$$\hat{P}_{ij}(a) = \frac{\sum_{1}^{t} \mathbb{1}_{\{X_s = i,\, A_s = a,\, X_{s+1} = j\}}}{\underbrace{\sum_{1}^{t} \mathbb{1}_{\{X_s = i,\, A_s = a\}}}_{T_t(i,a)}}$$

↳

$$M_t(i,a) = \left\{ P_{i,\cdot}(a) : \| P_{i,\cdot}(a) - \hat{P}_{i,\cdot}(a) \| \leq \sqrt{\frac{S \log\left(\frac{4SAT_t(i,a)}{\delta}\right) \times \frac{(i+T_t(i,a))}{\delta}}{T_t(i,a)}} \right\}$$

M_t: set of plausible MDPs.

In phase k, starting at T_k, play policy that is greedy wrt (ρ_k, h_k) where

$$\rho_k + h_k(i) = \min_a \left(c(i,a) + \sum \tilde{P}_{ij}(a)\, h(j) \right) \quad \text{for some } \tilde{P} \in M_{T_k} \qquad (1)$$

$$\rho_k = \min_{\pi \in \Pi \text{ stationary}} \min_{M \in M_{T_k}} \rho^{\pi}(M)$$

↳ Phase ends when $T_t(X_t, A_t) = 2\, T_{T_k}(X_t, A_t)$.

Extended MDP:

$$\tilde{S} = S$$
$$\tilde{A} = \{ (a, M) : a \in A,\ M \in M_{T_k} \}$$
$$\tilde{c}(i, (a,M)) = c(i,a)$$
$$\tilde{P}_{ij}(a,M) = P^{M}_{ij}(a).$$

↳ Solving above MDP, solves ①.

M_t: convex polyhedral set in higher dim.

Thm:

Under UCRL2,

$$E(R_n) \leq 1 + C\, D(M)\, S \sqrt{2An \log n}.$$

$D(M)$: diameter of MDP.

univ. constant.

—

* in case of BPI, if all mdps in M have same action for each state, stop.

Sampling: not as trivial.

①

EE 6106

16/4/24

↳ RL: Regret guarantees

Avg cost MDP: (S, A, P, c) (S, A finite)

Under "suitable" conditions, $\exists\ (\rho, h)$ satisfying

$$\rho + h(i) = \min_a \left(c(i,a) + \sum_j P_{ij}(a)\, h(j) \right)$$

Policy greedy wrt (ρ, h) is optimal.

↳ choose a which minimizes RHS; for every i.

↳ $\max_{(\rho,h)} \rho$

$$c(i,a) +$$

s.t. $\rho + h(i) \le \sum P_{ij}(a)\, h(j) \quad \forall i$

$h(0) = 0$.

↳ reference state. (can be translated additively so anchor one state to zero).

↳ LP solution solves B.E.

↳ RL:

P or C or both unknown.

Regret minimization

BPI (best policy identification).

↳ Regret minimisation:

Minimize $\mathbb{E}\left(\sum_{t} C(X_t, A_t)\right) - n\rho^*$

ρ^*: optimal time averaged cost (asymptotic)

↳ Turns out, Under optimal policy

$$\text{cost} = n\rho^* + O(1).$$

* UCRL2:

(Upper Confidence Bound For RL).

Assume C is known.

Unknown: P.

confidence set: (plausible MDPs)

$\mathcal{M}$ (set)

S, A, C: fixed.

P is the only unknown.

MDP with least ρ^*: optimal.

Play optimal action corresponding to best plausible MD