Course Content

Part I. Learning in Bandits

- Multi-armed bandits
- Linear bandits
- Contextual bandits
- Adversarial multi-armed bandits
- Adversarial linear bandits

Part II. Basics of MDPs

- Bellman (optimality) equations
- Value iteration
- Policy iteration

Part III. Learning in MDPs

- Approximate value iteration and variants
  - Least-square value iteration
  - Q-Learning
  - DQN
- Policy evaluation
  - Temporal difference
  - Monte Carlo
- Approximate policy iteration and variants
  - Least-square policy iteration
  - (Natural) policy gradient and actor-critic
  - REINFORCE, A2C, PPO
  - DDPG, SAC

Part IV. Offline RL

Student Project Presentation

Course Content

Part I. Learning in Bandits

- Multi-armed bandits
- Linear bandits
- Contextual bandits
- Adversarial multi-armed bandits
- Adversarial linear bandits

Part II. Basics of MDPs

- Bellman (optimality) equations
- Value iteration
- Policy iteration

Part III. Learning in MDPs

- Approximate value iteration and variants
  - Least-square value iteration
  - Q-Learning
  - DQN
- Policy evaluation
  - Temporal difference
  - Monte Carlo
- Approximate policy iteration and variants
  - Least-square policy iteration
  - (Natural) policy gradient and actor-critic
  - REINFORCE, A2C, PPO
  - DDPG, SAC

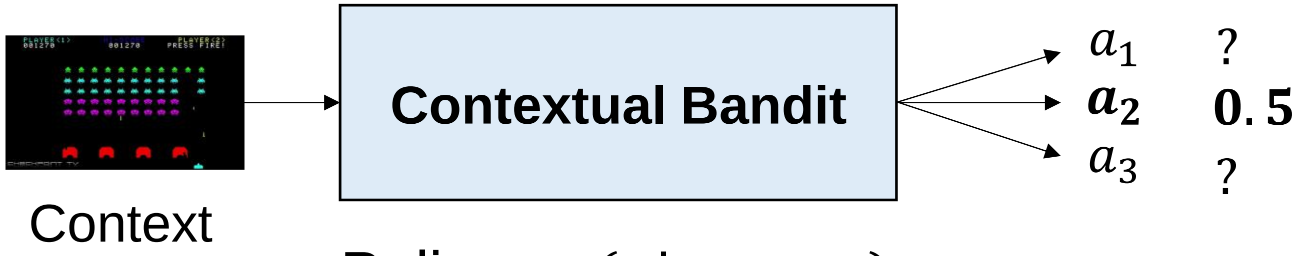
Part IV. Offline RL

Student Project Presentation

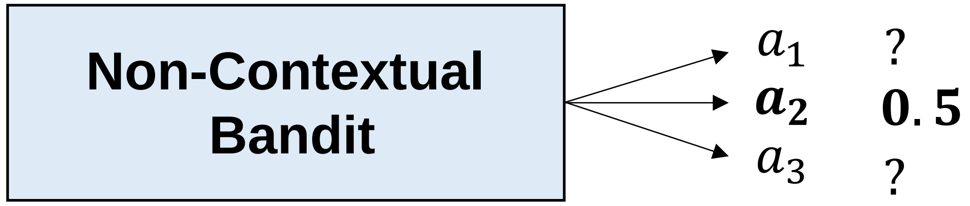
Bandits

Chen-Yu Wei

Contextual Bandits and Non-Contextual Bandits



Policy: $\pi(a|\text{context})$



Policy: $\pi(a)$

Multi-Armed Bandits

Multi-Armed Bandits

A slot machine

One-armed bandit

A row of slot machines

Multi-armed bandit

Multi-Armed Bandits

Given: arm set $\mathcal{A} = \{1, \dots, A\}$

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

Learner chooses an arm $a_t \in \mathcal{A}$

Learner observes $r_t = R(a_t) + w_t$

Arm = Action

Assumption: $R(a)$ is the (hidden) ground-truth reward function

w_t is a zero-mean noise

Goal: maximize the total reward $\sum_{t=1}^{T} R(a_t)$ (or $\sum_{t=1}^{T} r_t$)

How to Evaluate an Algorithm's Performance?

