Reinforcement learning for Quantum Tiq-Taq-Toe

Catalin Dinu and Thomas Moerland

Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

1 Introduction

Quantum Tiq-Taq-Toe [5] is a well-known benchmark and playground for both quantum computing and machine learning. Despite its popularity, no reinforcement learning (RL) methods have been applied to Quantum Tiq-Taq-Toe. Although there has been some research on Quantum Chess [15,3], this game is significantly more complex in terms of computation and analysis. Therefore, we study the combination of quantum computing and reinforcement learning in Quantum Tiq-Taq-Toe (code for our work can be found [1]), which may serve as an accessible testbed for the integration of both fields.

2 Methodology

Quantum games are challenging to represent classically due to their inherent partial observability and the potential for exponential state complexity. In Quantum Tiq-Taq-Toe, states are observed through Measurement (a 3x3 matrix of state probabilities) and Move History (a 9x9 matrix of entanglement relations), making strategy complex as each move can collapse the quantum state.

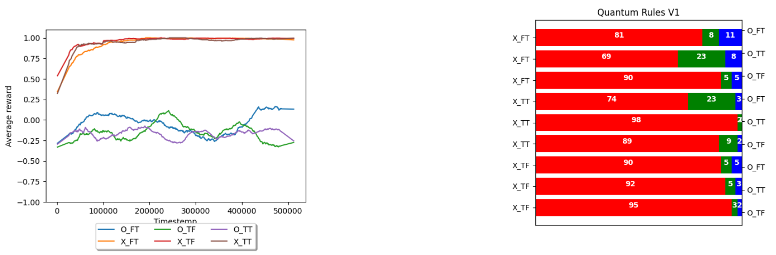
Our study examines two versions of Quantum Tiq-Taq-Toe from the *quantumlib* repository [6]. The first version restricts entanglement moves to include at least one empty cell, blending traditional rules with quantum mechanics. The second version lifts these restrictions, allowing more diverse quantum states and interactions, thereby increasing strategic depth.

3 Results

We conducted a comparative analysis of self-play PPO [12,13,10] agents in Quantum Tiq-Taq-Toe, exploring their performance with access to both measurement matrices and historical entanglement records (TT agent), as well as with access to only the measurement matrix (TF) or historical entanglement record (FT).

For the first set of rules, which imposes constraints on entanglement moves, we observe a tendency for the first player to gain an advantage (Fig. 1). This advantage is noticeable despite inherent randomness in the game, which prevents guaranteed wins, and therefore suggests the presence of discernible winning strategies.

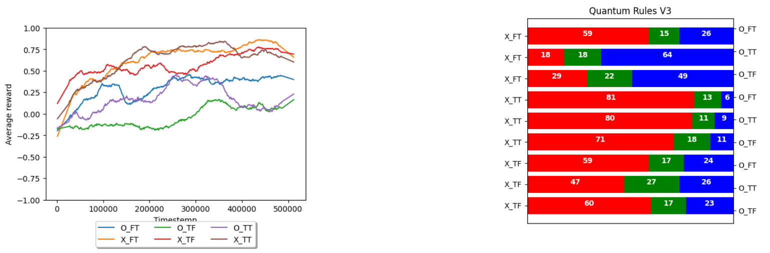
Fig. 1: RL results on Quantum Tiq-Tac-Toe Version 1



(a) Average reward on 100 games during the training of different agents

(b) Pitting best agents: Red → X-Wins, Blue → O-Wins, Green → Draws

Fig. 2: RL results on Quantum Tiq-Tac-Toe Version 3



(a) Average reward on 100 games during the training of different agents

(b) Pitting best agents: Red → X-Wins, Blue → O-Wins, Green → Draws

For the third set of rules, which allows for triple entanglement, the combined state of measurement matrix and historical entanglement yields optimal performance based on the pitting results (Fig. 2). This integrated approach enables agents to utilize real-time state probabilities and insights from past game interactions, leading to more equitable outcomes between players. It also underscores the importance of comprehensive information in strongly partially observable quantum environments.

