Bound
actions before we have enough visits
to unknown states.

⇒

Define $t_1, t_2, \ldots$ rounds s.t

$|t_{i+1} - t_i| \geq H$.

&

If π_i - policy used in time i, K_i - known states,
then

$$\mathbb{P}^{\pi_i}\left[\text{escape from } K_i \mid s_0 = s_{t_i}\right] \geq \varepsilon.$$

⇒ Set:

$$X_i = \mathbb{1}\left(\exists s \text{ in } \{s_{t_i}, s_{t_i+1}, \ldots s_{t_i+H}\}: s \notin K_i\right)$$

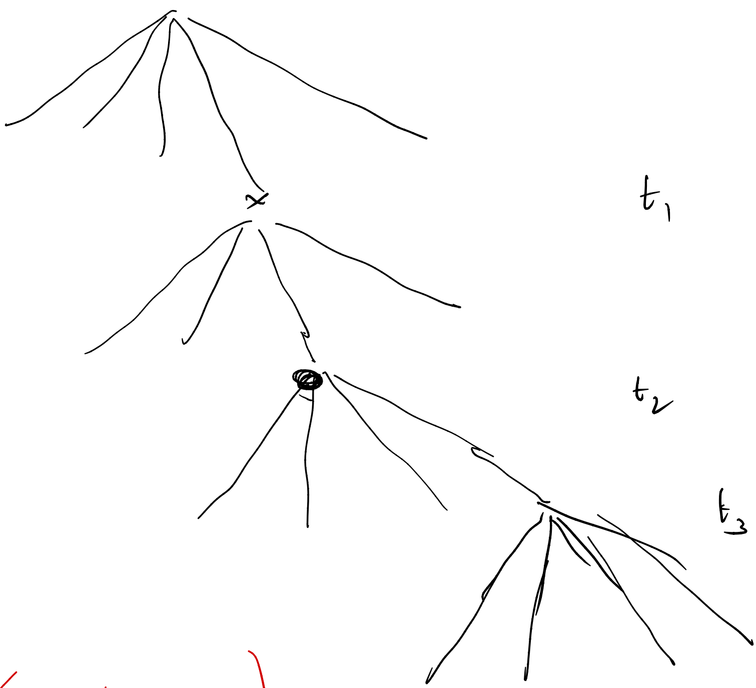
⇒ <u>By defin</u> $\mathbb{E}\left(X_i \mid s_{t_i}\right) \geq \frac{\varepsilon}{2}$.

Let $\mathcal{F}_i$ - values of all random variables prior to time t_i, including time t_i

Clearly $\mathcal{F}_\phi \subseteq \mathcal{F}_1 \subseteq \cdots$

a σ-field;

X_i is measurable w.r.t $\mathcal{F}_i$



t_1

t_2

t_3

What $\mathbb{E}(X_i \mid \mathcal{F}_{i-1})$

At time t_i — use policy π_i.

And $x_i = 1$ iff we escape from K_i in H steps; conditioned on S_{t_i}.

$$- \mathbb{E}\left[x_i \mid \mathcal{F}_{i-1} \right] \geq \frac{\varepsilon}{2};$$

Now

$$\mathbb{E}\left[\left(x_i - \mathbb{E}\left[x_i \mid \mathcal{F}_{i-1} \right] \right)^2 \,\middle|\, \mathcal{F}_{i-1} \right]$$

$$= \mathbb{E}\left[x_i^2 - 2 x_i \mathbb{E}\left[x_i \mid \mathcal{F}_{i-1} \right] + \mathbb{E}\left[x_i \mid \mathcal{F}_{i-1} \right]^2 \,\middle|\, \mathcal{F}_{i-1} \right]$$

$$= \mathbb{E}\left[x_i^2 \mid \mathcal{F}_{i-1} \right] - \mathbb{E}\left[x_i \mid \mathcal{F}_{i-1} \right]^2$$

$$\leq \mathbb{E}\left[x_i^2 \mid \mathcal{F}_{i-1} \right] = \mathbb{E}\left[x_i \mid \mathcal{F}_{i-1} \right] \quad \Leftarrow x_i = \begin{cases} 0 \\ 1 \end{cases}$$

