IMPROVED QUANTUM ALGORITHMS FOR REINFORCEMENT LEARNING UNDER A GENERATIVE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning is a subfield of machine learning that studies how an agent interacts with an environment in order to extract as large a reward as possible. A standard approach to study such interaction is through Markov Decision Processes (MDPs) and the task of choosing an optimal policy—a function that tells the agent which action to take. In this work, we study two types of MDPs—finite-horizon and infinite-horizon discounted—and propose new quantum algorithms for approximating optimal policies. Our quantum algorithms are based on a new combination of standard value iteration and quantum subroutines like quantum mean estimation and quantum maximum finding, overall enhanced with techniques from sample-optimal classical algorithms. Our query complexities improve on previous works, thus approaching already established quantum lower bounds.

1 INTRODUCTION

The paradigm of reinforcement learning (RL) (Sutton & Barto, 1998) has received much attention recently due to its success in modeling an agent interaction with a dynamical environment (Åström, 1965; Hu & Yue, 2007; Sato & Zouain, 2010; Bäuerle & Rieder, 2011; Feinberg & Shwartz, 2012; Bennett & Hauser, 2013; Chen et al., 2014; Steimle & Denton, 2017; Natarajan & Kolobov, 2022) and due to its applicability to a wide range of problems (Sutton & Barto, 1998; Szepesvári, 2010; Bertsekas, 2012; 2022). Such interaction is usually studied using Markov Decision Processes (MDPs) (Puterman, 2014), a simple yet powerful mathematical abstraction wherein an agent chooses actions given the environment's state in order to maximise some kind of reward. Despite their versatility, MDPs suffer a "curse of dimensionality" when the number of possible actions or states are sufficiently large, rendering a solution computationally impossible (Powell, 2007).

Parallel to RL developments, quantum computation (Nielsen & Chuang, 2010) has emerged as a new field promised to deliver substantial speedups to traditional problems like factoring (Shor, 1994) and unstructured search (Grover, 1996) by exploiting quantum mechanics. Quantum machine learning (Schuld et al., 2015; Biamonte et al., 2017; Allcock & Zhang, 2019; Alchieri et al., 2021), a subfield of quantum computation, has been consolidated as an area of study that can potentially provide quantum speedups to several traditional machine learning problems (Lloyd et al., 2013; Rebentrost et al., 2014; Wiebe et al., 2015; Kerenidis & Prakash, 2017; Kerenidis et al., 2019; Liu et al., 2021; Doriguello et al., 2023), including RL (Dong et al., 2008; Crawford et al., 2018; Lumbreras et al., 2022; Dai et al., 2023; Wan et al., 2023; Su et al., 2025; Nagy et al., 2025; Ambainis et al., 2025). At the same time, new advances in quantum hardware (Arute et al., 2019; Saggio et al., 2021; Acharya et al., 2025) point to a future when quantum computers can become a reality.

In this work, we propose novel quantum algorithms to tackle a central problem related to MDPs, obtaining approximately optimal policies, thus alleviating the curse of dimensionality.

2 PRELIMINARIES

For $n \in \mathbb{N} := \{0, 1, 2, \dots\}$, let $[n] := \{0, \dots, n-1\}$. Given a finite set $\mathcal{S}$, let $\mathscr{B}(\mathcal{S})$ be the space of all bounded Borel measurable real-valued functions on $\mathcal{S}$, which can be interpreted as $\mathbb{R}^S$ for $S = |\mathcal{S}|$, and let $\Delta(\mathcal{S})$ denote the probability simplex over $\mathcal{S}$. Let the ℓ_∞-norm of $u \in \mathscr{B}(\mathcal{S})$ be $\|u\|_\infty = \max_{s\in\mathcal{S}} |u(s)|$. We use $\mathbf{1} \in \mathscr{B}(\mathcal{S})$ to denote the all-ones function $\mathbf{1}(s) = 1$. Given $u, v \in \mathscr{B}(\mathcal{S})$, let $uv \in \mathscr{B}(\mathcal{S})$ be the function $(uv)(s) = u(s)v(s)$, and $u \leq v \iff u(s) \leq v(s)$.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2.1 Quantum Computation

Little background on quantum computation is needed for this paper and we refer to Nielsen & Chuang (2010) for more information. The state of a quantum system is described by a unit vector from a Hilbert space denoted by the ket notation $|\cdot\rangle$. An n-qubit system is described by a unit vector in $\mathbb{C}^{2^n}$. The evolution of a quantum state $|\psi\rangle \in \mathbb{C}^{2^n}$ is described by a unitary operator $U \in \mathbb{C}^{2^n \times 2^n}$, $UU^\dagger = I$ where $U^\dagger$ is the Hermitian conjugate of U. In order to extract classical information from a quantum system, a quantum measurement is performed, which is a set $\{E_m\}_m$ of positive operators $E_m \succ 0$ that sum to identity, $\sum_m E_m = I$. The probability of measuring E_m on $|\psi\rangle$ is $\langle\psi|E_m|\psi\rangle$. We use $|\bar{0}\rangle$ to denote the state $|0\rangle \otimes \cdots \otimes |0\rangle$ where the number of qubits is clear from context.

In this paper, we employ quantum oracles for functions and probability distributions. We say we have quantum access to $u \in \mathscr{B}(\mathcal{S})$ if we have access to the unitary $\mathcal{O}_u : |s\rangle|\bar{0}\rangle \mapsto |s\rangle|u(s)\rangle$ and its inverse, and we say we have quantum sampling access to $p \in \Delta(\mathcal{S})$ if we have access to the unitary $\mathcal{O}_p : |\bar{0}\rangle \mapsto \sum_{s\in\mathcal{S}} \sqrt{p(s)}|s\rangle$ and its inverse. Quantum access to a function is usually referred to as a quantum random access memory (QRAM) (Giovannetti et al., 2008a;b; Jaques & Rattew, 2025; Allcock et al., 2024). It is possible to build quantum access to $u \in \mathscr{B}(\mathcal{S})$ in $\widetilde{O}(S)$ time. We shall require the following well-known quantum subroutines.

Fact 1 (Quantum maximum finding Dürr & Høyer (1996))**.** *Given quantum access to $u \in \mathscr{B}(\mathcal{S})$ via oracle $\mathcal{O}_u$, there is a quantum algorithm that finds $\max_{s\in\mathcal{S}} u(s)$ and $\arg\max_{s\in\mathcal{S}} u(s)$ with probability $1-\delta$ using $O(\sqrt{S}\log\frac{1}{\delta})$ queries to $\mathcal{O}_u, \mathcal{O}_u^\dagger$.*

Fact 2 (Quantum mean estimation with variance (Kothari & O'Donnell, 2023, Theorem 1.1))**.** *Let $\epsilon > 0$ and $\delta \in (0,1)$. Assume quantum access to $u : \mathcal{S} \to \mathbb{R}$ via oracle $\mathcal{O}_u$ and quantum sampling access to $p \in \Delta(\mathcal{S})$ via oracle $\mathcal{O}_p$. Let $\sigma := \sum_{s\in\mathcal{S}} p(s)u(s)^2 - \big(\sum_{s\in\mathcal{S}} p(s)u(s)\big)^2$. There is a quantum algorithm that computes $\widetilde{\mu} \in \mathbb{R}$ such that $|\widetilde{\mu} - \sum_{s\in\mathcal{S}} p(s)u(s)| \le \sqrt{\sigma}\epsilon$ with success probability $1-\delta$ using $O\big(\frac{1}{\epsilon}\log\frac{1}{\delta}\big)$ queries to $\mathcal{O}_u, \mathcal{O}_p$, and their inverses.*

2.2 Background on Markov Decision Processes

A discrete-time MDP models the interaction between an agent and an environment. At each time step $t \in \mathbb{N}$, the environment is in some state s_t from a finite set $\mathcal{S}$ of size S and the agent chooses an action a_t from a finite set $\mathcal{A}$ of size A according to a *randomised decision rule* $\pi_t : \mathcal{S} \to \Delta(\mathcal{A})$ so that $a_t \sim \pi_t(\cdot|s_t)$. If $\pi_t : \mathcal{S} \to \mathcal{A}$, we say π_t is *deterministic*. After choosing action a_t, the agent receives a reward $r_t(s_t, a_t)$ given by the measurable bounded reward function $r_t : \mathcal{S}\times\mathcal{A} \to [0,1]$ and the environment transitions to a new state $s_{t+1} \sim p_t(\cdot|s_t, a_t)$ according to the stochastic kernel $p_t : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$. A *randomised policy* $\pi = (\pi_t)_t$ is a sequence of decision rules $\pi_t : \mathcal{S} \to \Delta(\mathcal{A})$, and similarly for a *deterministic policy*. A policy is *stationary* if $\pi_t = \pi_{t'}$ for all t, t'. Let $\mathcal{D}^{\mathrm{R}}$ and $\mathcal{D}^{\mathrm{D}}$ be the set of all randomised and deterministic decision rules and Π^{R} and Π^{D} the set of all randomised and deterministic policies, respectively. Given $d \in \mathcal{D}^{\mathrm{R}}$, define $r_t^d \in \mathscr{B}(\mathcal{S})$ and $p_t^d : \mathcal{S} \to \Delta(\mathcal{S})$ as

$$r_t^d(s) := \mathop{\mathbb{E}}_{a\sim d(\cdot|s)}[r_t(s,a)] \qquad \text{and} \qquad p_t^d(s'|s) := \mathop{\mathbb{E}}_{a\sim d(\cdot|s)}[p_t(s'|s,a)].$$

Given stochastic kernel $p_t : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, time step $t \in \mathbb{N}$, decision rule $d \in \mathcal{D}^{\mathrm{R}}$, and action $a \in \mathcal{A}$, let the operators $\mathcal{P}_t^d, \mathcal{P}_t^a : \mathscr{B}(\mathcal{S}) \to \mathscr{B}(\mathcal{S})$ be defined as $(\mathcal{P}_t^d u)(s) = \sum_{s'\in\mathcal{S}} p_t^d(s'|s)u(s')$ and $(\mathcal{P}_t^a u)(s) = \sum_{s'\in\mathcal{S}} p_t(s'|s,a)u(s')$ for all $u \in \mathscr{B}(\mathcal{S})$ and $s \in \mathcal{S}$.

Finite-horizon MDPs. Central to this work are *finite-horizon* MDPs where the agent interacts with the environment for a pre-determined number of steps $H \in \mathbb{N}$ called *horizon*. A finite-horizon MDP is described by the tuple $\langle \mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, H\rangle$. Given policy $\pi = (\pi_h)_{h\in[H]}$ and $h \in [H]$, let the *value function* $V_h^{\pi,H} : \mathcal{S} \to [0,H]$ be defined as

$$V_h^{\pi,H}(s) = \mathbb{E}\left[\sum_{h'=h}^{H-1} r_{h'}(s_{h'}, a_{h'}) \;\middle|\; s_h = s, a_{h'} \sim \pi_{h'}(s_{h'}), s_{h'+1} \sim p_{h'}(\cdot|s_{h'}, a_{h'})\right] \qquad \forall s \in \mathcal{S}.$$

Note that not all decision rules that make up π are necessarily employed in $V_h^{\pi,H}$. A policy $\pi_\varepsilon \in \Pi^{\mathrm{R}}$ is *ε-optimal* for $\varepsilon \ge 0$ if $V_0^{\pi_\varepsilon,H} \ge V_0^{\pi,H} - \varepsilon\mathbf{1}$ for all $\pi \in \Pi^{\mathrm{R}}$, and a policy π^* is *optimal* if

$V_0^{\pi^*,H} \geq V_0^{\pi,H}$ for all $\pi \in \Pi^{\mathrm{R}}$. The *optimal value function* is $V_h^{*,H} := \sup_{\pi\in\Pi^{\mathrm{R}}} V_h^{\pi,H}$. It is well known that $V_0^{*,H} = \sup_{\pi\in\Pi^{\mathrm{R}}} V_0^{\pi,H} = \sup_{\pi\in\Pi^{\mathrm{D}}} V_0^{\pi,H}$, i.e., the optimal policy is deterministic.

Let the *h-step Bellman operator* $\mathcal{L}_h^d : \mathscr{B}(\mathcal{S}) \to \mathscr{B}(\mathcal{S})$ associated with $d \in \mathcal{D}^{\mathrm{R}}$ be $\mathcal{L}_h^d u = r_h^d + \mathcal{P}_h^d u$ and the *h-step optimal Bellman operator* $\mathcal{L}_h : \mathscr{B}(\mathcal{S}) \to \mathscr{B}(\mathcal{S})$ be $\mathcal{L}_h u = \sup_{d\in\mathcal{D}^{\mathrm{R}}} \mathcal{L}_h^d u$. The Bellman operators $\mathcal{L}_h$ and $\mathcal{L}_h^d$ are monotonic, i.e., $u \leq v \implies \mathcal{L}_h u \leq \mathcal{L}_h v$, and non-expansive, i.e., $\|\mathcal{L}_h u - \mathcal{L}_h v\|_\infty \leq \|u - v\|_\infty$ for all $u, v \in \mathscr{B}(\mathcal{S})$ (similarly for $\mathcal{L}_h^d$). Finally, the value functions $V_h^{\pi,H}$ satisfy the recursive equation $V_h^{\pi,H} = \mathcal{L}_h^{\pi_h} V_{h+1}^{\pi,H} = r_h^{\pi_h} + \mathcal{P}_h^{\pi_h} V_{h+1}^{\pi,H}$.

