# Quantum Best Arm Identification: Query Complexity Lower Bound and Near-optimal Algorithm

**Xuchuang Wang[1], Yu-Zhen Janice Chen[1], Matheus Guedes de Andrade[1], Jonathan Allcock[2], Mohammad Hajiesmaili[1], John C.S. Lui[3], and Don Towsley[1]**

[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, Massachusetts, USA
[2]Tencent Quantum Laboratory, Tencent, Hong Kong
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
{xuchuangwang, yuzhenchen, mguedesdeand, hajiesmaili, towsley}@cs.umass.edu,
jonallcock@tencent.com, cslui@cse.cuhk.edu.hk

**Abstract**

Best arm identification is an important problem of multi-armed bandits, where each arm is associated with a reward distribution and, when querying one arm, one obtains a reward sampled from the queried arm's distribution. In this paper, we study this problem in the *quantum* multi-armed bandits model, where the reward distributions of classical bandits are replaced with *quantum oracles*, and the reward samples are replaced with quantum states generated by the quantum oracles. We first prove a query complexity lower bound for the quantum multi-armed bandits model, and then devise an algorithm whose query complexity upper bound matches the lower bound up to some logarithmic factors. Compared to the classical best arm identification sample complexity bound of $O(\Delta^{-2}\log(1/\delta))$, our quantum query complexity bounds is $O(\Delta^{-1}\log(1/\delta))$, improving the dependence of $\Delta$ from $\Delta^{-2}$ to $\Delta^{-1}$, where $\delta$ is the given confidence parameter and $\Delta\,(<1)$ is the mean reward gap between optimal and sub-optimal arms. We also extend our complexity bound analysis and algorithm design to the multiple top arms identification problem. Lastly, we conduct numerical experiments to corroborate our theoretical results.

## 1 Introduction

Recent progress in building quantum computers (Arute et al. 2019; Chow, Dial, and Gambetta 2021) and quantum networks (Wehner, Elkouss, and Hanson 2018; Azuma et al. 2022) envisages wide applications of quantum systems in the near future. With the advantage of quantum computation, one can speed up not only fundamental algorithms, e.g., unstructured search (Grover 1996) and factoring (Shor 1994), but recent machine learning algorithms (Biamonte et al. 2017) as well. In this paper, we study the quantum speedup of a canonical task of reinforcement learning in quantum system—best arm identification in multi-armed bandits with quantum oracles.

The multi-armed bandit (MAB) model—first studied by Lai and Robbins (1985)—is a well-established sequential decision making model (ref., (Lattimore and Szepesvári 2020; Slivkins et al. 2019)). In the stochastic case, a MAB consists of $K$ arms, each of which is associated with an unknown reward distribution. When *querying* an arm $k \in \mathcal{K} := \{1, 2, \ldots, K\}$, one obtains a reward drawn from a reward distribution $\mathcal{B}(\mu_k)$, i.e.,

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

which we assume to be a Bernoulli distribution with unknown mean $\mu_k$. This assumption can be straightforwardly extended, as shown in the MAB literature. We refer to Eq. (1) as the *classical oracle*.

Two kinds of quantum oracles have been proposed for MAB: Wang et al. (2021) proposed an oracle that encodes the Bernoulli reward distributions of all arms 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 with basis states $|1\rangle$ and $|0\rangle$. On the other hand, Wan et al. (2023) proposed another oracle that encodes the reward distribution of each arm $k$ into separate oracles as follows,

**(Weak quantum oracle)**

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

where the output is a single-qubit "bandit reward" superposition. The oracle in Eq. (2) is more powerful than the oracle in Eq. (3) since the former allows one to access all arms coherently in a single query, while the latter is restricted to individual arm access per query. For example, when the arm index register is queried in a uniform superposition of the arm indices $\sum_{k\in\mathcal{K}}(1/\sqrt{K})\,|k\rangle_I\,|0\rangle_R$, the oracle of Eq. (2) returns $\sum_{k\in\mathcal{K}}(1/\sqrt{K})\,|k\rangle_I \left(\sqrt{\mu_k}\,|1\rangle_R + \sqrt{1-\mu_k}\,|0\rangle_R\right)$ in which the qubit in register $R$ encodes the information of all arms' reward distributions. We refer to the oracle in Eq. (2) as the *strong quantum oracle*, and the one in Eq. (3) as the *weak quantum oracle*.

We emphasize that, in several interesting settings that can be modelled by quantum MAB, it may be unrealistic to assume coherent access to all arms, whereas the weak oracle assumption may still apply. For example, the weak oracle model can be used to select paths in a quantum network under the assumption that paths introduce unitary noise. In this setting, the paths are modelled by the weak oracle, and the objective is to find the path that introduces the least amount

of noise. On the other hand, the strong oracle model may be unrealistic in this case, as coherent path access would necessitate complex, macroscopic entanglement among network paths (which could, in principle, be operated by different network providers). Another motivating example is where the oracles correspond to quantum solvers offered by different providers (e.g., IBM, Google, etc.) to solve the same problem. One may wish to determine which algorithm succeeds with the highest probability, but it is not plausible to assume coherent access to the different backend quantum computers.

Prior works on the MAB problem typically focus on regret minimization and best arm identification (BAI, a.k.a., pure exploration). In this paper, we focus on the BAI setting introduced to MAB by Even-Dar, Mannor, and Mansour (2002); Even-Dar et al. (2006) and Mannor and Tsitsiklis (2004). The BAI setting can be further divided into two categories based on the learning objective: (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 samples as possible (Bubeck, Munos, and Stoltz 2011); and (2) BAI with fixed budget—given a fixed budget of query times $Q$, find the best arm with as large a probability as possible (Karnin, Koren, and Somekh 2013). In this paper, we focus on the former category, and for brevity, hereinafter refer to it as the BAI problem. We study BAI under the weak quantum oracle setting in Eq. (3).

### 1.1 Contributions

In Section 3, and inspired by quantum hypothesis testing (Kargin 2005), we first derive a query complexity lower bound of $\Omega\left(\sum_k(1/\Delta_k)\ln(1/\delta)\right)$ for BAI with the weak quantum oracle of Eq. (3), where $\Delta_k \coloneqq \mu_1-\mu_k$ is the difference in the rewards of the optimal arm and arm $k$. The proof of this lower bound is a non-trivial combination of three key components: a quantum hypothesis testing lower bound (Lemma 1) (Kargin 2005), a mapping of a lower bound on the failure probability of testing any two quantum states to the query complexity of distinguishing two quantum MAB instances (Lemma 2), and an application of hypothesis testing to best arm identification (following the classical BAI lower bound proof (Mannor and Tsitsiklis 2004)). We also derive another query complexity lower bound $\Omega(\sum_k(1/\Delta_k)(1-\sqrt{\delta}))$ based on the quantum adversary method (Ambainis 2000) (Appendix A). Both highlight the impact of the term $\sum_k(1/\Delta_k)$ on the query complexity lower bound. Further, we extend the above lower bounds for BAI to the top $m$ arms identification (TMI) problem.

In Section 4, we propose near-optimal algorithms to address BAI and TMI whose query complexity upper bounds are tight up to some logarithmic factors. We devise an elimination-based algorithm (Q-Elim) for BAI (Section 4.1) and prove that this algorithm enjoys a query complexity upper bound of $\tilde{O}\left(\sum_k(1/\Delta_k)\ln(1/\delta)\right)$. We also devise a gap-based exploration algorithm (Q-GapE) for TMI (Section 4.2) which enjoys an upper bound of $\tilde{O}(\sum_k(1/\Delta_k^{\langle m\rangle})\ln(1/\delta))$, where $\Delta_k^{\langle m\rangle}$ is the reward gap between arm $k$ and the $m^{\text{th}}$ (or $(m+1)^{\text{th}}$)-largest arm (see the definition in Eq. (5)). Both upper bounds match their corresponding lower bounds up to some logarithmic factors. We emphasize that existing BAI and TMI algorithms with classical oracles do not extend in a straightforward manner to address BAI and TMI with weak quantum oracles, as these classical algorithms rely on the flexible utilization of classical estimators, which are not feasible when utilizing quantum estimators (see Remark 2 for more details). Lastly, we conduct experiments to corroborate the superiority of our quantum algorithms over classical ones (Section 5).

In Table 1, we summarize the key results in this paper and compare them to prior works. Comparing the coefficient of these complexities, we have

$$\underbrace{\sqrt{\sum_k\frac{1}{\Delta_k^2}}}_{\text{Strong quantum oracle}}\leqslant\underbrace{\sum_k\frac{1}{\Delta_k}}_{\text{Weak quantum oracle}}\leqslant\underbrace{\sum_k\frac{1}{\Delta_k^2}}_{\text{Classical oracle}}. \tag{4}$$

Both quantum MAB models enjoy smaller query complexities than that of classical 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, it 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 $\mathcal{O}_{\text{stro}}$ in the sense that we cannot explore multiple arms at the same time.

