Final presentation schedule in the third sheet of sign-up google sheet.

Piazza poll @ 43.

Final report:

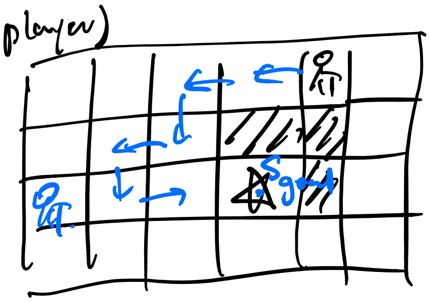
* Introduction
  - what is the research problem?
  - why is this qn interesting
  - prior work w/ citations
  - what are the qns your project tries to answers?
  - what you've accomplished (at a high level).
* Preliminaries (notation & setting)
  - define models / notations
  - basic math facts. Lemma / Fact.
* Body.
  - details.

§ Reinforcement learning

account for long term effects of actions.

Mathematical model: episodic Markov Decision Processes (MDPs).

Defns: state space: $\mathcal{S}$ (e.g. location of player)



action space: $\mathcal{A}$.

(left / right / up / down / stay)

episode length H: (# steps in episode) $\{1, \dots, H\}$

reward fn $r_1, \dots r_H : S \times A \to [0,1]$

(e.g. $\forall h$, $r_h(s_{goal}, a) = 1$

$r_h(s, a) = 0 \quad s \neq s_{goal}$)

transition fn: $P_1 \dots P_H : S \times A \to \Delta(S)$

Learning protocol:

Assume: $r_1 \dots r_H$ ~~known~~ unknown. $\quad P_1 \dots P_H$ unknown

(e.g. know the goal's location)

For episodes $t = 1, 2, \dots T$: $\quad \underline{s_h^t}$

— learner sees some initial state for this episode s_1^t

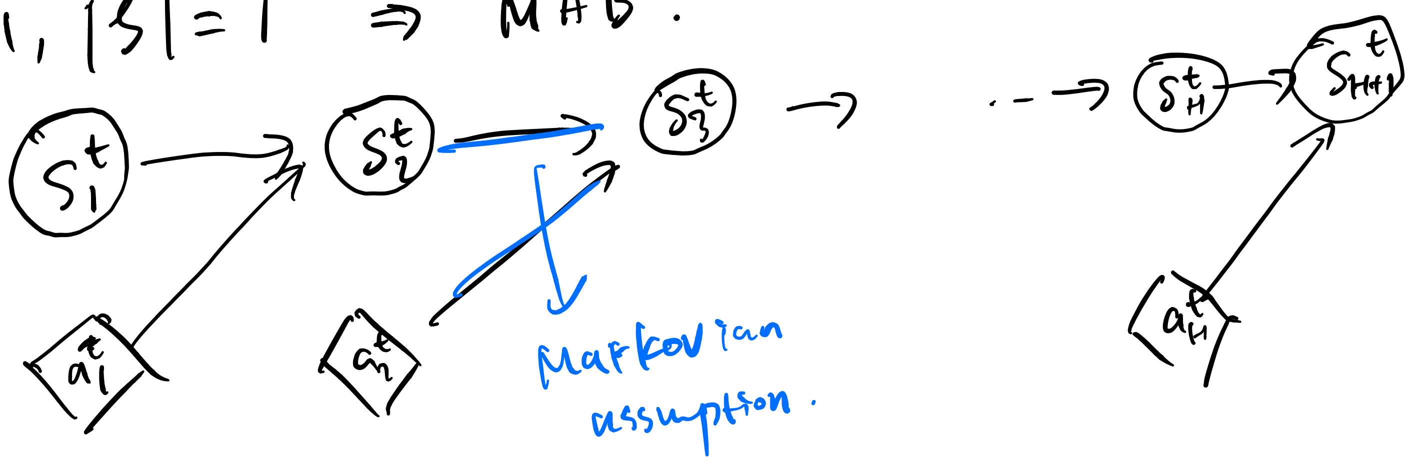
For steps $h = 1, \dots H$:

— learner takes action $a_h^t \in A$, receive reward $r_h(s_h^t, a_h^t)$,

observe next state $s_{h+1}^t \sim P_h(\cdot \mid s_h^t, a_h^t)$.

