HAL
open science

Quantum bandits

Balthazar Casalé, Giuseppe Di Molfetta, Hachem Kadri, Liva Ralaivola

To cite this version:

Balthazar Casalé, Giuseppe Di Molfetta, Hachem Kadri, Liva Ralaivola. Quantum bandits. Quantum Machine Intelligence, 2020, 2 (1), 10.1007/s42484-020-00024-8 . hal-03118185

HAL Id: hal-03118185

https://hal.science/hal-03118185

Submitted on 21 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quantum Bandits

Balthazar Casalé BALTHAZAR.CASALE@LIS-LAB.FR
Aix-Marseille Université, CNRS, LIS, Marseille, France

Giuseppe Di Molfetta GIUSEPPE.DIMOLFETTA@LIS-LAB.FR
Aix-Marseille Université, CNRS, LIS, Marseille, France and Quantum Computing Center, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama, 223-8522 Japan

Hachem Kadri HACHEM.KADRI@LIS-LAB.FR
Aix-Marseille Université, CNRS, LIS, Marseille, France

Liva Ralaivola L.RALAIVOLA@CRITEO.COM
Criteo AI Lab, Criteo, Paris

Abstract

We consider the quantum version of the bandit problem known as best arm identification (BAI). We first propose a quantum modeling of the BAI problem, which assumes that both the learning agent and the environment are quantum; we then propose an algorithm based on quantum amplitude amplification to solve BAI. We formally analyze the behavior of the algorithm on all instances of the problem and we show, in particular, that it is able to get the optimal solution quadratically faster than what is known to hold in the classical case.

Keywords: Bandits, Best Arm Identification, Quantum Amplitude Amplification

1. Introduction

Many decision-making problems involve learning by interacting with the environment and observing what rewards result from these interactions. In the field of machine learning, this line of research falls into what is referred as reinforcement learning (RL), and algorithms to train artificial agents that interact with an environment have been studied extensively (Sutton and Barto, 2018; Kaelbling et al., 1996; Bertsekas and Tsitsiklis, 1996). We are here interested in the best arm identification (BAI) problem from the family of bandit problems, which pertains the set of RL problems where the interactions with the environment give rise to immediate rewards and where long-term planning is unnecessary (see the survey of Lattimore and Szepesvári, 2020). More precisely, we are interested in a quantum version of the BAI problem, for which we design a quantum algorithm capable to solve it.

Quantum machine learning is a research field at the interface of quantum computing and machine learning where the goal is to use quantum computing paradigms and technologies to improve the speed and performance of learning algorithms (Wittek, 2014; Biamonte et al., 2017; Ciliberto et al., 2018; Schuld and Petruccione, 2018). A fundamental concept in quantum computing is quantum superposition, which is the means by which quantum algorithms like that of Grover (1996) —one of the most popular quantum algorithm— succeeds in solving the problem of finding one item from an unstructured database of N items in time $O(\sqrt{N})$, so beating the classical $O(N)$ time requirement. Recent works have investigated the use of Grover's quantum search algorithm to enhance machine learning and have proved its ability of providing non-trivial improvements not only

© B. Casalé, G.D. Molfetta, H. Kadri & L. Ralaivola.

in the computational complexity but also in the statistical performance of these models (Aïmeur et al., 2013; Wittek, 2014; Kapoor et al., 2016). Beyond Grover's algorithm, quantum algorithms for linear algebra, such as quantum matrix inversion and quantum singular value decomposition, were recently proposed and used in the context of machine learning (Rebentrost et al., 2014; Kerenidis and Prakash, 2017). Works on quantum reinforcement learning are emerging (Dong et al., 2008; Naruse et al., 2015; Dunjko et al., 2016; Lamata, 2017), and our paper aims at providing a new piece of knowledge in that area, by bringing two contributions: i) a formalization of the best arm identification problem in a quantum setting, and ii) a quantum algorithm to solve this problem that is quadratically faster than classical ones.

