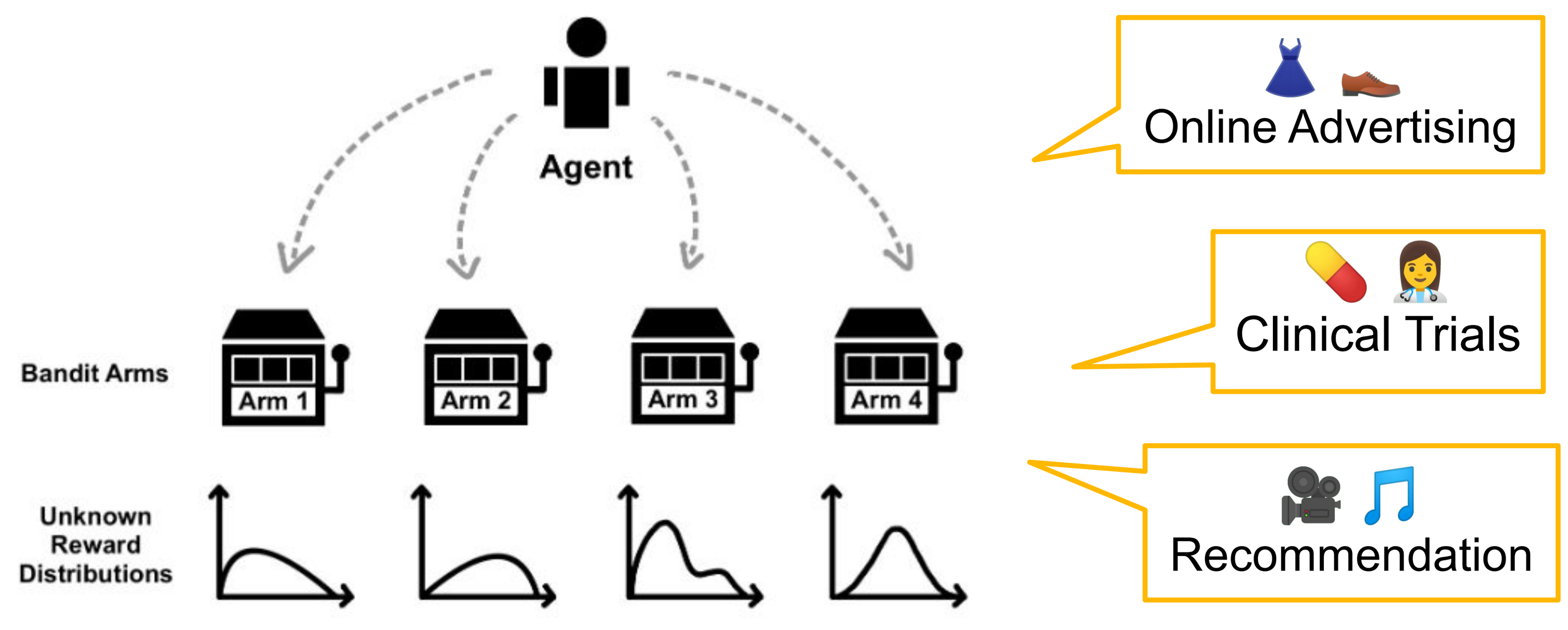
From 6235149080811616882909238708 to 29: Vanilla Thompson Sampling Revisited

Bingshan Hu (University of British Columbia)
Tianyue H. Zhang (Mila-Quebec AI Institute, Université de Montréal)

NEURAL INFORMATION PROCESSING SYSTEMS
OPT2023 Optimization for Machine Learning

THE UNIVERSITY OF BRITISH COLUMBIA
Mila
Université de Montréal

Motivation



Multi-Armed Bandits (MAB)

Stochastic Multi-Armed Bandits

A stochastic MAB instance $\Theta := ([K]; \mu_1, \mu_2, \ldots, \mu_K)$
In every round $t = 1, 2, \ldots, T$

1. Environment generates a reward vector $\left(X_1(t), \ldots, \underbrace{X_j(t)}_{\sim \mathrm{Ber}(\mu_j)}, \ldots, X_K(t)\right)$
2. Simultaneously, Learner pulls an arm $J_t \in [K]$
3. Environment reveals $X_{J_t}(t)$; Learner observes and obtains $X_{J_t}(t)$