### 1.2 Related Works

For MAB with the strong quantum oracle, Wang et al. (2021) proposed an algorithm that enjoys a quadratic speedup in the query complexity for best arm identification with a fixed confidence setting. For MAB with the weak quantum oracle, Wan et al. (2023) devised regret minimization algorithms for both multi-armed bandits and linear bandits with quantum reward oracles that achieve an $O(\ln T)$ problem-independent upper bound, while with classical oracle in Eq. (1), one only has $O(\sqrt{T})$ problem-independent bounds. Wu et al. (2023) extended the regret minimization results of Wan et al. (2023) to bandits with heavy-tailed quantum rewards. In this paper, we are the first to study the BAI problem with the weak quantum oracle.

Besides MAB with quantum oracles (Casalé et al. 2020; Wang et al. 2021; Wan et al. 2023; Wu et al. 2023) literature, there are other interdisciplinary works involved with multi-armed bandits and quantum computation (Lumbreras, Haapasalo, and Tomamichel 2022; Brahmachari, Lumbreras, and Tomamichel 2023; Ohno 2023; Buchholz, Kübler, and Schölkopf 2023; Naruse et al. 2019; Cho et al. 2022). For example, Lumbreras, Haapasalo, and Tomamichel (2022); Brahmachari, Lumbreras, and Tomamichel (2023) applied the classical bandit algorithms to learn properties of quantum states and recommend quantum states. Ohno (2023) applied quantum maximization and amplitude encoding to speed up the classical $\epsilon$-greedy algorithm in MAB. Cho et al. (2022) proposed quantum amplitude amplification exploration algorithm for adversarial MAB. Naruse et al. (2019) built a physical quantum system (based on photons) to implement the classical MAB algorithms.

Table 1: Comparison of query complexity bounds with classical and quantum BAI

| | Lower Bound | Upper Bound |
|---|---|---|
| Classical oracle (1) | $\Omega\left(\sum_k \frac{1}{\Delta_k^2} \ln \frac{1}{\delta}\right)$ (Mannor and Tsitsiklis 2004) | $O\left(\sum_k \frac{1}{\Delta_k^2} \ln \frac{1}{\delta}\right)$ (Karnin, Koren, and Somekh 2013) |
| Strong quantum oracle (2) | $\Omega\left(\sqrt{\sum_k \frac{1}{\Delta_k^2}}\right)$ (Wang et al. 2021) | $\tilde{O}\left(\sqrt{\sum_k \frac{1}{\Delta_k^2}} \ln\left(\frac{1}{\delta}\right)\right)$ (Wang et al. 2021) |
| Weak quantum oracle (3) (ours) | $\Omega\left(\sum_k \frac{1}{\Delta_k} \ln\left(\frac{1}{\delta}\right)\right)$ (Theorem 1) | $\tilde{O}\left(\sum_k \frac{1}{\Delta_k} \ln\left(\frac{1}{\delta}\right)\right)$ (Theorem 3) |

## 2 Model

**Bra-ket notation.** We make use of quantum bra-ket notation, where ket $|x\rangle := (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$ denotes a column vector, while the bra $\langle x| := |x\rangle^\dagger = (x_1^*, x_2^*, \ldots, x_n^*)$, a row vector, is the conjugate transpose of the ket $|x\rangle$. The inner product between $|x\rangle \in \mathbb{C}^n$ and $|y\rangle \in \mathbb{C}^n$ is defined as $\langle x|y\rangle := \langle x| \cdot |y\rangle = \sum_{i=1}^n x_i^* y_i \in \mathbb{C}$, and the tensor product between $|x\rangle \in \mathbb{C}^n$ and $|z\rangle \in \mathbb{C}^m$ is defined as $|x\rangle |z\rangle := |x\rangle \otimes |z\rangle = (x_1 z_1, x_1 z_2, \ldots, x_n z_m) \in \mathbb{C}^n \otimes \mathbb{C}^m$.

**Quantum query model.** In this model, one can access a black-box function implemented by a quantum oracle. The objective is to study the *query complexity*, i.e., the number of calls (denoted as $Q$) to this oracle that are needed to solve a given task; all other costs except for querying the oracle are ignored. This is a commonly used model for studying quantum algorithms (Childs 2017, §20) and can be used, for instance, to obtain algorithmic running time lower bounds. In this paper, we study the query complexity of best arm identification with fixed confidence using the weak quantum oracle of Eq. (3).

**Quantum MAB model.** Consider a $K$-armed bandit, where each arm $k$ is associated with a Bernoulli distribution with mean $\mu_k \in (0, 1)$, i.e., $\mathcal{B}(\mu_k)$. When querying one arm $k$, instead of obtaining a sample drawn from $\mathcal{B}(\mu_k)$, we consider the case that one obtains a qubit in state $|\psi_k\rangle := \sqrt{\mu_k}\,|1\rangle + \sqrt{1-\mu_k}\,|0\rangle$. Formally, each arm $k \in \mathcal{K}$ is associated with a weak quantum oracle $\mathcal{O}_k$ defined in Eq. (3). This weak quantum oracle models the query feedback of classical MAB in Eq. (1) as quantum superpositions. Measuring $|\psi_k\rangle$ yields $|1\rangle$ with probability $\mu_k$ and $|0\rangle$ otherwise, which is equivalent to classical oracle in Eq. (1), i.e., drawing a sample from a Bernoulli distribution $\mathcal{B}(\mu_k)$. Although this weak oracle does not provide an opportunity to simultaneously explore multiple arms coherently as the strong oracle in Eq. (2) does, it is still more informative than the classical oracle in Eq. (1) because its superposition output encodes the information of the whole reward distribution, instead of a single reward sample as in the classical oracle. In this paper, we show that the weak quantum oracle, followed by some quantum computations, outperforms classical MAB algorithms. For simplicity, we assume the $K$ arms are ordered in descending order of their means: $\mu_1 > \mu_2 > \cdots > \mu_K$, unknown to the learner. A quantum MAB instance is determined by the reward means of its arms, and we denote an arbitrary instance $\mathcal{I}$ with means $\mu_1, \ldots, \mu_K$ as $\mathcal{I} := \{\mu_1, \ldots, \mu_K\}$.

**Best arm (and top $m$ arms) identification.** Given $\delta \in (0, 1)$, design an algorithm that minimizes the number of queries required to correctly output the best arm (top $m$ arms) with a probability of at least $1 - \delta$. To express the query complexity, we denote the reward mean (suboptimality) gap as

$$\Delta_k^{\langle m \rangle} := \begin{cases} \mu_k - \mu_{m+1} & \text{if } k \leqslant m \\ \mu_m - \mu_k & \text{if } k > m \end{cases}, \tag{5}$$

where $m = 1$ corresponds to the BAI. For brevity, we denote $\Delta_k := \Delta_k^{\langle 1 \rangle}$.

## 3 Lower Bound

In this section, we start by reviewing a quantum hypothesis testing lower bound (Lemma 1), and then apply this lower bound to distinguish two quantum MAB models (Lemma 2). Next, based on the previous two lower bounds, we derive query complexity lower bounds for the best arm identification (BAI) (Theorem 1) and the top $m$ arms identification (TMI) problems (Theorem 2).

### 3.1 Preliminary Lower Bounds

Quantum hypothesis testing (Holevo 2003, §2.2) aims to solve the following problem: Given multiple copies of one of two known quantum states, $|\psi_0\rangle, |\psi_1\rangle$, determine which of both states has been given. We focus on pure states in this paper, and consider the case that the quantum superposition of both hypotheses are $|\psi_0\rangle$ and $|\psi_1\rangle$. Lemma 1 presents a failure probability lower bound of testing whether the quantum pure state is $|\psi_0\rangle$ or $|\psi_1\rangle$, given $Q$ quantum copies.

**Lemma 1** (Error probability lower bound for quantum pure state hypothesis testing (Kargin 2005))**.** *Given $Q$ copies of one of two pure quantum states, $|\psi_0\rangle$ or $|\psi_1\rangle$ (equal prior), the smallest error probability of deciding which state has been given is*

$$p_{error}^{(Q)} \geqslant \frac{1}{2}\left(1 - \sqrt{1 - |\langle\psi_0|\psi_1\rangle|^{2Q}}\right).$$

Next, we extend the hypothesis testing between two quantum pure states to distinguishing two quantum MAB instances $\mathcal{I}_0 = (\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_K^{(0)})$ and $\mathcal{I}_1 = (\mu_1^{(1)}, \mu_2^{(1)}, \ldots, \mu_K^{(1)})$. We consider the case where both instances have only one arm $\ell$ where the reward oracles differ:

$$\mathcal{O}_\ell^{(0)} : |0\rangle \to \sqrt{1-\mu_0}\,|0\rangle + \sqrt{\mu_0}\,|1\rangle,$$
$$\mathcal{O}_\ell^{(1)} : |0\rangle \to \sqrt{1-\mu_1}\,|0\rangle + \sqrt{\mu_1}\,|1\rangle,$$

that is, $\mu_\ell^{(0)} = \mu_0 \neq \mu_1 = \mu_\ell^{(1)}$, and all other arm reward means are the same, i.e., $\mu_k^{(0)} = \mu_k^{(1)}$ for all $k \neq \ell$. As we are interested in query complexity, we transfer the failure probability bound in Lemma 1 to query complexity in Lemma 2 as follows.

