08/09/23

UCB-VI Algorithm

Recap: so far, MDP w/ generative model (simulator)

Today: exploration in MDP

Setup - $M = (S, A, H, \{P_h\}_{h \in [H]}, \{r_h\}_{h \in [H]})$

- no assumption on generative model
- interaction protocol:

For each game (episode):

- start of the episode: $s_1 \sim d_1(s_1)$
- For $h = 1, 2, .., H$:
  - take $a_h \in A$
  - observe $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ and reward $r_h = r_h(s_h, a_h)$

regret minimization: Find sequence of policies $\{\pi_k\}_{k\in[K]}$

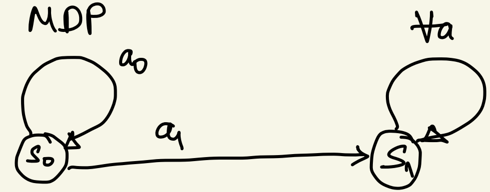
to minimize $\text{regret}(K) = K \cdot V_1^*(s_1) - \sum_{k=1}^{K} V_1^{\pi_k}(s_1)$

Review Story from MAB

Alg	Regret
explore-then-commit	$\tilde{O}(A^{1/3} T^{2/3})$
ε-greedy	$\tilde{O}(A^{1/3} T^{2/3})$
UCB	$\tilde{O}(\sqrt{AT})$
lower bound	$\Omega(\sqrt{AT})$

Lemma Random exploration requires $\Omega(2^H)$ samples to find an optimal policy

Proof: "Combinatorial lock" MDP



zero rewards everywhere except at (H, s_0) where $r_H(s_0, a) = 1 \quad \forall a$

Optimal policy: $\pi^*_h(s_0) = a_0$, $\quad V_1^{\pi^*}(s_0) = 1$

$\pi^*_h(s_1) = a_0 \text{ or } a_1$ $\quad V_1^{\pi^*}(s_1) = 0$

To discover π^*, action sequence must be $(a_0, \dots, a_0)$

$$\Pr_{\text{Unif}}\left((a_0, \dots, a_0)\right) = \frac{1}{2^H}$$

$\Rightarrow$ need $\Omega(2^H)$ episodes to discover $(a_0, \dots, a_0)$ for once

UCB - VI Let $b(N) = c\sqrt{\frac{H^2 L}{N}}$ where $L := \log\left(\frac{SAHK}{\delta}\right)$

- initialize $D = \emptyset$, $Q_h(s,a) = 0 \;\; \forall h \in [H]$ $Q_{H+1}(s) = 0$
- for $k = 1, 2, \dots, K$: (estimation phase)
  - $$\hat{P}_h(s'|s,a) = \frac{N_h(s,a,s')}{N_h(s,a)}$$

    where $N_h(s,a,s') = \left|\{(h,s,a,s') \in D\}\right|$

    $N_h(s,a) = \left|\{(h,s,a) : (h,s,a,s') \in D\}\right|$
  - $$Q_h(s,a) = \Big[r_h(s,a) + \big(\hat{P}_h V_{h+1}\big)(s,a) + \underbrace{b\big(N_h(s,a)\big)}_{\text{bonus}}\Big]_{[0,H]}$$
  - $$V_h(s) = \max_a Q_h(s,a)$$
- (Execution phase)

  For $h = 1, \dots, H$:

  Take $a_h = \arg\max_a Q_h(s_h, a)$

  Observe s_{h+1}, r_h

  $D = D \cup \{(h, s_h, a_h, s_{h+1})\}$

Theorem   w.p. at least $1-\delta$, the regret of UCB-VI is:

$$\text{regret}(K) \leq c \cdot \left(H^2 \sqrt{SAK\iota} + H^3 S^2 A \iota^3 \right)$$

Notations   Add superscript k to all quantities

- D^k: D up to k-th episode
- $N_h^k(s,a,s')$: number of (h,s,a,s') in D^k
- $\hat{P}_h^k$: empirical distribution
- $(s_h^k, a_h^k, s_{h+1}^k)$: tuple played at k-th episode

Lemma (optimism) wpal $1-\delta$:

$$Q_h^k(s,a) \geq Q_h^*(s,a), \quad V_h^k(s) \geq V_h^*(s) \quad \forall (k,h,s,a)$$

Proof (by induction)

- when $h = H+1 \longrightarrow$ trivial
- Assume by induction that it holds for some $h+1$.
- $$Q_h^k(s,a) - Q_h^*(s,a) = (\hat{P}_h^k V_{h+1}^k)(s,a) + b(N_h^k(s,a)) - (P_h V_{h+1}^*)(s,a)$$

$$= \hat{P}_h^k (V_{h+1}^k - V_{h+1}^*)(s,a)$$

$$+ \underbrace{(\hat{P}_h^k - P_h) V_{h+1}^*(s,a) + b(N_h^k(s,a))}_{\geq 0 \text{ by Hoefding's ineq}}$$

- $$V_h^k(s) = \max_a Q_h^k(s,a) \geq \max_a Q_h^*(s,a) = V_h^*(s)$$

$$\cdot\ \text{regret}(K) = \sum_{k=1}^{K} \left(V_1^{\pi^*}(s_1) - V_1^{\pi_k}(s_1) \right) \leq \sum_{k=1}^{K} \left(V_1^{k}(s_1) - V_1^{\pi_k}(s_1) \right)$$

