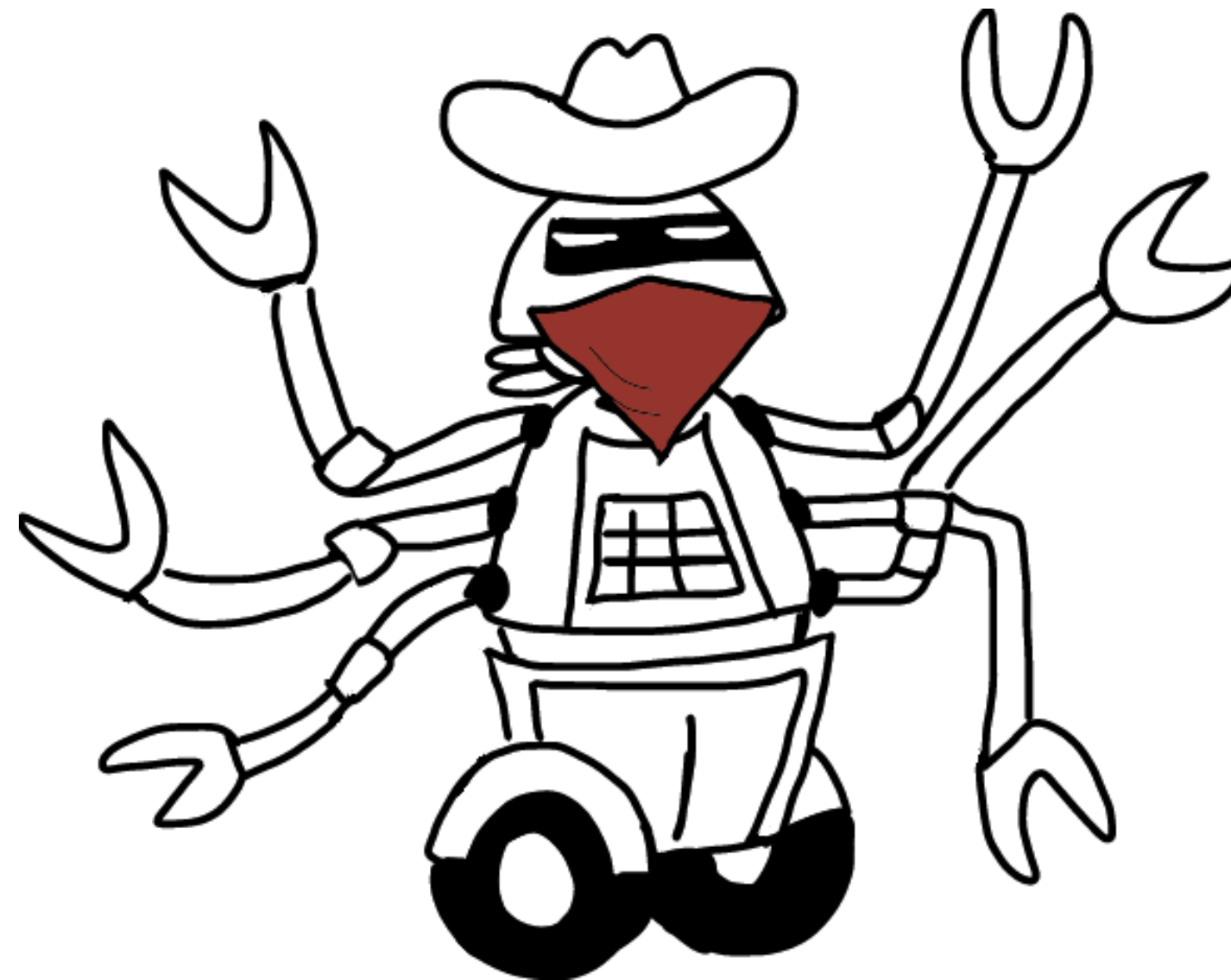


# Reinforcement Learning

HSE, autumn - winter 2022

## Lecture 7: Multi-armed Bandits



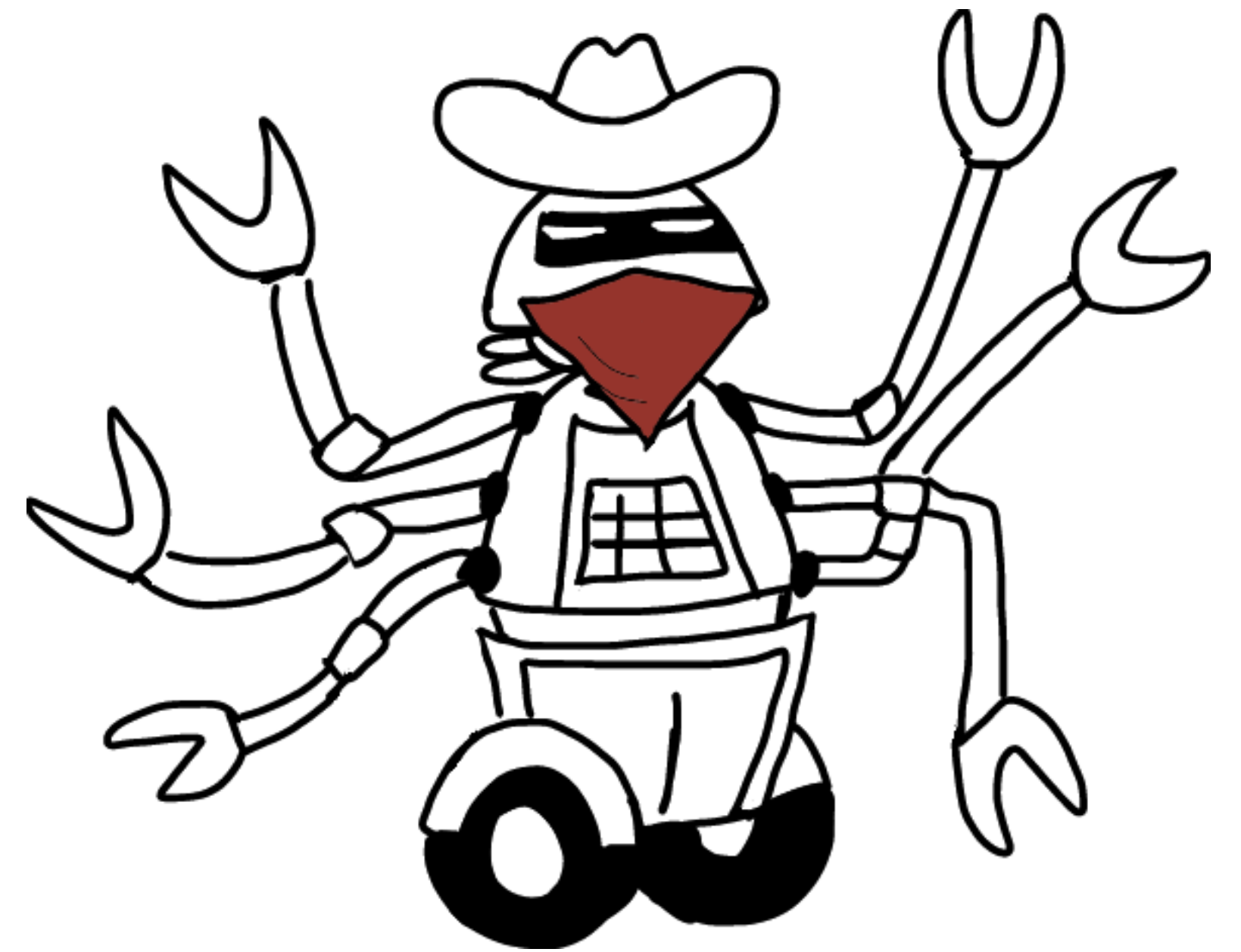
Sergei Laktionov  
[slaktionov@hse.ru](mailto:slaktionov@hse.ru)  
[LinkedIn](#)

# Background

1. Practical RL course by YSDA, week 5
2. Sutton & Barto, Chapter 2
3. DeepMind course, Lecture 2

