) - \underline{V}_1^t(s_1) \leq \varepsilon\right\} \quad \hat{\pi} = \text{greedy}(\underline{Q}^\tau)$$

where we define upper *and lower* bounds on the optimal values.

$$\begin{aligned} \underline{Q}_h^t(s,a) &= \min_{p \in \mathcal{C}_h^t(s,a)} \left[ r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_b \overline{Q}_{h+1}^t(s',b) \right] \\ \underline{V}_h^t(s) &= \max_a \underline{Q}_h^t(s,a) \end{aligned}$$

# An adaptive stopping rule

**BPI-UCRL** [Kaufmann et al., 2021] : UCB-VI with Hoeffding bonuses together with the stopping rule

$$\tau = \inf\left\{t \in \mathbb{N} : \overline{V}_1^t(s_1) - \underline{V}_1^t(s_1) \leq \varepsilon\right\} \quad \hat{\pi} = \text{greedy}(\underline{Q}^\tau)$$

where we define upper *and lower* bounds on the optimal values.

$$\begin{aligned} \underline{Q}_h^t(s,a) &= \min_{p \in \mathcal{C}_h^t(s,a)} \left[ r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_b \overline{Q}_{h+1}^t(s',b) \right] \\ \underline{V}_h^t(s) &= \max_a \underline{Q}_h^t(s,a) \end{aligned}$$

- Why is it $(\varepsilon, \delta)$-PAC ?

On some event of probability $\geq 1 - \delta$, all the Q-values are in their corresponding confidence intervals and

$$V_1^{\hat{\pi}}(s_1) \geq \underline{V}_1^{\tau,\hat{\pi}}(s_1) = \underline{V}_1^\tau(s_1) \geq \overline{V}_1^\tau(s_1) - \epsilon \geq V_1^\star(s_1) - \epsilon \ .$$

# An adaptive stopping rule

**BPI-UCRL** [Kaufmann et al., 2021] : UCB-VI with Hoeffding bonuses together with the stopping rule

$$\tau = \inf\left\{ t \in \mathbb{N} : \overline{V}_1^t(s_1) - \underline{V}_1^t(s_1) \leq \varepsilon \right\} \quad \hat{\pi} = \text{greedy}(\underline{Q}^\tau)$$

where we define upper *and lower* bounds on the optimal values.

$$\begin{aligned} \underline{Q}_h^{t,\pi}(s,a) &= \min_{p \in \mathcal{C}_h^t(s,a)} \left[ r_h(s,a) + \sum_{s'} p_h(s'|s,a) \overline{Q}_{h+1}^t(s', \pi(s')) \right] \\ \underline{V}_h^{t,\pi}(s) &= \underline{Q}_h^t(s, \pi(s)) \end{aligned}$$

- Why is it $(\varepsilon, \delta)$-PAC ?

On some event of probability $\geq 1 - \delta$, all the Q-values are in their corresponding confidence intervals and

$$V_1^{\hat{\pi}}(s_1) \geq \underline{V}_1^{\tau,\hat{\pi}}(s_1) = \underline{V}_1^\tau(s_1) \geq \overline{V}_1^\tau(s_1) - \epsilon \geq V_1^\star(s_1) - \epsilon \ .$$

# Towards a Minimax Optimal BPI Algorithm

**Theorem** [Kaufmann et al., 2021]

Using UCB-VI with Hoeffding bonuses together with

$$\tau = \inf\left\{t \in \mathbb{N} : \overline{V}_1^t(s_1) - \underline{V}_1^t(s_1) \le \varepsilon\right\} \quad \hat{\pi} = \text{greedy}(\underline{Q}^\tau)$$

yields an $(\varepsilon, \delta)$-PAC algorithm with $\mathbb{P}\left(\tau = \widetilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)\right) \ge 1 - \delta$.