Infinite-horizon discounted MDPs. Another important class of MDPs are *infinite-horizon discounted MDPs* where the agent-environment interaction lasts an infinite amount of time but at every time step the received reward is decreased by a *discount factor* $\gamma \in [0, 1)$. The parameter $\Gamma := (1-\gamma)^{-1}$ is usually called *effective horizon*. Here we shall focus on the time-independent setting where the reward functions and stochastic kernels are the same for all $t \in \mathbb{N}$, $r_t = r$ and $p_t = p$. A time-independent infinite-horizon discounted MDP is described by the tuple $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$. Given policy $\pi = (\pi_t)_{t\in\mathbb{N}}$ and $t \in \mathbb{N}$, let the *value function* $V_t^{\pi,\gamma} : \mathcal{S} \to [0, \Gamma]$ be defined as

$$V_t^{\pi,\gamma}(s) = \mathbb{E}\left[\sum_{t'=0}^{t} \gamma^{t'} r(s_{t'}, a_{t'}) \,\middle|\, s_0 = s, a_{t'} \sim \pi_{t-t'}(s_{t'}), s_{t'+1} \sim p(\cdot|s_{t'}, a_{t'})\right] \qquad \forall s \in \mathcal{S}.$$

Note that not all decision rules that make up π are necessarily employed in $V_t^{\pi,\gamma}$. We observe the unusual ordering of the decision rules $\pi_{t-t'}$ employed in choosing the t'-th action $a_{t'}$ instead of the more natural choice $a_{t'} \sim \pi_{t'}(s_{t'})$ (the reason is explained below). Define also the value function

$$V_\infty^{\pi,\gamma}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s, a_t \sim \pi_t(s_t), s_{t+1} \sim p(\cdot|s_t, a_t)\right] \qquad \forall s \in \mathcal{S}.$$

Here the decision rule ordering is the conventional one, so $a_t \sim \pi_t(s_t)$. A policy $\pi_\varepsilon \in \Pi^{\mathrm{R}}$ is *ε-optimal* for $\varepsilon \geq 0$ if $V_\infty^{\pi_\varepsilon,\gamma} \geq V_\infty^{\pi,\gamma} - \varepsilon\mathbf{1}$ for all $\pi \in \Pi^{\mathrm{R}}$, and a policy $\pi^* \in \Pi^{\mathrm{R}}$ is *optimal* if $V_\infty^{\pi^*,\gamma} \geq V_\infty^{\pi,\gamma}$ for all $\pi \in \Pi^{\mathrm{R}}$. The *optimal value function* is $V_\infty^{*,\gamma} := \sup_{\Pi^{\mathrm{R}}} V_\infty^{\pi,\gamma}$. It is well known that $V_\infty^{*,\gamma} = \sup_{\pi\in\Pi^{\mathrm{R}}} V_\infty^{\pi,\gamma} = \sup_{d\in\mathcal{D}^{\mathrm{D}}} V_\infty^{d^\infty,\gamma}$, i.e., the optimal policy is deterministic and stationary (here the notation d^∞ for $d \in \mathcal{D}^{\mathrm{R}}$ means the stationary policy $(d, d, \dots)$).

Let the *Bellman operator* $\mathcal{L}^d : \mathscr{B}(\mathcal{S}) \to \mathscr{B}(\mathcal{S})$ associated with $d \in \mathcal{D}^{\mathrm{R}}$ and *optimal Bellman operator* $\mathcal{L} : \mathscr{B}(\mathcal{S}) \to \mathscr{B}(\mathcal{S})$ be defined as $\mathcal{L}^d u = r^d + \gamma \mathcal{P}^d u$ and $\mathcal{L} u = \sup_{d\in\mathcal{D}^{\mathrm{R}}} \mathcal{L}^d u$. The Bellman operators $\mathcal{L}$ and $\mathcal{L}^d$ are monotonic i.e., $u \leq v \implies \mathcal{L}u \leq \mathcal{L}v$, and contractive, i.e., $\|\mathcal{L}u - \mathcal{L}v\|_\infty \leq \gamma\|u - v\|_\infty$ for all $u, v \in \mathscr{B}(\mathcal{S})$ (similarly for $\mathcal{L}^d$). Given our deliberate choice for ordering the decision rules $\pi_{t-t'}$ in the definition of $V_t^{\pi,\gamma}$, the value functions $V_t^{\pi,\gamma}$ satisfy the recursive equation $V_{t+1}^{\pi,\gamma} = \mathcal{L}^{\pi_{t+1}} V_t^{\pi,\gamma} = r^{\pi_{t+1}} + \gamma \mathcal{P}^{\pi_{t+1}} V_t^{\pi,\gamma}$, similarly to the finite-horizon setting. Finally, given $d \in \mathcal{D}^{\mathrm{R}}$, the value function $V_\infty^{d^\infty,\gamma}$ is the unique fixed point of the Bellman operator $\mathcal{L}^d$, i.e., $\mathcal{L}^d V_\infty^{d^\infty,\gamma} = V_\infty^{d^\infty,\gamma}$, while the optimal value function $V_\infty^{*,\gamma}$ is the unique fixed point of the Bellman operator $\mathcal{L}$, i.e., $\mathcal{L} V_\infty^{*,\gamma} = V_\infty^{*,\gamma}$. As a consequence, for any $d \in \mathcal{D}^{\mathrm{R}}$ and $u \in \mathscr{B}(\mathcal{S})$, $(\mathcal{L}^d)^\infty u = V_\infty^{d^\infty,\gamma}$ and $(\mathcal{L})^\infty u = V_\infty^{*,\gamma}$.

3 Problem Setup: Approximating Optimal Policies

One of the fundamental theoretical problems related to MDPs is that of learning an ε-optimal policy. In the finite-horizon setting, this means obtaining $\pi_\varepsilon \in \Pi^{\mathrm{D}}$ such that $V_0^{\pi_\varepsilon,H} \geq V_0^{*,H} - \varepsilon\mathbf{1}$, while in the infinite-horizon discounted setting, obtaining $d_\varepsilon \in \mathcal{D}^{\mathrm{D}}$ such that $V_\infty^{d_\varepsilon^\infty,\gamma} \geq V_\infty^{*,\gamma} - \varepsilon\mathbf{1}$. Among the several studied input models, arguably the most important — and the one considered in this work — is the so-called *generative model* (Kearns & Singh, 1998b; Kearns et al., 2002; Kakade, 2003) where one has full knowledge of state and action spaces and of the reward functions, but the transition probabilities $p_t(\cdot|s, a)$ can only be accessed through an oracle.

In the classical setting using standard classical computers, generative access to the stochastic kernel $p_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is given via an oracle or simulator $\mathcal{C}_{p_t}$ that, on input $(s, a) \in \mathcal{S} \times \mathcal{A}$, returns $s' \in \mathcal{S}$ with probability $p_t(s'|s, a)$. One is thus interested in computing an ε-optimal policy with the least amount of queries to oracles $\{\mathcal{C}_{p_t}\}_t$ (there is only one oracle for time-independent MDPs).

The quantum equivalent of generative access to p_t is done through an oracle $\mathcal{Q}_{p_t}$ (and its inverse $\mathcal{Q}_{p_t}^\dagger$) called quantum accessible-environment (Wang et al., 2021; Wiedemann et al., 2023; Jerbi et al., 2023; Zhong et al., 2024) which is the unitary operator

$$\mathcal{Q}_{p_t} : |s\rangle|a\rangle|\bar{0}\rangle \to \sum_{s' \in \mathcal{S}} \sqrt{p_t(s'|s,a)}|s\rangle|a\rangle|s'\rangle \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

The quantum oracle $\mathcal{Q}_{p_t}$ puts in superposition all the possible elements $s' \in \mathcal{S}$ with the corresponding weight $\sqrt{p_t(s'|s,a)}$ and is a proper quantum generalisation of $\mathcal{C}_{p_t}$ by noticing that a measurement on the third register $|s'\rangle$ returns an element $s' \in \mathcal{S}$ with probability $p_t(s'|s,a)$, just like in the classical setting. Moreover, if the classical oracle $\mathcal{C}_{p_t}$ is itself a computer program and we have access to its source code, then it is straightforward to transform $\mathcal{C}_{p_t}$ into its quantum version $\mathcal{Q}_{p_t}$ above: one first produces a Boolean circuit (of roughly the same size as the program's time complexity) that produces the samples $s' \sim p_t(\cdot|s,a)$ and then replaces all classical gates with their quantum counterparts as described, for example, in (Nielsen & Chuang, 2010, Section 1.4.1).

3.1 Previous Works

Classically computing ε-optimal policies under the generative model has been mostly solved for finite- and infinite-horizon discounted MDPs. A long list of works slowly improved the query complexity (Kearns & Singh, 1998a; Kearns et al., 2002; Azar et al., 2013; Wang, 2017; Sidford et al., 2018b) until Sidford et al. (2018a) obtained sample-optimal algorithms for *time-independent* finite- and infinite-horizon discounted MDPs with query complexity $\widetilde{O}\big(\frac{H^3SA}{\varepsilon^2}\big)$ and $\widetilde{O}\big(\frac{\Gamma^3SA}{\varepsilon^2}\big)$, respectively, thus matching the lower bounds from Azar et al. (2013); Sidford et al. (2018a). Later Li et al. (2020) obtained the query upper bound $\widetilde{O}\big(\frac{H^4SA}{\varepsilon^2}\big)$ for time-dependent finite-horizon MDPs.

The algorithms of Sidford et al. (2018a) are a modern and improved version of the standard value iteration algorithm (Puterman, 2014) (explained in Section 4) and work (e.g., for infinite-horizon discounted MDPs) by starting with a function V_0 such that $0 \leq V_\infty^{*,\gamma}(s) - V_0(s) \leq 2\epsilon$ and then producing function V_T such that $0 \leq V_\infty^{*,\gamma}(s) - V_T(s) \leq \epsilon$, thus halving the initial error. The initial choice $V_0(s) = 0$ yields a Γ-approximation and by repeating their halving procedure $O\big(\Gamma \log \frac{\Gamma}{\varepsilon}\big)$ times, an ε-optimal policy can be obtained.

In the quantum setting where one has access to the oracle $\mathcal{Q}_p$ and its inverse, Wang et al. (2021) were the first to tackle the problem of approximating optimal policies in the infinite-horizon discounted setting and obtained the query upper bound $\widetilde{O}\big(\min\big\{\frac{\Gamma^{1.5}SA}{\varepsilon}, \frac{\Gamma^3S\sqrt{A}}{\varepsilon}\big\}\big)$ and the query lower bound $\widetilde{\Omega}\big(\frac{\Gamma^{1.5}S\sqrt{A}}{\varepsilon}\big)$. The (time-dependent) finite-horizon setting, on the other hand, has been more recently addressed by Luo et al. (2025); Ambainis et al. (2025), both works proposing query upper bounds of $\widetilde{O}\big(\min\big\{\frac{H^{2.5}SA}{\varepsilon}, \frac{H^3S\sqrt{A}}{\varepsilon}\big\}\big)$. Luo et al. (2025) further proved a query lower bound of $\widetilde{\Omega}\big(\frac{H^{1.5}S\sqrt{A}}{\varepsilon}\big)$.

The quantum algorithms of Wang et al. (2021); Luo et al. (2025); Ambainis et al. (2025) are the combination of two different algorithms. The first one is a quantised version of the standard value iteration (Puterman, 2014). The quantum advantage comes from estimating quantities like $\sum_{s' \in \mathcal{S}} p(s'|s,a)V_t(s')$ with improved precision using a quantum mean estimation subroutine (Montanaro, 2015; Hamoudi & Magniez, 2019; Hamoudi, 2021b;a; Kothari & O'Donnell, 2023) and nesting it into the quantum maximum finding subroutine of Dürr & Høyer (1996) in order to find a maximum over $a \in \mathcal{A}$ quadratically faster.

