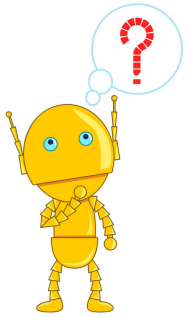
Introduction to bandit algorithms

Julia Olkhovskaya

Reinforcement Learning Summer School
Amsterdam, 2022

Bandits



a_1 a_2 a_3 a_4

For t $= 1, 2, \ldots, T$:

- The learner chooses **action** A_t,
- The learner gains and observes **reward** $r_t(A_t)$

Goal:

- minimize regret $R_T = \max_a \mathbb{E}[\sum_{t=1}^{T} r_t(a) - r_t(A_t)]$

Applications

- Ad placement
  - Arms – possible ads
  - Reward – a click
- Website optimization
  - Arms – possible website options
  - Reward – user engagement
- Clinical trials
  - Arms – possible medications
  - Reward – health outcomes
- Social networks
  - Arms – possible posts/ads
  - Reward – a click, time spended
- Where you can see it used:
  - Netflix, Amazon, eBay, Yahoo…

Why bandits are useful for more complex RL?

- Learning to balance exploration/exploitation
- The model is simple for the theoretical studies
- Dealing with uncertainty about the environment
- Easily incorporates adversarial data
- Useful algorithm design principles

Adversarial bandits

Adversarial bandits

For t $= 1, 2, \ldots, T$:

- The learner chooses action $A_t \in \{1, \ldots, K\}$,
- The adversary simultaneously picks losses $\ell_t \in [0,1]^K$
- The learner suffers loss $\ell_t(A_t)$

Adversarial bandits

For t $= 1, 2, \ldots, T$:

- The learner chooses action $A_t \in \{1, \ldots, K\}$,
- The adversary simultaneously picks losses $\ell_t \in [0,1]^K$
- The learner suffers loss $\ell_t(A_t)$

- Follow the leader: $A_t = argmin_a \sum_{s=1}^{t-1} \ell_t(a) I[A_t = a]$

Adversarial bandits

For t $= 1, 2, \ldots, T$:

- The learner chooses action $A_t \in \{1, \ldots, K\}$,
- The adversary simultaneously picks losses $\ell_t \in [0,1]^K$
- The learner suffers loss $\ell_t(A_t)$

- Follow the leader: $A_t = argmin_a \sum_{s=1}^{t-1} \ell_t(a) I[A_t = a]$

Caution: adversary can trick the the learner

Randomization is essential!

EXP3 Algorithm (Exponential-weight algorithm for Exploration and Exploitation)

Algorithm:

- Compute $p_{t,a} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,a}\right)}{\sum_{b=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,b}\right)}$
- Sample $A_t \sim p_t$
- Observe ℓ_{t,A_t}

$$R_T \leq \sqrt{KTlogK}$$

- Loss estimate $\hat{\ell}_{t,a} = \frac{\ell_t}{p_t(a)} \mathbb{I}\{A_t = a\}$

- Unbiased estimator: $\mathbb{E}\left[\hat{\ell}_{t,a} \,|\mathcal{F}_{t-1}\right] = \frac{\ell_{t,a}}{p_t(a)} \quad \mathbb{E}[\mathbb{I}\{A_t = a\}|\mathcal{F}_{t-1}] = \ell_{t,a}$

Follow the regularized leader

Idea: add regularization and apply algorithms for online linear optimization to compute distribution over arms:

$$p_t = argmin_{p \in \Delta^K} \left(\eta \left\langle p, \sum_{s=1}^{t-1} \hat{\ell}_t \right\rangle + F(p) \right)$$

- F is a convex function
- $\eta > 0$ is the learning rate
- Different choice of F lead to different algorithms

Follow the regularized leader

- $p_t = argmin_{p \in \Delta^K} \left(\eta \langle p, \sum_{s=1}^{t-1} \hat{\ell}_t \rangle + F(p) \right)$

Example:

- Unnormalised negentropy : $F(p) = \sum_a^K p_a \log p_a - p_a$
- Closed-form solution: $p_{t,a} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,a}\right)}{\sum_{b=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,b}\right)}$

EXP 3

Follow the regularized leader

- $p_t = argmin_{p \in \Delta^K} \left(\eta \langle p, \sum_{s=1}^{t-1} \hat{\ell}_t \rangle + F(p) \right)$

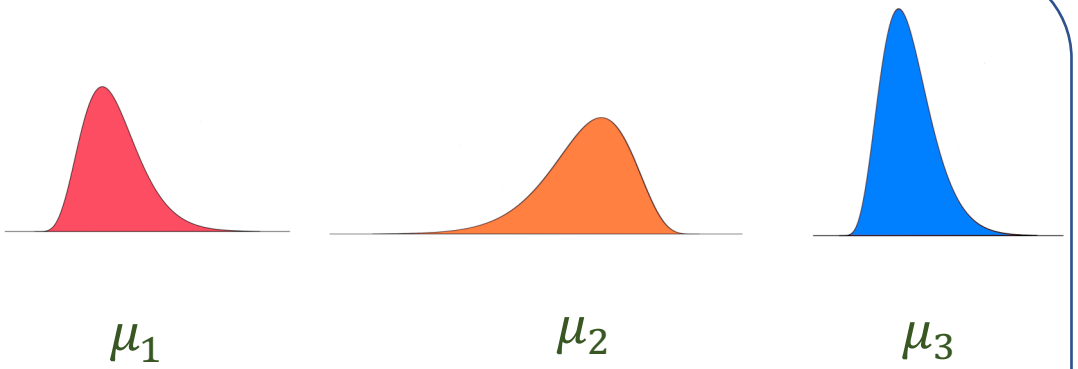
Example:

