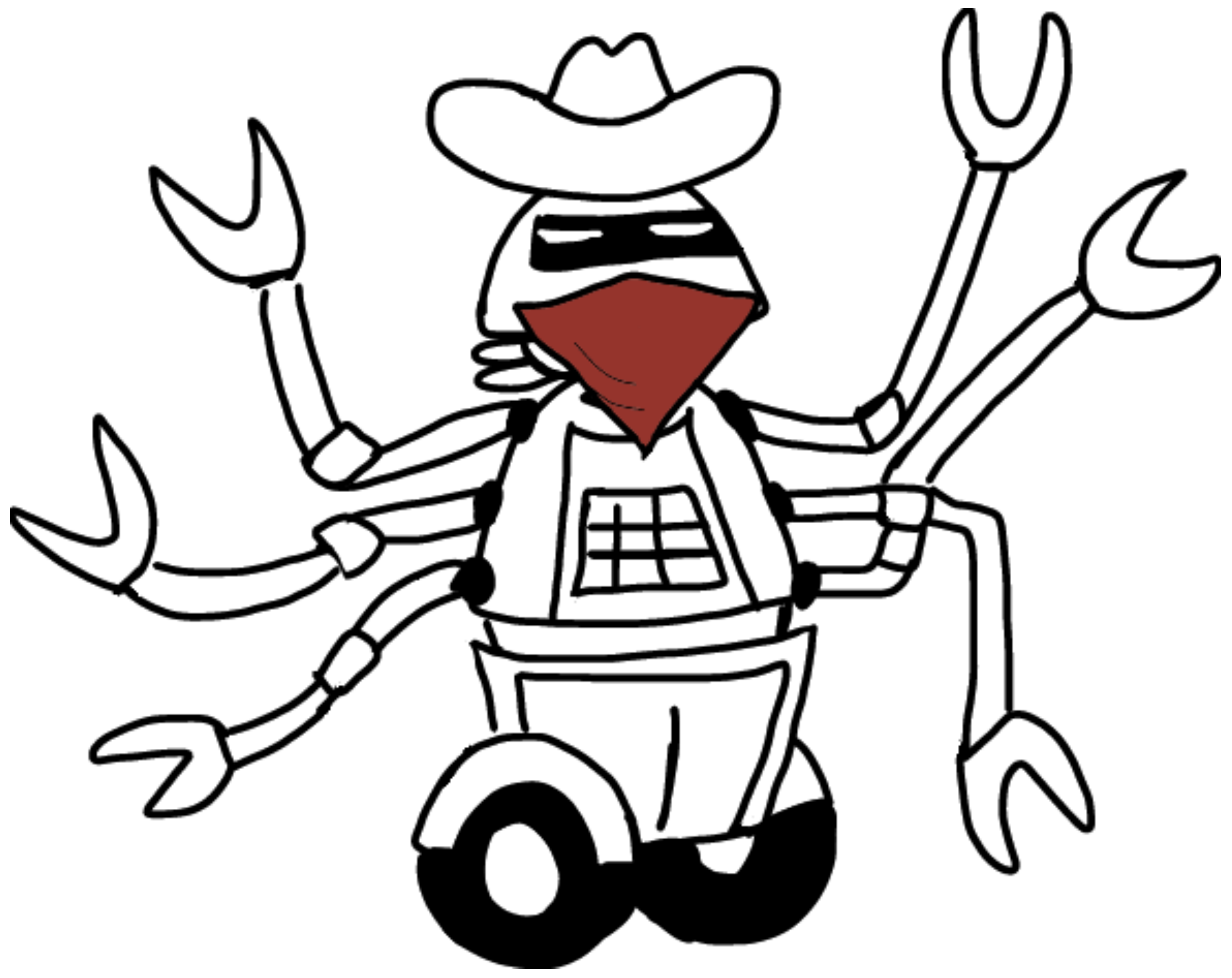
Reinforcement Learning

HSE, autumn - winter 2022

Lecture 7: Multi-armed Bandits



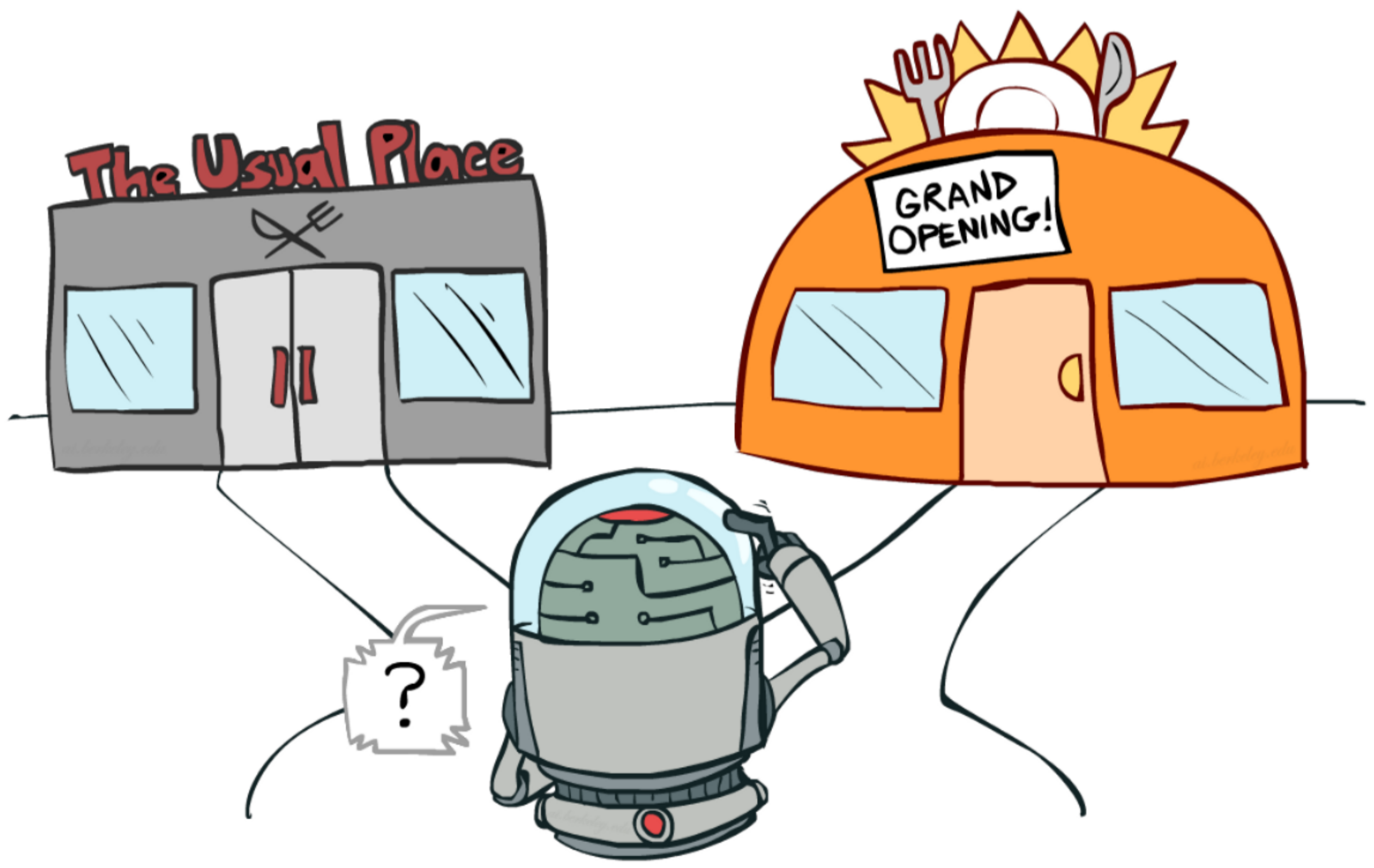
Sergei Laktionov
slaktionov@hse.ru
LinkedIn

Background

1. Practical RL course by YSDA, week 5
2. Sutton & Barto, Chapter 2
3. DeepMind course, Lecture 2
4. David Silver Course, Lecture 9

