Quantum Best Arm Identification

Xuchuang Wang
Department of Computer Science & Engineering
The Chinese University of Hong Kong
xcwang@cse.cuhk.edu.hk

Yu-Zhen Janice Chen
College of Information & Computer Sciences
University of Massachusetts Amherst
yuzhenchen@cs.umass.edu

Matheus Guedes de Andrade
College of Information & Computer Sciences
University of Massachusetts Amherst
mguedesdeand@cs.umass.edu

Mohammad Hajiesmaili
College of Information & Computer Sciences
University of Massachusetts Amherst
hajiesmaili@cs.umass.edu

John C.S. Lui
Department of Computer Science & Engineering
The Chinese University of Hong Kong
cslui@cse.cuhk.edu.hk

Don Towsley
College of Information & Computer Sciences
University of Massachusetts Amherst
towsley@cs.umass.edu

1. INTRODUCTION

Recent progress on building quantum computers [1] envisages wide applications of quantum algorithms in the near future. With the advantage of quantum computer, one can speed up not only fundamental algorithms, e.g., unstructured search [6] and factoring [11], but recent machine learning algorithms [3] as well. In this paper, we study the quantum speedup on a canonical task of reinforcement learning—best arm identification in multi-armed bandits.

Multi-armed bandit (MAB)—initiated from Lai and Robbins [8]—is an important sequential decision making model (ref., [9]). In the stochastic case, a MAB consists of K arms, each of which is associated with a reward distribution with unknown mean μ_k. When *querying* an arm $k \in \mathcal{K} := \{1, 2, \dots, K\}$, one obtains a reward drawn from its reward distribution, i.e.,

$$\textbf{(Classic oracle)} \qquad X_k \sim \mathcal{B}(\mu_k), \tag{1}$$

where we consider the Bernoulli distribution for simplicity, and this can be easily generalized as MAB literature showed. In this paper, we show that, the quantum oracle, followed by some quantum computations, allows one to outperform classic MAB algorithms. For simplicity, we assume these K arms are ordered in descending order of their means: $\mu_1 > \mu_2 > \cdots > \mu_K$, unknown to the learner.

Besides the regret minimization objective for studying exploration-exploitation trade-off, the best arm identification (BAI) is another important objective in MAB for studying pure exploration. BAI was introduced to MAB by Even-Dar and Mannor et al. [5]. This objective has two cases: (1) BAI with fixed confidence—find the best arm with a confidence of at least $1 - \delta\,(\delta \in (0, 1))$ with as small number of queries as possible, and (2) BAI with fixed budget—given a fixed budget of queries times T, find the best arm with a correct probability as high as possible. In this paper, we focus on the BAI with fixed confidence case, and, for brevity, hereinafter, refer to it as best arm identification (BAI). Formally, the objective of BAI is, given $\delta \in (0, 1)$, to design an algorithm that minimizes the number of queries Q required for outputting the best arm with a probability of at least $1 - \delta$. To express results, we denote a reward mean gap as follows,

$$\Delta_k := \begin{cases} \mu_1 - \mu_k & \text{if } k \neq 1 \\ \mu_1 - \mu_2 & \text{if } k = 1 \end{cases}.$$

In this paper, we study this BAI objective in the *quantum* multi-armed bandits model, where the reward distributions of classic bandits are replaced with *quantum oracles*, and the reward samples are replaced with quantum copies generated from quantum oracles. We devise an elimination-based algorithm for BAI with quantum oracles. Comparing to classic BAI's sample complexity bounds, our algorithm's query complexity bounds improve the coefficient before $\log(1/\delta)$ from Δ_k^{-2} to Δ_k^{-1}, where δ is the given confidence parameter.

2. TWO QUANTUM ORACLES

Wang et al. [13] and Casalé et al. [4] first studied BAI in a quantum MAB with a quantum oracle as follows,
(Strong quantum oracle)

$$\mathcal{O}_{\text{stro}} : |k\rangle_I |0\rangle_R \mapsto |k\rangle_I \left(\sqrt{\mu_k}\,|1\rangle_R + \sqrt{1-\mu_k}\,|0\rangle_R\right), \tag{2}$$

where I is the "arm index" register with K states corresponding to K arms, and R is a single-qubit "bandit reward" register. This oracle covers the reward feedback of classic MAB: inputting a basis state to the oracle, i.e., $|k\rangle_I |0\rangle_R$, and then measuring the oracle's output $|k\rangle_I (\sqrt{\mu_k}\,|1\rangle_R + \sqrt{1-\mu_k}\,|0\rangle_R)$ yields an observation equivalent to drawing a sample from a Bernoulli distribution with mean μ_k. A potential advantage of this oracle is that it can provide information about all arms when given an input that is a uniform superposition of all arm indices. Taking advantage of that, Wang et al. [13] proposed an algorithm that enjoys a quadratic speedup in query complexity for BAI. We list their query complexity results in Table 1.

