7.1&7.2 Strategic Exploration and The UCB-VI algorithm

Background:

In an unknow MDP, agents have to engage in exploration in order to reach new states and execute enough samples.

Setting:

We work with finite horizon MDPs with the fixed start state s_0; Agents learn in an episode setting; In every episode k, the learner acts for H step starting from s_0.

Goal:

To minimise the agents' expected regret over K episodes:

$$\text{Regret} := \mathbb{E}\left[K V^*(s_0) - \sum_{k=0}^{K-1} \sum_{h=0}^{H-1} r(s_h^k, a_h^k) \right]$$

7.1 On the need for strategic exploration.

First, we present a sublinear regret algorithm: UCB-Value Iteration.

Algorithm UCBVI

Input: reward r, confidence parameter

1: for $k = 0, \cdots, K-1$ do

2: Compute $\widehat{P}_h^k$ as the empirical estimates for all h

3: Compute reward bonus b_h^k for all h.

4: Run Value iteration on $\{\widehat{P}_h^k, r_h + b_h^k\}_{h=0}^{H-1}$

5: Set π^k as the returned policy of Value iteration.

6: end for

We leave the concrete setting in the next section.

7.2 The UCBVI algorithm

1.2 The UCB-VI algorithm

[Notation]

① $N_h^k(s,a,s') = \sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}$

② $N_h^k(s,a) = \sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\}$

Note: These statistics help to form an empiral model.

In terms of the UCB-VI, we define the transitions:

$$\hat{P}_h^k(s'|s,a) = N_h^k(s,a,s') / N_h^k(s,a)$$

We also define the reward bonus as

$$b_h^k(s,a) = 2H\sqrt{L / N_h^k(s,a)}$$

where $L := \ln(SAHK/\delta)$ and δ represents the failure prob..

Now we state the value iteration, we perform all the way to $h=0$:

① $\hat{V}_H^k(s) = 0$,

② $\hat{Q}_h^k(s,a) = \min\{ r_h(s,a) + b_h^k(s,a) + \hat{P}_h^k(\cdot|s,a) \cdot \hat{V}_{h+1}^k , H \}$

③ $\hat{V}_h^k(s) = \max_a \hat{Q}_h^k(s,a)$, $\pi_h^k(s) = \arg\max_a \hat{Q}_h^k(s,a)$

Note: (i) In ②, we truncate $\hat{Q}_h^k(s,a)$ by H, because with the assumption $r \in [0,1]$, $Q^\pi \leq H$;

(ii) Denote $\pi^k = \{\pi_0^k, \cdots, \pi_{H-1}^k\}$. Leaner then executes π^k to get a new trajectory τ^k.

In following sections we will learn that UCBVI gives a regret bound in $O(H^2 S\sqrt{AK})$, followed by a more refined analysis that in $O(H^2\sqrt{SAK} + H^3 S^2)$.