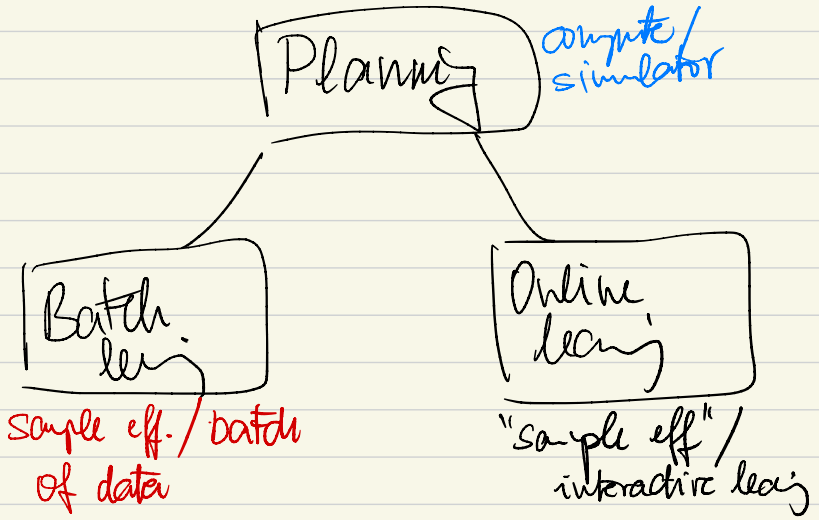
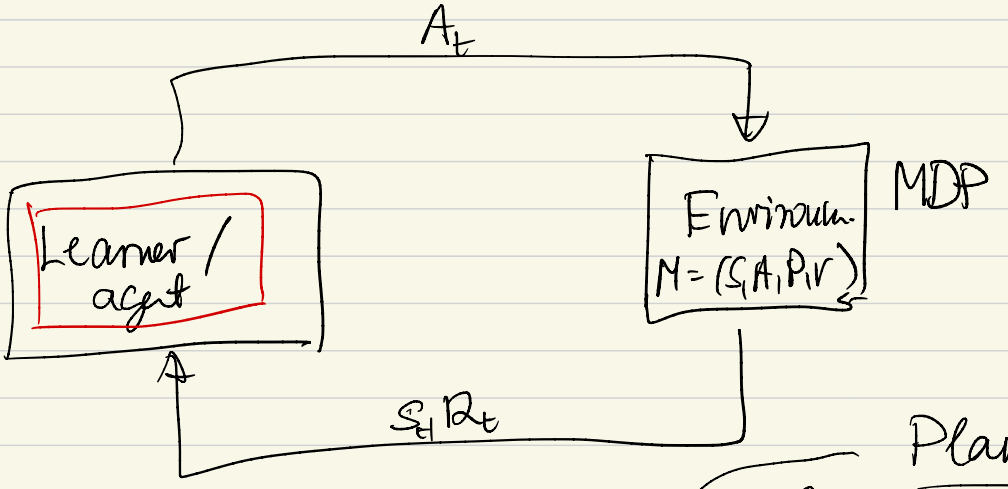
April 1

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

$$\mathbb{E}\left[\frac{R_K}{K}\right] \to 0, \; K \to \infty$$

PAC $\longleftrightarrow$ Regret

MDP

$H = 1$ $\Leftrightarrow$ contextual bandits

$|S| = 1$ finite-armed bandits

$n = KH$

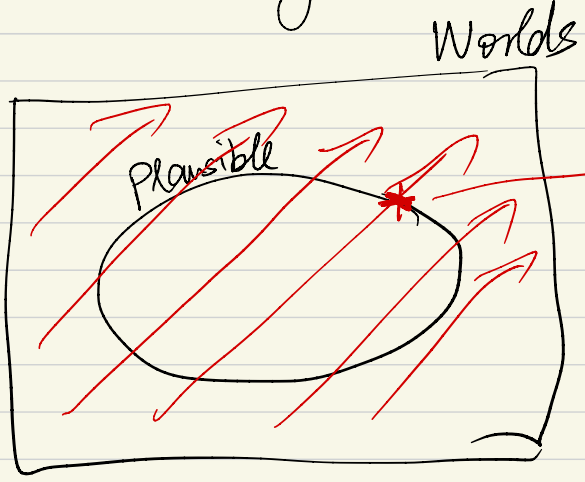
regret is unknown

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"

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

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

Follow π_k up to the end of the episode.

