Quantum Algorithms for Finite-horizon Markov Decision Processes

Bin Luo[1], Yuwen Huang[1], Jonathan Allcock[2], Xiaojun Lin[1], Shengyu Zhang[2], John C.S. Lui[1]
[1] The Chinese University of Hong Kong, [2] Tencent Quantum Laboratory

Introduction

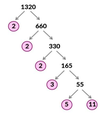
- Markov decision processes (MDPs) provide a framework for modelling decision-making in various environments.


Autonomous Driving


Robotics


Reinforcement Learning

- However, MDPs often face significant computational challenges in practice, e.g. curse of dimensionality.
- Quantum computing has shown a **significant speedup** over classical computing in terms of time complexity for certain problems.



Integer Factorization

$$\begin{aligned} 2x + 7y &= 34 \\ 5x - 4y &= -1 \end{aligned}$$
Solving Linear Systems

Unstructured Search

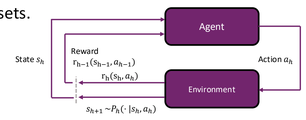
- **Question:** Can one design quantum algorithms that are more efficient than classical algorithms in solving time-dependent and finite-horizon MDPs?
- **Contributions:**
  1. Exact dynamics setting: (a) derive a new **classical lower bound**; (b) propose **Quantum Value Iteration algorithms QVI-1 and QVI-2**; (c) propose a new quantum subroutine, **Quantum Mean Estimation with Binary Oracle**.
  2. Generative model setting: (a) propose quantum algorithms **QVI-3 and QVI-4;** (b) derive **classical and quantum lower bounds**.

MDP Preliminaries

A time-dependent and finite-horizon MDP is defined as a 5-tuple

$$\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, \{P_h\}_{h=0}^{H-1}, \{r_h\}_{h=0}^{H-1}, H\right).$$

- State space $\mathcal{S}$ and action space $\mathcal{A}$ are discrete and finite sets.
- The total time horizon H is a positive integer.
- $P_h(s_{h+1}|s_h, a_h)$ is a transition probability.
- A reward $r_h(s_h, a_h)$ is a scalar in $[0,1]$.



Optimization goal

Denote $[H] = \{0, 1, \ldots, H-1\}$.

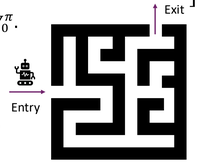
- Find a policy $\pi: \mathcal{S} \times [H] \to \mathcal{A}$ that maximizes the expected cumulative reward (V-value or Q-value function) over a finite number of time steps, where V-value and Q-value functions are defined as:

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^{H-1} r_t(s_t, a_t)|\pi, s_h = s\right],\ Q_h^\pi(s,a) := \mathbb{E}\left[\sum_{t=h}^{H-1} r_t(s_t, a_t)|\pi, s_h = s, a_h = a\right].$$

- $V_h^*(s) := \max_\pi V_h^\pi(s)$, $Q_h^*(s,a) := \max_\pi Q_h^\pi(s,a)$ and $\pi^* := \operatorname{argmax}_\pi V_0^\pi$.

A toy example: Robot-in-Maze

- States: positions in the maze.
- Actions: movements (up, down, left, right).
- Reward: $r_h(s_h, a_h) = 0$ if s_h is the exit; otherwise, $r_h(s_h, a_h) = -1$.
- Objective: minimize the expected number of steps to reach the exit.

