2023 International Symposium on Nonlinear Theory and Its Applications
NOLTA2023, September 26-29, 2023, Catania and Online

Quantum-Walk-Based Bandit Algorithm and Its Performance

Tomoki Yamagami†, Etsuo Segawa‡, Takatomo Mihana†,
André Röhm†, Ryoichi Horisaki†, and Makoto Naruse†

†Department of Information Physics and Computing, Graduate School of Information Science and Technology,
The University of Tokyo, 7–3–1 Hongo, Bunkyo, Tokyo 113-8656, Japan
‡Graduate School of Environment and Information Sciences, Yokohama National University,
79–1 Tokiwadai, Hodogaya, Yokohama, Kanagawa 240-8501, Japan

Email: yamagami-tomoki-qwb@g.ecc.u-tokyo.ac.jp

Abstract— This study presents a novel approach to address multi-armed bandit (MAB) problems through the utilization of quantum walks (QWs). QWs exhibit a distinctive characteristic known as the coexistence of linear spreading and localization. While this property has been utilized in various applications, its application to decision-making is almost untapped. This paper presents an algorithm that leverages the coexisting behaviors of QWs to tackle MAB problems, which are recognized as one of the fundamental models in decision-making. By associating the two fundamental operations of exploration and exploitation with the behaviors of QWs, this study demonstrates that the proposed policy outperforms the corresponding random-walk-based model.

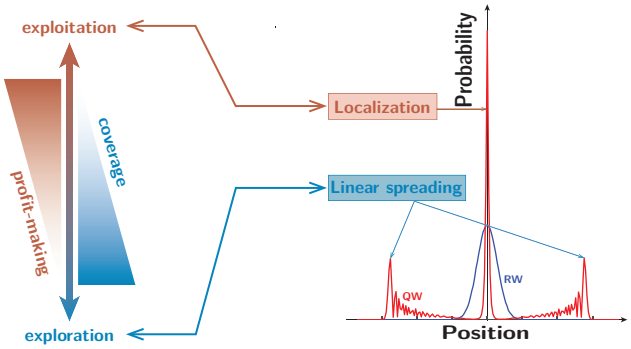

Figure 1: Association between the properties of quantum walks (linear spreading and localization) and the operations in MAB problems (exploration and exploitation).

1. Introduction

A quantum walk (QW) [1] is the quantum counterpart of the classical random walk (RW), encompassing quantum superposition and time evolution effects. In classical RWs, a random walker (RWer) probabilistically chooses the direction to move at each time step, allowing for the tracking of the RWer's position at any time step. In contrast, QWs do not reveal the precise location of a quantum walker (QWer) during the time evolution; the location becomes ascertainable only after conducting the measurement. QWs possess a unique characteristic absent in classical RWs: the coexistence of *linear spreading* and *localization*. Consequently, QWs exhibit probability distributions that significantly differ from the weak convergence to normal distributions observed in RWs. The former implies that the standard deviation of the probability distribution for QWs increases proportionally to the runtime t, which is quadratically faster than that for RWs. The latter signifies that the probability remains distributed at a specific position regardless of the duration of the walk. RWs, on the other hand, exhibit flattening probability distributions despite retaining a bell-shaped curve, thus lacking localization.

While this property has captivated various fields and has been considered for numerous applications, its potential application to decision-making remains largely untapped. This paper introduces novel solution schemes for multi-armed bandit (MAB) problems [2] employing QWs. MAB problems entail multiple slot machines, each with an assigned *success probability* representing the chance of a reward. The objective is for an agent, initially unaware of these probabilities, to maximize cumulative rewards by iteratively selecting machines and obtaining probabilistic rewards. To make a better decision, it is required to gather information on success probabilities through a certain number of selections, which we call *exploration*. On the other hand, it is also necessary to spend some rounds to bet on reliable machines based on acquired data, which we call *exploitation*. Balancing these operations, known as the *exploration–exploitation dilemma* [3], presents a challenge in MAB problems. This study addresses this challenge by leveraging the distinct property of QWs, namely the coexistence of linear spreading and localization. Specifically, we combine exploration with linear spreading and exploitation with localization, as illustrated in Figure 1.

