7.4 Summary

Why do we need strategic exploration in a finite horizon MDP?

Set $M = \{ \{r_h\}_{h=0}^{H-1}, \{P_h\}_{h=0}^{H-1}, H, \mu, S, A \}$, consider a chain:

$s_0 \to$

0	0	- - -	0	1

Length : H.

Then the prob. of random walk hitting reward 1 is $(\frac{1}{3})^{-H}$, which is extremly small. We need to find a better policy.

The next question is : How can we find the optimal policy in face with unknow MDPs?

There are two attempts:

① Treat it as a MAB and run UCB.

If we consider the policy as the arms, then we have A^{SH} unique arms. Let K be the number of episodes.

Run UCB and we have the regret of $O(\sqrt{A^{SH}K})$

Note: shouldn't treat policies as independent arms since they do share infos.

② We run a new algorithm: The UCB Value-Iteration.

UCBVI protocol:

1. estimate $\hat{P}_h^k$
2. design reward bonus $b_h^k(s,a)$
3. optimise with learned model: $\pi^n = \text{value-iter}(\{\hat{P}_h^k, r_h + b_h^k\}_{h=0}^{H-1})$
4. collect a new trajectory under π^n.

The setting of UCBVI is similar with UCB, while the VI is designed as :

$V^k(s) = 0$ $\forall$

$V_H(s) = 0 \quad \forall s,$

$Q_h^k(s,a) = \min\{r_h(s,a) + b_h^k(s,a) + P_h^k(\cdot|s,a)\cdot\hat{V}_{h+1}^k, H\} \quad \forall s,a$

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

$V_h^k(s) = \max_a \hat{Q}_h^n(s,a)$

The UCBVI achieves a regret of $\tilde{O}(H^2\sqrt{S^2AK})$. which comes from Th. 7.1:

[Theorem 7.1] Regret Bound of UCBVI.

UCBVI achieves the following regret bound:

$$\text{Regret} := E\left[\sum_{k=0}^{K-1}(V^*(s_0) - V^{\pi^k}(s_0))\right]$$
$$\leq 20H^2s\sqrt{AK\cdot\ln(SAH^2K^2)} = \tilde{O}(H^2s\sqrt{AK})$$

which is much better in contrast with treating it as MAB.