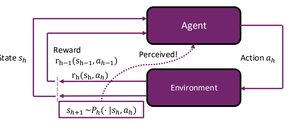


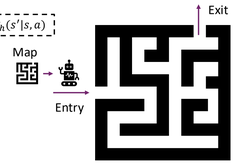
Exact Dynamics Setting

Background

- Assumption: Dynamics of the environment are **fully known** to the agent. (⇔ The robot has the map of the maze.)
- Classical algorithms assume they have access to a classical oracle $O_{\mathcal{M}}: (s,a,h,s') \mapsto \left(r_h(s,a), P_h(s'|s,a)\right)$.
- Classical value iteration (VI) algorithm can obtain π^* and V_0^* with $O(S^2AH)$ queries to $O_{\mathcal{M}}$.
  - Define Bellman optimality value operator $[\mathcal{T}^h(V_{h+1})]_s := \max_{a\in\mathcal{A}}\{r_h(s,a) + P_{h|s,a}^T V_{h+1}\}$.
  - VI algorithm repeatedly applies $\mathcal{T}^h$ on V_{h+1} in a backward manner with $V_H = \mathbf{0}$ and updates the policy following $\pi(s,h) = \operatorname{argmax}_{a\in\mathcal{A}}\{r_h(s,a) + P_{h|s,a}^T V_{h+1}\}$.

Define $P_{h|s,a}(s') = P_h(s'|s,a)$





Summary of the Results

Table 1. Classical and quantum query complexities for solving general finite-horizon MDPs

Goal	Classical query complexity		Quantum query complexity
	Upper bound	Lower bound	Upper bound
Optimal π^*, V_0^*	$O(S^2AH)$	$\boldsymbol{\Omega(S^2A)}$	$\boldsymbol{\widetilde{O}(S^2\sqrt{A}H)}$ **[QVI-1]**
ϵ-accurate estimate of π^* and $\{V_h^*\}_{h=0}^{H-1}$	$O(S^2AH)$	$\boldsymbol{\Omega(S^2A)}$	$\boldsymbol{\widetilde{O}(S^{1.5}\sqrt{A}H^3/\epsilon)}$ **[QVI-2]**

Classical Lower Bound

Theorem (informal version): Given access to the classical oracle $O_{\mathcal{M}}$, any algorithm, which outputs **ϵ-approximations of $\{V_h^*\}_{h=0}^{H-1}$ or π^*** with probability at least 0.9, must require at least $\boldsymbol{\Omega(S^2A)}$ queries to $O_{\mathcal{M}}$ on the worst case of input $\mathcal{M}$.

- Note that the above theorem implies that it also requires at least $\Omega(S^2A)$ queries to $O_{\mathcal{M}}$ to obtain $\{V_h^*\}_{h=0}^{H-1}$ or π^*.
- **Question: whether quantum algorithms can break this barrier in the dependence on action space size ($|\mathcal{A}| := A$) or state space size ($|\mathcal{S}| := S$)?**

Quantum Oracle of Finite-horizon MDPs

Definition: A quantum oracle of a finite-horizon MDP is a unitary operator $O_{Q\mathcal{M}}$ such that
$O_{Q\mathcal{M}}: |s\rangle|a\rangle|h\rangle|s'\rangle|0\rangle|0\rangle \mapsto |s\rangle|a\rangle|h\rangle|s'\rangle|r_h(s,a)\rangle|P_h(s'|s,a)\rangle$
for all $(s,a,h,s') \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}$.

Quantum Speedup on Action Space Size (A)

- Quantum Maximum Searching (QMS) Algorithm [Durr et al., 1999]

Problem: For an unsorted list $f \in \mathbb{R}^N$, one wants to find the index i such that $f(i) = \max_{j\in[N]} f(j)$.

- Classical algorithm: $\Theta(N)$ queries to the vector f.
- Quantum algorithm: $\Theta(\sqrt{N})$ queries to the vector f.
- Suppose $N = 1{,}000{,}000$ and it takes 1 second for each query, then the classical algorithm needs roughly **11. 5 days**, but QMS algorithm only needs roughly **17 minutes!**

Quantum Value Iteration QVI-1 Algorithm

- **Main idea:** apply QMS algorithm when taking the maximum over the whole action space in the classical value iteration algorithm.
- Output: optimal policy π^* and optimal V-value function V_0^*.
- Query complexity: $\boldsymbol{\widetilde{O}(S^2\sqrt{A}H)}$ queries to the quantum oracle $O_{Q\mathcal{M}}$.