The second algorithm of Wang et al. (2021); Luo et al. (2025); Ambainis et al. (2025) is a quantised version of the modern value iteration algorithm of Sidford et al. (2018a) mentioned above. The quantum advantage again comes from estimating mean quantities with improved precision using quantum mean estimation subroutines, which leads to a better H or Γ dependence. However, due to the complexity of Sidford et al.'s algorithm, it was not possible to employ quantum maximum finding in order to further improve the complexity on A. This is partially due to the fact that in Sidford et al.'s algorithm, several average quantities are estimated at the beginning of a halving step. Pre-processing all these quantities in the quantum setting already costs linearly in A, hindering the quadratic advantage on the action space size. The combination of both quantised algorithms just highlighted above is the reason why the query upper bounds of Wang et al. (2021); Luo et al. (2025); Ambainis et al. (2025) are a minimum of two different quantities, one with better H, Γ complexity and the other with better A complexity.

Table 1: Classical and quantum query bounds for computing an ε-optimal policy and value function for finite- and infinite-horizon discounted MDPs in the generative model setting. Here S and A are the size of the state and action spaces, respectively, H is the horizon and $\Gamma = (1-\gamma)^{-1}$ the effective horizon. All bounds are up to $\operatorname{poly}\log$ factors and assume a constant failure probability δ.

Setting	Classical query complexity		Quantum query complexity		
	Upper bound	Lower bound	Upper bound		Lower bound
Time-dependent finite-horizon	$\frac{H^3SA}{\varepsilon^2}$ (time-independent) $\frac{H^4SA}{\varepsilon^2}$ (time-dependent) Sidford et al. (2018a) Li et al. (2020)	$\frac{H^3SA}{\varepsilon^2}$ Sidford et al. (2018a)	$\min\left\{\frac{H^{2.5}SA}{\varepsilon}, \frac{H^3S\sqrt{A}}{\varepsilon}\right\}$ Ambainis et al. (2025) Luo et al. (2025)	$\frac{H^{2.5}S\sqrt{A}}{\varepsilon}$ Result 1	$\frac{H^{1.5}S\sqrt{A}}{\varepsilon}$ Luo et al. (2025)
Time-independent infinite-horizon	$\frac{\Gamma^3SA}{\varepsilon^2}$ Sidford et al. (2018a) Li et al. (2020)	$\frac{\Gamma^3SA}{\varepsilon^2}$ Azar et al. (2013)	$\min\left\{\frac{\Gamma^{1.5}SA}{\varepsilon}, \frac{\Gamma^3S\sqrt{A}}{\varepsilon}\right\}$ Wang et al. (2021)	$\frac{\Gamma^{2.5}S\sqrt{A}}{\varepsilon}$ Result 2	$\frac{\Gamma^{1.5}S\sqrt{A}}{\varepsilon}$ Wang et al. (2021)

4 Our Results

As just explained and summarised in Table 1, there is still a considerable gap between lower and upper quantum query complexities for the problem of approximating optimal policies. Interesting enough, the quantum algorithms of Wang et al. (2021); Luo et al. (2025); Ambainis et al. (2025) can only provide an improvement either in H, Γ or in A compared to their classical counterparts, but never jointly in both parameters. As our main contribution, we propose new quantum algorithms that improve upon the previous query upper bounds for finite- and infinite-horizon discounted MDPs.

Result 1. *Let $\langle \mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, H\rangle$ be a time-dependent finite-horizon MDP. There is a quantum algorithm that outputs an ε-optimal policy w.h.p. and query complexity $\widetilde{O}\big(\frac{H^{2.5}S\sqrt{A}}{\varepsilon}\big)$.*

Result 2. *Let $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma\rangle$ be a time-independent infinite-horizon discounted MDP. There is a quantum algorithm that outputs an ε-optimal policy w.h.p. and query complexity $\widetilde{O}\big(\frac{\Gamma^{2.5}S\sqrt{A}}{\varepsilon}\big)$.*

Our new query upper bounds provide a joint quantum advantage in both H, Γ and A parameters, thus breaking the barrier of previous algorithms (Wang et al., 2021; Luo et al., 2025; Ambainis et al., 2025) and approaching the corresponding query lower bounds. Our quantum algorithms employ the quantum mean estimation subroutine *with variance* from Kothari & O'Donnell (2023) (Fact 2) within the quantum maximum finding subroutine from Dürr & Høyer (1996) (Fact 1), plus techniques like monotonicity and total-variance from Sidford et al. (2018a). In contrast to previous quantum works that either (i) use quantum maximum finding but does not include any variance in the error propagation or (ii) take variance into account but not quantum maximum finding, here we perform both, which ultimately yields a combined dependence on $\sqrt{A}$ and a better dependence than Γ^3 or H^3. Our finite-horizon complexity thus subsumes the ones of Luo et al. (2025); Ambainis et al. (2025), while our infinite-horizon complexity improves the part $\widetilde{O}\big(\frac{\Gamma^3S\sqrt{A}}{\varepsilon}\big)$ from Wang et al. (2021), leading to a new overall complexity of $\widetilde{O}\big(\min\big\{\frac{\Gamma^{1.5}SA}{\varepsilon}, \frac{\Gamma^{2.5}S\sqrt{A}}{\varepsilon}\big\}\big)$. Our results are thus a measurable improvement on past works in tackling the curse of dimensionality of MDPs.

We now give a high-level summary of our quantum algorithms. The formal proof of Result 1 is given in Section 5, while the proof of Result 2 is postponed to Section A.

4.1 Review of Techniques

It is known (Puterman, 2014) that a solution to the equations below provides optimal value functions:

$$V_H \equiv 0 \quad \text{and} \quad V_h = \mathcal{L}_h V_{h+1} \quad \forall h \in [H] \quad \text{(finite-horizon)},$$
$$V_0 \equiv 0 \quad \text{and} \quad V_{t+1} = \mathcal{L} V_t \quad \forall t \in \mathbb{N} \quad \text{(infinite-horizon)}.$$

More precisely, for finite-horizon MDPs, $V_h = V_h^{*,H}$ for $h \in [H]$ and $\pi = (\pi_h)_{h\in[H]}$ with $\pi_h \in \arg\max_{d\in\mathcal{D}^{\mathrm{D}}}\{\mathcal{L}_h^d V_{h+1}\}$ is an optimal policy. For infinite-horizon discounted MDPs, the solution V_t converges to the optimal value, i.e., $\lim_{t\to\infty} V_t = V_\infty^{*,\gamma}$. More usefully, for $t \in \mathbb{N}$ such that $\|V_{t+1} - V_t\|_\infty \le \frac{\varepsilon}{2\gamma\Gamma}$, one have that $\|V_{t+1} - V_\infty^{*,\gamma}\|_\infty \le \frac{\varepsilon}{2}$ and d_ε^∞ with $d_\varepsilon \in \arg\max_{d\in\mathcal{D}^{\mathrm{D}}}\{\mathcal{L}^d V_{t+1}\}$ is an ε-optimal policy (Puterman, 2014, Theorem 6.3.1).

Given sampling access to p, the standard value iteration algorithm (Puterman, 2014) obtains an ε-optimal value function and policy for finite-horizon MDPs by approximating $(\mathcal{P}_h^a V_{h+1})(s) = \sum_{s' \in \mathcal{S}} p_h(s'|s,a) V_{h+1}(s')$ up to $\frac{\varepsilon}{H}$ precision for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H-1]$, for a total classical query complexity of $O\big(\frac{H^5 SA}{\varepsilon^2}\big)$. For infinite-horizon discounted MDPs, the standard value iteration algorithm iteratively computes V_t until $\|V_{t+1} - V_t\|_\infty \leq \frac{\varepsilon}{2\gamma\Gamma}$. By the contractive property of $\mathcal{L}$, one must solve $O\big(\Gamma \log \frac{\Gamma}{\varepsilon}\big)$ iteration steps. Therefore, approximating the quantities $(\mathcal{P}^a V_t)(s)$ up to additive error $\frac{\varepsilon}{2\Gamma}$ leads to an ε-optimal value function and $\Gamma\varepsilon$-optimal greedy policy (in the worst case) after $\widetilde{O}(\Gamma)$ iterations. The final classical query complexity is $\widetilde{O}\big(\frac{\Gamma^7 SA}{\varepsilon}\big)$.

In order to achieve the optimal query complexities $\widetilde{O}\big(\frac{H^3 SA}{\varepsilon^2}\big)$ and $\widetilde{O}\big(\frac{\Gamma^3 SA}{\varepsilon^2}\big)$, Sidford et al. (2018a) employs three crucial techniques: *monotonicity*, *variance reduction*, and *total-variance* techniques. The monotonicity technique means choosing an approximate value V_t and decision rule π_t such that the monotonicity condition $V_t \leq \mathcal{L}^{\pi_t} V_t$ is maintained throughout all time steps t of the algorithm.[1] This is vital to guarantee that $V_t(s) \leq V_\infty^{\pi_t^\infty,\gamma}(s)$ and to obtain an ε-optimal policy from an ε-optimal function, otherwise an ε-optimal function only yields an $\Gamma\varepsilon$-optimal greedy policy in the worst case (Bertsekas, 2022). The total-variance technique, on the other hand, is based on the fact that the true error accumulates as $\sqrt{\Gamma^3/m}$ given m samples, much less than the naive sum of estimation errors at each time step. This means that one does not require an update error of $\frac{\epsilon}{2\Gamma}$ at each time step and $m = O\big(\frac{\Gamma^3}{\epsilon^2}\big)$ samples suffices for a total error ϵ, thus shaving off a factor of Γ. The precise statements for total-variance are given in Fact 3 (finite-horizon) and Lemma 1 (infinite-horizon). Finally, we omit a description of the variance reduction technique since it will not be used in our quantum algorithms (variance reduction is actually what hinders the use of quantum maximum finding in Wan et al. (2023); Luo et al. (2025); Ambainis et al. (2025)).

Our new algorithms are a quantised version of the standard value iteration but further improved with the monotonicity and total-variance techniques. For simplicity, we review the infinite-horizon setting here. The finite-horizon case is similar, with Γ replaced with H. At each time step t, we employ a quantum mean estimation subroutine to compute an approximation $\widetilde{\mu}_t(s,a) \approx (\mathcal{P}^a V_t)(s)$. Since we want to leverage the total-variance technique, we use the subroutine of Kothari & O'Donnell (2023) (Fact 2) to approximate $(\mathcal{P}^a V_t)(s)$ up to error $O\big(\frac{\varepsilon\sqrt{\sigma_t(s,a)}}{\Gamma^{1.5}}\big)$ using its variance $\sigma_t(s,a) = (\mathcal{P}^a V_t^2)(s) - (\mathcal{P}^a V_t)^2(s)$. We note that no prior knowledge of $\sigma_t(s,a)$ is required for the quantum subroutine. However, since we ultimately need an approximation with one-side error, we also approximate $\widetilde{\sigma}_t(s,a) \approx \sigma_t(s,a)$ separately up to error $O(\varepsilon\sqrt{\Gamma})$.

What is crucial to understand is that all the aforementioned approximations and calls to quantum mean estimation are done in *superposition* over different values of $a \in \mathcal{A}$. This means that we treat these calculations as forming an oracle (for fixed $s \in \mathcal{S}$)

$$\mathcal{O}_s^{\text{mean}} : |a\rangle|\bar{0}\rangle \mapsto |a\rangle\big(\sqrt{1-\delta_a}|\widehat{\mu}_t(s,a)\rangle|\widetilde{\mu}_t(s,a)\rangle|\widetilde{\sigma}_t(s,a)\rangle|\text{garbage}(a)\rangle + \sqrt{\delta_a}|\perp_a\rangle\big),$$

which will serve as the base unitary for quantum maximum finding (see Fact 1). Here $\delta_a \in [0,1)$ is the failure probability of quantum mean estimation within the branch of the wave function described by $|a\rangle$, $\widetilde{\mu}_t(s,a)$ and $\widetilde{\sigma}_t(s,a)$ are good enough approximations of $\mu_t(s,a)$ and $\sigma_t(s,a)$, $|\text{garbage}(a)\rangle$ are "garbage" unit vectors, and $|\perp_a\rangle$ is a unit vector orthogonal to $|\widetilde{\mu}_t(s,a)\rangle|\widetilde{\sigma}_t(s,a)\rangle|\text{garbage}(a)\rangle$. The quantity $\widehat{\mu}_t(s,a)$ is a one-sided-error version of $\widetilde{\mu}_t(s,a)$, so that $\widehat{\mu}_t(s,a) \leq (\mathcal{P}^a V_t)(s)$, necessary to guarantee the approximate optimality of the derived greedy policies and the reason why we estimate $\sigma_t(s,a)$. The oracle $\mathcal{O}_s^{\text{mean}}$ can be constructed using $O\big(\frac{\Gamma^{1.5}}{\varepsilon}\big)$ calls to $\mathcal{Q}_p$ and its inverse.