Optimism remove the unknown π^* from our bound

$\longrightarrow$ make our job easier

$$V_h^{k}(s_1^{k}) - V_h^{\pi_k}(s_1^{k}) =$$

$$(Q_h^k - Q_h^{\pi_k})(s_h^k, a_h^k) \le (\hat{P}_h^k V_{h+1}^k - P_h V_{h+1}^{\pi_k})(s_h^k, a_h^k) + b_h^k$$

$$= (\hat{P}_h^k - P_h) V_{h+1}^k (s_h^k, a_h^k) + P_h (V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^k, a_h^k) + b_h^k$$

$$b(N_h^k(s_h^k, a_h^k))$$

$$= \underbrace{(\hat{P}_h^k - P_h) V_{h+1}^* (s_h^k, a_h^k)}_{\le b_h^k} + (\hat{P}_h^k - P_h)(V_{h+1}^k - V_{h+1}^*)(s_h^k, a_h^k) + P_h (V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^k, a_h^k) + b_h^k$$

$$\le \underbrace{(\hat{P}_h^k - P_h)(V_{h+1}^k - V_{h+1}^*)(s_h^k, a_h^k)}_{I} + \underbrace{P_h (V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^k, a_h^k)}_{J} + 2b_h^k$$

Note: $(\hat{P}_h^k - P_h) V(s_h^k, a_h^k) \leq b_h^k$ but V_{h+1}^k is data-dependent!

$$u(s') = (V_{h+1}^k - V_{h+1}^*)(s') \quad \forall s' \in S$$

$$I = \sum_{s' \in S} \left(\hat{P}_h^k(s' \mid s_h^k, a_h^k) - P_h(s' \mid s_h^k, a_h^k) \right) u(s')$$

$$\lesssim c \sum_{s' \in S} \left[\sqrt{\frac{P_h(s' \mid s_h^k, a_h^k)\, \iota}{N_h^k(s_h^k, a_h^k)}} + \frac{\iota}{N_h^k(s_h^k, a_h^k)} \right] \cdot u(s') \quad \text{(Bernstein's)}$$

$$\leq c \sum_{s' \in S} \left[\frac{P_h(s' \mid s_h^k, a_h^k)}{cH} + \frac{cH\iota}{N_h^k(s_h^k, a_h^k)} \right] u(s') \quad \text{(AM–GM)}$$

$$= \frac{1}{H} P_h (V_{h+1}^k - V_{h+1}^*)(s_h^k, a_h^k) + c \underbrace{\frac{SH^2 \iota}{N_h^k(s_h^k, a_h^k)}}_{\zeta_h^k}$$

$$\underbrace{V_h^k(s_h^k) - V_h^{\pi_k}(s_h^k)}_{\Delta_h^k} \le I + J + 2b_h^k$$

$$\le \frac{1}{H} P_h(V_{h+1}^k - V_{h+1}^{*})(s_h^k, a_h^k) + S_h^k$$

$$P_h(V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^k, a_h^k) + 2b_h^k$$

$$\le \left(1+\frac{1}{H}\right) \underbrace{P_h(V_{h+1}^k - V_{h+1}^{\pi_k})(s_h^k, a_h^k)}_{\xi_h^k + (V_{h+1}^k - V_{h+1}^{\pi_k})(s_{h+1}^k)} + 2b_h^k + S_h^k$$

where $\xi_h^k = P_h(V_h^k - V_{h+1}^{\pi_k})(s_h^k, a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi_k})(s_{h+1}^k)$

(martingale).

$$\Delta_h^k \leq \left(1+\frac{1}{H}\right)\left(\Delta_{h+1}^k + \xi_h^k\right) + 2b_h^k + S_h^k$$

$$= \left(1+\frac{1}{H}\right)\Delta_{h+1}^k + \left(1+\frac{1}{H}\right)\xi_h^k + S_h^k + 2b_h^k$$

$$\Delta_1^k \leq \underbrace{\left(1+\frac{1}{H}\right)^H}_{e} \Delta_H^k + \underbrace{\left(1+\frac{1}{H}\right)^H}_{\leq e} \sum_{h=1}^{H} \left(\xi_h^k + S_h^k + b_h^k\right)$$

$$\Rightarrow \text{regret}(K) = \sum_{k=1}^{K} \Delta_1^k \leq c \cdot \sum_{k=1}^{K} \sum_{h=1}^{H} \left(\xi_h^k + S_h^k + b_h^k\right)$$

$$\cdot \quad \sum_k \sum_h b_h^k = c' H\sqrt{\iota} \sum_k \sum_h \frac{1}{\sqrt{N_h^k(s_h^k, a_h^k)}}$$

$$= c \cdot H\sqrt{\iota} \sum_h \sum_{(s,a)} \sum_{i=1}^{N_h^K(s,a)} \frac{1}{\sqrt{i}}$$

$$\leq c H\sqrt{\iota} \sum_{(s,a,h)} \sqrt{N_h^K(s,a)}$$

$$= cH\sqrt{\iota} \sqrt{SAH} \sqrt{KH} = H^2 \sqrt{SAK\iota}$$

$$\cdot \quad \sum_k \sum_h \delta_h^k \lesssim c \cdot SH^2\iota \sum_{k,h} \frac{1}{N_h^k(s_h^k, a_h^k)}$$

$$\lesssim c \cdot SH^2\iota \sum_{(h,s,a)} \sum_{i=1}^{N_h^K(s,a)} \frac{1}{i}$$

$$\lesssim c \cdot SH^2\iota \sum_{(h,s,a)} \log N_h^K(s,a) \leq c \cdot SH^2\iota \log(KH)$$

- $\sum_k \sum_n \xi_n^k \leq H^2 \cdot \sqrt{K}$