Quantum machine learning research can be classified into four categories depending on whether the data, the learner, both, or none are quantum (Aïmeur et al., 2006; Dunjko and Briegel, 2018). Our work deals with the BAI problem when both the agent and the environment are quantum systems, and so falls into the Quantum-Quantum (QQ) setting. Although less studied, the QQ approach is particularly attractive because it would allow the exploitation of the full potential of quantum technologies in machine learning. In this setting, the interaction can be fully quantum, and the agent and the environment may become entangled (Dunjko et al., 2016). Recent progress in reinforcement learning has achieved very impressive results in games (Mnih et al., 2015) and robotics (Levine et al., 2016). The training process of these models is often done in a computer simulated environment, as it would require too much agent-environment interactions to be done with a physical system in a reasonable amount of time. Performing such simulations on a quantum computer or simulator should give rise to environment's internal states that are naturally quantum. The internal state of the environment may be hidden from the agent, and considering quantum interactions between the agent and the environment would lead to more efficient learning. This motivates the setting we are interested in: quantum agents and quantum environments.

The paper is organized as follows. In Section 2, we formulate the best arm identification (BAI) problem, briefly review the upper confidence bound, and illustrate how it can be used to solve the BAI problem. In Section 3, we describe the quantum amplitude amplification, at the core of Grover's algorithm, which forms the basis of our approach. Our main results are in Section 4: we provide our quantum modeling of the BAI problem, which assumes that both the learning agent and the environment are quantum; and then we proposes an algorithm based on quantum amplitude amplification to solve BAI, that it is able to get the optimal solution quadratically faster than what is known to hold in the classical case. Section 5 concludes the paper.

2. Best Arm Identification

2.1. Stochastic Multi-Armed Bandits and the BAI Problem

Bandit problems are RL problems where it is assumed an agent evolves in an environment with which it can interact by choosing at each time step an action (or arm), each action taken providing the agent with a reward, which values the quality of the chosen action (see function f below, and more generally, Lattimore and Szepesvári, 2020).

The bandit problem we want to study from a quantum point of view is that of best arm identification from stochastic multi-armed bandits (Audibert and Bubeck, 2010). It comes with the following assumptions: the set X of actions is finite and discrete, with $|X| = N$, and when action x_t is chosen at time t then the reward r_t depends upon the independent realisation (called y_t afterwards) of a random variable distributed according to some unknown (but fixed) law ν_{x_t}. The BAI problem

Data: A number of rounds T
Result: $\tilde{x}_T$ a recommended action
for $t \leftarrow 1$ **to** T **do**
  the agent chooses the action x_t
  the environment picks an internal state y_t following ν_{x_t}
  the agent perceives the reward $r_t = f(x_t, y_t)$
end
the agent return $\tilde{x}_T$ the recommended action

Algorithm 1: The best arm identification problem

is to devise a strategy of action selection for the agent such that, after a predefined number T of interactions, the agent is able to identify the best action with the best possible guarantees.

We may go one step further in the formal statement of the problem and, in the way, use a modelling that is both in line with the classical BAI problem and suitable for its quantum extension. In particular, in order to take the unknown distributions ν_x, $x \in X$, we will explicitly introduce Y, the set of all possible internal states y_t of the environment —this notion of internal state of the environment is uncommon in the classical bandit literature. The agent's action x_t sets the internal state of the environment to y_t, which is a random draw from distribution ν_{x_t}, unknown to the agent. The agent then receives a reward $r_t = f(x_t, y_t)$, indicating the fit of action x_t with the state of the environment; we here assume that f can only take values in $\{0, 1\}$ —this corresponds to the classical case where the reward r_t is drawn according to a Bernoulli distribution of unknown parameter $\theta_{x_t} \in [0, 1]$. With these assumptions, the average reward associated with action x is

$$a_x = \sum_{y \in Y} \nu_x(y) f(x, y), \tag{1}$$

and we may define the optimal action x^* as

$$x^* = \arg\max_{x \in X} a_x, \tag{2}$$

and $a^* = a_{x^*}$ the mean reward of the optimal action. After T interactions with the environment, the agent will choose an action $\tilde{x}_T$ as its recommendation (see Algorithm 1). The quality of the agent's decision $\tilde{x}_T$ is then evaluated as the regret $a^* - a_{\tilde{x}_T}$, i.e. the difference between a^* the mean reward of optimal action a^* and $a_{\tilde{x}_T}$ the mean reward of the recommended action.

Let us elaborate further on the regret; let

$$\Delta_x = a^* - a_x \tag{3}$$

be the difference between the value of the optimal action and the value of action x. If the agent recommends the action x with probability $P_T(x)$ after T rounds, then the average difference between the value of its recommendation and the value of the optimal action is

$$R_T = \sum_{x \in X} P_T(x) \Delta_x, \tag{4}$$

which is the average regret after T iterations of the agent's strategy. Our goal is to find an action selection strategy for which the value of R_T decreases quickly as the value of T increases.