Quantum Speedup on State Space Size (S)

- **New Quantum Subroutine**: Quantum Mean Estimation with Binary Oracles (QMEBO)
  - Goal: obtain an ϵ-estimate $\hat{\mu}$ of $\mu = \mathbb{E}[f(x) \in \mathbb{R}^N | x\sim p]$ with probability at least $1-\delta$.
  - Query complexity: $O\left(\left(\frac{\sqrt{N}}{\epsilon} + \sqrt{\frac{N}{\epsilon}}\right)\log\left(\frac{1}{\delta}\right)\right)$ queries to the function f.

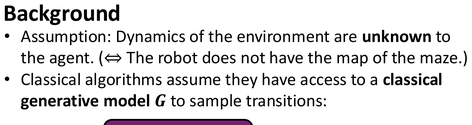
Quantum Value Iteration QVI-2 Algorithm

- **Main idea: apply QMEBO to obtain $\frac{\epsilon}{H}$-estimates of $P_{h|s,a}^T V_{h+1}$ based on QVI-1 algorithm.**
- Output: **ϵ-approximations of π^* and $\{V_h^*\}_{h=0}^{H-1}$.**
- Query complexity: $\boldsymbol{\widetilde{O}(S^{1.5}\sqrt{A}H^3/\epsilon)}$ queries to the quantum oracle $O_{Q\mathcal{M}}$.

Generative Model Setting

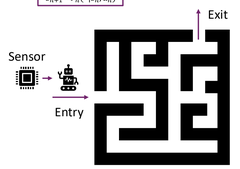
Background

- Assumption: Dynamics of the environment are **unknown** to the agent. (⇔ The robot does not have the map of the maze.)
- Classical algorithms assume they have access to a **classical generative model G** to sample transitions:



(s,a,h) → Generative Model → $s_h^i(s,a) \overset{\text{i.i.d.}}{\sim} P_{h|s,a}, i = 1, \ldots, N$

(⇔ The robot is equipped with a sensor.)



- Optimization goal: obtain ϵ-approximations of $\pi^*, \{V_h^*\}_{h=0}^{H-1}$ and $\{Q_h^*\}_{h=0}^{H-1}$.

Summary of the Results

Table 2. Classical and quantum sample complexities for solving general finite-horizon MDPs

Goal: obtain an ϵ-accurate estimate of	Classical query complexity		Quantum query complexity	
	Upper bound	Lower bound	Upper bound	Lower bound
$\{Q_h^*\}_{h=0}^{H-1}$	$\widetilde{O}\left(\frac{SAH^4}{\epsilon^2}\right)$ [Li et al., 2020]	$\widetilde{\Omega}\left(\frac{SAH^3}{\epsilon^2}\right)$	$\boldsymbol{\widetilde{O}\left(\frac{SAH^{2.5}}{\epsilon}\right)}$ **[QVI-4]**	$\boldsymbol{\widetilde{\Omega}\left(\frac{SAH^{1.5}}{\epsilon}\right)}$
$\pi^*, \{V_h^*\}_{h=0}^{H-1}$	$\widetilde{O}\left(\frac{SAH^4}{\epsilon^2}\right)$ [Li et al., 2020]	$\widetilde{\Omega}\left(\frac{SAH^3}{\epsilon^2}\right)$	$\boldsymbol{\widetilde{O}\left(\frac{SAH^{2.5}}{\epsilon}\right)}$ **[QVI-4]** $\boldsymbol{\widetilde{O}\left(\frac{S\sqrt{A}H^3}{\epsilon}\right)}$ **[QVI-3]**	$\boldsymbol{\widetilde{\Omega}\left(\frac{S\sqrt{A}H^{1.5}}{\epsilon}\right)}$

Quantum Generative Oracle of Finite-horizon MDPs

Definition: A quantum generative oracle of a finite-horizon MDP is a unitary operator $\mathcal{G}$ such that

$$\mathcal{G}: |s\rangle|a\rangle|h\rangle|0\rangle|0\rangle \mapsto |s\rangle|a\rangle|h\rangle\left(\sum_{s'\in\mathcal{S}}\sqrt{P_h(s'|s,a)}\,|s'\rangle|w_{s'}\rangle\right)$$

for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, where $|w_{s'}\rangle$ are arbitrary auxiliary states.