Exploration vs Exploitation Dilemma

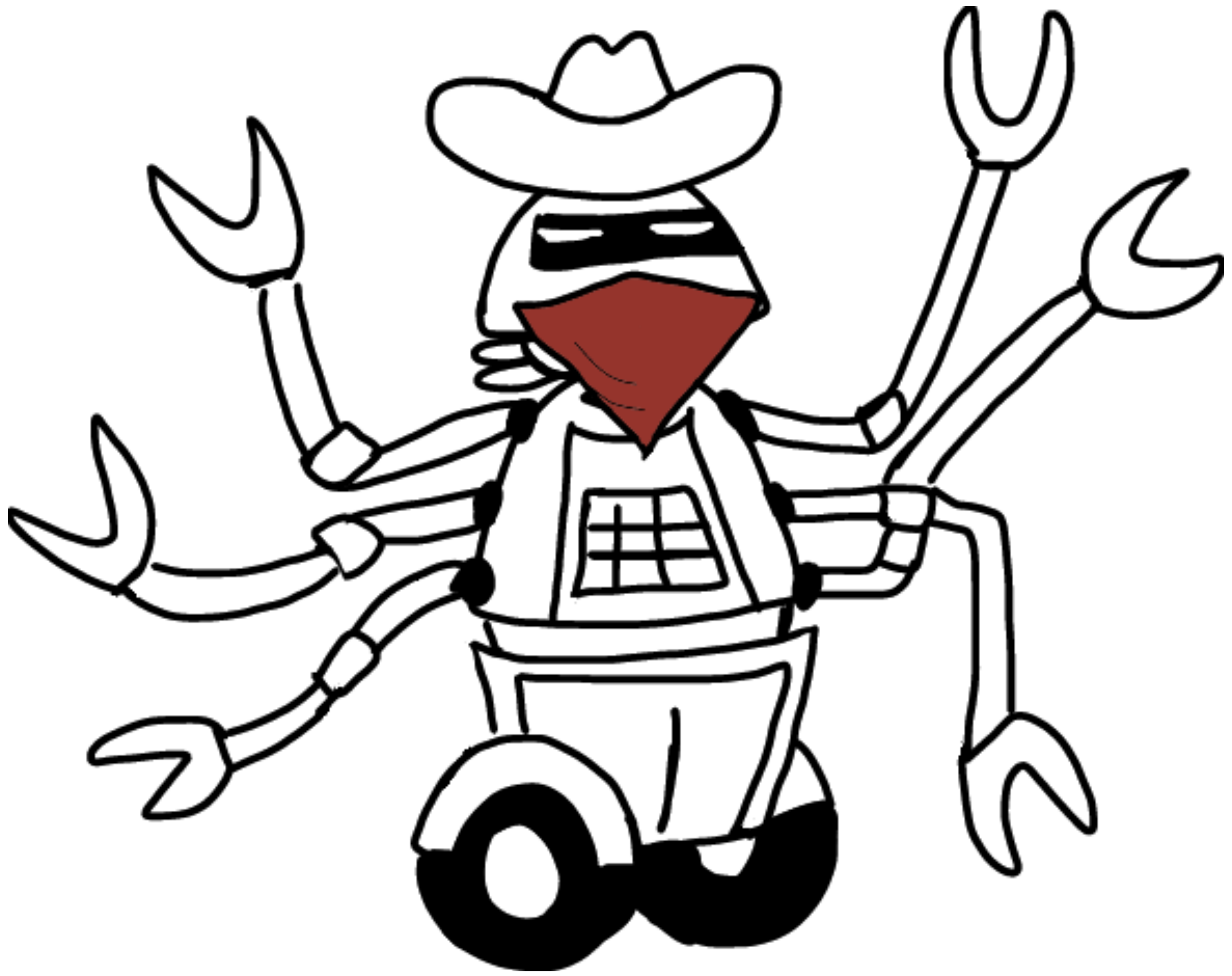
- Online decision-making involves a fundamental choice:
  - Exploitation Make the best decision given current information
  - Exploration Gather more information
- The best long-term strategy may involve short-term sacrifices
- Gather enough information to make the best overall decisions