Freedmans inequality:

Let $X_1, X_2, \ldots, X_T$ be a sequence of real valued random variables adapted to the filtration $\mathcal{F}_t$.
$\therefore X_i$ is measurable w.r.t $\mathcal{F}_i$ & further assume that $\mathbb{E}\left[X_i \mid \mathcal{F}_{i-1}\right] < \infty$;

Define $S = \sum_{t=1}^{T} X_t$, $V = \sum_{t=1}^{T} \mathbb{E}\left(X_t^2 \mid \mathcal{F}_{t-1}\right)$ and let

$X_t \le R$ almost surely $\forall t$;
$\forall \delta \in (0,1)$ & $\lambda \in [0, 1/R]$ with prob at least $1-\delta$,

$$S \le (e-2)\lambda V + \frac{\ln(1/\delta)}{\lambda}$$

Choosing: $\lambda = \min\left(\frac{1}{R}, \sqrt{\frac{\ln(1/\delta)}{V}}\right)$ we get

$$S \le 2\sqrt{V \ln(1/\delta)} + R \ln(1/\delta).$$

- Set $Y_i = \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i$

Applying Friedman's inequality:

$$\sum_{i=1}^{n} Y_i$$

$$= \sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i$$

$$\leq 2\sqrt{\ln\left(\frac{1}{\delta}\right) \sum_{i=1}^{n} \mathbb{E}\left[\left(\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i\right)^2 \mid \mathcal{F}_{i-1}\right]} + \ln(1/\delta)$$

$$\leq 2\sqrt{\ln(1/\delta) \sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]} + \ln(1/\delta)$$

$$\leq \frac{1}{2}\sum \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] + 3\ln(1/\delta)$$

$$\therefore \sum X_i \geq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - 3\ln(1/\delta)$$

$$\geq \frac{n\varepsilon}{2} \cdot \frac{1}{2} - 3\ln(1/\delta)$$

Want: $\displaystyle\sum_{i=1}^{n} x_i \geq mSA$

$$\therefore \frac{n\varepsilon}{4} - 3\ln\left(\frac{1}{\delta}\right) \geq mSA.$$

$$n \geq \frac{4}{\varepsilon}\left[mSA + 3\ln(1/\delta)\right]$$: Set $n = O\left(\frac{m|S||A|\log(1/\delta)}{\varepsilon}\right)$

- for rounds $t \in [t_{i+H+1}, t_{i+1} - 1]$ – prob of escape $\leq \varepsilon$

$\therefore$ value function is near optimal on those rounds.

So with prob $(1-\delta)$, $\sum x_i \geq mSA$ which means all states will be known.

- The # states which are not ε optimal?

Each of the rounds t_i, $i=1,\dots,n$

Gives us $O\left(\frac{m|S||A|\ln(1/\delta)}{\varepsilon}\right)$ states.

And for each of them H more rounds.

$\therefore$ At most $O\left(\frac{m|S||A|}{\varepsilon} H \ln(1/\delta)\right)$ states.

<u>Note</u>: This is independent of the # episodes.

Where an episode is that period where K remains the same.

Weaker bound

EXPLAINING ALL THIS WITHOUT FRIEDMAN.

- What we have shown is that the # rounds where $V_M^{\pi_t}(s_t) \leq V_M^{\pi^*(M)}(s_t) - \varepsilon$

is at most $O\left(\frac{mHSA}{\varepsilon} \ln(1/\delta)\right)$.

Note: Think of an episode as the rounds where K remains the same. (ie) policy does not change.

For the value H chosen, our t_i's are such that escape prob in an episode is at least $\varepsilon/2$ $[x_i = 1]$

- We want to estimate # episodes where we may fail;

\# state action pairs, $= |S||A|$.

and a state is known if we see each

each state action pair m times.

- $\therefore$ An upper bound on # state action pairs is $m|S||A|$.

