



Batched Multi-armed Bandits Problem

Zijun Gao[†], Yanjun Han^{*}, Zhimei Ren[†], Zhengqing Zhou[‡]

[†] Stanford STATS, ^{*} Stanford EE, [‡] Stanford MATH

Email: {zijungao,yjhan,zren,zqzhou}@stanford.edu

Introduction

Stochastic multi-armed bandits problem:

- Arms of a stochastic bandit $\mathcal{I} = \{1, 2, \dots, K\}$, $K \geq 2$.
- Reward of pulling arm i at time t : $r_{t,i} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu^{(i)}, 1)$.
- Time horizon T .
- A predictable process $\pi = (\pi_t)_{t=1}^T$ with regard to the filtration $\mathcal{F}_t = \{A_1, A_2, \dots, A_t, r_{1,A_1}, r_{2,A_2}, \dots, r_{t,A_t}\}$.

Batch constraints:

- Grid of M batches, $1 \leq t_1 < t_2 < \dots < t_M = T$.
- For $t_j < t \leq t_{j+1}$, π_t is \mathcal{F}_{t_j} measurable.

Two types of grid:

- Static grid:** Fix the grid beforehand.
- Adaptive grid:** Determine t_{j+1} based on \mathcal{F}_{t_j} .

Target: Minimize regret

$$R_T(\pi) \triangleq \sum_{t=1}^T (\mu^* - \mu^{(\pi_t)}) = T\mu^* - \sum_{t=1}^T \mu^{(\pi_t)}$$

under batched constraints, where $\mu^* = \max_{i \in [K]} \mu^{(i)}$.

Two Types of Regrets

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

$$R_{\min\text{-max}}^*(K, M, T) \triangleq \inf_{\pi \in \Pi_{M,T}} \sup_{\{\mu^{(i)}\}_{i=1}^K, \Delta_i \leq \sqrt{K}} \mathbb{E}[R_T(\pi)],$$

$$R_{\text{pro-dep}}^*(K, M, T) \triangleq \inf_{\pi \in \Pi_{M,T}} \sup_{\Delta > 0} \Delta \cdot \sup_{\{\mu^{(i)}\}_{i=1}^K, \Delta_i \in \{0\} \cup [\Delta, \sqrt{K}]} \mathbb{E}[R_T(\pi)].$$

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

Related Works

Without batch constraint [1, 2]:

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

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

Required number of batches [3]:

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

Two-armed bandit with static grid [4]:

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

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

Main Results

Theorem 1 (Upper Bound): There exist policies π^1, π^2 such that

$$\mathbb{E}[R(\pi^1)] \leq \text{polylog}(K, T) \cdot \sqrt{KT}^{2^{-2^{1-M}}},$$

$$\mathbb{E}[R(\pi^2)] \leq \text{polylog}(K, T) \cdot \frac{KT^{1/M}}{\min_{i \neq *}} \Delta_i.$$

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

$$R_{\min\text{-max}}(K, M, T) = \Omega(\sqrt{KT}^{2^{-2^{1-M}}}),$$

$$R_{\text{pro-dep}}(K, M, T) = \Omega(KT^{1/M}).$$

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

$$R_{\min\text{-max}}(K, M, T) = \Omega(M^{-2} \cdot \sqrt{KT}^{2^{-2^{1-M}}}),$$

$$R_{\text{pro-dep}}(K, M, T) = \Omega(M^{-2} \cdot KT^{1/M}).$$

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 $\Theta(K \log T)$.
- 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 $M = \Omega(\log T / \log \log T)$ to achieve the optimal problem-dependent regret $\Theta(K \log T)$.
- It is an open problem to remove the M^{-2} factor in the adaptive lower bound.

BaSE Policy

Key Idea: Sequentially drop the arms which are “significantly” worse than the “best” one.