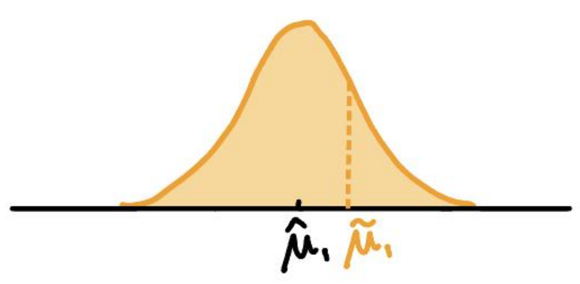
Goal: pull arms sequentially to maximize cumulative reward

$$\text{Regret: } \mathcal{R}(T;\Theta) = \mathbb{E}\left[\sum_{t=1}^{T}\left(\max_{j\in[K]} \mu_j - \mu_{J_t}\right)\right]$$

Vanilla Thompson Sampling

"Randomly take action according to the probability you believe it is the optimal action" - Thompson 1933

TS uses a data-dependent distribution to model the mean of the reward distribution for each arm.



Vanilla TS uses Gaussian distributions to model the mean reward:

- Compute the **empirical mean** of each arm and build the posterior distribution;
- Draw a **random sample** as a proxy for goodness of arm.

Algorithm 1 Thompson Sampling with Gaussian Priors [1]

Initialization: for each $i \in [K]$: pull it once to initialize n_i and the empirical mean $\hat{\mu}_{i,n_i}$
for $t = K+1, K+2, \cdots$ **do**
  Draw $\theta_i(t) \sim \mathcal{N}\left(\hat{\mu}_{i,n_i}, \frac{1}{n_i}\right)$ for all $i \in [K]$
  Pull arm $i_t \in \arg\max_{i\in[K]} \theta_i(t)$ and observe $X_{i_t}(t)$
  Set $n_{i_t} \leftarrow n_{i_t} + 1$ and update the empirical mean $\hat{\mu}_{i_t,n_{i_t}}$ of the pulled arm i_t accordingly.
end for

Existing Regret Bound of Vanilla TS

[Agrawal and Goyal, 2017]

$$\sum_{i:\Delta_i>0} \frac{288\left(\boxed{e^{64}}+6\right)\ln\left(T\Delta_i^2+e^{32}\right)}{\Delta_i} + \frac{10.5}{\Delta_i} + \Delta_i$$

$$\Delta_i = \mu_* - \mu_i$$

Sub-optimality gap

- The coefficient for the leading term is at least

$$\boxed{288 \cdot e^{64} \approx 1.8 \times 10^{30}}$$

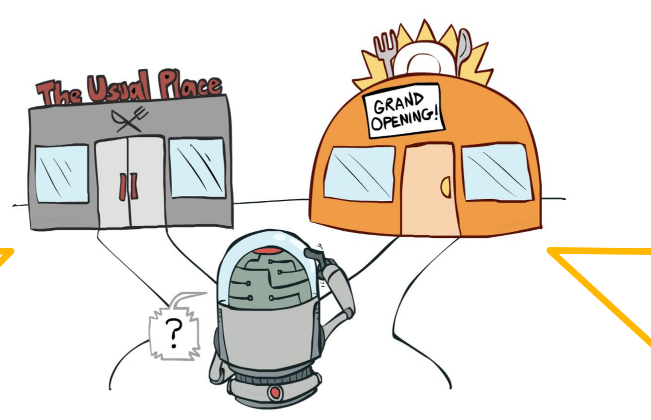
- Since the regret is at most T, the regret bound is vacuous for learning problems when