$H = 1 \Rightarrow$ contextual bandits

$H = 1, |S| = 1 \Rightarrow$ MAB.

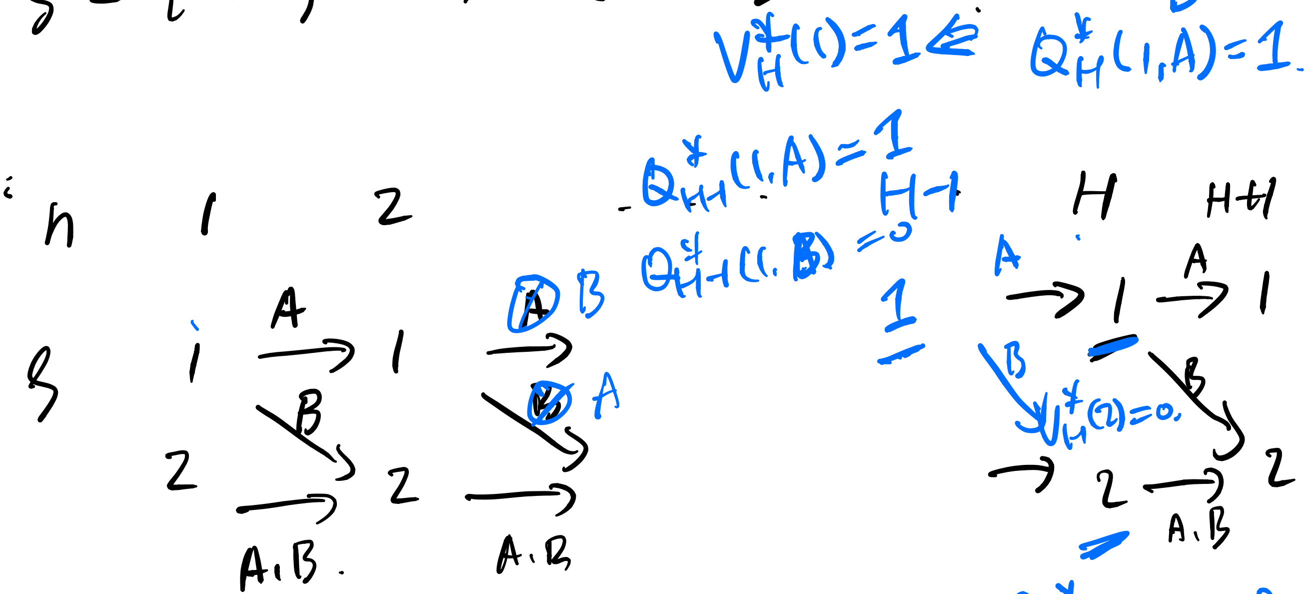


What makes RL uniquely challenging?

Ex: combination lock.

$S = \{1, 2\} \qquad A = \{A, B\} . \qquad H . \qquad B$

$(P_h)_{h=1}^{H}$:



$Q_H^*(2,A)=0$, B

$\forall h,s,a,$
$r_h(s,a)=0$. except that $r_H(1,A)=1$
$r_H(1,B)=1$

① If random actions are taken, chance of receiving nonzero reward is $\frac{1}{2^{H-1}}$ (exponentially small)

② " $\underbrace{A \ldots A}_{H-1}$ " as password.

If we view this problem as a MAB problem [w/ 2^H actions], then need to try $\Omega(2^H)$ "actions" to get nonzero reward.

Optimal value & policy in MDPs. $\sum_{h'=h}^{H} r_{h'}$.

