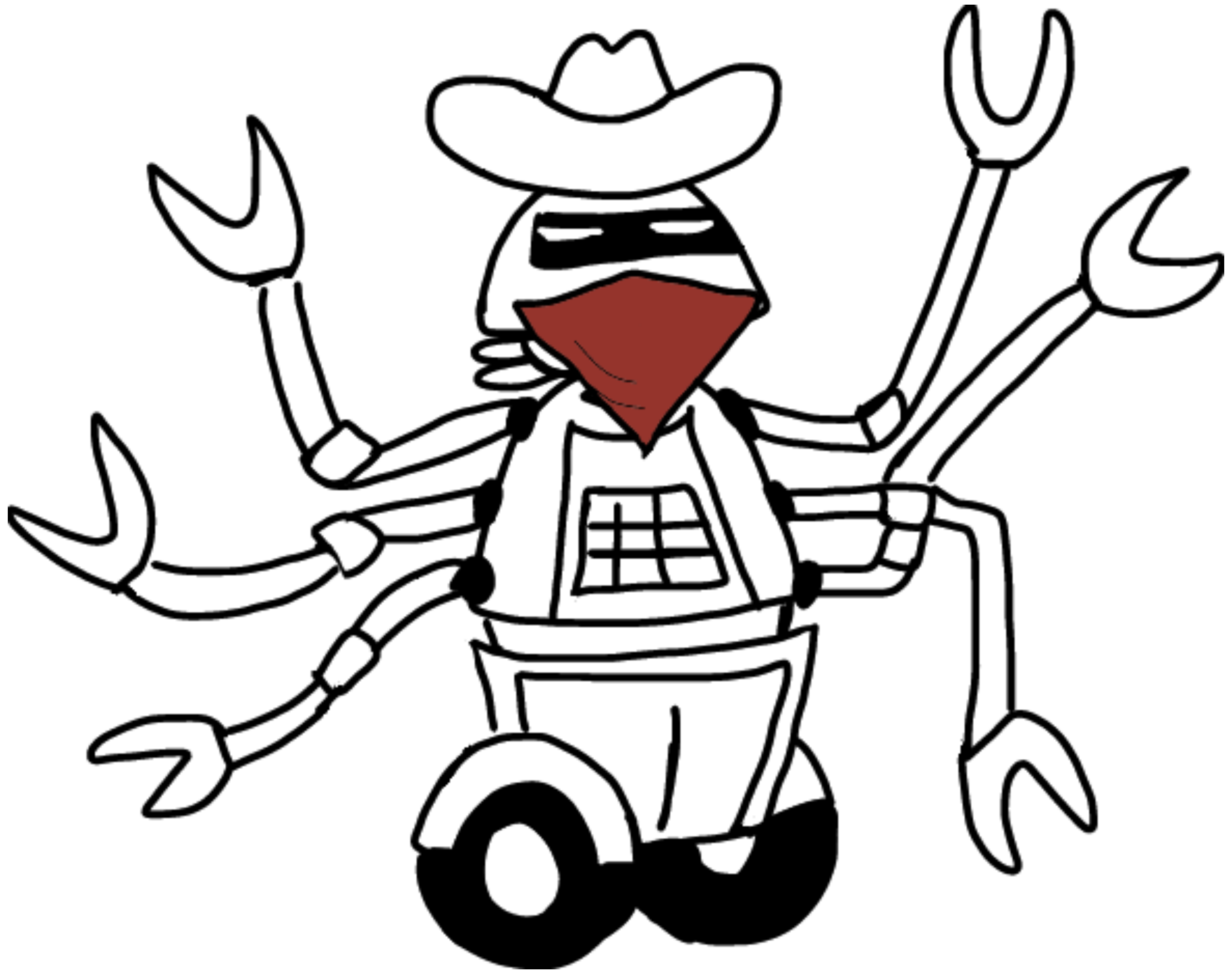
Reinforcement Learning

HSE, winter - spring 2024

Lecture 8: Multi-armed Bandits



Sergei Laktionov
slaktionov@hse.ru
LinkedIn

Multi-armed Bandit

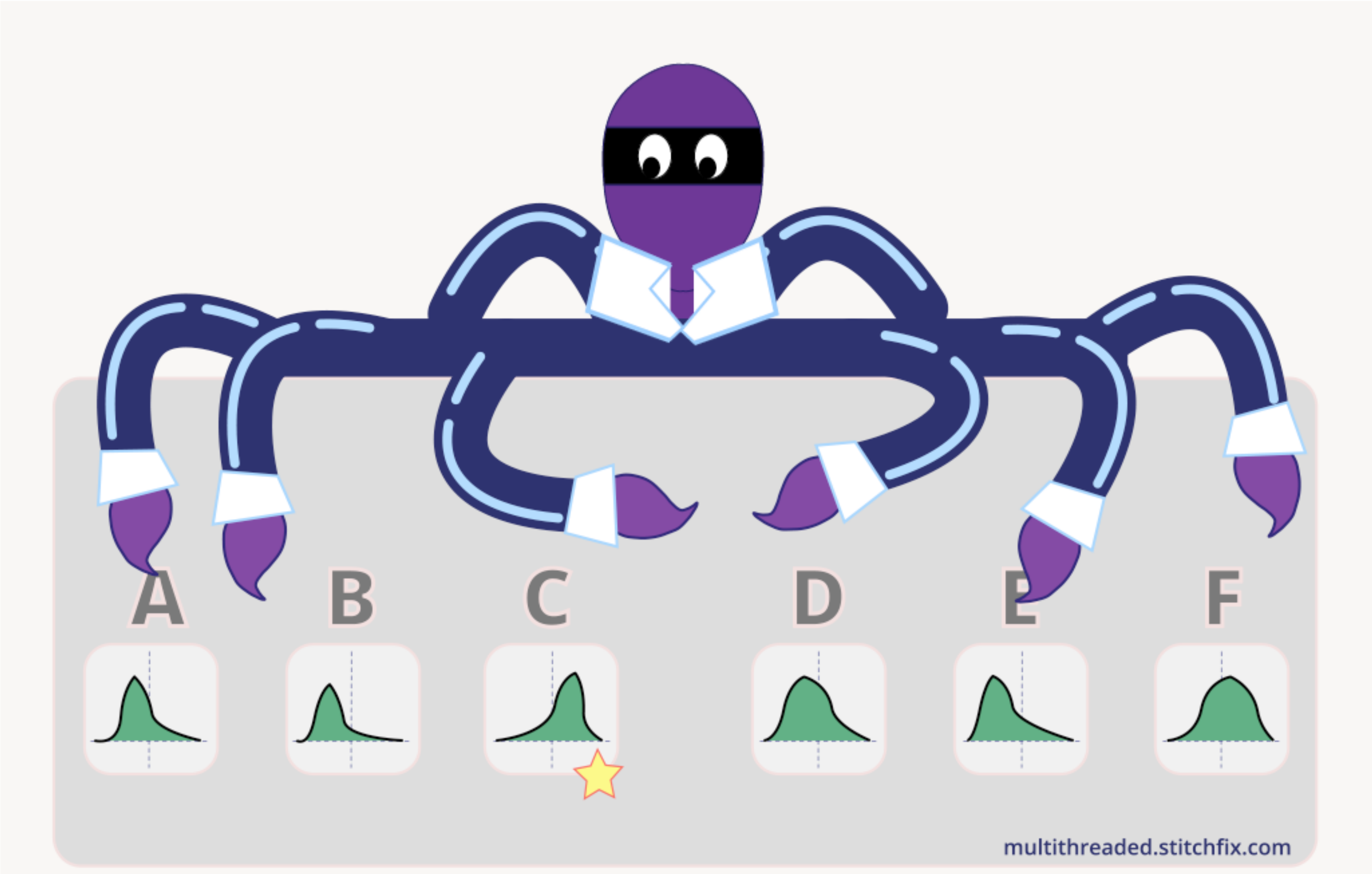
- The episode ends after the first step so we have only one state in the environment.

