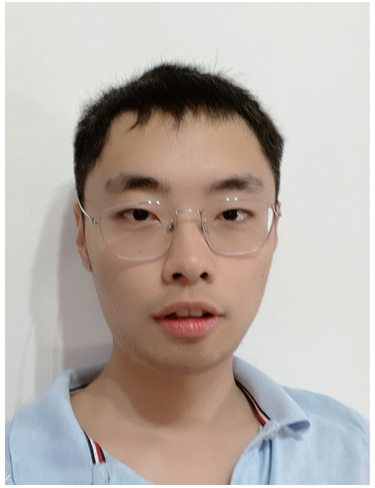
MOTS: Minimax Optimal Thompson Sampling



Tianyuan Jin

Pan Xu

Jieming Shi

Xiaokui Xiao

Quanquan Gu

NUS National University of Singapore

Caltech

The Hong Kong Polytechnic University 香港理工大學

UCLA

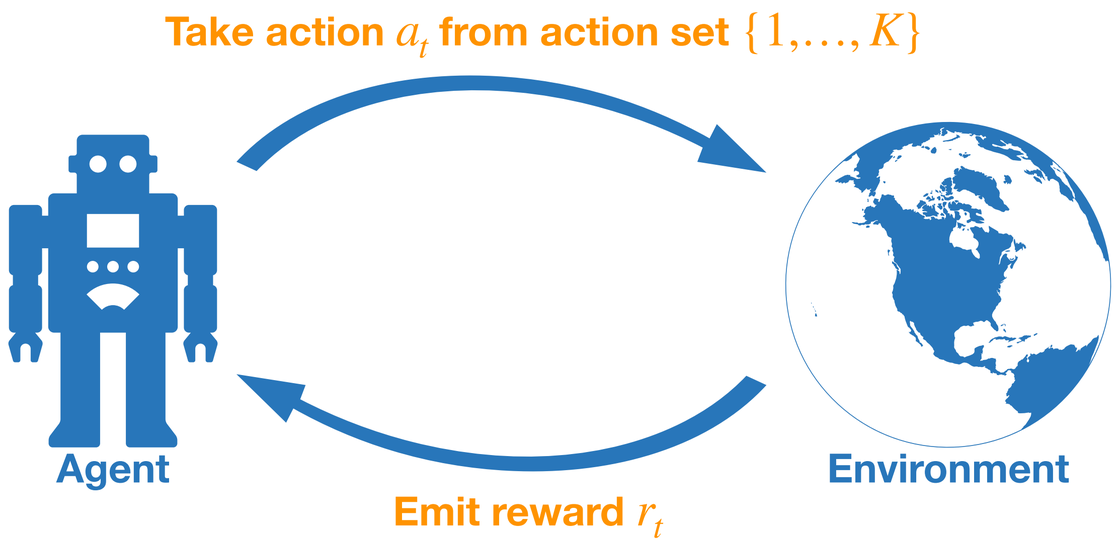
Multi-Armed Bandits (MAB)

- Sequential decision process with an agent, a set of arms $\{1,\ldots,K\}$, and an environment that emits rewards.

- Each arm $a \in \{1,\ldots,K\}$ leads to a reward following a 1-subGuassian distribution with mean μ_a.

- **Goal**: **minimize the regret** due to not knowing the best arm

$$R_\mu(T) = T \times \max_{a \in \{1,\ldots,K\}} \mu_a - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$



Thompson Sampling Type Algorithms

The general prototype of TS-type algorithms:

For each arm $i \in [K]$, maintain a prior distribution for the mean reward

For $t = 1,2,\ldots$ **do**

For each arm $i \in [K]$, sample $\theta_i(t)$ independently from the prior

Play arm $A_t = \arg\max_{i \in [K]} \theta_i(t)$

For each arm $i \in [K]$, update the mean and variance of the prior

end

Thompson Sampling

The original Thompson Sampling algorithm [Thompson, 1933][Li & Chapelle, 2012][Agrawal & Goyal, 2013][Agrawal & Goyal, 2017]

For each arm $i \in [K]$, set $\hat{\mu}_i(0) = 0$, $T_i(0) = 0$. Play each arm once.