$V_h^*(s)$ = highest expected reward we can get, given that we are at state s at step h.
optimal value fn

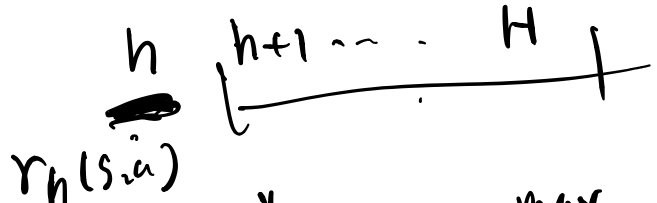
$Q_h^*(s,a)$ = highest expected ~~r~~ " " ", given that we are at state s. and take action a at step h.
optimal action-value fn
Q-fn

Solve V_h^* Q_h^*'s using dynamic programming.

$h = H+1$. Q^*_{H+1} undefined $V^*_{H+1}(s) = 0$.

$h \le H$. $Q_h^*(s,a) = r_h(s,a) + \sum_{s' \in S} P_h(s'|s,a) \cdot V^*_{h+1}(s')$

Bellman optimality eqn.



$$V_h^*(s) = \max_{a \in A} Q_h^*(s,a)$$

$$V_1^* \leftarrow \cdots \leftarrow V_H^* \leftarrow Q_H^* \leftarrow V_{H+1}^*$$

Optimal policy: given we are at state s at h.

$$a = \arg\max_{a \in A} Q_h^*(s,a) =: \pi_h^*(s).$$

$\pi^* = (\pi_1^* \dots \pi_H^*)$ In general, π_h^*'s are not equal.

$\pi = (\pi_1 \dots \pi_h \dots \pi_H)$.

$V_h^\pi(s)$ = expected reward given we are at state s at step h, and execute policy π.

Solve V_h^π using DP:

$V^\pi(s')$

$$V_h^\pi(s) = r_h(s, \pi_h(s)) + \sum_{s' \in S} P_h(s' | s, \pi_h(s)) V_{h+1}^\pi(s')$$

$$V_{H+1}^\pi(s) = 0.$$

performance measure of online RL: $V_1^{\pi^t}(s_1^t)$

$$R_T = \mathbb{E}\left[\sum_{t=1}^T V_1^*(s_1^t) - \underline{\sum_{t=1}^T \sum_{h=1}^H r_h(s_h^t, a_h^t)} \right]$$

learner takes policy π^t for episode t

$$= \mathbb{E}\left[\sum_{t=1}^T \left(V_1^*(s_1^t) - V_1^{\pi^t}(s_1^t) \right) \right].$$

would like $R_T = o(T)$.

RL algorithms
- Model Based
- Model Free
  - Value based — $\underline{S \times A}$, Q_h^*
  - policy search : gradient descent
  - actor critic

This lecture: value-based method.

optimistic Q-learning (Jin, Allen-Zhu, Bubeck, Jordan, 2018).

maintain optimistic estimate Q_h for Q_h^* online.

$V_h(s) = H$. $\le H$.

Alg: Initialization: $Q_h(s,a) = H$. $\forall h, s, a$

$V_{H+1}(s) \equiv 0$.

$m_h(s,a) = 0$ $\forall s, a, h$.

For episodes $t = 1, 2, \dots T$:

initial state s_1^t

execute $\pi_h^t(s) = \arg\max_a Q_h(s,a)$.

For step $h = 1 \dots H$: s_h^t.

take action $a_h^t = \arg\max_{a \in A} Q_h(s_h^t, a)$.

transition to s_{h+1}^t. $r_h(s_h^t, a_h^t)$

$m = m_h(s_h^t, a_h^t) \leftarrow m(s_h^t, a_h^t) + 1$

update:

$\alpha_m = \frac{1}{m}$

$\alpha_m = \frac{H+1}{H+m} > \frac{1}{m}$

$$Q_h(s_h^t, a_h^t) \leftarrow (1-\alpha_m) Q_h(s_h^t, a_h^t) + \alpha_m \left(r_h(s_h^t, a_h^t) + V_{h+1}(s_{h+1}^t) + b_m \right)$$

