

Exploration in MDPs

Chen-Yu Wei

- Review of what we talked about
 - Failure example
 - Missing exploration over states
 - Regret bound decomposition
-
- Revisiting UCB and uncertainty quantification again

We have addressed all 3 main challenges in online RL

Data + Function approximation

Generalization

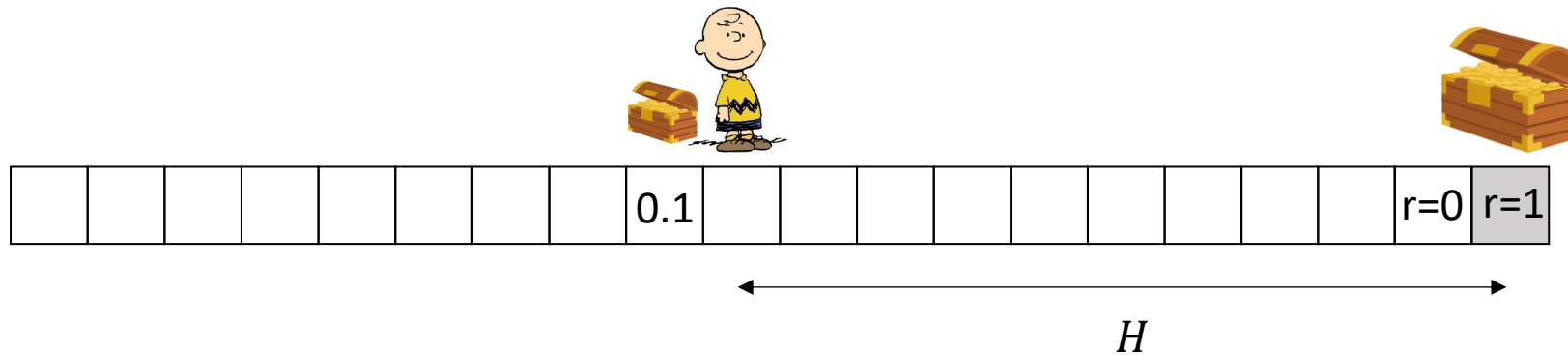
EG
BE
IGW
UCB
TS

Exploration

Credit
Assignment

VI
PI

We have addressed all 3 main challenges in online RL (?)



Environment:

- Fixed-horizon MDP with episode length H
- Initial state at 0
- A single rewarding state at state H
- Actions: Go LEFT or RIGHT

Suppose we perform DQN with ϵ -greedy with random initialization

⇒ On average, we need 2^H episodes to see the reward

(before that, we won't make any meaningful update and will just do random walk around state 0)

Regret Analysis for MDPs?

- We have done regret analysis for several bandit algorithms:
 - Regression oracle + (ϵ -greedy or inverse gap weighting)
 - UCB
 - EXP3
- We did not really establish regret bounds for MDPs
 - Partially – DQN under 2 assumptions: the data in replay buffer is exploratory and Bellman completeness
 - Not for policy-iteration-based algorithms

Regret Analysis for MDPs?

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

$$\textcircled{1} = \sum_{s,a} d_{\rho}^{\pi}(s) (\pi^*(a|s) - \pi(a|s)) \underbrace{Q^*(s,a)} = \sum_{s,a} d_{\rho}^{\pi}(s,a) (V^*(s) - Q^*(s,a))$$

For VI-based algorithm (approximating Q^*)

Approximating $Q^*(s,a)$ requires the replay buffer to cover **wide range of** state-actions.