- log barrier : $F(p) = -\sum_a^K \log p_a$
- No closed-form solution
- Regret bound: $R_T = \tilde{O}\left(\sqrt{K \min_a \sum_{t=1}^T \ell_{t,a}} \right)$

Adaptive bound

Stochastic bandits

Stochastic bandits



Mean reward in **unknown**

For t $= 1, 2, \ldots, T$:

- The learner chooses **action** A_t,
- The learner observes **reward** $r_t(A_t) \sim P_{A_t}$

Goal: minimize regret

$$R_T = \max_A \sum_{t=1}^{T} \left(\mu_A - \mathbb{E}\left[\mu_{A_t}\right]\right)$$

The price paid by the learner for not knowing μ

Algorithm idea: optimism

- Estimate the mean of each arm
- Act as you are in the nicest possible world

Algorithm idea: optimism

- Estimate the mean of each arm
- Act as you are in the nicest possible world

- Nicest – we want mean to be large
- Possible – with high probability true mean is within the confidence interval

Action is either optimal or you learn something new.

Action is either optimal or you learn something new.

Upper Confidence Bound Algorithm

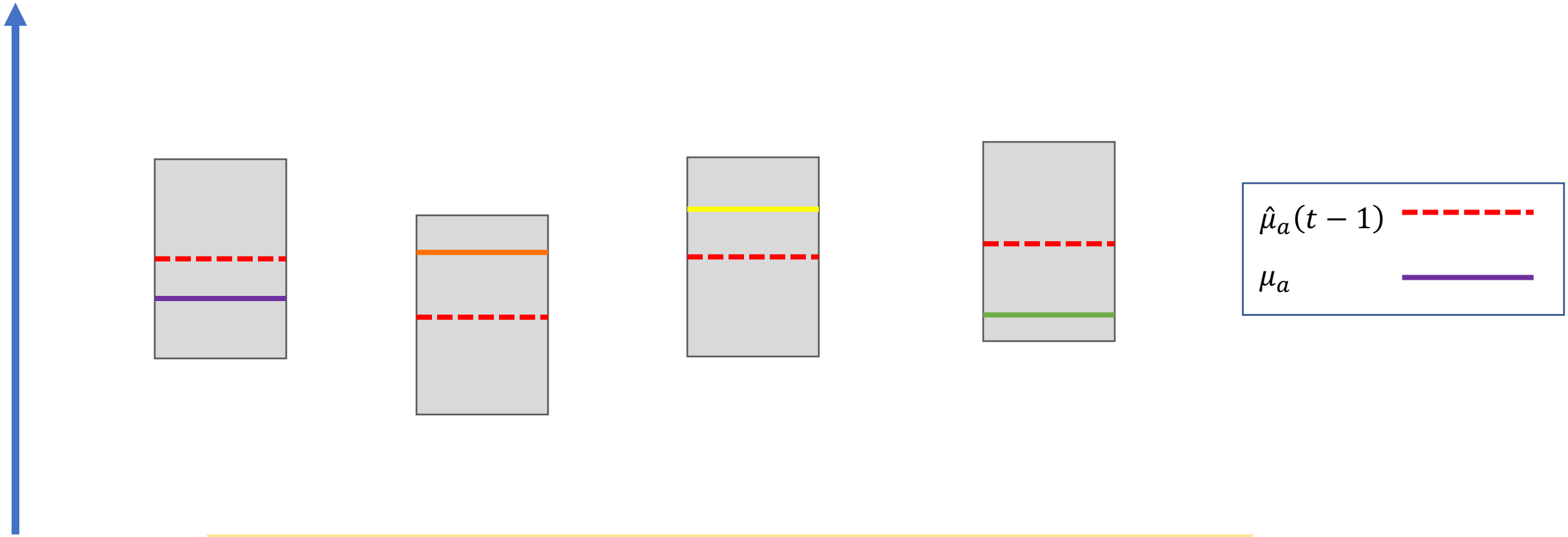
Choose each arm once and then

- $A_t = argmax_a \, \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$

$\hat{\mu}_a(t)$ = empirical mean
$N_a(t)$ = number of plays of arm a
δ = confidence level (we will use $1/T^2$)