For $t = K+1, K+2, \ldots$ **do**

For each arm $i \in [K]$, sample $\theta_i(t)$ independently from $\mathcal{N}(\hat{\mu}_i(t), 1/T_i(t))$

Play arm $A_t = \arg\max_{i \in [K]} \theta_i(t)$

For each arm $i \in [K]$, update $\hat{\mu}_i(t+1) = \dfrac{T_i(t) \cdot \hat{\mu}_i(t) + r_t \cdot \mathbf{1}\{i = A_t\}}{T_i(t) + \mathbf{1}\{i = A_t\}}$,

$$T_i(t+1) = T_i(t) + \mathbf{1}\{i = A_t\}$$

end

Regret Bound of Thompson Sampling

[Agrawal & Goyal, 2017]: TS with Beta priors has an $O(\sqrt{KT\log T})$ regret, and TS with Gaussian priors has an $O(\sqrt{KT\log K})$ regret.

- The existing best regret bound for Thompson sampling does not exactly match the minimax optimal regret $\Omega(\sqrt{KT})$ for MAB [Auer et al., 2002]
- It remains an open problem [Li & Chapelle]: whether TS type algorithms can achieve the minimax optimal regret bound $\Omega(\sqrt{KT})$ for MAB problems

MOTS: Minimax Optimal Thompson Sampling

Use a clipped Gaussian distribution as the prior distribution for $\theta_i(t)$

For each arm $i \in [K]$, set $\hat{\mu}_i(0) = 0$, $T_i(0) = 0$

For $t = 1,2,\ldots$ **do**

For each arm $i \in [K]$, sample $\theta_i(t)$ from $\mathcal{N}^{\text{clipped}}(\hat{\mu}_i(t), 1/T_i(t))$

Play arm $A_t = \arg\max_{i \in [K]} \theta_i(t)$

For each arm $i \in [K]$, update $\hat{\mu}_i(t+1) = \dfrac{T_i(t) \cdot \hat{\mu}_i(t) + r_t \cdot \mathbf{1}\{i = A_t\}}{T_i(t) + 1}$,

$$T_i(t+1) = T_i(t) + \mathbf{1}\{i = A_t\}$$

end

MOTS: Minimax Optimal Thompson Sampling

Use a clipped Gaussian distribution as the prior distribution for $\theta_i(t)$

Clipped Gaussian distribution

sample $\theta_i(t)$ from

$\mathcal{N}^{\text{clipped}}(\hat{\mu}_i(t), 1/T_i(t))$

=

Step 1: sample $\tilde{\theta}_i(t)$ from $\mathcal{N}(\hat{\mu}_i(t), 1/(\rho T_i(t)))$

Inflation parameter ρ is used to avoid the underestimation of the optimal arm

Step 2: obtain $\theta_i(t)$ as $\min\{\tilde{\theta}_i(t), \tau_i(t)\}$

Clipping threshold $\tau_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{\alpha}{T_i(t)} \log^+\left(\frac{T}{KT_i(t)}\right)}$

is used to avoid the overestimation of suboptimal arms

Regret Bound of MOTS

Theorem: For any fixed $\rho \in (1/2, 1)$ and $\alpha \geq 4$, the regret of MOTS is

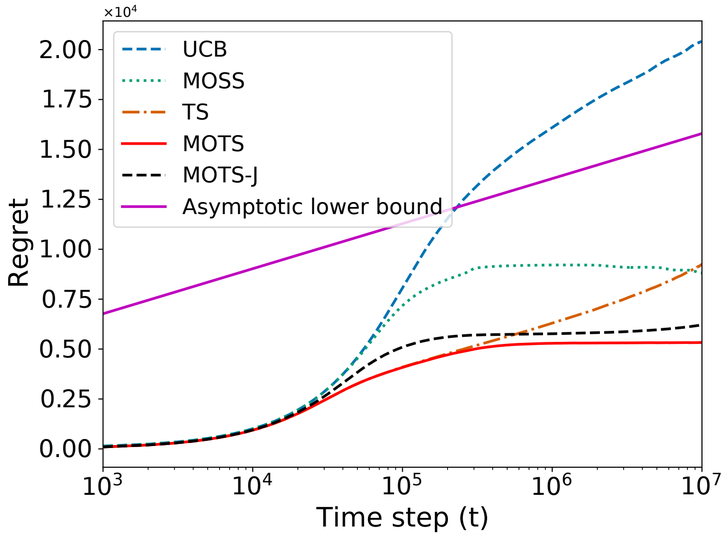
$$R_\mu(T) = O\left(\sqrt{KT} + \sum_{i=1}^{K} \Delta_i\right)$$