If $e_T = 1 - P_T(x^*)$ is the probability that the agent does not recommend the best action after T iterations, then, as $\forall x \in X, \Delta_x \leq 1$, the (average) regret is so that $R_T < e_T$. In the following, we recall how a tight upper bound for e_T can be derived.

Data: a number of trials T
an exploration parameter p
Result: $\tilde{x}_n$ a recommended action
let $B_{x,t} = \tilde{a}_x(t) + \sqrt{\frac{p}{t-1}}$
for $t \leftarrow 1$ **to** T **do**
the agent chooses the action $x_t \in \arg\max_{x \in X} B_{x,t}$
the environment picks an internal state y_t according to ν_{x_t}
the agent perceives the reward $r_t = f(x_t, y_t)$
the agent updates the values of $B_{x,t}$ to take r_t into account
end
the agent return $\tilde{x}_T = \arg\max_{x \in X} \tilde{a}_x(T)$
Algorithm 2: UCB-E algorithm

2.2. Upper Confidence Bound Exploration-based strategy

Part of the difficulty in the BAI problem comes from the fact that the value of each action is the mean of random variable that depends on an unknown probability distribution. The only way for an agent to estimate the value a_x of action x is to repeatedly interact with the environment to obtain a sample of rewards associated to x. Thus, a good strategy needs to find a balance between sampling the most promising actions, and sampling the actions for which we lack information. The Upper Confidence Bound Exploration (UCB-E) depicted in Algorithm 2, first described in Audibert and Bubeck (2010), is an efficient strategy to solve the best arm identification problem. It is based on a very well known and used family of UCB strategies (Lai and Robbins, 1985; Auer et al., 2002), which were proven to be optimal for solving the multi-armed bandit problem (Thompson, 1933).

Let $\Omega_x(T)$ be the set of rounds for which the agent picked action x until time T, and

$$\tilde{a}_x(T) = \frac{1}{|\Omega_x(T)|} \sum_{t \in \Omega_x(T)} r_t \tag{5}$$

be the empirical average of the reward for action x. We know from Hoeffding (1963) that a_x and $\tilde{a}_x(T)$ are tied by the relation

$$\mathbb{P}(|\tilde{a}_x(T) - a_x| > \epsilon) < 2\exp\big(-2\epsilon^2 |\Omega_x(T)|\big).$$

This means that, for all $\delta \in [0, 1]$, there is a range of value centered around $\tilde{a}_x(T)$ in which a_x lies with probability at least $1 - \delta$. The more the agent interacts with the environment with action x, the smaller this range of values is. The principle behind UCB is to choose, at each iteration, the action x for which the upper bound of this range is the highest.

Audibert and Bubeck (2010) showed that UCB-E admits the following upper bound on e_T, when the exploration parameter p is well tuned :

$$e_T < 2TN \exp\left(-\frac{T-N}{18H_1}\right), \text{ where } H_1 = \sum_{x \in X \setminus \{x^*\}} \frac{1}{\Delta_x^2}.$$

From this inequality, we can deduce a lower bound of the number of iterations to recommend the optimal arm with probability at least $1 - \delta$, for any $\delta \in (0, 1)$:

$$e_T < \delta \Rightarrow T > 18H_1 \ln\left(\frac{2N}{\delta}\right) + N.$$

The quantum modelling and accompanying algorithm proposed in this paper come with a theoretical result that quadratically improves this bounds.

3. Quantum Amplitude Amplification

If we dispose of an unstructured, discrete set X of N elements and we are interested in finding one marked element x_0, a simple probability argument shows that it takes an average of N/2 (exhaustive) queries to find the marked element. While it is well known that $O(N)$ is optimal with classical means, Grover (1996) proved that a simple quantum search algorithm speeds up any brute force $O(N)$ problem into a $O(\sqrt{N})$ problem. This algorithm comes in many variants and has been rephrased in many ways, including in terms of resonance effects (Grover, 1996) and quantum walks (Childs and Goldstone, 2004; Roget et al., 2020). The principle behind the original Grover search algorithm is the amplitude amplification (Brassard et al., 2000; Grover, 1998) in contrast with the techniques called probability amplification used in classical randomized algorithms.

