Quantum-enhanced Reinforcement Learning

André Sequeira
andresequeira401@gmail.com

Universidade do Minho
Escola de Engenharia

Department of Informatics
University of Minho
Doctoral Program in Informatics (PDInf)

May 14, 2021

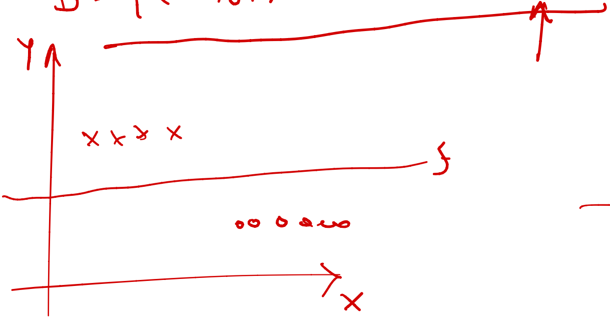
Contents

Contents

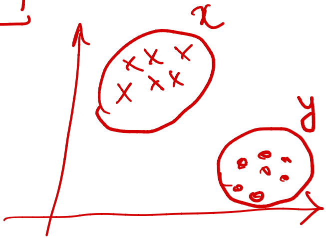
SL vs UL vs RL

SL

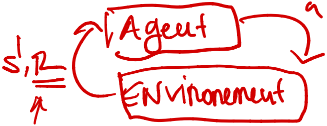
$$D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$



UL

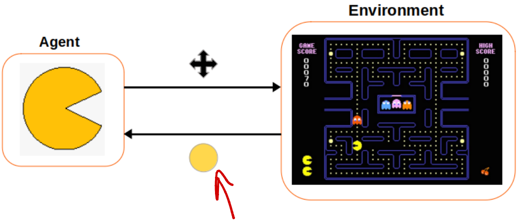


RL



Reinforcement Learning

Agent-Environment Paradigm (A-E Paradigm): Two party system. Learning by interaction.



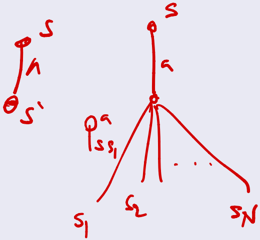
- **Agent**: $\pi : S \mapsto A$
- **Environment**: Characterized by a Markov Decision Process (MDP)

Markov Decision Process

Definition

A *Markov Decision process* is a tuple $\langle S, A, P, R, \gamma \rangle$

- S - Finite set of states
- A - Finite set of actions
- P - state transition probability matrix
- R - Reward function for being in state s
- $\gamma \in [0, 1)$ - Discount factor



	s_1	s_2	s_3	s_4	s_5
(s, a)	0.3	0	0	0.7	0

(a) Stochastic MDP

	s_1	s_2	s_3	s_4	s_5
(s, a)	0	0	0	1	0

(b) Deterministic MDP

MDP's

RL AGENTS GOAL: For a given number of transitions in the environment (*horizon*), the goal of the agent is to find the *optimal policy* π^*, that maximizes the *expected* discounted cumulative reward

$$R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots + \gamma^{h-1} R_{h-1} = \sum_{t=0}^{h-1} \gamma^t R_t$$

- Convergence
- Immediate VS Delayed Rewards

REINFORCE

More on MDP's and RL

MDP: Fully observable (MDP) VS Partially Observable (POMDP)

Solving the MDP for the optimal policy π^*:

Model-based RL:

- The agent knows the dynamics of the environment (P,R)
- Dynamic Programming - Value Iteration / Policy Iteration

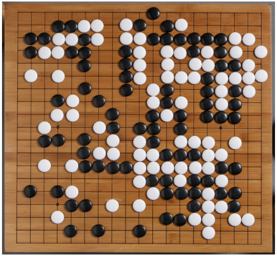
Model-Free RL:

- Unknown dynamics - Resort to Sampling techniques
- *Exploration-Exploitation dillema* ⟶
- MC Learning , TD-Learning (SARSA, Q-Learning) Sutton (2018)

A complicated problem

Problem:

- Solve the optimal policy problem for the entire state-space of the MDP.
- Real world problems - Large State-Space MDPs
- Planning may be intractable

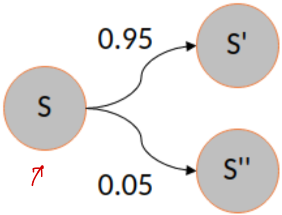


MuZero

A rather simple solution

Approach:

- Perform planning on states that we actually care about
- Not solving for the optimal state action mapping for the entire state space, but only for the most visited states.

