

Improved Quantum Algorithms for Reinforcement Learning Under a Generative Model

Joao F. Doriguello¹

¹HUN-REN Alfréd Rényi Institute of Mathematics, Budapest, Hungary

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 choice for 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 computing approximate optimal policies. Our quantum algorithms are based on a new combination of standard value iteration and quantum subroutines like quantum mean estimation and quantum minimum finding, overall enhanced with techniques from the sample-optimal classical algorithms. Our resulting query complexities improve upon previous works, thus approaching already established quantum lower bounds.

The paradigm of reinforcement learning (RL) [35] has received much attention recently due to its success in modeling an agent interaction with a dynamical environment [2, 14, 31, 3, 12, 4, 8, 34, 27] and due to its applicability to a wide range of problems [35, 36, 5, 6]. Such agent-environment interaction is usually studied using Markov decision processes (MDPs) [29].

In its most general form, a (discrete-time) MDP is described by a finite set of states $\mathcal{S}$ that the environment can assume and which the agent observes, a finite set of actions $\mathcal{A}$ that the agent can take, a reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ that describes the amount of reward the agent receives when choosing action a under the environment's state s , and stochastic kernels $p(\cdot | s, a)$ that denote the probability of transitioning to a new state given the current environment's state s and the agent's chosen action a after the reward is obtained. Beyond this general formalisation, there are different subtypes of MDPs depending on the reward criteria the agent must maximise. Central to this work are *finite-horizon* MDPs, where the agent interacts with the environment for a pre-determined number of steps H called *horizon*, and *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* $\gamma \in [0, 1)$. Finite- and infinite-horizon discounted MDPs are thus described by tuples $\langle \mathcal{S}, \mathcal{A}, p, r, H \rangle$ and $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, respectively.

The main agent's objective when interacting with the environment is maximising some reward criteria, which is done by choosing an appropriate *policy* $\pi = (\pi_t)_t$, which is a sequence of probability distributions over $\mathcal{A}$ given s , $\pi_t : \mathcal{S} \rightarrow \Delta(\mathcal{A})$. The reward criteria for finite-horizon MDPs is the expected sum of rewards under policy $\pi = (\pi_h)_{h \in [H]}$ and initial state s ,

$$V_0^\pi(s) = \mathbb{E} \left[\sum_{h=0}^{H-1} r(s_h, a_h) \mid s_0 = s, a_i \sim \pi_i(s_i), s_{i+1} \sim p(\cdot | s_i, a_i) \right].$$

A policy π^ε is said to be ε -*optimal* if $V_0^{\pi^\varepsilon}(s) \geq V_0^\pi(s) - \varepsilon$ for all π and $s \in \mathcal{S}$, and the optimal value is $V_0^*(s) = \sup_\pi V_0^\pi(s)$. On the other hand, the reward criteria for infinite-horizon MDPs is the discounted sum of rewards under policy $\pi = (\pi_t)_{t \in \mathbb{N}}$ and initial state s ,

$$V_\gamma^\pi(s) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_i \sim \pi_i(s_i), s_{i+1} \sim p(\cdot | s_i, a_i) \right].$$

Once again, a policy π^ε is said to be ε -optimal if $V_{\gamma^\varepsilon}^\pi(s) \geq V_\gamma^\pi(s) - \varepsilon$ for all π and $s \in \mathcal{S}$, and the optimal value is $V_\gamma^*(s) = \sup_\pi V_\gamma^\pi(s)$. Let also $\Gamma := (1 - \gamma)^{-1}$.

One of the fundamental theoretical problems related to MDPs is that of learning an ε -optimal policy. Among the several studied input models, arguably the most important — and the one considered in this work — is the so-called *generative model* [18, 17, 16] where one has full knowledge of state and action spaces and of the reward function, but the transition probabilities $p(\cdot|s, a)$ can only be accessed through an oracle. In the classical setting, this means access to an oracle $\mathcal{C}_p$ that, on input $(s, a) \in \mathcal{S} \times \mathcal{A}$, returns $s' \in \mathcal{S}$ with probability $p(s'|s, a)$. One is thus interested in computing an ε -optimal policy with the least amount of queries to oracle $\mathcal{C}_p$.