2. Quantum Walk on Cycles

We introduce discrete-time quantum walks on a cycle $\mathcal{C}_N$ with vertices labeled by the set $V_N := \{0, 1, \cdots, N-1\}$ in clockwise order, where N is a natural number. It should be noted that addition and subtraction in V_N are modulo N;

ORCID iDs T.Y.: 0000-0003-3003-1935, E.S.: 0000-0001-8279-9108, T.M.: 0000-0002-4390-710X, A.R.: 0000-0002-8552-7922, R.H.: 0000-0002-2280-5921, M.N.: 0000-0001-8982-9824

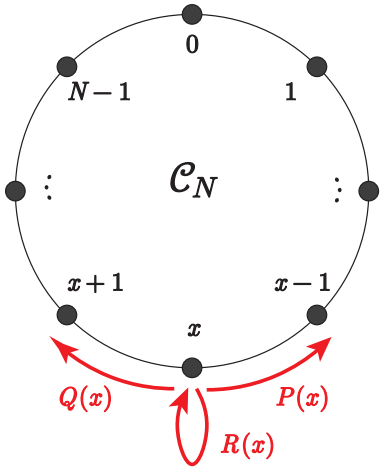

Figure 2: Quantum walks on cycle $\mathcal{C}_N$ and matrices $P(x)$, $Q(x)$, and $R(x)$ for time evolution.

i.e., $(N-1)+1 \equiv 0$ and $0-1 \equiv N-1$.

Our QW model operates in a compound Hilbert space comprising the position Hilbert space $\mathcal{H}_P := \mathrm{span}\{|x\rangle \mid x \in V_N\}$ spanned by the computational basis states representing vertices on $\mathcal{C}_N$, and the coin Hilbert space $\mathcal{H}_C := \mathrm{span}\{|-\rangle, |\mathrm{O}\rangle, |+\rangle\}$ spanned by the computational basis states corresponding to the *internal states*. Here, this model assumes the existence of three internal states: clockwise (+), anti-clockwise (−), and staying (O), and the states are represented by the vectors $|-\rangle = [1\ 0\ 0]^\mathsf{T}$, $|\mathrm{O}\rangle = [0\ 1\ 0]^\mathsf{T}$, and $|+\rangle = [0\ 0\ 1]^\mathsf{T}$, where a superscript T on a matrix represents its transpose. Based on $\mathcal{H}_P$ and $\mathcal{H}_C$, the whole system is described by

$$\begin{aligned}\mathcal{H}_{PC} &= \mathcal{H}_P \otimes \mathcal{H}_C \\ &= \mathrm{span}\{|x\rangle \otimes |\varepsilon\rangle \mid x \in V_N,\ \varepsilon \in \{\pm, \mathrm{O}\}\}.\end{aligned}$$

Then the total state of our QW at time step $t \in \mathbb{N} \cup \{0\}$ is represented as follows: for each $x \in V_N$, there exists $|\psi^{(t)}(x)\rangle \in \mathcal{H}_C$ such that

$$|\Psi^{(t)}\rangle = \sum_{x \in V_N} |x\rangle \otimes |\psi^{(t)}(x)\rangle \in \mathcal{H}_{PC}.$$

Here, $|\psi^{(t)}(x)\rangle$ is called the *probability amplitude vector* at position $x \in V_N$ at time step t. The initial state is set as

$$|\Psi^{(0)}\rangle = |\Phi\rangle := |s\rangle \otimes |\mathrm{O}\rangle$$

with $s \in V_N$. It indicates that QWers start from position s with probability amplitude vector $|\mathrm{O}\rangle$.

Let us now discuss the time evolution of $|\Psi^{(t)}\rangle$. First, we define a site-dependent unitary matrix $C(x)$ as

$$C(x) = \begin{bmatrix} -\frac{1+\cos\theta(x)}{2} & \frac{\sin\theta(x)}{\sqrt{2}} & \frac{1-\cos\theta(x)}{2} \\ \frac{\sin\theta(x)}{\sqrt{2}} & \cos\theta(x) & \frac{\sin\theta(x)}{\sqrt{2}} \\ \frac{1-\cos\theta(x)}{2} & \frac{\sin\theta(x)}{\sqrt{2}} & -\frac{1+\cos\theta(x)}{2} \end{bmatrix}$$