Quantum Mean Estimation (QME) Algorithms [Montanaro, 2015]

- **QME1 is a quantum version of Hoeffding's inequality.**

Problem: for a random variable $X \in [0,u]$, one needs to obtain an ϵ-estimate of $\mathbb{E}[X]$.

  - **Hoeffding's inequality implies that** $O(u^2/\epsilon^2)$ classical samples are required.
  - **QME1** only requires $O(u/\epsilon)$ quantum samples.

- **QME2 is a quantum version of Chebyshev's inequality.**

Problem: for a random variable X with finite non-zero variance σ^2, one needs to obtain an ϵ-estimate of $\mathbb{E}[X]$.

  - **Chebyshev's inequality implies that** $O(\sigma^2/\epsilon^2)$ classical samples are required.
  - **QME2** only requires $\widetilde{O}(\sigma/\epsilon)$ quantum samples.

Quantum Value Iteration QVI-3 Algorithm

- **Main idea:** QVI-3 is similar to QVI-2, but it applies QME1 to obtain $\frac{\epsilon}{H}$-estimates of $P_{h|s,a}^T V_{h+1}$.
- Output: **ϵ-approximations of π^* and $\{V_h^*\}_{h=0}^{H-1}$.**
- Sample complexity: $\boldsymbol{\widetilde{O}\left(\frac{S\sqrt{A}H^3}{\epsilon}\right)}$ **queries to the quantum generative oracle $\mathcal{G}$.**

Quantum Value Iteration QVI-4 Algorithm

- **Main idea:** QVI–4 incorporates quantum adaptations of **"variance reduction"** and **"total variance"** techniques from [Sidford et al., 2018] by applying QME1 and QME2 algorithms.
- Output: **ϵ-approximations of π^*, $\{V_h^*\}_{h=0}^{H-1}$ and $\{Q_h^*\}_{h=0}^{H-1}$.**
- Sample complexity: $\boldsymbol{\widetilde{O}\left(\frac{SAH^{2.5}}{\epsilon}\right)}$ **queries to the quantum generative oracle $\mathcal{G}$.**

Classical and Quantum Lower Bounds

- **Reduction:** We show that solving a finite-horizon MDP can be reduced to solving an infinite-horizon MDP. Therefore, **the lower bounds of solving finite-horizon MDP inherits from those of the infinite-horizon MDP** [Wang et al., 2021].
- **Implication:** QVI-3 and QVI-4 are **asymptotically optimal**, up to log terms, for computing ϵ-approximations of $\pi^*, \{V_h^*\}_{h=0}^{H-1}$ and $\{Q_h^*\}_{h=0}^{H-1}$, provided a constant time horizon.

Contact
Bin Luo
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR
binluo@link.cuhk.edu.hk
WeChat: RobinBinLuo

Notations

S: state space size
A: action space size
H: total time horizon
ϵ: error term
δ: failure probability
$\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ ignore logarithmic factors

References

1. [Durr et al., 1999] *A quantum algorithm for finding the minimum.* Christoph Durr, Peter Hoyer.
2. [Li et al., 2020] *Breaking the sample size barrier in model-based reinforcement learning with a generative model.* Gen Li, Yuting Wei, Yuejie Chi, Yuxin Chen
3. [Montanaro, 2015] *Quantum speedup of Monte Carlo methods.* Ashley Montanaro.
4. [Sidford et al., 2018] *Near-optimal time and sample complexities for solving Markov decision processes with a generative model.* Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, Yinyu Ye.
5. [Wang et al., 2021] *Quantum algorithms for reinforcement learning with a generative model.* Daochen Wang, Aarthi Sundaram, Robin Kothari, Ashish Kapoor, Martin Roetteler