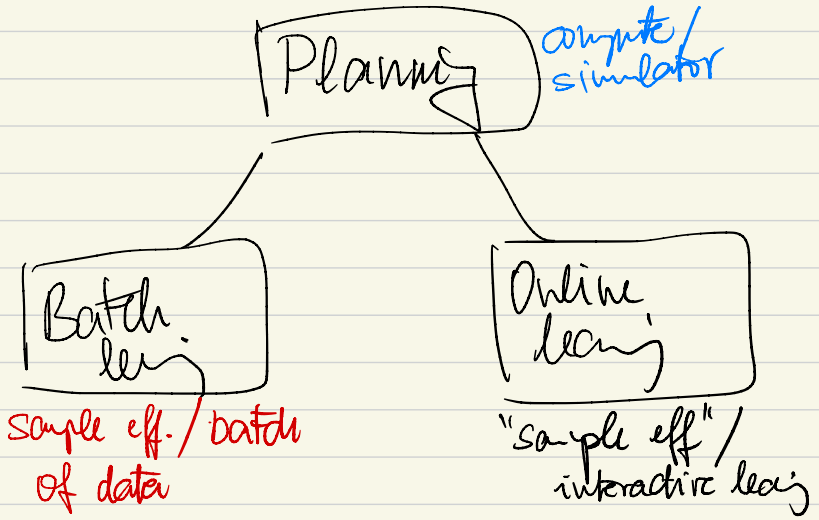
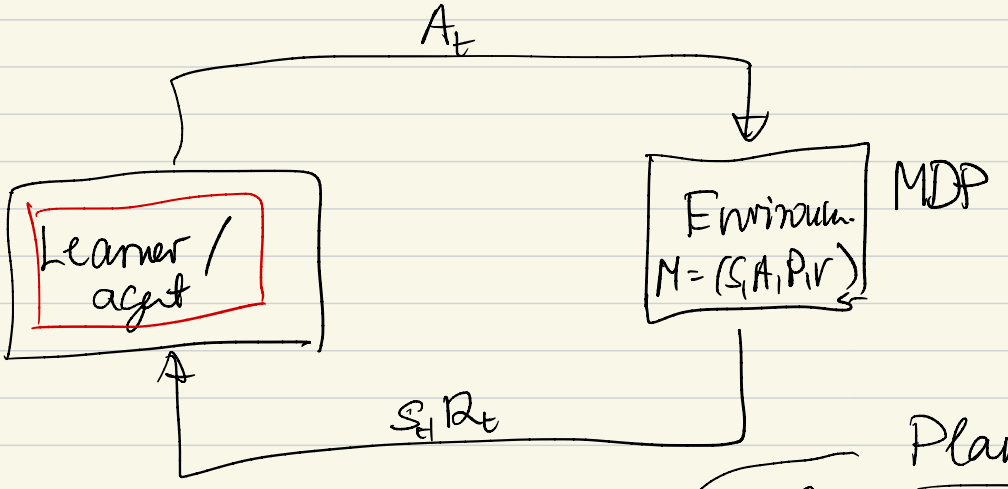
April 1

Online learning / Online RL





How good is a learner?

"Regret"

Cumulative regret

over time

undiscounted

|

fixed-horizon

infinite horizon

$$R_n = v_n^*(S_0) - \sum_{t=0}^{n-1} R_t$$

$$\rightarrow R_n = g^* n - \sum_{t=0}^{n-1} R_t$$

Σ opt. avg. reward

Planning

$\hat{v}$

$v^* - v^{\hat{\pi}}$

"simple" regret

Episodes of length $H > 0$, $h = 0, 1, \dots, H-1$

$1 \le k \le K$: # episodes.

$$S_0^{(k)} \sim \mu, \; A_0^{(k)}, S_1^{(k)}, A_1^{(k)}, \dots, S_{H-1}^{(k)}, A_{H-1}^{(k)}, S_H^{(k)}$$

$$S_{h+1}^{(k)} \sim P_{A_h^{(k)}}(S_h^{(k)})$$

$M = (S, A, P, r)$

For simplicity : r is known

$$R_K = \sum_{k=1}^{K} \left(V_0^*(S_0^{(k)}) - \sum_{h=0}^{H-1} r_{A_h^{(k)}}(S_h^{(k)}) \right)$$

↑ regret

$\frac{R_K}{K}$

PAC $\longleftrightarrow$ Regret

MDP

$H = 1$ $\Leftrightarrow$ contextual bandits

$|S| = 1$ finite-armed bandits

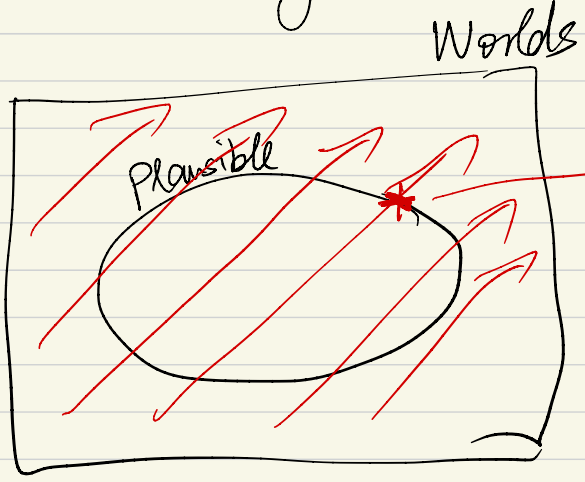
regret is unknown

ε-greedy $R_n = \Theta(n^{2/3})$, $R_n = O(\sqrt{n})$

Principle: Optimism in the Face of Uncertainty

Necessary: Try s_2 gives info about how good 's_2' is

Suff. for opt. to be "opt": try x does not give "much" info about y



follow policy that gives best value in the best world

Lai & Robbins '85

P.R. Kumar '83

$H > 0$ fixed horizon, episodic setting, $0 \le \delta < 1$

$M = (S, A, P^*, r)$

$k = 1, \dots, K$

knows / does not know

$a \vee b = \max(a, b)$

$$P_a^{(k)}(s, s') = \frac{N_k(s, a, s')}{1 \vee N_k(s, a)}$$

$$C_k = \left\{ P = (P_a(s)) \;\middle|\; \forall s,a : \left\| P_a^{(k)}(s) - P_a(s) \right\|_1 \le \beta(N_k(s,a)) \right\}$$

$\beta(0) = 1$

$\beta : \mathbb{N} \to (0, \infty)$

Goal of choosing β:

1. $P^* \in C_k \;\forall k$ w.p. $1-\delta$
2. C_k "not too large"

$$\pi_k = \operatorname*{argmax}_{\pi} v^{\pi}_{\tilde{P}_k}(s_0^{(k)})$$

$$\tilde{P}_k = \operatorname*{argmax}_{P \in C_k} v^{*}_{P}(s_0^{(k)})$$

Follow π_k up to the end of the episode.

Step 1: $C_k = ?$

Step 2: $R_k \le ?$