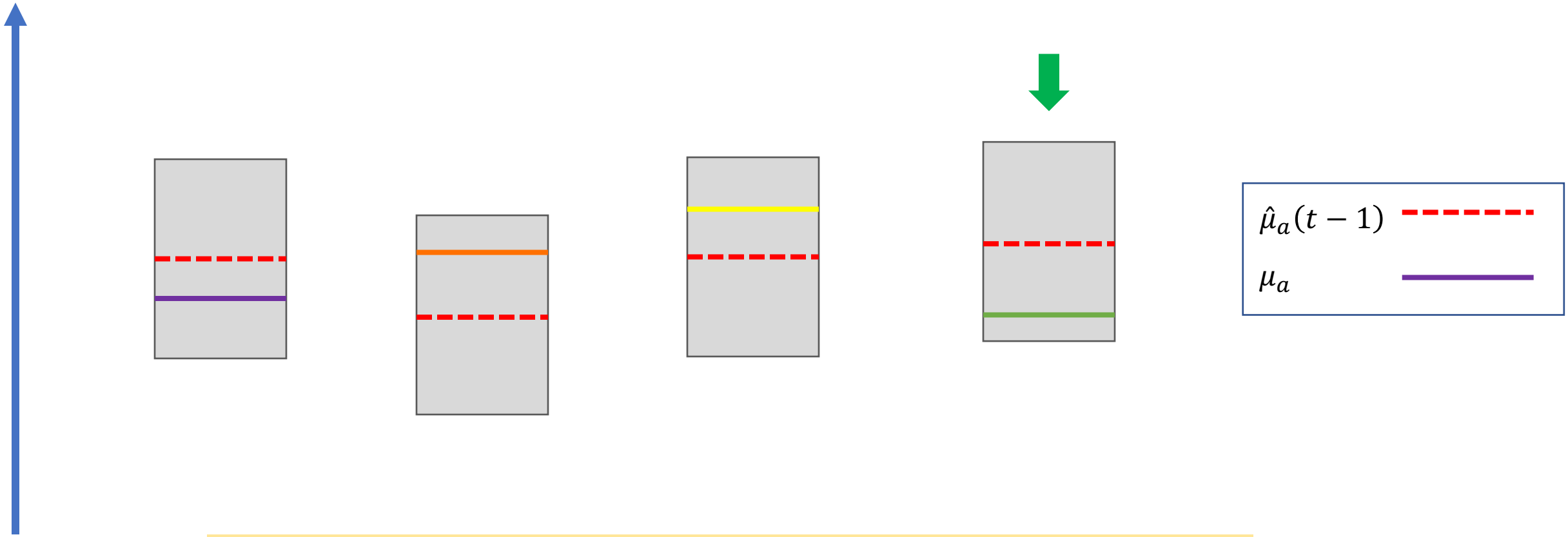
Assumption: distributions P_a- *1 sub-Gaussian* for all a: $\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\lambda^2/2}$

Upper Confidence Bound Algorithm



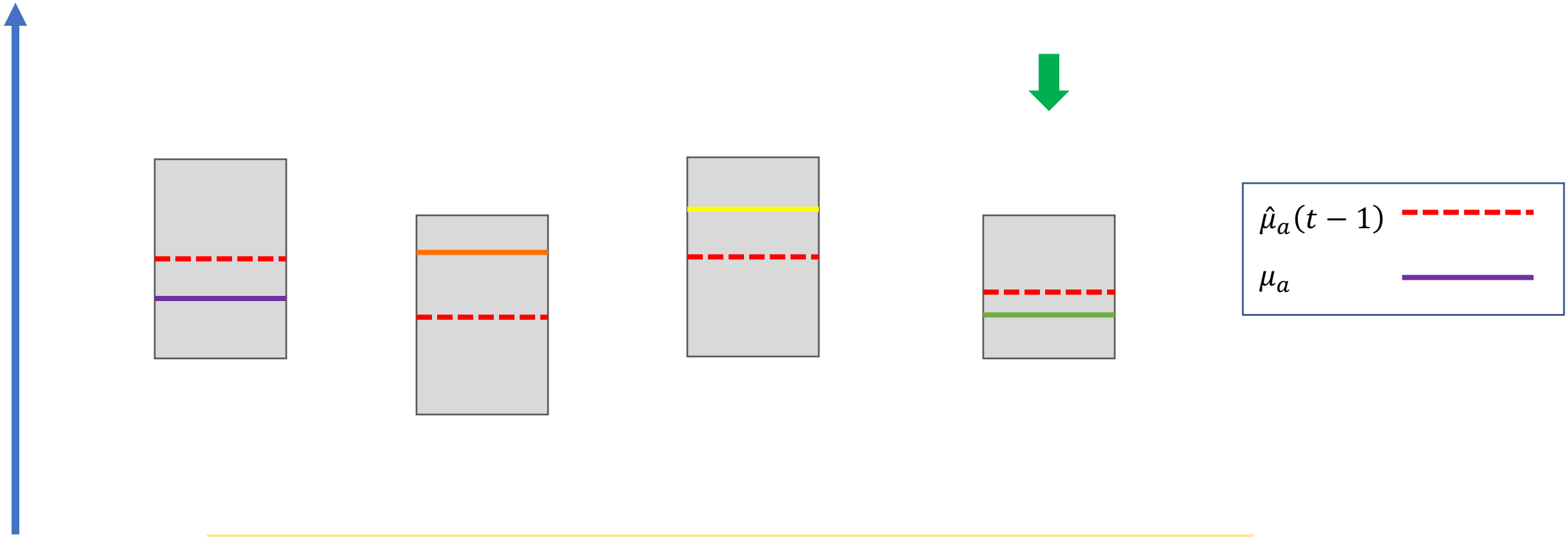
$$A_t = argmax_a\, \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$$