**Lemma 2** (Query complexity lower bound for distinguishing two quantum MAB instances differing by exactly one arm's reward mean)**.** *Given $\mu_0, \mu_1 \in (0, 1/2)$[1], the necessary number of queries $Q$ to distinguish the quantum* MAB *instances $\mathcal{I}_0$ and $\mathcal{I}_1$, with a probability of at least $1-\delta$, is lower bounded by*

$$Q \geqslant \frac{1}{4|\mu_0 - \mu_1|} \ln \frac{1}{4\delta}.$$

*Proof of Lemma 2.* **Step 1. Relax the task to quantum hypothesis testing.** We begin with an easier task than distinguishing two quantum MAB instances. We assume that the reward mean parameters of both instances are known a priori, that is, the arm index $\ell$ whose reward mean is different in instance $\mathcal{I}_0$ and $\mathcal{I}_1$ and the values of $\mu_0$ and $\mu_1$ are known. To address the relaxed task, one only needs to pull arm $\ell$ and test whether the reward mean of this arm is $\mu_0$ or $\mu_1$. We note that with the above additional information, the task becomes easier, and, hence, the query complexity lower bound of this relaxed task also serves as a lower bound for the original task.

**Step 2. Calculate the query complexity lower bound from the quantum hypothesis testing result.** Let $\sqrt{\mu_0} = \sin\theta_0$ and $\sqrt{\mu_1} = \sin\theta_1$ for $\theta_0, \theta_1 \in (0, \pi/4)$. We can rewrite the quantum states as

$$|\psi_0\rangle \coloneqq \sqrt{1-\mu_0}\,|0\rangle + \sqrt{\mu_0}\,|1\rangle = \cos\theta_0\,|0\rangle + \sin\theta_0\,|1\rangle\,,$$
$$|\psi_1\rangle \coloneqq \sqrt{1-\mu_1}\,|0\rangle + \sqrt{\mu_1}\,|1\rangle = \cos\theta_1\,|0\rangle + \sin\theta_1\,|1\rangle\,.$$

Lemma 1 shows that to differentiate both oracles with a probability of a least $1-\delta$, one needs

$$\delta \geqslant p_{\text{error}}^{(Q)} \geqslant \frac{1}{2}\left(1 - \sqrt{1 - |\langle\psi_0|\psi_1\rangle|^{2Q}}\right).$$

After rearranging the above inequality, we have

$$Q \geqslant \frac{\ln\big(1-(1-2\delta)^2\big)}{2\ln|\langle\psi_0|\psi_1\rangle|} = \frac{1}{-2\ln|\langle\psi_0|\psi_1\rangle|} \ln\frac{1}{4\delta(1-\delta)} \tag{6}$$

Algebraic calculations (see Appendix B.1) yield $\ln|\langle\psi_0|\psi_1\rangle|^{-1} \leqslant (\theta_0-\theta_1)^2/2$ and $|\mu_0-\mu_1| \geqslant (\theta_0-\theta_1)^2/4$. Substituting both inequalities into Eq. (6) concludes the proof. □

## 3.2 Lower Bounds for Best Arm Identification and Top *m* Arms Identification

Next, we extend the lower bound for distinguishing two quantum MAB instances in Lemma 2 to the BAI problem. The proof follows the approach used to prove the classical BAI sample complexity lower bound (Mannor and Tsitsiklis 2004).

**Theorem 1** (Query complexity lower bound for best arm identification)**.** *Given a quantum multi-armed bandits instance $\mathcal{I}_0 = \{\mu_1, \ldots, \mu_K\}$ where $\mu_k \in (0, 1/2)$ for all $k$ and $\mu_1 > \mu_2 \geqslant \mu_k$ for any $k \neq 1$, the necessary query times $Q$ of any algorithm—that identifies the optimal arm with a given confidence $1-\delta$, $\delta \in (0,1)$—satisfies the following inequality,*

$$Q \geqslant \sum_{k\in\mathcal{K}} \frac{1}{4\Delta_k} \ln\frac{1}{4\delta}.$$

*Proof of Theorem 1.* For any suboptimal arm $k \neq 1$ in instance $\mathcal{I}_0$, we consider another instance $\mathcal{I}_k = \{\mu_1^{(k)}, \ldots, \mu_K^{(k)}\}$ whose reward means are the same as instance $\mathcal{I}_0$ except for the reward mean of arm $k$ which assumes the form $\mu_k^{(k)} = \mu_1 + \epsilon$ for $0 < \epsilon < 1/2 - \mu_1$. Therefore, in instance $\mathcal{I}_k$, the optimal arm is $k \neq 1$.

Because instances $\mathcal{I}_0$ and $\mathcal{I}_k$ have different optimal arms, any feasible policy must be able to distinguish these two instances with a confidence of at least $1-\delta$. Given the additional information that all other arms have the same means, this task reduces to distinguishing two instances $\mathcal{I}_0$ and $\mathcal{I}_k$ as in Lemma 2. Therefore, the query complexity of distinguishing both instances is at least $1/4(\Delta_k+\epsilon) \ln 1/4\delta$. Note that these queries are all on arm $k$.

For the optimal arm $k = 1$ in instance $\mathcal{I}_0$, we consider another instance $\mathcal{I}_1 = \{\mu_1^{(1)}, \ldots, \mu_K^{(1)}\}$ whose oracles are the same as instance $\mathcal{I}_0$ except that the reward mean of the arm 1 in $\mathcal{I}_1$ is $\mu_1^{(1)} = \mu_2 - \epsilon$ for $0 < \epsilon < \mu_2$ and recall that arm 2 is the second best arm in $\mathcal{I}_0$. Therefore, in instance $\mathcal{I}_1$, the optimal arm is $2 \neq 1$. Similarly, applying Lemma 2 to distinguishing instances $\mathcal{I}_0$ and $\mathcal{I}_1$, the query complexity is lower bounded by $\big(1/4(\Delta_2+\epsilon)\big) \ln 1/4\delta = \big(1/4(\Delta_1+\epsilon)\big) \ln 1/4\delta$.

Last, summing the least query complexity spent on each arm and letting $\epsilon$ go to zero yield $Q \geqslant \sum_{k\in\mathcal{K}} 1/4\Delta_k \ln 1/4\delta$. □

We further extend the result to top $m$ arms identification in Theorem 2 with a proof in Appendix B.2.

**Theorem 2** (Query complexity lower bound for top $m$ arms identification)**.** *Given a quantum multi-armed bandits instance $\mathcal{I}_0 = \{\mu_1, \mu_2, \ldots, \mu_K\}$ where $\mu_k \in (0, 1/2)$ for all $k$ and $\mu_{k_1} > \mu_{k_2}$ for any $k_1 \leqslant m$ and $k_2 > m$, any algorithm—that identifies the set of top $m$ arms with a given confidence $1-\delta$ where $\delta \in (0,1)$—satisfies the following inequality,*

$$Q \geqslant \sum_{k\in\mathcal{K}} \frac{1}{2\Delta_k^{\langle m\rangle}} \ln\frac{1}{4\delta}.$$

*Remark* 1 (Comparison to the lower bounds of classical and strong quantum oracles). Compared to the classical sample complexity lower bound $\Omega(\sum_{k\in\mathcal{K}}(1/\Delta_k)^2 \ln(1/\delta))$ (Mannor and Tsitsiklis 2004), the query complexity lower bounds in Theorems 1 and 2 transform the quadratic dependence on $1/\Delta_k$ to a linear dependence. Compared to the strong quantum oracle's sample complexity lower bound

[1]This assumption is needed in the algebraic calculations in Appendix B.1, and such a constant constraint assumption is common for lower bound results, e.g., (Mannor and Tsitsiklis 2004, Theorem 1).

$\Omega(\sqrt{\sum_k 1/\Delta_k^2}(1-\sqrt{\delta(1-\delta)}))$ (Wang et al. 2021, Theorem 5), the coefficient of the query complexity lower bound of the weak oracle is larger than that of the strong oracle, as shown by the first inequality of Eq. (4), and is, in the worst case, $\sqrt{K}$ times larger. Nevertheless, our lower bound has a better dependence on $\delta$, since $\ln(1/\delta) \gg (1-\sqrt{\delta(1-\delta)})$ when $\delta$ is small.