# Multi-armed Bandit Problem Statement

Assume that the episode ends after the first step so we have only one state in the environment. You as an agent are facing repeatedly with a choice among  $k$  different actions.

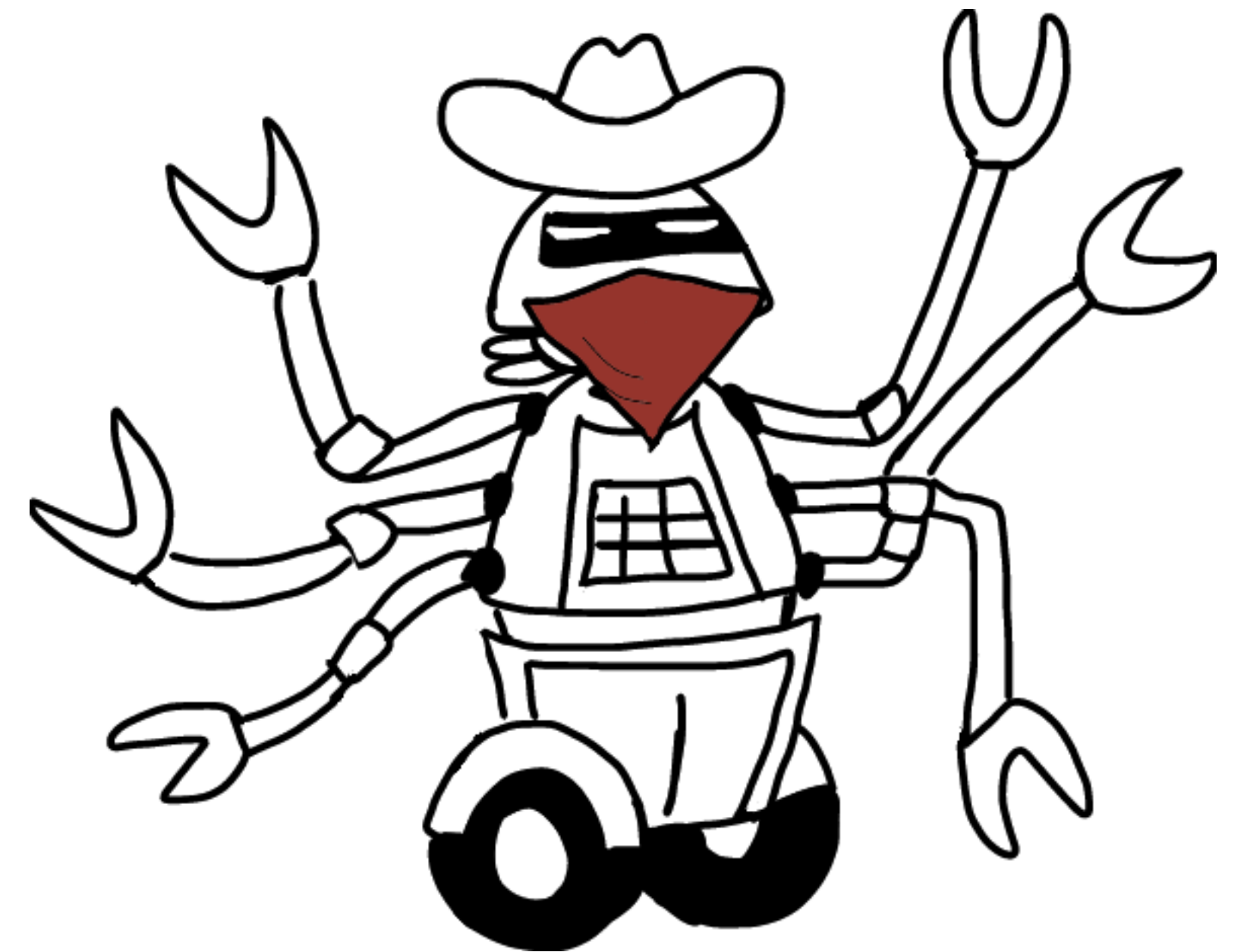


Source

# Multi-armed Bandit Problem Statement

Assume that the episode ends after the first step so we have only one state in the environment. You as an agent are facing repeatedly with a choice among  $k$  different actions.