$V_h(s_h^t) \leftarrow \min\left(\max_{a \in A} Q_h(s_h^t, a), H\right)$

$b_m = c\sqrt{\frac{1}{m}}$.. $c \lesssim \sqrt{H^3 \ln\left(\frac{SAHT}{\delta}\right)}$

s_h^t, a_h^t

$$Q_h^*(s,a) = r_h(s,a) + \langle P_h(\cdot \mid s,a), V_{h+1}^* \rangle$$

$\mathbb{E}$

$$r_h(s_h^t, a_h^t) + V_{h+1}(s_{h+1})$$

$$\uparrow$$

$$r_h(\quad) + V_{h+1}(s_{h+1}^t)$$

Thm: optimistic Q-learning has:

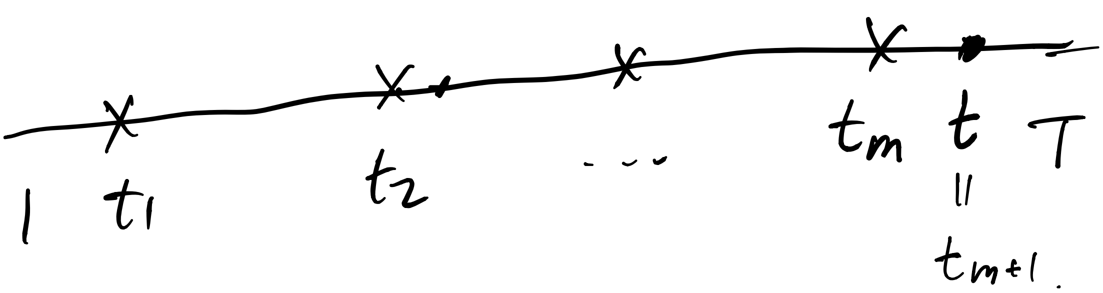
$$R_T \le O\left(\sqrt{T \cdot S \cdot A \cdot H^5}\right)$$

(Lower bound: $\forall$ alg, $\exists$ episodic RL environment, $R_T \geq \Omega\left(\sqrt{T \cdot S \cdot A \cdot H^3}\right)$)

defns: $Q_h^t(s,a) = Q_h(s,a)$, at the beginning of episode t

$V_h^t(s) = V_h(s)$ " " " "

$m_h^t(s,a) = m_h(s,a)$ " " " " " ".



$$Q_h^{t_1}(s,a) = H.$$

$\forall i = 1, 2, \dots m.$

$$Q_h^{t_{i+1}}(s,a) = (1-\alpha_i)\, Q_h^{t_i}(s,a) + \alpha_i \cdot \underbrace{\left(r_h(s,a) + V_{h+1}^{t_i}(s_{h+1}^{t_i}) + b_i \right)}_{\tilde{Q}^i}.$$

$$\begin{tikzcd}
 & & & Q_h^{t_{m+1}} \arrow[dl, "1-\alpha_m"'] \arrow[dr, "\alpha_m"] & \\
 & & Q_h^{t_m} \arrow[dl, "1-\alpha_{m-1}"'] \arrow[dr, "\alpha_{m-1}"] & & \tilde{Q}^m \\
 & Q_h^{t_{m-1}} \arrow[dl] \arrow[dr, "\alpha_i"] & & \tilde{Q}^{m-1} & \\
  & & \tilde{Q}^i & &
\end{tikzcd}$$

$$\alpha_m = \frac{H+1}{m+H}.$$

$1-\alpha_{i+1}$

$$\begin{tikzcd}
 & Q_h^{t_2} \arrow[dl, "1-\alpha_1 = 0"'] \arrow[dr, "\alpha_1"] & \\
 Q_h^{t_1} & & \tilde{Q}^1
\end{tikzcd}$$

