Problem set 2

$R: S \times A \times S \to \mathbb{R}$.

MDP: $\langle S, A, R, T, \gamma \rangle$. $M' = \langle S, A, R', T, \gamma \rangle$

where $R'(s,a,s') = R(s,a,s') + F(s,a,s')$

$F: S \times A \times S \to \mathbb{R}$ = feedback function.

$F(s,a,s') = \gamma\phi(s') - \phi(s)$; $\phi: S \to \mathbb{R}$;

$\forall (s,a,s') \in S \times A \times S$.

① $Q^*_M, Q^*_{M'}$

$$Q^*_{M'}(s,a) - \phi(s) = Q^*_M(s,a)$$

$$Q^*_{M'}(s,a) = \max_a R'(s,a,s') + \gamma \sum P(s'|s,a) V(s')$$

$$= \max_a (R(s,a) + F(s,a,s'))$$

$$Q^*_{M'} = R'(s,a,s') + \gamma \max_{a'} Q^*_{M'}(s',a')$$

$$= R(s,a,s') + F(s,a,s') + \gamma \max_{a'} Q_{M'}(s',a')$$

$$= R(s,a,s') + \gamma\phi(s') - \phi(s) + \gamma \max Q_{M'}(s',a')$$

$$= R(s,a,s')$$

$$Q^*_{M'} + \phi(s) = R(s,a,s') + \gamma \max [Q^*_{M'}(s',a') + \phi(s')]$$

$$= R(s,a,s') + \gamma \max Q_M(s',a') = Q_M(s,a)$$

$$\pi^*_{M'}(s) = \arg\max_{a'} Q^*_{M'}(s,a) = \arg\max Q^*_M(s,a) - \phi(s)$$

$$= \arg\max Q^*_{M^*}(s,a)$$

$\rightsquigarrow \pi^*_{M'}(s) = \pi^*_M(s)$.

(2) $Q^0_\mu(s,a)$ & $Q^0_{\mu'}(s,a)$; $q_{init} \in \mathbb{R}$.

$Q^0_\mu(s,a) = q_{init} + \phi(s)$; $Q^0_{\mu'}(s,a) = q_{init}$
$= Q^0_\mu(s,a) - \phi(s)$

$Q_M(s,a) = Q^0_\mu(s,a) + \Delta Q_\mu(s,a)$

$Q'_\mu(s,a) = Q^0_{\mu'}(s,a) + \Delta Q'_\mu(s,a)$

$\Delta Q_\mu(s,a) = \Delta Q'_\mu(s,a)$

$\Rightarrow Q_\mu(s,a) = Q'_\mu(s,a) + \phi(s)$

$Q_\mu(s,a) = R(s,a,s') + \gamma \max_{a \in A'} Q_\mu(s',a')$.

$= R(s,a,s') + \gamma \max_{a \in A'} \left[Q'_\mu(s',a') + \phi(s') \right]$

$\Rightarrow \underline{\text{argmax } Q_\mu(s,a) \equiv \text{argmax } Q'_\mu(s,a)} \Rightarrow$ update coincides.

Review policy update process:

$\pi(s) = 0$
while $\pi(s)$ and $\pi'(s) \leq \varepsilon$:
- Policy Eval (π) $\Rightarrow$ get V; Q; ...
- Policy Update $\Rightarrow$ update using policy Iteration
$\Rightarrow \pi = \text{argmax}_a Q^{\pi_i}(s,a) \geq Q^{\pi_i}(s,a) = \pi_i(s)$.

Policy Evaluation:
- Analytical Method (know model)
- Monte carlo Method (doesn't know model but can sample).
- Temporal Difference.

$V^\pi_t = \alpha V^\pi_t + \alpha(R_t + \gamma V^\pi_{t+1} - V^\pi_t)$

$= \alpha(R + \gamma V(s_{t+1})) + (1-\alpha) V(s_t)$

SARSA
Q-Learning :
Other sampling methods.

⊕ Also use sampling.
⊕ Same with $Q(s,a)$: $S \times A \to \mathbb{R}$.

With exploration: $\pi(a|s) = \begin{cases} 1-\varepsilon & \text{argmax} \\ \frac{\varepsilon}{|A|-1} & \text{otherwise.} \end{cases}$

$\Rightarrow$ policy update with probability.

CamScanner