Upper Confidence Bound Algorithm



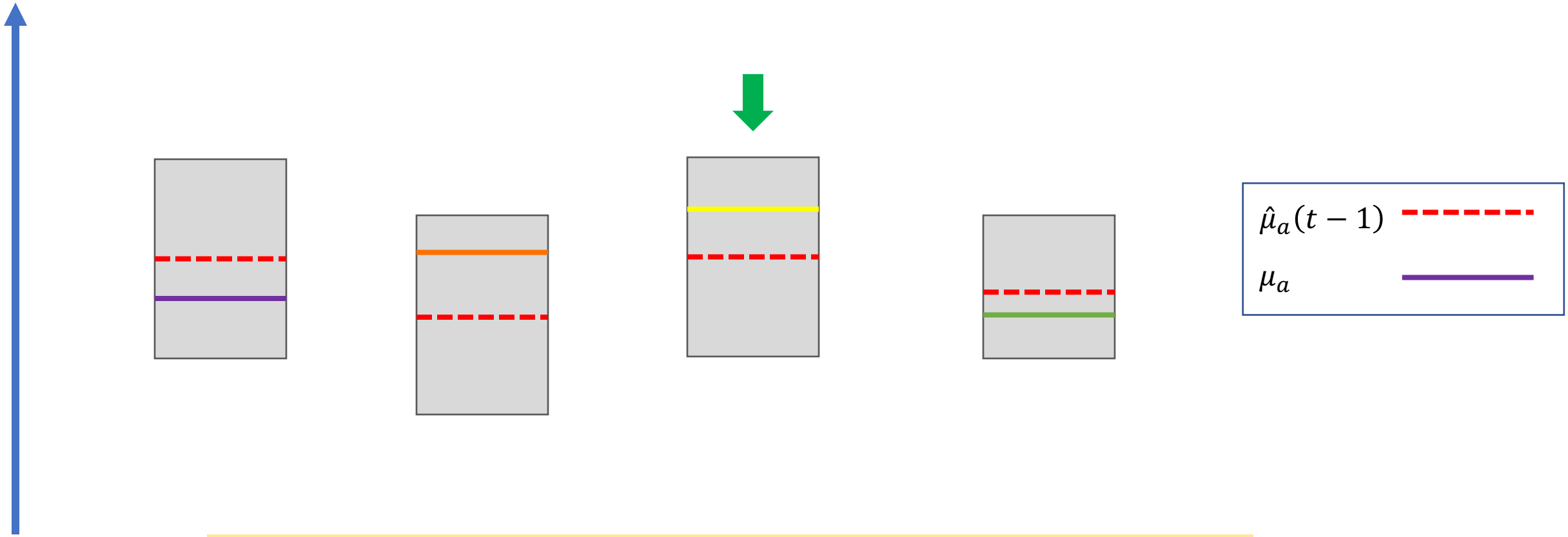
$$A_t = argmax_a\, \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$$

Upper Confidence Bound Algorithm



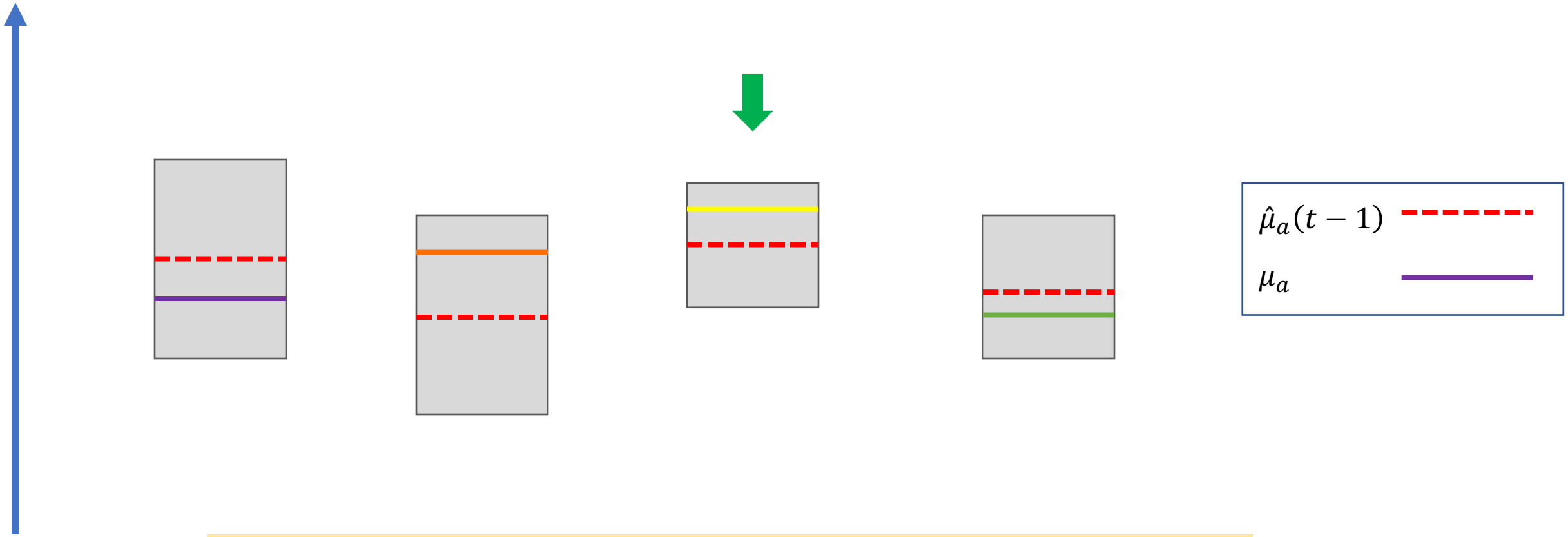
$$A_t = argmax_a\, \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$$

Upper Confidence Bound Algorithm



$$A_t = argmax_a\, \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$$

Upper Confidence Bound Algorithm



$$A_t = argmax_a\, \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$$

UCB analysis

- Denote $\Delta_a = \mu_{a^*} - \mu_a$

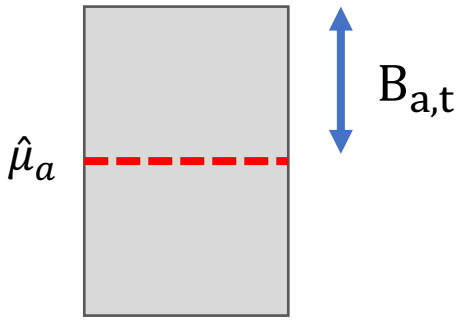
- $R_T = \sum_{t=1}^{T}\left(\mu_{a^*} - \mathbb{E}\left[\mu_{A_t}\right]\right) = \mathbb{E}\sum_{t=1}^{T}\sum_{a=1}^{K}\Delta_a\, I[A_t = a]$
$= \sum_{a=1}^{K}\Delta_a \mathbb{E}[N_a(T)]$

