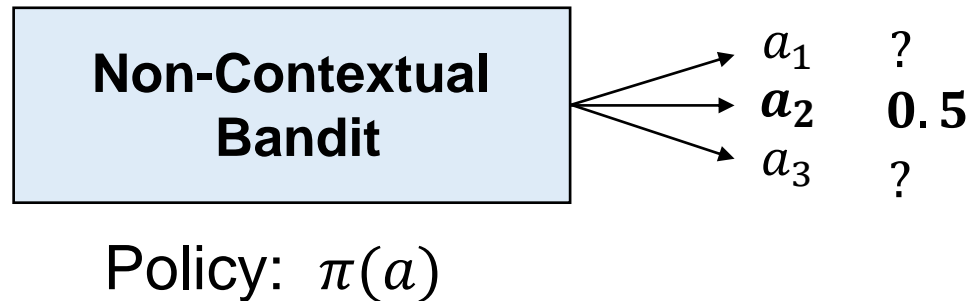
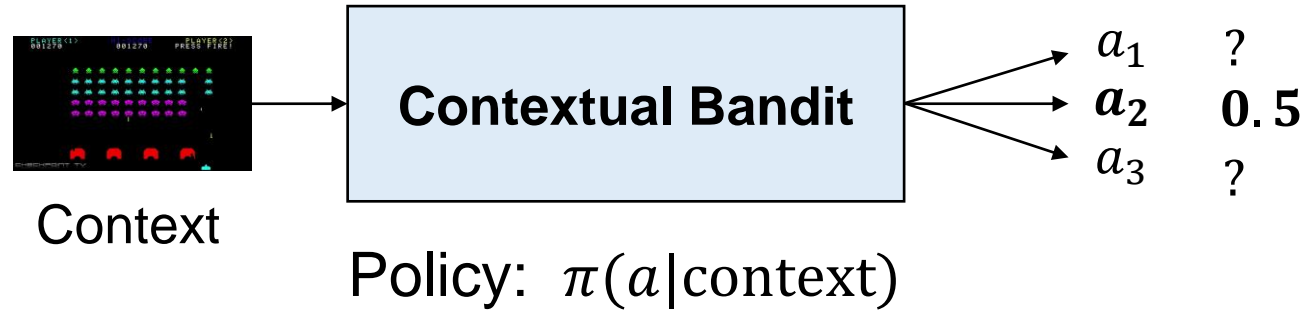


Bandits

Chen-Yu Wei

Contextual Bandits and Non-Contextual Bandits



Multi-Armed Bandits

Multi-Armed Bandits

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

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

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

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

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?
- Benchmarking the problem

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

- “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

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 $\hat{a} = \operatorname{argmax}_a \hat{R}(a)$ **(Exploit)**

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

Quantifying the Estimation Error

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

Key Question:

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

for some decreasing function of N

Empirical mean
of N i.i.d. samples

True mean

Explore-then-Exploit Regret Bound Analysis

Explore-then-Exploit Regret Bound Analysis

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 $1 - \delta$,

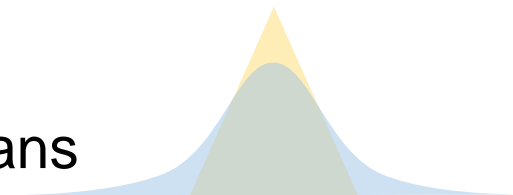
$$\left| \frac{1}{N} \sum_{i=1}^N X_i - \frac{1}{N} \sum_{i=1}^N \mathbb{E}[X_i] \right| \leq \sigma \sqrt{\frac{2 \log(2/\delta)}{N}} .$$

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

Fact 1. $\mathcal{N}(\mu, \sigma)$ 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



Regret Bound of Explore-then-Exploit

Theorem. Regret Bound of Explore-then-Exploit

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

Then with probability at least $1 - A\delta$, Explore-then-Exploit ensures

$$\text{Regret} \leq T_0 + 2(T - T_0) \sqrt{\frac{2A \log(2/\delta)}{T_0}} .$$

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 $t > A$,

Draw

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

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 $t - 1$.