$\alpha_1 = 1$

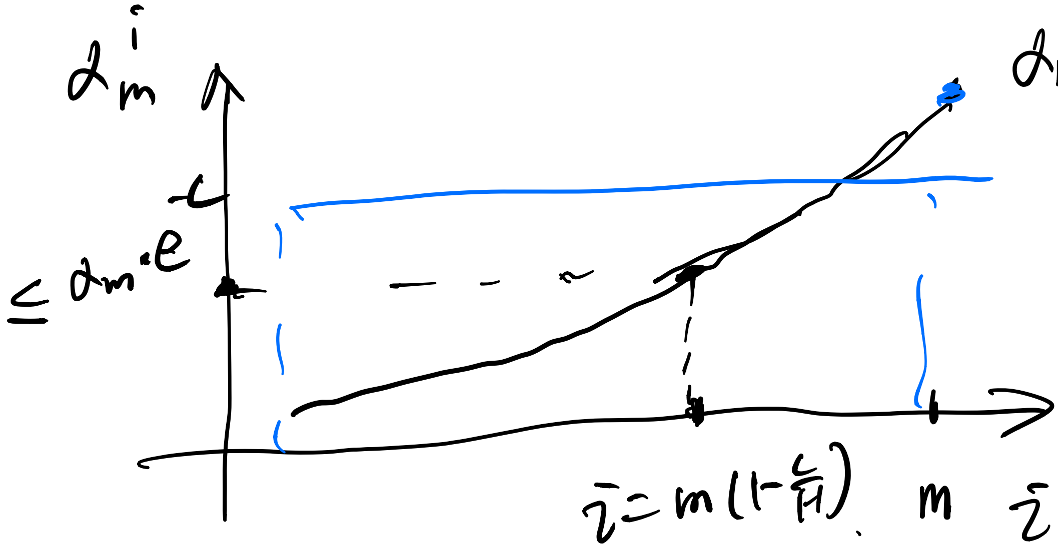
$$\frac{1}{m-1} \times \left(1 - \frac{1}{m}\right) = \frac{1}{m}$$

H

$$Q_h^t(s,a) = \sum_{i=1}^{m} \alpha_m^i \hat{Q}^i \qquad m \geq 1.$$

$$\alpha_m^i = \alpha_i \cdot \prod_{j=i+1}^{m} (1-\alpha_j).$$

For $\alpha_i = \frac{H+1}{H+i}$. $\quad \alpha_m^i$?



$$\alpha_m^m = \frac{H+1}{m+1} = \alpha_m \leq \frac{2H}{m}.$$

properties of α_m^i:

① $\sum_{i=1}^{m} \alpha_m^i = 1 \qquad \alpha_m^i \geq 0. \ \forall i.$

② $\sum_{i=1}^{m} \alpha_m^i \cdot \frac{1}{\sqrt{i}} \in \left[\frac{1}{\sqrt{m}}, \frac{2}{\sqrt{m}}\right].$

③ $\sum_{i=1}^{m} (\alpha_m^i)^2 \leq \frac{2H}{m}.$

$\leq \left(\sum^{m} \alpha_m^i\right) \cdot \left(\max \alpha_m^i\right)$

④ $\sum_{m=i}^{\infty} \alpha_m^i = 1 + \frac{1}{H}$.

$m \le i(1+\frac{1}{H})$, $\sum_{m=i}^{i(1+\frac{1}{H})} \frac{H}{m}$

$$Q_h^t(s,a) = \sum_{i=1}^{m} \alpha_m^i \left(r_h(s,a) + V_{h+1}^{t_i}(s_{h+1}^{t_i}) + b_i \right) + \mathbb{I}(m=0)\cdot H$$

Lemma: $\exists$ event E, $P(E) \ge 1-\delta$, $\forall t, \forall h$

$\forall s, a$.

$$0 \overset{①}{\le} Q_h^t(s,a) - Q_h^*(s,a)$$

$$\overset{②}{\le} \mathbb{I}(\underbrace{m}_{= m_h^t(s,a)}=0)\cdot H + 3\cdot\frac{c}{\sqrt{m}}\,\mathbb{I}(m>0) + \underbrace{\sum_{i=1}^{m} \alpha_m^i \phi_{h+1}^{t_i}}_{}$$

unique to RL.