➜ using Bernstein bonuses and a more sophisticated stopping rule yields a minimax optimal $\widetilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ sample complexity [Ménard et al., 2021]

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Reward-Free UCRL

### RF-UCRL [Kaufmann et al., 2021]

$$\pi_h^{t+1} = \text{greedy}\left(\overline{Q}_h^t\right) \text{ for}$$

$$\begin{aligned} \overline{Q}_h^t(s,a) &= \left[\xcancel{r_h^t(s,a)} + \mathrm{B}_h^t(s,a) + \sum_{s'\in\mathcal{S}} \hat{p}_h^t(s'|s,a)\max_b \overline{V}_{h+1}^t(s')\right] \wedge (H-h) \\ \overline{V}^t{}_h(s) &= \max_b \overline{Q}_h^t(s,b). \end{aligned}$$

# Reward-Free UCRL

RF-UCRL [Kaufmann et al., 2021]

$$\pi_h^{t+1} = \text{greedy}\left(\overline{E}_h^t\right) \text{ for}$$

$$\begin{aligned}
\overline{E}_h^t(s,a) &= \left[\xcancel{r_h^t(s,a)} + \mathrm{B}_h^t(s,a) + \sum_{s'\in\mathcal{S}} \hat{p}_h^t(s'|s,a)\max_b \overline{V}_{h+1}^t(s')\right] \wedge (H-h) \\
\overline{V}^t{}_h(s) &= \max_b \overline{E}_h^t(s,b).
\end{aligned}$$

**Why does it work ?** It greedily reduces the estimation error of the value of any policy for any reward function :

$$\forall\pi, \forall r, \forall h, s, a, t \quad |\hat{Q}_h^{t,\pi}(s,a;r) - Q_h^{\pi}(s,a;r)| \leq \overline{E}_h^t(s,a)$$

holds with high probability for some Hoeffding-type bonus $B$

# Reward-Free UCRL

## Reward-Free UCRL

- **exploration policy** : $\pi^{t+1}$ is the greedy policy wrt $\overline{E}^t(s,a)$ :

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \mathrm{argmax}_{a\in\mathcal{A}} \ \overline{E}_h^t(s,a).$$

- **stopping rule** : $\tau = \inf\left\{t \in \mathbb{N} : \overline{E}_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2\right\}$
- **prediction** : $\forall r$, $\widehat{\pi}_r = \pi^\star(\hat{P}^\tau, r)$

# Reward-Free UCRL

## Reward-Free UCRL

- **exploration policy** : $\pi^{t+1}$ is the greedy policy wrt $\overline{E}^t(s,a)$ :

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \text{argmax}_{a\in\mathcal{A}} \ \overline{E}_h^t(s,a).$$

- **stopping rule** : $\tau = \inf\left\{t \in \mathbb{N} : \overline{E}_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2\right\}$
- **prediction** : $\forall r$, $\widehat{\pi}_r = \pi^\star(\hat{P}^\tau, r)$

For a given reward function $r$

$$\begin{aligned} V_1^\star(s_1) - V_1^{\widehat{\pi}_r}(s_1) \quad &= \quad V_1^{\pi^\star}(s_1) - \widehat{V}_1^{\tau,\pi^\star}(s_1) + \underbrace{\widehat{V}_1^{\tau,\pi^\star}(s_1) - \widehat{V}_1^{\tau,\widehat{\pi}_r}(s_1)}_{\leq 0} + \widehat{V}_1^{\tau,\widehat{\pi}_r}(s_1) - V_1^{\widehat{\pi}_r}(s_1) \\ &\leq \quad 2\max_a \overline{E}_1^\tau(s_1,a) \\ &\leq \quad \varepsilon \end{aligned}$$

# Reward-Free UCRL

## Reward-Free UCRL

- **exploration policy** : $\pi^{t+1}$ is the greedy policy wrt $\overline{E}^t(s,a)$ :

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \text{argmax}_{a \in \mathcal{A}} \ \overline{E}_h^t(s,a).$$

- **stopping rule** : $\tau = \inf \left\{ t \in \mathbb{N} : \overline{E}_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2 \right\}$
- **prediction** : $\forall r$, $\widehat{\pi}_r = \pi^\star(\hat{P}^\tau, r)$