By employing quantum maximum finding with $\widetilde{O}(\sqrt{A})$ calls to the oracle $\mathcal{O}_s^{\text{mean}}$, one then obtains $\max_{a\in\mathcal{A}}\{r(s,a) + \gamma\widehat{\mu}_t(s,a)\}$ with high probability, which is an one-side approximation of $V_{t+1}(s)$ with additive error $O\big(\frac{\varepsilon\sqrt{\sigma_t(s,\pi_{t+1}(s))}}{\Gamma^{1.5}}\big)$. At this point, the monotonicity condition must be enforced by picking $V_{t+1}(s) \leftarrow \max\{V_{t+1}(s), V_t(s)\}$, i.e., the maximum between the current and previous iteration. After $T = O\big(\Gamma \log \frac{\Gamma}{\varepsilon}\big)$ time steps, the total-variance technique guarantees that the error does not accumulate too rapidly and that function V_T is ε-close to $V_\infty^{*,\gamma}$. By the monotonicity condition, its associated greedy policy is ε-optimal. The final query complexity across all $s \in \mathcal{S}$ and time steps $T = \widetilde{O}(\Gamma)$ is $\widetilde{O}\big(\frac{\Gamma^{2.5} S\sqrt{A}}{\varepsilon}\big)$.

[1] Finite-horizon MDPs do not require monotonicity since the interaction lasts for a finite number of steps.

Algorithm 1 Quantum algorithm for optimal policies for finite-horizon MDPs

Require: Finite state space $\mathcal{S}$ and action space $\mathcal{A}$, horizon H, quantum sampling access to $\{p_h\}_{h\in[H]}$, quantum access to $\{r_h\}_{h\in[H]}$, failure probability $\delta \in (0,1)$, error $\varepsilon \in (0, \sqrt{H}]$.

Ensure: ε-optimal deterministic policy π.

$\theta \leftarrow \frac{\varepsilon}{7H^{3/2}}$

$V_{H-1}(s) \leftarrow \max_{a\in\mathcal{A}} r_{H-1}(s,a)$ and $\pi_{H-1}(s) \leftarrow \arg\max_{a\in\mathcal{A}} r_{H-1}(s,a)$ with probability $1-\frac{\delta}{HS}$ $\forall s \in \mathcal{S}$ (Fact 1)

for $h = H-2, H-3, \ldots, 0$ **do**

  Let $\mu_{h+1}(s,a) = \sum_{s'\in\mathcal{S}} p_h(s'|s,a) V_{h+1}(s')$

  Let $\sigma_{h+1}(s,a) = \sum_{s'\in\mathcal{S}} p_h(s'|s,a) V_{h+1}(s')^2 - \left(\sum_{s'\in\mathcal{S}} p_h(s'|s,a) V_{h+1}(s')\right)^2$

  for $s \in \mathcal{S}$ **do**

    Use Fact 2 to obtain a unitary

$$\mathcal{U}_s^{(h)}\colon |a\rangle|\bar{0}\rangle \mapsto |a\rangle\left(\sqrt{1-\delta_a}|\widehat{\mu}_{h+1}(s,a)\rangle|\widetilde{\mu}_{h+1}(s,a)\rangle|\widetilde{\sigma}_{h+1}(s,a)\rangle|\mathrm{garbage}(a)\rangle + \sqrt{\delta_a}|\perp_a\rangle\right)$$

with $\max_{a\in\mathcal{A}} \delta_a = \widetilde{O}\left(\frac{\delta^2}{H^2S^2A}\right)$ and

$$|\widetilde{\mu}_{h+1}(s,a) - \mu_{h+1}(s,a)| \le \theta\sqrt{\sigma_{h+1}(s,a)},$$
$$|\widetilde{\sigma}_{h+1}(s,a) - \sigma_{h+1}(s,a)| \le \theta H^2,$$
$$\widehat{\mu}_{h+1}(s,a) := \widetilde{\mu}_{h+1}(s,a) - \theta\sqrt{\widetilde{\sigma}_{h+1}(s,a)} - \theta^{3/2}H.$$

    Use quantum maximum finding with unitary $\mathcal{U}_s^{(h)}$ (Fact 1 with $\mathcal{U}_s^{(h)}$) to obtain $V_h(s)$ and $\pi_h(s)$ such that, with probability $1-\frac{\delta}{HS}$,

$$V_h(s) = \max_{a\in\mathcal{A}}\{r_h(s,a) + \widehat{\mu}_{h+1}(s,a)\} \text{ and } \pi_h(s) = \arg\max_{a\in\mathcal{A}}\{r_h(s,a) + \widehat{\mu}_{h+1}(s,a)\}$$

  end for

end for

return $\pi = (\pi_0, \ldots, \pi_{H-1})$

5 Optimal Policies for Finite-Horizon MDPs

In this section, we formally describe and analyse our quantum algorithm (Algorithm 1) for computing ε-optimal policies for finite-horizon MDPs. We shall need the following technical result (total-variance technique).

Fact 3 ((Sidford et al., 2018a, Lemma F.4)). *Given policy $\pi = (\pi_h)_{h\in[H]}$, let $\sigma_{h+1}^{\pi} \in \mathscr{B}(\mathcal{S})$ be $\sigma_{h+1}^{\pi} = \mathcal{P}_h^{\pi_h}(V_{h+1}^{\pi,H})^2 - (\mathcal{P}_h^{\pi_h} V_{h+1}^{\pi,H})^2$ for $h \in [H]$. For any policy $\pi = (\pi_h)_{h\in[H]}$ and $h \in [H]$,*[2]

$$\left\| \sum_{h'=h}^{H-1} \left(\prod_{i=h}^{h'-1} \mathcal{P}_i^{\pi_i} \right) \sqrt{\sigma_{h'+1}^{\pi}} \right\|_\infty \le H^{3/2}.$$

Theorem 1. *Let $\langle \mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, H\rangle$ be a time-dependent finite-horizon MDP. Let $\delta \in (0,1)$ and $\varepsilon \in (0, \sqrt{H}]$. Algorithm 1 computes functions $\{V_h\}_{h\in[H]} \subset \mathscr{B}(\mathcal{S})$ and deterministic policy $\pi = (\pi_h)_{h\in[H]}$ such that, with probability $1-\delta$,*

$$V_h^{*,H} - \varepsilon\mathbf{1} \le V_h \le V_h^{\pi,H} \le V_h^{*,H} \qquad \forall h \in [H].$$

Its query complexity is (up to poly log log *factors)*

$$\widetilde{O}\left(\frac{H^{2.5}S\sqrt{A}}{\varepsilon} \log\left(\frac{HSA}{\delta}\right) \log\left(\frac{HS}{\delta}\right)\right).$$

[2] We note the typo in (Sidford et al., 2018a, Lemma F.4) where $\|\cdot\|_\infty^2$ should be $\|\cdot\|_\infty$.

Proof. The proof is by induction on $h \in [H]$, with the base case $h = H$ being trivial since $V_H \equiv 0$. Assume then that (the inequality is entry-wise)

$$V_{h'}^{*,H} - \varepsilon \mathbf{1} \leq V_{h'} \leq V_{h'}^{\pi,H} \leq V_{h'}^{*,H} \qquad \forall h' = h+1, h+2, \ldots, H.$$

The main idea is to employ Fact 2 to compute approximations for the true quantities

$$\mu_{h+1}(s,a) := (\mathcal{P}_h^a V_{h+1})(s) \qquad \text{and} \qquad \sigma_{h+1}(s,a) := (\mathcal{P}_h^a V_{h+1}^2)(s) - (\mathcal{P}_h^a V_{h+1})^2(s).$$

Let $\theta := \frac{\varepsilon}{7H^{3/2}}$. We claim that Fact 2 can be adapted to perform all of its steps in superposition without any need for intermediary measurements. This means that we can invoke Fact 2 twice to construct a unitary

$$\mathcal{O}_s^{(h)} : |a\rangle|\bar{0}\rangle \mapsto |a\rangle\big(\sqrt{1-\delta_a}|\widetilde{\mu}_{h+1}(s,a)\rangle|\widetilde{\sigma}_{h+1}(s,a)\rangle|\text{garbage}(a)\rangle + \sqrt{\delta_a}|\perp_a\rangle\big), \tag{1}$$

where $\delta_a \in [0,1)$ is the failure probability of quantum mean estimation within the branch $|a\rangle$ of the wave function, $|\text{garbage}(a)\rangle$ is a "garbage" unit complex vector accumulated through the computation, $|\perp_a\rangle$ is a unit complex vector orthogonal to $|\widetilde{\mu}_{h+1}(s,a)\rangle|\widetilde{\sigma}_{h+1}(s,a)\rangle|\text{garbage}(a)\rangle$, and

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \qquad \begin{aligned} &|\widetilde{\mu}_{h+1}(s,a) - \mu_{h+1}(s,a)| \leq \theta\sqrt{\sigma_{h+1}(s,a)}, \\ &|\widetilde{\sigma}_{h+1}(s,a) - \sigma_{h+1}(s,a)| \leq 2\theta H^2, \end{aligned} \tag{2}$$

using that the standard deviation of V_{h+1}^2 is at most H^2 and thus one can compute approximations of $(\mathcal{P}_a V_{h+1}^2)(s)$ and $(\mathcal{P}_a V_{h+1})^2(s)$ with error at most θH^2 each. Using Equation (2),

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \qquad |\widetilde{\mu}_{h+1}(s,a) - \mu_{h+1}(s,a)| \leq \theta\sqrt{\widetilde{\sigma}_{h+1}(s,a)} + \sqrt{2}\theta^{3/2}H.$$

Define then the quantities

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \qquad \widehat{\mu}_{h+1}(s,a) := \widetilde{\mu}_{h+1}(s,a) - \theta\sqrt{\widetilde{\sigma}_{h+1}(s,a)} - \sqrt{2}\theta^{3/2}H,$$

which have one-side error. We can express the above quantity using the optimal-policy variance $\sigma_{h+1}^*(s,a) := (\mathcal{P}_h^a (V_{h+1}^{*,H})^2)(s) - (\mathcal{P}_h^a V_{h+1}^{*,H})^2(s)$ since

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \qquad \sqrt{\widetilde{\sigma}_{h+1}(s,a)} \leq \sqrt{\sigma_{h+1}(s,a)} + \sqrt{2\theta}H \leq \sqrt{\sigma_{h+1}^*(s,a)} + \varepsilon + \sqrt{2\theta}H,$$

using Equation (2) and that $\sqrt{\text{Var}[V_{h+1}]} \leq \sqrt{\text{Var}[V_{h+1}^{*,H}]} + \sqrt{\text{Var}[V_{h+1}^{*,H} - V_{h+1}]} \leq \sqrt{\text{Var}[V_{h+1}^{*,H}]} + \varepsilon$ if $\|V_{h+1}^{*,H} - V_{h+1}\|_\infty \leq \varepsilon$ according to the induction hypothesis. This means that

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \qquad \begin{aligned} &\widehat{\mu}_{h+1}(s,a) \leq \mu_{h+1}(s,a), \\ &\widehat{\mu}_{h+1}(s,a) \geq \mu_{h+1}(s,a) - 2\theta\sqrt{\sigma_{h+1}^*(s,a)} - 2\theta\varepsilon - 4\sqrt{2}\theta^{3/2}H. \end{aligned}$$

By subtracting the content of register $|\widetilde{\sigma}_{h+1}(s,a)\rangle$ from the register $|\widetilde{\mu}_{h+1}(s,a)\rangle$ in Equation (1), it is then possible to effectively construct a black-box oracle $\mathcal{U}_s^{(h)} : |a\rangle|\bar{0}\rangle \mapsto |a\rangle(\sqrt{1-\delta_a}|\widehat{\mu}_{h+1}(s,a)\rangle + \sqrt{\delta_a}|\perp_a\rangle)$ using oracles $\mathcal{Q}_{p_h}, \mathcal{Q}_{p_h}^\dagger$ (up to garbage states). The maximum over $a \in \mathcal{A}$ of $r_h(s,a) + \widehat{\mu}_{h+1}(s,a)$ can thus be found by using quantum maximum finding (Fact 1) with unitary $\mathcal{U}_s^{(h)}$, leading to $V_h(s) = \max_{a\in\mathcal{A}}\{r_h(s,a) + \widehat{\mu}_{h+1}(s,a)\}$ and $\pi_h(s) = \arg\max_{a\in\mathcal{A}}\{r_h(s,a) + \widehat{\mu}_{h+1}(s,a)\}$ with probability $1 - \frac{\delta}{HS}$. This means that

$$\forall s \in \mathcal{S} : \qquad \begin{aligned} &V_h(s) \leq (\mathcal{L}_h^{\pi_h} V_{h+1})(s), \\ &V_h(s) \geq (\mathcal{L}_h^{\pi_h^*} V_{h+1})(s) - 2\theta\sqrt{\sigma_{h+1}^*(s,\pi_h^*(s))} - 2\theta\varepsilon - 4\sqrt{2}\theta^{3/2}H. \end{aligned}$$