- An agent is facing repeatedly with a choice among K different actions.

Multi-armed Bandit

- $\{p(r\,|\,a)\,|\,a \in \mathscr{A}\}$ is a set of reward's distributions;
- On each step, t an agent chooses a_t and get reward $r_t \sim p(\,.\,|\,a_t)$

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

RL Formalism

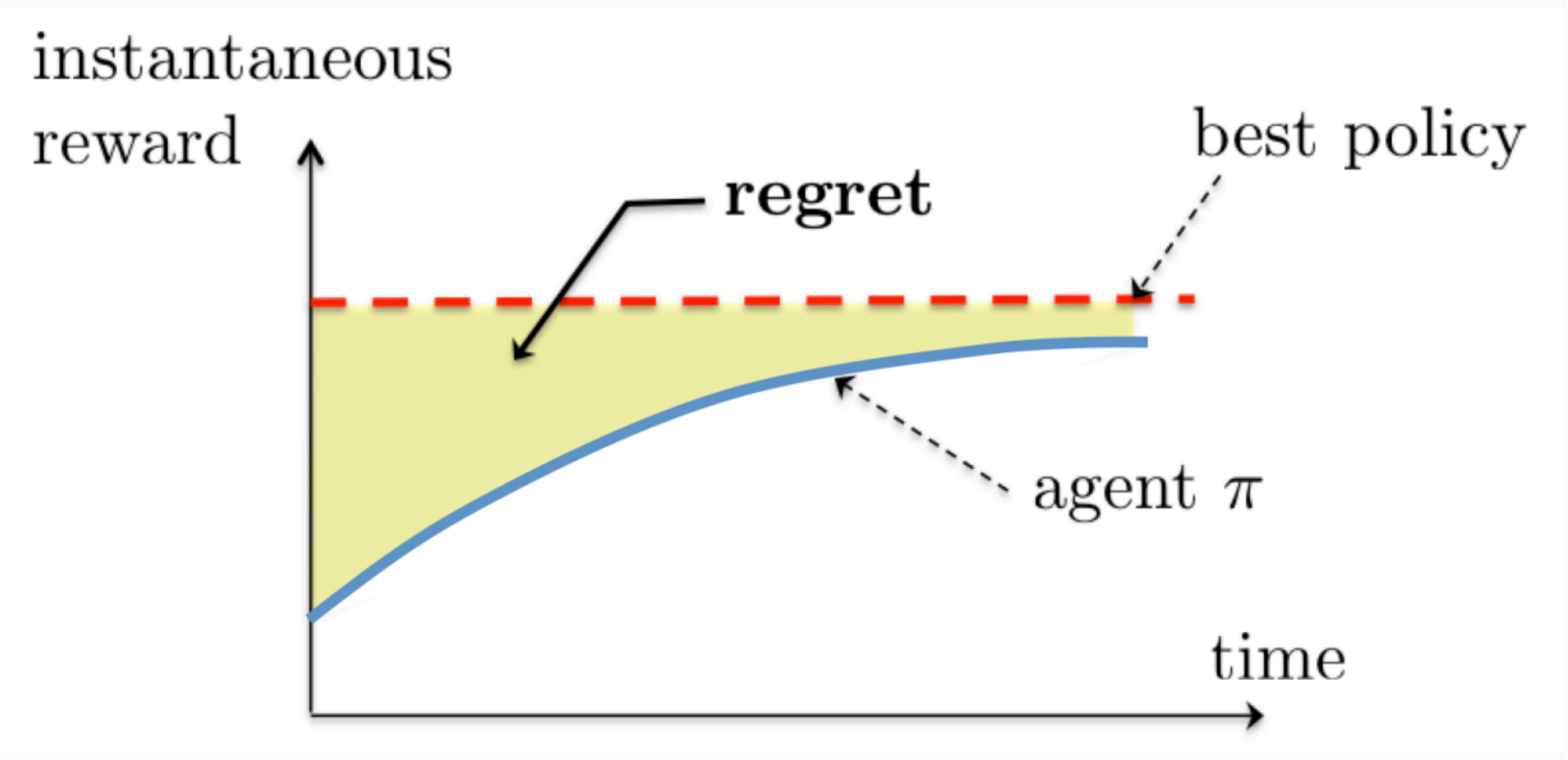
- **Policy**: π is just a rule of making decisions on each step
- **Action value function**: $Q(a) = \mathbb{E}[r_t \mid a_t = a]$
- **Optimal value**: $V^* = \max_a Q(a)$
- **Gap**: $V^* - Q(a) \geq 0$
- **Total Regret:** $\mathbb{E} \sum_{t=1}^{T} [V^* - Q(a_t)] \rightarrow \min_{\pi}$