Source

Multi-armed Bandit Problem Statement

Assume that the episode ends after the first step so we have only one state in the environment. An agent is facing repeatedly with a choice among k different actions.



Source

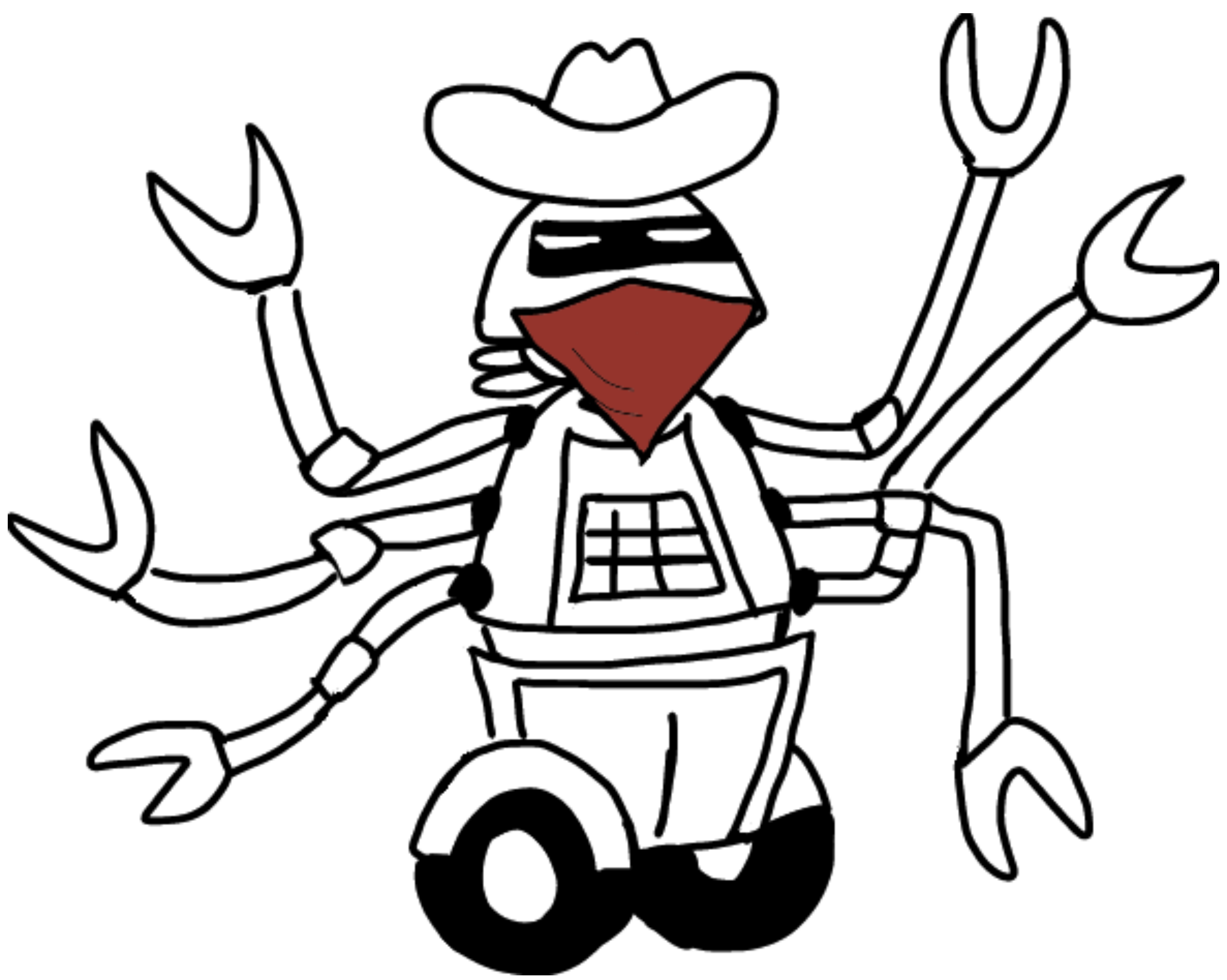
Multi-armed Bandit Problem Statement

Assume that the episode ends after the first step so we have only one state in the environment. An agent is facing repeatedly with a choice among k different actions.

A multi-armed bandit is a tuple $\langle \mathcal{R}, \mathcal{A} \rangle$ s.t. :

- $\{\mathcal{R}_a \,|\, a \in \mathcal{A}\}$ set of reward distributions;
- On each step t an agent chooses A_t and get reward $R_t \sim \mathcal{R}_{A_t}$

The agent's goal is to maximise $\mathbb{E}_{p(r|a)}[\sum_{t=1}^{T} R_t]$ by choosing an action on each step.