# 4 Algorithm

In this section, we propose two algorithms for BAI and top $m$ arms identification (TMI) respectively. We first recall the following useful result.

**Lemma 3** (Adapted from Montanaro (2015))**.** *For any weak quantum oracle $\mathcal{O}_k$ in Eq.* (3)*, there is a constant $C_1 > 1$ and a quantum estimate algorithm $\mathrm{QE}(\mathcal{O}, \epsilon, \delta)$ which returns an estimate $\hat{\mu}_k$ of $\mu_k$ such that $\mathbb{P}(|\hat{\mu}_k - \mu_k| \geqslant \epsilon) \leqslant \delta$ using at most $(C_1/\epsilon)\log(1/\delta)$ queries to $\mathcal{O}_k$ and $\mathcal{O}_k^\dagger$.*

*Remark* 2 (Comparison to classical estimator). To achieve the $\mathbb{P}(|\hat{\mu}_k - \mu_k| \geqslant \epsilon) \leqslant \delta$ claim in Lemma 3, a classical estimator (e.g., empirical mean) needs $O((1/\epsilon^2)\log(1/\delta))$ (e.g., via Hoeffding's inequality). Compared to the classical estimator, the quantum estimator QE enjoys a quadratic speedup in query complexity regarding parameter $\epsilon$. However, QE is not as flexible as the classical estimator. Before the QE procedure runs to completion, one cannot obtain any partial information of the reward mean, since that information is stored in quantum states that can only be accessed through measurements. In the case of a classical estimator, one can improve the estimate gradually as samples accumulate, and these samples can be reused freely.

*Remark* 3 (Generality of the quantum estimator algorithm and our BAI and TMI algorithms). We note that Lemma 3 is a special case of (Montanaro 2015, Theorems 3 and 5). With more general quantum estimators, we can extend algorithms in this section that apply to Bernoulli random variable rewards to any random variables with bounded variance.

## 4.1 Algorithm for Best Arm Identification

We devise a quantum elimination algorithm (Q-Elim) for BAI (Algorithm 1). Recall that the main idea of an elimination algorithm is to maintain a candidate arm set $\mathcal{C}$ (initiated as the full arm set $\mathcal{K}$), gradually identify and eliminate suboptimal arms from $\mathcal{C}$ as learning proceeds, and stop when $\mathcal{C}$ contains only one arm, which is then output as the optimal arm. Note that, while several elimination algorithms for BAI using classical oracles have been proposed, such as successive elimination (Even-Dar et al. 2006), it is not feasible to apply these algorithms by directly replacing classical estimators with the quantum estimator of Lemma 3. This limitation arises from the inherent rigidity of the quantum estimator discussed in Remark 2.

One key challenge in designing our quantum algorithm is to decide when to execute quantum estimation QE and arm elimination. To address this challenge, we propose a batch-based exploration and elimination scheme, where we use $p \in \{1, 2, \dots\}$ to denote batch number. In each batch,

Algorithm 1: Q-Elim: Quantum elimination for BAI

1: **Input:** fixed confidence parameter $\delta$ and number of arms $K$
2: **Initialize:** empirical mean $\hat{\mu}_k \leftarrow 0$, candidate arm set $\mathcal{C} \leftarrow \mathcal{K}$, batch number $p \leftarrow 1$
3: **while** $|\mathcal{C}| > 1$ **do**
4: Query each arm in $\mathcal{C}$ for $C_1 2^p \log(2^p|\mathcal{C}|/\delta)$ times
5: Run $\mathrm{QE}(\mathcal{O}, 2^{-p}, \delta/2^p|\mathcal{C}|)$ 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: $\mathcal{C} \leftarrow \mathcal{C} \setminus \{k \in \mathcal{C} : \hat{\mu}_k + 2\cdot 2^{-p} < \hat{\mu}_{\max}\}$ ▷ Elimination
8: $p \leftarrow p + 1$
9: **Output:** the remaining arm in $\mathcal{C}$.

we uniformly explore (query) all the remaining arms in candidate arm set $\mathcal{C}$ a number of times depending on the batch number $p$ (Line 4), conduct QE to estimate reward means of arms in $\mathcal{C}$ based on queries in this batch (Line 5), and eliminate the newly identified suboptimal arms (Line 7) at the end of the batch. As the batch number $p$ increases, we gradually increase the number of queries (Line 8) and the estimation accuracy of QE (Lines 4 and 5).

In Theorem 3, we analyze the query complexity upper bound of Q-Elim. The upper bound theorem means, with a probability of at least $1-\delta$, Q-Elim terminates before the proved query complexity upper bound and outputs the true optimal arm.

**Theorem 3** (Query complexity upper bound of Algorithm 1)**.** *Given a confidence parameter $\delta \in (0,1)$, the query complexity of* Q-Elim *is upper bounded as follows,*

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

*Proof of Theorem 3.* **Correctness:** Note that if all estimates of QE are correct, i.e., $\mu_k \in (\hat{\mu}_k - 2^{-p}, \hat{\mu}_k + 2^{-p})$ for all arms in $\mathcal{C}$, then the final output arm must be the true optimal arm. Hence, we only need to show that the probability that any of these QE fails is upper bounded by $\delta$.

In the $p^{\text{th}}$ round, the probability that any of the $|\mathcal{C}|$ quantum estimates fails is upper bounded by $|\mathcal{C}| \times {}^{2^{-p}\delta}/_{|\mathcal{C}|} = 2^{-p}\delta$. Therefore, the total failure probability over all rounds is upper bounded by $\sum_{p=1}^{\infty} 2^{-p}\delta = \delta$. This fulfills the fixed confidence requirement.

**Query Complexity:** Since failure of the QE procedures are accounted for by the fixed confidence above, we assume that $\mu_k \in (\hat{\mu}_k - 2^{-p}, \hat{\mu}_k + 2^{-p})$ holds for all arms in $\mathcal{C}$ and prove an upper bound of query times that Q-Elim needs to output the optimal arm.

Consider a complete execution of Algorithm 1. Fix a suboptimal arm $k$. Denote $s_k$ as the batch during which arm $k$ is eliminated. We show that this arm must have been eliminated when $4\cdot 2^{-p} < \Delta_k$. Otherwise this arm is not removed, which implies that

$$\mu_k + 4\cdot 2^{-p} \overset{(a)}{\geqslant} \hat{\mu}_k + 3\cdot 2^{-p} \overset{(b)}{\geqslant} \hat{\mu}_{\max} + 2^{-p} \geqslant \hat{\mu}_1 + 2^{-p} \overset{(c)}{\geqslant} \mu_{k_*},$$

where inequalities (a) and (c) are due to the confidence interval $\mu_k \in (\hat{\mu}_k - 2^{-p}, \hat{\mu}_k + 2^{-p})$, and inequality (b) stems from the fact that the elimination condition of Line 7 does not hold. That is, if the arm is not eliminated, we have $4 \cdot 2^{-p} \geqslant \mu_{k_*} - \mu_k = \Delta_k$, which contradicts $4 \cdot 2^{-p} < \Delta_k$. Therefore, assuming the last batch that the arm $k$ is queried is $s_k$, and we have $4 \cdot 2^{-s_k} \geqslant \Delta_k$. After rearrangement, we have $2^{s_k} \leqslant 4/\Delta_k$. So, we can bound the query times of this arm $k$ as follows,

$$\sum_{p=1}^{s_k} C_1 2^{-p} \ln \frac{K}{2^{-p}\delta} \leqslant C_1 \ln \frac{2K}{\delta} \sum_{p=1}^{s_k} 2^p (p+1)$$
$$\leqslant C_1 \ln \frac{2K}{\delta} \cdot (s_k+1) 2^{s_k+1} \leqslant C_1 \ln \frac{2K}{\delta} \log_2 \left(\frac{4}{\Delta_k}\right) \frac{16}{\Delta_k}.$$

Last, summing the query times of all arms concludes the proof. □

*Remark* 4 (Optimality of query complexity upper bound). Compared to the query complexity lower bound in Theorem 1, our upper bound in Theorem 3 is tight up to some logarithmic factor.

*Remark* 5 (Comparison to upper bounds of classical and strong quantum oracles). Compared to the classical oracle sample complexity upper bound $O(\sum_{k\in\mathcal{K}}(1/\Delta_k)^2 \log(1/\delta))$ (Karnin, Koren, and Somekh 2013), the query complexity upper bound in Theorem 3 has a quadratic improvement in the dependence on $1/\Delta_k$ for each individual arm. In contrast, the strong quantum oracle sample complexity upper bound $\tilde{O}(\sqrt{\sum_k 1/\Delta_k^2}\log(1/\delta))$ (Wang et al. 2021) 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 the weak quantum oracle is larger than that of the strong oracle, and is, in the worst case, $\sqrt{K}$ times larger.