However, the quantum oracle in Eq.(2) is too powerful in the sense that it provide information regarding the reward distributions of all arms in a single query. For example, when

Table 1: Comparison of BAI's query complexity bounds of different oracles

Oracle	Query complexity upper bound
Classic oracle (Eq.(1))	$O\left(\sum_k (1/\Delta_k^2)\log(1/\delta)\right)$ [7]
Strong quantum oracle (Eq.(2))	$\tilde{O}(\sqrt{\sum_k (1/\Delta_k^2)}\log\left(1/\delta\right))$ [13]
Weak quantum oracle (Eq.(3), ours)	$\tilde{O}\left(\sum_k (1/\Delta_k)\log\left((1/\delta)\right)\right)$ (Theorem 1)

the arm index register is queried in uniform superposition of the arm indices $\sum_{k\in\mathcal{K}}(1/\sqrt{K})\left|k\right\rangle_I\left|0\right\rangle_R$, the oracle returns $\sum_{k\in\mathcal{K}}(1/\sqrt{K})\left|k\right\rangle_I\left(\sqrt{p_k}\left|1\right\rangle_R+\sqrt{1-p_k}\left|0\right\rangle_R\right)$ in which the qubit in register R encodes the information of all arms' reward distributions. Such a query is impossible in a setting where the arms are separated and can only be queried individually. Therefore, algorithms based on the oracle in Eq.(2) cannot give feasible insights about pure exploration in MAB.

Indeed, Wan et al. [12] proposed a more reasonable quantum oracle to clearly separate arm exploration for regret minimization. Associated with each arm k is an oracle $\mathcal{O}_k$, **(Weak quantum oracle)**

$$\mathcal{O}_k : \left|0\right\rangle \mapsto \sqrt{\mu_k}\left|1\right\rangle + \sqrt{1-\mu_k}\left|0\right\rangle, \quad k\in\mathcal{K}. \tag{3}$$

This quantum oracle models the query feedback of classic MAB as quantum superposition as the strong oracle $\mathcal{O}_{\text{stro}}$ in Eq.(2), but it does not provide the opportunity to simultaneously explore multiple arms as the strong oracle $\mathcal{O}_{\text{stro}}$ allows. Although one cannot query multiple arms simultaneously, this oracle in Eq.(3) is still more informative than the classic oracle in Eq.(1) because its superposition output encodes the information of the whole reward distribution, instead of a single reward sample as in Eq.(1). Therefore, this weak oracle is a reasonable choice for studying the quantum version of BAI. Wan et al. [12] devised regret minimization algorithms for both multi-armed bandits and linear bandits with this weak oracle that achieve $O(\log T)$ problem-independent upper bounds, while in classic MAB, one only has $O(\sqrt{T})$ bounds. In this paper, we look into BAI with the weak quantum oracle.

2.1 Main Result and Comparison

In Table 1, we compare our result to prior works. Comparing the coefficient of these complexities, we have

$$\underbrace{\sqrt{\sum_{k\in\mathcal{K}}\frac{1}{\Delta_k^2}}}_{\text{Strong quantum oracle}} \leqslant \underbrace{\sum_{k\in\mathcal{K}}\frac{1}{\Delta_k}}_{\text{Weak quantum oracle}} \leqslant \underbrace{\sum_{k\in\mathcal{K}}\frac{1}{\Delta_k^2}}_{\text{Classic oracle}}. \tag{4}$$

Both quantum MAB models enjoy smaller query complexities than that of traditional MAB. Secondly, our quantum query complexity (via the weak quantum oracle Eq.(3)) is larger than that of the strong quantum oracle Eq.(2); in the worst case ours can be $\sqrt{K}$ times larger by the Cauchy-Schmidt inequality. This echoes the fact that our quantum oracle (Eq.(3)) is weaker than the strong oracle in Eq.(2) in the sense that we cannot query multiple arms at the same time.

3. ALGORITHM AND ANALYSIS

In this section, we propose an elimination-based algorithm for BAI with weak quantum oracle in Algorithm 1 and prove its query complexity upper bound in Theorem 1. Before presenting our algorithms, we recall a useful quantum estimator in Lemma 1 adapted from Montanaro [10] and compare it to classic estimators in Remark 1.