- "My algorithm obtains $0.3T$ total reward within T rounds"
  – Is my algorithm good or bad?

$\Rightarrow \max_a R(a) - \frac{1}{T}\sum_{t=1}^{T} R(a_t) \leq \frac{1}{\sqrt{T}}$

- Benchmarking the problem

$\leq \sqrt{T}$

$$\text{Regret} := \underbrace{\max_{\pi} \sum_{t=1}^{T} R(\pi)}_{\text{The total reward of the best policy}} - \sum_{t=1}^{T} R(a_t) \quad \underset{\text{In MAB}}{=} \quad \max_a TR(a) - \sum_{t=1}^{T} R(a_t)$$

- "My algorithm ensures $\text{Regret} \leq 5T^{\frac{3}{4}}$ "
- $\text{Regret} = o(T) \;\Rightarrow\;$ the algorithm is as good as the optimal policy asymptotically

Multi-Armed Bandits

- Key challenge: Exploration
- The other three challenges we will discuss for RL
  - Generalization (there is no input in MAB)
  - Temporal credit assignments (there is no delayed feedback)
  - Distribution mismatch (there is no pre-collected data)

- We will discuss about three categories of exploration strategies
  - Non-adaptive
  - Mean-adaptive
  - (Mean & Uncertainty)-adaptive

Multi-Armed Bandits

Non-Adaptive Exploration

The Exploration and Exploitation Trade-off in MAB

- To perform as well as the best policy (i.e., best arm) asymptotically, the learner has to pull the best arm most of the time

  ⇒ need to **exploit**

- To identify the best arm, the learner has to try every arm sufficiently many times

  ⇒ need to **explore**

A Simple Strategy: Explore-then-Exploit

Explore-then-exploit (Parameter: T_0)

In the first T_0 rounds, sample each arm T_0/A times. **(Explore)**

Compute the **empirical mean** $\hat{R}(a)$ for each arm a

In the remaining $T - T_0$ rounds, draw $\hat{a} = \mathrm{argmax}_a \hat{R}(a)$ **(Exploit)**

What is the *right* amount of exploration (T_0)?

Another Simple Strategy: ϵ-Greedy

Mixing exploration and exploitation in time

ϵ-Greedy (Parameter: ϵ)

In the first A rounds, draw each arm once.

In the remaining rounds $t > A$,

Take action

$$a_t = \begin{cases} \text{uniform}(\mathcal{A}) & \text{with prob. } \epsilon & \textbf{(Explore)} \\ \text{argmax}_a \hat{R}_t(a) & \text{with prob. } 1-\epsilon & \textbf{(Exploit)} \end{cases}$$

where $\hat{R}_t(a) = \frac{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}\, r_s}{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}}$ is the empirical mean of arm a using samples up to time $t-1$.

Comparison

- ϵ-Greedy is more **robust to non-stationarity** than Explore-then-Exploit

Multi-Armed Bandits

Mathematical analysis for Explore-then-Exploit & ϵ-Greedy

Quantifying the Estimation Error

In the exploration phase, we obtain $N = T_0/A$ i.i.d. samples of each arm.

Key Question:

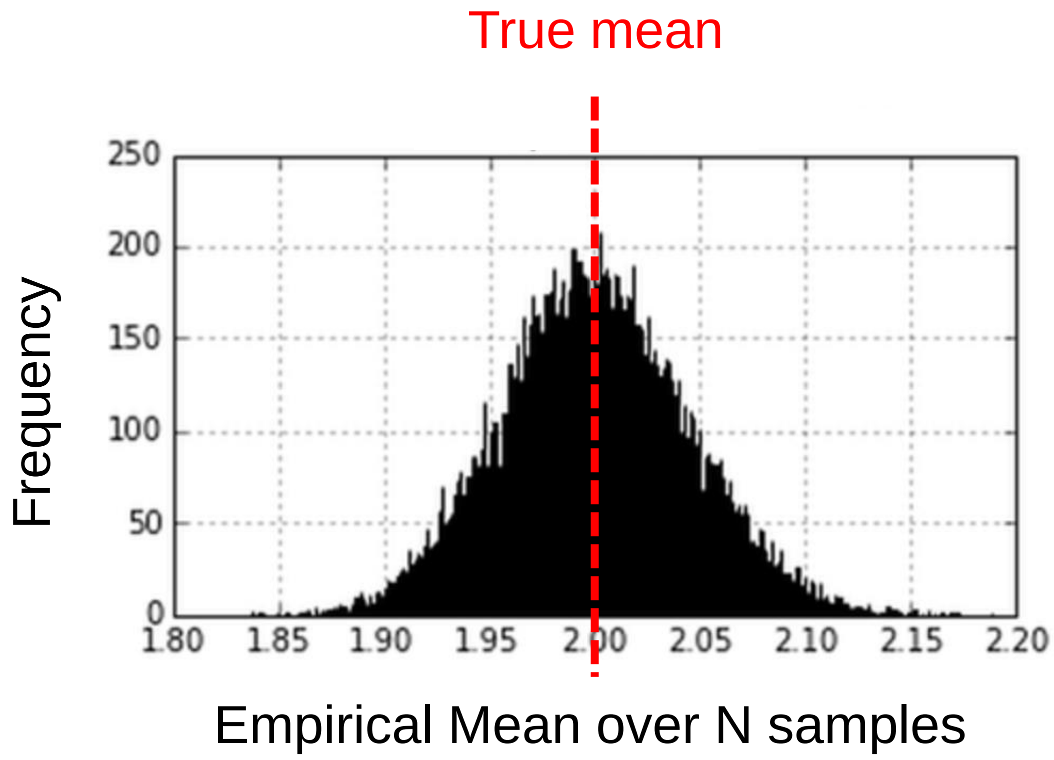
$$\left| \hat{R}(a) - R(a) \right| \leq ?\ f(N)$$