UCB analysis

- Denote $\Delta_a = \mu_{a^*} - \mu_a$

- $R_T = \sum_{t=1}^{T}\left(\mu_{a^*} - \mathbb{E}\left[\mu_{A_t}\right]\right) = \mathbb{E}\sum_{t=1}^{T}\sum_{a=1}^{K}\Delta_a\, I[A_t = a]$
$= \sum_{a=1}^{K}\Delta_a \mathbb{E}[N_a(T)]$

- Denote $M_a = \left\lceil \frac{16\log T}{\Delta_a^2} \right\rceil$ be a moment when $\mathrm{B_{a,t}} = \sqrt{\frac{4\log(T)}{N_a(t-1)}} \leq \frac{\Delta_a}{2}$

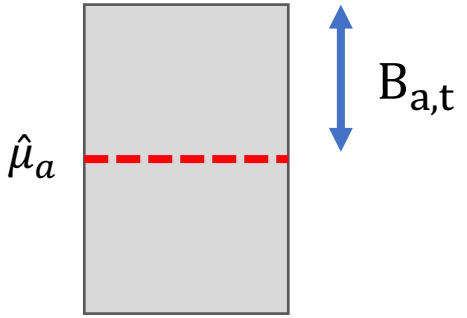


UCB analysis

- Denote $\Delta_a = \mu_{a^*} - \mu_a$

- $R_T = \sum_{t=1}^{T}\left(\mu_{a^*} - \mathbb{E}\left[\mu_{A_t}\right]\right) = \mathbb{E}\sum_{t=1}^{T}\sum_{a=1}^{K}\Delta_a\, I[A_t = a]$
$= \sum_{a=1}^{K}\Delta_a \mathbb{E}[N_a(T)]$

- Denote $M_a = \left\lceil \frac{16\log T}{\Delta_a^2} \right\rceil$ be a moment when $\mathrm{B_{a,t}} = \sqrt{\frac{4\log(T)}{N_a(t-1)}} \leq \frac{\Delta_a}{2}$

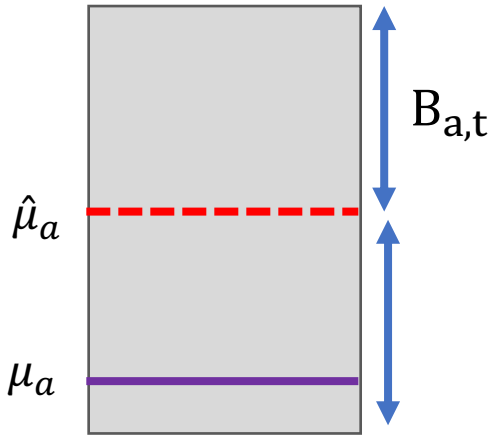


- $\mathbb{E}[N_a(T)] = \sum_{t=1}^{T}\mathbb{E}[I\{A_t = a\}]$
$\leq M_a + \sum_{t=M_a+1}^{T}\mathbb{E}[I\{A_t = a, N_a(t) > M_a + 1\}]$

UCB analysis

- $\mathrm{B_{a,t}} = \sqrt{\frac{4\log(T)}{N_a(t-1)}} \leq \frac{\Delta_a}{2}$
- $\{A_t = a^*, N_a(t) > M_a + 1\}$:

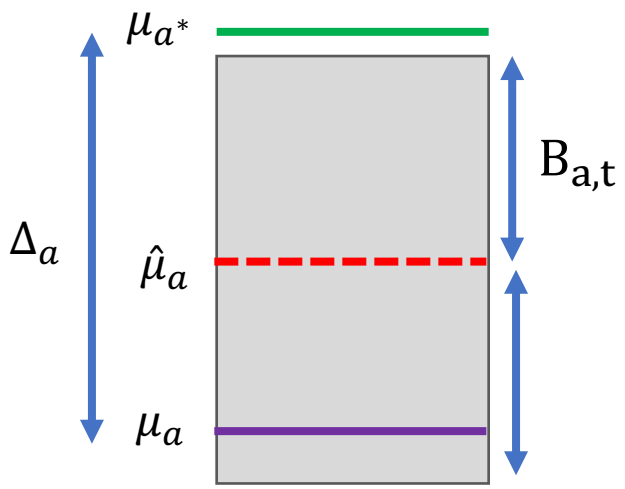
$$\hat{\mu}_{a,t} + B_{a,t} < \mu_a + 2B_{a,t} \leq \mu_a + \Delta_a$$



UCB analysis

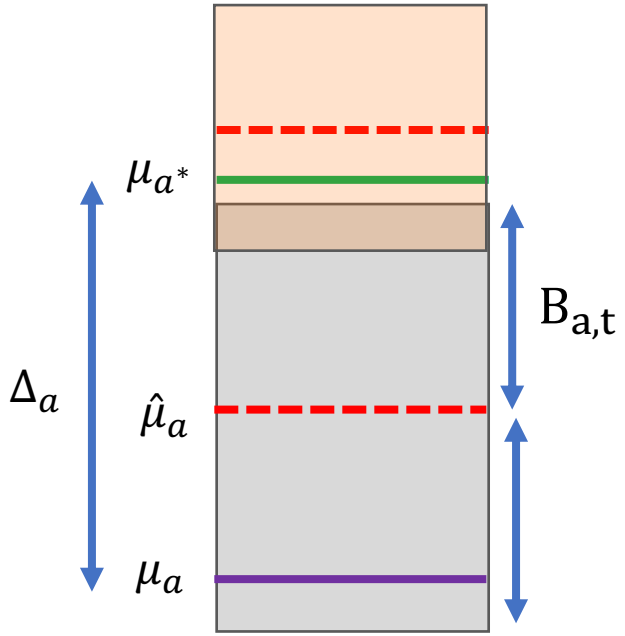
- $\mathrm{B_{a,t}} = \sqrt{\frac{4\log(T)}{N_a(t-1)}} \leq \frac{\Delta_a}{2}$
- $\{A_t = a^*, N_a(t) > M_a + 1\}$:

$$\hat{\mu}_{a,t} + B_{a,t} < \mu_a + 2B_{a,t} \leq \mu_a + \Delta_a = \mu_{a^*}$$