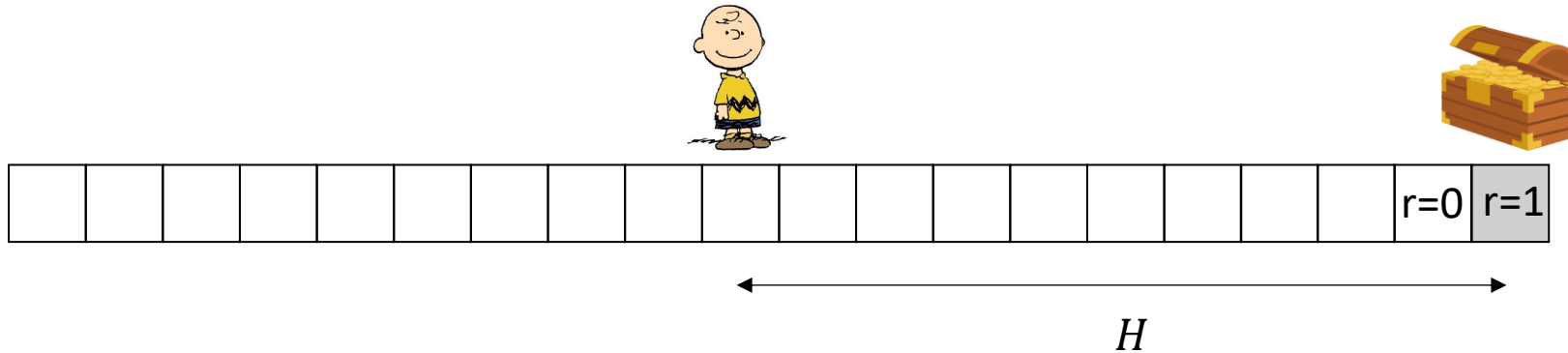
$$\textcircled{2} = \sum_{s,a} \boxed{d_{\rho}^{\pi^*}(s)} (\pi^*(a|s) - \pi(a|s)) \underbrace{Q^{\pi}(s,a)} = \sum_{s,a} \boxed{d_{\rho}^{\pi^*}(s,a)} (Q^{\pi}(s,a) - V^{\pi}(s))$$

For PI-based algorithm (approximating Q^{π})

Approximating $Q^{\pi}(s,a)$ only requires state-actions generated from current policy

But...

Regret Analysis for MDPs?



$$\sum_{s,a} d_{\rho}^{\pi}(s,a) (V^*(s) - Q^*(s,a))$$

$\exists s,a, \quad d_{\rho}^{\pi^*}(s,a) \text{ large} \quad Q^{\pi^*}(s,a) - V^{\pi^*}(s) \text{ large}$

$d_{\rho}^{\pi^*}(s,a)$

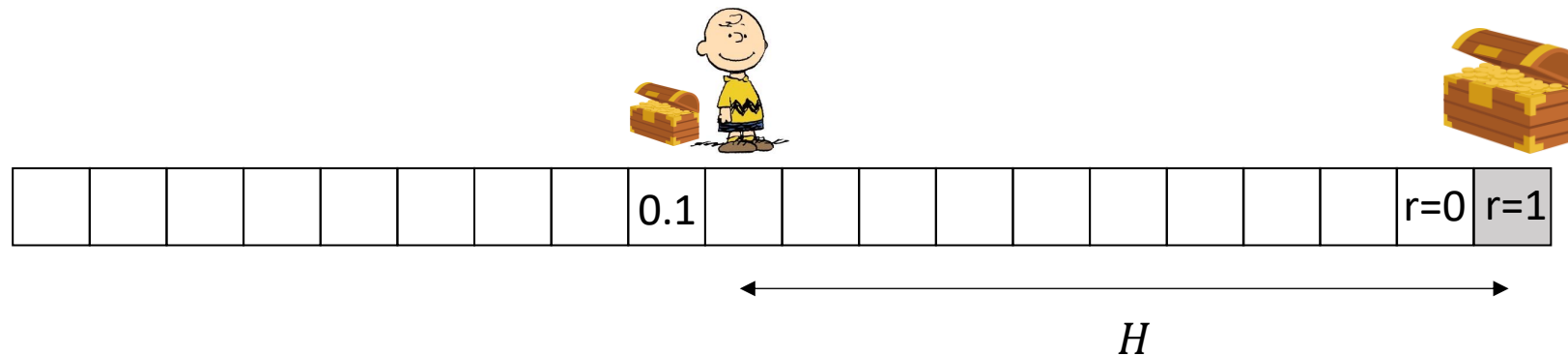
$$\sum_{s,a} d_{\rho}^{\pi^*}(s,a) (Q^{\pi}(s,a) - V^{\pi}(s)) \quad \text{large}$$

PI-based algorithm only tries to make $\sum_{s,a} d_{\rho}^{\pi^k}(s,a) (Q^{\pi}(s,a) - V^{\pi}(s))$ small.