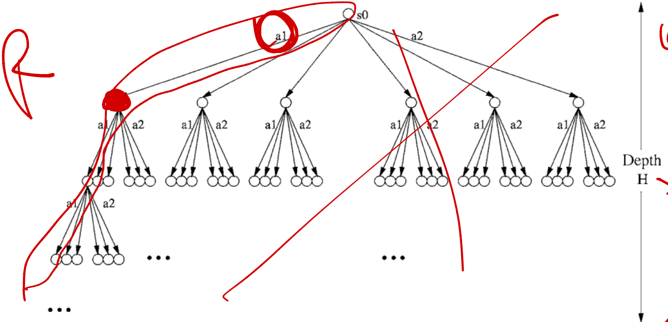


- This will be the approach taken in the quantum setting as well

Sparse Sampling

- From an initial state, we can sample enough trajectories that enables us to decide what action to take at that particular state
- Sample every possible action m times for every $m|A|$ generated states, for a given horizon
- Could be viewed as the agent to be *thinking*

$(m|A|)^h$



$(m|A|)^h << |S|$

effective horizon

γ^t

OpenAI
Deepmind

Figure: Look-ahead tree from Kearns et al. (1999)

Sparse Sampling Complexity

- ϵ-approximation of the optimal action

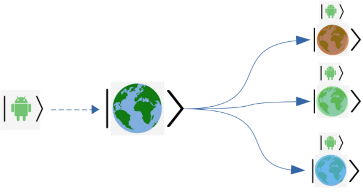
$$\mathcal{O}\left(\left(\frac{|A|H}{\epsilon}\right)^{Hlog(\frac{H}{\epsilon})}\right) \tag{1}$$

- $\ominus$ Complexity **exponential** in the horizon
- $\oplus$ Complexity **independent** of the # of states of the MDP

Contents

Quantum Agent-Environment Paradigm

- We need a notion of a quantum Agent/Environment
- Agent can evolve in parallel in the environment, performing actions in superposition



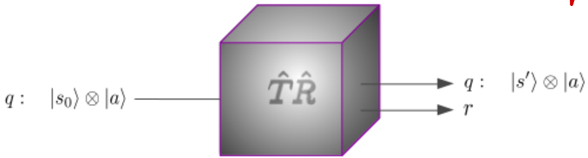
$$|\;\rangle = \alpha|\uparrow\rangle + \beta|\downarrow\rangle + \eta|\leftarrow\rangle + \gamma|\rightarrow\rangle$$

- How to collapse the superposition into something meaningful?

Quantum Agent-Environment Paradigm (2)

γ episodic

h passos



- **qAgent**: $|s\rangle \mapsto |a\rangle$
- **qEnvironment**: Oraculization of task environments Dunjko et al. (2016)
- $\hat{T}$,$\hat{R}$ will be dependent on the nature of the environment itself

State Transition Dynamics

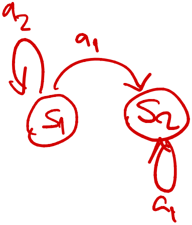
- States and actions are *basis encoded*

State transition operator - T

$$T : |s\rangle \otimes |a\rangle \otimes |0\rangle^{\otimes n_s} \mapsto |s\rangle \otimes |a\rangle \otimes \sum_{s' \in S} \sqrt{P^a_{ss'}} |s'\rangle$$

- The action register is the uniform superposition over the set of admissible actions for a given state, A_s.

$$|a\rangle = \frac{1}{\sqrt{|A_s|}} \sum_{i \in |A_s|} |a_i\rangle \tag{2}$$



Reward Function

- Rewards are *angle encoded*

$|r\rangle = |0\rangle$

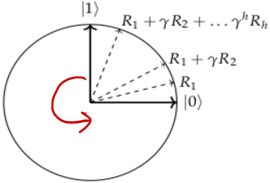
Reward operator - R

$$R : |s\rangle \otimes |r\rangle \mapsto |s\rangle \otimes e^{ir\hat{\sigma_y}}|r\rangle$$

- Why angle encoding?

  Agents goal is to maximize the expected cumulative reward
- Iteratively tweaking the angle, we can sum rewards essentially for free.

$$R_y(r_1)R_y(r_0)|r\rangle = R_y(r_1)R_y(r_0)|0\rangle = cos(r_0 + r_1)|0\rangle + sin(r_0 + r_1)|1\rangle$$



$$\begin{pmatrix} \cos\theta & -\text{sen}\,\theta \\ \text{sen}\,\theta & \cos\theta \end{pmatrix}$$

$$\mathcal{R} = \sum_{t=0}^{H-1} \gamma^t R_t \leq \frac{\pi}{2}$$

Maximum Expected Cumulative Reward

The oracularized environment, O, will be the product of the State transition and reward operators acting on the respective transition step quantum registers

$$O = \prod_{i=0}^{H-1} R_i T_i \tag{3}$$

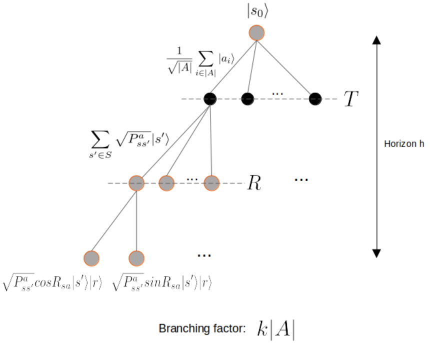
$$O|\psi_0\rangle = \sum_{s*} \sqrt{P^a_{s_0 s_1} P^a_{s_1 s_2} \dots P^a_{s_{H-1} s_H}} . \mathcal{R}|\psi\rangle$$