LEMMA 1. *For any weak quantum oracle $\mathcal{O}_k$ in Eq.(3), there is a constant $C_1 > 1$ and a quantum estimator $\text{QE}(\mathcal{O}_k,\epsilon,\delta)$ which returns an estimate $\hat{\mu}$ of μ_k such that $\mathbb{P}(\left|\hat{\mu}_k-\mu_k\right|\geqslant\epsilon)\leqslant\delta$ using at most $(C_1/\epsilon)\log(1/\delta)$ queries to $\mathcal{O}_k$ and $\mathcal{O}_k^\dagger$.*

REMARK 1. *To achieve the $\mathbb{P}(\left|\hat{\mu}_k-\mu_k\right|\geqslant\epsilon)\leqslant\delta$ claim in Lemma 1, a classic estimator (e.g., empirical mean) needs $O((1/\epsilon^2)\log(1/\delta))$ (e.g., via Hoeffding's inequality). The quantum estimator* QE *enjoys a quadratic speedup in query complexity regarding parameter ϵ. However, this* QE *is not as flexible as a classic estimator: Before applying the* QE*, one cannot get any classic information of the reward mean (since all are in quantum superpositions), and after applying the* QE*, all utilized quantum superpositions collapse and cannot be reused anymore, while, for classic samples, one can improve the estimation gradually as in the sample accumulation process, and these samples can be reused freely.*

Recall that the main idea of elimination algorithms is to maintain a candidate arm set $\mathcal{C}$ (initiated as the full arm set $\mathcal{K}$), gradually eliminate identified suboptimal arms from the candidate arm set $\mathcal{C}$ as the learning proceeds, and stop when the candidate arm set $\mathcal{C}$ only containing one arm which is the output optimal arm. We note that although there were some elimination algorithms proposed for BAI with classic oracle, e.g., successive elimination [5], one cannot obtain a feasible quantum algorithm for BAI with the weak oracle by replacing these known elimination algorithms' classic estimator with the quantum estimator in Lemma 1 due to the inconvenience of quantum estimator mentioned in Remark 1.

One key challenge of designing our quantum algorithm is to decide when to execute quantum estimation QE and arm elimination. To address the challenge, we propose a phase-based (batched) exploration and elimination scheme. In each phase, we uniformly explore (query) all remaining arms in candidate arm set $\mathcal{C}$ for a number of times (Line 4), conduct QE to estimate reward means of these arms based on these queries (Line 5), and eliminate the newly identified suboptimal arms (Line 7) at end of the phase. As phase increases, we gradually increase the number of queries and the estimation accuracy of QE (Lines 4 and 8). We analyze the query complexity upper bound of Algorithm 1 in Theorem 1.

THEOREM 1. *Given a confidence parameter $\delta\in(0,1)$, the query complexity of Algorithm 1 is upper bounded as follows,*

$$\mathbb{E}[Q]\leqslant\sum_{k\in\mathcal{K}}\log_2\left(\frac{4}{\Delta_k}\right)\frac{16C_1}{\Delta_k}\ln\frac{K}{\delta}.$$

PROOF OF THEOREM 1. **Correctness:** Note that if all estimates of QE are correct, i.e., $\mu_k\in(\hat{\mu}_k-r,\hat{\mu}_k+r)$ always hold for all arms in $\mathcal{C}$, then the final output arm must be

Algorithm 1 Quantum elimination for BAI

```
1: Input: confidence parameter $\delta$ and arm number $K$
2: Initialize: empirical mean $\hat{\mu}_k \leftarrow 0$, candidate arm set $\mathcal{C} \leftarrow \mathcal{K}$, confidence width $r \leftarrow 1/2$
3: while $|\mathcal{C}| > 1$ do
4:     Query each arm in $\mathcal{C}$ for $\frac{C_1}{r} \log \frac{|\mathcal{C}|}{r\delta}$ times
5:     Run QE$\left(\mathcal{O}, r, \frac{r\delta}{|\mathcal{C}|}\right)$ for each arm in $\mathcal{C}$ and update these arms' estimates $\hat{\mu}_k$
6:     $\hat{\mu}_{\max} \leftarrow \max_{k\in\mathcal{C}} \hat{\mu}_k$
7:     $\mathcal{C} \leftarrow \mathcal{C} \setminus \{k \in \mathcal{C} : \hat{\mu}_k + 2r < \hat{\mu}_{\max}\}$        ▷ Elimination
8:     $r \leftarrow r/2$
9: Output: the remaining arm in $\mathcal{C}$.
```

the true optimal arm. Hence, we only need the probability that any of these QEs fail to be upper bounded by δ. Denote phase index p as the number of times that the while loop of Algorithm 1 runs. Then, we have $r = 2^{-p}$. In p^{th} round, the probability that any QE estimate fails is upper bounded by $|\mathcal{C}| \times \frac{r\delta}{|\mathcal{C}|} = 2^{-p}\delta$. Therefore, the total failure probability over all rounds is upper bounded as follows $\sum_{p=1}^{P} 2^{-p}\delta \leqslant \sum_{p=1}^{\infty} 2^{-p}\delta = \delta$. This fulfills the fixed confidence requirement.