We note that `Q-Elim` in Algorithm 1 for `BAI` cannot be extended to address `TMI` by simply changing the while-loop condition $|\mathcal{C}| > 1$ in Line 3 to $|\mathcal{C}| > m$, because such an algorithm cannot guarantee to output the top $m$ arms with a confidence of at least $1-\delta$. For example, when the reward means of the $m^{\text{th}}$-best arm and the $(m+1)^{\text{th}}$-best arm are close, the extended `Q-Elim` can easily make a mistake and eliminate the $m^{\text{th}}$-best arm instead of $(m+1)^{\text{th}}$-best arm, whose failure probability is not taken into account by the current analysis and can be significantly larger than $\delta$.

## 4.2 Algorithm for Top *m* Arms Identification

In this subsection, we propose the quantum gap-based exploration algorithm (`Q-GapE`) for `TMI` using the weak quantum oracle in Algorithm 2. This algorithm adapts gap-based exploration (Gabillon, Ghavamzadeh, and Lazaric 2012) to a batched version so that one can apply the quantum estimator for reward mean estimate.

We illustrate the gap-based exploration. Let $B_k := \max_{i\neq k}^m(\hat{\mu}_i + 2^{-p_i}) - (\hat{\mu}_k - 2^{-p_k})$ denote the gap, where $\max_{i\neq k}^m$ means outputting the $m^{\text{th}}$-largest quantity among all quantities referred to in index range $\mathcal{K}\setminus\{k\}$, and $p_k \in \{1,2,\dots\}$ denotes the batch number for exploring the arm $k$. $B_k \leqslant 0$ means that the lower confidence bound $\hat{\mu}_k - 2^{-p_k}$ of arm $k$ is no less than the $m^{\text{th}}$-largest upper confidence bound among arms other than $k$, which implies that arm $k$ is among the top $m$ arms with high probability; the smaller $B_k$ is, the higher the assurance. Let $\mathcal{H} := \arg\min_{k\in\mathcal{K}}^{1\dots m} B_k$ denote the set of $m$ arm indexes with the smallest $B_k$'s, where $\arg\min_{k\in\mathcal{K}}^{1\dots m}$ represents the $m$ indices with the minimum $m$ quantities referred to in the index range $\mathcal{K}$. Arm set $\mathcal{H}$ serves as the estimated top $m$ arms set in the algorithm. If, for all $k \in \mathcal{H}$, $B_k \leqslant 0$ holds (i.e., $\max_{k\in\mathcal{H}} B_k \leqslant 0$), then we output arm set $\mathcal{H}$ as the identified top $m$ arms fulfilling the required confidence. This leads to the while-loop condition $\max_{k\in\mathcal{H}} B_k > 0$ in Line 3 of Algorithm 2.

Within each iteration of the while-loop, we let arm $h$ with the smallest lower confidence bound among arms in $\mathcal{H}$ denote the estimated $m^{\text{th}}$-best arm, and arm $\ell$ with the largest upper confidence bound among arms in $\mathcal{K}\setminus\mathcal{H}$ denote the estimated $(m+1)^{\text{th}}$-best arm (Line 7). Since both arms are critical for separating the top $m$ arms from the rest, we pick the arm with largest confidence interval width among $h$ and $\ell$ (Line 4) to conduct a batched exploration (Line 5). We present `Q-GapE` in Algorithm 2 and its query complexity upper bound in Theorem 4 with a proof in Appendix B.3.

Algorithm 2: `Q-GapE` : Quantum gap-based exploration for top $m$ arms identification

1: **Input:** fixed confidence parameter $\delta$, number of arms $K$, and top arm set size $m$
2: **Initialize:** empirical means $\hat{\mu}_k \leftarrow 0$ and batch number $p_k \leftarrow 1$ for all arm $k$, empirical top $m$ arms set $\mathcal{H} \leftarrow \{1,2,\dots,m\}$, empirical $m^{\text{th}}$-best arm index $h \leftarrow m$, empirical $(m+1)^{\text{th}}$-best arm index $\ell \leftarrow m+1$, and exploration arm index $u \leftarrow 1$.
3: **while** $\max_{k\in\mathcal{H}} B_k > 0$ **do**
4: $\quad u \leftarrow \arg\max_{k\in\{h,\ell\}} 2^{-p_k}$, break ties arbitrarily
5: $\quad$ Query arm $u$ for $C_1 2^{-p_u}\log(2^{p_u}K/\delta)$ times and use $\mathrm{QE}(\mathcal{O}, 2^{-p_u}, \delta/2^{p_u}K)$ to update $\hat{\mu}_u$
6: $\quad \mathcal{H} \leftarrow \arg\min_{k\in\mathcal{K}}^{1\dots m} B_k$.
7: $\quad h \leftarrow \arg\min_{k\in\mathcal{H}}(\hat{\mu}_k - 2^{-p_k})$ and $\ell \leftarrow \arg\max_{k\notin\mathcal{H}}(\hat{\mu}_k + 2^{-p_k})$
8: $\quad p_u \leftarrow p_u + 1$
9: **Output:** the arm set $\mathcal{H}$

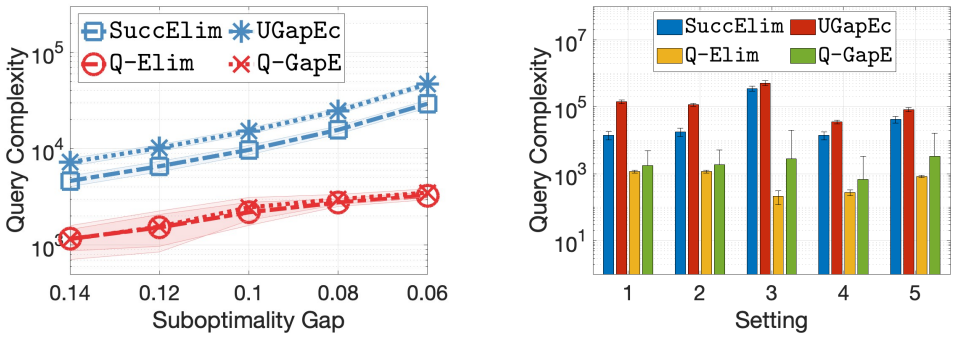
Within each iteration of the while-loop, we let arm $h$ with the smallest lower confidence bound among arms in $\mathcal{H}$ denote the estimated $m^{\text{th}}$-best arm, and arm $\ell$ with the largest upper confidence bound among arms in $\mathcal{K}\setminus\mathcal{H}$ denote the estimated $(m+1)^{\text{th}}$-best arm (Line 7). Since both arms are critical for separating the top $m$ arms from the rest, we pick the arm with largest confidence interval width among $h$ and $\ell$ (Line 4) to conduct a batched exploration (Line 5). We present `Q-GapE` in Algorithm 2 and its query complexity upper bound in Theorem 4 with a proof in Appendix B.3.

**Theorem 4** (Query complexity upper bound of Algorithm 2)**.** *Given a confidence parameter $\delta \in (0,1)$, the query complexity of* `Q-GapE` *is upper bounded as follows,*

$$Q \leqslant \sum_{k\in\mathcal{K}} \log_2\left(\frac{4}{\Delta_k^{\langle m\rangle}}\right) \frac{16C_1}{\Delta_k^{\langle m\rangle}} \cdot \ln\frac{2K}{\delta}. \tag{7}$$

*where $\Delta_k^{\langle m\rangle}$ is defined in Eq.* (5)*.*

*Remark* 6 (Optimality of this query complexity upper bound). Similar to Remark 4, compared to the query complexity lower bound for `TMI` in Theorem 2, the upper bound in Theorem 4 is tight up to some logarithmic factor.




(a) Five different suboptimality gaps

(b) Five different scenarios

Figure 1: Comparing `Q-Elim` and `Q-GapE` with `SuccElim` and `UGapEc`

*Remark* 7 (Comparison to upper bounds of classical oracles). Compared to the classical oracle query upper bound for `TMI` $O(\sum_{k\in\mathcal{K}}(1/(\Delta_k^{\langle m\rangle})^2)\log(1/\delta))$ (Gabillon, Ghavamzadeh, and Lazaric 2012), the query complexity bound of Algorithm 2 in Eq. (7) improves the coefficient before $\log(1/\delta)$ from $1/(\Delta_k^{\langle m\rangle})^2$ to $1/(\Delta_k^{\langle m\rangle})$ (ignoring the logarithmic factor). Additionally, whether one can design a `TMI` algorithm with the strong quantum oracle in Eq. (2) that enjoys a $\tilde{O}(\sqrt{\sum_{k\in\mathcal{K}}1/(\Delta_k^{\langle m\rangle})^2}\log(1/\delta))$ query complexity is an open problem. We conjecture that its design is possible.

*Remark* 8 (Comparison between `Q-GapE` and `Q-Elim`). Letting $m = 1$, `Q-GapE` becomes a `BAI` algorithm as `Q-Elim`, and both algorithms have the same query complexity upper bound (notice that $\Delta_k^{\langle 1\rangle} = \Delta_k$). Numerical simulations in Figure 1 show that both algorithms also have similar empirical performance.