Note that $V_h(s) \leq (\mathcal{L}_h^{\pi_h} V_{h+1})(s) \leq (\mathcal{L}_h^{\pi_h} V_{h+1}^{\pi,H})(s) = V_h^{\pi,H}(s)$ by using the induction hypothesis to argue that $V_{h+1} \leq V_{h+1}^{\pi,H}$. We are left with proving that $V_h^*(s) - \varepsilon \leq V_h(s)$. For such,

$$V_h^{*,H} - V_h \leq \mathcal{L}_h^{\pi_h^*} V_{h+1}^{*,H} - \mathcal{L}_h^{\pi_h^*} V_{h+1} + \xi_h = \mathcal{P}_h^{\pi_h^*}(V_{h+1}^{*,H} - V_{h+1}) + \xi_h,$$

where we defined the function $\xi_h(s) = 2\theta\sqrt{\sigma^*_{h+1}(s,\pi^*_h(s))} + 2\theta\varepsilon + 4\sqrt{2}\theta^{3/2}H$. Solving the above recursion with the boundary condition $V_{H-1} \equiv V^{*,H}_{H-1}$,

$$V^{*,H}_h - V_h \le \sum_{h'=h}^{H-1} \left(\prod_{i=h}^{h'-1} \mathcal{P}_i^{\pi^*_i} \right) \xi_{h'}.$$

Using that (where $\sigma^*_{h+1} = \mathcal{P}_h^{\pi^*_h}(V^{*,H}_{h+1})^2 - (\mathcal{P}_h^{\pi^*_h} V^{*,H}_{h+1})^2 \in \mathscr{B}(\mathcal{S})$)

$$\left\| \sum_{h'=h}^{H-1} \left(\prod_{i=h}^{h'-1} \mathcal{P}_i^{\pi^*_i} \right) \sqrt{\sigma^*_{h'+1}} \right\|_\infty \le H^{3/2}, \qquad \text{(by Fact 3)}$$

and that $\| \sum_{h'=h}^{H} \prod_{i=h}^{h'-1} \mathcal{P}_i^{\pi^*_i} \|_\infty \le H$, we get that

$$V^{*,H}_h(s) - V_h(s) \le 2\theta H^{3/2} + 2\theta H\varepsilon + 4\sqrt{2}\theta^{3/2}H^2 = \varepsilon\left(\frac{2}{7} + \frac{2\varepsilon}{7\sqrt{H}} + \frac{4\sqrt{2}}{7^{3/2}H^{1/4}}\right) < \varepsilon.$$

We now analyse the success probability of Algorithm 1.[3] Ideally, we would like to implement the unitary $\mathcal{U}^{(h)}_{s,\mathrm{ideal}} : |a\rangle|\bar{0}\rangle \mapsto |a\rangle|\widehat{\mu}_{h+1}(s,a)\rangle$. In practice, however, we implement the unitary $\mathcal{U}^{(h)}_s : |a\rangle|\bar{0}\rangle \mapsto |a\rangle(\sqrt{1-\delta_a}|\widehat{\mu}_{h+1}(s,a)\rangle + \sqrt{\delta_a}|\perp_a\rangle)$ (up to garbage states) for $\{\delta_a\}_{a\in\mathcal{A}} \subset [0,1)$, where the register $|\widehat{\mu}_{h+1}(s,a)\rangle$ holds the desired approximation $\widehat{\mu}_{h+1}(s,a)$ and $|\perp_a\rangle$ is a normalised quantum state orthogonal to $|\widehat{\mu}_{h+1}(s,a)\rangle$. Let $\delta_2 := \max_{a\in\mathcal{A}} \delta_a$. Consider a unitary $\mathcal{V}^{(h)}_s$ such that, for all $a \in \mathcal{A}$,

1. $\mathcal{V}^{(h)}_s|a\rangle(\sqrt{1-\delta_a}|\widehat{\mu}_{h+1}(s,a)\rangle + \sqrt{\delta_a}|\perp_a\rangle) = |a\rangle|\widehat{\mu}_{h+1}(s,a)\rangle$;
2. $\mathcal{V}^{(h)}_s|a\rangle(\sqrt{\delta_a}|\widehat{\mu}_{h+1}(s,a)\rangle - \sqrt{1-\delta_a}|\perp_a\rangle) = |a\rangle|\perp_a\rangle$ (note that $\sqrt{\delta_2}|\widehat{\mu}_{h+1}(s,a)\rangle - \sqrt{1-\delta_a}|\perp_a\rangle$ is orthogonal to $\sqrt{1-\delta_a}|\widehat{\mu}_{h+1}(s,a)\rangle + \sqrt{\delta_a}|\perp_a\rangle$);
3. For every $|\phi\rangle$ orthogonal to $\mathrm{span}\{|\widehat{\mu}_{h+1}(s,a)\rangle, |\perp_a\rangle\}$, $\mathcal{V}^{(h)}_s|a\rangle|\phi\rangle = |a\rangle|\phi\rangle$.

Then $\mathcal{V}^{(h)}_s\mathcal{U}^{(h)}_s|a\rangle|\bar{0}\rangle = \mathcal{V}^{(h)}_s|a\rangle(\sqrt{1-\delta_a}|\widehat{\mu}_{h+1}(s,a)\rangle + \sqrt{\delta_a}|\perp_a\rangle) = |a\rangle|\widehat{\mu}_{h+1}(s,a)\rangle$. Therefore, we can just set $\mathcal{U}^{(h)}_{s,\mathrm{ideal}} = \mathcal{V}^{(h)}_s\mathcal{U}^{(h)}_s$ without loss of generality. As a consequence,

$$\|\mathcal{U}^{(h)}_s - \mathcal{U}^{(h)}_{s,\mathrm{ideal}}\| = \|I - \mathcal{V}^{(h)}_s\| \le \sqrt{(1-\sqrt{1-\delta_2})^2 + \delta_2} = \sqrt{2 - 2\sqrt{1-\delta_2}} \le \sqrt{2\delta_2},$$

using that $\sqrt{1-\delta_2} \ge 1 - \delta_2$. Moving on, the success probability of quantum maximum finding (Fact 1) is $1-\delta_1$ when employing $\mathcal{U}^{(h)}_{s,\mathrm{ideal}}$, for some $\delta_1 \in [0,1)$. However, since it employs $\mathcal{U}^{(h)}_s$ instead, the success probability decreases by at most the spectral norm of the difference between the "real" and the "ideal" total unitaries. To be more precise, the "ideal" quantum maximum finding is a sequence of gates $W = U_1E_1U_2E_2\cdots U_NE_N$, where $U_i \in \{\mathcal{U}^{(h)}_{s,\mathrm{ideal}}, \mathcal{U}^{(h)\dagger}_{s,\mathrm{ideal}}\}$, E_i is a circuit of elementary gates, and $N = c\sqrt{A}\log\frac{1}{\delta_1}$ with c constant is the number of queries to $\mathcal{U}^{(h)}_{s,\mathrm{ideal}}$. The "real" implementation, on the other hand, is $\widetilde{W} = \widetilde{U}_1E_1\widetilde{U}_2E_2\cdots\widetilde{U}_NE_N$, where $\widetilde{U}_i \in \{\mathcal{U}^{(h)}_s, \mathcal{U}^{(h)\dagger}_s\}$. Then $\|W - \widetilde{W}\| \le c\sqrt{A}\log\left(\frac{1}{\delta_1}\right)\|\mathcal{U}^{(h)}_{s,\mathrm{ideal}} - \mathcal{U}^{(h)}_s\| \le c\sqrt{2\delta_2 A}\log\frac{1}{\delta_1}$ and the failure probability is $\delta_1 + c\sqrt{2\delta_2 A}\log\frac{1}{\delta_1}$. By taking $\delta_1 = O\left(\frac{\delta}{HS}\right)$ and $\delta_2 = O\left(\frac{\delta_1^2}{A\log^2(1/\delta_1)}\right)$, the failure probability in outputting $\max_{a\in\mathcal{A}}\{r_h(s,a) + \widehat{\mu}_{h+1}(s,a)\}$ is at most $\frac{\delta}{HS}$. By a usual union bound over all $s \in \mathcal{S}$ and $h \in [H]$, the failure probability is at most δ.

Regarding the query complexity, one call to the unitary $\mathcal{U}^{(h)}_s$ in order to compute $\widetilde{\mu}_{h+1}(s,a)$ and $\widetilde{\sigma}_{h+1}(s,a)$ in Equation (1) uses $O\left(\frac{1}{\theta}\log\frac{1}{\delta_2}\right) = \widetilde{O}\left(\frac{H^{3/2}}{\varepsilon}\log\frac{HSA}{\delta}\right)$ queries to $\{\mathcal{Q}_{p_h}, \mathcal{Q}^\dagger_{p_h}\}_{h\in[H]}$, while quantum maximum finding makes $O\left(\sqrt{A}\log\frac{1}{\delta_1}\right) = O\left(\sqrt{A}\log\frac{HS}{\delta}\right)$ queries to $\mathcal{U}^{(h)}_s$. Summing over all $\mathcal{S}\times[H]$, we obtain the stated query complexity. □

[3] See (Chen & de Wolf, 2021, Appendix A) for a similar argument.

AI USE STATEMENT

In this work, we used generative AI tools for checking for typos and mistakes. We have not used generative AI tools for proving the main results and writing the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rajeev Acharya, Dmitry A. Abanin, Laleh Aghababaie-Beni, Igor Aleiner, Trond I. Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Nikita Astrakhantsev, Juan Atalaya, Ryan Babbush, Dave Bacon, Brian Ballard, Joseph C. Bardin, Johannes Bausch, Andreas Bengtsson, Alexander Bilmes, Sam Blackwell, Sergio Boixo, Gina Bortoli, Alexandre Bourassa, Jenna Bovaird, Leon Brill, Michael Broughton, David A. Browne, Brett Buchea, Bob B. Buckley, David A. Buell, Tim Burger, Brian Burkett, Nicholas Bushnell, Anthony Cabrera, Juan Campero, Hung-Shen Chang, Yu Chen, Zijun Chen, Ben Chiaro, Desmond Chik, Charina Chou, Jahan Claes, Agnetta Y. Cleland, Josh Cogan, Roberto Collins, Paul Conner, William Courtney, Alexander L. Crook, Ben Curtin, Sayan Das, Alex Davies, Laura De Lorenzo, Dripto M. Debroy, Sean Demura, Michel Devoret, Agustin Di Paolo, Paul Donohoe, Ilya Drozdov, Andrew Dunsworth, Clint Earle, Thomas Edlich, Alec Eickbusch, Aviv Moshe Elbag, Mahmoud Elzouka, Catherine Erickson, Lara Faoro, Edward Farhi, Vinicius S. Ferreira, Leslie Flores Burgos, Ebrahim Forati, Austin G. Fowler, Brooks Foxen, Suhas Ganjam, Gonzalo Garcia, Robert Gasca, Élie Genois, William Giang, Craig Gidney, Dar Gilboa, Raja Gosula, Alejandro Grajales Dau, Dietrich Graumann, Alex Greene, Jonathan A. Gross, Steve Habegger, John Hall, Michael C. Hamilton, Monica Hansen, Matthew P. Harrigan, Sean D. Harrington, Francisco J. H. Heras, Stephen Heslin, Paula Heu, Oscar Higgott, Gordon Hill, Jeremy Hilton, George Holland, Sabrina Hong, Hsin-Yuan Huang, Ashley Huff, William J. Huggins, Lev B. Ioffe, Sergei V. Isakov, Justin Iveland, Evan Jeffrey, Zhang Jiang, Cody Jones, Stephen Jordan, Chaitali Joshi, Pavol Juhas, Dvir Kafri, Hui Kang, Amir H. Karamlou, Kostyantyn Kechedzhi, Julian Kelly, Trupti Khaire, Tanuj Khattar, Mostafa Khezri, Seon Kim, Paul V. Klimov, Andrey R. Klots, Bryce Kobrin, Pushmeet Kohli, Alexander N. Korotkov, Fedor Kostritsa, Robin Kothari, Borislav Kozlovskii, John Mark Kreikebaum, Vladislav D. Kurilovich, Nathan Lacroix, David Landhuis, Tiano Lange-Dei, Brandon W. Langley, Pavel Laptev, Kim-Ming Lau, Loïck Le Guevel, Justin Ledford, Joonho Lee, Kenny Lee, Yuri D. Lensky, Shannon Leon, Brian J. Lester, Wing Yan Li, Yin Li, Alexander T. Lill, Wayne Liu, William P. Livingston, Aditya Locharla, Erik Lucero, Daniel Lundahl, Aaron Lunt, Sid Madhuk, Fionn D. Malone, Ashley Maloney, Salvatore Mandrà, James Manyika, Leigh S. Martin, Orion Martin, Steven Martin, Cameron Maxfield, Jarrod R. McClean, Matt McEwen, Seneca Meeks, Anthony Megrant, Xiao Mi, Kevin C. Miao, Amanda Mieszala, Reza Molavi, Sebastian Molina, Shirin Montazeri, Alexis Morvan, Ramis Movassagh, Wojciech Mruczkiewicz, Ofer Naaman, Matthew Neeley, Charles Neill, Ani Nersisyan, Hartmut Neven, Michael Newman, Jiun How Ng, Anthony Nguyen, Murray Nguyen, Chia-Hung Ni, Murphy Yuezhen Niu, Thomas E. O'Brien, William D. Oliver, Alex Opremcak, Kristoffer Ottosson, Andre Petukhov, Alex Pizzuto, John Platt, Rebecca Potter, Orion Pritchard, Leonid P. Pryadko, Chris Quintana, Ganesh Ramachandran, Matthew J. Reagor, John Redding, David M. Rhodes, Gabrielle Roberts, Eliott Rosenberg, Emma Rosenfeld, Pedram Roushan, Nicholas C. Rubin, Negar Saei, Daniel Sank, Kannan Sankaragomathi, Kevin J. Satzinger, Henry F. Schurkus, Christopher Schuster, Andrew W. Senior, Michael J. Shearn, Aaron Shorter, Noah Shutty, Vladimir Shvarts, Shraddha Singh, Volodymyr Sivak, Jindra Skruzny, Spencer Small, Vadim Smelyanskiy, W. Clarke Smith, Rolando D. Somma, Sofia Springer, George Sterling, Doug Strain, Jordan Suchard, Aaron Szasz, Alex Sztein, Douglas Thor, Alfredo Torres, M. Mert Torunbalci, Abeer Vaishnav, Justin Vargas, Sergey Vdovichev, Guifre Vidal, Benjamin Villalonga, Catherine Vollgraff Heidweiller, Steven Waltman, Shannon X. Wang, Brayden Ware, Kate Weber, Travis Weidel, Theodore White, Kristi Wong, Bryan W. K. Woo, Cheng Xing, Z. Jamie Yao, Ping Yeh, Bicheng Ying, Juhwan Yoo, Noureldin Yosri, Grayson Young, Adam Zalcman, Yaxing Zhang, Ningfeng Zhu, Nicholas Zobrist, Google Quantum AI, and Collaborators. Quantum error correction below the surface code threshold. *Nature*, 638(8052):920–926, Feb 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08449-y. URL https://doi.org/10.1038/s41586-024-08449-y. 1