If we have $\frac{m|S||A|}{\varepsilon}$ episodes, we expect more

than $\frac{m|S||A|}{\varepsilon} \times \varepsilon = m|S||A|$ successes. In which

case we are done

Set $n = \frac{m|S||A|}{\varepsilon} \log\left(\frac{|S||A|}{\delta}\right)$

Then the expected # successes $\geq m|S||A| \log\left(\frac{|S||A|}{\delta}\right)$

- All $m|S||A|$ explorations succeed with probability $\geq 1-\delta$:

Main theorem:

Let s_t be the state visited at round t, & let $m = O\left(\frac{SH^2}{\varepsilon^2}\log\left(\frac{S^2A}{\delta}\right)\right)$. For any $\varepsilon \geq 0$, $\delta < 1$, w.p $1-\delta$, $V_M^{\pi_t}(s_t) \geq V_M^*(s_t) - \varepsilon$ for all but

$O\left(\frac{H^3 S^2 A}{\varepsilon^3}\log\left(\frac{S^2A}{\delta}\right)\right)$ rounds in the MDP.

Proof:

Assume m is large.

And we have an approximation $\hat{M}_k$ of the induced chain.

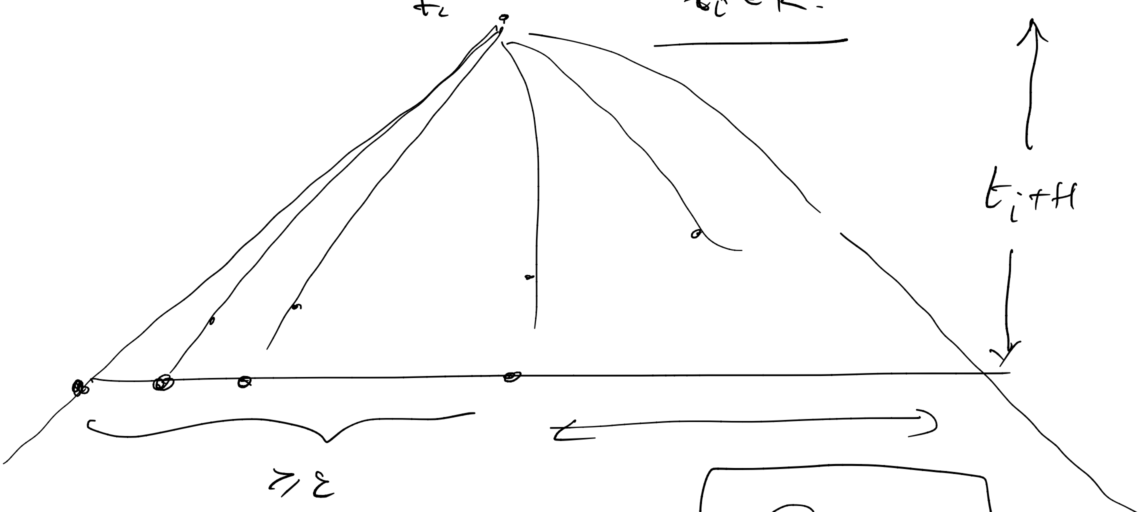
Assume that $\sum |P_{M_k}(s'|s,a) - P_{\hat{M}_k}(s'|s,a)| \leq \varepsilon_1$

so that $\| V_{M_k}^{\pi} - V_{\hat{M}_k}^{\pi} \| \leq \frac{\gamma}{1-\gamma}\varepsilon_1 = \frac{\varepsilon}{2}$.

$(*)_{t' \le t_i - 1}$

value of all random variable.

t_i s_i $s_{t_i} \in K$?



$\geq \varepsilon$

$$\underbrace{V^{\pi_t}_{M_K}(s) \geq V^{\pi_t}_{\hat{M}_K}(s)}_{\varepsilon/2 \text{ closed}} - \frac{\varepsilon}{2}$$

$\boxed{1/\varepsilon \ \# \text{number of steps}}$

Visits unknown state with prob $\geq \varepsilon/\,$

$$\underbrace{\varepsilon \leq P[s \notin K] + \gamma \Pr[s \in K, s_1 \notin K] + \gamma^2 P - -}$$