Terminal Estimation

willer

June 8, 2020

In general cases:

$$Q(s,a) = \sum_{s' \neq \widetilde{s}} p(s'|s,a)V(s') + p(\widetilde{s}\,|s,a)V(\widetilde{s})$$

where $\widetilde{s}$ means terminal state
so we can set:

$$\begin{aligned}
\hat{Q}(s_t,a_t) &= \sum_{s' \neq \widetilde{s}} p(s'|s,a)V(s') \\
f(s_t,a_t) &= p(\widetilde{s}\,|s,a)V(\widetilde{s}) \\
V(s_t) &= \sum_{a} \pi(a_t|s_t)Q(s_t,a_t) \\
Q(s_t,a_t) &= \hat{Q}(s_t,a_t) + f(s_t,a_t)
\end{aligned}$$

In Network, all output is random and doesn't equal zero. for some cases: if $p(s'|s,a) = 0$, $f(s,a)$ should be zero always and if $p(s'|s,a) = 1$, $\hat{Q}(s,a)$ should be zero always, too.

so for all cases

$$\begin{aligned}
\hat{Q}(s_t,a_t) &= \alpha * \hat{Q}(s_t,a_t) + (1-\alpha) * (R_{t+1} - \gamma Q(s_{t+1},a_{t+1})) \\
f(s_t,a_t) &= \beta * f(s_t,a_t) + (1-\beta) * \widetilde{R}_{t+1}
\end{aligned}$$

It has a shrinking effect in network and correct the former bias. And it works well in experiment.