Leonardo Alchieri, Davide Badalotti, Pietro Bonardi, and Simone Bianco. An introduction to quantum machine learning: from quantum logic to quantum deep learning. *Quantum Machine Intelligence*, 3(2):28, Nov 2021. ISSN 2524-4914. doi: 10.1007/s42484-021-00056-8. URL https://doi.org/10.1007/s42484-021-00056-8. 1

Jonathan Allcock and Shengyu Zhang. Quantum machine learning. *National Science Review*, 6(1): 26–28, 01 2019. ISSN 2095-5138. doi: 10.1093/nsr/nwy149. URL https://doi.org/10.1093/nsr/nwy149. 1

Jonathan Allcock, Jinge Bao, Joao F. Doriguello, Alessandro Luongo, and Miklos Santha. Constant-depth circuits for Boolean functions and quantum memory devices using multi-qubit gates. *Quantum*, 8:1530, November 2024. ISSN 2521-327X. doi: 10.22331/q-2024-11-20-1530. 2

Andris Ambainis, Joao F Doriguello, and Debbie Lim. A bit of freedom goes a long way: Classical and quantum algorithms for reinforcement learning under a generative model. *arXiv preprint arXiv:2507.22854*, 2025. 1, 4, 5, 6

Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C. Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando G. S. L. Brandao, David A. Buell, Brian Burkett, Yu Chen, Zijun Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Gidney, Marissa Giustina, Rob Graff, Keith Guerin, Steve Habegger, Matthew P. Harrigan, Michael J. Hartmann, Alan Ho, Markus Hoffmann, Trent Huang, Travis S. Humble, Sergei V. Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Paul V. Klimov, Sergey Knysh, Alexander Korotkov, Fedor Kostritsa, David Landhuis, Mike Lindmark, Erik Lucero, Dmitry Lyakh, Salvatore Mandrà, Jarrod R. McClean, Matthew McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen, Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Ostby, Andre Petukhov, John C. Platt, Chris Quintana, Eleanor G. Rieffel, Pedram Roushan, Nicholas C. Rubin, Daniel Sank, Kevin J. Satzinger, Vadim Smelyanskiy, Kevin J. Sung, Matthew D. Trevithick, Amit Vainsencher, Benjamin Villalonga, Theodore White, Z. Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven, and John M. Martinis. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, Oct 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1666-5. URL https://doi.org/10.1038/s41586-019-1666-5. 1

Karl Johan Åström. Optimal control of Markov processes with incomplete state information I. *Journal of mathematical analysis and applications*, 10:174–205, 1965. 1

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3): 325–349, Jun 2013. ISSN 1573-0565. doi: 10.1007/s10994-013-5368-1. 4, 5

Nicole Bäuerle and Ulrich Rieder. *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011. 1

Casey C Bennett and Kris Hauser. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial intelligence in medicine*, 57(1):9–19, 2013. 1

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012. 1

Dimitri Bertsekas. *Abstract dynamic programming*. Athena Scientific, 2022. 1, 6

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, Sep 2017. ISSN 1476-4687. doi: 10.1038/nature23474. 1

Liuhua Chen, Haiying Shen, and Karan Sapra. Distributed autonomous virtual resource management in datacenters using finite-Markov decision process. In *Proceedings of the ACM Symposium on Cloud Computing*, pp. 1–13, 2014. 1

Yanlin Chen and Ronald de Wolf. Quantum algorithms and lower bounds for linear regression with norm constraints. *arXiv preprint arXiv:2110.13086*, 2021. 9

Daniel Crawford, Anna Levit, Navid Ghadermarzy, Jaspreet S. Oberoi, and Pooya Ronagh. Reinforcement learning using quantum boltzmann machines. *Quantum Info. Comput.*, 18(1–2):51–74, February 2018. ISSN 1533-7146. 1

Zhongxiang Dai, Gregory Kang Ruey Lau, Arun Verma, Yao Shu, Bryan Kian Hsiang Low, and Patrick Jaillet. Quantum Bayesian optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 20179–20207. Curran Associates, Inc., 2023. 1

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. 1

João F Doriguello, Alessandro Luongo, and Ewin Tang. Do you know what q-means? *arXiv preprint arXiv:2308.09701*, 2023. 1

Christoph Dürr and Peter Høyer. A quantum algorithm for finding the minimum. *arXiv preprint quant-ph/9607014*, 1996. 2, 4, 5

Eugene A Feinberg and Adam Shwartz. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012. 1

Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Architectures for a quantum random access memory. *Phys. Rev. A*, 78:052310, Nov 2008a. doi: 10.1103/PhysRevA.78.052310. 2

Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum random access memory. *Phys. Rev. Lett.*, 100:160501, Apr 2008b. doi: 10.1103/PhysRevLett.100.160501. 2

Lov K Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 212–219, 1996. 1

Yassine Hamoudi. *Quantum Algorithms for the Monte Carlo Method*. Theses, Université Paris Cité, July 2021a. URL https://cnrs.hal.science/tel-05370224. 4

Yassine Hamoudi. Quantum Sub-Gaussian Mean Estimator. In Petra Mutzel, Rasmus Pagh, and Grzegorz Herman (eds.), *29th Annual European Symposium on Algorithms (ESA 2021)*, volume 204 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 50:1–50:17, Dagstuhl, Germany, 2021b. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-204-4. doi: 10.4230/LIPIcs.ESA.2021.50. URL https://drops.dagstuhl.de/opus/volltexte/2021/14631. 4

Yassine Hamoudi and Frédéric Magniez. Quantum Chebyshev's Inequality and Applications. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi (eds.), *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 69:1–69:16, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-109-2. doi: 10.4230/LIPIcs.ICALP.2019.69. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2019.69. 4

Qiying Hu and Wuyi Yue. *Markov decision processes with their applications*, volume 14. Springer Science & Business Media, 2007. 1

Samuel Jaques and Arthur G. Rattew. QRAM: A Survey and Critique. *Quantum*, 9:1922, December 2025. ISSN 2521-327X. doi: 10.22331/q-2025-12-02-1922. URL https://doi.org/10.22331/q-2025-12-02-1922. 2

Sofiene Jerbi, Arjan Cornelissen, Maris Ozols, and Vedran Dunjko. Quantum Policy Gradient Algorithms. In Omar Fawzi and Michael Walter (eds.), *18th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2023)*, volume 266 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 13:1–13:24, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-283-9. doi: 10.4230/LIPIcs.TQC.2023.13. 4

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003. 3

Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In M. Kearns, S. Solla, and D. Cohn (eds.), *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998a. 4

Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998b. 3

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49(2):193–208, Nov 2002. ISSN 1573-0565. doi: 10.1023/A:1017932429737. 3, 4

Iordanis Kerenidis and Anupam Prakash. Quantum Recommendation Systems. In Christos H. Papadimitriou (ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 49:1–49:21, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-029-3. doi: 10.4230/LIPIcs.ITCS.2017.49. 1

Iordanis Kerenidis, Jonas Landman, Alessandro Luongo, and Anupam Prakash. q-means: A quantum algorithm for unsupervised machine learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

Robin Kothari and Ryan O'Donnell. Mean estimation when you have the source code; or, quantum Monte Carlo methods. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1186–1215, 2023. doi: 10.1137/1.9781611977554.ch44. 2, 4, 5, 6

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12861–12872. Curran Associates, Inc., 2020. 4, 5

Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, Sep 2021. ISSN 1745-2481. doi: 10.1038/s41567-021-01287-z. URL https://doi.org/10.1038/s41567-021-01287-z. 1

Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum algorithms for supervised and unsupervised machine learning. *arXiv preprint arXiv:1307.0411*, 2013. 1

Josep Lumbreras, Erkka Haapasalo, and Marco Tomamichel. Multi-armed quantum bandits: Exploration versus exploitation when learning properties of quantum states. *Quantum*, 6:749, June 2022. ISSN 2521-327X. doi: 10.22331/q-2022-06-29-749. 1

Bin Luo, Yuwen Huang, Jonathan Allcock, Xiaojun Lin, Shengyu Zhang, and John C.S. Lui. Quantum algorithms for finite-horizon Markov decision processes. In *Forty-second International Conference on Machine Learning*, 2025. 4, 5, 6

Ashley Montanaro. Quantum speedup of Monte Carlo methods. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2181):20150301, 2015. 4

Dániel T. R. Nagy, Csaba Czabán, Bence Bakó, Péter Hága, Zsófia Kallus, and Zoltán Zimborás. Hybrid quantum-classical reinforcement learning in latent observation spaces. *Quantum Machine Intelligence*, 7(2):88, Sep 2025. ISSN 2524-4914. doi: 10.1007/s42484-025-00306-z. URL https://doi.org/10.1007/s42484-025-00306-z. 1

Mausam Natarajan and Andrey Kolobov. *Planning with Markov decision processes: An AI perspective*. Springer Nature, 2022. 1

Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010. 1, 2, 4

Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007. 1

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. 1, 4, 5, 6

Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Phys. Rev. Lett.*, 113:130503, Sep 2014. doi: 10.1103/PhysRevLett.113.130503. 1

V. Saggio, B. E. Asenbeck, A. Hamann, T. Strömberg, P. Schiansky, V. Dunjko, N. Friis, N. C. Harris, M. Hochberg, D. Englund, S. Wölk, H. J. Briegel, and P. Walther. Experimental quantum speed-up in reinforcement learning agents. *Nature*, 591(7849):229–233, Mar 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03242-7. URL https://doi.org/10.1038/s41586-021-03242-7. 1

Renato Cesar Sato and Désirée Moraes Zouain. Markov models in health care. *Einstein (São Paulo)*, 8:376–379, 2010. 1

Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015. doi: 10.1080/00107514.2014.964942. URL https://doi.org/10.1080/00107514.2014.964942. 1

P.W. Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pp. 124–134, 1994. doi: 10.1109/SFCS.1994.365700. 1

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. 4, 5, 6, 7

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pp. 770–787, USA, 2018b. Society for Industrial and Applied Mathematics. ISBN 9781611975031. 4

Lauren N Steimle and Brian T Denton. Markov decision processes for screening and treatment of chronic diseases. *Markov Decision Processes in Practice*, pp. 189–222, 2017. 1

Yuexin Su, Ziyi Yang, Peiyuan Huang, Tongyang Li, and Yinyu Ye. Quantum algorithms for bandits with knapsacks with improved regret and time complexities. *arXiv preprint arXiv:2507.04438*, 2025. 1

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998. 1

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer Cham, 2010. doi: 10.1007/978-3-031-01551-9. 1

Zongqi Wan, Zhijie Zhang, Tongyang Li, Jialin Zhang, and Xiaoming Sun. Quantum multi-armed bandits and stochastic linear bandits enjoy logarithmic regrets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):10087–10094, Jun. 2023. doi: 10.1609/aaai.v37i8.26202. 1, 6

Daochen Wang, Aarthi Sundaram, Robin Kothari, Ashish Kapoor, and Martin Roetteler. Quantum algorithms for reinforcement learning with a generative model. In *International Conference on Machine Learning*, pp. 10916–10926. PMLR, 2021. 4, 5

Mengdi Wang. Randomized linear programming solves the discounted Markov decision problem in nearly-linear (sometimes sublinear) running time. *arXiv preprint arXiv:1704.01869*, 2017. 4

Nathan Wiebe, Ashish Kapoor, and Krysta M. Svore. Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning. *Quantum Info. Comput.*, 15(3–4):316–356, March 2015. ISSN 1533-7146. 1

Simon Wiedemann, Daniel Hein, Steffen Udluft, and Christian B. Mendl. Quantum policy iteration via amplitude estimation and Grover search – towards quantum advantage for reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. 4

Han Zhong, Jiachen Hu, Yecheng Xue, Tongyang Li, and Liwei Wang. Provably efficient exploration in quantum reinforcement learning with logarithmic worst-case regret. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. 4

A Optimal Policies for Infinite-Horizon Discounted MDPs

In this section, we describe our quantum algorithm (Algorithm 2) for computing ε-optimal policies for infinite-horizon discounted MDPs. We shall need the following technical result (total-variance technique).