4 Discussion

Most quantum problems require both precise control and mitigation of partial observability, making machine learning particularly suitable. Indeed, RL has already shown promise in fields like quantum error correction [2,11]. We identify Quantum Tiq-Taq-Toe, with its various subtypes, as an accessible testbed for the development of RL methods in the quantum setting. Future work could investigate other methods to mitigate partial observability, such as the use of state windowing [9], recurrent neural networks [8], recurrent state space models [7], or transformers [14].

References

1. Our work. `https://github.com/Dinu23/Quantum-Tiq-Taq-Toe.git`
2. Andreasson, P., Johansson, J., Liljestrand, S., Granath, M.: Quantum error correction for the toric code using deep reinforcement learning. Quantum **3**, 183 (2019)
3. Cantwell, C.: Quantum chess: Developing a mathematical framework and design methodology for creating quantum games. arXiv preprint arXiv:1906.05836 (2019)
4. Chiofalo, M.L., Foti, C., Michelini, M., Santi, L., Stefanel, A.: Games for teaching/learning quantum mechanics: a pilot study with high-school students. Education Sciences **12**(7), 446 (2022)
5. Goff, A.: Quantum tic-tac-toe: A teaching metaphor for superposition in quantum mechanics. American Journal of Physics **74**(11), 962–973 (2006)
6. Google: Quantumlib. `https://github.com/quantumlib/unitary.git`
7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
9. Lin, L.J., Mitchell, T.M.: Reinforcement learning with hidden states (1993)
10. Liu, S., Cao, J., Wang, Y., Chen, W., Liu, Y.: Self-play reinforcement learning with comprehensive critic in computer games. Neurocomputing **449**, 207–213 (2021)
11. Nautrup, H.P., Delfosse, N., Dunjko, V., Briegel, H.J., Friis, N.: Optimizing quantum error correction codes with reinforcement learning. Quantum **3**, 215 (2019)
12. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
13. Tang, C.Y., Liu, C.H., Chen, W.K., You, S.D.: Implementing action mask in proximal policy optimization (ppo) algorithm. ICT Express **6**(3), 200–203 (2020)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
15. Youvan, D.C.: Sequential superposition collapse in quantum chess: A novel approach to quantum strategies and decision-making (2024)

Appendix

A. Environment

Quantum Tiq-Taq-Toe [4] is an altered variant of the classic Tic-Tac-Toe game. In the traditional game, each cell on a 3x3 board can be empty, X, or O. In the quantum version, each cell is a qutrit, existing in a superposition of three quantum states: empty ($|\_\_\rangle$), X ($|X\rangle$), or O ($|O\rangle$). Given the full board state ($|\eta\rangle$), the probability of collapsing to a specific state $|c_1 c_2 ... c_8 c_9\rangle$ ($c_i \in \{\_\_, X, O\}$) is:

$$|\eta\rangle = \sum_{\phi \in \{\_\_, X, O\}^9} \alpha_\phi \, |\phi\rangle \, , \sum_{\phi \in \{\_\_, X, O\}^9} \alpha_\phi^2 = 1$$

$$\mathbb{P}(\eta = c_1 c_2 ... c_8 c_9) = || \, \langle c_1 c_2 ... c_8 c_9 | \eta \rangle \, ||^2 = \alpha^2_{c_1 c_2 ... c_8 c_9}$$

where $|\eta\rangle$ is the quantum state of the system, that can be express as a linear combination of all possible classical states ($|\phi\rangle$) with α_ϕ^2 being the probability to observe the state $|\phi\rangle$.

Action Space In classical Tic-Tac-Toe, a move changes an empty cell to X or O (left of Figure 3). In the quantum version, these moves are $X_{NOT} \, |\_\_\rangle \rightarrow |X\rangle$ and $O_{NOT} \, |\_\_\rangle \rightarrow |O\rangle$.

Additionally, quantum moves involve entangled pairs of cells. These moves create two possible states: one with X/O in the first cell and another with X/O in the second cell, leading to complex quantum states. The simplest case entangles two empty cells with X/O, resulting in a 50% probability of X/O appearing in either cell (right of Fig. 3).