In the classical case it is known that, if we know the procedure which verifies the output, then we can amplify the success probability n times, and the probability to recover the good result is approximately np where p is the probability to return the searched value. Thus in order to amplify the probability to 1 we need to multiply the runtime by a factor $1/p$. In the quantum case, the basic principle is the same and we amplify amplitudes instead of probabilities. Grover's algorithms and all its generalisations have shown that in order to achieve a maximum probability close to 1, we amplify for a number of rounds which is $O(\sqrt{1/p})$, then quadratically faster than the classical case. Before we show how to apply this result to the best arm identification problem, let us briefly recall how the amplitude-amplification algorithms works. First, we need to introduce a N-dimensional state space H, which can be supplied by $n = \log_2 N$ qubits, spanned by the orthonormal set of states $|x\rangle$, with $x \in X$. In general, we say that, after the application of an arbitrary quantum operator, the probability to find the marked element x_0 is p, where this element is a point in the domain of a generic Boolean function $f : \{0,1\}^n \to \{0,1\}$ such that $f(x_0) = 1$. This function induces a partition of $\mathcal{H}$ into two subspaces, $\mathcal{H}_1$ and $\mathcal{H}_0$, and each of them can be seen respectively as the good subspace spanned by the set of basis states for which $f(x) = 1$ and the bad subspace, which is its orthogonal. Any arbitrary state $|\Psi\rangle$ belonging to $\mathcal{H}$ can be decomposed on the basis $\{|\Psi_1\rangle, |\Psi_0\rangle\}$ as follows

$$|\Psi\rangle = \sin\theta\, |\Psi_1\rangle + \cos\theta\, |\Psi_0\rangle ,$$

where $\{|\Psi_1\rangle, |\Psi_0\rangle\}$ are the normalised projections of $|\Psi\rangle$ in the two subspaces $\mathcal{H}_1$ and $\mathcal{H}_0$:

$$|\Psi_1\rangle = \frac{1}{\sqrt{p}} \sum_{f(x)=1} \alpha_x |x\rangle , \quad |\Psi_0\rangle = \frac{1}{\sqrt{1-p}} \sum_{f(x)=0} \alpha_x |x\rangle ,$$

where α_x is a complex number and $\sin\theta = \sqrt{p}$ denotes the probability that measuring $|\Psi\rangle$ produces a marked state (for which $f(x) = 1$). In general terms, one step of the algorithm is composed by two operators: (i) the oracle, as in the original Grover results; (ii) and the generalised Grover diffusion operator. The oracle O_f is built using f and reads:

$$O_f |x\rangle = (-1)^{f(x)} |x\rangle ,$$

which essentially *marks* the searched state with minus sign. The diffusion operator is defined as:

$$R_\Psi = AS_0A^{-1} = 2\,|\Psi\rangle\langle\Psi| - \mathbb{I},$$

where $S_0 = 2\,|0\rangle\,\langle 0| - \mathbb{I}$ is the usual reflection operator around $|0\rangle$ and $|\Psi\rangle = A\,|0\rangle$. The composition of both operators leads to one evolution step of the amplitude-amplification algorithm:

$$Q = R_\Psi O_f.$$

Notice that when $A = H^{\otimes n}$, the Walsh-Hadamard transform, the above algorithm reduces to the original Grover algorithm, where the initial state is an uniform superposition of states. The repetitive application of Q after n iterations leads to:

$$Q^n\,|\Psi\rangle = \sin((2n+1)\theta)\,|\Psi_1\rangle + \cos((2n+1)\theta)\,|\Psi_0\rangle\,. \tag{6}$$

As in the Grover algorithm for $n \approx \frac{\pi}{4\theta}$ and $\theta \ll 1$, the number of call to Q needed to find the desired element is in $O(\frac{1}{\sqrt{p}})$, leading to a quadratic speedup over classical algorithms.

4. Quantum Best Arm Identification

Efficiently Solving the best arm identification problem is generally limited by the amount of information the agent needs to recover from a single interaction with the environment. This is also the case in the unstructured classical search problem, as a single call to the indication function f, the oracle, gives us information on a single element of the set. In general terms, the idea is to apply the same basic principle of the amplitude-amplification quantum algorithm to the best arm identification problem, where the reward function introduced in Section 2 now plays the role of the oracle. Indeed, in the same way that the boolean function f in a searching problem *recognises* whether x is the marked element we are looking for, the reward $r_t = f(x_t, y_t)$, indicates whether $\{x_t, y_t\}$ corresponds to a desirable outcome (in that case, $f(x_t, y_t) = 1$) or not (then $f(x_t, y_t) = 0$), where x_t is the action of the agent and y_t the state of the environment. Thus, our strategy in the following is to apply the amplitude-amplification quantum algorithm to recover the desirable outcome, i.e., the optimal action of the agent.