In the classical setting, 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 [19, 17, 13, 38, 33, 32, 23], until Sidford et al. [32] obtained sample-optimal algorithms for finite- and infinite-horizon discounted MDPs with query complexity $\tilde{O}(\frac{H^3 SA}{\varepsilon^2})$ and $\tilde{O}(\frac{\Gamma^3 SA}{\varepsilon^2})$, respectively, thus matching the lower bounds from [13, 32].

Parallel to RL developments, quantum machine learning [7], a subfield of quantum computation [28], has been consolidated as an area of study that can provide quantum speedups to several traditional machine learning problems [24, 30, 21, 20, 10], including RL [9, 1]. The equivalent version of generative access to p in the quantum setting is through an oracle $\mathcal{Q}_p$ (and its inverse) called quantum accessible-environment [37, 39, 15, 40] which is a unitary operator

$$\mathcal{Q}_p : |s\rangle|a\rangle|\bar{0}\rangle \rightarrow \sum_{s' \in \mathcal{S}} \sqrt{p(s'|s, a)} |s'\rangle|a\rangle|s'\rangle \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \quad (1)$$

The same fundamental problem of learning an ε -optimal policy can thus be proposed in a quantum setting where one has access to the oracle $\mathcal{Q}_p$. The first work to consider quantum algorithms for ε -optimal policies in the infinite-horizon discounted setting was due to Wang et al. [37], who obtained the query upper bound $\tilde{O}(\min\{\frac{\Gamma^{1.5} SA}{\varepsilon}, \frac{\Gamma^3 S\sqrt{A}}{\varepsilon}\})$ and the query lower bound $\Omega(\frac{\Gamma^{1.5} S\sqrt{A}}{\varepsilon})$. The finite-horizon setting, on the other hand, has been more recently addressed by Luo et al. [25] and Ambainis et al. [1], both works proposing query upper bounds of $\tilde{O}(\min\{\frac{H^{2.5} SA}{\varepsilon}, \frac{H^3 S\sqrt{A}}{\varepsilon}\})$. Luo et al. [25] further proved a query lower bound of $\Omega(\frac{H^{1.5} S\sqrt{A}}{\varepsilon})$.

The quantum algorithms of [37, 25, 1] are the combination of two different algorithms. The first one is a quantised version of the standard value iteration [29], where for finite-horizon MDPs one solves the equations

$$V_t(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \sum_{s' \in \mathcal{S}} p(s'|s, a) V_{t+1}(s') \right\} \quad \forall s \in \mathcal{S} \quad \text{and} \quad t = H, H-1, \dots, 0 \quad (2)$$

with initial condition $u_H \equiv 0$, while for infinite-horizon discounted MDPs one solves

$$V_{t+1}(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V_t(s') \right\} \quad \forall s \in \mathcal{S} \quad \text{and} \quad t \in \mathbb{N} \quad (3)$$

with initial condition $V_0 \equiv 0$. One can show that the sequence of maximums $a \in \mathcal{A}$ in the above equations form an optimal policy, i.e., $\pi_t(s) = \arg \max_{a \in \mathcal{A}} \{r(s, a) + \sum_{s' \in \mathcal{S}} p(s'|s, a) V_{t+1}(s')\}$ and similarly for infinite-horizon discounted MDPs. The quantum advantage comes from estimating $\sum_{s' \in \mathcal{S}} p(s'|s, a) V_{t+1}(s')$ with error $\frac{\varepsilon}{H}$ using quantum mean estimation [26, 22] with $O(\frac{H^2}{\varepsilon})$ queries to $\mathcal{Q}_p$ and nesting it into the quantum minimum finding subroutine of Dürr and Høyer [11] in order to find $V_t(s)$, which uses $O(\frac{H^2 \sqrt{A}}{\varepsilon})$ queries to $\mathcal{Q}_p$.

The second algorithm of [37, 25, 1] is a quantisation of the more advanced value iteration of Sidford et al. [32] (which is optimal as mentioned above). The advanced value iteration of [32] work in *epochs*¹: by starting the k -th epoch with functions $V_0^{(k-1)}, \dots, V_{H-1}^{(k-1)}$ such that

¹ Here explained for finite-horizon MDPs. The infinite-horizon discounted case is similar.

$0 \leq V_t^*(s) - V_t^{(k-1)}(s) \leq 2\epsilon_k$ for all $s \in \mathcal{S}$ and $t \in [H]$, their algorithms produce $V_0^{(k)}, \dots, V_{H-1}^{(k)}$ such that $0 \leq V_t^*(s) - V_t^{(k)}(s) \leq \epsilon_k$ by the end of the k -th epoch, thus halving the initial error.