BaSE (Batched Successive Elimination)

Input: number of arms K , number of batches M , time horizon T , time grid \mathcal{T} , tuning parameter $\gamma > 0$

Output: policy π

initialize the set of active arms $\mathcal{A} \leftarrow [K]$;

for $m = 1$ to M **do**

pull all active arms for same number of times in m -th batch;

for $i \in \mathcal{A}$ **do**

compute the mean reward \bar{r}_i for arm i ;

end for

compute the maximum mean reward $r_{\max} = \max_{i \in \mathcal{A}} \bar{r}_i$ and the number of pullings τ_m for each active arm;

eliminate all active arms with $r_{\max} - \bar{r}_i \geq \sqrt{\gamma \log(TK) / \tau_m}$ from \mathcal{A} ;

end for

Optimal Grid Design

Minimax grid: $t_1 = a$, $t_m = \lfloor a\sqrt{t_{m-1}} \rfloor$, where $a = \Theta(T^{1/(2-2^{1-M})})$.

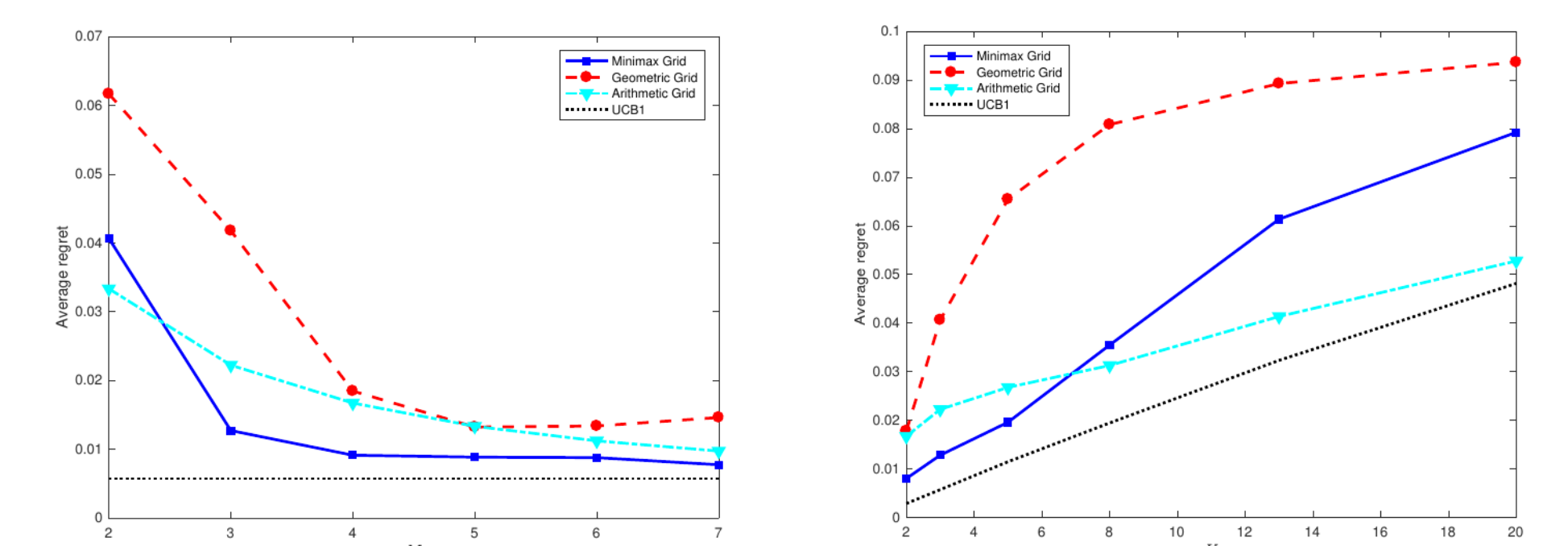
Geometric grid: $t'_1 = b$, $t'_m = \lfloor bt'_{m-1} \rfloor$, where $b = \Theta(T^{1/M})$.