- A multi-armed bandit is a set of distributions  $\{\mathcal{R}_a \mid a \in \mathcal{A}\}$
- On each step  $t$  an agent chooses  $A_t$  and get reward  $R_t \sim \mathcal{R}_{A_t}$
- The 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:**  $Q(a) = \mathbb{E}[R_t | A_t = a]$ ,  $Q(a_t) = Q_t$

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

**Regret:**  $V^* - Q(a) \geq 0$

# 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:**  $V^* - Q(a) \geq 0$

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

# Multi-armed Bandit Problem Statement

**Exploration:** find the best action which maximises expected reward

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

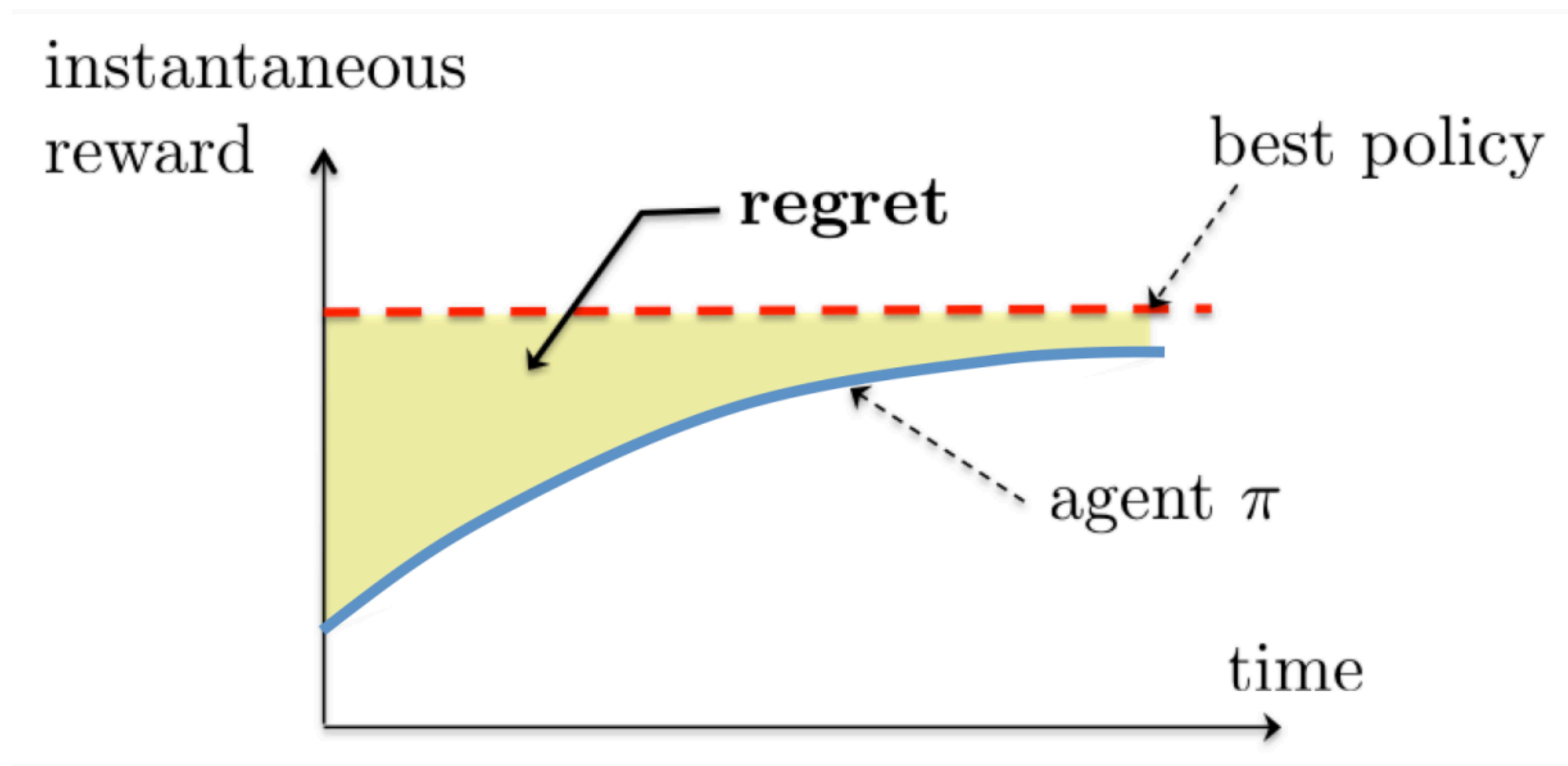
**Optimal value:**  $V^* = \max_a Q(a)$

**Regret:**  $V^* - Q(a) \geq 0$

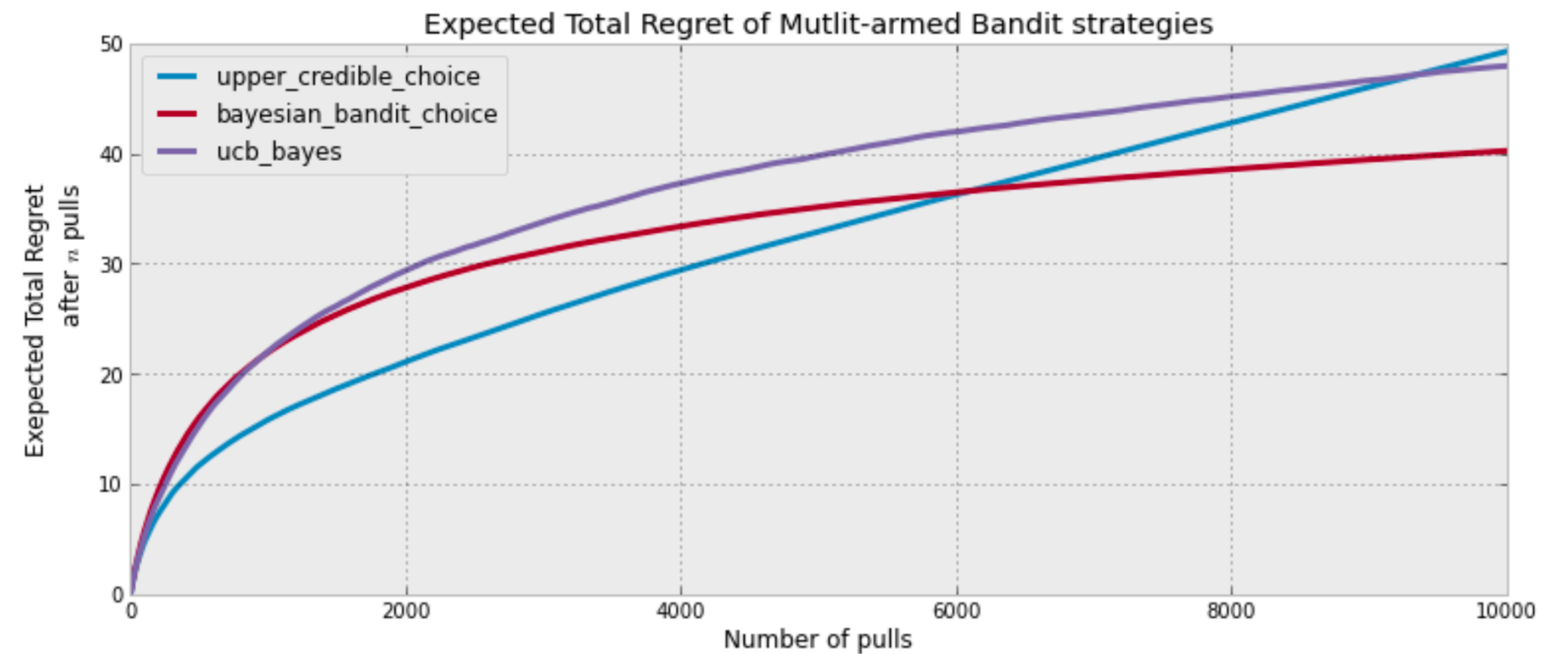
**Total Regret:**  $\sum_{t=1}^T [V^* - Q(a_t)] \rightarrow \min_{\pi} \iff \mathbb{E}_{p(r|a)} \left[ \sum_{t=1}^T R_t \right]$

# Regret Minimisation

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



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

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

# $\epsilon$ -greedy Policy

$$\pi_t(a) = \begin{cases} (1 - \epsilon) + \frac{\epsilon}{|\mathcal{A}|}, & \text{if } Q_t(a) = \max_b Q_t(b) \\ \frac{\epsilon}{|\mathcal{A}|}, & \text{otherwise} \end{cases}$$

- Greedy can stuck on a suboptimal action forever
- $\epsilon$ -greedy continues to explore

# $\epsilon$ -greedy Policy

$$\pi_t(a) = \begin{cases} (1 - \epsilon) + \frac{\epsilon}{|\mathcal{A}|}, & \text{if } Q_t(a) = \max_b Q_t(b) \\ \frac{\epsilon}{|\mathcal{A}|}, & \text{otherwise} \end{cases}$$

- Greedy can stuck on a suboptimal action forever
- $\epsilon$ -greedy continues to explore

