EEML "Bits of truths about RL"

1. $\Omega(\sqrt{SAn})$ contextual bandits
2. Optimism
3. Function approx
4. Batch RL $\Omega(A^{\min(S,H)}/\varepsilon^2)$

---

## Contextual bandit

Bandit = 1 state MDP

Contextual bandits = $S$ states, $S \sim \underline{P}$ i.i.d. ← $\varepsilon$

$A > 0$ # actions

Online

$1 \le i \le A$ : actions

$\underline{\mu_i}(S, \cdot)$ reward distr. for action $i$.

$t = 1, 2, \ldots$

- $I_t$ : choice of learner ←
- $S_t \sim p(\cdot)$
- $R_t \sim \mu_{I_t}(S_t, \cdot)$

} observed by learner

$$r_i(s) = \int r \, \mu_i(s, r) \, dr$$

$R_t \in [0,1]$

$$\text{Regret}_n = \sum_{t=1}^{n} \max_i r_i(s_t) - R_t$$

$$\mathbb{E}[\text{Regret}_n] = \mathcal{R}_n(A, \mathcal{E})$$

$$\mathcal{R}_n^*(\mathcal{C}) = \inf_A \sup_{\mathcal{E} \in \mathcal{C}} \mathcal{R}_n(A, \mathcal{E})$$

$\mathcal{C}$ $\mu_i(s_i)$ $\in [0,1]$ gaussien 1-variance

$$\mathcal{R}_n^*(\mathcal{C}) = \Theta(\sqrt{SAn})$$

$$\frac{\mathcal{R}_n^*(\mathcal{C})}{n} = \Theta\left(\sqrt{\frac{SA}{n}}\right)$$

# Part II — Explore – exploit dilemma

## Optimism

Optimism in the face of uncertainty

$S = 1$

$r_1, \dots, r_A$ : means

$\hat{r}_1, \dots, \hat{r}_A$ : empirical mean

$n_1, \dots, n_A > 0$ : counts.

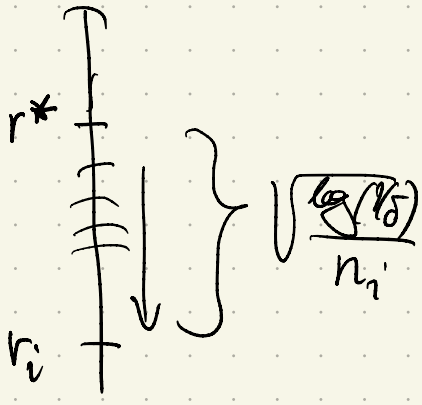
UCB

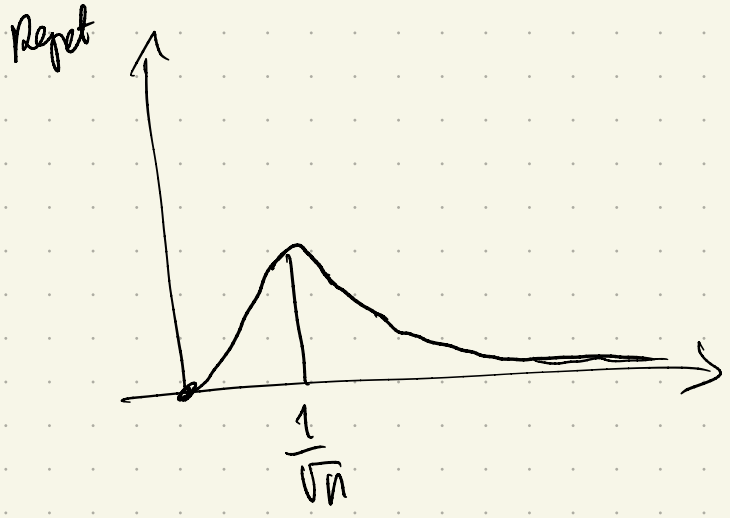
$$\underset{1 \le i \le A}{\operatorname{argmax}} \underbrace{\hat{r}_i + c\sqrt{\frac{\log(1/\delta)}{n_i}}}_{\tilde{r}_i}$$

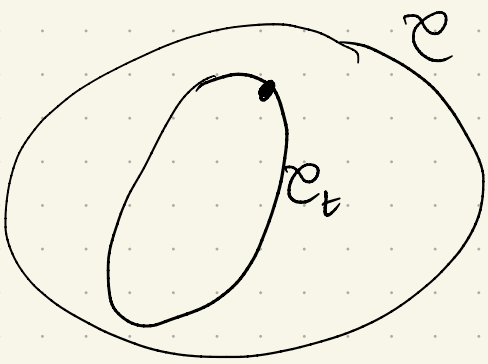
$c > 0$

$\delta = 1/n$

$$\mathbb{P}(\tilde{r}_i < r_i) \le \frac{1}{n}$$

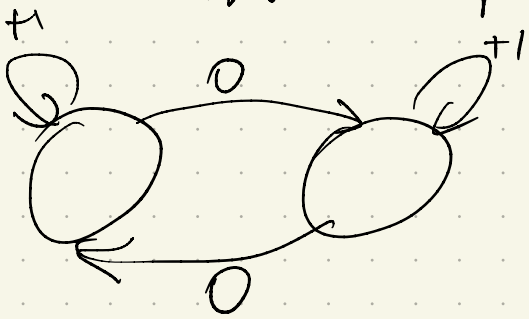





Controlled state

copy policy has to be used for an extended period



Part 3) Function approximation

$\Omega(\sqrt{SAn})$

$([S],[A],Q)$

Bellman

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P(s,a,s') \max_{a'} Q^*(s',a')$$

$$\arg\max_a Q^*(s,a)$$

$[S] = \{1,\dots,S\}$

$[A] = \{1,\dots,A\}$

$Q \approx Q^*$

$\arg\max_a Q(s,a)$

$\varphi : [S] \times [A] \to \mathbb{R}^d$

$\exists\, \theta^* \in \mathbb{R}^d$

$$Q^*(s,a) \approx \varphi(s,a)^T \theta^*$$

$\theta^* = ?$

$\hat{\theta} \approx \theta^*$

Simulation optimization / planning

$$\arg\max_a \varphi(s,a)^T \hat{\theta}$$

---

Thm: $\forall$ planner $\mathcal{P}$ $\exists$ (MDP, $\varphi$) $Q^* = \varphi^T \theta^*$

s.t. #queries used by $\mathcal{P}$

$\geq 2^{\Omega(\min(H,d))}$

$$H = \frac{1}{1-\gamma}$$

$V^* = \varphi^T \theta^*$ + positive

$$O\left(\left(\frac{dH}{\delta}\right)^{\frac{A}{2}}\right)$$

$\delta$ Suboptimality of $\pi$ induced by planner

TD

$A = 2$

---

$Q^\pi \approx \varphi^T \theta_\pi$

policy iteration

policy eval

TD methods?

---

value iteration

$\mathcal{T}\mathcal{F} \subseteq \mathcal{F}$

"closedness"

(Part IV.) Batch RL / Off-policy policy opt.

**Thm:** Data generated by following some policy.

query complexity $= \Theta\left(A^{\min(H,S)}/\varepsilon^2\right)$