RL Formalism

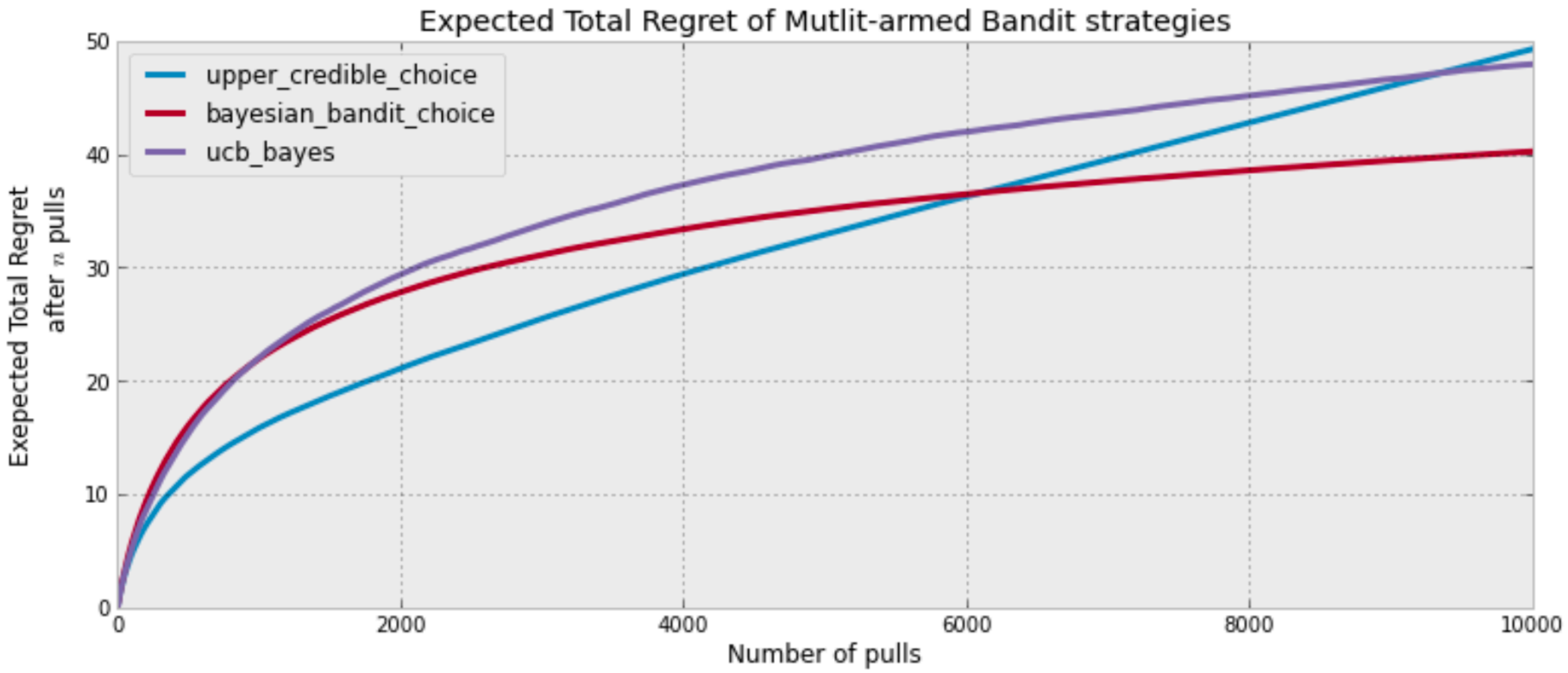
- **Policy**: π is just a rule of making decisions on each step
- **Action value function**: $Q(a) = \mathbb{E}[r_t \,|\, a_t = a]$
- **Optimal value**: $V^* = \max_a Q(a)$
- **Gap**: $V^* - Q(a) \geq 0$
- **Total Regret:** $\mathbb{E} \sum_{t=1}^{T} [V^* - Q(a_t)] \to \min_\pi \iff \mathbb{E}_{p(r|a)}[\sum_{t=1}^{T} r_t] \to \max_\pi$

Regret Minimisation

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



Source



Source

Regret

$$\sum_{t=1}^{T} [V^* - Q(a_t)] = TV^* - \sum_{t=1}^{T} Q(a_t)$$

Realised regret: $R(T) = TV^* - \sum_{t=1}^{T} Q(a_t)$

Expected regret: $\mathbb{E}[R(T)]$

We are interested in the minimising of $\lim_{T \to \infty} \mathbb{E}[R(T)]$

Regret Bounds

$$R(T) = \sum_a n_T(a)[V^* - Q(a)]$$

Upper bound: $\mathbb{E}[R(T)] \leq T(V^* - \min_a Q(a))$

Regret Bounds

$$R(T) = \sum_a n_T(a)[V^* - Q(a)]$$

Upper bound: $\mathbb{E}[R(T)] \leq T(V^* - \min_a Q(a))$

<u>**Lower bound**</u>: $\mathbb{E}[R(T)] \geq \log T \sum_{a|V^*>Q(a)} \frac{V^* - Q(a)}{D_{KL}(p(r|a)||p(r|a^*))}$

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

$$Q_t(a) = \frac{\sum_{n=1}^{t} \mathbb{I}(a_n = a) r_n}{\sum_{n=1}^{t} \mathbb{I}(a_n = a)} = \frac{\sum_{n=1}^{t} \mathbb{I}(a_n = a) r_n}{n_t(a)} \iff Q_t(a) = Q_{t-1}(a) + \alpha_t(a)[r_t - Q_{t-1}(a)]$$

$$\alpha_t(a) = \frac{\mathbb{I}[a_t = a]}{n_t(a)}, n_t(a) = n_{t-1}(a) + \mathbb{I}[a_t = a]$$