It can only quickly find optimal policy when $d_{\rho}^{\pi^k} \approx d_{\rho}^{\pi^*}$

Insufficiency of algorithms we have discussed for MDPs

- Lack of **exploration over the state space** (we need **deep exploration**)
- This issue is particular critical if
 - Local reward does not provide any information (*sparse reward*)
 - Local reward provide misleading information



- Solution
 - Try to make the data (i.e., state-action) distribution close to d^{π^*}
 - Try to visit as many states as possible

Exploration Bonus (Optimism Principle)

- We have discussed this idea for action exploration – UCB.

Upper Confidence Bound

$$a_t = \operatorname{argmax}_a \hat{R}_t(a) + \sqrt{\frac{2 \log(2/\delta)}{N_t(a)}}$$

$\hat{R}_t(a)$ = the empirical mean of arm a up to time $t - 1$.

$N_t(a)$ = the number of times we draw arm a up to time $t - 1$.

Exploration Bonus (Optimism Principle)

$$a_t = \operatorname{argmax}_a \hat{R}_t(a) + \sqrt{\frac{2 \log(2/\delta)}{N_t(a)}}$$

$$\tilde{R}_t(a)$$

$$\sum_t \sqrt{\frac{1}{N_t(a_t)}} \leq \sqrt{AT}$$

$$R_{\text{regret}} = \sum_t (R(a^*) - R(a_t))$$

$$= \sum_t \underbrace{(\tilde{R}(a^*) - \tilde{R}(a_t))}_{\leq 0} + \sum_t \underbrace{(R(a^*) - \tilde{R}(a^*))}_{\leq 0} + \sum_t (\tilde{R}(a_t) - R(a_t))$$

① $\tilde{R}_t(a) + \overset{\text{bonus}}{\downarrow} b_t(a) \geq R(a)$

② $\sum_t b_t(a_t) \leq \text{sub-linear}(T)$

$$\lesssim \sqrt{\frac{1}{N_t(a_t)}}$$

Exploration Bonus for MDPs

UCB Value Iteration (UCBVI) *(finite state - action)*

For episode $1, 2, \dots, T$:

$$\tilde{Q}_{H+1}(s, a) = 0 \quad \forall s, a$$

For step $H, H-1, \dots, 1$:

$$\tilde{Q}_h(s, a) \triangleq \hat{R}(s, a) + \sum_{s'} \hat{P}(s'|s, a) \max_{a'} \tilde{Q}_{h+1}(s', a') + H \sqrt{\frac{2 \log(2/\delta)}{N_t(s, a)}} \quad \forall s, a$$

Receive $s_1 \sim \rho$

For step $1, 2, \dots, H$:

Take action $a_h = \operatorname{argmax}_a \tilde{Q}_h(s_h, a)$

Receive $r_h = R(s_h, a_h) + \text{noise}$, $s_{h+1} \sim P(\cdot | s_h, a_h)$

Exploration Bonus for MDPs

$$\tilde{Q}_h(s, a) \triangleq \hat{R}(s, a) + \sum_{s'} \hat{P}(s'|s, a) \max_{a'} \tilde{Q}_{h+1}(s', a') + H \sqrt{\frac{2 \log(2/\delta)}{N_t(s, a)}} \quad \forall s, a$$

$$\tilde{Q}_h(s, a) \geq Q_h^*(s, a) \quad \forall s, a, h \quad \text{w.h.p.}$$

$$\tilde{Q}_H(s, a) = \hat{R}(s, a) + H \sqrt{\frac{2 \log(2/\delta)}{N_t(s, a)}} \geq Q_H^*(s, a) = R(s, a)$$

For $h < H$

$$\begin{aligned} \tilde{Q}_h(s, a) &= \hat{R}(s, a) + \sum_{s'} \hat{P}(s'|s, a) \tilde{V}_{h+1}(s') \geq \hat{R}(s, a) + \sum_{s'} \hat{P}(s'|s, a) V_{h+1}^*(s') + H \sqrt{\frac{2 \log(2/\delta)}{N_t(s, a)}} \\ &\quad + H \sqrt{\frac{2 \log(2/\delta)}{N_t(s, a)}} \geq R(s, a) + \sum_{s'} P(s'|s, a) V_{h+1}^\pi(s') = Q_h^*(s, a) \end{aligned}$$