Step 1: $C_k = ?$

Step 2: $R_k \le ?$

s,a fixed

$u \in \mathbb{N}$ # visits to s,a

$$P_{u,a}(s,s') = \frac{\sum_{v=1}^{u} \mathbb{I}(X_v = s')}{u}$$

$X_v \in S$: next state observed upon visiting (s,a) the v^{th} time

Markov property $\Rightarrow$ $(X_v)_{v=1}^{n}$ i.i.d.

$P_a^*(s)$

$\|p\|_1 = \sup_{\|x\|_\infty \le 1} \langle p, x \rangle$

$$\| P_{u,a}(s) - P_a^*(s) \|_1 =$$

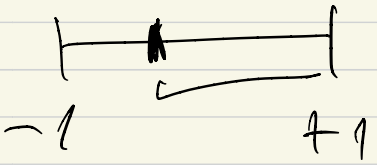
$$= \max_{x \in \{\pm 1\}^S} \langle P_{u,a}(s) - P_a^*(s), x \rangle$$

Fix $x \in \{\pm 1\}^S$

$$\langle P_{u,a}(s) - P_a^*(s), x \rangle =$$

$$\frac{1}{u} \sum_{v=1}^{u} \underbrace{\sum_{s'=1}^{S} x_{s'} \left(\mathbb{I}(X_v = s') - P_a^*(s, s') \right)}_{\Delta_v}$$

$$= \frac{1}{u} \sum_{v=1}^{u} \Delta_v$$



$$\mathbb{E}\Delta_v = 0, \quad |\Delta_v| \leq 2$$

$$(\Delta_v)_{v=1}^{u} \text{ i.i.d.}$$

Hoeffding's $\leq$: $\forall \delta \in (0,1)$

w.p. $1-\delta$ $\quad \frac{1}{u} \sum_{v=1}^{u} \Delta_v \leq 2 \sqrt{\frac{\log(1/\delta)}{2u}}$

$x \in \{\pm 1\}^S \qquad |\{\pm 1\}^S| = 2^S$

$\delta := \frac{\delta}{2^S}$ / union bound $\Rightarrow$

wp $1-\delta$

$$\|P_{n,a}(s) - P^*_a(s)\|_1 \le 2\sqrt{\frac{S\log 2 + \log(1/\delta)}{2n}}$$

$\forall k$: $\forall s,a$

$$\|P^{(k)}_a(s) - P^*_a(s)\|_1 \le \beta(N_k(s,a))$$

$$P^{(k)}_a(s) = P_{N_k(s,a),a}(s)$$

Take union bound over n, s, a

$$\sum_{n=1}^{\infty} \frac{\delta}{n(n+1)} = \delta \qquad \sum \frac{1}{n^2} < \infty$$

$$\frac{1}{n} - \frac{1}{n+1}$$

$\to$ wp $1-\delta$ $\forall n \ge 1$, $\forall s,a$

$$\|P_{n,a}(s) - P^*_a(s)\|_1 \le 2\sqrt{\frac{S\log 2 + \log\frac{n(n+1)SA}{\delta}}{2n}}$$

$$\frac{\delta}{n(n+1)SA}$$

$$\beta(n) = 2\sqrt{\frac{S\log 2 + \log\frac{n(n+1)SA}{\delta}}{2n}}$$

$\beta(u) \leq c_5/\sqrt{u}$, $C_\delta = 2\sqrt{\frac{1}{2}(S\log 2 + \log\frac{|\mathcal{H}|(KHA)^{SA}}{\delta})}$

Lemma 1: With β as above

$\forall \delta \in (0,1)$ wp $1-\delta$ $P^* \in \bigcap_{k \geq 1} C_k$

Fix δ, choose β as above.

Consider $\mathcal{E} = \{P^* \in \bigcap_{k \geq 1} C_k\}$

By Lemma 1: $P(\mathcal{E}) \geq 1-\delta$

In what follows assume that we are on $\mathcal{E}$.

$$R_K = \sum_{k=1}^{K} V_0^*(S_0^{(k)}) - V_k$$

$$V_k = \sum_{h=0}^{H-1} r_{A_h^{(k)}}(S_h^{(k)})$$

Fix $k \geq 1$.