## Theorem [Kaufmann et al. 2020]

RF-UCRL is $(\varepsilon, \delta)$-PAC for Reward-Free Exploration and

$$\mathbb{P}\left(\tau^{\text{RF-UCRL}} = \tilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\left[\log\left(\frac{1}{\delta}\right) + S\right]\right)\right) \geq 1 - \delta.$$

# Reward-Free UCRL

## Reward-Free UCRL

- **exploration policy** : $\pi^{t+1}$ is the greedy policy wrt $\overline{E}^t(s,a)$ :

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \text{argmax}_{a \in \mathcal{A}} \, \overline{E}_h^t(s,a).$$

- **stopping rule** : $\tau = \inf\left\{t \in \mathbb{N} : \overline{E}_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2\right\}$
- **prediction** : $\forall r$, $\widehat{\pi}_r = \pi^\star(\hat{P}^\tau, r)$

➜ To get a near-optimal $\widetilde{O}\left(\frac{SAH^3}{\varepsilon^2}(\log(1/\delta) + S)\right)$ sample complexity the algorithm structure and bonus type has to be changed a bit [Ménard et al., 2021]

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Instance dependent results

## Goal

Design $(\varepsilon, \delta)$-PAC algorithms that adapt to the difficulty of each specific MDP $\mathcal{M}$ and get

$$\tau_\delta = \mathcal{O}\left(C_\varepsilon(\mathcal{M}) \log\left(1/\delta\right)\right)$$

where $C_\varepsilon(\mathcal{M})$ is some appropriate complexity term.

The lower bound is complex...

## Theorem [Al Marjani et al., 2023b]

Any PAC RL algorithm that is $(\epsilon, \delta)$-PAC on MDP with Gaussian rewards with variance 1 satisfies for any $\mathcal{M}$,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\mathcal{M}}[\tau]}{\log(1/\delta)} \geq \mathcal{C}_{\mathrm{LB}}(\mathcal{M}, \epsilon)$$

where

$$\mathcal{C}_{\mathrm{LB}}(\mathcal{M}, \epsilon) := 2 \min_{\pi^\epsilon \in \Pi^\epsilon} \min_{\rho \in \Omega} \max_{\pi \in \Pi^D} \sum_{s,a,h} \frac{\left(p_h^\pi(s,a) - p_h^{\pi^\epsilon}(s,a)\right)^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\epsilon) + \epsilon)^2}.$$

# Algorithmic attempts

For $\varepsilon = 0$ (exact BPI is there is a unique optimal policy), a recent algorithm based on posterior sampling matches the lower bound [Kone and Jameison, 2026]

Prior attempts for $\varepsilon > 0$

- EPRL [Tirinzoni et al., 2022] for deterministic MDPs
- MOCA [Wagenmaker et al., 2022] (gap-visitation complexity)
- PEDEL [Wagenmaker and Jamieson, 2022]
- PRINCIPLE [Al Marjani et al., 2023a]

These algorithms feature different complexity measures, and some mechanisms to visit certain triplets $(h, s, a)$ proportionally to some instance-dependent quantity ("policy gap" or "value gap").

# A Coverage Algorithm

PRINCIPLE [Al Marjani et al., 2023a] relies on optimal coverage

## $\delta$-correct $c$-coverage

Let $c : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ be a target function.
An algorithm $(\pi^t)_{t\in\mathbb{N}}$ is a $\delta$-correct $c$-coverage if it interacts with $\mathcal{M}$ and return a dataset $\mathcal{D}_t$ such that

$$\mathbb{P}\left(\exists t \geq 1,\ \forall (h,s,a),\ n_h^t(s,a) \geq c_h(s,a)\right) \geq 1-\delta.$$

where $n_h^t(s,a)$ is the number of visits of $(h,s,a)$ in $\mathcal{D}_t$

**Sample complexity :**

$$\tau = \inf\left\{t \in \mathbb{N} : \forall h,s,a, n_h^t(s,a) \geq c_h(s,a)\right\}$$