Source

Multi-armed Bandit Problem Statement

Exploration: find the best action which maximises expected reward

Action value function: $Q(a) = \mathbb{E}[R_t | A_t = a]$

Optimal value: $V* = \max_a Q(a)$

Regret: $\mathbb{E}_\pi[V* - Q(a)] \geq 0$

Multi-armed Bandit Problem Statement

Exploration: find the best action which maximises expected reward

Action value function: $Q(a) = \mathbb{E}[R_t | A_t = a]$

Optimal value: $V* = \max_a Q(a)$

Regret: $\mathbb{E}_\pi[V* - Q(a)] \geq 0$

Total Regret: $\mathbb{E}_\pi \sum_{t=1}^{T} [V* - Q(a_t)] \to \min_\pi$

Note: as we don't have states policy π is just a rule of making decision on each step.

Multi-armed Bandit Problem Statement

Exploration: find the best action which maximises expected reward

Action value function: $Q(a) = \mathbb{E}[R_t \,|\, A_t = a]$

Optimal value: $V* = \max_a Q(a)$

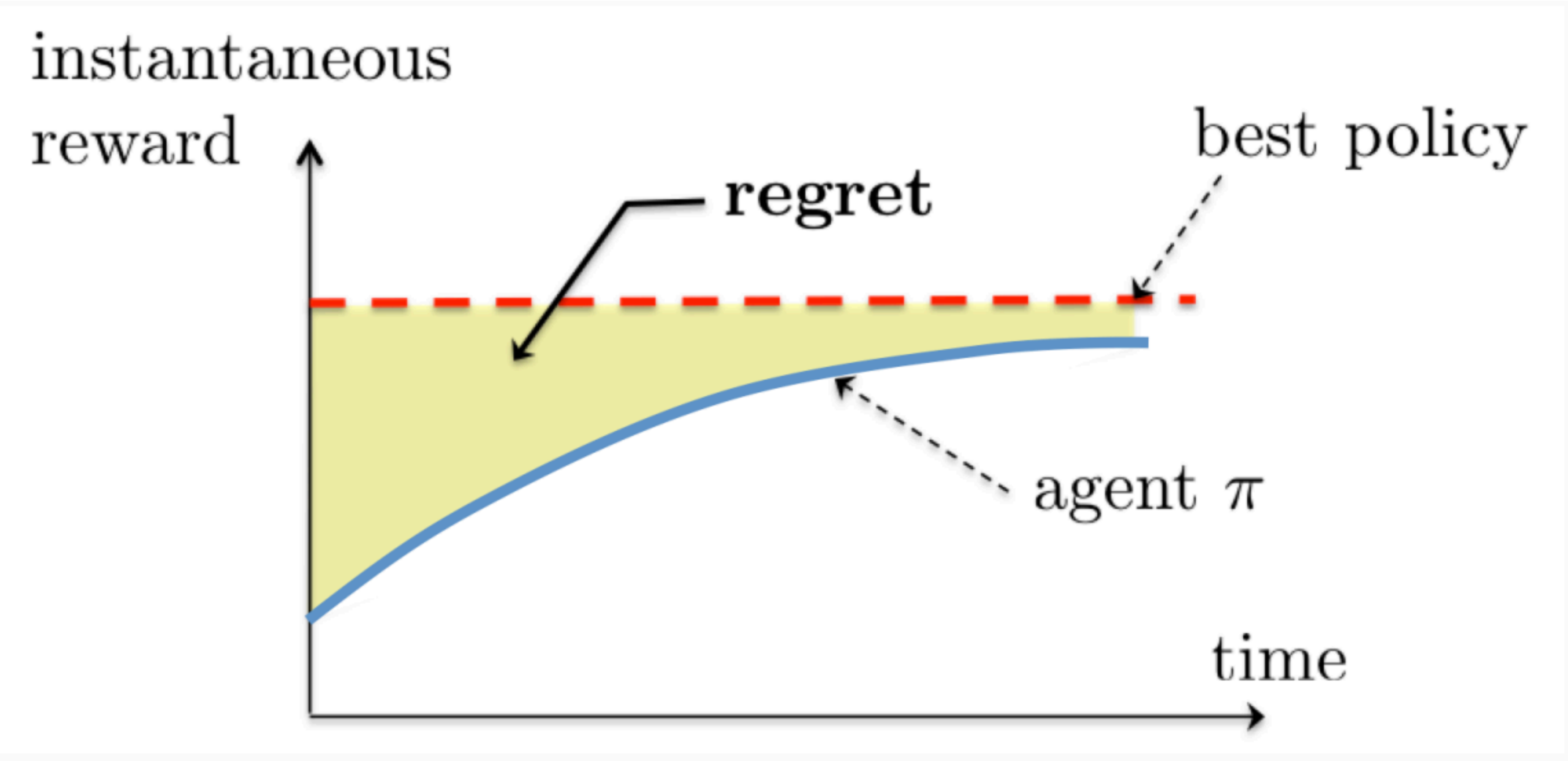
Regret: $\mathbb{E}_\pi[V* - Q(a)] \geq 0$

Total Regret: $\mathbb{E}_\pi \sum_{t=1}^{T} [V* - Q(a_t)] \to \min_\pi \iff \mathbb{E}_{p(r|a)}[\sum_{t=1}^{T} R_t] \to \max_\pi$