The algorithm of [32] further uses three crucial techniques: *monotonicity*, *variance reduction*, and *total-variance* techniques. The monotonicity technique means maintaining the monotonicity condition $V_0^{(k)}(s) \leq V_0^{\pi^{(k)}}(s) \leq V_0^*(s)$, which yields an ϵ -optimal policy. Variance reduction rewrites the standard value iteration as

$$V_t^{(k)}(s) \leftarrow \max_{a \in \mathcal{A}} \left\{ r(s, a) + \sum_{s' \in \mathcal{S}} p(s'|s, a) V_{t+1}^{(k-1)}(s') + \sum_{s' \in \mathcal{S}} p(s'|s, a) (V_{t+1}^{(k)}(s') - V_{t+1}^{(k-1)}(s')) \right\}.$$

The main idea is that $\sum_{s' \in \mathcal{S}} p(s'|s, a) V_t^{(k-1)}(s')$ for all $t \in [H]$ can be computed at the beginning of the epoch using the same batch of $\tilde{O}(\frac{H^4}{\epsilon_k^2})$ samples, which saves a factor of H . Regarding the quantity $\sum_{s' \in \mathcal{S}} p(s'|s, a) (V_{t+1}^{(k)}(s') - V_{t+1}^{(k-1)}(s'))$, since $\|V_{t+1}^{(k)} - V_{t+1}^{(k-1)}\|_\infty \leq \epsilon_k$, it can be approximated up to error $\frac{\epsilon_k}{2H}$ using $\tilde{O}(H^2)$ samples for each $t \in [H]$, leading to $\tilde{O}(H^3)$ samples.

Finally, in order to reduce the sample complexity dependence on H down to $O(H^3)$, Ref. [32] show that the true error accumulates as $\sqrt{H^3/m}$ given m samples, much less than the naive sum of estimation errors at each time step. This means that one does not require the update error to be $\frac{\epsilon_k}{H}$ at each time step and $m = O(\frac{H^3}{\epsilon_k^2})$ samples suffices for a total error ϵ_k .

In the quantum algorithms of [25, 1], the quantities $\sum_{s' \in \mathcal{S}} p(s'|s, a) (u_{t+1}^{(k)}(s') - u_{t+1}^{(k-1)}(s'))$ are estimated using quantum mean estimation [26, 22] with $\tilde{O}(H)$ queries to the oracle $\mathcal{Q}_p$, for a total of $\tilde{O}(H^2)$ queries. Similarly for the quantities $\sum_{s' \in \mathcal{S}} p(s'|s, a) u_t^{(k-1)}(s')$. However, unlike the classical case where the same batch of samples can be reused to approximate all these quantities, in the quantum setting we must start each calculation anew, which incurs an extra H factor in the query complexity and hinders a full quadratic advantage in H .

In this work, we propose new quantum algorithms that improve upon the query upper bounds of [37, 25, 1] both for finite- and infinite-horizon discounted MDPs!

Result 1. *Let $\langle \mathcal{S}, \mathcal{A}, p, r, H \rangle$ be a finite-horizon MDP. There is a quantum algorithm that outputs an ϵ -optimal policy with high probability and query complexity $\tilde{O}(\frac{H^{2.5}S\sqrt{A}}{\epsilon})$.*

Result 2. *Let $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$ be an infinite-horizon discounted MDP. There is a quantum algorithm that outputs an ϵ -optimal policy with high probability and query complexity $\tilde{O}(\frac{\Gamma^{2.5}S\sqrt{A}}{\epsilon})$.*

Our new algorithms are a mixture of both approaches highlighted above. More precisely, we revisit and improve the standard value iteration of (2) and (3) by using the monotonicity and total-variance techniques from [32]. This includes using quantum mean estimation *with variance* [22] to approximate the mean $\mu_{t+1}(s, a) = \sum_{s' \in \mathcal{S}} p(s'|s, a) V_{t+1}(s')$ together with its variance $\sigma_{t+1}(s, a)$ on the flight as in

$$\sum_{a \in \mathcal{A}} \alpha_a |s, a\rangle |\bar{0}\rangle \mapsto \sum_{a \in \mathcal{A}} \alpha_a |s, a\rangle |\tilde{\mu}_{t+1}(s, a)\rangle |\tilde{\sigma}_{t+1}(s, a)\rangle |\text{garbage}(a)\rangle, \quad (4)$$