In order to properly apply the above quantum strategy, we define a composite Hilbert space $\mathcal{H} = \mathcal{H}_X \otimes \mathcal{H}_Y$, where $\mathcal{H}_X$ is the space of the quantum actions of the agent, spanned by the orthonormal basis $\{|x\rangle\}_{x\in X}$ and $\mathcal{H}_Y$ is the space of the quantum environment states, spanned by the orthonormal basis $\{|y\rangle\}_{y\in Y}$. All vector $|\Psi\rangle$, representing the whole composite system, decomposes on the basis $\{|xy\rangle\}_{x\in X, y\in Y}$. Notice that in the classical context, the agent's action sets the internal state of the environment to y_t, according to a random distribution ν_{x_t}, which is unknown to the agent. A straightforward way to recover the same condition, is to prepare the state of the environment in a superposition $|\psi_x\rangle = \sum_{y\in Y} \sqrt{\nu_x(y)}\,|y\rangle$, where $\nu_x(y)$ depends on the action x chosen by the agent. This is achieved preparing the initial state of the environment as follows:

$$\forall x \in X, \quad O_e\,|x0\rangle = |x\psi_x\rangle : |\langle y|\psi_x\rangle|^2 = \nu_x(y),$$

where O_e is a unitary operator acting on the composite Hilbert space $\mathcal{H}$. Moreover, the initial state of the agent is prepared in an arbitrary superposition state, applying an unitary operator A on the state space of the agent $\mathcal{H}_X$:

$$A\,|0\rangle = |\phi\rangle = \sum_{x\in X} \alpha_x\,|x\rangle.$$

Data: a unitary operator A acting on $\mathcal{H}_X$
a unitary operator O_e acting on the composite system agent-environment
n number of rounds
Result: the recommended action $\tilde{x}_n$
prepare a quantum register to the state $|00\rangle$
apply $O_e(A \otimes \mathbb{I}_e)$ to the state of the register
for $t \leftarrow 1$ **to** n **do**
  apply $G = (O_e(A \otimes \mathbb{I}))(S_0^{(X)} \otimes S_0^{(Y)})(O_e(A \otimes \mathbb{I}))^{-1} O_f$ to the state of the register
end
return $\tilde{x}_n$

Algorithm 3: Quantum Best Arm Identification (QBAI)

Once the initial state is prepared, we build the oracle O_f on the composite Hilbert space of the agent and the environment, the action of which is:

$$\forall x \in X, y \in Y, \quad O_f |xy\rangle = \begin{cases} -|xy\rangle \text{ if } f(x,y) = 1, \\ |xy\rangle \text{ otherwise.} \end{cases}$$

As for a search problem, we propose a quantum procedure that allows us to a find the optimal action (for which $r_t = f(x_t, y_t) = 1$) using $O(1/\sqrt{p})$ application of O_f, with probability approaching 1. The quantum amplitude amplification algorithm and its analysis is then reminiscent of what was presented in Section 3. One round of the algorithm is defined by the composition of the above three operators and the resulting algorithm QBAI (Quantum Best Arm Identification) is depicted in Algorithm 3. As shown in Algorithm 3, our strategy is based on applying, at each iteration, the operator G, computed from O_e, O_f and A. It is worth noting that although G does not vary as a function of time/iteration, our strategy is able to take into account the reward at each time step. This is achieved by means of the environment's internal state which can be in a quantum superposition that evolves with time according to the reward obtained after performing an action.

Defining $|\Psi\rangle = O_e(A \otimes \mathbb{I}) |00\rangle$, iterating n times the ~~above algorithm~~ operator G we recover

$$G^n |\Psi\rangle = \sin((2n+1)\theta) |\Psi_1\rangle + \cos((2n+1)\theta) |\Psi_0\rangle ,$$

which is of the same form of Equation 6, where now $\{|\Psi_1\rangle , |\Psi_0\rangle\}$ are the normalised projections of $|\Psi\rangle$ in the two subspaces $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively the good subspace spanned by the set of basis states for which $r_t = f(x,y) = 1$ and the bad subspace, which is its orthogonal. We know from Section 3, that to recover the optimal action we need to maximise the sinus. Let us choose an alternative, but equivalent, path. Let us compute the recommendation probability $P_n(x) = \sum_{y \in Y} |\langle xy| G^n |\Psi\rangle|^2$. After a straightforward computation and few simplifications, it results:

$$P_n(x) = |\langle x| A |0\rangle|^2 (1 + (a_x - p) C(p,n)),$$

where $C(p,n) = \frac{\sin((2n+1)\theta)^2 - p}{p(1-p)}$, $p = \sin(\theta)^2$ and $a_x = \sum_{y: f(x,y)=1} |\langle y|\psi_x\rangle|^2$. The recommendation probability $P_n(\tilde{x})$ for the optimal action $\tilde{x}$ is then recovered when $\sin((2n+1)\theta)^2 = 1$, i.e. when $n \approx \frac{\pi}{4}\sqrt{1/p} - \frac{1}{2}$.

Summarizing the results so far:

Theorem 1 *The probability $P_n(\tilde{x})$ that QBAI will recommend the optimal action $\tilde{x}$ is maximized when $n \approx \frac{\pi}{4}\sqrt{1/p} - \frac{1}{2}$. It follows that $P_n(\tilde{x}) = |\langle \tilde{x}| A |0\rangle|^2 \frac{a^*}{p}$.*

In order to compare this result with the classical bounds, we need to define A. For sake of simplicity, let consider A so that $\forall x \in X, |\langle x| A |0\rangle|^2 = \frac{1}{N}$, which translates in $p = \mathbb{E}_X[a_x]$. From Theorem 1, we need

$$n = \frac{\pi}{4}\sqrt{\mathbb{E}_X[a_x]^{-1}} - \frac{1}{2}$$

rounds to recommend the optimal action with probability $1 - (1 - \frac{a^*}{N\mathbb{E}_X[a_x]})$. Let us recall that UCB-E needs at least $18H_1 \ln(\frac{2N}{\delta}) + N$ rounds to recommend the optimal action with the same probability. The ratio between ~~both probabilities~~ both number of rounds is of order $O(\sqrt{\mathbb{E}_X[a_x]}H_1 \ln\left(\frac{2N^2\mathbb{E}_X[a_x]}{N\mathbb{E}_X[a_x]-a^*}\right) + \sqrt{\mathbb{E}_X[a_x]}N)$. In the case $\mathbb{E}_X[a_x] > \frac{1}{N}$, then $\sqrt{\mathbb{E}_X[a_x]}N > \sqrt{N}$ and the complexity gain for the quantum algorithm results quadratic in respect of the number of actions. Otherwise, since $H_1 > (N-1)a^{*-2}$, we get that $\sqrt{\mathbb{E}_X[a_x]}H_1 > a^{*-\frac{3}{2}}\sqrt{N}$, and the speedup is once again quadratic in respect of the number of actions. This result is sufficient to prove that QBAI is quadratically faster than a classical algorithm to recommend the optimal arm with probability at least $\frac{a^*}{N\mathbb{E}_X[a_x]} = (N - \sum_{x \neq x^*} \Delta_x/a^*)^{-1}$.

We know from Theorem 1 that QBAI cannot identify the best arm with better probability without modifying the operator A during the learning process. As such, QBAI does not allow one to identify the best action with arbitrary small margin of error, as can be done in the classical approach. However, because it is able to attain the same level of confidence in fewer interactions with the environment than classical strategies, it is reasonable to think that an algorithm based on QBAI could identify the best action with arbitrarily small margin of error while keeping a quantum advantage. Devising such an algorithm is out of the scope of this paper and we leave this possibility for future research.

5. Conclusion

We studied the problem of Best Arm Identification (BAI) in a quantum setting. We proposed a quantum modeling of this problem when both the learning agent and the environment are quantum. We introduced a quantum bandit algorithm based on quantum amplitude amplification to solve the quantum BAI problem and showed that is able to get the optimal solution quadratically faster than what is known to hold in the classical case. Our results confirm that quantum algorithms can have a significant impact on reinforcement learning and open up new opportunities for more efficient bandit algorithms.