# Lower bound

**Theorem** [Al Marjani et al., 2023a]

For any target function $c$ and $\delta \in [0,1)$, the stopping time $\tau$ of any $\delta$-correct $c$-coverage algorithm satisfies $\mathbb{E}[\tau] \geq (1-\delta)\varphi^\star(c)$, where

$$\varphi^\star(c) = \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{(s,a,h)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)} ,$$

with $\mathcal{X} := \{(h,s,a) : c_h(s,a) > 0\}$.

**Intuition :** $\frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$ is the expected number of episodes needed before getting $c_h(s,a)$ visits from $(h,s,a)$ using exploration policy $\pi_{\text{exp}}$.

# Towards a coverage algorithm

$$\varphi^\star(c) = \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{(h,s,a)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$$

with $\mathcal{X} = \{(h,s,a) : c_h(s,a) > 0\}$

$$\frac{1}{\varphi^\star(c)} = \sup_{\pi_{\text{exp}} \in \Pi_S} \min_{(s,a,h)\in\mathcal{X}} \frac{p_h^{\pi_{\text{exp}}}(s,a)}{c_h(s,a)}$$

# Towards a coverage algorithm

$$\varphi^{\star}(c) = \inf_{\pi_{\exp}\in\Pi_S} \max_{(h,s,a)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\exp}}(s,a)}$$

with $\mathcal{X} = \{(h,s,a) : c_h(s,a) > 0\}$

$$\begin{aligned}\frac{1}{\varphi^{\star}(c)} &= \sup_{\pi_{\exp}\in\Pi_S} \min_{(s,a,h)\in\mathcal{X}} \frac{p_h^{\pi_{\exp}}(s,a)}{c_h(s,a)} \\ &= \sup_{\pi_{\exp}\in\Pi_S} \inf_{\lambda\in\Delta_{\mathcal{X}}} \sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)}\end{aligned}$$

where $\Delta_{\mathcal{X}}$ is the simplex over $\mathcal{X}$.

# Towards a coverage algorithm

$$\varphi^\star(c) = \inf_{\pi_{\exp}\in\Pi_S} \max_{(h,s,a)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\exp}}(s,a)}$$

with $\mathcal{X} = \{(h,s,a) : c_h(s,a) > 0\}$

$$\frac{1}{\varphi^\star(c)} = \sup_{\pi_{\exp}\in\Pi_S} \min_{(s,a,h)\in\mathcal{X}} \frac{p_h^{\pi_{\exp}}(s,a)}{c_h(s,a)}$$

$$= \sup_{\pi_{\exp}\in\Pi_S} \inf_{\lambda\in\Delta_{\mathcal{X}}} \sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)}$$

$$= \text{value of a game!}$$

where $\Delta_{\mathcal{X}}$ is the simplex over $\mathcal{X}$.

# CovGame

$$\frac{1}{\varphi^\star(c)} = \sup_{\pi_{\exp}\in\Pi_S} \inf_{\lambda\in\Delta_{\mathcal{X}}} \sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)}$$

- $\sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)} = V^{\pi_{\exp}}(s_1;\widetilde{r})$
  value function for the reward function $\widetilde{r}_h(s,a) = \frac{\lambda_h(s,a)}{c_h(s,a)}$
- $\sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)} = \lambda^\top \left(p^{\pi_{\exp}}/c\right)$
  linear loss function

unknown MDP : $V^{\pi_{\exp}}$ and $p^{\pi_{\exp}}$ cannot be computed

➜ combine UCB-VI and an online learning algorithm : CovGame

# The form of CovGame

CovGame can be written

$$\pi^t(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \overline{Q}_h^t\left(s, a; \widetilde{r}_h^t\right)$$

$$\overline{Q}_h^t(s, a; r) = \left[r_h(s,a) + B_h^t(s,a) + \sum_{s' \in \mathcal{S}} \widehat{p}_h^t(s'|s,a) \max_b \overline{Q}_{h+1}^t(s, b; r)\right] \wedge 1$$

for the a time-varying reward $\widetilde{r}^t \in \Delta_{\mathcal{X}}$

$$\widetilde{r}_h^t(s,a) \propto \exp\left(-\eta_t \left[n_h^t(s,a) - n_h^{m_t}(s,a)\right]\right) \mathbb{1}\left(c_h(s,a) > c_{\min} 2^{k_t}\right)$$