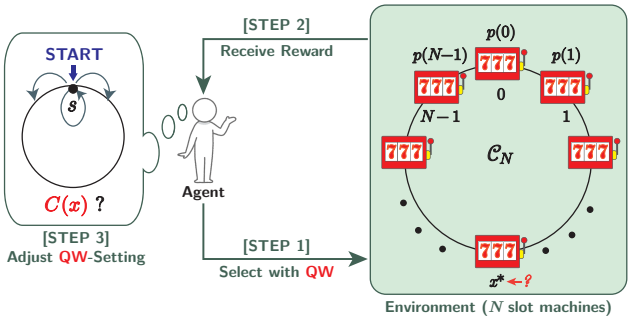

Figure 3: Single decision on the quantum-walk-based model for MAB problems

with $\theta(x) \in [0, 2\pi)$ for all $x \in V_N$. This matrix $C(x)$ is called a *coin matrix*, which governs the variation of the internal states on $x \in V_N$. Next, we define matrices $P(x) = |-\rangle\langle -|\, C(x)$, $Q(x) = |+\rangle\langle +|\, C(x)$, and $R(x) = |\mathrm{O}\rangle\langle \mathrm{O}|\, C(x)$. Using them, the time evolution of $|\Psi^{(t)}\rangle$ is defined by

$$|\Psi^{(t+1)}\rangle = U\,|\Psi^{(t)}\rangle, \tag{2.1}$$

where

$$\begin{aligned}U = \sum_{x \in V_N} \big(&|x-1\rangle\langle x| \otimes P(x) \\ &+ |x\rangle\langle x| \otimes R(x) + |x+1\rangle\langle x| \otimes Q(x)\big).\end{aligned} \tag{2.2}$$

The matrices $P(x)$, $Q(x)$, and $R(x)$ are considered to be the decomposition elements of $C(x)$; i.e., $P(x)+Q(x)+R(x) = C(x)$. They correspond to the transition clockwise, the transition anti-clockwise, and remaining in place, as shown in Figure 2. By Eqs. (2.1) and (2.2), we obtain

$$\begin{aligned}|\psi^{(t+1)}(x)\rangle = {}& P(x+1)\,|\psi^{(t)}(x+1)\rangle \\ &+ R(x)\,|\psi^{(t)}(x)\rangle + Q(x-1)\,|\psi^{(t)}(x-1)\rangle.\end{aligned}$$

Finally, the measurement probability of the particle at position x at time step t, denoted by $\mu^{(t)}(x)$, is given by

$$\mu^{(t)}(x) := \|\psi^{(t)}(x)\|^2.$$

This definition is based on the Born rule of quantum mechanics and satisfies the requirements for the probability measure for any $t \in \mathbb{N} \cup \{0\}$.

3. Proposed Method

We consider the N-armed bandit problem with cycle $\mathcal{C}_N$; where N slot machines with probabilistic rewards are bijectively linked to vertices on $\mathcal{C}_N$. The principle involves the initialization of the QW-setting and three additional steps shown in Figure 3. These three steps are repeated iteratively, wherein we call a round a *decision*. A run consists of multiple decisions, typically performed J times.

[STEP 0] QW-setting initialization

For the first decision, the settings of the quantum walk are determined as follows:

- The initial position s_1 is probabilistically determined by the uniform distribution on V_N, resulting in the initial state $|\Phi_1\rangle = |s_1\rangle \otimes |\mathrm{O}\rangle$.
- The parameter of coin matrices is $\theta_1(x) = \theta^\circ \in [0,\ 2\pi)$ for all $x \in V_N$.

After this step, the run proceeds to the next three steps.

[STEP 1] Quantum walk

Quantum walks are executed for T time steps with the initial position s_j and the parameter $\theta_j(x)$. After running T steps of time evolution, the QWer is measured, yielding the value $\hat{x}_j \in V_N$ according to the probability distribution $\mu^{(T)}(x)$.

[STEP 2] Slot machine play

The slot machine $\hat{x}_j \in V_N$ obtained in **[STEP 1]** is played, and the reward is probabilistically obtained. Consequently, the empirical success probability $\hat{p}_j(x)$ of slot machine $x = \hat{x}_j$ is updated.

[STEP 3] QW-setting adjustment

Using the new empirical success probability $\hat{p}_j(x)$, the QW-setting is updated for the next decision:

- The new initial position s_{j+1} is the one whose $\hat{p}_j(x)$ is provisionally highest among all N. Thus, the new initial state is $|\Phi_{j+1}\rangle = |s_{j+1}\rangle \otimes |\mathrm{O}\rangle$.
- The new parameter of the coin matrices is

$$\theta_{j+1}(x) = \theta^\circ \exp(-a \cdot \hat{p}_j(x)^b),$$

where $a,\ b \geq 1$, and θ° is defined in **[STEP 0]**.

