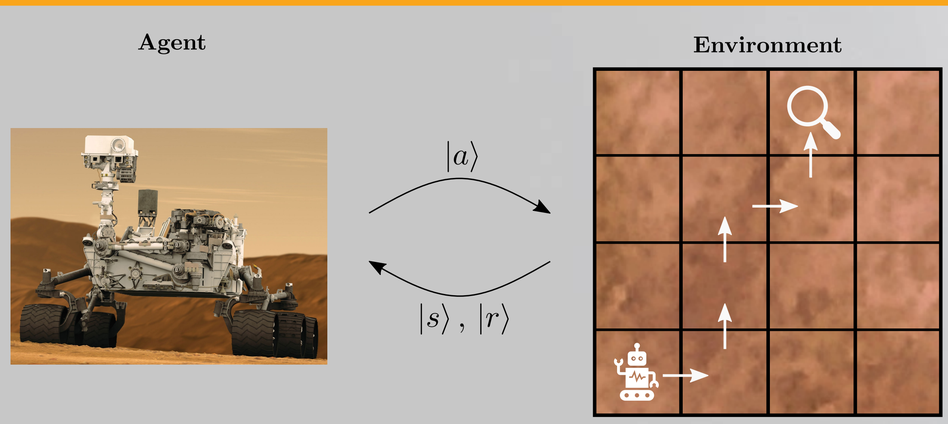
A Quantum Enhanced Learning Algorithm for Maze Problems

Deutsches Zentrum für Luft- und Raumfahrt

Oliver Sefrin and Sabine Wölk

Motivation



Reinforcement Learning (RL):

- Interaction between an *agent* and an *environment*: the agent performs, at timestep t, an action a_t and receives the new state s_{t+1} of the environment and a reward r_{t+1} in return
- Agent's behavior is determined by its current *policy* Π
- Goal: learn optimal policy which maximizes the expected cumulative reward $G_t \equiv \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, where $0 \leq \gamma \leq 1$ is a discount factor
- The agent learns by interacting with the environment and updating its policy with the received information

Grid-World:

- Environment: 2D grid with a start and finish cell; cell positions as states; finish cell is the only rewarded state
- Agent's actions: $\{\text{left}, \text{right}, \text{up}, \text{down}\}$
- Goal: find the shortest path from start to finish
- *Episodic*: the agent is reset to its starting cell after finding a reward or after a certain number of steps (strictly episodic)

Hybrid RL Algorithm

- Quadratic speedup in terms of sample complexity compared to a classical algorithm [3, 4]
- *Hybrid*: enhancement of a classical RL agent (e.g., SARSA, Q-Learning, Projective Simulation) with amplitude amplification

Prerequisites:

- Quantum registers for actions $|a\rangle_A$, states $|s\rangle_S$, and rewards $|r\rangle_R$
- In a deterministic and strictly episodic environment, the environment's response on an entire action sequence $\vec{a}$ of length L can be described as

$$U_{\text{Env}} |\vec{a}\rangle_A |0\rangle_S |0\rangle_R = |\vec{a}\rangle_A |\vec{s}(\vec{a})\rangle_S |r(\vec{a})\rangle_R . \tag{5}$$

The unitary U_{Env} can be used to create a phase kick-back oracle O_E (see eq. (3)).

Algorithm:

1. For a chosen *search length* L, prepare the weighted superposition of all sequences: $|\psi\rangle = \sum_{\vec{a}} \sqrt{\Pi(\vec{a})} |\vec{a}\rangle$
2. Prepare the reward register in the $|-\rangle_R$ state.
3. Apply the Grover operator $G = (2 |\psi\rangle \langle\psi| - \mathbb{1}) O_E$ (k times). Here, we choose k according to [1].
4. Measure the action register to receive a candidate sequence $\vec{a'}$.
5. Play $\vec{a'}$ with the classical agent and update the agent's policy Π according to the outcome.
6. If the estimated reward probability exceeds the quantum profitability threshold, continue training classically; else, return to step 1.

Grover Algorithm

Task: Out of N items in an unstructured database, M are marked. Find a marked item. [2]

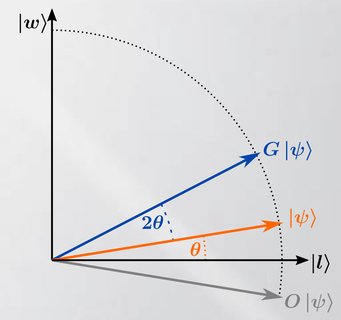


Figure 1: Illustration of the Grover operation in the 2D space spanned by $|w\rangle$ and $|l\rangle$.

Algorithm:

1. Prepare the superposition of all items:

$$|\psi\rangle = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} |i\rangle = \sin\theta |w\rangle + \cos\theta |l\rangle \tag{1}$$

with $\sin\theta = \sqrt{\frac{M}{N}}$, $|w\rangle = \frac{1}{\sqrt{M}} \sum_{i \in \mathcal{M}} |i\rangle$, $|l\rangle = \frac{1}{\sqrt{N-M}} \sum_{i \notin \mathcal{M}} |i\rangle$.