ε-greedy Policy

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

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

ε-greedy Policy

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

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

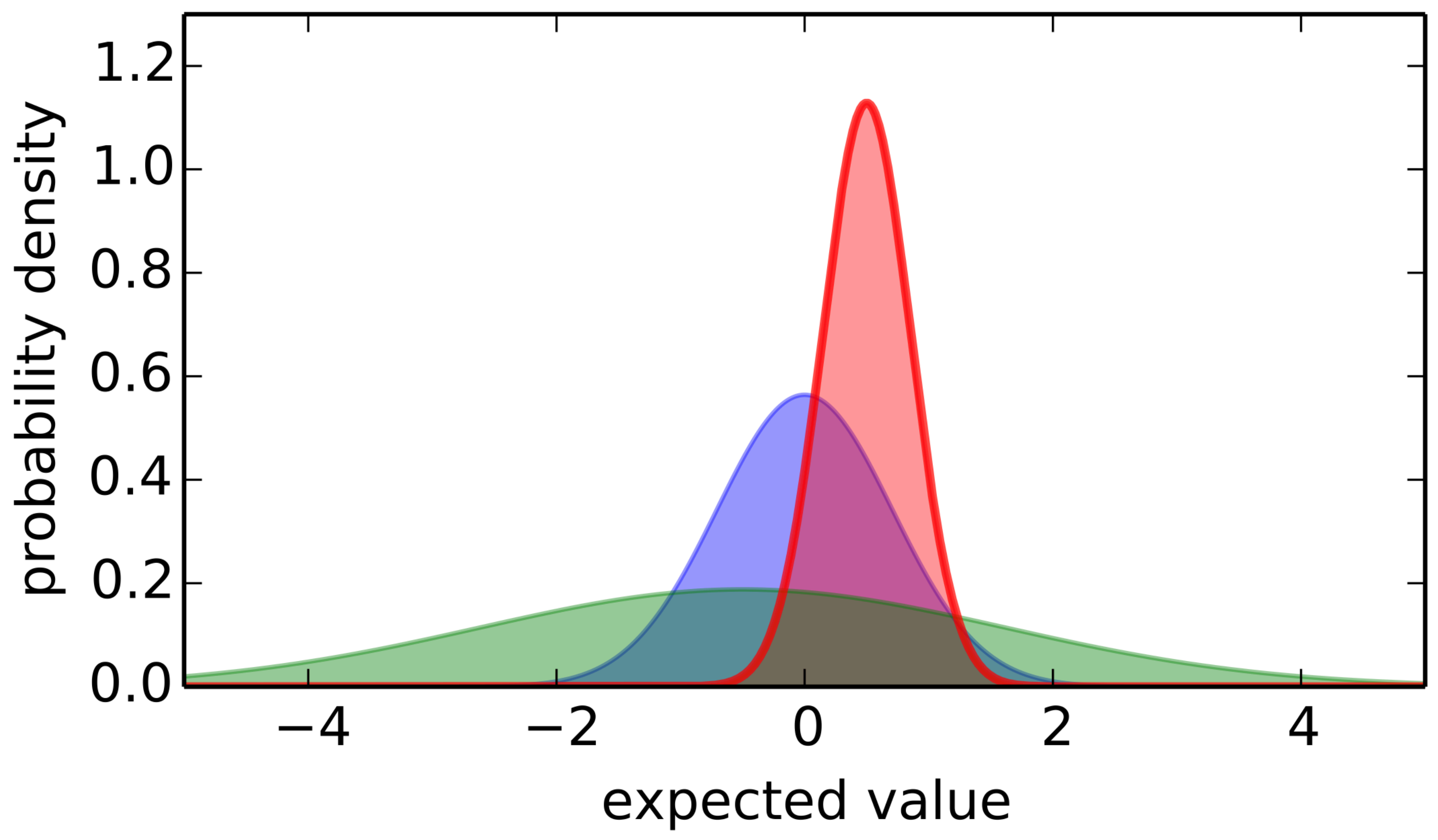
ε-greedy policy has linear regret

Adaptive Exploration

Epsilon-greedy algorithm with exploration probabilities $\varepsilon_t = t^{-\frac{1}{3}}(K \log T)^{\frac{1}{3}}$ achieves regret bound:

$$\mathbb{E}[R(T)] \leq T^{\frac{2}{3}} O(K \log T)^{\frac{1}{3}}$$

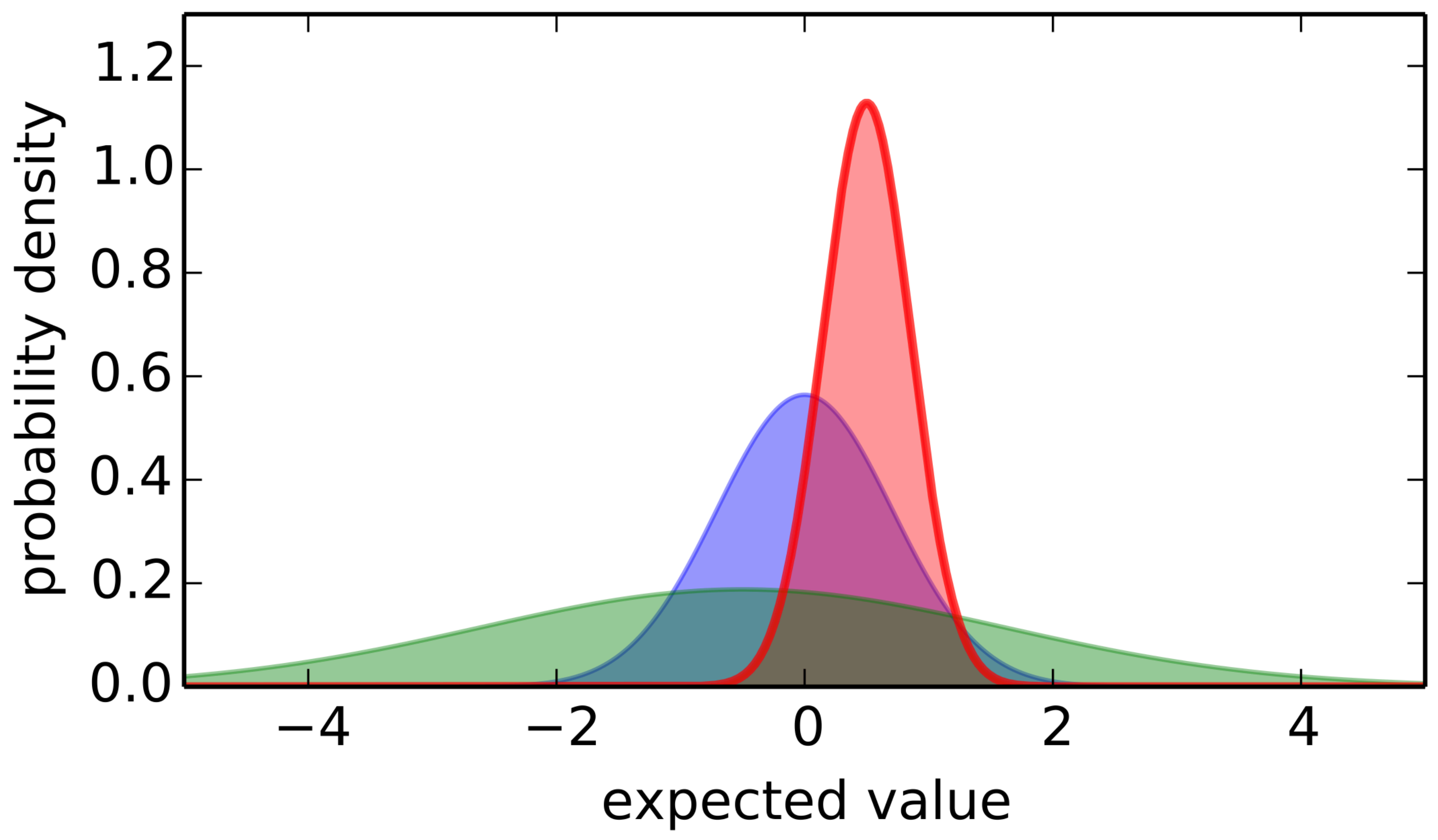
Optimism in the Face of Uncertainty

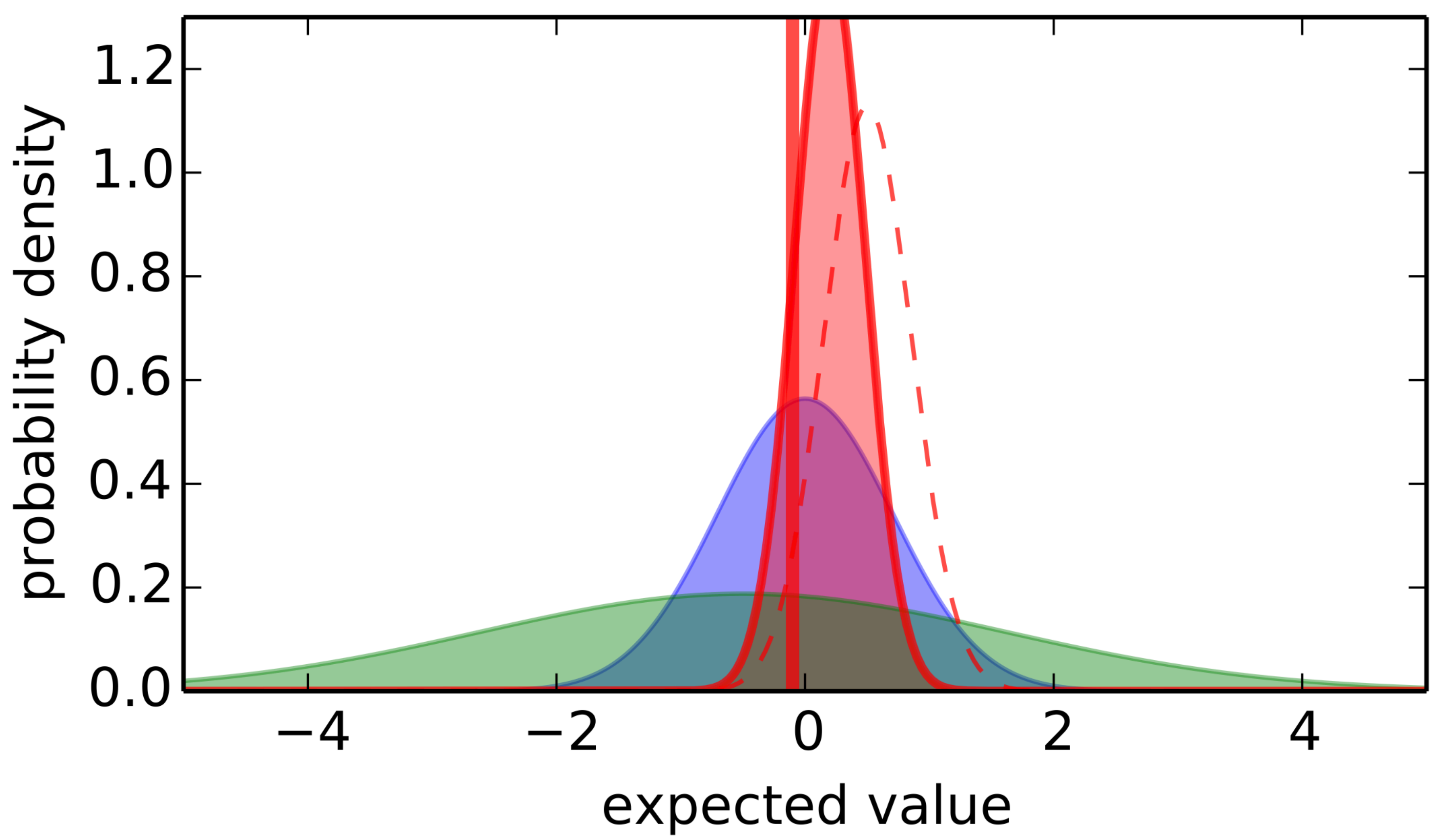


- Which action should we pick?

Source

Optimism in the Face of Uncertainty

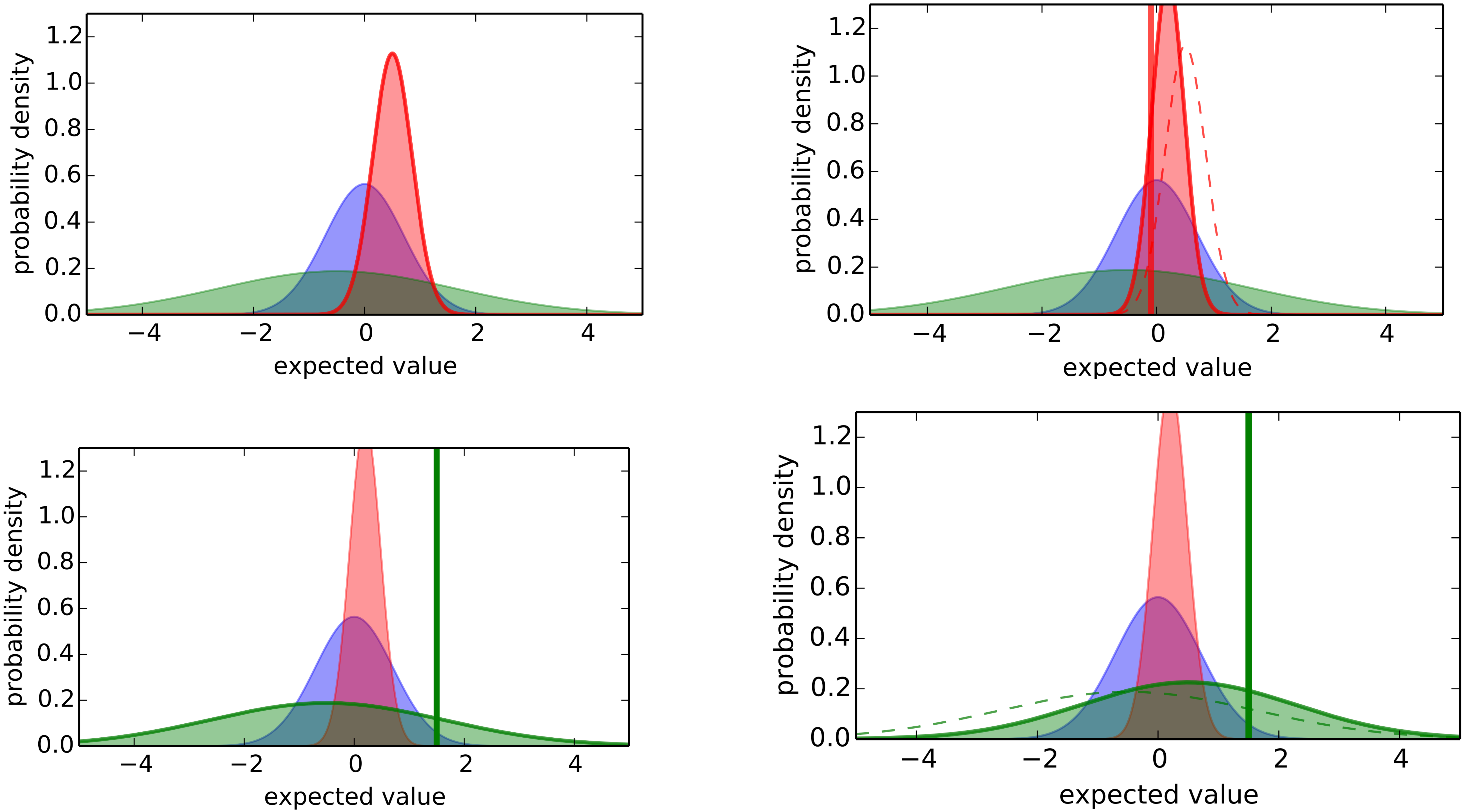




- Which action should we pick?
- The more uncertain we are about an action-value, the more critical it is to explore that action.
- It could be the best action.