$\epsilon$ -greedy policy has linear regret

# Gradient Policy

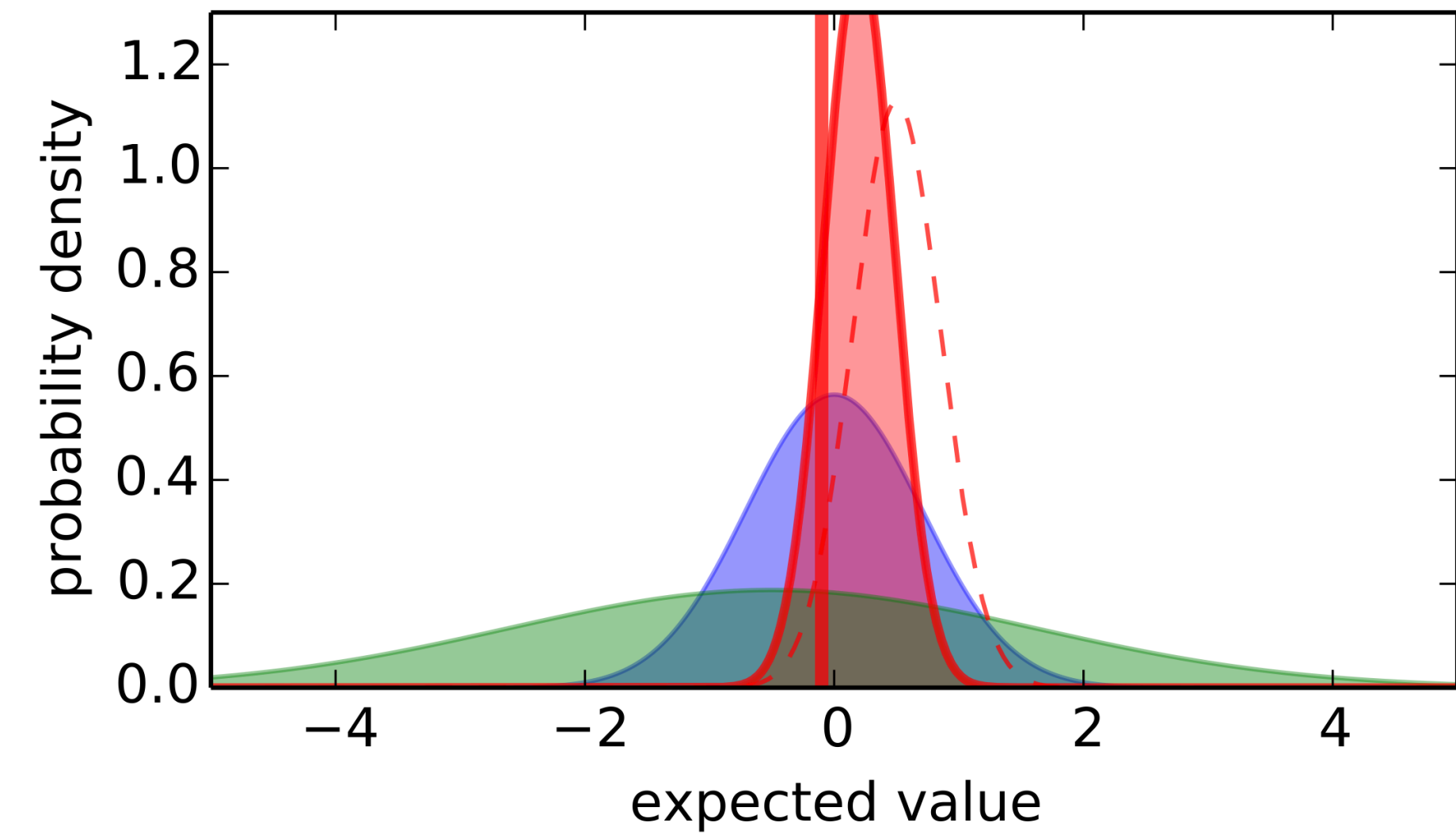
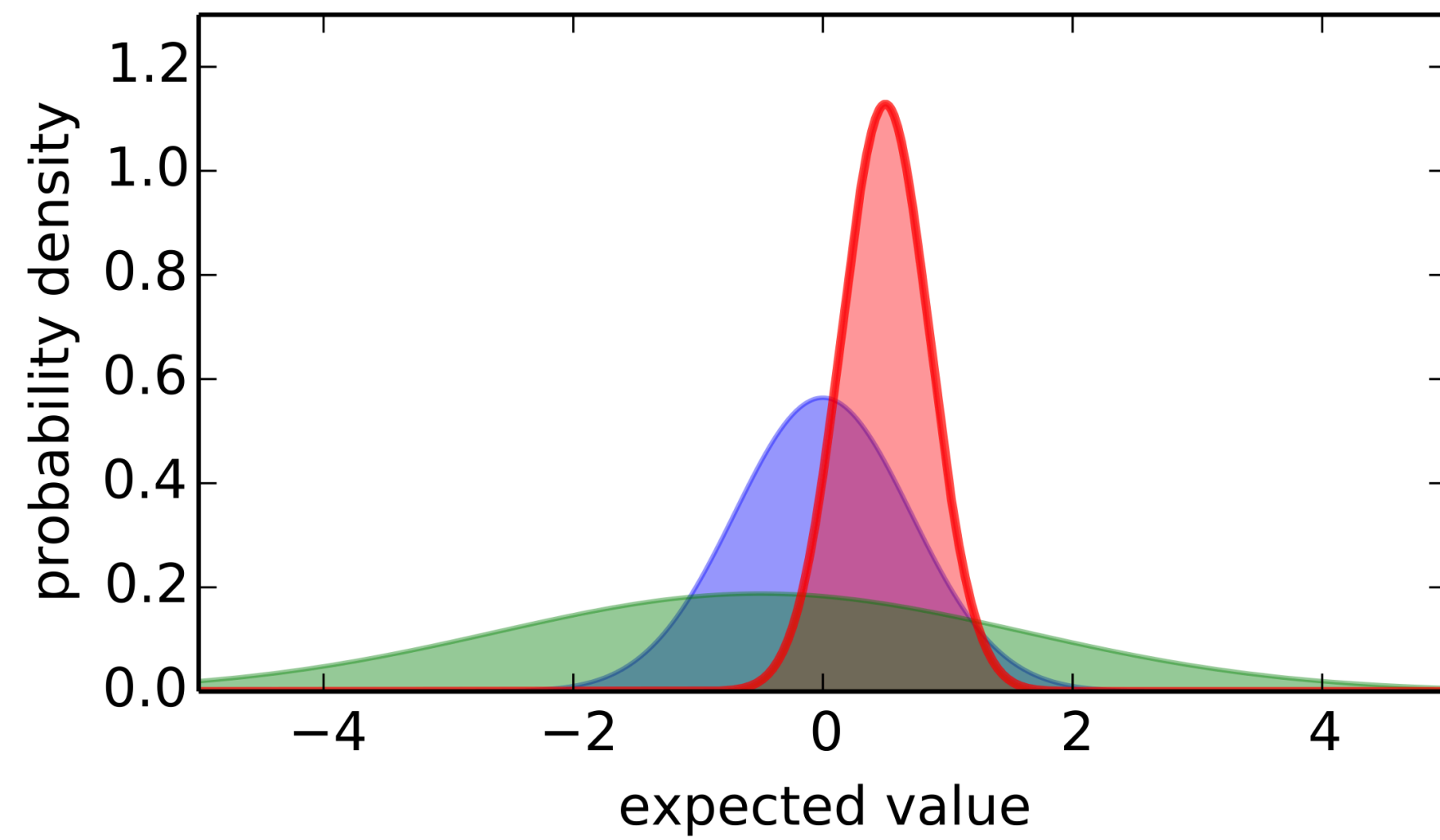
We can learn softmax policy using REINFORCE via gradient ascent, but there is no still guarantees for convergence to global optimum.

