Reinforcement Learning

Comparison between supervised learning and reinforcement learning :

- supervised learning learns from labelled dataset, which is a set of (usually independent) input-output pairs
  reinforcement learning learns from a simulator, which is sequential interaction between agent and environment
- supervised learning is regression mathematically
  reinforcement learning is dynamic programming mathematically

Bellman equation and pricing tree

The principal of optimality states that the maximum reward can be obtained by :

- behaving optimally for one step
- behaving optimally for the rest, which is defined by value function at the state you end up in

Different favors of Bellman equation :

- Bellman equation for value function (MRP)
- Bellman expectation equation for state-value function (MDP)
- Bellman expectation equation for action-value function (MDP)
- Bellman optimality equation for state-value function (MDP)
- Bellman optimality equation for action-value function (MDP) $\rightarrow$ identical to shortest path, knight/queen move problem

$v_\pi(s)$	=	$\sum_{a\in A}[\pi(a\mid s)q_\pi(s,a)]$	depends on agent's policy (no discount in this half-step)
$q_\pi(s,a)$	=	$R_{sa}+\gamma\sum_{s'\in S}P_{sas'}v_\pi(s')$	depends on environment's transition and reward function (with discount in this half-step)
			this is the same as the linear Bellman in MRP

Mountain car
Random walk
Puddle world
Cart and pole

Feature vector, state aliasing
- partially obserable MDP or
- equivalently selected features limited the view of this world (i.e. observation of state

Connection between critic actor vs Generative adversarial network

Beta distribution
Conjugate gradient
Prior prob vs posterior prob
Alpha beta search truncating search tree

Combining the two half-steps, we have the Bellman for MDP, which can be written in either v-domain or q-domain :

1.
$$
\begin{aligned}
v_\pi(s) &= \textstyle\sum_{a\in A}[\pi(a|s)q_\pi(s,a)] \\
&= \textstyle\sum_{a\in A}[\pi(a|s)(R_{sa}+\gamma\sum_{s'\in S}P_{sas'}v_\pi(s'))] && \text{Bellman of value function for MDP} \\
&= \underbrace{\textstyle\sum_{a\in A}[\pi(a|s)R_{sa}]}_{R_{s\pi}}+\gamma\textstyle\sum_{s'\in S}[\underbrace{\textstyle\sum_{a\in A}[\pi(a|s)P_{sas'}]}_{P_{s\pi s'}}v_\pi(s')] \\
&= R_{s\pi}+\gamma\textstyle\sum_{s'\in S}[P_{s\pi s'}v_\pi(s')]
\end{aligned}
$$

where
$$
\begin{aligned}
R_{s\pi} &= \textstyle\sum_{a\in A}[\pi(a|s)R_{sa}] \\
P_{s\pi s'} &= \textstyle\sum_{a\in A}[\pi(a|s)P_{sas'}]
\end{aligned}
$$

2.
$$
\begin{aligned}
q_\pi(s,a) &= R_{sa}+\gamma\textstyle\sum_{s'\in S}P_{sas'}v_\pi(s') \\
&= R_{sa}+\gamma\textstyle\sum_{s'\in S}[P_{sas'}\sum_{a'\in A}\pi(a'|s')q_\pi(s',a')] && \text{Bellman of action-value function for MDP}
\end{aligned}
$$

For optimum policy (there always exists a deterministic optimal policy for all MDP), we have :

$$
\pi_{opt}(a|s) = \begin{cases} 1 & if \quad a=\arg\max_{a\in A} q_{opt}(s,a) \\ 0 & otherwise \end{cases} \qquad \text{theorem : there always exists a optimal deterministic policy}
$$

Therefore the <u>Bellman equation</u> in value domain and action-value domain becomes :

1.
$$
\begin{aligned}
v_{opt}(s) &= \textstyle\sum_{a\in A}[\pi_{opt}(a|s)q_{opt}(s,a)] \\
&= \textstyle\sum_{a\in A}[1_{a=\arg\max_{a\in A} q_{opt}(s,a)}\times q_{opt}(s,a)] \\
&= \max_{a\in A} q_{opt}(s,a) && \text{Bellman equation expresses v in terms of expectation of q.}
\end{aligned}
$$

2.
$$
q_{opt}(s,a) = R_{sa}+\gamma\textstyle\sum_{s'\in S}P_{sas'}v_{opt}(s') \qquad \text{Bellman optimality equation expresses v in terms of maximum of q.}
$$

Finally we have a <u>Bellman optimality equation</u> in value domain and action-value domain, which are both non-linear :

$$
\begin{aligned}
v_{opt}(s) &= \max_{a\in A} q_{opt}(s,a) \\
&= \max_{a\in A}[R_{sa}+\gamma\textstyle\sum_{s'\in S}P_{sas'}v_{opt}(s')] \\
q_{opt}(s,a) &= R_{sa}+\gamma\textstyle\sum_{s'\in S}P_{sas'}v_{opt}(s') \\
&= R_{sa}+\gamma\textstyle\sum_{s'\in S}[P_{sas'}\max_{a\in A} q_{opt}(s',a)]
\end{aligned}
$$

Graphical representation of MDP

$$
G = (S,A,P,\pi) \quad \text{such that } \forall s,a\in P,\ s\in S \text{ and } a\in A
$$
$$
\text{whereas } \forall s,a\in\pi,\ s\in S \text{ and } a\in A
$$

Notation for MRP and MDP

		graph nodes	value at nodes	environment	agent decision
• $v_\pi(s) = R_s+\gamma\sum_{s'\in S}P_{ss'}v_\pi(s')$	MRP	s	v	R_s $P_{ss'}$	
• $v_\pi(s) = \sum_{a\in A}[\pi(a|s)q_\pi(s,a)]$	MDP	s, a	v_π, q_π	R_{sa} $P_{sas'}$	$\pi(a|s)$
$q_\pi(s,a) = R_{sa}+\gamma\sum_{s'\in S}P_{sas'}v_\pi(s')$			(with π index)	(with a index, hence action matters)	

Intuition of Bellman equation
- optimal decision is the optimal combo of this step and future optimality