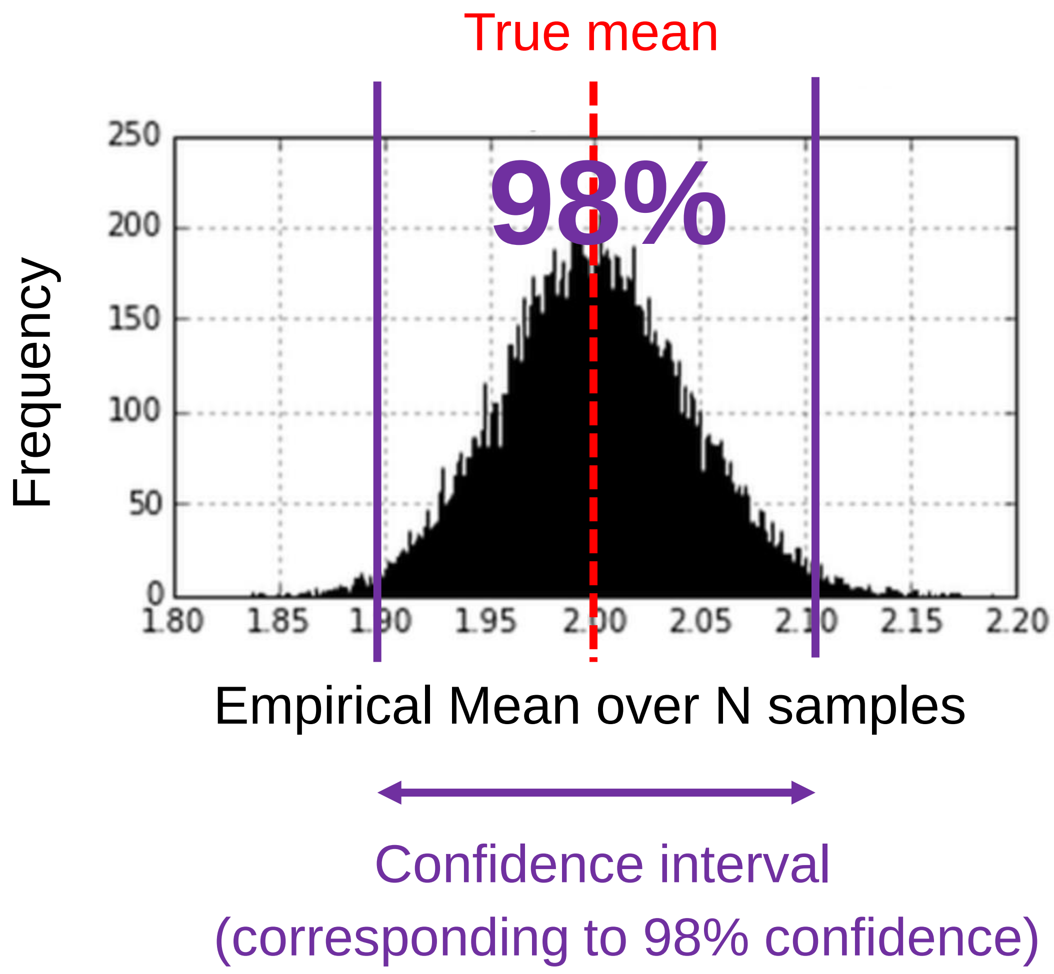
some decreasing function of N

Empirical mean of N i.i.d. samples

True mean

Quantifying the Estimation Error





Quantifying the Estimation Error

In the exploration phase, we obtain $N = T_0/A$ i.i.d. samples of each arm.

Key Question:

$$\left| \hat{R}(a) - R(a) \right| \leq ? \ f(N)$$

Empirical mean of N i.i.d. samples

True mean

some decreasing function of N

Quantifying the Estimation Error

In the exploration phase, we obtain $N = T_0/A$ i.i.d. samples of each arm.

Key Question:

With probability at least $1 - \delta$, $\quad \delta = 0.02$, $= 0.98$

$$\left| \hat{R}(a) - R(a) \right| \leq ? \;\; f(N, \delta)$$

some decreasing function of N

Empirical mean of N i.i.d. samples

True mean

Quantifying the Error: Concentration Inequality

Theorem. Hoeffding's Inequality

Let $X_1, \dots, X_N$ be independent σ-**sub-Gaussian** random variables.

Then with probability at least $1 - \delta$, $\mathbb{E}[X]$

$\delta = 0.0001$

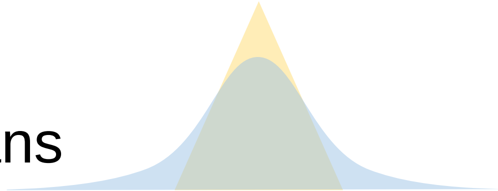
$$\left| \frac{1}{N} \sum_{i=1}^{N} X_i - \overbrace{\cancel{\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[X_i]}}^{\mathbb{E}[X]} \right| \leq \sigma \sqrt{\frac{2 \log(2/\delta)}{N}} \;.$$

A random variable is called σ-sub-Gaussian if $\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq e^{\lambda^2 \sigma^2 / 2} \quad \forall \lambda \in \mathbb{R}$.

Fact 1. $\mathcal{N}(\mu, \sigma^2)$ is σ-sub-Gaussian.

Fact 2. A random variable $\in [a, b]$ is $(b - a)$-sub-Gaussian.

Intuition: tail probability $\Pr\{|X - \mathbb{E}[X]| \geq z\}$ bounded by that of Gaussians



Quantifying the Estimation Error

With probability at least $1-\delta$, $\left|\hat{R}(a)-R(a)\right| = O\left(\sqrt{\frac{\log(1/\delta)}{N}}\right)$

Omit constants

With high probability, $\left|\hat{R}(a)-R(a)\right| = \tilde{O}\left(\sqrt{\frac{1}{N}}\right)$

$\left|\hat{R}(a)-R(a)\right| \lesssim \sqrt{\frac{1}{N}}$

Omit constants and $\log(1/\delta)$ factors

Explore-then-Exploit Regret Bound Analysis

In the first T_0 rounds, sample each arm T_0/A times.