Regret Bound of ϵ -Greedy

Theorem. Regret Bound of ϵ -Greedy

With proper choice of ϵ , the expected regret of ϵ -Greedy is bounded by

$$\mathbb{E}[\text{Regret}] \leq \tilde{O}(A^{1/3} T^{2/3}).$$

Can We Do Better?

In explore-then-exploit and ϵ -greedy, every arm receives the same amount of exploration.

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

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

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

Boltzmann Exploration

Boltzmann Exploration (Parameter: λ_t)

In each round, sample a_t according to

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

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

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

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

Another adaptive exploration $p_t(a) = \frac{1}{\gamma - \lambda_t \hat{R}_t(a)}$ will work! (later in the course)

Another Idea: “Optimism in the Face of Uncertainty”

In words:

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

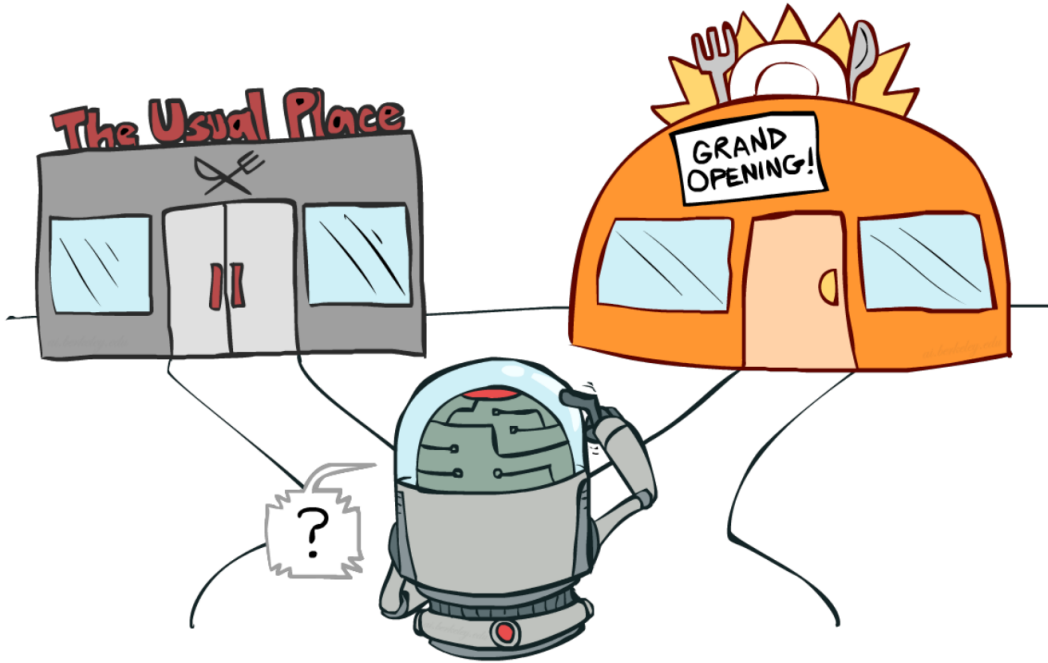


Image source: UC Berkeley AI course [slide](#), [lecture 11](#).

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 $R(a)$?

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

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

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

UCB

UCB (Parameter: δ)

In the first A rounds, draw each arm once.

For the remaining rounds: in round t , draw

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

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 $t - 1$.

Regret Bound of UCB

Theorem. Regret Bound of UCB

With probability at least $1 - AT\delta$,

$$\text{Regret} \leq O\left(\sqrt{AT \log(1/\delta)}\right) = \tilde{O}(\sqrt{AT}) .$$

UCB Regret Bound Analysis

Comparison

	Regret Bound	Exploration
Explore-then-Exploit ϵ -Greedy	$A^{1/3} T^{2/3}$	Non-adaptive
Boltzmann Exploration	---	Adaptive
UCB Thompson Sampling	\sqrt{AT}	Adaptive

Bayesian Setting for MAB

Assumptions:

- At the beginning, the environment draws a parameter θ from some prior distribution $\theta \sim P_{\text{prior}}$
- In every round, the reward vector $\mathbf{r}_t = (r_t(1), \dots, r_t(A))$ is generated from $\mathbf{r}_t \sim P_{\theta}$

E.g., Gaussian Case

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

For the learner, P_{prior} is known; θ 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 $\mathcal{H}_t = (a_1, r_1(a_1), a_2, r_2(a_2), \dots, a_{t-1}, r_{t-1}(a_{t-1}))$, the learner has a **posterior distribution** for θ :

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

In math:

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

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

Thompson Sampling in the Gaussian Case

More on Thompson Sampling

Besides Gaussian, the **Beta** prior is a prior distribution commonly used with Bernoulli reward.

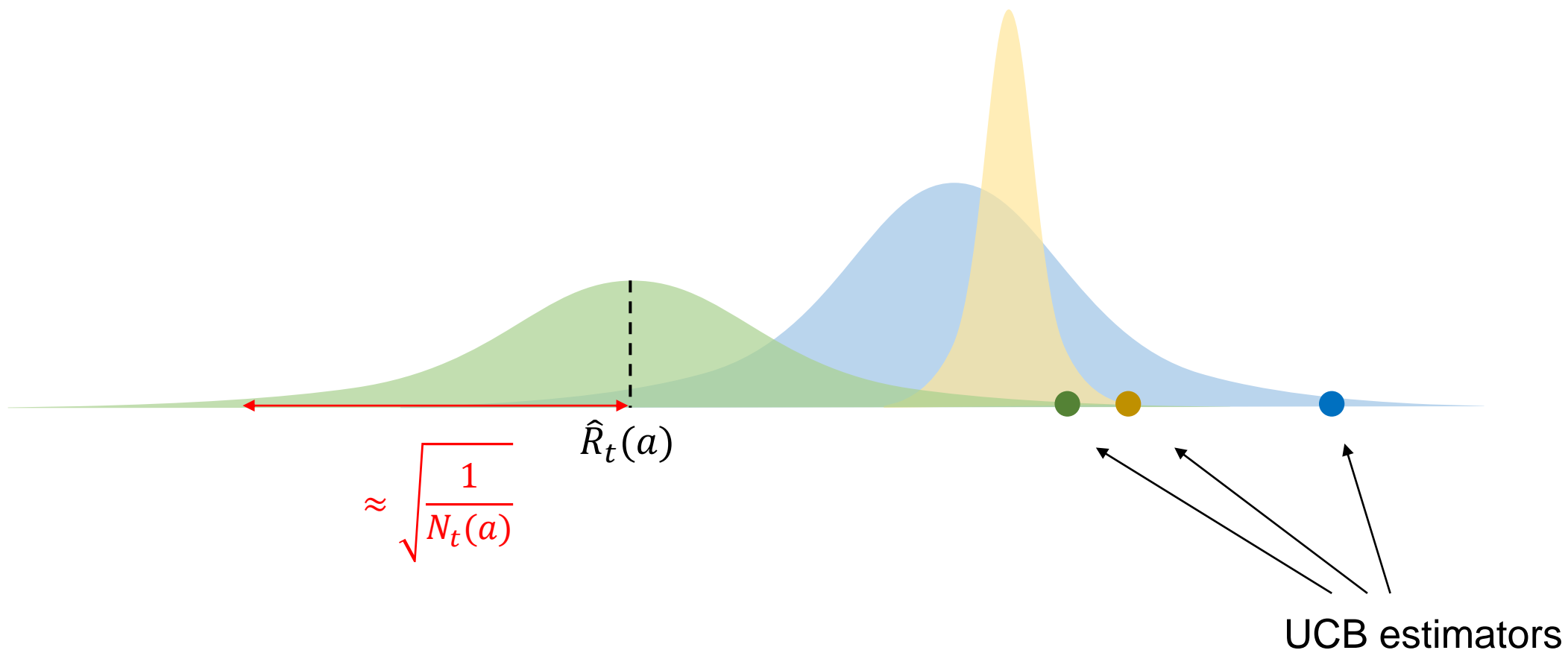
Regret bound analysis of Thompson sampling

Shipra Agrawal, Navin Goyal. [Near-optimal Regret Bounds for Thompson Sampling](#). 2017.