➜ another form of intrisic reward

# Outline

1. Regret vs Sample Complexity in the Bandit Case ($H = 1$)
   - Is UCB enough ?
   - Fixed-confidence algorithms
   - Optimal Best Arm Identification

2. Back to $H > 1$
   - Recap : Optimistic Algorithms
   - Optimism for Best Policy Identification
   - ... and Reward-Free Exploration

3. Beyond Minimax Guarantees

4. Beyond episodic MDPs

# Average Rewards MDPs

$$\text{gain :} \quad g^{\pi}(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\pi}\left[\sum_{t=1}^{T} r_t \middle| s_1 = s\right]$$

Poisson equations :

$$g^{\star} + b^{\star}(s) = \max_{a}\left\{r(s,a) + \sum_{s'} p(s'|s,a)b^{\star}(s')\right\}$$

$$\pi^{\star}(s) = \arg\max_{a}\left\{r(s,a) + \sum_{s'} p(s'|s,a)b^{\star}(s')\right\}$$

(in communicating MDPs, $g^{\star}(s) = g^{\star}$)

# Best Policy Identification Algorithm

At step $t = 1, 2, \ldots,$ the agent

- selects an action $a_t$ in its current state $s_t$ based on past observations
- observes $s' \sim p(\cdot|s_t, a_t)$ and
  - ➜ generative model : select $s_{t+1}$
  - ➜ online model : set $s_{t+1} = s'$
- can decide to stop exploration → adaptive stopping time $\tau$
- if so, can output a guess for $\pi_\star$, $\widehat{\pi}$

**$(\varepsilon, \delta)$-PAC algorithm**

$$\mathbb{P}\left(\tau < \infty, \exists s \in \mathcal{S} : g^{\widehat{\pi}}(s) < g^\star - \varepsilon\right) \leq \delta.$$

# State-of-the-art

This problem has been mostly studies in the generative model setting.

**Lower bound** : [Wang et al., 2022] for any $(\varepsilon, \delta)$-PAC algorithm, there exists an MDP $\mathcal{M}$ such that

$$\mathbb{E}_{\mathcal{M}}[\tau_\delta] = \Omega\left(\frac{SAH}{\varepsilon^2}\log(1/\delta)\right)$$

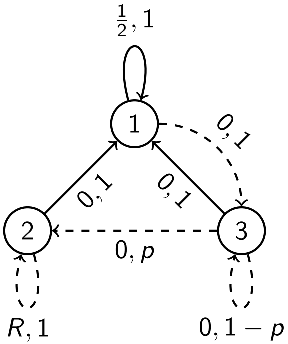
**Upper bound** : [Zurek and Chen, 2023] there exists an algorithm such that, for all MDPs,

$$\mathbb{E}_{\mathcal{M}}[\tau_\delta] = \widetilde{\mathcal{O}}\left(\frac{SAH}{\varepsilon^2}\log(1/\delta)\right)$$

... but it requires the knowledge of the optimal span bias, $H$

$$H \quad = \quad \max_s b^\star(s) - \min_s b^\star(s)$$

# Estimating $H$ is hard



- $R = 1/2 - \varepsilon \;\Rightarrow\; \pi^\star = \rightarrow \;\Rightarrow\; H = 1/2$
- $R = 1/2 + \varepsilon \;\Rightarrow\; \pi^\star = \dashrightarrow \;\Rightarrow\; H = (1/2 + \varepsilon)\frac{1+p}{p}$

# Estimating $H$ is hard

## Theorem

For any $\delta < \frac{1}{2e^4}$, $T > 0$, $\Delta$, there exists an MDP $\mathcal{M}$ with $H = 1/2$ such that any algorithm that computes a $\hat{H}$ satisfying $H \leq \hat{H} \leq H + \Delta$ with probability greater than $1 - \delta$ needs (in expectation) more than $T$ samples in $\mathcal{M}$.