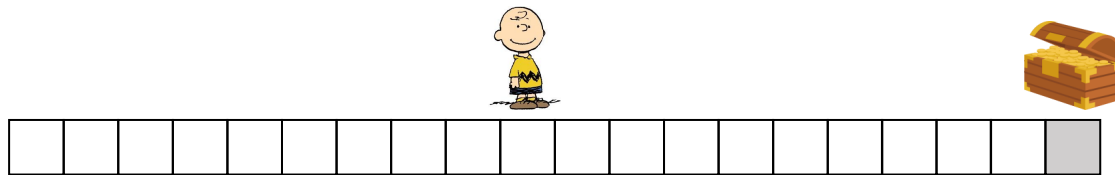
Exploration Bonus for MDPs

Theorem. Regret Bound of UCBVI

UCBVI ensures with high probability,

$$\text{Regret} = \sum_{t=1}^T (V^*(s_{t,1}) - V^{\pi_t}(s_{t,1})) \lesssim H\sqrt{SAT}.$$

$$\frac{1}{T} \sum_{t=1}^T (V^*(s_{t,1}) - V^{\pi_t}(s_{t,1})) \lesssim \frac{H\sqrt{SA}}{\sqrt{T}} \leq \varepsilon$$
$$\Rightarrow T \gtrsim \frac{H^2 SA}{\varepsilon^2}$$



Improving the required number of episodes from 2^H to $\text{poly}(H)$

Jaksch, Ortner, Auer. Near-Optimal Regret Bounds for Reinforcement Learning. 2010.

Azar, Osband, Munos. Minimax Regret Bounds for Reinforcement Learning. 2017.

Thompson Sampling (Posterior Sampling)

$$\mathcal{H}_t = (\underline{a_1}, \underline{r_1}, \underline{a_2}, \underline{r_2}, \dots, \underline{a_{t-1}}, \underline{r_{t-1}})$$

Bayesian interpretation:

Assume the reward mean $(\theta(1), \dots, \theta(A))$ is drawn from a Gaussian distribution (prior distribution).

Then the **posterior distribution** is

$$P(\theta(a) | \mathcal{H}_t) = \mathcal{N} \left(\hat{R}_t(a), \frac{1}{N_t(a)} \right)$$

we want to find $\arg\max_a \theta(a)$

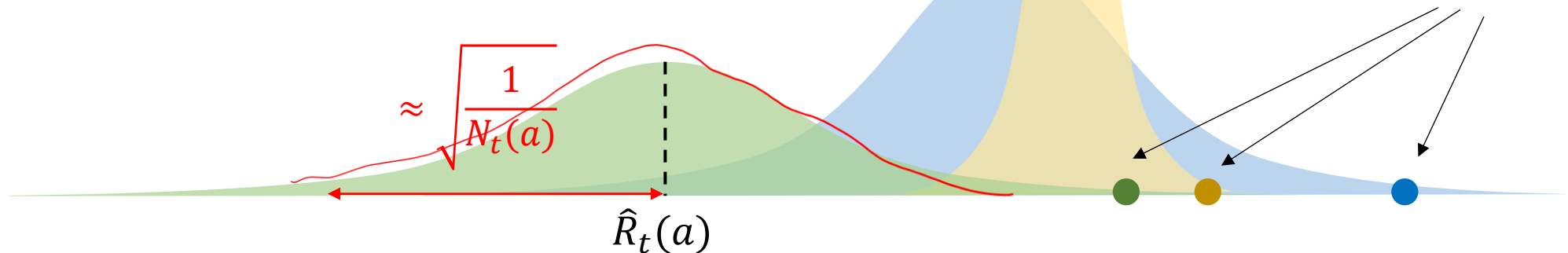
$$\text{UCB: } a_t \approx \arg\max_a \hat{R}_t(a) + c \sqrt{\frac{1}{N_t(a)}}$$

$$\text{TS: } a_t \approx \arg\max_a \hat{R}_t(a) + c \sqrt{\frac{1}{N_t(a)}} n_t(a) \text{ with } n_t(a) \sim \mathcal{N}(0,1)$$

a sample of $\theta(a)$

TS: sample $\theta \sim P(\cdot | \mathcal{H}_t)$
pick $a_t = \arg\max_a \theta(a)$

UCB estimators



Randomized Exploration for MDPs

Randomized Value Iteration

For episode $1, 2, \dots, T$:

$$\tilde{V}_{H+1}(s) = 0$$

For step $H, H - 1, \dots, 1$:

$$\tilde{Q}_h(s, a) \triangleq \hat{R}(s, a) + \sum_{s'} \hat{P}(s'|s, a) \max_{a'} \tilde{Q}_{h+1}(s', a') + H \sqrt{\frac{2 \log(2/\delta)}{N_t(s, a)}} \underbrace{n_t(s, a)}_{\sim \mathcal{N}(0,1)}$$