After this step, the process returns to **[STEP 1]**.

4. Numerical Simulation

In this section, we give simulation results for our proposed model. We conduct runs in parallel K times to assess the efficiency of our model. To evaluate the impact of linear spreading and localization, we consider a scenario where these elements are removed from the model. Specifically, we construct an RW-based model for MAB problems corresponding to the QW-based model and compare the performance of the QW-based and RW-based models.

4.1. Random-Walk-Based Algorithm

Here we introduce discrete-time random walks (RWs) on cycle $\mathcal{C}_N$. The position of walkers is determined as follows:

- Initially, a walker exists at position $s \in V_N$.
- At each time step, a walker at position x moves one unit clockwise with probability $q(x)$, moves one unit anti-clockwise with probability $q(x)$ or stays on the current position with probability $1 - 2q(x)$.

Note that the probabilities of moving clockwise and anti-clockwise are equal in this paper. Additionally, the condition $0 \leq q(x) \leq 1/2$ must be satisfied. Denoting the existence probability of walkers by $v^{(t)}(x)$, the rules above are represented by the following equations:

Table 1: Parameter values used for numerical simulations of the QW- and RW-based models.

Parameter	Symbol	Value
# of slot machines	N	32
# of runs	K	500
# of decisions for a run	J	5000
Params. (QW-based)	$(a,\ b,\ \theta^\circ)$	$(5,\ 6,\ 5\pi/16)$
Params. (RW-based)	$(a,\ b,\ q^\circ)$	$(9,\ 6,\ 0.5)$

$$v^{(0)}(s) = 1, \quad v^{(0)}(x) = 0 \ \ (x \neq s);$$
$$v^{(t+1)}(x) = q(x+1)v^{(t)}(x+1) + (1 - 2q(x))v^{(t)}(x) + q(x-1)v^{(t)}(x-1).$$

Recall that addition and subtraction are modulo N.

The RW-based algorithm for the MAB problem is constructed based on the analogy of the QW-based one. Herein, similar to the QW-based case, the initial position is updated, and instead of the coin parameter $\theta(x)$, the transition probability $q(x)$ varies using an exponential function that depends on the empirical success probability:

$$q_{j+1}(x) = q^\circ \exp(-a \cdot \hat{p}_j(x)^b).$$

4.2. Comparison with RW-Based Algorithm

Here we compare the QW- and RW-based models. The parameter values used for this series of simulations are summarized in Table 1. The parameters are selected based on their superior performances within the range of $a = 1,\ 3,\ 5,\ 7,\ 9$ and $b = 2,\ 4,\ 6$ for the respective models.

The blue and orange curves in Figure 4 demonstrate the performances of QW- and RW-based models as the variations of the mean of total reward M, respectively. For $T \geq 4$, we see that the mean of M for the QW-based model is larger than that of the RW-based model. This result indicates that the performance of the QW-based model is superior to that of the RW-based model under some settings.

Linear spreading and localization's impact on decision-making can be observed by analyzing variations. Figures 5(a) and (b) illustrate the relationship between the

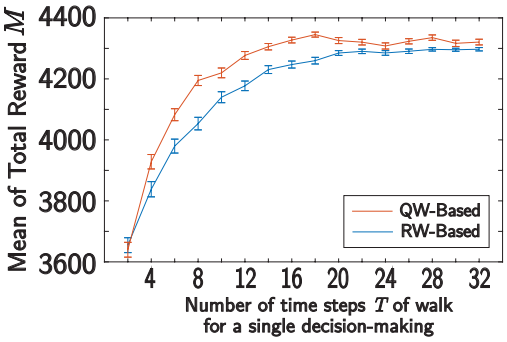

Figure 4: Mean of total reward M over the variation of final time step T of walks of the QW- and RW-based models. Parameters are determined as shown in Table 1.

