Quantile Multi-Armed Bandits with 1-bit Feedback

Ivan Lau Jonathan Scarlett

National University of Singapore

SCAN ME

Summary

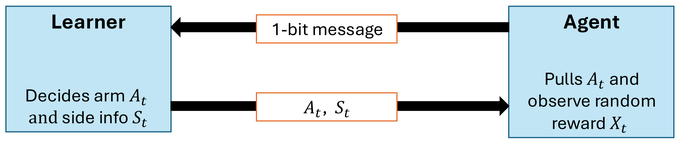
We study a variant of best-arm identification where

- the goal is to identify arm with the highest quantile (e.g., median) reward (instead of highest mean reward)
- the communication from agent (observing rewards) and learner (deciding actions) is restricted to one bit of feedback per arm pull

Motivation:

- Quantile reward is robust to heavy-tailed distributions
- Reward observations are usually done by agent (sensor) before being communicated to the learner (central server), and uplink communication bandwidth is usually limited

Problem Setup



Arms and Rewards:

- A set of arms $\mathcal{A} = \{1, \dots, K\}$
- Stochastic reward: Observations of its reward are i.i.d. random variables from unknown reward distribution F_k
- Each arm has q-quantile $F_k^{-1}(q) \in [0, \lambda]$

1-bit Communication Constraint:
At round $t = 1, \dots$

- Learner decides an arm $A_t \in \mathcal{A}$ and sends agents side info S_t
- Agent pulls A_t and observes a random reward $X_t \sim F_{A_t}$.
- Agent transmits a 1-bit message to the learner

Goal:

- Identify arm with ϵ-optimal q-quantile:

$$F_k^{-1}(q) \geq F_{k^*}^{-1}(q) - \epsilon$$

with high probability while using as few arm pulls as possible

Contributions

- Introduce arm gap Δ_k where $\max_k \Delta_k > 0 \iff$ "solvable"
- Provide an algorithm with an instance-dependent upper bound that scales logarithmically with λ/ϵ – in contrast, existing upper bounds for mean-based bandits with 1-bit constraints scales linearly with λ [1, 2]
- Provide a worst-case lower bound showing upper bound is tight to within logarithmic factors
- Lower bound is applicable even in the absence of communication constraints – restricting to 1-bit feedback has a minimal impact on the scaling of the sample complexity

Algorithm Idea: Part I (Elimination)

- Computes confidence intervals $[\mathrm{LCB}_t(k), \mathrm{UCB}_t(k)]$ containing q-quantile (with high probability)
- Eliminates suboptimal arms based on confidence interval
- Terminates when found arm satisfying

$$\mathrm{LCB}_t(k) \geq \max_{a \neq k} \mathrm{UCB}_t(a) - \epsilon$$

- Also used in quantile bandit problem with no communication constraint [3, 4]
- But how does learner compute $\mathrm{LCB}_t(k), \mathrm{UCB}_t(k)$ with no access to empirical quantiles of the observed rewards?? (due to communication constraint)

Algorithm Idea: Part II (Noisy binary search)

- Learner sends 'Is $X_t \leq \gamma_t$?" to the agent as side info S_t
- Based on the 1-bit comparison feedback $\mathbf{1}(X_t \leq \gamma_t)$ received from agent, the learner uses **noisy binary search** [5] to compute $\mathrm{LCB}_t(k)$ and $\mathrm{UCB}_t(k)$
- Computing using noisy binary search costs additional $O(\log(\lambda/\epsilon))$ compared to using empirical quantiles (see upper bound)

Arm Gap (Grossly Simplified)

- If k is not ϵ-optimal, then

$$\Delta_k := \sup\left\{\Delta : Q_k(q + \Delta) \leq \max_{a \in \mathcal{A}} Q_a(q - \Delta)\right\}$$

- If k is ϵ-optimal, then

$$\Delta_k := \sup\left\{\Delta : Q_k(q - \Delta) \geq \max_{a \neq k} Q_a(q + \Delta) - \epsilon\right\}$$

Solvable Instances

- If an instance satisfies $\max_{k \in \mathcal{A}} \Delta_k > 0$, then it is solvable by our algorithm (see upper bound below)
- If an instance satisfies $\max_{k \in \mathcal{A}} \Delta_k = 0$, then it is "unsolvable" (i.e., **any PAC-style algorithm** cannot solve reliably)

Upper and Lower Bounds

- With probability at least $1 - \delta$, our algorithm returns an ϵ-optimal arm and uses a total number of arm pulls

$$O\Bigg(\sum_k \Delta_k^{-2} \cdot \Big(\log\left(\delta^{-1}\right) + \underbrace{\log\left(\lambda K/\epsilon\right)}_{\text{noisy binary search}} + \log\left(\Delta_k^{-1}\right)\Big)\Bigg)$$

- There exists an instance such that for sufficiently small ϵ, the number of arm pulls τ by any (ϵ, δ)-PAC algorithm satisfies

$$\mathbb{E}[\tau] \geq \Omega\left(\sum_k \Delta_k^{-2} \cdot \log\left(\delta^{-1}\right)\right),$$

References

[1] D. Vial, S. Shakkottai, and R. Srikant. One-bit feedback is sufficient for upper confidence bound policies. *arXiv:2012.02876*, 2020.

[2] O. A. Hanna, L. Yang, and C. Fragouli. Solving multi-arm bandit using a few bits of communication. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 11215–11236, 2022.

[3] B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1660–1668, 2015.

[4] K. E. Nikolakakis, D. S. Kalogerias, O. Sheffet, and A. D. Sarwate. Quantile multi-armed bandits: Optimal best-arm identification and a differentially private scheme. *IEEE Journal on Selected Areas in Information Theory*, 2(2):534–548, 2021.

[5] L. Gretta and E. Price. Sharp Noisy Binary Search with Monotonic Probabilities. In *51st International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 297, pages 75:1–75:19, 2024.

[6] I. Lau and J. Scarlett. Quantile multi-armed bandits with 1-bit feedback. In *36th International Conference on Algorithmic Learning Theory*.