Compute the **empirical mean** $\hat{R}(a)$ for each arm a

In the remaining $T - T_0$ rounds, draw $\hat{a} = \operatorname{argmax}_a \hat{R}(a)$

$a^* = \operatorname{argmax}_a R(a)$ (true best arm)

After the exploration phase, we have $\boxed{|\hat{R}(a) - R(a)| \lesssim \sqrt{\frac{1}{N}} = \sqrt{\frac{A}{T_0}}}$

In the exploitation phase,

At any time $t \in$ exploitation phase, $R(a^*) - R(\hat{a})$

$$= \underbrace{\hat{R}(a^*) - \hat{R}(\hat{a})}_{\leq 0} + \underbrace{\left[R(a^*) - \hat{R}(a^*)\right]}_{\lesssim \sqrt{\frac{A}{T_0}}} + \underbrace{\left(\hat{R}(\hat{a}) - R(\hat{a})\right)}_{\sqrt{\frac{A}{T_0}}}$$

Regret $\lesssim$ cost of exploration $+ \sum_{t \in \text{second phase}} \left(R(a^*) - R(\hat{a})\right) \lesssim T_0 + (T - T_0) \cdot 2\sqrt{\frac{A}{T_0}}$

Regret Bound of Explore-then-Exploit and ϵ-Greedy

Theorem. Regret Bound of Explore-then-Exploit

Suppose that $R(a) \in [0,1]$ and w_t is 1-sub-Gaussian.

Then Explore-then-Exploit ensures with high probability.

$$\text{Regret} \lesssim T_0 + T\sqrt{\frac{A}{T_0}} \approx A^{1/3}T^{2/3} \qquad \left(T_0^* \approx A^{1/3}T^{2/3}\right)$$

Theorem. Regret Bound of ϵ-Greedy (Your Exercise)

Suppose that $R(a) \in [0,1]$ and w_t is 1-sub-Gaussian.

Then ϵ-Greedy ensures

$$\text{Regret} \lesssim \epsilon T + \sqrt{\frac{AT}{\epsilon}} \approx A^{1/3}T^{2/3}$$

Can We Do Better?

→ every arm receives the same amount of exploration

In explore-then-exploit and ϵ-greedy, our exploration strategy is **non-adaptive.**

… Maybe, for those arms that look worse, the amount of exploration on them can be reduced?

One Solution: Refine the amount of exploration for each arm **based on the current mean estimation**.

(Has to do this carefully to avoid **under-exploration**)

Multi-Armed Bandits

Mean-Adaptive Exploration

Mean-Adaptive Exploration

Boltzmann Exploration (Parameter: λ_t)

In each round, sample a_t according to

$$p_t(a) \propto \exp\big(\lambda_t \hat{R}_t(a)\big)$$

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

Inverse Gap Weighting (Parameter: λ_t)

$$p_t(a) = \frac{1}{\gamma_t - \lambda_t \hat{R}_t(a)}$$

γ_t is a normalization factor that makes $\sum_a p_t(a) = 1$

Mean-Adaptive Exploration

- Boltzmann Exploration
  - A quite commonly used exploration strategy (like ϵ-greedy)
  - There is no good regret bound we can prove
  - There are bad examples where it suffers from **under-exploration**

  Cesa-Bianchi, Gentile, Lugosi, Neu. **Boltzmann Exploration Done Right**, 2017.

  Bian and Jun. **Maillard Sampling: Boltzmann Exploration Done Optimally**. 2021.

- Inverse Gap Weighting
  - Not very well-known
  - We can show a regret bound for it (we'll do this when talking about contextual bandits)

  Foster and Rakhlin. **Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles**

Multi-Armed Bandits

(Mean and Uncertainty)-Adaptive Exploration

Another Idea: “Optimism in the Face of Uncertainty”

In words:

Act according to the **best plausible world**.