Our aim with this paper has been to provide a direct application of amplitude amplification to the best arm identification problem, and to show that it exhibits the same behavior it did in other problems of the same nature in term of efficiency. It has been proposed a direct quantum analogue of the multi-armed bandit problem, and an analytical proof that amplitude amplification can find the best action quadratically faster than the best known classical algorithm with respect to the number of actions. Future extensions of this work might include the following topics: (i) Could this algorithm be adapted to recommend the optimal action with arbitrarily small margin of error? (ii) Can it be possible to treat the case where the reward function have value in $\mathbb{N}$? (iii) Can this algorithm be adapted to solve more complex decision making problems? (iv) Can it be proven or disproven that amplitude amplification is optimal for this problem, as it is for other unstructured search problems?

Acknowledgements

This work has been funded by the French National Research Agency (ANR) project QuantML (grant number ANR-19-CE23-0011), the Pépinière d'Excellence 2018, AMIDEX fondation, project DiTiQuS, and the ID #60609 grant from the John Templeton Foundation, as part of the "The Quantum Information Structure of Spacetime (QISS)" Project.

References

Esma Aïmeur, Gilles Brassard, and Sébastien Gambs. Machine learning in a quantum world. In Conference of the Canadian Society for Computational Studies of Intelligence, pages 431–442, 2006.

Esma Aïmeur, Gilles Brassard, and Sébastien Gambs. Quantum speed-up for unsupervised learning. Machine Learning, 90(2):261–287, 2013.

Jean-Yves Audibert and Sébastien Bubeck. Best Arm Identification in Multi-Armed Bandits. In COLT - 23th Conference on Learning Theory - 2010, page 13 p., Haifa, Israel, June 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2):235–256, May 2002.

Dimitri P. Bertsekas and John N. Tsitsiklis. Neuro-dynamic programming. Athena Scientific, Belmont, MA, 1996.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. Nature, 549(7671):195–202, 2017.

Gilles Brassard, Peter Hoyer, Michele Mosca, and Alain Tapp. Quantum Amplitude Amplification and Estimation. arXiv e-prints, art. quant-ph/0005055, May 2000.

Andrew M. Childs and Jeffrey Goldstone. Spatial search by quantum walk. Physical Review A, 70 (2):022314, 2004.

Carlo Ciliberto, Mark Herbster, Alessandro Davide Ialongo, Massimiliano Pontil, Andrea Rocchetto, Simone Severini, and Leonard Wossnig. Quantum machine learning: a classical perspective. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 474 (2209):20170551, 2018.

Daoyi Dong, Chunlin Chen, Hanxiong Li, and Tzyh-Jong Tarn. Quantum reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38(5):1207–1220, 2008.

Vedran Dunjko and Hans J. Briegel. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. Reports on Progress in Physics, 81(7):074001, 2018.

Vedran Dunjko, Jacob M. Taylor, and Hans J. Briegel. Quantum-enhanced machine learning. Physical review letters, 117(13):130501, 2016.

Lov 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.

Lov K. Grover. Quantum computers can search rapidly by using almost any transformation. Physical Review Letters, 80(19):4329, 1998.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58(301):13–30, 1963.

Leslie Pack Kaelbling, Michael L. Littman, and Andrew P. Moore. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4:237–285, 1996.

Ashish Kapoor, Nathan Wiebe, and Krysta Svore. Quantum perceptron models. In Advances in Neural Information Processing Systems, pages 3999–4007, 2016.

Iordanis Kerenidis and Anupam Prakash. Quantum recommendation systems. 2017.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1):4 – 22, 1985.

Lucas Lamata. Basic protocols in quantum reinforcement learning with superconducting circuits. Scientific reports, 7(1):1–10, 2017.

Tor Lattimore and Csaba Szepesvári. Bandit Algorithms. Cambridge University Press, 2020.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 17(1):1334–1373, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.

Makoto Naruse, Martin Berthel, Aurélien Drezet, Serge Huant, Masashi Aono, Hirokazu Hori, and Song-Ju Kim. Single-photon decision maker. Scientific reports, 5(1):1–9, 2015.

Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. Physical review letters, 113(13):130503, 2014.

Mathieu Roget, Stéphane Guillet, Pablo Arrighi, and Giuseppe Di Molfetta. Grover search as a naturally occurring phenomenon. Physical Review Letters, 124(18):180501, 2020.

Maria Schuld and Francesco Petruccione. Supervised learning with quantum computers, volume 17. Springer, 2018.

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. The MIT Press, second edition, 2018.

William R. Thompson. on the likelihood that one unknown probability distribution exceeds another in view of the evidence ot two samples. Biometrika, 25(3-4):285–294, 12 1933.

Peter Wittek. Quantum machine learning: what quantum computing means to data mining. Academic Press, 2014.