Numerical Experiments

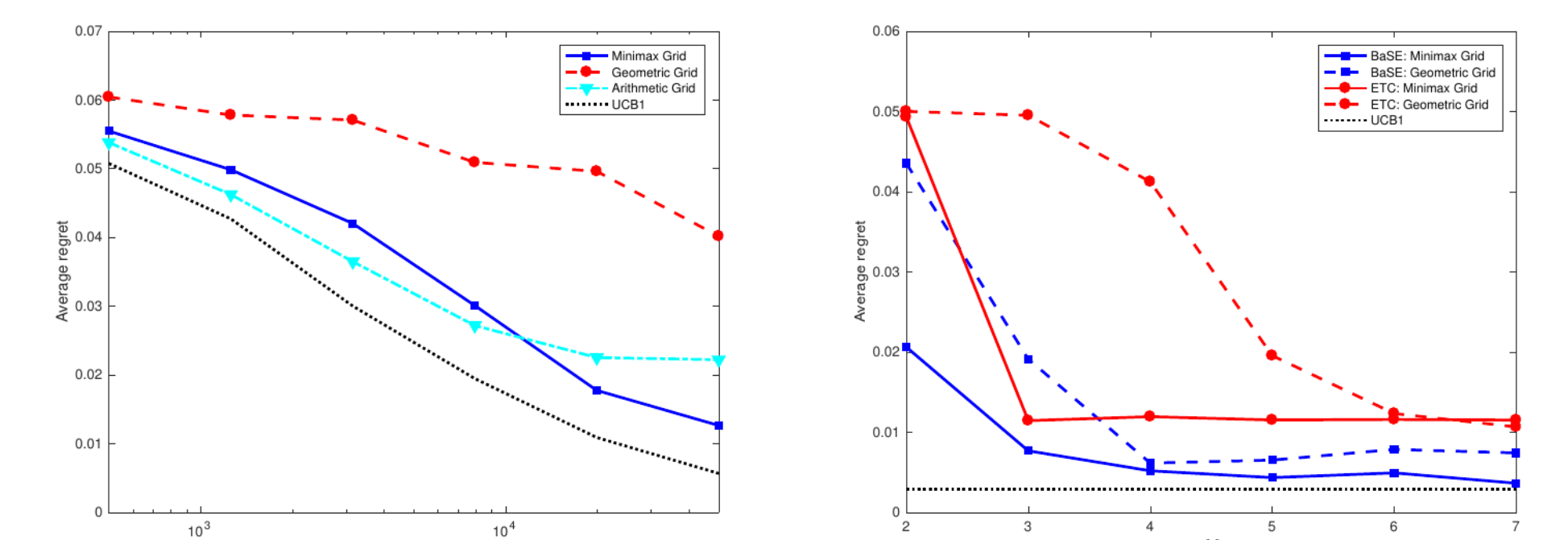
Setting:

- Parameters: $T = 5 \times 10^4$, $K = 3$, $M = 3$ and $\gamma = 1$.
- Mean reward: $\mu^* = 0.6$ for the optimal arm and $\mu = 0.5$ for all other arms.
- Implement minimax grid, geometric grid and the arithmetic grid with $t_j = jT/M$ for $j \in [M]$.
- Baseline: UCB1 algorithm [5] without any batch constraints.

Experimental results:



(a) Average regret vs. the number of batches M . (b) Average regret vs. the number of arms K .



(c) Average regret vs. the time horizon T . (d) Comparison of BaSE and ETC.

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

- [1] Walter Vogel. *A sequential design for the two armed bandit*. The Annals of Mathematical Statistics, 31(2):430–443, 1960.
- [2] Tze Leung Lai and Herbert Robbins. *Asymptotically efficient adaptive allocation rules*. Advances in applied mathematics, 6(1):4–22, 1985.
- [3] Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. *Online learning with switching costs and other adaptive adversaries*. In Advances in Neural Information Processing Systems, pages 1160–1168, 2013.
- [4] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. *Batched bandit problems*. The Annals of Statistics, 44(2):660–681, 2016.
- [5] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. *Finite-time analysis of the multiarmed bandit problem*. Machine learning, 47(2-3):235–256, 2002.