**Introduction**

Stochastic multi-armed bandits problem:
- Arms of a stochastic bandit $I = \{1, 2, \cdots, K\}$, $K \geq 2$.
- Reward of pulling arm $i$ at time $t$: $r_{i,t} \sim N(\mu_i, 1)$.
- Time horizon $T$.
- A predictable process $\pi = (\pi_t)_{t \geq 1}$ with regard to the filtration $\mathcal{F}_t = \{A_1, A_2, \cdots, A_t, \pi_1, \pi_2, \cdots, \pi_t\}$.

**Batch constraints:**
- Grid of $M$ batches, $1 \leq t_1 < t_2 < \cdots < t_M = T$.
- For $t_j < t < t_{j+1}$, $\pi_t$ is $\mathcal{F}_t$ measurable.

We aim to characterize the following problem-dependent regret under the batched setting:

- Reward of pulling arm
- Arms of a stochastic bandit
- Time horizon
- Static grid: Fix the grid beforehand.
- Grid of $I_t \in \{I_1, I_2, \cdots, I_{M-1}\}$, and $\Delta_t < T$

**Main Results**

**Theorem 1 (Upper Bound):** There exist policies $\pi_1, \pi_2$ such that

$$E[R(\pi_1)] \leq \text{polylog}(K, T) \cdot \sqrt{KT},$$

$$E[R(\pi_2)] \leq \text{polylog}(K, T) \cdot \frac{KT}{\min_i \Delta_i}.$$  

**Theorem 2 (Static Lower Bound):** Under any static grid,

$$R_{\text{min-max}}(K, M, T) = \Omega(\sqrt{KT}),$$

$$R_{\text{pro-dep}}(K, M, T) = \Omega(KT^{\gamma}).$$

**Theorem 3 (Adaptive Lower Bound):** Under any adaptive grid,

$$R_{\text{min-max}}(K, M, T) = \Omega(\sqrt{KT}),$$

$$R_{\text{pro-dep}}(K, M, T) = \Omega(KT^{\gamma}).$$

Remark:
- It is sufficient to have $M = O(\log \log T)$ batches to achieve the optimal minimax regret $\Theta(\sqrt{KT})$, and $M = O(\log T)$ to achieve the optimal problem-dependent regret $\Theta(KT^{\gamma})$.
- With either static or adaptive grids, it is necessary to have $M = \Omega(\log \log T)$ batches to achieve the optimal minimax regret $\Theta(\sqrt{KT})$, and $M = \Omega(\log T)$ to achieve the optimal problem-dependent regret $\Theta(KT^{\gamma})$.
- It is an open problem to remove the $M^{-2}$ factor in the adaptive lower bound.

**Two Types of Regrets**

We aim to characterize the following minimax regret and problem-dependent regret under the batched setting:

$$R_{\text{min-max}}(K, M, T) \equiv \inf_{\pi} \sup_{\mu \in \Pi} E[R(\pi)],$$

$$R_{\text{pro-dep}}(K, M, T) \equiv \inf_{\pi} \sup_{\mu \in \Pi} \sup_{\Delta \geq 0} E[R(\pi)].$$

where $\Pi_{M,T}$ is the set of policies with $M$ batches and horizon $T$, and $\Delta = \mu^* - \mu_0$.

**Related Works**

Without batch constraint [1, 2]:

$$R_{\text{min-max}}(K, T, T) = \Theta(\sqrt{KT}),$$

$$R_{\text{pro-dep}}(K, T, T) = \Theta(K \log T).$$

Required number of batches [3]:

$$R_{\text{min-max}}(K, \log T, T) = \Theta(\sqrt{KT}).$$

Two-batch grid with static bandit [4]:

$$R_{\text{min-max}}(2, M, T) = \Theta(T^{1/2 - 2^{-m}}),$$

$$R_{\text{pro-dep}}(2, M, T) = \Theta(T^{1/2 - m}).$$

**Optimal Grid Design**

Minimax grid: $t_a = a, t_b = [a \sqrt{T}]$, where $a = \Theta \left( \frac{T}{(1/2 - m)} \right)$.

Geometric grid: $t_a = \bar{a}, t_b = \lfloor \bar{a} \sqrt{T} \rfloor$, where $\bar{a} = \Theta \left( \frac{T}{(1/2 - m)} \right)$.

**Numerical Experiments**

**Baseline:** UCB1 algorithm [5] without any batch constraints.

**Observations:**
- The minimax grid typically results in a smallest regret among all grids.
- $M = 4$ batches appear to be sufficient for the BaSE performance to approach the centralized performance.

**References**


