Optimal Policies for Quantum Markov Decision Processes

量子马尔可夫决策系统的优化策略

Ming-Sheng Ying, Yuan Feng, Sheng-Gang Ying

International Journal of Automation and Computing, vol. 18, no. 3, pp.410-421, 2021.

yingmsh@tsinghua.edu.cn, Yuan.Feng@uts.edu.au, yingsg@ios.ac.cn

Markov decision process (MDP) offers a general framework for modelling sequential decision making where outcomes are random. In particular, it serves as a mathematical framework for reinforcement learning. This paper introduces an extension of MDP, namely quantum MDP (qMDP), that can serve as a mathematical model of decision making about quantum systems. We develop dynamic programming algorithms for policy evaluation and finding optimal policies for qMDPs in the case of finite-horizon. The results obtained in this paper provide some useful mathematical tools for reinforcement learning techniques applied to the quantum world.

Basic definitions

Definition 1. *A* qMDP is a 7-tuple

$$\mathcal{P} = (\mathcal{T}, \mathcal{H}, \rho, \mathcal{A}, \{\mathcal{E}_t(\cdot|a) : t \in \mathcal{T}, a \in \mathcal{A}\}, \mathcal{M}, \{r_t : t \in \mathcal{T}\})$$

where:

1) $\mathcal{T} = \{1, 2, \cdots, N\}$ is the set of decision epochs.

2) $\mathcal{H} = \mathbf{C}^n$ is the state space of an n-level quantum system.

3) ρ is a density matrix in $\mathcal{H}$, called the starting state.

4) $\mathcal{A}$ is a set of action names.

5) For each $t \in \mathcal{T}$ and $a \in \mathcal{A}$, $\mathcal{E}_t(\cdot|a)$ is a super-operators in $\mathcal{H}$.

6) $\mathcal{M}$ is a set of quantum measurements in $\mathcal{H}$. We write:

$$\mathcal{O} = \bigcup_{M \in \mathcal{M}} [\{M\} \times O(M)] .$$

7) For each $1 \le t \le N-1$, $r_t : \mathcal{O} \times \mathcal{A} \to \mathbf{R}$ (real numbers) is the reward function at decision epoch t, and $r_N : \mathcal{O} \to \mathbf{R}$ is the reward function at the final decision epoch N.

Definition 2. *Let* $1 \le t \le N$. Then a sequence

$$h_t = (M_1, m_1, a_1, \cdots, M_{t-1}, m_{t-1}, a_{t-1}, M_t, m_t)$$

is called a history of t epochs if $(M_1, m_1), \cdots, (M_{t-1}, m_{t-1}), (M_t, m_t) \in \mathcal{O}$ and $a_1, \cdots, a_{t-1} \in \mathcal{A}$.

History h_t records the activities of the decision maker: For each $j \le t$, she/he performed measurement M_j on the system, got outcome m_j, and then took action a_j on it. It is assumed that measurement M_j happened before action a_j. If a_j was taken before M_j, then the result would be different because a measurement usually changes the state of a quantum system. We write $tail(h_t) = (M_t, m_t)$. The set of histories of t epochs is denoted H_t. Obviously, if $h_t \in H_t$, $a_t, a_{t+1}, \cdots, a_{t+(k-1)} \in \mathcal{A}$ and (M_{t+1}, m_{t+1}), $(M_{t+2}, m_{t+2}), \cdots, (M_{t+k}, m_{t+k}) \in \mathcal{O}$, then

$$(h_t, a_t, M_{t+1}, m_{t+1}, a_{t+1}, M_{t+2}, m_{t+2}, ..., a_{t+(k-1)}, M_{t+k}, m_{t+k}) \in H_{t+k}$$

for $1 \le k \le N-t$.

Definition 3. *A*randomised history-dependent policy is a sequence $\pi = (\alpha_0, \beta_1, \alpha_1, \cdots, \beta_{N-1}, \alpha_{N-1})$, where:

1) $\alpha_0 \in \mathcal{D}(\mathcal{M})$.