which will serve as the base unitary for quantum minimum finding. Therefore, in contrast to [37, 25, 1] that either (i) use quantum minimum finding but does not include any variance in the error propagation or (ii) take variance into account but not quantum minimum finding, here we do 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 [25, 1], while our infinite-horizon complexity improves the part $\tilde{O}(\frac{\Gamma^3 S\sqrt{A}}{\epsilon})$ from [37], thus leading a new overall complexity of $\tilde{O}(\min\{\frac{\Gamma^{1.5}SA}{\epsilon}, \frac{\Gamma^{2.5}S\sqrt{A}}{\epsilon}\})$.

Our new complexities are therefore a measurable improvement and a new step into the lower bounds $\Omega(\frac{H^{1.5}S\sqrt{A}}{\epsilon})$ and $\Omega(\frac{\Gamma^{1.5}S\sqrt{A}}{\epsilon})$.

References

- [1] 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. 2, 3
- [2] Karl Johan Åström. Optimal control of Markov processes with incomplete state information I. *Journal of mathematical analysis and applications*, 10:174–205, 1965. 1
- [3] Nicole Bäuerle and Ulrich Rieder. *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011. 1
- [4] 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
- [5] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012. 1
- [6] Dimitri Bertsekas. *Abstract dynamic programming*. Athena Scientific, 2022. 1
- [7] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, Sep 2017. 2
- [8] 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*, pages 1–13, 2014. 1
- [9] 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. 2
- [10] João F Doriguello, Alessandro Luongo, and Ewin Tang. Do you know what q-means? *arXiv preprint arXiv:2308.09701*, 2023. 2
- [11] Christoph Dürr and Peter Høyer. A quantum algorithm for finding the minimum. *arXiv preprint quant-ph/9607014*, 1996. 2
- [12] Eugene A Feinberg and Adam Shwartz. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012. 1
- [13] 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. 2
- [14] Qiying Hu and Wuyi Yue. *Markov decision processes with their applications*, volume 14. Springer Science & Business Media, 2007. 1
- [15] Sofiene Jerbi, Arjan Cornelissen, Maris Ozols, and Vedran Dunjko. Quantum Policy Gradient Algorithms. In Omar Fawzi and Michael Walter, editors, *18th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2023)*, volume 266 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 13:1–13:24, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. 2
- [16] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003. 2

- [17] 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. 2
- [18] Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998. 2
- [19] Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998. 2
- [20] 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, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2
- [21] Iordanis Kerenidis and Anupam Prakash. Quantum Recommendation Systems. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 49:1–49:21, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. 2
- [22] 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)*, pages 1186–1215, 2023. 2, 3
- [23] 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, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872. Curran Associates, Inc., 2020. 2
- [24] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum algorithms for supervised and unsupervised machine learning. *arXiv preprint arXiv:1307.0411*, 2013. 2
- [25] 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. 2, 3
- [26] Ashley Montanaro. Quantum speedup of Monte Carlo methods. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2181):20150301, 2015. 2, 3
- [27] Mausam Natarajan and Andrey Kolobov. *Planning with Markov decision processes: An AI perspective*. Springer Nature, 2022. 1
- [28] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010. 2
- [29] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. 1, 2
- [30] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Phys. Rev. Lett.*, 113:130503, Sep 2014. 2
- [31] Renato Cesar Sato and Désirée Moraes Zouain. Markov models in health care. *Einstein (São Paulo)*, 8:376–379, 2010. 1

- [32] 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, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2, 3
- [33] 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, page 770–787, USA, 2018. Society for Industrial and Applied Mathematics. 2
- [34] Lauren N Steimle and Brian T Denton. Markov decision processes for screening and treatment of chronic diseases. *Markov Decision Processes in Practice*, pages 189–222, 2017. 1
- [35] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998. 1
- [36] Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer Cham, 2010. 1
- [37] 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*, pages 10916–10926. PMLR, 2021. 2, 3
- [38] Mengdi Wang. Randomized linear programming solves the discounted Markov decision problem in nearly-linear (sometimes sublinear) running time. *arXiv preprint arXiv:1704.01869*, 2017. 2
- [39] Simon Wiedemann, Daniel Hein, Steffen Udfluft, 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. 2
- [40] 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. 2