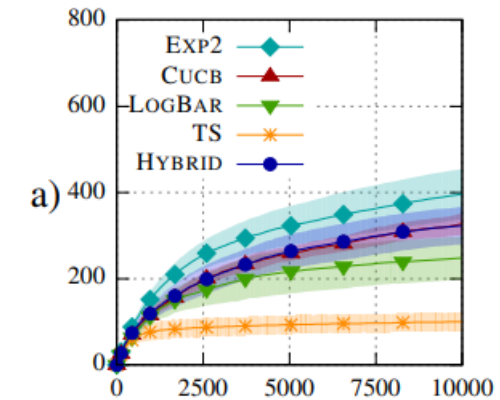
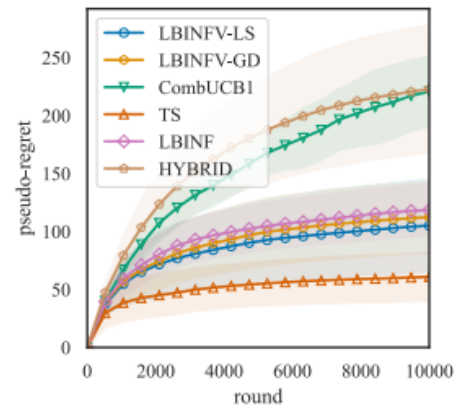
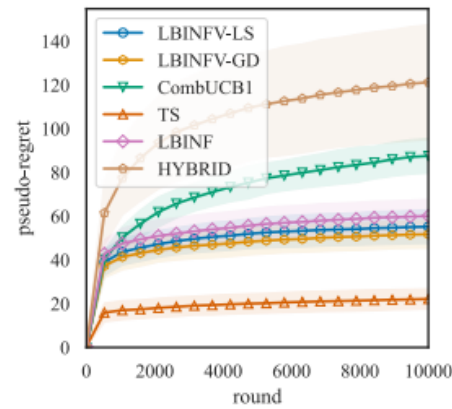
Daniel Russo and Ben Van Roy. [An Information-Theoretic Analysis of Thompson Sampling](#). 2016.

Tor Lattimore and Saba Szepesvári. [Bandit Algorithms](#) Chapter 36. 2020.

TS, UCB, and BE

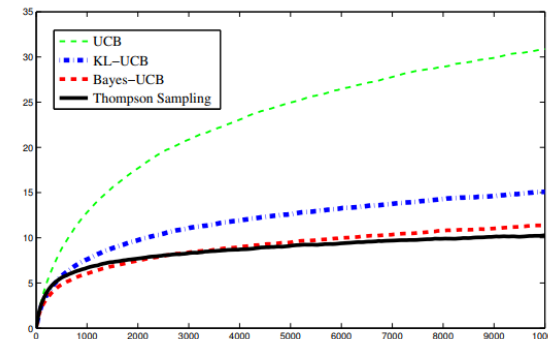
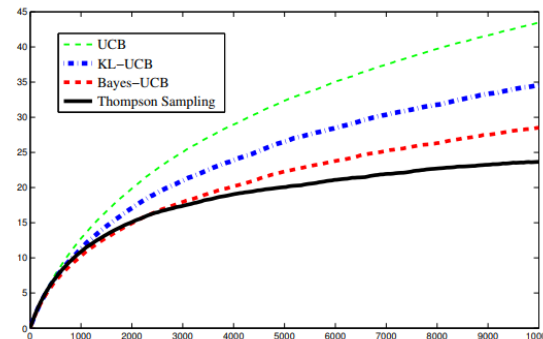


Superior Empirical Performance of TS



Tsuchiya, Ito, Honda. Further Adaptive Best-of-Both-Worlds Algorithm for Combinatorial Semi-Bandits. 2023

Zimmert, Luo, Wei. Beating Stochastic and Adversarial Semi-bandits Optimally and Simultaneously. 2019.



Kaufmann, Korda Munos. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis. 2012.