2. Apply the Grover operator G (k times):

$$G = (2 |\psi\rangle \langle\psi| - \mathbb{1}) O \tag{2}$$

with the *oracle* O:

$$O |i\rangle = \begin{cases} - |i\rangle , & \text{if item } i \text{ is marked, } i \in \mathcal{M} \\ |i\rangle , & \text{else.} \end{cases} \tag{3}$$

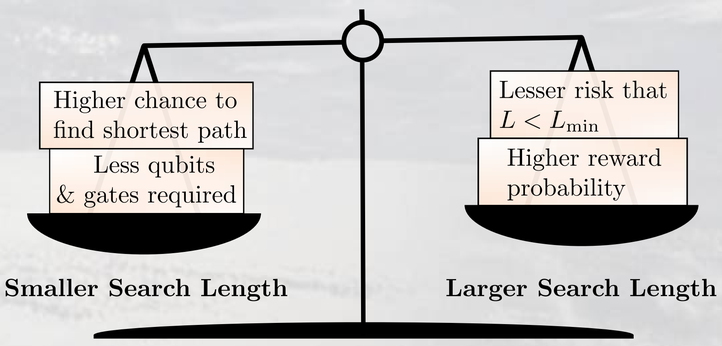
3. Measuring the final state

$$G^k |\psi\rangle = \sin([2k+1]\theta) |w\rangle + \cos([2k+1]\theta) |l\rangle \tag{4}$$

yields success probability $p = \sin^2([2k+1]\theta)$.

The Search Length Dilemma

Which search length L should we choose for the hybrid RL algorithm?



Research Questions:

1. Given that we know the length $L_{\min}$ of the shortest winning path, how do different strategies with varying L perform?
2. Do we still have a quantum advantage in terms of performed actions if we release the *strictly episodic* condition for the classical agent?
3. Knowing only L_1 and L_2 such that $L_1 < L_{\min} < L_2$, which strategies are optimal?

Fixed Goal Scenario: $L_{\min}$ exactly known

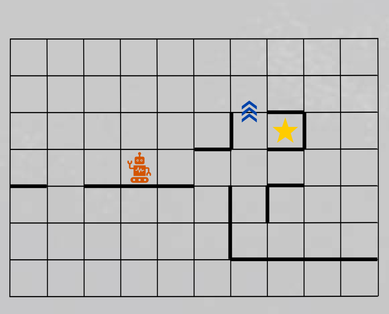
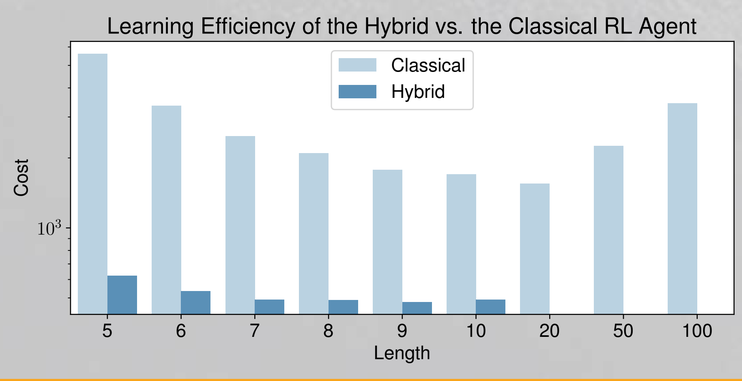
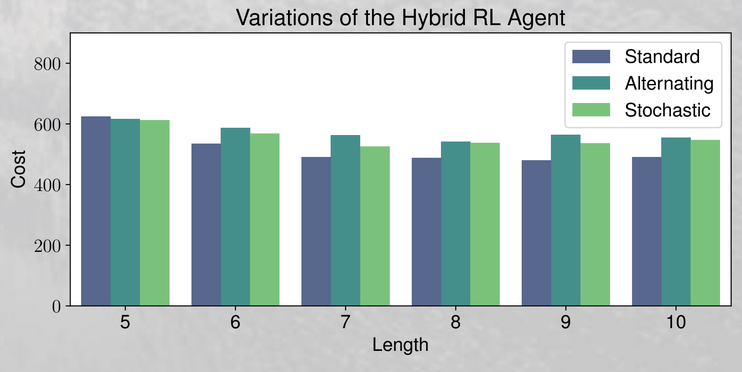


Figure 2: Cutout of the Grid-World layout used in the simulations. (Symbols: thick lines → walls, star → goal, robot → start position, triple arrow → one-way door)

Simulation:

- Grid-World with $L_{\min} = 5$
- Performance metric: *cost* ≡ total number of steps taken
- Training is stopped after five consecutive rewards





Strategies with varying search length:

- *Stochastic*: at the start of an episode, randomly choose between using $L_{\min}$ or the actual length L
- *Alternating*: use the oracle O_E with $L_{\min}$ for the first $k/2$ Grover iterations and the oracle with the actual length L for the second $k/2$ iterations

Bibliography

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

[2] L. K. Grover. Quantum mechanics helps in searching for a needle in a haystack. *Physical review letters*, 79(2):325, 1997.

[3] A. Hamann and S. Wölk. Performance analysis of a hybrid agent for quantum-accessible reinforcement learning. *arXiv preprint arXiv:2107.14001*, 2021. Accepted at New Journal of Physics.

[4] V. Saggio, B. E. Asenbeck, A. Hamann, T. Strömberg, P. Schiansky, V. Dunjko, N. Friis, N. C. Harris, M. Hochberg, D. Englund, et al. Experimental quantum speed-up in reinforcement learning agents. *Nature*, 591(7849):229–233, 2021.

oliver.sefrin@dlr.de

DLR, Institut für Quantentechnologien, 89081 Ulm