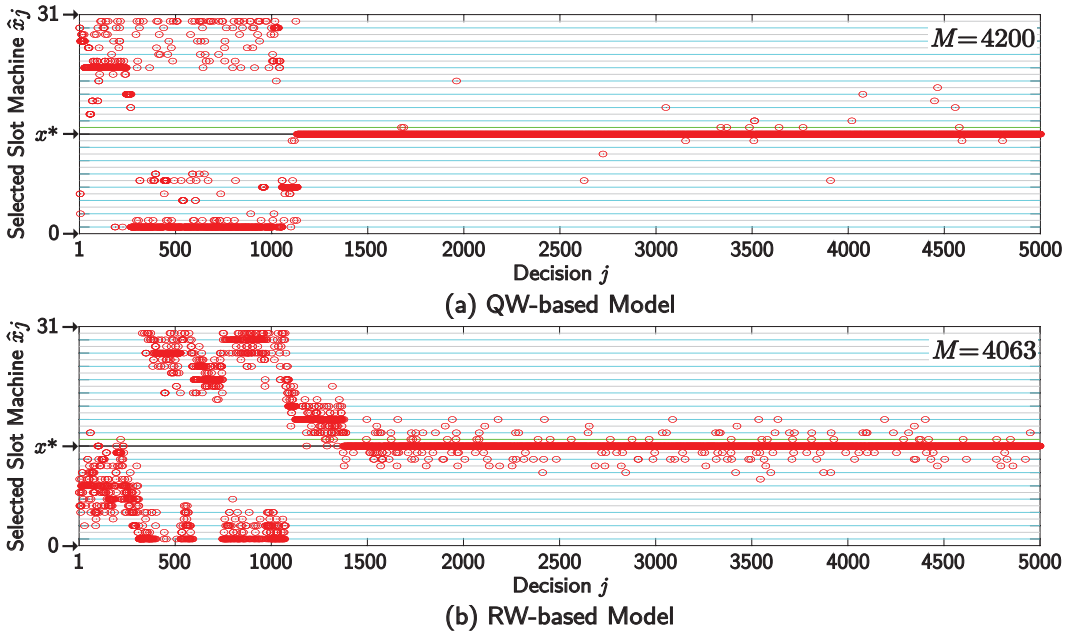

Figure 5: The red markers show selected slot machine $\hat{x}_j$ over decision j in single runs of the (a) QW- and (b) RW-based models. The number of time steps T per decision is 8, and other parameters are as in Table 1. Each run has total reward M almost equal to the average values on each model. The black, sky blue, gray, and light green lines indicate the slot machines with success probabilities 0.9, 0.7, 0.5, and 0.1, respectively. In this paper, only $x = x^* = 14$ has the success probability 0.9, so x^* is regarded as the *best slot machine*.

decision j and the selected slot machine $\hat{x}_j$ in single runs of the QW- and RW-based models, respectively. Therein, you see that decision-making in the QW-based model converges to the best slot machine x^* around $j = 1200$, while the RW-based one does around $j = 1400$. This indicates that the QW-based model exhibits a higher level of exploration than the RW-based one. The phenomenon of linear spreading leads to a wider probability distribution of QWs compared to RWs, facilitating the faster exploration of the QW-based model. Additionally, the behavior of the QW-based model after discovering x^* is more stable than that of the RW-based model. This suggests that the QW-based model achieves more effective exploitation than the RW-based model with this parameter set. After beginning the concentrated investments to x^*, strong localization occurs on vertex x^* in the QW-based model, contributing to this behavior.

5. Conclusion

This paper has introduced a novel approach for multi-armed bandit (MAB) problems using quantum walks (QWs). We have demonstrated that the QW-based model can outperform the random-walk-based one by effectively addressing the exploration–exploitation dilemma via the unique property of QWs: the coexistence of linear spreading and localization. Our approach combines exploration with linear spreading and exploitation with localization. By using linear spreading, the QWs cover the entire environment, minimizing the risk of missing the best slot machine. Simultaneously, localization helps continue using the promising slot machine with a higher probability distribution. Indeed, we showed that, under some settings, linear spreading contributes to exploring the environment and quickly finding the best slot machine, and localization contributes to exploiting the best slot machine more frequently.

Acknowledgments

This work was supported by the SPRING program (JPMJSP2108), the CREST project (JPMJCR17N2) funded by the Japan Science and Technology Agency, and Grant-in-Aid for JSPS Fellows (JP23KJ0384), Grants-in-Aid for Scientific Research (JP20H00233), and Transformative Research Areas (A) (JP22H05197) funded by the Japan Society for the Promotion of Science.

References

[1] N. Konno, "Quantum walks," in *Quantum potential theory*, pp. 309–452, Springer, 2008.

[2] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[3] N. D. Daw, J. P. O'doherty, P. Dayan, B. Seymour, and R. J. Dolan, "Cortical substrates for exploratory decisions in humans," *Nature*, vol. 441, no. 7095, pp. 876–879, 2006.