- Interacting with the quantum environment for h steps, creates superpositions with *approximate* expected utility of each action encoded into the amplitude
- Superposition term with highest amplitude corresponds to the optimal action to take

Contents

Quantum Sparse Sampling

- Applying the oracles for an horizon h is the same as computing a lookahead tree with depth h



- Small/Null rewarded sequence of actions maximize cosine term of reward $\mapsto$ Measuring state does not guarantee optimal action

Amplitude Amplification

Let $P = P^{a}_{s_0 s_1} P^{a}_{s_1 s_2} \dots P^{a}_{s_{H-1} s_H}$

- Amplify amplitude of good states $\mapsto |r\rangle = |1\rangle$

$$|\psi\rangle = \sqrt{P} cos(\mathcal{R})|0\rangle + \sqrt{P} sin(\mathcal{R})|1\rangle$$
$$\hat{\sigma}_z|\psi\rangle = \sqrt{P} cos(\mathcal{R})|0\rangle - \sqrt{P} sin(\mathcal{R})|1\rangle = |\psi'\rangle$$

- For j iterations of the Grover Operator, $\mathcal{G}$

$$\begin{aligned}\mathcal{G}^j|\psi\rangle &= \left[(2|\psi\rangle\langle\psi| - \mathbb{1})\hat{\sigma}_z\right]^j |\psi\rangle \\ &= \sqrt{P} cos((2j+1)\mathcal{R})|0\rangle + \sqrt{P} sin((2j+1)\mathcal{R})|1\rangle\end{aligned}$$

- How many iterations?

Exponential Search

$O(\sqrt{N})$ $O\left(\sqrt{\frac{N}{h}}\right)$

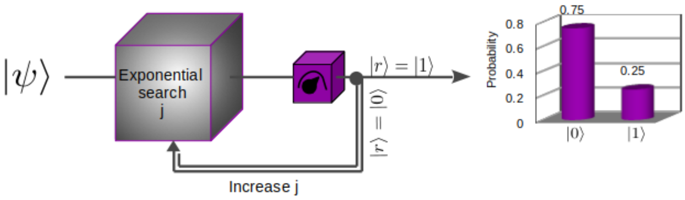
Problem:

- Initial distribution is **non**-uniform
- Unknown number of superposition terms
- Unknown number of marked states
- Measuring a good state, does not guarantee optimal action

Solution:

- Perform Exponential Search Boyer et al. (1998)
- Exponentially increase the number of Grover Iterations

Exponential Search



- Sampling the state, we achieve a distribution from which we can extract the optimal action to take
- How many samples ?

Contents

Complexity will be dictated by two separate components:

- Number of samples, $\mathcal{S}$
- Runtime per sample , C

Complexity

For any initial state $s \in S$, the algorithm computes an ϵ-approximation of the optimal action with complexity:

$$\mathcal{S} \times C$$

Runtime per sample

The complexity of each execution will be dominated by the runtime of the exponential search algorithm.

$$\mathcal{O}\left(\sqrt{\frac{N}{n}}\right)$$

- Worst-case scenario: Single marked state, $n = 1$
- Search space will be dependent on the dynamics of the environment, k and the branching factor $|A|$.

Runtime per sample (2)

$2^h \rightarrow O(|A|^h)$

$(k.|A|)^h$

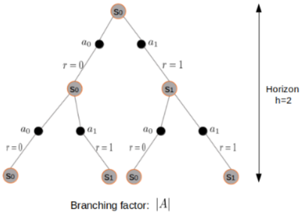


Figure: Deterministic MDP

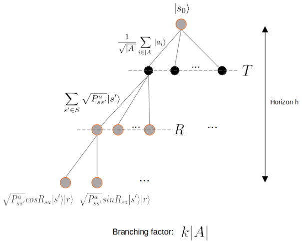


Figure: Stochastic MDP

$$C = \mathcal{O}\left(\sqrt{\frac{N}{n}}\right) = \mathcal{O}\left(\sqrt{k|A|^h}\right) \quad (4)$$

Number of samples

We can estimate the number of samples needed to estimate the probability of measuring a single qubit basis states $\{|0\rangle, |1\rangle\}$, using the *Wilson Interval* Schuld et al. (2018):

$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$

$$\mathcal{S} = \mathcal{O}\left(\frac{\sigma^2}{8\epsilon^2}(\sqrt{16\epsilon^2+1}+1)\right)$$

$$\begin{cases} \alpha' - \alpha = \varepsilon \\ \beta' - \beta = \varepsilon \end{cases}$$

- For an action register with $log|A|$ qubits:

$$\mathcal{S} = \mathcal{O}\left(\frac{\sigma^2}{8\epsilon^2}log|A|(\sqrt{16\epsilon^2+1}+1)\right)$$

where σ is the sample confidence interval and ϵ is the prediction associated error.