## 5 Experiments

We report experimental results from comparing our quantum algorithms, `Q-Elim` (Algorithm 1) and `Q-GapE` (Algorithm 2 with $m = 1$), to classical successive elimination algorithm (Even-Dar et al. 2006, Algorithm 3), `SuccElim`, and classical unified gap-based exploration with fixed confidence algorithm (Gabillon, Ghavamzadeh, and Lazaric 2012, Section 3), `UGapEc`, in `BAI` objective with different suboptimality gaps $\Delta$.

We experiment with confidence $\delta = 0.1$ and $K = 8$ arms with reward means $0.99 - i \times \Delta$, $i \in \{0, ..., K-1\}$, where suboptimality gap $\Delta$ is *reduced* from $0.14$ to $0.06$ with step size $0.02$ to study the impact of the suboptimality gap on query complexity. For classical `BAI` algorithms, we set $c = 4$ for `SuccElim`, $b = 1, c = 0.5$ for `UGapEc` according to the default in (Even-Dar et al. 2006; Gabillon, Ghavamzadeh, and Lazaric 2012). For our quantum algorithms, we set $C_1 = 10$ and simulate the output of the quantum estimator in Lemma 3 via the estimator's analytical output distribution (Brassard et al. 2002, Theorem 11). We report values averaged over $50$ independent trials as markers and their standard deviations as shaded regions.

Figure 1a shows that (1) our quantum algorithms outperform the classical ones; (2) as the suboptimality gap decreases (along the right direction of $x$-axis), query complexities of classical algorithms grow faster than that of quantum algorithms, which corroborates the quantum improvement of the dependence on $\Delta$ from $\Delta^{-2}$ to $\Delta^{-1}$; (3) both quantum algorithms have similar empirical performance.

We further evaluate our `Q-Elim` and `Q-GapE` in some settings considered in prior best arm identification work (Audibert, Bubeck, and Munos 2010):

- Setting 1: one group of bad arms, $K = 20$, $\mu_1 = 0.5$, $\mu_i = 0.4, \forall i \in \{2, ..., 20\}$
- Setting 2: two groups of bad arms, $K = 20$, $\mu_1 = 0.5$, $\mu_i = 0.42, \forall i \in \{2, ..., 6\}$, $\mu_i = 0.38, \forall i \in \{7, ..., 20\}$
- Setting 3: geometric progression, $K = 4$, $\mu_1 = 0.5$, $\mu_i = 0.5 - (0.37)^i, \forall i \in \{2, 3, 4\}$
- Setting 4: three groups of bad arms, $K = 6$, $\mu_1 = 0.5$, $\mu_2 = 0.42$, $\mu_3 = \mu_4 = 0.4$, $\mu_5 = \mu_6 = 0.35$
- Setting 5: arithmetic progression, $K = 15$, $\mu_1 = 0.5$, $\mu_i = 0.5 - 0.025i, \forall i \in \{2, ..., 15\}$

We report values averaged over $50$ independent trials and their standard deviations in Figure 1b.

## 6 Future Directions

In the `BAI` literature, `Q-Elim` and `Q-GapE` use two different mechanisms, known as *"uniform exploration and elimination"* and *"adaptive sampling"* respectively. Previous research (Kaufmann and Kalyanakrishnan 2013) suggests that adaptive sampling outperforms uniform exploration and elimination in classical `BAI`, but our quantum algorithms, despite employing different mechanisms, show similar theoretical and empirical performance. Thus, further research is needed to compare quantum algorithms based on both mechanisms to determine if adaptive sampling maintains its superiority in the quantum setting.

In addition to the best arm identification with fixed confidence setting studied in this paper, the best arm identification with fixed budget setting remains significantly less explored and understood, even within the existing literature on classical `MAB` (Kaufmann, Cappé, and Garivier 2016; Barrier,

Garivier, and Stoltz 2023). Another compelling future direction is to investigate this fixed budget setting utilizing the quantum oracles.

## Acknowledgments

The work of Mohammad Hajiesmaili is supported by NSF CAREER-2045641, CCF-2325956, CNS-2102963, CNS-2106299, and CPS-2136199. The work of John C.S. Lui is supported in part by SRFS2122-4S02. The work of Don Towsley is supported in part by the NSF- ERC Center for Quantum Networks grant EEC-1941583.

## References

Ambainis, A. 2000. Quantum lower bounds by quantum arguments. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 636–643.

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

Audibert, J.-Y.; Bubeck, S.; and Munos, R. 2010. Best arm identification in multi-armed bandits. In *COLT*, 41–53. Citeseer.

Azuma, K.; Economou, S. E.; Elkouss, D.; Hilaire, P.; Jiang, L.; Lo, H.-K.; and Tzitrin, I. 2022. Quantum repeaters: From quantum networks to the quantum internet. *arXiv preprint arXiv:2212.10820*.

Barrier, A.; Garivier, A.; and Stoltz, G. 2023. On Best-Arm Identification with a Fixed Budget in Non-Parametric Multi-Armed Bandits. In *International Conference on Algorithmic Learning Theory*, 136–181. PMLR.

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

Brahmachari, S.; Lumbreras, J.; and Tomamichel, M. 2023. Quantum contextual bandits and recommender systems for quantum data. *arXiv preprint arXiv:2301.13524*.

Brassard, G.; Hoyer, P.; Mosca, M.; and Tapp, A. 2002. Quantum amplitude amplification and estimation. *Contemporary Mathematics*, 305: 53–74.

Bubeck, S.; Munos, R.; and Stoltz, G. 2011. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19): 1832–1852.

Buchholz, S.; Kübler, J. M.; and Schölkopf, B. 2023. Multi armed bandits and quantum channel oracles. *arXiv preprint arXiv:2301.08544*.

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

Childs, A. M. 2017. Lecture notes on quantum algorithms. *Lecture notes at University of Maryland*.

Cho, B.; Xiao, Y.; Hui, P.; and Dong, D. 2022. Quantum bandit with amplitude amplification exploration in an adversarial environment. *arXiv preprint arXiv:2208.07144*.

Chow, J.; Dial, O.; and Gambetta, J. 2021. IBM Quantum breaks the 100-qubit processor barrier. *IBM Research Blog*.

Even-Dar, E.; Mannor, S.; and Mansour, Y. 2002. PAC bounds for multi-armed bandit and Markov decision processes. In *COLT*, volume 2, 255–270. Springer.

Even-Dar, E.; Mannor, S.; Mansour, Y.; and Mahadevan, S. 2006. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of machine learning research*, 7(6).

Gabillon, V.; Ghavamzadeh, M.; and Lazaric, A. 2012. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25.

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

Holevo, A. S. 2003. *Statistical structure of quantum theory*, volume 67. Springer Science & Business Media.

Kargin, V. 2005. ON THE CHERNOFF BOUND FOR EFFICIENCY OF QUANTUM HYPOTHESIS TESTING. *The Annals of Statistics*, 33(2): 959–976.

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

Kaufmann, E.; Cappé, O.; and Garivier, A. 2016. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17: 1–42.

Kaufmann, E.; and Kalyanakrishnan, S. 2013. Information complexity in bandit subset selection. In *Conference on Learning Theory*, 228–251. PMLR.

Lai, T. L.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Lumbreras, J.; Haapasalo, E.; and Tomamichel, M. 2022. Multi-armed quantum bandits: Exploration versus exploitation when learning properties of quantum states. *Quantum*, 6: 749.

Mannor, S.; and Tsitsiklis, J. N. 2004. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun): 623–648.

Montanaro, A. 2015. Quantum speedup of Monte Carlo methods. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2181): 20150301.

Naruse, M.; Chauvet, N.; Uchida, A.; Drezet, A.; Bachelier, G.; Huant, S.; and Hori, H. 2019. Decision making photonics: solving bandit problems using photons. *IEEE Journal of Selected Topics in Quantum Electronics*, 26(1): 1–10.

Ohno, H. 2023. Quantum greedy algorithms for multi-armed bandits. *Quantum Information Processing*, 22(2): 101.

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

Slivkins, A.; et al. 2019. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2): 1–286.

Wan, Z.; Zhang, Z.; Li, T.; Zhang, J.; and Sun, X. 2023. Quantum Multi-Armed Bandits and Stochastic Linear Bandits Enjoy Logarithmic Regrets. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

Wehner, S.; Elkouss, D.; and Hanson, R. 2018. Quantum internet: A vision for the road ahead. *Science*, 362(6412): eaam9288.

Wu, Y.; Guan, C.; Aggarwal, V.; and Wang, D. 2023. Quantum Heavy-tailed Bandits. *arXiv preprint arXiv:2301.09680*.