$$T \le 288 \cdot e^{64}$$

Challenge: Exploitation vs Exploration Trade-Off

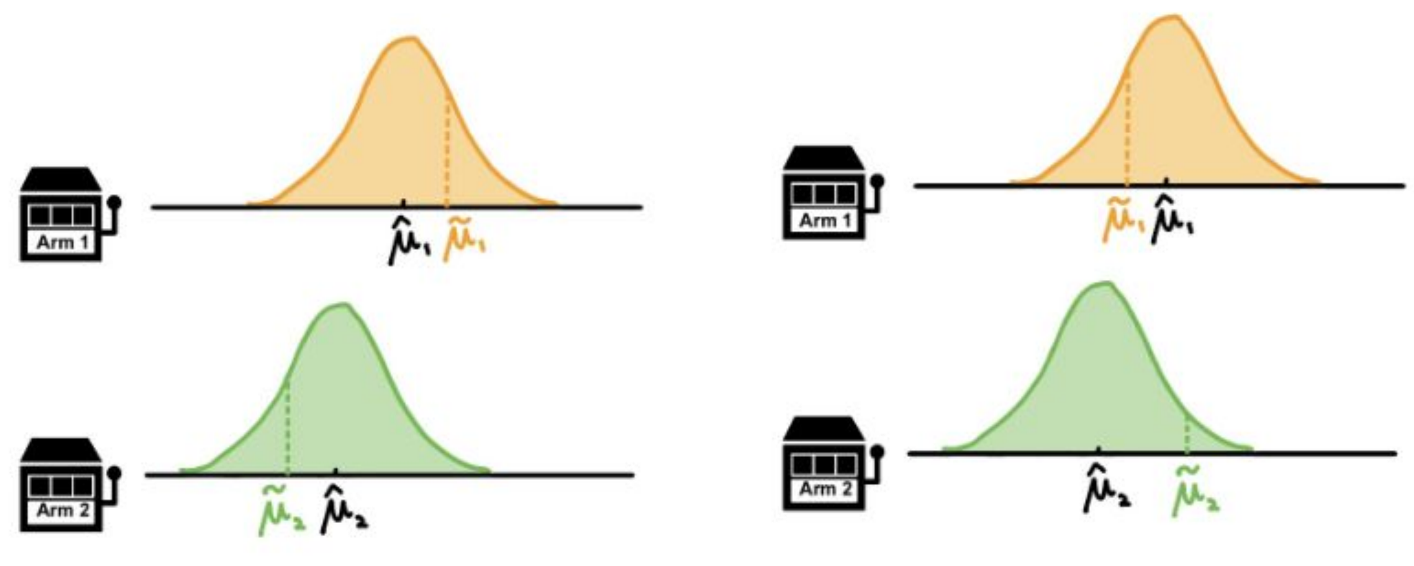
Exploitation

Take actions with high empirical reward to gain pay-off



Exploration

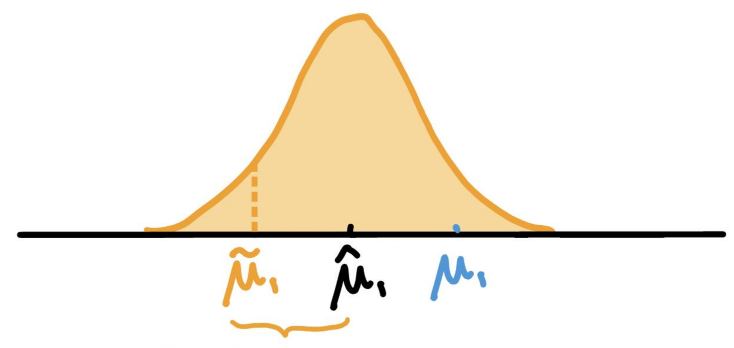
Take less observed actions to gather information



Exploitation   Exploration

Objective

When the posterior distribution of the optimal arm is not concentrated, that is, the optimal arm has not been sufficiently observed,
What is the the expected number of rounds needed before the optimal arm has a good posterior sample?



Example when the **true mean** of the optimal arm is **underestimated**, and the **sample** is also "bad".

Our Improved Bound

First, we show that the expected number of rounds is at most 29 for us to have a good sample for the optimal arm

Lemma 2 *Let $\tau_s^{(1)}$ be the round when the s-th pull of the optimal arm 1 occurs and $\theta_{1,s} \sim \mathcal{N}\left(\hat{\mu}_{1,s}, \frac{1}{s}\right)$. Then, for any integer $s \ge 1$, we have*

$$\mathbb{E}_{\mathcal{F}_{\tau_s^{(1)}}}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{\tau_s^{(1)}}\right\}} - 1\right] \le \boxed{29}.$$

Also, for any integer $s \ge L_{1,i} := \frac{4\left(\sqrt{2}+\sqrt{3.5}\right)^2 \ln\left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)}{\Delta_i^2}$, we have

$$\mathbb{E}_{\mathcal{F}_{\tau_s^{(1)}}}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{\tau_s^{(1)}}\right\}} - 1\right] \le \frac{180}{T\Delta_i^2}.$$

Then, our improved bound is:

$$\sum_{i:\Delta_i>0} \frac{\boxed{1252}\ln\left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)}{\Delta_i} + \frac{18\ln\left(T\Delta_i^2\right)}{\Delta_i} + \frac{182.5}{\Delta_i} + \Delta_i$$

Note that our improved problem-dependent regret bound also implies an improved worst-case regret bound.

Acknowledgement

This work was supported by Alberta Machine Intelligence Institute (Amii), the Canada CIFAR AI Program and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants.
Part of the work was done when Bingshan Hu was at Amii and University of Alberta, and when Tianyue Zhang was at University of British Columbia.