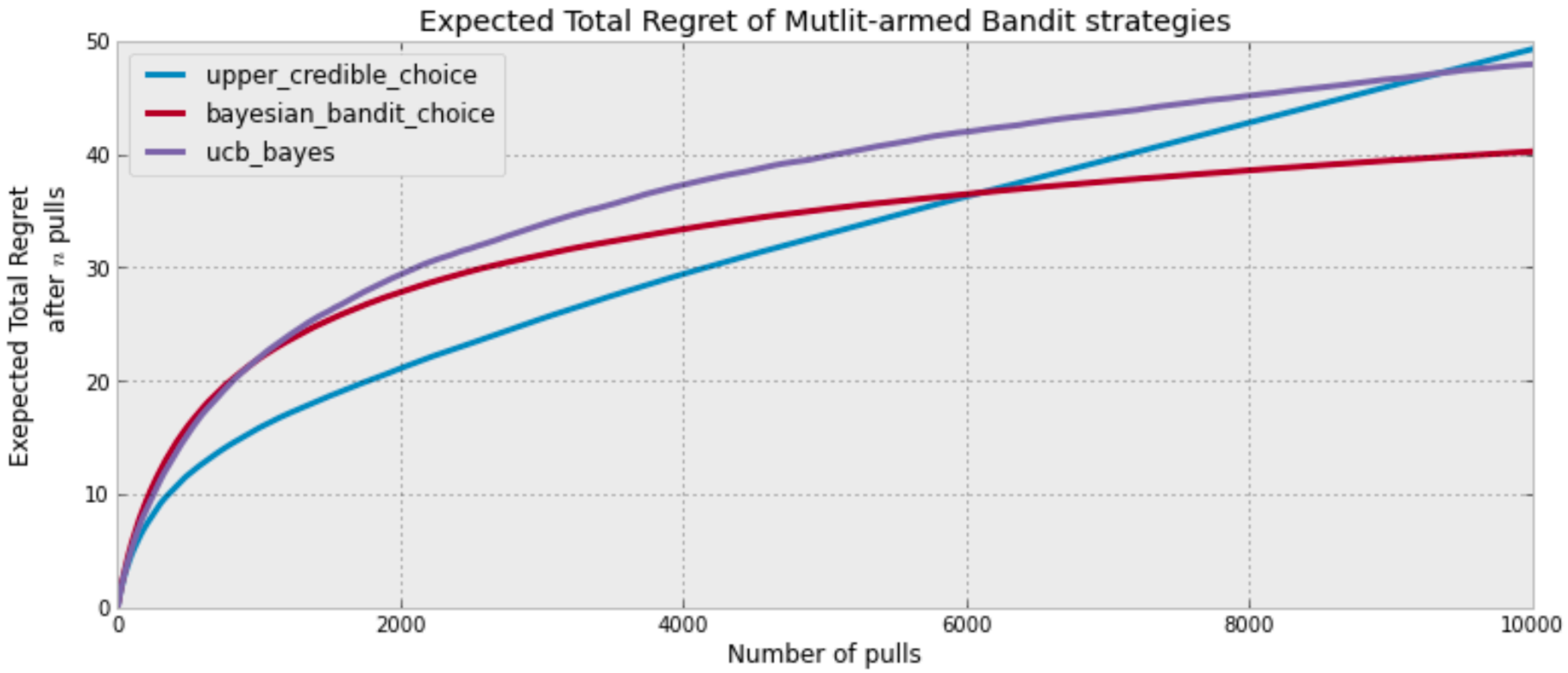
Note: as we don't have states policy π is just a rule of making decision on each step.

Regret Minimisation

$$\textbf{Total Regret: } \mathbb{E}_{\pi} \sum_{t=1}^{T} [V^* - Q(a_t)] \rightarrow \min_{\pi} \iff \mathbb{E}_{p(r|a)}[\sum_{t=1}^{T} R_t] \rightarrow \max_{\pi}$$



Source



Source

Action Values

$$Q_t(a) = \frac{\sum_{n=1}^{t} \mathbb{I}(A_n = a) r_n}{\sum_{n=1}^{t} \mathbb{I}(A_n = a)} = \frac{\sum_{n=1}^{t} \mathbb{I}(A_n = a) r_n}{N_t(a)} \iff$$

Action Values

$$Q_t(a) = \frac{\sum_{n=1}^{t} \mathbb{I}(A_n = a) r_n}{\sum_{n=1}^{t} \mathbb{I}(A_n = a)} = \frac{\sum_{n=1}^{t} \mathbb{I}(A_n = a) r_n}{N_t(a)} \iff \begin{gathered} Q_t(A_t) = Q_{t-1}(A_t) + \alpha_t [r_t - Q_{t-1}(A_t)] \\ \alpha_t = \frac{1}{N_t}, N_t(A_t) = N_{t-1}(A_t) + 1 \end{gathered}$$

ε-greedy Policy

$$\pi_t(a) = \begin{cases} (1-\varepsilon)+\dfrac{\varepsilon}{|\mathscr{A}|}, \text{ if } Q_t(a) = max_{a'}Q_t(a') \\ \dfrac{\varepsilon}{|\mathscr{A}|}, \text{ otherwise} \end{cases}$$

- Greedy policy can stuck in a suboptimal action forever
- ε-greedy continues to explore

ε-greedy Policy

$$\pi_t(a) = \begin{cases} (1-\varepsilon)+\dfrac{\varepsilon}{|\mathscr{A}|}, \text{ if } Q_t(a) = max_{a'}Q_t(a') \\ \dfrac{\varepsilon}{|\mathscr{A}|}, \text{ otherwise} \end{cases}$$

- Greedy policy can stuck on a suboptimal action forever
- ε-greedy continues to explore