Query Complexity: Since the failure of QE estimate are all taken account by the fixed confidence above, in this part we assume that $\mu_k \in (\hat{\mu}_k - r, \hat{\mu}_k + r)$ always holds for all arms in $\mathcal{C}$. Fix a suboptimal arm k. Denote p_k as the phase that arm k is eliminated. We show that this arm must have been eliminated when $4r < \Delta_k$. Otherwise (this arm is not removed), it would mean that

$$\mu_k + 4r \overset{(a)}{\geqslant} \hat{\mu}_k + 3r \overset{(b)}{\geqslant} \hat{\mu}_{\max} + r \geqslant \hat{\mu}_{k_*} + r \overset{(c)}{\geqslant} \mu_{k_*},$$

where inequalities (a) and (c) are due to the confidence interval $\mu_k \in (\hat{\mu}_k - r, \hat{\mu}_k + r)$, and inequality (b) is because the elimination condition of Line 7 does not hold. That is, if the arm is not eliminated, we have $4r \geqslant \mu_{k_*} - \mu_k = \Delta_k$, which contradicts $4r < \Delta_k$. Therefore, we know that in the phase before the arm k eliminated, i.e., phase $(p_k - 1)$, we have $4r = 4 \cdot 2^{-(p_k-1)} \geqslant \Delta_k$.

After rearrangement, we have $2^{p_k} \leqslant 8/\Delta_k$. So, we can bound the query times of this arm k as follows,

$$\sum_{p=1}^{p_k} \frac{C_1}{r} \ln \frac{|\mathcal{C}|}{r\delta} \leqslant \sum_{p=1}^{p_k} \frac{C_1}{2^{-p}} \ln \frac{K}{2^{-p}\delta} \leqslant C_1 \ln \frac{K}{\delta} \log_2\left(\frac{4}{\Delta_k}\right) \frac{16}{\Delta_k}.$$

Summing the query times of all arms concludes the proof. □

Remark 2. *Compared to the classic oracle's sample complexity upper bound $O(\sum_{k\in\mathcal{K}}(1/\Delta_k)^2 \log(1/\delta))$ [7], the query complexity upper bounds in Theorems 1 has a quadratic improvement of the dependence on $1/\Delta_k$ for each individual arm. For another thing, the strong quantum oracle's sample complexity upper bound $\tilde{O}(\sqrt{\sum_k 1/\Delta_k^2} \log(1/\delta))$ [13] enjoys an overall quadratic speedup over all arms. That is, as the first inequality of Eq.(4) shows, the coefficient of query complexity lower bound of weak oracle is larger than that of strong oracle, and in the worst case, can be $\sqrt{K}$ times larger.*

4. FUTURE DIRECTIONS

Besides BAI with fixed confidence studied in this paper and Wang et al. [13], the BAI with fixed budget with the quantum oracles is another interesting objective to study, which is less understood even in classic MAB [2]. For another thing, although the query complexity of weak oracle is worse than that of strong oracle, the implementation cost of a strong oracle ($O(\log_2 K)$ qubits) can be far more expensive than that of a weak oracle (1 qubit) due to the difficulty of building large quantum circuits. Therefore, it is also worth to studying the query cost complexity of both oracles

Acknowledgement

The work of John C.S. Lui was supported in part by the RGC's GRF 14207721 and SRFS2122-4S02.

5. REFERENCES

[1] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[2] A. Barrier, A. Garivier, and G. Stoltz. On best-arm identification with a fixed budget in non-parametric multi-armed bandits. In *International Conference on Algorithmic Learning Theory*. PMLR, 2023.

[3] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

[4] B. Casalé, G. Di Molfetta, H. Kadri, and L. Ralaivola. Quantum bandits. *Quantum Machine Intelligence*, 2(1):1–7, 2020.

[5] E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

[6] L. K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, pages 212–219, 1996.

[7] Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

[8] T. L. Lai, H. Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[9] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[10] A. Montanaro. Quantum speedup of monte carlo methods. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2181):20150301, 2015.

[11] P. W. Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 124–134. IEEE, 1994.

[12] Z. Wan, Z. Zhang, T. Li, J. Zhang, and X. Sun. Quantum multi-armed bandits and stochastic linear bandits enjoy logarithmic regrets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[13] D. Wang, X. You, T. Li, and A. M. Childs. Quantum exploration algorithms for multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10102–10110, 2021.