# But estimating $D$ is easy

**Diameter (D) versus Optimal Bias Span (H) : $H \leq D$**

$$\begin{aligned} D &= \max_{s \neq s'} \min_{\pi : \mathcal{S} \to \mathcal{A}} \mathbb{E}^{\pi}[\min\{t > 0, s_t = s'\} | s_0 = s] \\ H &= \max_s b^{\star}(s) - \min_s b^{\star}(s) \end{aligned}$$

A two-stage algorithm :

- ➜ Use an algorithm from [Tarbouriech et al., 2021] that outputs $\widehat{D}$ such that $\mathbb{P}(D \leq \hat{D} \leq 4D) \geq 1 - \delta/2$ using $\widetilde{\mathcal{O}}(D^2 \log(1/\delta) + S)$ samples
- ➜ Use the algorithm of [Zurek and Chen, 2023] with $\widehat{D}$ as an upper bound on $H$, which uses $\widetilde{\mathcal{O}}\left(\frac{SA\widehat{D}}{\varepsilon^2} \log(1/\delta)\right)$ samples

# Diameter Free Exploration

**Entry : Accuracy** $\varepsilon \in (0,1)$**, confidence level** $\delta \in (0,1)$

- $\widehat{D} = \mathrm{DiameterEstimation}(\delta/2)$
- $\hat{\pi} = \mathrm{BPI}(\widehat{D}, \varepsilon, \delta/2)$
- **Return** $\hat{\pi}$

### Theorem

The algorithm above is $(\varepsilon, \delta)$-PAC and

$$\mathbb{P}\left(\tau \leq \widetilde{\mathcal{O}}\left(\left[\frac{SAD}{\varepsilon^2} + D^2 SA\right] \log(1/\delta) + D^2 S^2 A\right)\right) \geq 1 - \delta.$$

# Diameter Free Exploration

**Entry : Accuracy $\varepsilon \in (0,1)$, confidence level $\delta \in (0,1)$**

- $\widehat{D} = \mathrm{DiameterEstimation}(\delta/2)$
- $\hat{\pi} = \mathrm{BPI}(\widehat{D}, \varepsilon, \delta/2)$
- **Return** $\hat{\pi}$

**Theorem**

The algorithm above is $(\varepsilon, \delta)$-PAC and

$$\mathbb{P}\left(\tau \leq \widetilde{\mathcal{O}}\left(\left[\frac{SAD}{\varepsilon^2} + D^2SA\right]\log(1/\delta) + D^2S^2A\right)\right) \geq 1 - \delta.$$

→ optimal in the regime of small $\varepsilon$ as the lower bound of [Wang et al., 2022] is for an instance with $H = D$ !

# Without a generative model ?

Little is known in the online setting !

- we prove that $H$ is definitely not the right complexity measure there
- using an online diameter estimation procedure, we propose an algorithm with a $\widetilde{O}_{\delta}\left(\frac{SAD^2}{\varepsilon^2} + S^2AD^3\right)$ sample complexity

➜ ... but more adaptive algorithms are needed

# Theory and Practise in RL

- **Practise :** in many case, we train our RL algorithm and hope that the last policies perform well

➜ small sample complexity

- **Theory :** the optimal sample complexity of Best Policy Identification
  - is still un-resolved beyond the worse-case for $H > 1$
  - leads to complex algorithm
  - ... that share the overall algorithmic component of using some "intrisic rewards"

➜ in practise, intrisic rewards are learnt as well

e.g., [Burda et al., 2019, Badia et al., 2020]

Al Marjani, A., Tirinzoni, A., and Kaufmann, E. (2023a).
Active coverage for PAC reinforcement learning.
In *Proceedings of the 36th Conference On Learning Theory (COLT)*.