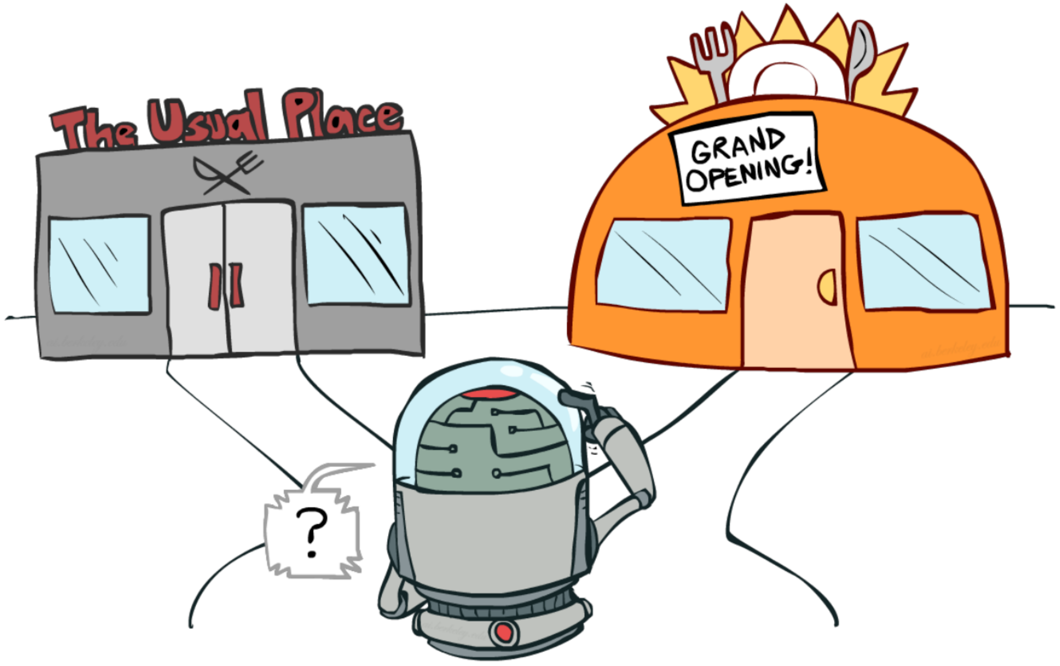


Image source: UC Berkeley CS188

Another Idea: "Optimism in the Face of Uncertainty"

In words:

Act according to the **best plausible world**.

At time t, suppose that arm a has been drawn for $N_t(a)$ times, with empirical mean $\hat{R}_t(a)$.

What can we say about the true mean $R(a)$?

$$\left| R(a) - \hat{R}_t(a) \right| \leq \sqrt{\frac{2\log(2/\delta)}{N_t(a)}} \quad \text{w.p.} \geq 1 - \delta$$

What's the most optimistic mean estimation for arm a?

$$\hat{R}_t(a) + \sqrt{\frac{2\log(2/\delta)}{N_t(a)}}$$

Upper Confidence Bound (UCB)

UCB (Parameter: δ)

In the first A rounds, draw each arm once.

For the remaining rounds: in round t, draw

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

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

$N_t(a)$ is the number of samples of arm a up to time $t-1$.

P Auer, N Cesa-Bianchi, P Fischer. **Finite-time analysis of the multiarmed bandit problem**, 2002.

Regret Bound of UCB

Theorem. Regret Bound of UCB

UCB ensures with high probability,

$$\text{Regret} \lesssim \sqrt{AT} \; .$$

UCB Regret Bound Analysis

Visualizing UCB

True mean: [0.2, 0.4, 0.6, 0.7]

Multi-Armed Bandits

Brief Summary for Exploration Strategies

Summary: Exploration

$\hat{R}_t(a)$: mean estimation for arm a at time t

$N_t(a)$: number of samples for arm a at time t

Explore-then-Exploit

$$a_t = \begin{cases} \mathrm{uniform}(\mathcal{A}) & t \le T_0 \\ \mathrm{argmax}_a \hat{R}_{T_0}(a) & t > T_0 \end{cases}$$