ε-greedy policy has linear regret

Gradient Policy

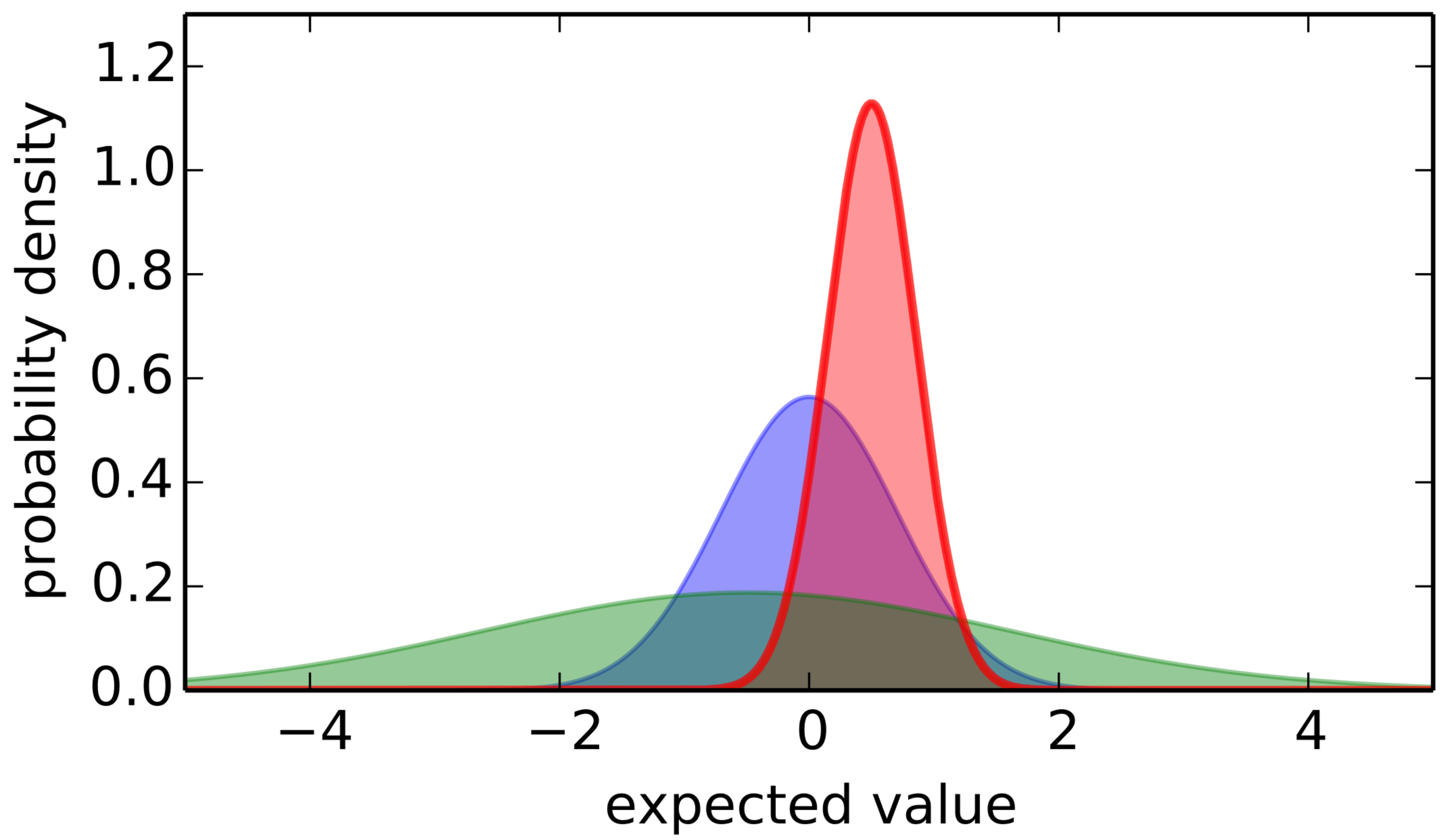
We can learn softmax policy using REINFORCE via gradient ascent, but there is no still guarantees for convergence to a global optimum.