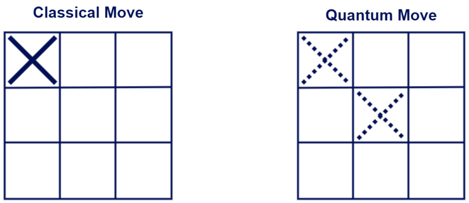


Fig. 3: Most simple classical/quantum moves allowed during the game

State Collapsing A pivotal phenomenon in the quantum game is State Collapsing, as illustrated in Fig.4. This occurrence occurs when the game board becomes saturated with moves, utilizing both quantum and classical moves that impact all cells on the board. Upon the state collapsing, a specific state is selected from the multitude of possible states. This selection is determined by the probability distribution outlined by the existing quantum state ($|\eta\rangle = \sum_{\phi \in \{\_\_, X, O\}^9} \alpha_\phi \, |\phi\rangle$).

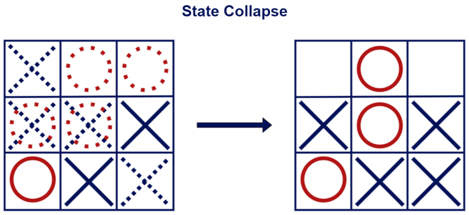


Fig. 4: State collapsing after filling all the cells with anything (quantum/classical moves)

In addition, we distinguish two sets of game rules (Fig.5) that affect the game play:

- Version 1 (V1) - reduce the list of available entanglements moves only to pairs of cells that contain at least one free cell (we consider a free cell as a cell that was not used for any previous moves).
- Version 3 (V3) - any combination of two cells is a valid entanglement move.

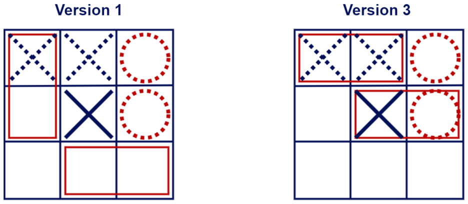


Fig. 5: Available moves considering the two sets of game rules

B. Observation Space

A paramount challenge in this game revolves around effectively representing quantum information in a classical format to facilitate an Agent's learning process. The most straightforward method involves classically storing the quantum state and simulating the game. However, this approach is deemed unreliable due to the impracticality of saving a 9-qutrit quantum state, which requires complex numbers. This not only contradicts the essence of a quantum game but also poses a significant computational burden.
To address this challenge, two classical pieces of information are explored in this report: Measurements and Moves History. These representations offer a more manageable way to capture and convey quantum aspects within the framework of the game, enabling effective learning for an Agent.

Measurements A piece of essential information that can help an Agent learn to play the game would be the probability of each cell being in either _/X/O state. However, to compute the real values of those would imply the access to the quantum state. To overcome this, we can estimate those probabilities. Given the quantum state of the game board $|\eta\rangle$, we can simulate the state collapsing a number of times (N) and estimate the probabilities $\mathbb{P}(\widehat{c_i = \_})$, $\mathbb{P}(\widehat{c_i = X})$and $\mathbb{P}(\widehat{c_i = O})$ as the number of appearances of _/X/O on each cell divided by N. The relation between the estimates and real values is: $\mathbb{P}(\widehat{c_i = \_}) = \mathbb{P}(c_i = \_) + \mathcal{N}(0, \frac{1}{N})$, $\mathbb{P}(\widehat{c_i = X}) = \mathbb{P}(c_i = X) + \mathcal{N}(0, \frac{1}{N})$ and $\mathbb{P}(\widehat{c_i = O}) = \mathbb{P}(c_i = O) + \mathcal{N}(0, \frac{1}{N})$

So, these estimations provide a practical means for an Agent to learn and make decisions based on approximated probabilities, offering a computationally feasible approach in the absence of direct access to the precise quantum state.

Moves History An additional informative resource for an agent's learning process involves maintaining a history of past moves. This data is structured using two matrices, one for X and one for O, each with dimensions of 9 × 9 (MH^X and MH^O). The matrices are defined as follows:

- $MH_{i,j}^{X/O}$: Represents the number of moves entangling c_i and c_j using X/O, where $i, j \in \{1, ..., 9\}$ and $i \neq j$.
- $MH_{i,i}^{X/O}$: Indicates the number of classical moves using X/O on c_i, where $i \in \{1, ..., 9\}$.

This fixed-dimension representation efficiently captures the historical moves in a structured manner, providing valuable information for the agent's learning process.