ϵ-Greedy

$$a_t = \begin{cases} \mathrm{uniform}(\mathcal{A}) & \text{with prob. } \epsilon \\ \mathrm{argmax}_a \hat{R}_t(a) & \text{with prob. } 1-\epsilon \end{cases}$$

Boltzmann Exploration

$$p_t(a) \propto \exp\big(\lambda_t \hat{R}_t(a)\big)$$

Inverse Gap Weighting

$$p_t(a) = \frac{1}{\gamma_t - \lambda_t \hat{R}_t(a)}$$

UCB

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

Summary: Exploration

	Regret Bound	Exploration
Explore-then-Exploit ϵ-Greedy	$A^{1/3}\,T^{2/3}$	Non-adaptive
Boltzmann Exploration Inverse Gap Weighting	None for BE $\sqrt{AT}$ for IGW	Mean-adaptive
Upper Confidence Bound Thompson Sampling	$\sqrt{AT}$	(Mean and uncertain)-adaptive

Bayesian Setting for MAB

Assumptions:

- At the beginning, the environment draws a parameter $\theta^\star$ from some prior distribution $\theta^\star \sim P_{\text{prior}}$
- In every round, the reward vector $\boldsymbol{r_t} = (r_t(1), \ldots, r_t(A))$ is generated from $\boldsymbol{r_t} \sim P_{\theta^\star}$

E.g., Gaussian Case

- At the beginning, $\theta^\star(a) \sim \mathcal{N}(0, 1)$ for all $a \in \{1, \ldots, A\}$.
- In every round, the reward of arm a is generated by $r_t(a) \sim \mathcal{N}(\theta^\star(a), 1)$.

For the learner, P_{prior} is known; $\theta^\star$ is unknown; P_θ is known for any θ.

Thompson Sampling

William Thompson. **On the likelihood that one unknown probability exceeds another in view of the evidence of two samples**, 1933.

In words:

Randomly pick an arm according to the probability you **believe** it is the optimal arm.

At time t, after seeing $\mathcal{H}_t = (a_1, r_1(a_1), a_2, r_2(a_2), \ldots, a_{t-1}, r_{t-1}(a_{t-1}))$, the learner has a **posterior distribution** for $\theta^\star$:

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

In math:

Sample a_t according to $p_t(a) = \int_\theta P(\theta|\mathcal{H}_t)\, \mathbb{I}\{a^\star(\theta) = a\} = \mathbb{E}_{\theta \sim P(\cdot | \mathcal{H}_t)}[\mathbb{I}\{a^\star(\theta) = a\}]$

Implementation: Sample $\theta_t \sim P(\cdot \mid \mathcal{H}_t)$, and choose $a_t = a^\star(\theta_t)$.