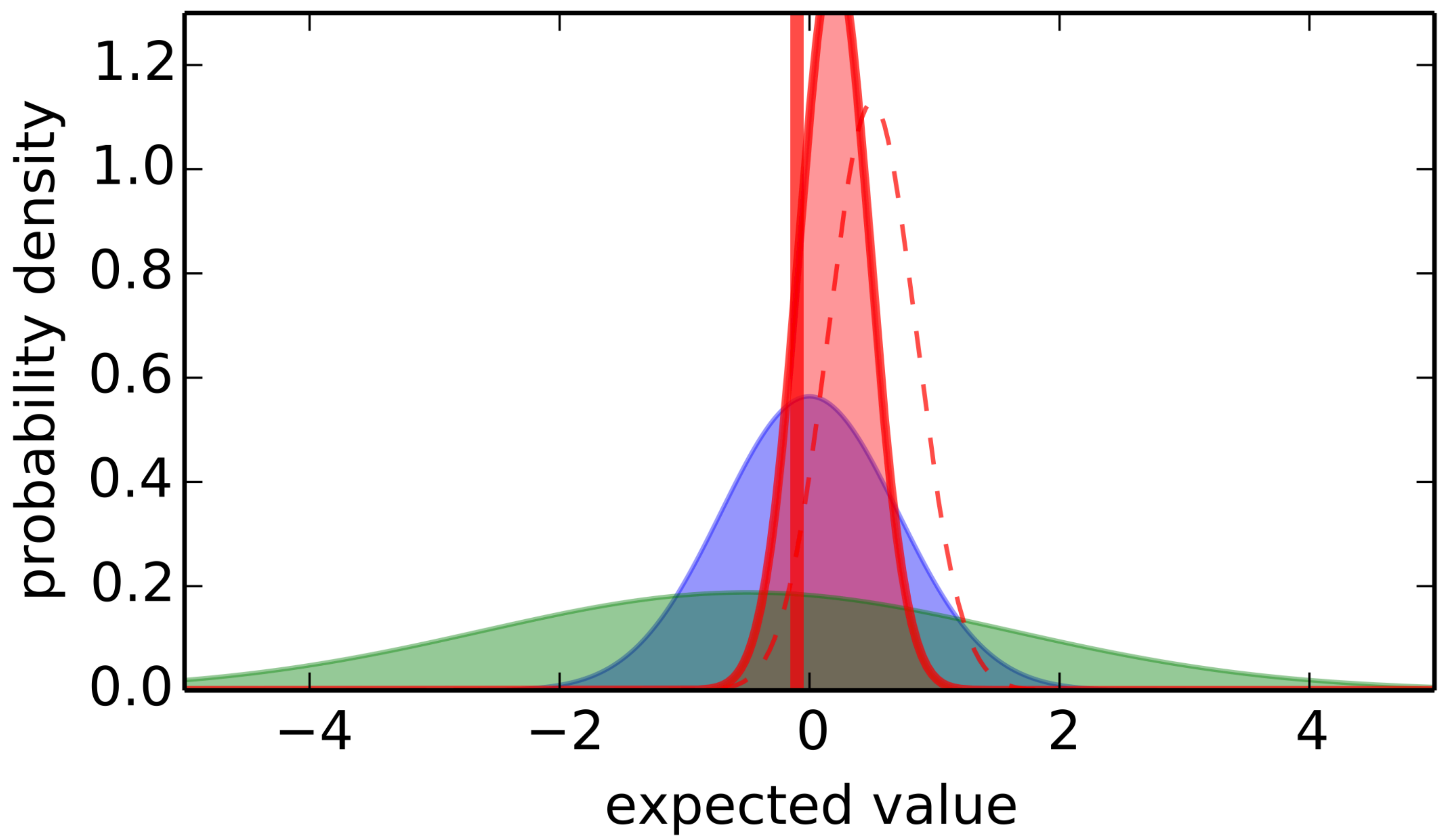
Regret Lower Bound

Theorem:

$$\mathbb{E}_\pi \sum_{t=1}^{T} [V^* - Q(a_t)] \geq \log T \sum_{a | V^* > Q(a)} \frac{V^* - Q(a)}{KL(\mathscr{R}_a || \mathscr{R}_{a^*})}$$

Optimism in the Face of Uncertainty

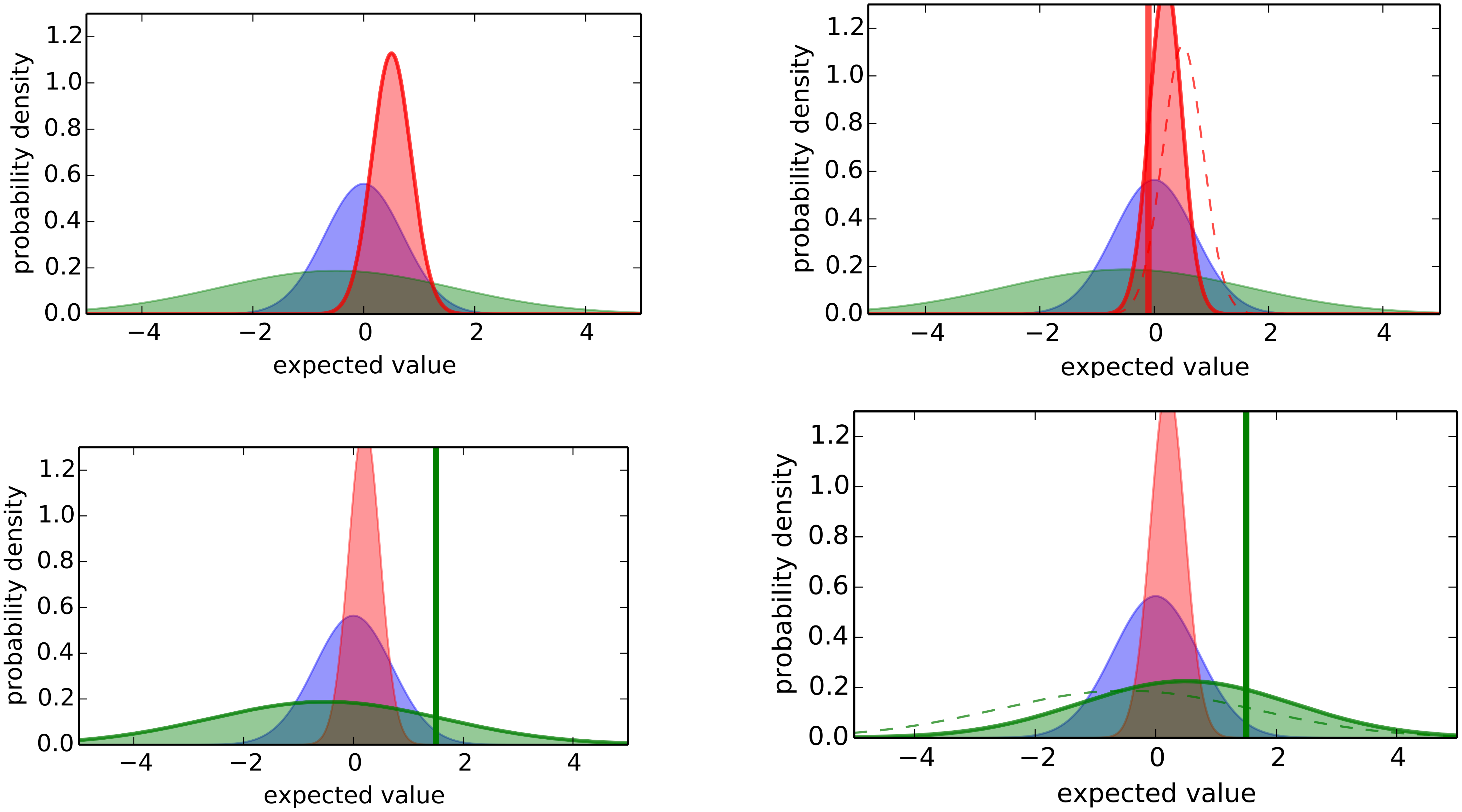




- Which action should we pick?
- The more uncertain we are about an action-value the more important it is to explore that action
- It could turn out to be the best action

