Markov Decision Processes (MDP)

Finite state space S, $|S| < \infty$

" action space A, $|A| < \infty$

$P_h(\cdot \mid s, a) \in \Delta_S \quad \forall s, a, h$

$r_h(s, a) \in [0,1] \quad \forall s, a, h$

Horizon $H < \infty$

Agent starts at $h=1$ in state $s_1 \in S$.

Agent takes action $a_h \in A$

Nature transports agent to state $s_2 \sim P_1(\cdot \mid s_1, a_1)$

and recieves reward $r_1(s_1, a_1)$

Repeat for $h = 2, 3, \ldots, H$.

We are given a set of policies Π where

$\pi \in \Pi$ defines $\pi_h(s) \in A \quad \forall s, h$.

$$V^{\pi} = \mathbb{E}\left[\sum_{h=1}^{H} r_h(s_h, \pi_h(s_h)) \mid s_{h+1} \sim P_h(\cdot \mid s_h, \pi(s_h)) \right].$$

$$Q_h^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \mid s_h = s, a_h = s, a_t = \pi_t(s_t) \ \forall t \right]$$

Theorem $\forall s, a, h$ define

$$Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a).$$

There exists $Q_h(s,a)$ s.t. $Q = Q^*$ if and only if

$$Q_h(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} Q_{h+1}(s',a')\right].$$

where $Q_{H+1}(s,a) = 0$. Furthermore $\pi_h(s) = \arg\max_a Q_h(s,a)$ is optimal.

How do you find Q^*?

$\Rightarrow$ $Q_H^*(s,a) = r_h(s,a)$.

for $h = H-1, H-2, \ldots$

$$Q_h^* = r_h(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} Q_{h+1}^*(s,a)\right].$$

We don't know r_h or P_h. Use empirical estimates.

Construct $\overline{Q}_h$ s.t. $\overline{Q}_h(s,a) \geq Q_h^*(s,a)$ $\forall h, s, a$.

$$\overline{Q}_h(s,a) = \hat{r}_h(s,a) + \mathbb{E}_{s' \sim \hat{P}_h(s,a)}\left[\max_{a'} \overline{Q}_{h+1}(s',a')\right] + b_h(s,a)$$

Confidence bound