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 $n_t(a)$ has been selected
  - Small $n_t(a) \Rightarrow$ large $U_t(a)$ (estimated value is uncertain)
  - Large $n_t(a) \Rightarrow$ small $U_t(a)$ (estimated value is accurate)
- Select action maximising upper confidence bound (UCB): $a_t = argmax_{a \in A}[Q_t(a) + U_t(a)]$

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 $\mathbb{P}(\mu \geq \bar{X}_t + u) \leq e^{-2tu^2}$.

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

UCB

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

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 get more information: $p = \frac{1}{t^{2C}}$

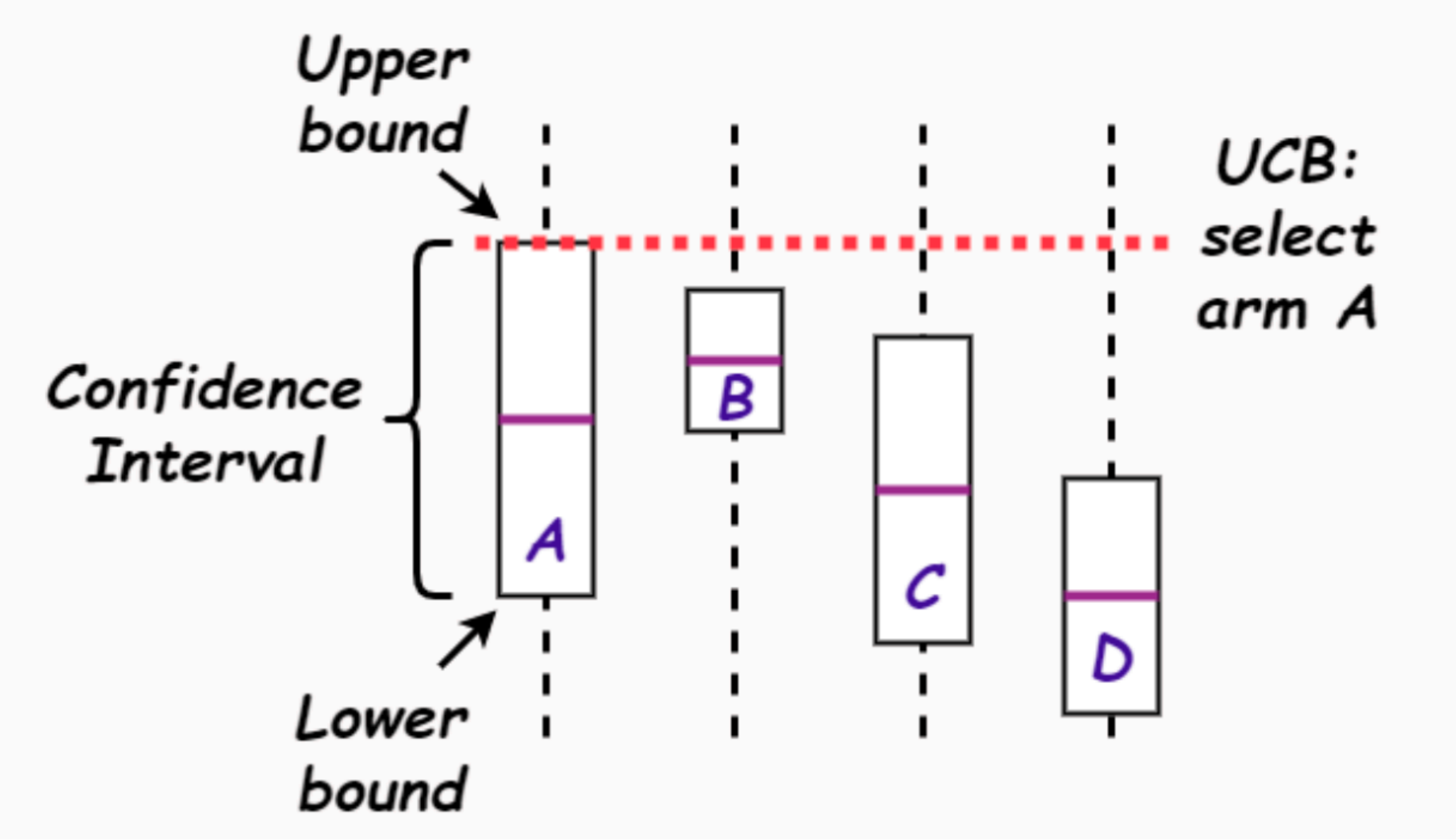
$$U_t(a) = C\sqrt{\frac{\log t}{2N_t(a)}}$$