# Regret Lower Bound

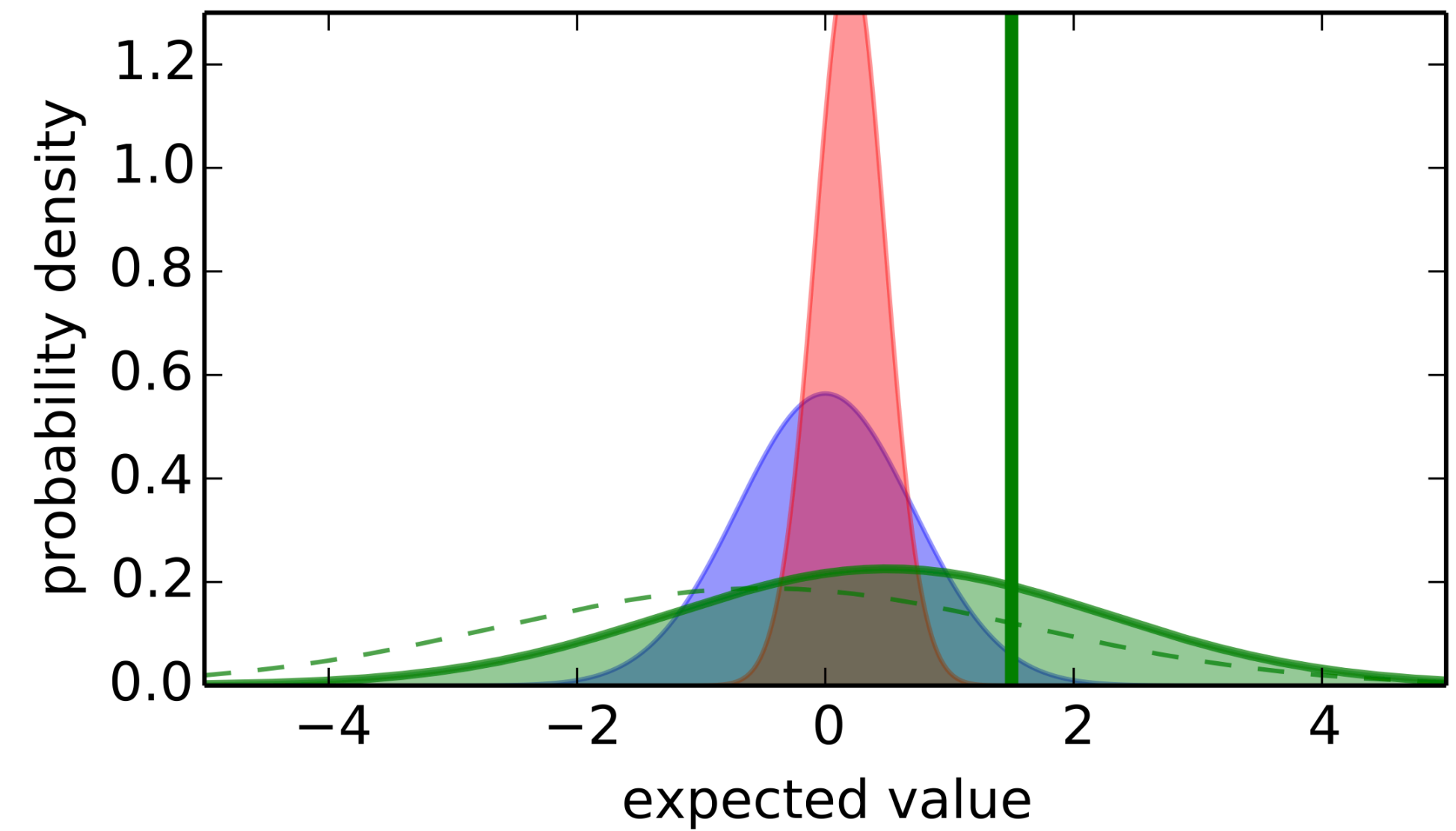
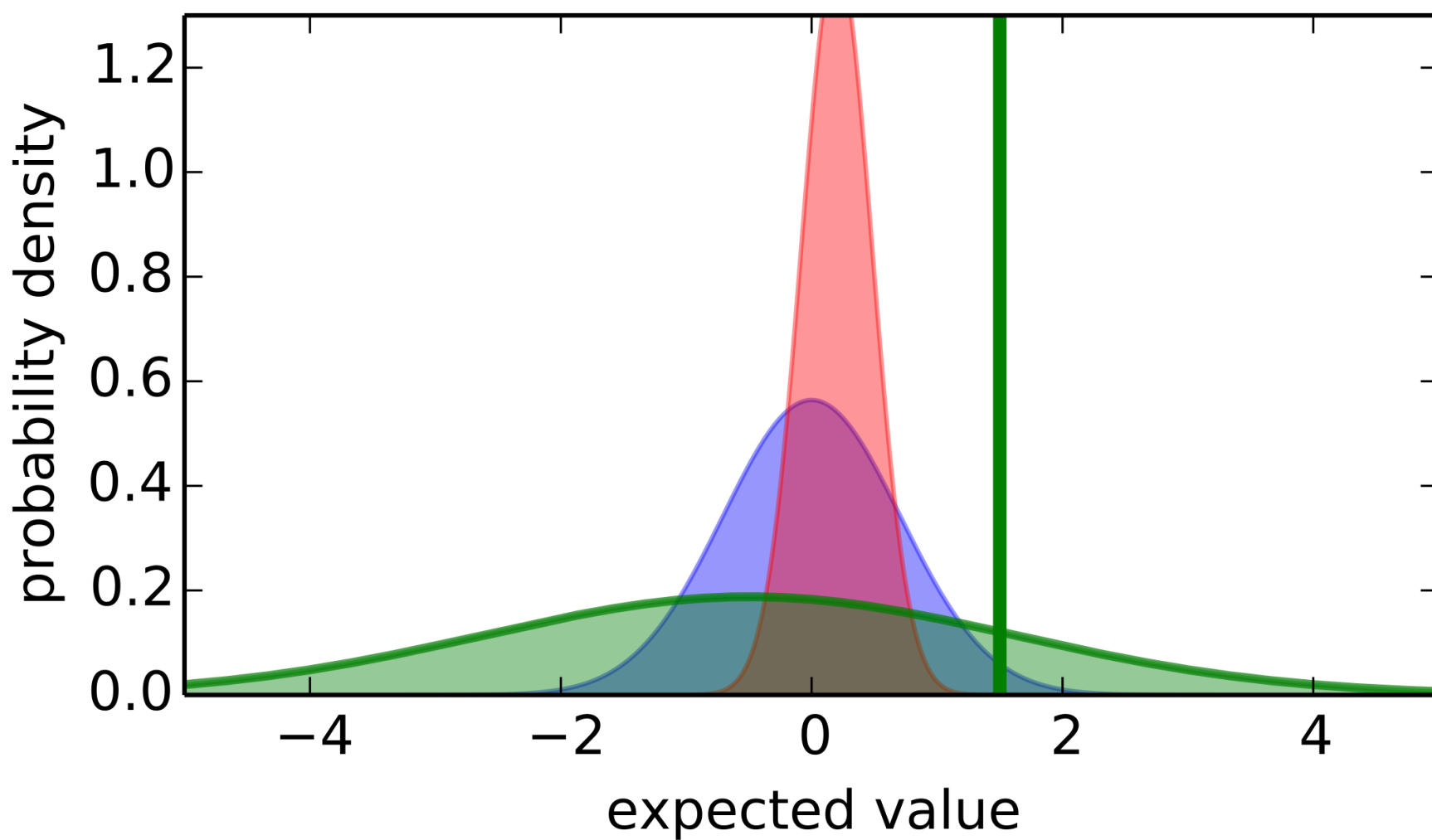
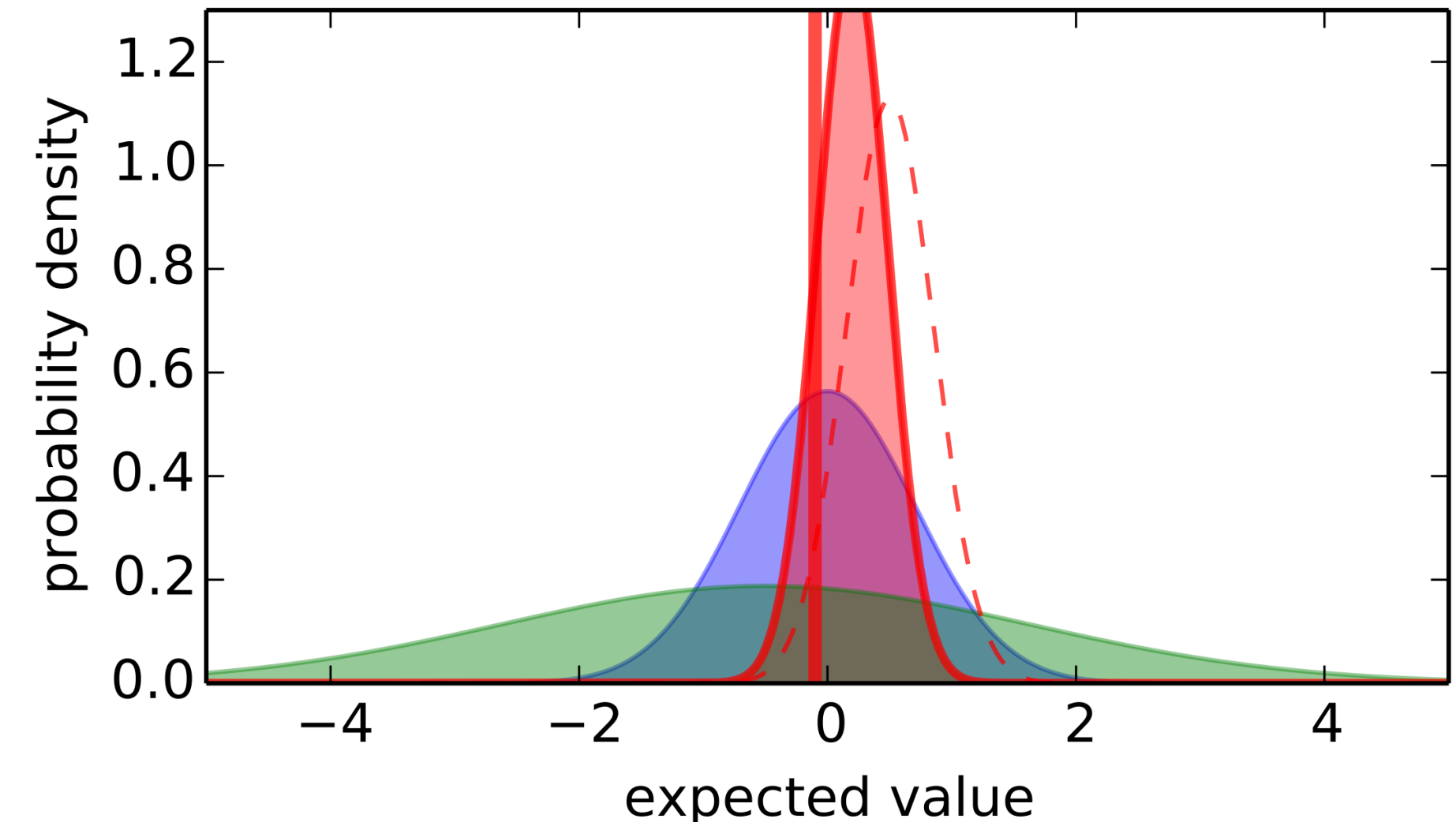
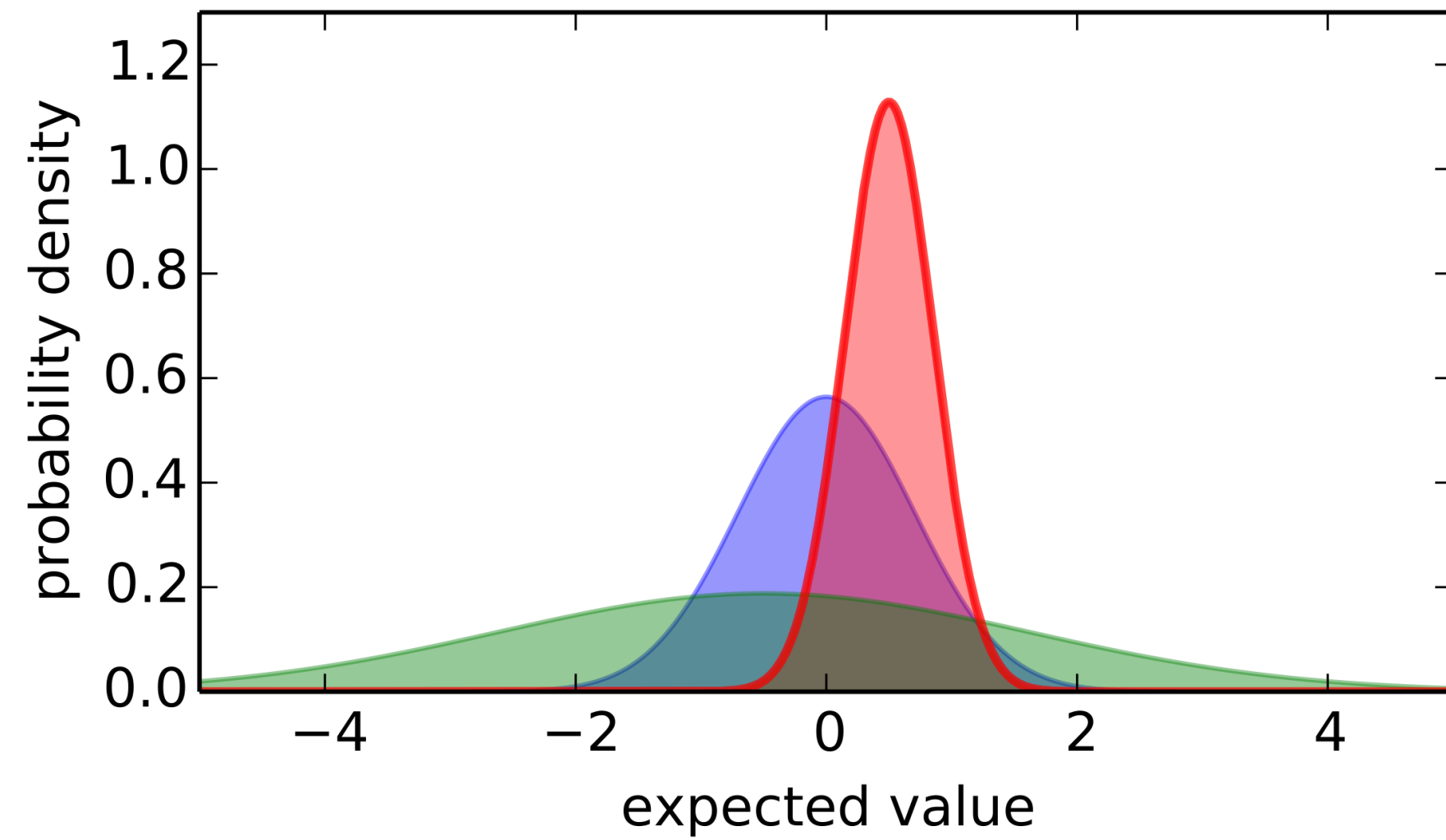
Theorem:

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

# Optimism in the Face of Uncertainty



# 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.
- Select action maximising upper confidence bound (UCB):  
$$a_t = \operatorname{argmax}_{a \in A} [Q_t(a) + U_t(a)]$$

# Optimality of UCB

Hoeffding's Inequality:

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

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

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

# Optimality of UCB

Hoeffding's Inequality:

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

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

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

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

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

# Optimality of UCB

Hoeffding's Inequality:

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

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

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

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

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

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

# UCB

- Select action maximising upper confidence bound (UCB):

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

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

# Bayesian Approach

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

# Thompson Sampling

- Priors  $p(\theta_a), a \in \mathcal{A}$
- $p(\theta_a) \leftarrow p(\theta_a | r_t) \propto p(r_t | \theta_a)p(\theta_a)$  is a bayesian update
- We can choose an action with maximal expected reward under the known distributions:  $a_{t+1} = \operatorname{argmax}_a \mathbb{E}_{\theta_a \sim p(\theta_a)} \mathbb{E}_{p(r|\theta_a)} r$
- 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  $\theta_a \sim p(\theta_a), a \in \mathcal{A}$  and choose action with the maximal expected reward under the  $\theta_a$

**Thank you for your attention!**