$$V_0^*(s_0^{(k)}) - V_k$$

$$= V_{0,\tilde{p}^*}^{*}(s_0^{(k)}) - V_{0,\tilde{p}^{(k)}}^{*}(s_0^{(k)}) \quad \text{I.}$$

$$V_{0,\tilde{p}^{(k)}}^{\pi_k}(s_0^{(k)}) - V_{0,p^*}^{\pi_k}(s_0^{(k)}) \quad \text{II.}$$

$$V_{0,p^*}^{\pi_k}(s_0^{(k)}) - V_k \quad \text{III.}$$

I. ≤ 0 $\quad p^* \in C_k$

III. $= V_{0,p^*}^{\pi_k}(s_0^{(k)}) - \sum_{h=0}^{H-1} r_{A_h^{(k)}}(s_h^{(k)})$

$\mathbb{E}$

ξ_k

$\sum_{k=1}^{K} \xi_k$

$\mathbb{E}_{k-1}\xi_k = 0$

$\text{Rad}(\xi_k) \leq H$

$$\sum_{t=1}^{K} \xi_t \le H\sqrt{\frac{K}{2}\log(1/\delta)}$$

martingale difference seq ($\approx$ i.i.d.)

+ union bound ($1-2\delta$)

Hoeffding - Azuma

$$\text{I.} = V_{0,P^*}^{\overline{\pi}_k}(S_0^{(k)}) - V_{0,\tilde{P}^{(k)}}^{\overline{\pi}_k}(S_0^{(k)}) \quad (= \delta_0^{(k)})$$

$$V_{h,P}^{\pi} = r^{\pi} + M_{\pi} P V_{h+1,P}^{\pi}, \quad 0 \le h \le H-1$$

$$\boxed{V_{H,P}^{\pi} = 0}$$

$\to \|P^* - P\|_\infty$

$$V_{h,P^*}^{\pi} - V_{h,P}^{\pi} = M_{\pi}(P^* - P) V_{h+1,P^*}^{\pi}$$

$$M_\pi P \underbrace{\left(V^{\pi}_{h+1,P^*} - V^{\pi}_{h+1,P}\right)}$$

$$\| \qquad \|_\infty$$

$$(*)\quad V^{\pi}_{h,P^*} - V^{\pi}_{h,P} = M_\pi P_* \left(V^{\pi}_{h+1,P^*} - V^{\pi}_{h+1,P}\right) + M_\pi \left(P_* - P\right) V^{\pi}_{h+1,P}$$

$$V^{\pi/k}_{h,P^*}\left(S_h^{(k)}\right) - V^{\pi/k}_{h,\tilde{P}^{(k)}}\left(S_h^{(k)}\right) = \delta_h^{(k)} - \underbrace{\mathbb{E}\left[\delta_{h+1}^{(k)} \mid \mathcal{F}_{h,k}\right]}$$

$$\delta_h^{(k)} = \left(M_{\pi^*} P^* \left(V^{\pi/k}_{h+1,P^*} - V^{\pi_k}_{h+1,\tilde{P}^{(k)}}\right)\right)\left(S_h^{(k)}\right) + \left(M_{\pi^k}\left(P^* - \tilde{P}^{(k)}\right) V^{\pi_k}_{h+1,\tilde{P}^{(k)}}\right)\left(S_h^{(k)}\right)$$

$$\leq \delta_{h+1}^{(k)} + \underbrace{\left(\mathbb{E}\left[\delta_{h+1}^{(k)} \mid \mathcal{F}_{h,k}\right] - \delta_{h+1}^{(k)}\right)}_{\gamma_{h+1}^{(k)}} + \left\| P^*_{A_h^{(k)}}\left(S_h^{(k)}\right) - \tilde{P}^{(k)}_{A_h^{(k)}}\left(S_h^{(k)}\right) \right\|_1 \cdot H$$

$$\left(M_\pi (P^* - P) V\right)(s) \qquad \pi(s) = a$$

$$= \langle P_a^*(s) - P_a(s), V \rangle$$

$$\leq \| P_a^*(s) - P_a(s) \|_1 \, \| V \|_\infty$$

$$\leq \delta_{h+1}^{(k)} + \eta_{h+1}^{(k)} + 2\beta\left(N_k(s_h^{(k)}, a_h^{(k)})\right)$$

$$\delta_H^{(k)} = 0$$

$$\delta_0^{(k)} \leq \eta_1^{(k)} + \dots + \eta_{H-1}^{(k)} + H 2 \sum_{h=0}^{H-1} \beta\left(N_k(s_h^{(k)}, a_h^{(k)})\right)$$