UCB analysis

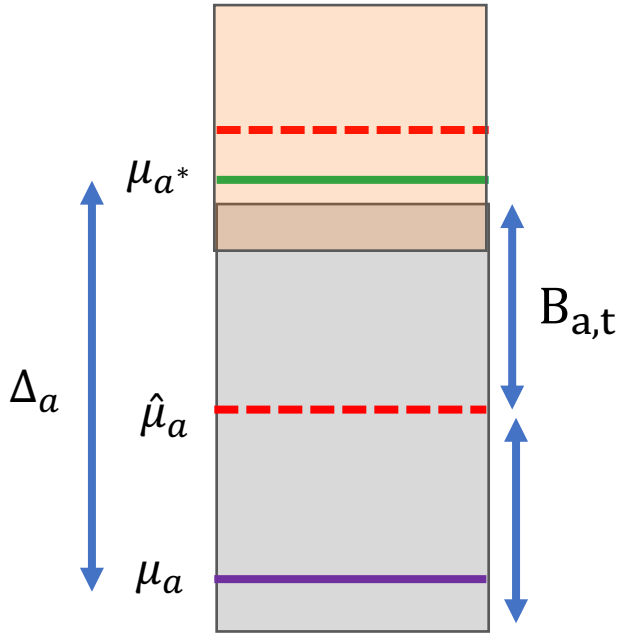
- $\mathrm{B_{a,t}} = \sqrt{\frac{4\log(T)}{N_a(t-1)}} \leq \frac{\Delta_a}{2}$
- $\{A_t = a^*, N_a(t) > M_a + 1\}$:



$$\hat{\mu}_{a,t} + B_{a,t} < \mu_a + 2B_{a,t} \leq \mu_a + \Delta_a = \mu_{a^*} \leq \hat{\mu}_{a^*,t} + B_{a^*,t}$$

UCB analysis

- $\mathrm{B_{a,t}} = \sqrt{\frac{4\log(T)}{N_a(t-1)}} \leq \frac{\Delta_a}{2}$
- $\{A_t = a^*, N_a(t) > M_a + 1\}$:



$$\hat{\mu}_{a,t} + B_{a,t} < \mu_a + 2B_{a,t} \leq \mu_a + \Delta_a = \mu_{a^*} \leq \hat{\mu}_{a^*,t} + B_{a^*,t}$$

If $\{A_t = a \ , N_a(t) > M_a\}$, at least one of two inequalities must be wrong, so

$$P(\{A_t = a, N_a(t) > M_a + 1\})$$
$$\leq P\left(\hat{\mu}_{a,t} > \mu_a + B_{a,t}\right) + P\left(\hat{\mu}_{a^*,t} + B_{a^*,t} \leq \mu_{a^*}\right)$$

- Using the Hoeffding inequality,

$$P\left(\hat{\mu}_{a,t} > \mu_a + B_{a,t}\right) \leq P\left(\exists s \leq t : \hat{\mu}_{a,s} - \mu_a > \sqrt{\frac{4\log(T)}{N_a(s)}}\right)$$

$$\leq \sum_{s=1}^{t} P\left(\hat{\mu}_{a,s} - \mu_a > \sqrt{\frac{4\log(T)}{N_a(s)}}\right)$$

$$\leq \sum_{s=1}^{t} \exp\left\{-\frac{2\,s\log(T)}{N_a(s)}\right\} \leq \sum_{s=1}^{t} T^{-2} \leq \frac{1}{T}$$

- By collecting all the terms, we get

$$R_T = \sum_{t=1}^{T}\left(\mu_{A^*} - \mathbb{E}\left[\mu_{A_t}\right]\right) = \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_a(T)] \leq \sum_{a} \Delta_a \left(\frac{8\log T}{\Delta_a^2} + 2\,T \cdot \frac{1}{T}\right)$$

UCB regret bound

Theorem

The UCB algorithm on 1-sub-Gaussian data has

$$R_T \leq \sum_a \frac{16 \log T}{\Delta_a} + 2\,\Delta_a$$

If Δ_a are small:

$$R_T \leq \sum_{a:\Delta_a \leq \Delta} \Delta T + \sum_{a:\Delta_a > \Delta} \frac{16 \log T}{\Delta_a} + 2\,\Delta_a = O(\sqrt{TK log\, T})$$

Thomson sampling

- Suppose reward for arm a is $r_a(t) \sim N(\mu_a, 1)$
- Prior distribution $P_a(\mu) = N(\mu_0, 1)$
- History of observations $\mathcal{D}_t = \left\{\left(A_1, r_{A_1}(1)\right), \ldots, \left(A_{t-1}, r_{A_{t-1}}(t-1)\right)\right\}$
- Posterior distribution $P_{t,a}(\mu|\mathcal{D}_t) \propto P_t(\mathcal{D}_t|\mu)P(\mu)$, which is $N(\hat{\mu}_a(t), \frac{1}{N_a(t)+1})$

Algorithm:

- Draw $\mu_{a,t} \sim P_{t,a}$ for all a
- Choose $A_t = argmax_a \mu_{a,t}$
- Observe $r_t(A_t)$
- Update the posterior P_{t+1,A_t}

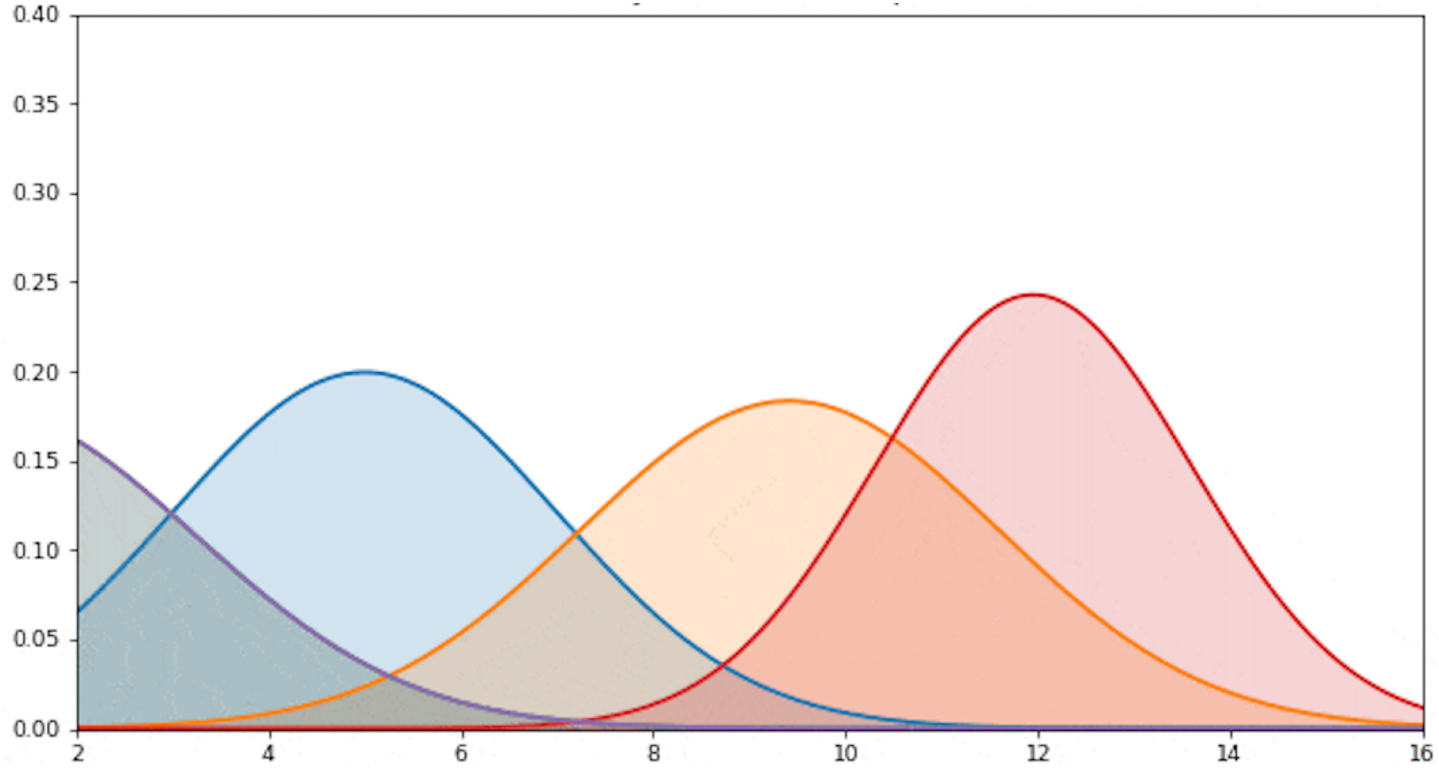
Thomson sampling

- Suppose reward for arm a is 1-sub Gaussian
- Prior distribution $P_a(\mu) = N(\mu_0, 1)$
- History of observations $\mathcal{D}_t = \left\{\left(A_1, r_{A_1}(1)\right), \ldots, \left(A_{t-1}, r_{A_{t-1}}(t-1)\right)\right\}$
- Posterior distribution $P_{t,a}(\mu|\mathcal{D}_t) \propto P_t(\mathcal{D}_t|\mu)P(\mu)$, which is $N(\hat{\mu}_a(t), \frac{1}{N_a(t)+1})$

Algorithm:

- Draw $\mu_{a,t} \sim P_{t,a}$ for all a
- Choose $A_t = argmax_a \mu_{a,t}$
- Observe $r_t(A_t)$
- Update the posterior P_{t+1,A_t}

Thomson sampling



- Arms with small $N_{a,t}$ implies a wide posterior, so a good probability of being selected
- Empirically perform better than UCB

Contextual bandits

Example: Ad placement

Repeat:

- Website is visited by a **user** (with cookies, profile, etc),
- website chooses **ad** to present to user,
- user **responds** (clicks, leaves, ..)