# Supplementary Material

## A Lower bound Proof via quantum adversary method

In addition to Lemma 2 based on quantum hypothesis testing, we provide an alternative lower bound based on the quantum adversary method (Ambainis 2000).

**Theorem 5.** *Given $\mu_0, \mu_1 \in (\mu, 1-\mu)$, the necessary number of queries to distinguish the quantum* MAB *instances $\mathcal{I}_0$ and $\mathcal{I}_1$, with a probability of at least $1-\delta$, has the following lower bound,*

$$Q \geqslant \frac{1}{|\mu_0 - \mu_1|} \cdot \frac{1 - 2\sqrt{\delta(1-\delta)}}{1 + 2\sqrt{1/\mu(1-\mu)}}.$$

Replacing Lemma 2 by Theorem 5 in the proofs of Theorems 1 and 2, one can obtain another two query complexity lower bounds for BAI and TMI as follows,

$$\texttt{BAI}: \quad Q \geqslant \sum_{k\in\mathcal{K}} \frac{1}{\Delta_k} \cdot \frac{1 - 2\sqrt{\delta(1-\delta)}}{1 + 2\sqrt{1/\mu(1-\mu)}},$$

$$\texttt{TMI}: \quad Q \geqslant \sum_{k\in\mathcal{K}} \frac{1}{\Delta_k^{\langle m \rangle}} \cdot \frac{1 - 2\sqrt{\delta(1-\delta)}}{1 + 2\sqrt{1/\mu(1-\mu)}}.$$

*Proof of Theorem 5.* Without loss of generality we assume $\mu_1 > \mu_0$ and denote $\Delta = \mu_1 - \mu_0$. We denote $\left|\psi_a^{(t)}\right\rangle$ as the output after querying the oracle $\mathcal{O}_a$ for $t$ times for $a = 0, 1$. In the adversary method, we consider a weight function as follows,

$$s_t = \frac{1}{\Delta} \left\langle \psi_0^{(t)} \middle| \psi_1^{(t)} \right\rangle.$$

Note that $s_0 = \frac{1}{\Delta}$ and, to distinguish both oracles' output after $T$ queries, we require that $s_T \leqslant \frac{1}{\Delta}\sqrt{2\delta(1-\delta)}$.

After the $t^{\text{th}}$ query, we have

$$\left|\psi_0^{(t)}\right\rangle = \alpha_{0,0}\left|0\right\rangle + \alpha_{0,1}\left|1\right\rangle, \quad \left|\psi_1^{(t)}\right\rangle = \alpha_{1,0}\left|0\right\rangle + \alpha_{1,1}\left|1\right\rangle.$$

Denote the action of the quantum oracles by the following two unitary matrices,

$$\boldsymbol{A}_0 = \begin{bmatrix} \sqrt{1-\mu_0} & -\sqrt{\mu_0} \\ \sqrt{\mu_0} & \sqrt{1-\mu_0} \end{bmatrix},$$

$$\boldsymbol{A}_1 = \begin{bmatrix} \sqrt{1-\mu_1} & -\sqrt{\mu_1} \\ \sqrt{\mu_1} & \sqrt{1-\mu_1} \end{bmatrix}.$$

Then, we have

$$\begin{aligned} &\left\langle \psi_0^{(t+1)} \middle| \psi_1^{(t+1)} \right\rangle - \left\langle \psi_0^{(t)} \middle| \psi_1^{(t)} \right\rangle \\ =& \left\langle \psi_0^{(t)} \middle| \boldsymbol{A}_0^\dagger \boldsymbol{A}_1 \middle| \psi_1^{(t)} \right\rangle - \left\langle \psi_0^{(t)} \middle| \psi_1^{(t)} \right\rangle \\ =& \left\langle \psi_0^{(t)} \middle| \boldsymbol{A}_0^\dagger \boldsymbol{A}_1 - \boldsymbol{I} \middle| \psi_1^{(t)} \right\rangle. \end{aligned}$$

Denote

$$\begin{aligned} &\begin{bmatrix} u & v \\ -v & u \end{bmatrix} \coloneqq \boldsymbol{A}_0^\dagger \boldsymbol{A}_1 - \boldsymbol{I} \\ =& \begin{bmatrix} \sqrt{\mu_0\mu_1} + \sqrt{(1-\mu_0)(1-\mu_1)} - 1 & \sqrt{(1-\mu_0)\mu_1} - \sqrt{\mu_0(1-\mu_1)} \\ \sqrt{\mu_0(1-\mu_1)} - \sqrt{(1-\mu_0)\mu_1} & \sqrt{\mu_0\mu_1} + \sqrt{(1-\mu_0)(1-\mu_1)} - 1 \end{bmatrix}. \end{aligned}$$

We have

$$\begin{aligned} &|s_{t+1} - s_t| \\ \leqslant& \frac{|u|}{\Delta^2}|\alpha_{0,0}\alpha_{1,0} + \alpha_{0,1}\alpha_{1,0}| + \frac{|v|}{\Delta^2}|\alpha_{0,0}\alpha_{1,1} - \alpha_{0,1}\alpha_{1,0}| \\ \overset{(a)}{\leqslant}& \frac{|u| + |v|}{\Delta^2} \\ \overset{(b)}{\leqslant}& \frac{1 + 2\sqrt{1/\mu(1-\mu)}}{\Delta}, \end{aligned}$$

where the inequality (a) is due to the Cauchy-Schwartz inequality, and the inequality (b) is due to that

$$\begin{aligned} |u| &= \left|1 - \sqrt{\mu_0\mu_1} - \sqrt{(1-\mu_0)(1-\mu_1)}\right| \\ &\leqslant |1 - \mu_0 - (1-\mu_1)| = \Delta, \\ |v| &= \frac{\mu_1 - \mu_0}{\sqrt{(1-\mu_0)\mu_1} + \sqrt{\mu_0(1-\mu_1)}} \leqslant \frac{\Delta}{2\sqrt{\mu(1-\mu)}}. \end{aligned}$$

At last, we have

$$\begin{aligned} \frac{1}{\Delta^2}\left(1 - 2\sqrt{\delta(1-\delta)}\right) &\leqslant |s_T - s_0| \\ &\leqslant T \cdot \frac{1 + 2\sqrt{1/\mu(1-\mu)}}{\Delta}. \end{aligned}$$

That is,

$$T \geqslant \frac{1}{\Delta} \cdot \frac{1 - 2\sqrt{\delta(1-\delta)}}{1 + 2\sqrt{1/\mu(1-\mu)}}.$$

□

## B Deferred Proofs

### B.1 Algebraic Details of Proof of Lemma 2

Then, we prove that $\log |\left\langle \psi_0 | \psi_1 \right\rangle|^{-1} \leqslant \frac{(\theta_0 - \theta_1)^2}{2}$ as follows,

$$\begin{aligned} &\log |\left\langle \psi_0 | \psi_1 \right\rangle| \\ =& \log(\cos(\theta_0 - \theta_1)) \\ \overset{(a)}{\geqslant}& \log\left(1 - \frac{(\theta_0 - \theta_1)^2}{2}\right) \overset{(b)}{\geqslant} -\frac{(\theta_0 - \theta_1)^2}{2} \end{aligned} \tag{8}$$

where inequality (a) is due to $\cos x \geqslant 1 - \frac{x^2}{2}$, and inequality (b) is due to to $\log(1-x) \geqslant -x$ for $x \in (0, 0.85)$.

Next, we upper bound $|\mu_0 - \mu_1|$ with an expression of $\theta_0$ and $\theta_1$. With trigonometric identities, we have

$$\begin{aligned} &\mu_0 - \mu_1 \\ =& \sin^2\theta_0 - \sin^2\theta_1 \\ =& \sin^2((\theta_0 - \theta_1) + \theta_1) - \sin^2\theta_1 \\ =& \sin^2(\theta_0 - \theta_1)\cos^2\theta_1 + \cos^2(\theta_0 - \theta_1)\sin^2\theta_1 \\ &+ 2\sin(\theta_0 - \theta_1)\cos(\theta_0 - \theta_1)\sin\theta_1\cos\theta_1 - \sin^2\theta_1 \\ =& \sin\theta_1\cos\theta_1\sin 2(\theta_0 - \theta_1) + (1 - 2\sin^2\theta_1)\sin^2(\theta_0 - \theta_1). \end{aligned}$$

Taking the absolute values of both sides, we obtain

$$\begin{aligned}|\mu_0 - \mu_1| &\geqslant \left|(1-2\sin^2\theta_1)\sin^2(\theta_0-\theta_1)\right| \\ &\quad - |\sin\theta_1\cos\theta_1\sin 2(\theta_0-\theta_1)| \\ &\geqslant \left|(1-2\sin^2\theta_1)\sin^2(\theta_0-\theta_1)\right| \\ &\overset{(a)}{\geqslant} |(1-2\mu_1)|\frac{(\theta_0-\theta_1)^2}{4} \\ &\geqslant \frac{(\theta_0-\theta_1)^2}{4}\end{aligned} \tag{9}$$

where inequality (a) is due to $\sin x \geqslant \frac{x}{2}$ for $x \in (0, 1.8)$.