Notation

$$\left[\ \phi_h^t = V_h^t(s_h^t) - V_h^*(s_h^t)\ \right]$$

Pf: Recall:

$$Q_h^t(s,a) = \sum_{i=1}^m \alpha_m^i \left(r_h(s,a) + V_{h+1}^{t_i}(s_{h+1}^{t_i}) + b_i \right)$$

$$+ \mathbb{I}(m=0)\cdot H$$

$$Q_h^*(s,a) = \underline{r_h(s,a) + \langle P(\cdot|s,a), V_{h+1}^* \rangle}$$

$$= \sum_{i=1}^m \alpha_m^i \left(\quad\quad \right)$$

$$+ \mathbb{I}(m=0)\, Q_h^*(s,a).$$

$$Q_h^t(s,a) - Q_h^*(s,a)$$

$$= \sum_{i=1}^m \alpha_m^i \Big(\underbrace{V_{h+1}^{t_i}(s_{h+1}^{t_i}) - V_{h+1}^*(s_{h+1}^{t_i})}_{a} \underbrace{+ V_{h+1}^*(s_{h+1}^{t_i}) - \langle P_h(\cdot|s,a), V_{h+1}^* \rangle}_{b} \Big)$$

$$+ \underbrace{\sum_{i=1}^m \alpha_m^i b_i}_{c},\ b_i \le \frac{c}{\sqrt{i}} \quad + \mathbb{I}(m=0)\underbrace{(H - Q_h^*(s,a))}_{d \ \geq 0},\ Q_h^*(s,a) \le H.$$

$$a = \sum_{i=1}^m \alpha_m^i \phi_{h+1}^{t_i} \quad \in [0,H] \quad\quad \in [0,H]$$

$$(b) = \sum_{i=1}^{m} \alpha_m^i \underbrace{\left(V_{h+1}^*(s_{h+1}^{t_i}) - \langle P_h(\cdot|s,a), V_{h+1}^* \rangle \right)}_{\in[-H,H]}$$

Azuma's ineq: Martingale seq. $X_1 \sim X_n$.

$|X_i| \le B_i$;

$$\left| \sum_{i=1}^{n} X_i \right| \le \sqrt{\sum_{i=1}^{n} B_i^2 \cdot \ln\frac{1}{\delta}},$$

w.p. $1-\delta$.

$$B_i = H \cdot \alpha_m^i$$

$$\le \sqrt{H^2 \cdot \underbrace{\sum_{i=1}^{m} (\alpha_m^i)^2}_{\frac{H}{m}} \ln\frac{1}{\delta}}.$$

union bound

$$(b) \in \left[-\frac{c}{\sqrt{m}}, \frac{c}{\sqrt{m}} \right].$$

$$(c) \in \left[\frac{c}{\sqrt{m}}, \frac{2c}{\sqrt{m}} \right].$$

$$(d) \in \left[0, \mathbb{I}(m=0) \cdot H \right].$$

$$Q_h^t(s,a) - Q_h^*(s,a)$$

$$\geq \sum_{i=1}^{m} \alpha_m^i \left(V_{h+1}^{t_i}(s_{h+1}^{t_i}) - V_{h+1}^*(s_{h+1}^{t_i}) \right)$$

$$+ \cancel{\left(-\frac{c}{\sqrt{m}}\right) + \frac{c}{\sqrt{m}}} + 0$$

$V_{H+1} \geq V_{H+1}^*$

$\quad \| \qquad \|$

$\quad 0 \qquad 0$

Pf of ineq ①:

backward induction on h.

base case: $h = H$, $V_{H+1} \equiv V_{H+1}^* \equiv 0$.

$$Q_H^t \geq Q_H^*.$$

inductive case: suppose for $h+1$, $\forall t$.

$$Q_{h+1}^t(s,a) \geq Q_{h+1}^*(s,a).$$

max over a on both $\Rightarrow$

$$\max_a Q_{h+1}^t(s,a) \geq \underline{V_{h+1}^*(s)}$$

$$H \geq V_{h+1}^*(s)$$

$$V_{h+1}^t(s) \geq V_{h+1}^*(s).$$

$$\Rightarrow \quad Q_h^t \geq Q_h^* . \qquad \boxtimes .$$

Pf of ②:

$$Q_h^t(s,a) - Q_h^*(s,a) \leq I(m=0)\,H + \frac{3c}{\sqrt{m}} I(m>0) + \sum_{i=1}^{m} \alpha_m^i \phi_{h+1}^{t_i} . \qquad \boxtimes .$$