UCB

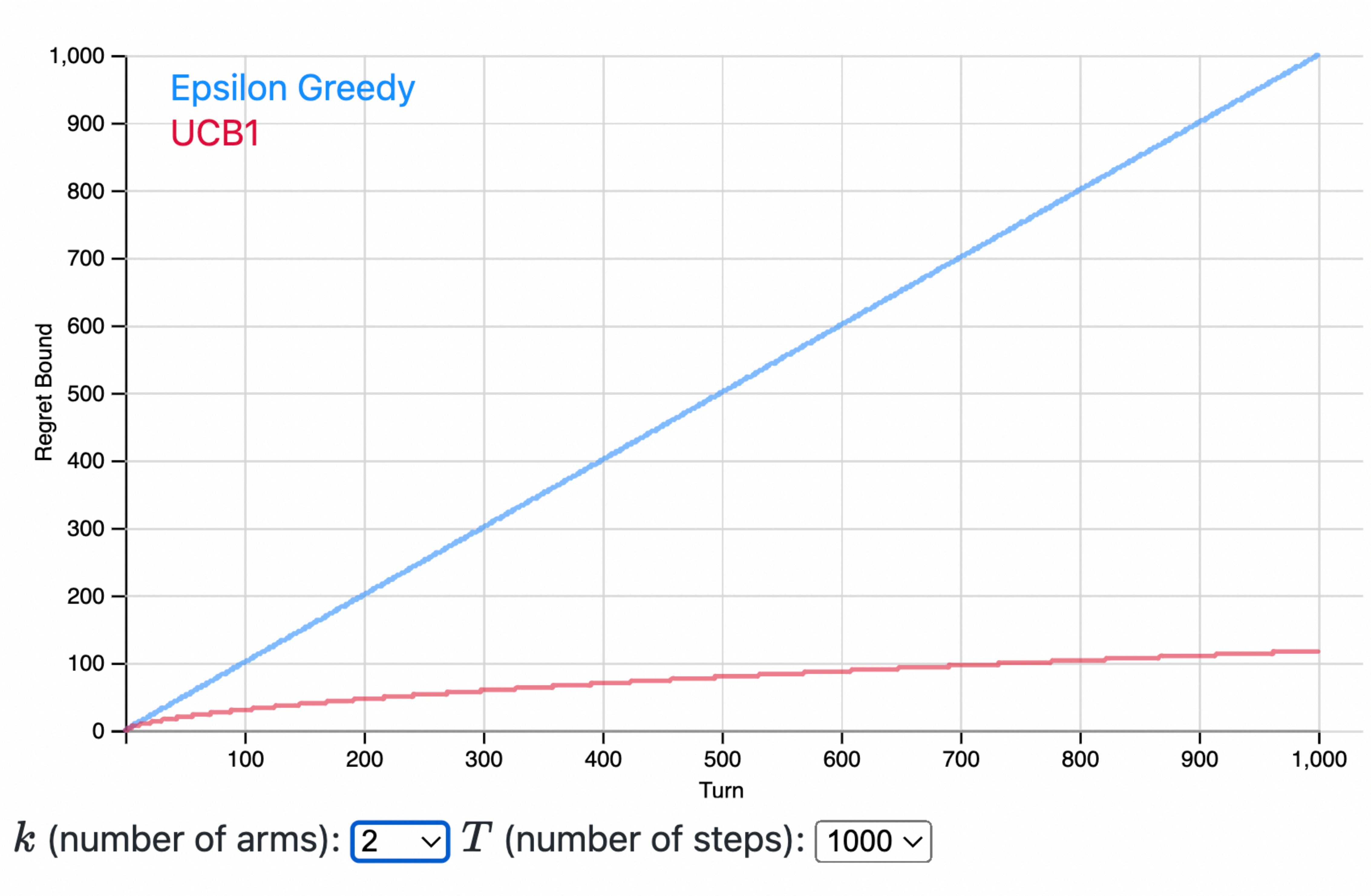
- Select action maximising upper confidence bound (UCB):

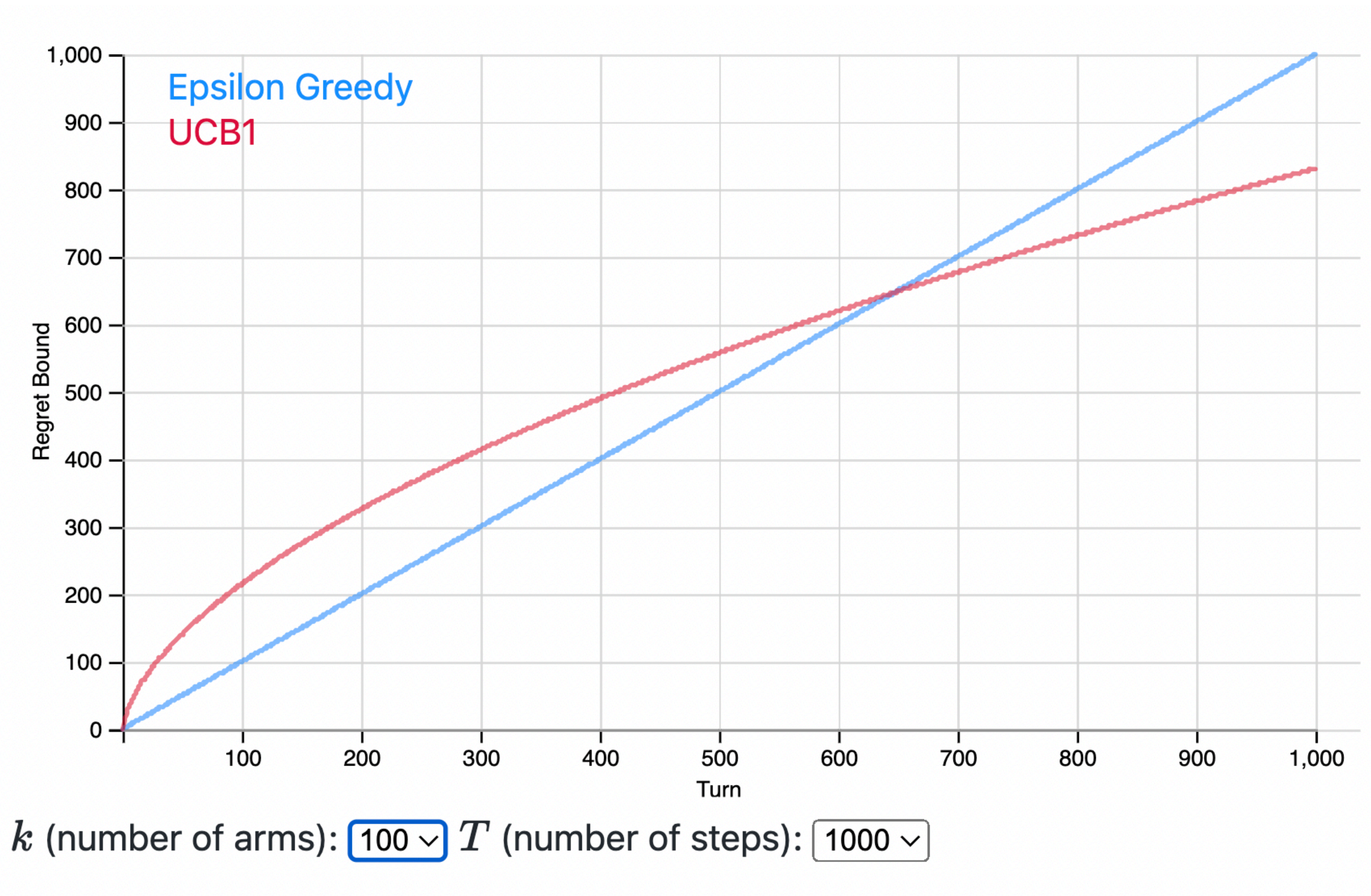
$$a_t = argmax_{a \in A}[Q_t(a) + c\sqrt{\frac{\log t}{2N_t(a)}}]$$

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



Comparison





Source

Model-based Approach

- Learn the environment's model: $p(r \mid a) \approx p(r \mid \theta_a)$
- It allows us to inject rich prior knowledge θ_a^0
- We can then use posterior belief to guide exploration: $p(\theta_a) \leftarrow p(\theta_a \mid r) \propto p(r \mid \theta_a)p(\theta_a)$
- $\mathbb{E}p(\,.\mid \theta_a)$ is a random variable