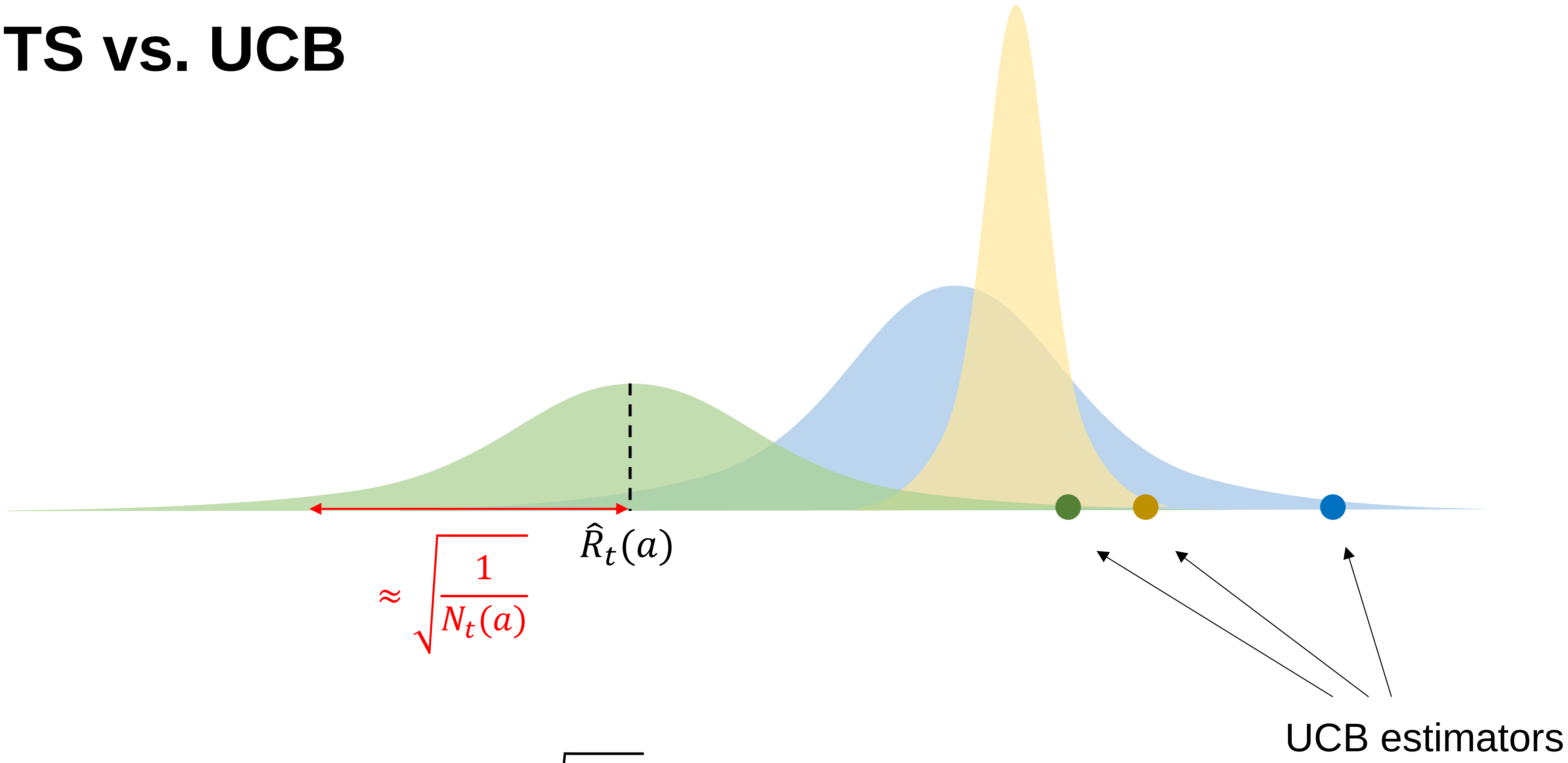
Gaussian Thompson Sampling

Gaussian prior $\theta^\star(a) \sim \mathcal{N}(0,1)$ + **Gaussian reward** $r_t(a) \sim \mathcal{N}(\theta^\star(a), 1)$:

$$P(\theta^\star(a) = \theta(a) \mid \mathcal{H}_t) = \mathcal{N}\left(\hat{R}_t(a), \frac{1}{N_t(a)+1}\right) \quad \text{where } \hat{R}_t(a) = \frac{\sum_{s=1}^{t-1} \mathbb{I}\{a_t = a\} r_t(a)}{N_t(a)+1}$$

Empirical mean assuming 1 fake sample with reward 0

TS vs. UCB



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

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

More on Thompson Sampling

For **Bernoulli** reward, we assume the **Beta** prior:
https://gdmarmerola.github.io//ts-for-bernoulli-bandit/

Regret bound analysis for Thompson sampling

Agrawal and Goyal. Near-optimal Regret Bounds for Thompson Sampling. 2017.

Russo and Van Roy. An Information-Theoretic Analysis of Thompson Sampling. 2016.

Thompson sampling is empirically strong

Chapelle and Li. An Empirical Evaluation of Thompson Sampling. 2011.

Yang. A Study on Multi-Arm Bandit Problem with UCB and Thompson Sampling Algorithm. 2024.

Wang and Chen. Thompson Sampling for Combinatorial Semi-Bandits. 2018.