Lemma 1. *Given policy $\pi = (\pi_t)_{t\in\mathbb{N}}$, define $\sigma_t^\pi \in \mathscr{B}(\mathcal{S})$ as $\sigma_t^\pi = \mathcal{P}^{\pi_{t+1}}(V_t^{\pi,\gamma})^2 - (\mathcal{P}^{\pi_{t+1}}V_t^{\pi,\gamma})^2$ for $t \in \mathbb{N}$. For any policy $\pi = (\pi_t)_{t\in\mathbb{N}}$ and $t \in \mathbb{N}$,*

$$\left\| \sum_{t'=0}^{t} \left(\prod_{i=t'+2}^{t+1} \gamma \mathcal{P}^{\pi_i} \right) \gamma \sigma_{t'}^\pi \right\|_\infty \leq 2\Gamma^2.$$

Proof. Observe that

$$\begin{aligned}
\gamma\sigma_t^\pi &\leq \gamma\mathcal{P}^{\pi_{t+1}}(V_t^{\pi,\gamma})^2 - (\gamma\mathcal{P}^{\pi_{t+1}}V_t^{\pi,\gamma})^2 \\
&= \gamma\mathcal{P}^{\pi_{t+1}}(V_t^{\pi,\gamma})^2 - (V_{t+1}^{\pi,\gamma} - r^{\pi_{t+1}})^2 && (V_{t+1}^{\pi,\gamma} = r^{\pi_{t+1}} + \gamma\mathcal{P}^{\pi_{t+1}}V_t^{\pi,\gamma}) \\
&\leq \gamma\mathcal{P}^{\pi_{t+1}}(V_t^{\pi,\gamma})^2 - (V_{t+1}^{\pi,\gamma})^2 + 2V_{t+1}^{\pi,\gamma}. && (r^{\pi_{t+1}} \leq \mathbf{1})
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{t'=0}^{t} \left(\prod_{i=t'+2}^{t+1} \gamma\mathcal{P}^{\pi_i} \right) \gamma\sigma_{t'}^\pi &\leq \sum_{t'=0}^{t} \left(\prod_{i=t'+2}^{t+1} \gamma\mathcal{P}^{\pi_i} \right) \left(\gamma\mathcal{P}^{\pi_{t'+1}}(V_{t'}^{\pi,\gamma})^2 - (V_{t'+1}^{\pi,\gamma})^2 + 2V_{t'+1}^{\pi,\gamma} \right) \\
&\leq 2\Gamma^2(1-\gamma^{t+1})\mathbf{1} + \sum_{t'=0}^{t} \left(\prod_{i=t'+1}^{t+1} \gamma\mathcal{P}^{\pi_i} \right) (V_{t'}^{\pi,\gamma})^2 - \sum_{t'=0}^{t} \left(\prod_{i=t'+2}^{t+1} \gamma\mathcal{P}^{\pi_i} \right) (V_{t'+1}^{\pi,\gamma})^2 \\
&\qquad (\|V_{t'+1}^{\pi,\gamma}\|_\infty \leq \Gamma \text{ and } \|\textstyle\sum_{t'=0}^{t} \prod_{i=t'+1}^{t} \gamma\mathcal{P}^{\pi_i}\|_\infty \leq \Gamma(1-\gamma^{t+1})) \\
&= 2\Gamma^2(1-\gamma^{t+1})\mathbf{1} + \left(\prod_{i=1}^{t+1} \gamma\mathcal{P}^{\pi_i} \right) (V_0^{\pi,\gamma})^2 - (V_{t+1}^{\pi,\gamma})^2 \quad \text{(telescope sum)} \\
&\leq (2\Gamma^2(1-\gamma^{t+1}) + \gamma^{t+1})\mathbf{1} && (\|V_0^{\pi,\gamma}\|_\infty \leq 1) \\
&\leq 2\Gamma^2\mathbf{1}. && \square
\end{aligned}$$

Theorem 2. *Let $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$ be a time-independent infinite-horizon discounted MDP with $\gamma \in [0,1)$. Let $\delta \in (0,1)$, $\varepsilon \in (0,1]$, and $T := \Gamma\lceil \ln(\Gamma/\varepsilon) \rceil$. Algorithm 2 computes functions $\{V_t\}_{t\in[T]} \subset \mathscr{B}(\mathcal{S})$ and deterministic policy $\pi = (\pi_t)_{t\in[T]}$ such that, with probability $1-\delta$,*

$$V_t^{*,\gamma} - \varepsilon\mathbf{1} \leq V_t \leq V_t^{\pi,\gamma} \leq V_t^{*,\gamma} \text{ and } V_\infty^{*,\gamma} - (\varepsilon + \Gamma\gamma^{t+1})\mathbf{1} \leq V_t \leq V_\infty^{\pi_t^\infty,\gamma} \leq V_\infty^{*,\gamma} \quad \forall t \in [T].$$

In particular, $V_\infty^{*,\gamma} - 2\varepsilon\mathbf{1} \leq V_\infty^{\pi_{T-1}^\infty,\gamma} \leq V_\infty^{*,\gamma}$. Its query complexity is (up to $\operatorname{poly}\log\log$ *factors)*

$$\widetilde{O}\left(\frac{\Gamma^{2.5} S\sqrt{A}}{\varepsilon} \log\left(\frac{\Gamma SA}{\delta}\right) \log\left(\frac{\Gamma S}{\delta}\right) \log\left(\frac{\Gamma}{\varepsilon}\right) \right).$$

Proof. The proof is similar to Theorem 1. We shall prove by induction on $t \in \mathbb{N}$ that

$$V_t^{*,\gamma} - \varepsilon\mathbf{1} \leq V_t \leq V_t^{\pi,\gamma} \leq V_t^{*,\gamma} \quad \text{and} \quad V_t \leq V_\infty^{\pi_t^\infty,\gamma} \leq V_\infty^{*,\gamma} \qquad \forall t \in [T].$$

Algorithm 2 Quantum algorithm for optimal policies for infinite-horizon discounted MDPs

Require: Finite state space $\mathcal{S}$ and action space $\mathcal{A}$, discount factor $\gamma \in [0,1)$, quantum sampling access to p, quantum access to r, failure probability $\delta \in (0,1)$, error $\varepsilon \in (0, \sqrt{\Gamma}]$.

Ensure: 2ε-optimal stationary deterministic policy π_{T-1}^{∞}.

Let $T \leftarrow \Gamma\lceil \ln(\Gamma/\varepsilon)\rceil$ and $\theta \leftarrow \frac{\varepsilon}{7\Gamma^{3/2}}$

$V_0(s) \leftarrow \max_{a\in\mathcal{A}} r(s,a)$ and $\pi_0(s) \leftarrow \arg\max_{a\in\mathcal{A}} r(s,a)$ with probability $1-\frac{\delta}{TS}$ $\forall s \in \mathcal{S}$ (Fact 1)

for $t \in [T-1]$ **do**

  Let $\mu_t(s,a) = \sum_{s'\in\mathcal{S}} p(s'|s,a)V_t(s')$

  Let $\sigma_t(s,a) = \sum_{s'\in\mathcal{S}} p(s'|s,a)V_t(s')^2 - \big(\sum_{s'\in\mathcal{S}} p(s'|s,a)V_t(s')\big)^2$

  for $s \in \mathcal{S}$ **do**

    Use Fact 2 to obtain a unitary

$$\mathcal{U}_s^{(t)} : |a\rangle|\bar{0}\rangle \mapsto |a\rangle\big(\sqrt{1-\delta_a}|\widehat{\mu}_t(s,a)\rangle|\widetilde{\mu}_t(s,a)\rangle|\widetilde{\sigma}_t(s,a)\rangle|\mathrm{garbage}(a)\rangle + \sqrt{\delta_a}|\perp_a\rangle\big)$$

with $\max_{a\in\mathcal{A}} \delta_a = \widetilde{O}\big(\frac{\delta^2}{T^2S^2A}\big)$ and

$$|\widetilde{\mu}_t(s,a) - \mu_t(s,a)| \le \theta\sqrt{\sigma_t(s,a)},$$
$$|\widetilde{\sigma}_t(s,a) - \sigma_t(s,a)| \le \theta\Gamma^2,$$
$$\widehat{\mu}_t(s,a) := \widetilde{\mu}_t(s,a) - \theta\sqrt{\widetilde{\sigma}_t(s,a)} - \theta^{3/2}\Gamma.$$

    Use quantum maximum finding with unitary $\mathcal{U}_s^{(t)}$ (Fact 1 with $\mathcal{U}_s^{(t)}$) to obtain $V'_{t+1}(s)$ and $\pi'_{t+1}(s)$ such that, with probability $1-\frac{\delta}{TS}$,

$$V'_{t+1}(s) = \max_{a\in\mathcal{A}}\{r(s,a) + \gamma\widehat{\mu}_t(s,a)\} \quad \text{and} \quad \pi'_{t+1}(s) = \arg\max_{a\in\mathcal{A}}\{r(s,a) + \gamma\widehat{\mu}_t(s,a)\}$$

    if $V'_{t+1}(s) \ge V_t(s)$ **then** $V_{t+1}(s) \leftarrow V'_{t+1}(s)$ and $\pi_{t+1}(s) \leftarrow \pi'_{t+1}(s)$

    else $V_{t+1}(s) \leftarrow V_t(s)$ and $\pi_{t+1}(s) \leftarrow \pi_t(s)$

    end if

  end for

end for

return $(\pi_0, \ldots, \pi_{T-1})$

The inequality $V_\infty^{*,\gamma} - (\varepsilon + \Gamma\gamma^{t+1})\mathbf{1} \le V_t$ then follows from the above and the fact that (let $d^* \in \arg\max_{d\in\mathcal{D}^{\mathrm{D}}} V_\infty^{d^\infty,\gamma}$ be an optimal deterministic decision rule)

$$\begin{aligned} V_\infty^{*,\gamma}(s) &= \mathbb{E}\left[\sum_{t'=0}^{\infty}\gamma^{t'}r(s_{t'},d^*(s_{t'})) \,\middle|\, s_0 = s, s_i \sim p(\cdot|s_i,d^*(s_i))\right] \\ &\le \mathbb{E}\left[\sum_{t'=0}^{t}\gamma^{t'}r(s_{t'},d^*(s_{t'})) \,\middle|\, s_0 = s, s_i \sim p(\cdot|s_i,d^*(s_i))\right] + \sum_{t'=t+1}^{\infty}\gamma^{t'} \qquad (r \le \mathbf{1}) \\ &\le V_t^{*,\gamma}(s) + \gamma^{t+1}\Gamma. \end{aligned}$$

Moreover, $\gamma^T\Gamma \le e^{-T(1-\gamma)}\Gamma \le \varepsilon$ if $T = \Gamma\lceil\ln(\Gamma/\varepsilon)\rceil$, using that $x \le e^{-(1-x)}$ for all $x \in \mathbb{R}$.