Probability Matching

- We can choose an action in the following way:

$$a = argmax_a \mathbb{E}_{\theta_a \sim p(\theta_a)} \mathbb{E}p(\,.\,|\theta_a)$$

- However, now there is a probability that it's not optimal
- Let's choose an action with the probability of being optimal

$$\pi(a) = \mathbb{P}[\mathbb{E}p(\,.\,|\theta_a) > \mathbb{E}p(\,.\,|\theta_{a'}), a \neq a']$$

Thompson Sampling

- Thompson sampling implements probability matching:

$$\pi(a) = \mathbb{P}[\mathbb{E}p(\,.\,|\theta_a) > \mathbb{E}p(\,.\,|\theta_{a'}), a \neq a']$$

1. Use Bayes's law to compute the posterior distribution $p(\theta_a | r)$
2. Sample parameters θ_a from these distributions
3. Select action maximising value on sample: $a = argmax_{a'} \mathbb{E}p(r | \theta_{a'})$

Thompson Sampling

- Thompson sampling implements probability matching:

$$\pi(a) = \mathbb{P}[\mathbb{E}p(\,.\,|\theta_a) > \mathbb{E}p(\,.\,|\theta_{a'}), a \neq a']$$

1. Use Bayes's law to compute the posterior distribution $p(\theta_a | r)$
2. Sample parameters θ_a from these distributions:
3. Select action maximising value on sample: $a = argmax_{a'}\mathbb{E}p(r|\theta_{a'})$

Thompson sampling achieves a logarithmic bound!

Contextual Bandits

1. The algorithm observes a "context" x_t;

2. The algorithm picks an arm a_t from the K possible actions;

3. The reward $r_t \sim p(\,.\,|\,x_t, a_t)$ is realised.

Example: a user with a known "user profile" arrives in each round, and the context is the user profile.

Disjoint LinUCB

- Let $x_{t,a} \in \mathbb{R}^d$ the context summarise information of both the global context x_t (e.g. for user u_t) and arm a.
- The main assumption: $\mathbb{E}[r_{t,a} | x_t, a] = x_{t,a}^T \theta_a$
- Apply ridge regression to derive $\hat{\theta}_{t,a}$ on each step.

Disjoint LinUCB

- $D_a \in \mathbb{R}^{m \times d}$ is a matrix containing contexts which were observed previously for action a.
- $A_a = D_a^T D_a + I_d$
- $b_a = r_t x_{t,a}$
- Expected payoff on each step: $x_{t,a} \hat{\theta}_a$ with variance $x_{t,a}^T A_a^{-1} x_{t,a}$.

Algorithm 1 LinUCB with disjoint linear models.

```
Inputs: α ∈ ℝ₊
for t = 1, 2, 3, ..., T do
    Observe features of all arms a ∈ 𝒜_t: x_{t,a} ∈ ℝ^d
    for all a ∈ 𝒜_t do
        if a is new then
            A_a ← I_d (d-dimensional identity matrix)
            b_a ← 0_{d×1} (d-dimensional zero vector)
        end if
        θ̂_a ← A_a^{-1} b_a
        p_{t,a} ← θ̂_a^⊤ x_{t,a} + α sqrt(x_{t,a}^⊤ A_a^{-1} x_{t,a})
    end for
    Choose arm a_t = arg max_{a∈𝒜_t} p_{t,a} with ties broken arbi-
    trarily, and observe a real-valued payoff r_t
    A_{a_t} ← A_{a_t} + x_{t,a_t} x_{t,a_t}^⊤
    b_{a_t} ← b_{a_t} + r_t x_{t,a_t}
end for
```

Regret bound: $\tilde{O}(\sqrt{KdT})$,

where $\tilde{O}(\,.\,)$ is is the same as $O(\,.\,)$ but suppresses logarithmic factors.

Bayesian Interpretation

- Gaussian prior: $p(\theta_a) = \mathcal{N}(0, I_d)$
- On step t for action a:
    - m noisy measurements: $\mathbf{r}_a \sim \mathcal{N}(D_a\theta_a, I_m)$
    - Posterior distribution: $\theta_a \sim \mathcal{N}(\hat{\theta}_a, A_a^{-1})$
    - $A_a = D_a^T D_a + I_d, \hat{\theta}_a = A_a^{-1} D_a^T \mathbf{r}_a$
    - $x_{t,a}^T \theta_a \sim \mathcal{N}(x_{t,a}^T \hat{\theta}_a, x_{t,a}^T A_a^{-1} x_{t,a})$
    - UCB: $x_{t,a}^T \hat{\theta}_a + \alpha \sqrt{x_{t,a}^T A_a^{-1} x_{t,a}}$

Neural UCB

Algorithm 1 NeuralUCB

Input: Number of rounds T, regularization parameter λ, exploration parameter ν, confidence parameter δ, norm parameter S, step size η, number of gradient descent steps J, network width m, network depth L.

Initialization: Randomly initialize $\boldsymbol{\theta}_0$ as described in the text

Initialize $\mathbf{Z}_0 = \lambda\mathbf{I}$

for $t = 1, \ldots, T$ **do**

  Observe $\{\mathbf{x}_{t,a}\}_{a=1}^{K}$

  for $a = 1, \ldots, K$ **do**

    Compute $U_{t,a} = f(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\sqrt{\mathbf{g}(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1})^\top \mathbf{Z}_{t-1}^{-1}\mathbf{g}(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1})/m}$

    Let $a_t = \mathrm{argmax}_{a\in[K]} U_{t,a}$

  end for

  Play a_t and observe reward r_{t,a_t}

  Compute $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1})\mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1})^\top/m$

  Let $\boldsymbol{\theta}_t = \mathrm{TrainNN}(\lambda, \eta, J, m, \{\mathbf{x}_{i,a_i}\}_{i=1}^{t}, \{r_{i,a_i}\}_{i=1}^{t}, \boldsymbol{\theta}_0)$

  Compute

$$\gamma_t = \sqrt{1 + C_1 m^{-1/6}\sqrt{\log m}L^4 t^{7/6}\lambda^{-7/6}} \cdot \left(\nu\sqrt{\log\frac{\det \mathbf{Z}_t}{\det \lambda\mathbf{I}} + C_2 m^{-1/6}\sqrt{\log m}L^4 t^{5/3}\lambda^{-1/6} - 2\log\delta} + \sqrt{\lambda}S\right)$$
$$+ (\lambda + C_3 tL)\left[(1-\eta m\lambda)^{J/2}\sqrt{t/\lambda} + m^{-1/6}\sqrt{\log m}L^{7/2}t^{5/3}\lambda^{-5/3}(1+\sqrt{t/\lambda})\right].$$

end for

Regret bound: $\tilde{O}(\tilde{d}\sqrt{T})$

Bandits in Practice

- Recommender systems: Spotify, Netflix
- Adaptive clinical trials

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
5. Introduction to Multi-Armed Bandits

Thank you for your attention!