Al Marjani, A., Tirinzoni, A., and Kaufmann, E. (2023b).
Towards instance-optimality in online PAC reinforcement learning.
*arXiv :2311.05638*.

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010).
Best Arm Identification in Multi-armed Bandits.
In *Proceedings of the 23rd Conference on Learning Theory*.

Azar, M. G., Osband, I., and Munos, R. (2017a).
Minimax regret bounds for reinforcement learning.
In *Proceedings of the 34th International Conference on Machine Learning, (ICML)*.

Azar, M. G., Osband, I., and Munos, R. (2017b).
Minimax regret bounds for reinforcement learning.
In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 263–272.

Badia, A. P., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., Piot, B., Kapturowski, S., Tieleman, O., Arjovsky, M., Pritzel, A., Bolt, A., and Blundell, C. (2020).
Never give up : Learning directed exploration strategies.
In *ICLR*.

Bubeck, S., Munos, R., and Stoltz, G. (2011).
Pure Exploration in Finitely Armed and Continuous Armed Bandits.
*Theoretical Computer Science 412, 1832-1852*, 412 :1832–1852.

Burda, Y., Edwards, H., Storkey, A. J., and Klimov, O. (2019).
Exploration by random network distillation.
In *ICLR*.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021).
Episodic reinforcement learning in finite mdps : Minimax lower bounds revisited.
In *Algorithmic Learning Theory (ALT)*.

Even-Dar, E., Mannor, S., and Mansour, Y. (2006).
Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems.
*Journal of Machine Learning Research*, 7 :1079–1105.

Garivier, A. and Kaufmann, E. (2016).
Optimal best arm identification with fixed confidence.
In *Proceedings of the 29th Conference On Learning Theory*.

Jaksch, T., Ortner, R., and Auer, P. (2010).
Near-Optimal regret bounds for reinforcement learning.
*Journal of Machine Learning Research*, 11 :1563–1600.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018).
Is Q-learning provably efficient ?
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jonsson, A., Kaufmann, E., Ménard, P., Domingues, O. D., Leurent, E., and Valko, M. (2020).
Planning in markov decision processes with gap-dependent sample complexity.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kakade, S. (2003).
*On the Sample Complexity of Reinforcement Learning*.
PhD thesis, University College London.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012).

PAC subset selection in stochastic multi-armed bandits.
In *International Conference on Machine Learning (ICML)*.

Kaufmann, E. and Koolen, W. (2021).
Mixture martingales revisited with applications to sequential tests and confidence intervals.
*Journal of Machine Learning Research*, 22(246).

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. (2021).
Adaptive reward-free exploration.
In *Algorithmic Learning Theory (ALT)*.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. (2021).
Fast active learning for pure exploration in reinforcement learning.
In *International Conference on Machine Learning (ICML)*.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. (2021).
Sample complexity bounds for stochastic shortest path with a generative model.
In *32nd International conference on Algorithmic learning theory*, volume 132. PMLR.

Tirinzoni, A., Marjani, A. A., and Kaufmann, E. (2022).
Near instance-optimal PAC reinforcement learning for deterministic mdps.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wagenmaker, A. and Jamieson, K. (2022).
Instance-dependent near-optimal policy identification in linear mdps via online experiment design.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wagenmaker, A. J., Simchowitz, M., and Jamieson, K. (2022).
Beyond no regret : Instance-dependent PAC reinforcement learning.
In *Conference On Learning Theory (COLT)*.

Wang, J., Wang, M., and Yang, L. F. (2022).
Near sample-optimal reduction-based policy learning for average reward MDP.
`https://arxiv.org/abs/2212.00603`.

Zurek, M. and Chen, Y. (2023).
Span-based optimal sample complexity for average reward MDPs.
`https://arxiv.org/abs/2311.13469`.