Pf of thm:

$$\sum_{t=1}^{T} \left(V_1^*(s_1^t) - V_1^{\pi^t}(s_1^t) \right)$$

Lemma 1

$$\leq \sum_{t=1}^{T} \left(\underbrace{\boxed{V_1^t(s_1^t) - V_1^{\pi^t}(s_1^t)}}_{\delta_1^t} \right)$$

$\min(H, Q_h^t(s_h^t, a_h^t))$ $\qquad$ $Q_h^{\pi^t}(s_h^t, a_h^t)$ $\qquad$ ||

define $\quad \delta_h^t = \underline{V_h^t(s_h^t)} - \underline{V_h^{\pi^t}(s_h^t)}$

Recall $\quad \phi_h^t = V_h^t(s_h^t) - V_h^*(s_h^t)$

V_h^t $\quad \} \ \phi_h^t$

V_h^* $\quad \} \ \delta_h^t - \phi_h^t$ $\qquad \} \ \delta_h^t$

$V_h^{\pi^t}$

$V_n^{\pi} +$

Bounding $\sum_{t=1}^{T} \delta_1^t$?

relate $\sum_{t=1}^{T} \delta_h^t$ to $\sum_{t=1}^{T} \delta_{h+1}^t$:

$\sum_{t=1}^{T} \delta_{H+1}^t \equiv 0.$

$$\delta_h^t \leq Q_h^t(s_h^t, a_h^t) - Q_h^{\pi^t}(s_h^t, a_h^t)$$

$$= \underbrace{(Q_h^t - Q_h^*)(s_h^t, a_h^t)}_{a_t} + \underbrace{(Q_h^* - Q_h^{\pi^t})(s_h^t, a_h^t)}_{b_t}$$

$$\sum_{t=1}^{T} a_t \leq \sum_{t=1}^{T} I(\underbrace{m_h^t}_{m_h^t(s_h^t, a_h^t)} = 0) \cdot H \qquad a.1$$

$$+ \sum_{t=1}^{T} \frac{3c}{\sqrt{m_h^t}} I(m_h^t > 0) \qquad a.2$$

$$+ \sum_{t=1}^{T} \sum_{i=1}^{m_h^t} \alpha^i_{m_h^t} \cdot \phi_{h+1}^{t_i} \qquad a.3$$

$$(a.1) \le H \times S \times A \times H.$$

$$(a.2) \le \sum_{s,a} \underbrace{\sum_{t:(s_h^t, a_h^t)=(s,a)}}_{m_h^{T+1}(s,a) \text{ terms}} \frac{3c}{\sqrt{m_h^t}} I(m_h^t > 0)$$