2) $\alpha_t : H_t \times \mathcal{A} \to \mathcal{D}(\mathcal{M})$ for $t = 1, \cdots, N-1$.

3) $\beta_t : H_t \to \mathcal{D}(\mathcal{A})$ for $t = 1, \cdots, N-1$.

For each $M \in \mathcal{M}$, $\alpha_0(M)$ is the probability that M is chosen at the beginning of the decision process. For each $1 \le t \le N-1$, $h_t \in H_t$, $a \in \mathcal{A}$ and $M \in \mathcal{M}$, $\beta_t(h_t)(a)$ is the probability that action a is chosen to take between decision epoch t and $t+1$ given history h_t, and $\alpha_t(h_t, a)(M)$ is the probability that measurement M is chosen to perform at epoch $t+1$ given history h_t and that action a was taken between epoch t and $t+1$. In particular, π is a deterministic (history-dependent) policy if α_0, $\alpha_t(h_t, a)$ and $\beta_t(h_t)$ are all single-point distributions; that is, $\alpha_0 \in \mathcal{M}$, and

$$\alpha_t : H_t \times \mathcal{A} \to \mathcal{M},\ \beta_t : H_t \to \mathcal{A}$$

for $t = 1, \cdots, N-1$.

Policy evaluation

As in the case of MDPs, a direct computation of the reward in a qMDP based on defining emuation (6) is very inefficient. In this section, we establish a backward recursion for the reward function so that dynamic programming can be used in policy evaluation for qMDPs. To this end, we first introduce a conditional probability function. Let π be a randomised history-dependent policy, $1 \le t \le N$ and

$$h_t = (M_1, m_1, a_1, \cdots, M_{t-1}, m_{t-1}, a_{t-1}, M_t, m_t) \in H_t$$
$$f_t = (a_t, M_{t+1}, m_{t+1}, \cdots, a_{N-1}, M_N, m_N) \in (\mathcal{A} \times \mathcal{O})^{N-t}.$$

Using the conditional probability function $p^\pi(\cdot|h_t)$, we can compute the expected reward in the tail of a decision process. More precisely, for each randomised history-dependent policy π, function

$$u_t^\pi : H_t \to R$$

is defined to be the expected total reward obtained by using policy π at decision epochs $t, t+1, \cdots, N$; i.e., for every $h_t \in H_t$,

$$u_t^\pi(h_t) = \sum_{f_t \in (\mathcal{A} \times \mathcal{O})^{N-t}} p^\pi(f_t|h_t) \times r(f_t) \tag{9}$$

where

$$r(f_t) = \sum_{j=t}^{N-1} r_j(M_j, m_j, a_j) + r_N(M_N, m_N).$$

Theorem 1. (Backward Recursion) For each $1 \le t \le N-1$, we have:

$$u_t^\pi(h_t) = \sum_{a_t \in \mathcal{A}} \sum_{M_{t+1} \in \mathcal{M}} \beta_t(h_t)(a_t) \times \alpha_t(h_t, a_t)(M_{t+1}) \times \Big[r_t(M_t, m_t, a_t) + \sum p_{t+1} \times u_{t+1}^\pi(h_t, a_t, M_{t+1}, m_{t+1}) \Big] \tag{10}$$

where the third $\sum$ is over $m_{t+1} \in O(M_{t+1})$.

Optimality of policies

Now we turn to consider how to compute optimal policies. The optimal expected total reward over the decision making horizon is defined by

$$v_N^* = \sup_\pi v_N^\pi .$$

Theorem 2. (The principle of optimality) Let $u_t : H_t \to R (t = 1, \cdots, N)$ be a solution of the optimality equations (13) and (14). Then,

$$u_t(h_t) = u_t^*(h_t)$$

for all $t = 1, \cdots, N$ and $h_t \in H_t$.