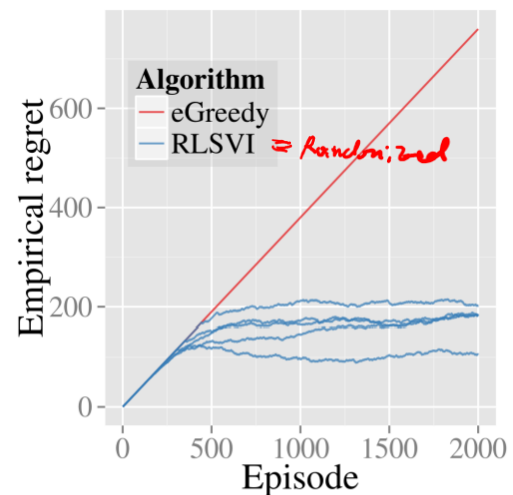
Receive $s_1 \sim \rho$

For step $1, 2, \dots, H$:

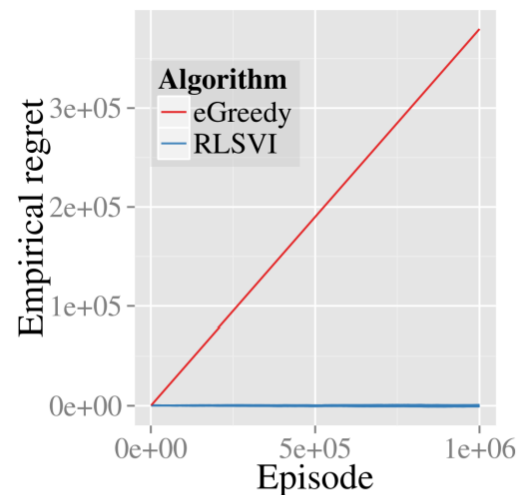
Take action $a_h = \operatorname{argmax}_a \tilde{Q}_h(s_h, a)$

Receive $r_h = R(s_h, a_h) + \text{noise}$, $s_{h+1} \sim P(\cdot | s_h, a_h)$

Randomized Exploration for MDPs



(a) First 2000 episodes



(b) First 10^6 episodes

Figure 2. Efficient exploration on a 50-chain

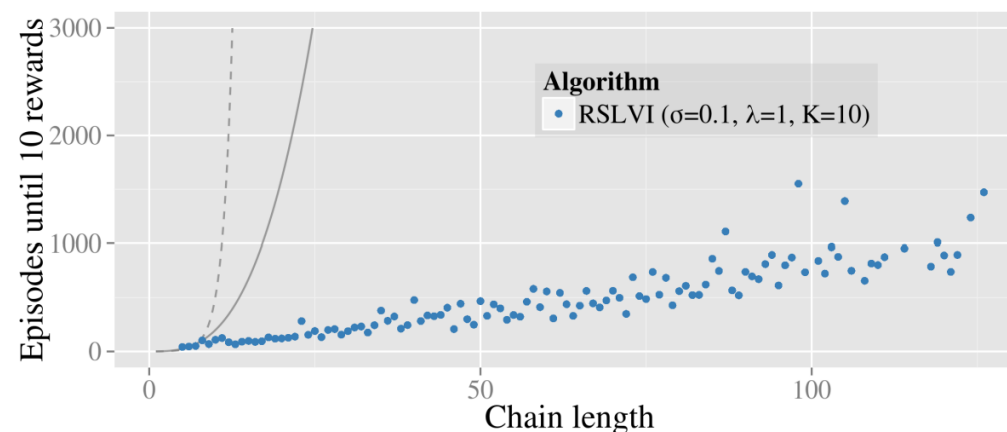
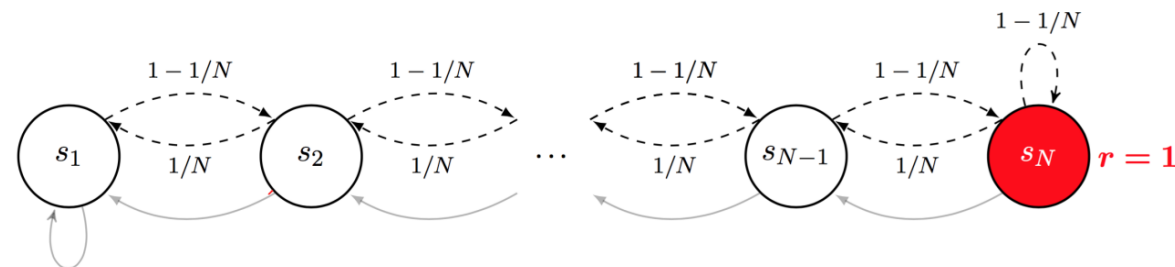