- In comparison, TS with Beta priors [Agrawal & Goyal, 2017] has an $O(\sqrt{KT \log T})$ regret, and TS with Gaussian priors [Agrawal & Goyal, 2017] has an $O(\sqrt{KT \log K})$ regret

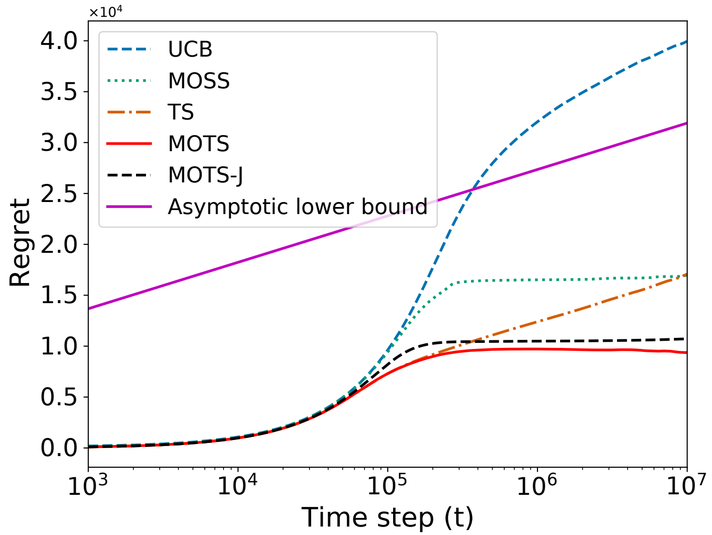
- MOTS is the first TS-type algorithm that matches the minimax optimal regret $\Omega(\sqrt{KT})$ for MAB [Auer et al., 2002]

Experimental Results

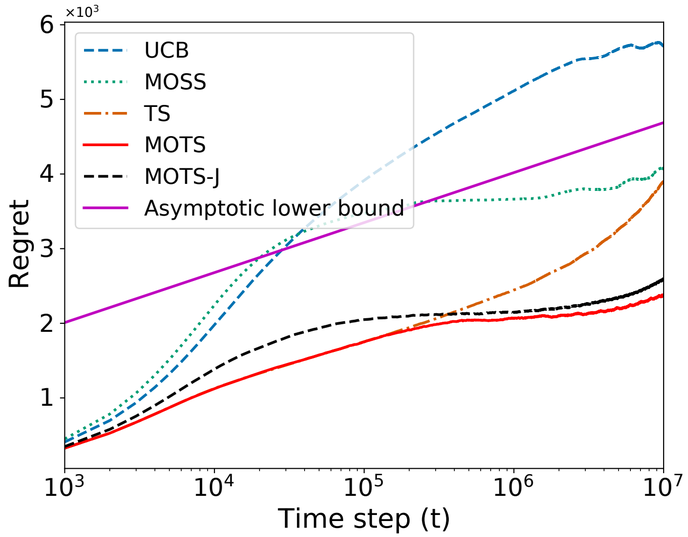
Set up: K arms, each one has a reward distribution $\mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, K$



Setting 1: K=50
{1, 0.9, 0.9, …, 0.9}



Setting 2: K=100
{1, 0.9, 0.9, …, 0.9}



Setting 3: K=50
{1, 0.9, 0.9, 0.9, 0.9, 0.9,
0.8, 0.8, 0.8, 0.8, 0.8,
…
0.1, 0.1, 0.1, 0.1, 0.1,
0, 0, 0, 0}

* the set of numbers in each setting is the mean reward for each arm respectively

Thank You

Paper: **MOTS: Minimax Optimal Thompson Sampling**