ON OPTIMAL Q-VALUE FUNCTIONS FOR DEC-POMDP

Shuo Liu
Computer Science
Northeastern University
shuo.liu2@northeastern.edu

ABSTRACT

This article discusses the optimal Q-value function definition in Dec-POMDP.

1 NOTIONS

s^t	the state at t with problem horizon h
o^t	the joint observation of agents $o^t = \langle o_1^t, \ldots, o_n^t \rangle$ at t
$\mathcal{O}$	the joint observation space
$\vec{\theta}^t$	the joint observation-action history until t, $\vec{\theta}^t = (o^0, a^0, \ldots, o^t)$
$\vec{\Theta}^t$	the joint observation space
$\vec{\Theta}_\pi^t$	the set of $\vec{\theta}^t$ consistent with policy π
δ^t	the decision rule (a temporal structure of policy) at t
$\delta^{t,*}$	the optimal decision rule at t following $\psi^{t-1,*}$
$\delta_\psi^{t,\circledast}$	the optimal decision rule at t following ψ^{t-1}
Δ^t	the decision rule space at t
ψ^t	the past joint policy until t, $\psi^t = \delta^{[0,t)}$
$\psi^{t,*}$	the optimal past joint policy until t, $\psi^{t,*} = \delta^{[0,t),*}$
$\psi^{t,\circledast}$	the past joint policy until t with non-optimal ψ^{t-1} and optimal $\delta_\psi^{t-1,\circledast}$
Ψ^t	the past joint policy space at t
ξ^t	the subsequent joint policy from t, $\xi^t = \delta^{[t,h)}$
$\xi^{t,*}$	the optimal subsequent joint policy from t, $\xi^t = \delta^{[t,h),*}$
$\xi_\psi^{t,\circledast}$	the optimal subsequent joint policy from t following non-optimal ψ^t
π	the joint pure policy $\pi = \delta^{[0,h)}$
π^*	the joint optimal pure policy $\pi^* = \delta^{[0,h),*}$
$R(\vec{\theta}^t, \psi^{t+1})$	the immediate reward function following ψ^{t+1}
$Q(\vec{\theta}^t, \psi^{t+1})$	the history-policy value function following ψ^{t+1}
$Q^*(\vec{\theta}^t, \psi^{t+1})$	the optimal history-policy value function following ψ^{t+1}
$Q^\circledast(\vec{\theta}^t, \psi^{t+1})$	the sequentially rational optimal history-policy value function following ψ^{t+1}

2 NORMATIVE OPTIMAL Q-VALUE FUNCTION

Definition 1. *The optimal Q-value function Q^* in Dec-POMDP, the expected cumulative reward over time steps $[t, h)$ induced by optimal joint policy π^*, $\forall \vec{\theta}^t \in \vec{\Theta}^t_{\psi^{t,*}}, \forall \psi^{t+1} \in (\psi^{t,*}, \Delta^t)$, is defined as,*

$$Q^*(\vec{\theta}^t, \psi^{t+1}) = \begin{cases} R(\vec{\theta}^t, \psi^{t+1}), & t = h-1 \\ R(\vec{\theta}^t, \psi^{t+1}) + \sum_{o^{t+1} \in \mathcal{O}} P(o^{t+1}|\vec{\theta}^t, \psi^{t+1}) Q^*(\vec{\theta}^{t+1}, \pi^*(\vec{\theta}^{t+1})). & 0 \leqslant t < h-1 \end{cases} \tag{1}$$

Here, $\pi^*(\vec{\theta}^{t+1}) \equiv \psi^{t+2,*}$ because of the consistent optimality of policy.

Proposition 1. *In Dec-POMDP, deriving an optimal policy from the normative optimal history-policy value function defined in Equ. 1 is impractical (clarifying Sec. 4.3.3, Oliehoek et al. (2008)).*

Proof. We check the optima in 2 steps. The independent and dependent variables are marked in red.