Lastly, we conclude the proof as follows,

$$\begin{aligned}Q &\overset{(a)}{\geqslant} \frac{1}{-2\log|\langle\psi_0|\psi_1\rangle|}\log\frac{1}{4\delta(1-\delta)} \\ &\overset{(b)}{\geqslant} \frac{1}{(\theta_0-\theta_1)^2}\log\frac{1}{4\delta(1-\delta)} \\ &\overset{(c)}{\geqslant} \frac{1}{4|\mu_0-\mu_1|}\log\frac{1}{4\delta(1-\delta)} \\ &\geqslant \frac{1}{4|\mu_0-\mu_1|}\log\frac{1}{4\delta},\end{aligned}$$

where inequalities (a), (b), and (c) are due to Eq. (6), Eq. (8), and Eq. (9) respectively.

### B.2 Proof of Theorem 2

For any top arm $k \leqslant m$, i.e., among top $m$ arms set, in instance $\mathcal{I}_0$, we consider another instance $\mathcal{I}_k = \{\mu_1^{(k)}, \ldots, \mu_K^{(k)}\}$ whose reward means are the same as instance $\mathcal{I}_0$ except for the reward mean of arm $k$ which assumes the form $\mu_k^{(k)} = \mu_{m+1} - \epsilon$ for $0 < \epsilon < \mu_{m+1}$. Therefore, in instance $\mathcal{I}_k$, the top $m$ arms set is $\{1, 2, \ldots, m, m+1\} \setminus \{k\}$.

Because the instances $\mathcal{I}_0$ and $\mathcal{I}_k$ have different top $m$ arms, any feasible policy must be able to distinguish these two instances with a confidence of at least $1-\delta$. Given the additional information that all other arms have the same means, this task reduces to distinguishing two instances $\mathcal{I}_0$ and $\mathcal{I}_k$ as in Lemma 2. Therefore, the query complexity of distinguishing both instances is at least $1/4(\Delta_k^{\langle m\rangle}+\epsilon)\log 1/4\delta$. Note that these queries are all on arm $k$.

For any other suboptimal arm $k > m$, i.e., not among top $m$ arms set, in instance $\mathcal{I}_0$, we consider another instance $\mathcal{I}_k$ whose oracles are the same as instance $\mathcal{I}_0$ except that the reward mean of the arm $k$ in $\mathcal{I}_k$ is $\mu_k^{(k)} = \mu_m + \epsilon$ for $0 < \epsilon < 1/2 - \mu_m$. Therefore, in instance $\mathcal{I}_k$, the top $m$ arms set is $\{1, 2, \ldots, m-1\} \cup \{k\}$. Similarly, applying Lemma 2 to distinguishing instances $\mathcal{I}_0$ and $\mathcal{I}_k$, the query complexity is lower bounded by $1/4(\Delta_k^{(m)}+\epsilon)\log 1/4\delta$.

Last, summing the least query complexity spent on each arm and letting $\epsilon \to 0$ yields

$$Q \geqslant \sum_{k\in\mathcal{K}} \frac{1}{4\Delta_k^{\langle m\rangle}}\log\frac{1}{4\delta}.$$

### B.3 Proof of Theorem 4

**Correctness:** If all estimates of QE are correct, i.e., $\mu_k \in (\hat{\mu}_k - 2^{-p_k}, \hat{\mu}_k + 2^{-p_k})$ holds for all arms, then the final output arm set $\mathcal{H}$ must be the true top $m$ arms. We show that the probability that any of these QEs fails, i.e., $\mu_k \notin (\hat{\mu}_k - 2^{-p_k}, \hat{\mu}_k + 2^{-p_k})$, is upper bounded by $\delta$.

Fix an arm $k$. Its $p^{\text{th}}$ reward mean estimate QE fails with a probability of at most $\frac{2^{-p_k}\delta}{K} = \frac{\delta}{2^p K}$. Consequently, the probability of any of this arm's reward mean estimates failing are upper bounded as follows,

$$\sum_{p=1}^{s_k}\frac{\delta}{2^p K} < \sum_{p=1}^{\infty}\frac{1}{2^p}\frac{\delta}{K} = \frac{\delta}{K},$$

where $s_k$ is the total number of times that arm $k$'s reward mean is estimated in Algorithm 2. Applying the union bound to all arms shows that the total failure probability of this algorithm is upper bounded by $\delta$.

**Query complexity:** Since the above correctness analysis rules out all failures of confidence interval constructions (with probability at most $\delta$), we assume in this part of the proof that $\mu_k \in (\hat{\mu}_k - 2^{-p_k}, \hat{\mu}_k + 2^{-p_k})$ holds for all arms' reward mean estimates in the algorithm.

**Lemma 4** (Adapted from (Gabillon, Ghavamzadeh, and Lazaric 2012, Lemma 2))**.** *When arm $k \in \{h, \ell\}$ is queried, we have*

$$\max_{k\in\mathcal{H}} B_k \leqslant \min\{0, -\Delta_k^{\langle m\rangle} + 2\cdot 2^{-p_k}\} + 2\cdot 2^{-p_k}.$$

For any queried arm $k$, Lemma 4 and the while-loop condition yields

$$\begin{aligned}0 \leqslant \max_{k\in\mathcal{H}} B_k &\leqslant \min\{0, -\Delta_k^{\langle m\rangle} + 2\cdot 2^{-p_k}\} + 2\cdot 2^{-p_k} \\ &\implies \Delta_k^{\langle m\rangle} \leqslant 4\cdot 2^{-p_k}.\end{aligned}$$

Denote $s_k$ as the last round in which the arm $k$ is queried in Algorithm 2. Then, we have $\Delta_k^{\langle m\rangle} \leqslant 4\cdot 2^{-s_k}$. Hence, the number of times arm $k$ is queried in the algorithm is upper bounded as follows,

$$\begin{aligned}&\sum_{p=1}^{s_k}\frac{C_1}{2^{-p}}\ln\frac{K}{2^{-p}\delta} \leqslant C_1\ln\frac{2K}{\delta}\cdot\sum_{p=1}^{s_k}(p+1)2^p \\ &\leqslant C_1\ln\frac{2K}{\delta}\cdot(s_k+1)2^{s_k+1} \leqslant C_1\ln\frac{2K}{\delta}\cdot\frac{16}{\Delta_k^{\langle m\rangle}}\log_2\frac{4}{\Delta_k^{\langle m\rangle}}.\end{aligned}$$

Lastly, summing over all arms yields the following upper bound to the total number of queries performed in Algorithm 2:

$$\begin{aligned}\mathbb{E}[Q] &\leqslant C_1\ln\frac{K}{\delta}\cdot\sum_{k\in\mathcal{K}}\frac{16}{\Delta_k^{\langle m\rangle}}\log_2\frac{4}{\Delta_k^{\langle m\rangle}} \\ &= \ln\frac{K}{\delta}\sum_{k\in\mathcal{K}}\frac{16C_1}{\Delta_k^{\langle m\rangle}}\log_2\frac{4}{\Delta_k^{\langle m\rangle}}.\end{aligned}$$

## C Experimental Setup Detail

Following the convention of prior quantum bandit works (Wan et al. 2023; Wu et al. 2023), we implement the

quantum estimate algorithm in Lemma 2 according to (Brassard et al. 2002, Theorem 11). Specifically, $\mathrm{QE}(\mathcal{O}, \epsilon, \delta)$ outputs $\hat{\mu}$, where $\hat{\mu} = \text{median}(\hat{\mu}'_1, ..., \hat{\mu}'_i, ..., \hat{\mu}'_\delta)$, and

$$\begin{cases} \mathbb{P}\left(\frac{\arcsin\left(\sqrt{\hat{\mu}'_i}\right)}{\epsilon\pi} = x\right) = \frac{\sin\left(\gamma\pi/\epsilon\right)^2}{\sin\left(\gamma\pi\right)^2/\epsilon^2}, \text{ if } \sin\left(\gamma\pi\right)^2 \neq 0 \\ \mathbb{P}\left(\frac{\arcsin\left(\sqrt{\hat{\mu}'_i}\right)}{\epsilon\pi} = x\right) = \mu, \text{ if } \sin\left(\gamma\pi\right)^2 = 0 \end{cases},$$
$$\forall x \in \left[\frac{1}{\epsilon}\right], i \in [\delta],$$

where $\gamma = \min\{(\arcsin\left(\sqrt{\mu}\right)/\pi - (x-1)\epsilon)\%1, 1 - (\arcsin\left(\sqrt{\mu}\right)/\pi - (x-1)\epsilon)\%1)\}$. Our experiments are executed on a computer equipped with Xeon E5-2680 CPU and 128GB RAM.