- **Goal:** maximize clicks

Contextual Bandits

In each round $t = 1, 2, \ldots, T$

- Nature reveals the context $X_t \in \mathbb{R}^d$
- the learner chooses action $A_t \in [K]$
- the learner observed the reward $r_t(X_t, A_t)$

Goal: minimize regret

$$R_T(\pi) = \mathbb{E} \sum_t \left(r_t\left(X_t, \pi(X_t)\right) - r_t(X_t, A_t) \right)$$

against best policy $\pi: \mathcal{X} \to [K]$

Bandits with expert advice

Idea:

- M policies, K actions
- Each policy m produces the “advise” $\pi_{t,m} \in \Delta^K$
- The algorithm learns which policy to trust

Exp4 algorithm

For $t = 1,2,\ldots,T$:

- Get expert advise $\pi_{t,1},\ldots,\pi_{t,M} \in \Delta^K$
- Draw $A_t \sim p_t$, where $p_{a,t} = \sum_{m=1}^{M} q_{t,m}\pi_{t,m}$
- Compute estimated losses for actions: $\hat{\ell}_{a,t} = \frac{\ell_{a,t}}{p_{a,t}} \mathbb{I}[A_t = a]$
- Compute estimated losses for policies $\hat{y}_{t,m} = \langle \pi_{t,m}, \hat{\ell}_t \rangle$
- Compute $q_{t+1,m} = \frac{\exp(-\eta \sum_{s=1}^{t} \hat{y}_{s,m})}{\sum_{m=1}^{M} \exp(-\eta \sum_{s=1}^{t} \hat{y}_{s,m})}$

Exp4: theoretical guarantees

Theorem:

EXP4 with $\eta = \sqrt{\frac{2 \log M}{TK}}$ gives $R_T \leq \sqrt{2TKlogM}$

What if we want to compare to the infinite family of policies?

Need to make some assumptions on the loss!

Linear contextual bandits

For $t = 1,2, \dots, T$,

- The learner observes the context c_t,
- The learner picks action $A_t \in \{1, \dots, K\}$,
- The learner receives reward $X_t = \langle c_t, \theta_a \rangle + \xi_{t,a}$.

Regret is defined w.r.t. an agent that knows true $\theta_1, \dots, \theta_K$:

$$R_T = \mathbb{E} \sum_t \left(\max_a \langle c_t, \theta_a \rangle - \langle c_t, \theta_{A_t} \rangle \right)$$

Confidence set

- Reward of action a:

$$X_t = \langle c_t, \theta_a \rangle + \xi_{t,a}$$

- Estimate of θ_a: regularized least squares estimator

$$\hat{\theta}_{t,a} = V_{t,a}^{-1} \left(\sum_{s=1}^{t} c_s X_{t,a} \, I[A_s = a] \right)$$

$$V_{t,a} = \sum_{s=1}^{t} c_s c_s^T I[A_s = a] + \lambda I$$

- Confidence ellipsoids with radius $r_{t,a} = \sqrt{\log \frac{\det(V_{t,a})}{\delta^2 \lambda^d}}$:

$$B_{t,a} = \{\theta : (\theta - \hat{\theta}_{t,a})^T V_{t,a} (\theta - \hat{\theta}_{t,a}) \leq r_{t,a}\}$$

LinUCB algorithm

For $t = 1,2, \ldots, T$:

- Receive context c_t,
- Choose $A_t = argmax_a \max_{\theta \in B_{t-1,a}} \langle \theta, c_t \rangle$ (optimism)
- Observe $X_{t,A_t} = \langle c_t, \theta_{A_t} \rangle + \xi_{t,A_t}$
- Update $B_{t,a} = \{\theta : (\theta - \hat{\theta}_{t,a})^T V_{t,a} (\theta - \hat{\theta}_{t,a}) \leq r_{t,a}\}$

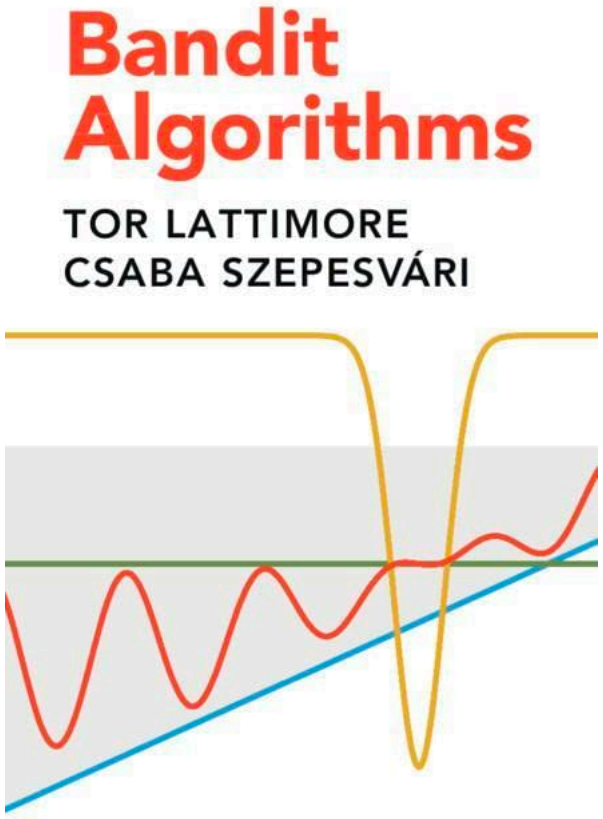
$B_{t,a}$ are confidence ellipsoids with $P(\forall t > 0 : \theta_{t,a} \in B_{t,a}) \geq 1 - \delta$

Review

- Setting: adversarial bandits
  - Online linear optimization
  - Exp3
- Setting: stochastic bandits
  - UCB
  - Thomson sampling
- Setting: Contextual bandits
  - Exp4
  - LinUCB - linear contextual bandits

Review

- Setting: adversarial bandits
  - Online linear optimization
  - Exp3
- Setting: stochastic bandits
  - UCB
  - Thomson sampling
- Setting: Contextual bandits
  - Exp4
  - LinUCB - linear contextual bandits



Thanks!

Questions?

julia.olkhovskaya@gmail.com