$$= \sum_{s,a} 3c \cdot \sum_{m=1}^{m_h^{T+1}(s,a)} \sqrt{\frac{1}{m}}$$

$$\sqrt{\sum_{i=1}^{m} \frac{1}{i}}$$

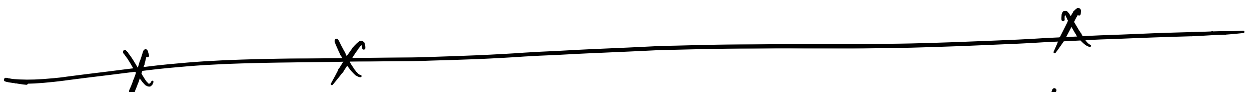
$$\le 2\sqrt{m_h^{T+1}(s,a)}$$

$$\le 6c \cdot \sum_{s,a} \sqrt{m_h^{T+1}(s,a)}.$$

Cauchy-Schwarz

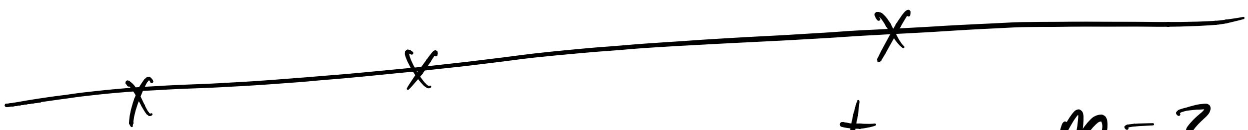
$$\le 6c \cdot \sqrt{\frac{\sum_{s,a} 1}{S \cdot A} \frac{\sum_{s,a} m_h^{T+1}(s,a)}{T}}.$$

(a.3): regroup according to $(s_h^t, a_h^t) = (s, a)$



t_1 t_2 t_m

$m = m_h^{T+1}(s,a)$.



$t_1 = 2$ $\quad t_2 = 5$ $\quad t_3 = 9$ $\quad m = 3$

$\alpha_1^1 \cdot \phi_{h+1}^2$ $\quad \alpha_2^1 \phi_{h+1}^2 + \alpha_2^2 \phi_{h+1}^5$ $\quad \alpha_3^1 \phi_{h+1}^2 + \alpha_3^2 \phi_{h+1}^5 + \alpha_3^3 \phi_{h+1}^9$

$$\sum_{s,a} \sum_{t:(s_h^t, a_h^t)=(s,a)} \sum_{i=1}^{m_h^t(s,a)} \alpha_m^i \phi_{h+1}^i$$

coeff of ϕ_{h+1}^2: $\alpha_1^1 + \alpha_2^1 + \alpha_3^1 \le 1 + \frac{1}{H}$.

ϕ_{h+1}^5: $\alpha_2^2 + \alpha_3^2 \le 1 + \frac{1}{H}$.

$$\phi^{q}_{h+1}: \quad \alpha^3_3 \le 1 + \frac{1}{H}.$$

$$\le \left(1+\frac{1}{H}\right) \cdot \sum_{t=1}^{T} \phi^t_{h+1}.$$

$$\sum_t (a_t) \le \underline{H^2 SA + 6c\sqrt{SAT}} + \left(1+\frac{1}{H}\right)\sum_{t=1}^{T} \phi^t_{h+1}$$

$$\sum_t (b_t) = \sum_t Q^*_h(s^t_h, a^t_h) - Q^{\pi_t}_h(s^t_h, a^t_h)$$

$$= \sum_t \langle P_h(\cdot \mid s^t_h, a^t_h), \underbrace{V^*_{h+1} - V^{\pi_t}_{h+1}}_{\le H\cdot\sqrt{T}} \rangle$$

$\in [-H, H]$

$$\le \sum_t \langle P_h(\cdot \mid s^t_h, a^t_h) - e_{s^t_{h+1}}, V^*_{h+1} - V^{\pi_t}_{h+1} \rangle$$

$$+ \underbrace{V^*_{h+1}(s^t_{h+1}) - V^{\pi_t}(s^t_{h+1})}_{\left(1+\frac{1}{H}\right)\left(\delta^t_{h+1} - \phi^t_{h+1}\right)}.$$

combining:

$$\sum \delta^t_h \le \left(1+\frac{1}{H}\right) \sum_t \delta^t_{h+1} + H^2 SA$$

$$+\ 6c\sqrt{SAT} + H\sqrt{T}.$$

$\sqrt{H^3}$

$$\sum_t \delta_1^t \le \left(1+\frac{1}{H}\right)^H \sum_t \delta_{H+1}^t$$

0 $\quad \le H. \quad \approx \sqrt{H^3}$

$$+\ \sum_{h=1}^{H} \left(1+\frac{1}{H}\right)^{H-h} \left(H^2SA + c\sqrt{SAT}\right)$$

$$\left(1+\frac{1}{H}\right)^H \le e.$$

$$\le O\left(H^3SA + \sqrt{H^5SAT}\right)$$

$\Rightarrow$ expected regret $\le O(\sqrt{H^5SAT})$.

$\boxtimes$.