Complexity Analysis

Complexity

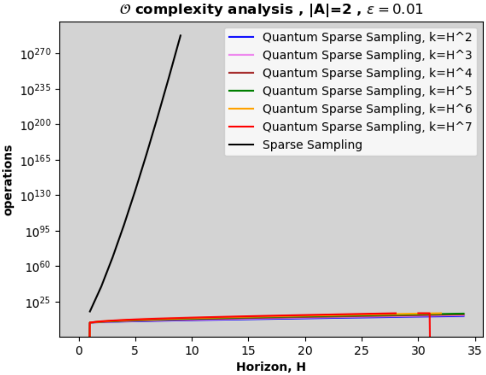
For any initial state $s \in S$, the algorithm computes an ϵ-approximation of the optimal action with complexity:

$$\mathcal{O}\left(\frac{\sigma^2}{8\epsilon^2} log|A|(\sqrt{16\epsilon^2+1}+1)\sqrt{k|A|^h}\right)$$

Without further assumptions on the environment dynamics, we cannot say anything about k

Complexity Separation

Varying k exponentially with the horizon. Binary action MDP with $\sigma = 99\%$ and $\epsilon = 1\%$



Contents

Non-uniform tree expansion

GROVER

$\frac{N}{4}$

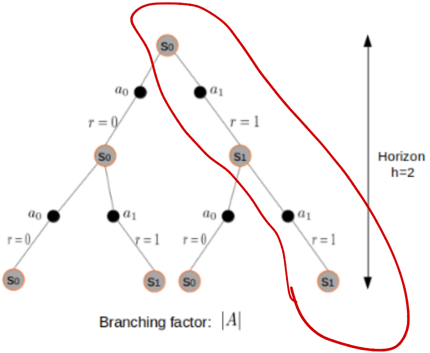


Figure: Deterministic MDP

Oraculo

25%

- Exploit information to reduce search space
- Informed tree search

- Variational Circuits as Policy generators
- Hybrid algorithms - Classical optimization
- Supervised Learning with Quantum Computers - Schuld and Petruccione

Agent

s, a

s, r

Envin

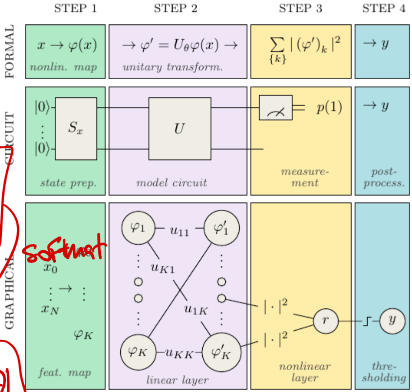
s = s'

softmax

ENviron

s, r

$L(\theta) = R \cdot \pi(\theta)$



$R_x(\theta)$

$R_y(\theta)$

softmax

π

$\pi^* \leftarrow \pi$

Bibliography I

Michel Boyer, Gilles Brassard, Peter HØyer, and Alain Tapp. Tight bounds on quantum searching. *Fortschritte der Physik*, 46(4-5):493–505, 1998. ISSN 00158208. doi: 10.1002/(SICI)1521-3978(199806)46:4/5⟨493::AID-PROP493⟩3.0.CO;2-P.

Vedran Dunjko, Jacob M. Taylor, and Hans J. Briegel. Quantum-Enhanced Machine Learning. *Physical Review Letters*, 117(13):1–19, 2016. ISSN 10797114. doi: 10.1103/PhysRevLett.117.130501.

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. In *IJCAI International Joint Conference on Artificial Intelligence*, 1999.

Maria Schuld and Francesco Petruccione. *Quantum Science and Technology Supervised Learning with Quantum Computers*. ISBN 9783319964232. URL http://www.springer.com/series/10039.

Bibliography II

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware, 2018. ISSN 23318422.

Sutton. *Reinforcement Learning Book*. 2018. ISBN 9780262039246.

Quantum-enhanced Reinforcement Learning

André Sequeira
andresequeira401@gmail.com

Universidade do Minho
Escola de Engenharia

Department of Informatics
University of Minho
Doctoral Program in Informatics (PDInf)

May 14, 2021