Source

Optimism in the Face of Uncertainty



Source

Upper Confidence Bound

- Estimate an upper confidence $U_t(a)$ for each action value, such that $Q(a) \leq Q_t(a) + U_t(a)$ with high probability.
- This depends on the number of times $N(a)$ has been selected
  - Small $N(a) \Rightarrow$ large $U_t(a)$ (estimated value is uncertain)
  - Large $N(a) \Rightarrow$ small $U_t(a)$ (estimated value is accurate)
- Select action maximising upper confidence bound (UCB):
  $a_t = argmax_{a \in A}[Q_t(a) + U_t(a)]$

Optimality of UCB

Hoeffding's Inequality:

Let $X_1, \ldots, X_t$ be i.i.d. random variables in $[0,1]$ with true mean μ, and let $\bar{X}_t$ be the sample mean. Then $\mathbb{P}(\bar{X}_t + u \leq \mu) \leq e^{-2tu^2}$.

$$\mathbb{P}(Q_t(a) + U_t(a) \leq Q(a)) \leq e^{-2N_t(a)U_t(a)^2}$$

Optimality of UCB

Hoeffding's Inequality:

Let $X_1, \ldots, X_t$ be i.i.d. random variables in $[0,1]$ with true mean μ, and let $\bar{X}_t$ be the sample mean. Then $\mathbb{P}(\bar{X}_t + u \leq \mu) \leq e^{-2tu^2}$.

$$\mathbb{P}(Q_t(a) + U_t(a) \leq Q(a)) \leq e^{-2N_t(a)U_t(a)^2}$$

If $U_t(a) = \sqrt{\frac{-\log p}{2N_t(a)}}$ then $e^{-2N_t(a)U_t(a)^2} = p$

Reduce p as we observe more rewards, e.g. $p = \frac{1}{t}$

Optimality of UCB

Hoeffding's Inequality:

Let $X_1, \ldots, X_t$ be i.i.d. random variables in $[0,1]$ with true mean μ, and let $\bar{X}_t$ be the sample mean. Then $\mathbb{P}(\bar{X}_t + u \leq \mu) \leq e^{-2tu^2}$.

$$\mathbb{P}(Q_t(a) + U_t(a) \leq Q(a)) \leq e^{-2N_t(a)U_t(a)^2}$$

If $U_t(a) = \sqrt{\frac{-\log p}{2N_t(a)}}$ then $e^{-2N_t(a)U_t(a)^2} = p$

$$U_t(a) = \sqrt{\frac{\log t}{2N_t(a)}}$$

Reduce p as we observe more rewards, e.g. $p = \frac{1}{t}$

UCB

- Select action maximising upper confidence bound (UCB):

$$a_t = argmax_{a \in A}[Q_t(a) + c\sqrt{\frac{\log t}{2N_t(a)}}]$$

- Theorem: if $c = \sqrt{2}$ then UCB achieves logarithmic expected total regret

Bayesian Approach

- We could adopt Bayesian interpretation and model distributions over values $\mathscr{R}_a \approx p(r \mid \theta_a)$ and use model-based approach
- E.g., θ_a could contain the means and variances of Gaussian belief distributions
- Allows us to inject rich prior knowledge θ_a^0
- We can then use posterior belief to guide exploration

Probability Matching

- Probability matching selects action a according to probability that a is the optimal action
- $\pi(a \mid h_t) = \mathbb{P}[Q(a) > Q(a'), a \neq a' \mid h_t], h_t = \{a_1, r_1, \ldots, a_{t-1}, r_{t-1}\}$ is history;
- Probability matching is optimistic in the face of uncertainty: Uncertain actions have higher probability of being max
- Can be difficult to compute analytically from posterior

Thompson Sampling

- Thompson sampling implements probability matching:

$$\pi(a \mid h_t) = \mathbb{P}[Q(a) > Q(a'), a \neq a' \mid h_t] = \mathbb{E}_{\mathscr{R}|h_t}[\mathbb{I}[a = argmax_{a'} Q(a')]]$$

- Use Bayes law to compute posterior distribution $p[\mathscr{R} \mid h_t]$
- Sample a reward distribution $\mathscr{R}$ from posterior
- Compute action-value function $Q(a) = \mathbb{E}[\mathscr{R}_a]$
- Select action maximising value on sample: $a = argmax_{a'} Q(a')$
- Thompson sampling achieves logarithmic lower bound!

.

Thompson Sampling

- Priors $p(\theta_a), a \in \mathscr{A}$
- $p(\theta_a) \leftarrow p(\theta_a \mid r_t) \propto p(r_t \mid \theta_a)p(\theta_a)$ is a bayesian update
- We can choose an action with maximal expected reward under the known distributions: $a_{t+1} = argmax_a \mathbb{E}_{\theta_a \sim p(\theta_a)} \mathbb{E}_{p(r|\theta_a)} r$
- However there is a probability that the chosen action will be suboptimal: $\mathbb{E}_{p(r|\theta_b)} r > \mathbb{E}_{p(r|\theta_a)} r$
- Let's choose action with the probability of being optimal: $\pi(a) = \mathbb{P}(\mathbb{E}_{p(r|\theta_a)} r = \max_b \mathbb{E}_{p(r|\theta_b)} r)$
- We only have to sample $\theta_a \sim p(\theta_a), a \in \mathscr{A}$ and choose action with the maximal expected reward under the θ_a

Thank you for your attention!