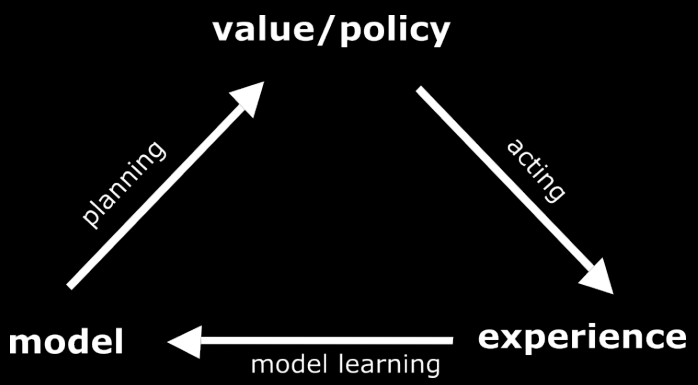
Regret analysis of UCRL with post-decision state

Yijun Wan

Département de Mathématiques et Applications
École Normale Supérieure

February 8, 2021

Upper Confidence Reinforcement Learning



- To optimistically fit a model to collected samples

$$(s_t, a_t, r_t, s_{t+1})_{t \geq 1}.$$

- Use the learned model for making decisions.

Questions:

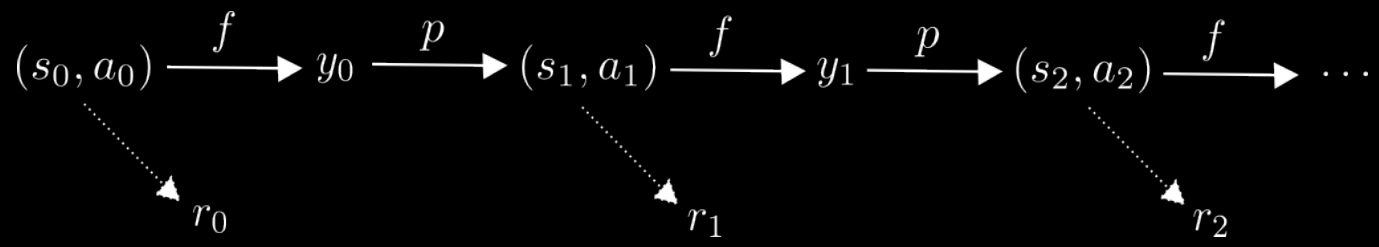
- How to optimistically "fit a model"?
  - Extended value iteration (EVI) with confidence bounds
- Regret guarantee?
  -

MDP with post-decision state

MDP $(\mathcal{S}, \mathcal{A}, p, r)$ such that the transition probabilities

$$p(s, a, s') = p(f(s, a), s')$$

$f : \mathcal{S} \times \mathcal{A} \to \mathcal{Y}$ is a known deterministic transition function.



Examples: Inventory problem, Board games etc.

Advantages: reduces the dimension of transition kernel and the dimension of value function

Algorithm

In each episode,

- Condidence set construction: construct a confidence set

$$B_k^p(y) := [p(\cdot|y) \in \mathcal{P}(\mathcal{S}) : \|p(\cdot|y) - \hat{p}_k(\cdot|y)\|_1 \leq \beta_k^p(y)]$$
$$B_k^r(s,a) := [\hat{r}_k(s,a) - \beta_k^r(s,a), \hat{r}_k(s,a) + \beta_k^r(s,a)]$$

of the unknown transition model, based on experiences $(s_t, a_t, y_t, s_{t+1}, r_t)$

$$\beta_k^p(y) := \sqrt{\frac{2S + 4\log(2Yt_k/\delta)}{N_k^+(y)}} \text{ and } \beta_k^r(s,a) := \sqrt{\frac{\log(2SAt_k/\delta)}{N_k^+(s,a)}}$$

- Optimistic planning:

$$\tilde{\pi}_k(s) = \arg\max_{a \in \mathcal{A}} \max_{r,p \in B} \{r(s,a) + V_{p,r}(f(s,a))\}$$

- Use the optimistic policy in the next episode.

Regret analysis

Definition

Given an MDP M and a stationary policy $\pi : \mathcal{Y} \to \mathcal{A}$, consider the Markov process for which the transition probability is

$$p(y, y') = \sum_{f(s,\pi(y))=y'} p_A(y, s) \text{ for all } y, y' \in \mathcal{Y}.$$

Define $T(y'|M, \pi, y)$ to be the first time when hitting y'' for the above Markov process starting from y. Then the *diameter* of M is defined as

$$D(M) := \max_{y \neq y' \in \mathcal{Y}} \min_{\pi:\mathcal{Y}\to\mathcal{A}} \mathbb{E}[T(y'|M, \pi, y)],$$

MDPs with finite diameter are called communicating.