Figure 3. RLSVI learning time against chain length.

Common Approaches of Exploration

- Optimistic-Value Exploration
 - Upper Confidence Bound
- Randomized-Value Exploration
 - Thompson Sampling (Posterior Sampling)
- Information-Directed Exploration

Information-Directed Exploration (1/3)

Another Bayesian approach – like Thompson sampling.

Assume the parameter of the world (e.g., the mean reward of the arms) is drawn as $\theta \sim P_{\text{prior}}$

After observing history $\mathcal{H}_t = (a_1, r_1, a_2, r_2, \dots, a_{t-1}, r_{t-1})$, we can calculate the posterior distribution of θ :

$$P(\theta|\mathcal{H}_t) = \frac{P(\mathcal{H}_t, \theta)}{P(\mathcal{H}_t)} = \frac{P(\mathcal{H}_t|\theta)P_{\text{prior}}(\theta)}{P(\mathcal{H}_t)} \propto P(\mathcal{H}_t|\theta)P_{\text{prior}}(\theta)$$

Key question: Based on the posterior estimation of the world $P(\theta|\mathcal{H}_t)$, what action should we pick next?

Information-Directed Exploration (2/3)

Thompson Sampling: Sample $\theta_t \sim P(\cdot | \mathcal{H}_t)$ and choose $a_t = a^*(\theta_t) = \operatorname{argmax}_a \theta_t(a)$

Equivalently, execute $\pi(a) = \operatorname{argmax}_{\pi} \mathbb{E}_{\theta \sim P(\cdot | \mathcal{H}_t)} [\mathbb{I}\{a^*(\theta) = a\}]$

The optimal action
in the world of θ_t



Information-directed Sampling: Select an arm that tradeoffs **regret** and **information gain**

$$\text{Regret}_{\theta}(\pi) = \max_{a^*} \theta(a^*) - \theta(\pi)$$

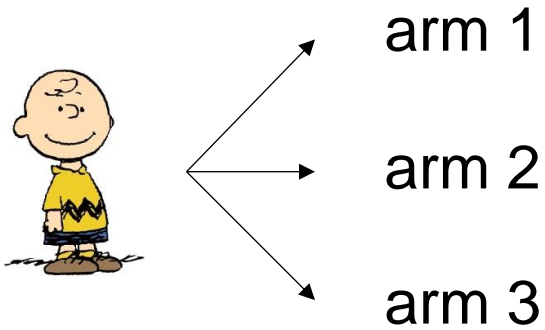
$$\text{InfoGain}_{\theta}(\pi) = \mathbb{E}_{r \sim \theta(\pi)} [\text{KL}(P(\cdot | \mathcal{H}_t, \pi, r), P(\cdot | \mathcal{H}_t))]$$

How much will the posterior change after
obtaining a new sample from π ?

$$\text{Execute } \pi = \operatorname{argmin}_{\pi} \mathbb{E}_{\theta \sim P(\cdot | \mathcal{H}_t)} [\text{Regret}_{\theta}(\pi) - \lambda \text{InfoGain}_{\theta}(\pi)]$$

Information-Directed Exploration (3/3)

When is information-directed exploration better than optimistic / posterior exploration?



Suppose we know there are two possible worlds, where the three arms follow $\{\text{Bernoulli}(0.5), \text{Bernoulli}(0.6), 0.1\}$ or $\{\text{Bernoulli}(0.6), \text{Bernoulli}(0.5), 0\}$

⇒ Although we know arm 3 is definitely not the best arm, we still want to sample it (once), so we can easily tell which world we're in.

Next

We will see how to generalize

- Optimistic-Value Exploration
- Posterior-Value Exploration
- Information-Directed Exploration

to large state space by incorporating **function approximation**