To calculate the Pareto optima of Bayesian game at t,

$$\delta^{t,*} = \operatorname*{argmax}_{\delta^t} \sum_{\vec{\theta}^t \in \vec{\Theta}^t_{\psi^{t,*}}} P(\vec{\theta}^t|\psi^{t,*}) Q^*(\vec{\theta}^t, (\psi^{t,*}, \delta^t)), \tag{2}$$

note that calculating $\delta^{t,*}$ depends on $\psi^{t,*} = \delta^{[0,t),*}$ and $Q^*(\vec{\theta}^t, \cdot)$.

According to Definition. 1, the optimal Bellman equation can be written as,

$$Q^*(\vec{\theta}^t, \psi^{t+1}) = R(\vec{\theta}^t, \psi^{t+1}) + \sum_{o^{t+1} \in \mathcal{O}} P(o^{t+1}|\vec{\theta}^t, \psi^{t+1}) \max_{\delta^{t+1}} Q^*(\vec{\theta}^{t+1}, (\psi^{t+1,*}, \delta^{t+1})), \tag{3}$$

when $0 \leqslant t < h-1$. This indicates that $Q^*(\vec{\theta}^t, \cdot)$ depends on $\psi^{t+1,*}$.[1] Consequently, calculating $\delta^{t,*}$ inherently depends on $\delta^{[0,t],*}$ (includes itself), making it self-dependent and impractical to solve.[2] □

3 SEQUENTIALLY RATIONAL OPTIMAL Q-VALUE FUNCTION

To make optimal Q-value in Dec-POMDP computable, Oliehoek et al. (2008) defined another form of Q-value function and eliminated the dependency on past optimality.

Definition 2. *The sequentially rational optimal Q-value function $Q^{\circledast}$ in Dec-POMDP, the expected cumulative reward over time steps $[t, h)$ induced by optimal subsequent joint policy $\xi^{t,\circledast}_{\psi}$, $\forall \vec{\theta}^t \in \vec{\Theta}^t_{\Psi^t}, \forall \psi^{t+1} \in \Psi^{t+1}$, is defined as,*

$$Q^{\circledast}(\vec{\theta}^t, \psi^{t+1}) = \begin{cases} R(\vec{\theta}^t, \psi^{t+1}), & t = h-1 \\ R(\vec{\theta}^t, \psi^{t+1}) + \sum_{o^{t+1} \in \mathcal{O}} P(o^{t+1}|\vec{\theta}^t, \psi^{t+1}) Q^{\circledast}(\vec{\theta}^{t+1}, \psi^{t+2,\circledast}), & 0 \leqslant t < h-1 \end{cases} \tag{4}$$

where $\psi^{t+2,\circledast} = (\psi^{t+1}, \delta^{t+1,\circledast}_{\psi}), \forall\, \psi^{t+1} \in \Psi^{t+1}$.

Note that the only difference of $Q^{\circledast}$ from Q^* is $\psi^{t+2,\circledast}$, consequently expanding Q^*'s candidates of history from $\vec{\theta}^t \in \vec{\Theta}^t_{\psi^{t,*}}$ to $\vec{\theta}^t \in \vec{\Theta}^t_{\Psi^t}$ and policy from $\psi^{t+1} \in (\psi^{t,*}, \Delta^t)$ to $\psi^{t+1} \in (\Psi^t, \Delta^t)$.

Beyond solving the problem of Proposition 1, another advantage of $Q^{\circledast}$ is that it allows for the computation of optimal subsequent policy $\xi^{t,*}_{\psi}$ following any past policy ψ^t. This is beneficial in online applications where agents may occasionally deviate from the optimal policy.

[1] The dependency of $P(o^{t+1}|\vec{\theta}^t, \psi^{t+1})$ is not a problem and can be solved just like how the stochasticity $P(s^{t+1}|s^t, a)$ tackled by double learning in Sec. 6.7, Sutton & Barto (2018).

[2] Single-agent MDP (even partially observable) does not have such a problem because the Q-value function is not history-dependent, thanks to Markovian property.

4 Open Questions

- We have seen some advantages of defining the optimal Q-value function as $Q^{\circledast}$, what are the downsides to defining it this way (e.g., high computational costs)?

References

Frans A. Oliehoek, Matthijs T. J. Spaan, and Nikos Vlassis. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, 32:289–353, 2008.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.