The base case $t = 0$ of the induction is trivial since $V_0(s) = \max_{a\in\mathcal{A}}\{r(s,a)\} = V_0^{*,\gamma}(s)$ and $\pi_0(s) = \arg\max_{a\in\mathcal{A}}\{r(s,a)\}$. Assume then that Algorithm 2 has computed functions $V_0, \ldots, V_t$ and decision rules $\pi_0, \ldots, \pi_t$ (which form policy π) such that

$$V_{t'}^{*,\gamma} - \varepsilon\mathbf{1} \le V_{t'} \le V_{t'}^{\pi,\gamma} \le V_{t'}^{*,\gamma} \quad \text{and} \quad V_{t'} \le V_\infty^{\pi_{t'}^\infty,\gamma} \le V_\infty^{*,\gamma} \qquad \forall t' = 0, \ldots, t,$$

and consider the time step $t+1$. Let the true quantities

$$\mu_t(s,a) := (\mathcal{P}^a V_t)(s) \qquad \text{and} \qquad \sigma_t(s,a) := (\mathcal{P}^a V_t^2)(s) - (\mathcal{P}^a V_t)^2(s).$$

Let $\theta := \frac{\varepsilon}{7\Gamma^{3/2}}$. Once again we invoke Fact 2 twice to construct the unitary

$$\mathcal{O}_s^{(t)} : |a\rangle|\bar{0}\rangle \mapsto |a\rangle\big(\sqrt{1-\delta_a}|\widetilde{\mu}_t(s,a)\rangle|\widetilde{\sigma}_t(s,a)\rangle|\mathrm{garbage}(a)\rangle + \sqrt{\delta_a}|\perp_a\rangle\big), \tag{3}$$

where $\delta_a \in [0,1)$ is the failure probability of quantum mean estimation within the branch $|a\rangle$ of the wave function, $|\text{garbage}(a)\rangle$ is a "garbage" unit complex vector accumulated through the computation, $|\perp_a\rangle$ is a unit complex vector orthogonal to $|\widetilde{\mu}_t(s,a)\rangle|\widetilde{\sigma}_t(s,a)\rangle|\text{garbage}(a)\rangle$, and

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad \begin{aligned} |\widetilde{\mu}_t(s,a) - \mu_t(s,a)| &\leq \theta \sqrt{\sigma_t(s,a)}, \\ |\widetilde{\sigma}_t(s,a) - \sigma_t(s,a)| &\leq 2\theta\Gamma^2, \end{aligned} \tag{4}$$

using that the standard deviation of V_t^2 is at most Γ^2 and thus one can compute approximations of $(\mathcal{P}^a V_t^2)(s)$ and $(\mathcal{P}^a V_t)^2(s)$ with error at most $\theta\Gamma^2$ each. Using Equation (4),

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad |\widetilde{\mu}_t(s,a) - \mu_t(s,a)| \leq \theta \sqrt{\widetilde{\sigma}_t(s,a)} + \sqrt{2}\theta^{3/2}\Gamma.$$

Define then the quantities

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad \widehat{\mu}_t(s,a) := \widetilde{\mu}_t(s,a) - \theta \sqrt{\widetilde{\sigma}_t(s,a)} - \sqrt{2}\theta^{3/2}\Gamma,$$

which have one-side error. We can express the above quantity using the optimal-policy variance $\sigma_t^*(s,a) := (\mathcal{P}^a (V_t^{*,\gamma})^2)(s) - (\mathcal{P}^a V_t^{*,\gamma})^2(s)$ since

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad \sqrt{\widetilde{\sigma}_t(s,a)} \leq \sqrt{\sigma_t(s,a)} + \sqrt{2\theta}\Gamma \leq \sqrt{\sigma_t^*(s,a)} + \varepsilon + \sqrt{2\theta}\Gamma,$$

where we used Equation (4) and that $\sqrt{\text{Var}[V_t]} \leq \sqrt{\text{Var}[V_t^{*,\gamma}]} + \sqrt{\text{Var}[V_t^{*,\gamma} - V_t]} \leq \sqrt{\text{Var}[V_t^{*,\gamma}]} + \varepsilon$ if $\|V_t^{*,\gamma} - V_t\|_\infty \leq \varepsilon$ according to the induction hypothesis. Hence

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad \begin{aligned} \widehat{\mu}_t(s,a) &\leq \mu_t(s,a), \\ \widehat{\mu}_t(s,a) &\geq \mu_t(s,a) - 2\theta\sqrt{\sigma_t^*(s,a)} - 2\theta\varepsilon - 4\sqrt{2}\theta^{3/2}\Gamma. \end{aligned}$$

By subtracting the content of register $|\widetilde{\sigma}_t(s,a)\rangle$ from the register $|\widetilde{\mu}_t(s,a)\rangle$ in Equation (3), it is then possible to effectively construct a black-box unitary $\mathcal{U}_s^{(t)} : |a\rangle|\bar{0}\rangle \mapsto |a\rangle(\sqrt{1-\delta_a}|\widehat{\mu}_t(s,a)\rangle + \sqrt{\delta_a}|\perp_a\rangle)$ using oracle $\mathcal{Q}_p, \mathcal{Q}_p^\dagger$ (up to garbage states). The maximum over $a \in \mathcal{A}$ of $r(s,a) + \gamma\widehat{\mu}_t(s,a)$ can thus be found by using quantum maximum finding (Fact 1) with unitary $\mathcal{U}_s^{(t)}$, leading to $V'_{t+1}(s) = \max_{a\in\mathcal{A}}\{r(s,a) + \gamma\widehat{\mu}_t(s,a)\}$ and $\pi'_{t+1}(s) = \arg\max_{a\in\mathcal{A}}\{r(s,a) + \gamma\widehat{\mu}_t(s,a)\}$ with probability $1 - \frac{\delta}{HS}$. We then let

$$V_{t+1}(s) = \max\{V'_{t+1}(s), V_t(s)\} \quad \text{and} \quad \pi_{t+1}(s) = \begin{cases} \pi'_{t+1}(s) & \text{if } V'_{t+1}(s) \geq V_t(s), \\ \pi_t(s) & \text{if } V'_{t+1}(s) < V_t(s). \end{cases}$$

This means that

$$\forall s \in \mathcal{S}: \qquad V_{t+1}(s) \geq (\mathcal{L}_{\pi^*_{t+1}} V_t)(s) - 2\gamma\theta\sqrt{\sigma_t^*(s, \pi^*_{t+1}(s))} - 2\gamma\theta\varepsilon - 4\sqrt{2}\gamma\theta^{3/2}\Gamma. \tag{5}$$

We now prove that $V_{t+1} \leq \mathcal{L}^{\pi_{t+1}} V_t = V_{t+1}^{\pi,\gamma} \leq \mathcal{L}^{\pi_{t+1}} V_{t+1}$. There are two cases to analyse.

Case I: $V'_{t+1}(s) \geq V_t(s)$. Then

$$V_{t+1}(s) = r(s, \pi_{t+1}(s)) + \gamma\widehat{\mu}_t(s, \pi_{t+1}(s)) \leq (\mathcal{L}^{\pi_{t+1}} V_t)(s) \leq (\mathcal{L}^{\pi_{t+1}} V_{t+1})(s),$$

since $\widehat{\mu}_t(s,a) \leq \mu_t(s,a)$ and $V_{t+1} \geq V_t$.

Case II: $V'_{t+1}(s) < V_t(s)$. Then

$$V_{t+1}(s) = V_t(s) \leq (\mathcal{L}^{\pi_t} V_t)(s) = (\mathcal{L}^{\pi_{t+1}} V_t)(s) \leq (\mathcal{L}^{\pi_t} V_{t+1})(s),$$

using the induction hypothesis in $V_t(s) \leq (\mathcal{L}^{\pi_t} V_t)(s)$ and that $V_{t+1} \geq V_t$ and $\pi_{t+1}(s) = \pi_t(s)$. This proves that $V_{t+1} \leq V_{t+1}^{\pi,\gamma} \leq \mathcal{L}^{\pi_{t+1}} V_{t+1}$, which by the monotonicity of the Bellman operator $\mathcal{L}^{\pi_{t+1}}$ and the fact that $V_\infty^{\pi^\infty_{t+1},\gamma}$ is its fixed point, leads to $V_{t+1} \leq (\mathcal{L}^{\pi_{t+1}})^\infty V_{t+1} = V_\infty^{\pi^\infty_{t+1},\gamma} \leq V_\infty^{*,\gamma}$.

We move on to proving that $V_{t+1}^{*,\gamma} - \varepsilon\mathbf{1} \leq V_{t+1}$. For such,

$$V_{t+1}^{*,\gamma} - V_{t+1} \leq \mathcal{L}^{\pi^*_{t+1}} V_t^{*,\gamma} - \mathcal{L}^{\pi^*_{t+1}} V_t + \gamma\xi_{t+1} = \gamma\mathcal{P}^{\pi^*_{t+1}}(V_t^{*,\gamma} - V_t) + \gamma\xi_{t+1},$$
(by Equation (5) and $V_{t+1}^{*,\gamma} = \mathcal{L}^{\pi^*_{t+1}} V_t^{*,\gamma}$)

where we defined $\xi_{t+1}(s) = 2\theta\sqrt{\sigma_t^*(s,\pi_{t+1}^*(s))} + 2\theta\varepsilon + 4\sqrt{2}\theta^{3/2}\Gamma$. Let $\sigma_t^* = \mathcal{P}^{\pi_{t+1}^*}(V_t^{*,\gamma})^2 - (\mathcal{P}^{\pi_{t+1}^*}V_t^{*,\gamma})^2 \in \mathscr{B}(\mathcal{S})$. Solving the above recursion with the boundary condition $V_0^{*,\gamma} \equiv V_0$,

$$V_{t+1}^{*,\gamma} - V_{t+1} \le \sum_{t'=1}^{t+1}\left(\prod_{i=t'+1}^{t+1}\gamma\mathcal{P}^{\pi_i^*}\right)\gamma\xi_{t'}.$$

Using that

$$\left\|\sum_{t'=0}^{t}\left(\prod_{i=t'+2}^{t+1}\gamma\mathcal{P}^{\pi_i^*}\right)\gamma\sqrt{\sigma_{t'}^*}\right\|_\infty \le \left\|\Gamma\sum_{t'=0}^{t}\left(\prod_{i=t'+2}^{t+1}\gamma\mathcal{P}^{\pi_i^*}\right)\gamma^2\sigma_{t'}^*\right\|_\infty^{1/2} \le \sqrt{2}\Gamma^{3/2},$$

(by Cauchy-Schwarz and Lemma 1)

and that $\|\sum_{t'=1}^{t+1}\prod_{i=t'+1}^{t+1}\gamma\mathcal{P}^{\pi_i^*}\|_\infty \le \sum_{t'=0}^{t}\gamma^{t'} \le \Gamma$, then finally

$$V_{t+1}^{*,\gamma}(s) - V_{t+1}(s) \le 2\sqrt{2}\theta\Gamma^{3/2} + 2\theta\Gamma\varepsilon + 4\sqrt{2}\theta^{3/2}\Gamma^2 = \varepsilon\left(\frac{2\sqrt{2}}{7} + \frac{2\varepsilon}{7\sqrt{\Gamma}} + \frac{4\sqrt{2}}{7^{3/2}\Gamma^{1/4}}\right) < \varepsilon.$$

($\theta = \frac{\varepsilon}{7\Gamma^{3/2}}$ and $\varepsilon \le \sqrt{\Gamma}$)

The error analysis of Algorithm 2 is very similar to Algorithm 1. The failure probability is $\delta_1 + c\sqrt{2\delta_2 A}\log\frac{1}{\delta_1}$, where δ_1 is the error associated with quantum maximum finding (Fact 1) and δ_2 is the error behind quantum mean estimation (Fact 2). Choosing $\delta_1 = O\big(\frac{\delta}{TS}\big)$ and $\delta_2 = O\big(\frac{\delta_1^2}{A\log^2(1/\delta_1)}\big)$ leads to outputting $\max_{a\in\mathcal{A}}\{r(s,a) + \gamma\widehat{\mu}_t(s,a)\}$ with probability at least $1 - \frac{\delta}{TS}$. A union bound over all $(s,t) \in \mathcal{S}\times[T]$ leads to a failure probability at most δ.

Regarding the query complexity, for each $s \in \mathcal{S}$, one call to the unitary $\mathcal{U}_s^{(t)}$ in order to compute $\widetilde{\mu}_t(s,a)$ and $\widetilde{\sigma}_t(s,a)$ in Equation (3) uses $O\big(\frac{1}{\theta}\log\frac{1}{\delta_2}\big) = \widetilde{O}\big(\frac{\Gamma^{3/2}}{\varepsilon}\log\frac{\Gamma SA}{\delta}\big)$ queries to $\mathcal{Q}_p$, while quantum maximum finding makes $O\big(\sqrt{A}\log\frac{1}{\delta_1}\big) = \widetilde{O}\big(\sqrt{A}\log\frac{\Gamma S}{\delta}\big)$ queries to $\mathcal{U}_s^{(t)}$. Summing over all $\mathcal{S}\times[T]$, we obtain the stated query complexity. □