Recall R-max-γ algorithm.

- $\hat{M}_K$ - induced MDP:

Def: Let $M = (S, A, P, r, \gamma)$; $K \subseteq S$ - known.

$\forall s \in K$

$$P_{M_K}(s' | s,a) = P_M(s' | s,a) \;\&\; r_{M_K}(s,a) = r_M(s,a)$$

$\forall s \notin K$:

$$P_{M_K}(s' | s,a) = \mathbb{1}(s' = s) \;\&\; r_{M_K}(z | s,a) = \mathbb{1}(z = 1)$$

Stay back in the same state

Reward = 1.

Set $H = \dfrac{\log\left(2/\varepsilon(1-\gamma) \right)}{1-\gamma}$

WAS THE # iterations for Q-value-iteration to converge with V^{π^k} with ε ball of V^* (in l_∞ norm)

THM: Let s_t be the state visited by R-Max-γ algorithm in round t and set $m = \left(\frac{SH^2}{\varepsilon^2}\log\left(\frac{S^2A}{\delta}\right)\right)$

for $\varepsilon > 0$, $\delta < 1$,

$$V_M^{\pi_t}(s_t) \geq V_M^*(s_t) - \varepsilon \quad \forall \text{ but } O\left(\frac{H^3SA}{\varepsilon^3}\log\left(\frac{S^2A}{\delta}\right)\right)$$

rounds in the MDP.

We're only talking of $V(s_t)$ in round t.

- We may want to prove that it finds a near-opt policy, (i.e.) a policy whose expected reward is within ε of the optimal, when taking expectations over the start state as well.

WE DON'T PROVIDE SUCH A GUARANTEE!

- OPTIMISM in the face of uncertainty.

*: In the proof below r also refers to expected reward!

Lemma:

Let M, M' be MDPs on same action & state spaces. If

$$\sum_{s' \in S} \left| P_M(s'|s,a) - P_{M'}(s'|s,a) \right| \le \varepsilon_1 \quad \forall s, a.$$

&

$$\left| r_M(s,a) - r_{M'}(s,a) \right| \le \varepsilon_2 \quad \forall s, a$$

then

$\forall \pi$, stationary policies, $\| V_M^{\pi} - V_{M'}^{\pi} \|_\infty \le \frac{\gamma \varepsilon_1}{1-\gamma} + \varepsilon_2$

Proof:

$$\left| V_M^{\pi}(s) - V_{M'}^{\pi}(s) \right| \le$$

$$(1-\gamma)\varepsilon_2 + \gamma \left| \sum_{s' \in S} P_M(s'|s,\pi(s)) V_M^{\pi}(s') - P_{M'}(s'|s,\pi(s)) V_{M'}^{\pi}(s') \right|$$

$$\le (1-\gamma)\varepsilon_2 + \gamma \Big(\sum_{s' \in S} P_M(s'|s,\pi(s)) \left[V_M^{\pi}(s') - V_{M'}^{\pi}(s') \right] + \sum_{s'} \left[P_M(s'|s,\pi(s)) - P_{M'}(s'|s,\pi(s)) \right] V_{M'}^{\pi}(s') \Big)$$

$$\leq (1-\gamma)\varepsilon_2 + \gamma \left\| V_M^{\pi}(s) - V_{M'}^{\pi}(s) \right\|_\infty + \gamma \varepsilon_1$$

$$\therefore \left\| V_M^{\pi}(s) - V_{M'}^{\pi}(s) \right\|_\infty \leq \varepsilon_2 + \frac{\gamma}{1-\gamma} \varepsilon_1$$

NOTATION:

$$\mathbb{P}_M^{\pi}\left[\text{escape from } K \mid s_0 = s\right] = \mathbb{1}(s \notin K) + \sum_{t=1}^{\infty} \gamma^t \, \mathbb{P}_M^{\pi}\left(s_t \notin K, s_0, \ldots, s_{t-1} \in K\right)$$

discounted

The above term is the discounted probability of reaching an unknown state when executing π, starting from state s

Lemma 3.4:

Let M be an MDP, K - known state & M_K as before. $\forall \pi$, stationary & $s \in S$,

$$V^{\pi}_{M_K}(s) \geq V^{\pi}_{M}(s) \quad \&$$

$$V^{\pi}_{M}(s) \geq V^{\pi}_{M_K}(s) - \mathbb{P}^{\pi}_{M}\left[\text{escape from } K \,\middle|\, s_0 = s\right]$$

- M_K is an optimistic version of M since it gives greater value to all states $\forall$ stationary policies.

- But value of optimistic policy not too high! The difference is high only if there is a large probability of escaping to an unknown state!

- $V^{\pi}_{M_K}(s) \geq V^{\pi}_{M}(s)$ is clear.

- If $s \notin K$, get a max reward of 1
  $(1-\gamma)(1+\gamma+\gamma^2+\dots) = 1$.

- If $s \in K$ - immediate reward is exactly that in M
  - subsequent steps - remain in K and get same reward
  - leaving K & get maximum reward!

- $$\left| V^{\pi}_{M}(s) - V^{\pi}_{M_K}(s) \right| \leq$$
$$\mathbb{1}(s \notin K) + \mathbb{1}(s \in K)\gamma \left[\sum_{s' \in S} P_M(s' | s, \pi(s)) V^{\pi}_{M}(s') - P_{M_K}(s' | s, \pi(s)) V^{\pi}_{M_K}(s') \right]$$

Here $\because s \in K$, $P_M(s' | s, \pi(s)) = P_{M_K}(s' | s, \pi(s))$

$$\leq \mathbb{1}(s \notin K) + \mathbb{1}(s \in K)\gamma \left[\sum_{s'} P_M(s' | s, \pi(s)) \left(V^{\pi}_{M}(s') - V^{\pi}_{M_K}(s') \right) \right]$$

$$\leq \mathbb{1}(s \notin K) + \mathbb{1}(s \in K)\gamma P_M(s' \notin K \mid s, \pi(s))$$

$$+ \mathbb{1}(s \in K)\gamma \Big| \sum_{s' \in K} P_M(s' \mid s, \pi(s)) \underbrace{\left(V_M^\pi(s') - V_{M_K}^\pi(s')\right)}_{\text{expand this again}} \Big|$$

$\underline{s' \in K}$, get

$$\gamma \sum_{s'' \in S} \mathbb{1}(s' \in K) \left[P_M(s'' \mid s', \pi(s')) V_M^\pi(s'') - P_{M_K}(s'' \mid s', \pi(s')) V_{M_K}^\pi(s'') \right]$$

which;

if $s'' \notin K$ gives $\gamma P_M(s'' \notin K \mid \underbrace{s'}_{\in K}, \pi(s'))$

If $s'' \in K$ gives

$$\gamma \sum_{s''' \in S} P_M(s''' \mid s'', \pi(s'')) \left[V_M^\pi(s''') - V_{M_K}^\pi(s''') \right]$$

$\underline{\text{Coeff of } \gamma^2},$ $\quad \gamma^2 \mathbb{1}(s \in K) \mathbb{1}(s' \in K) \left[\text{Prob of reaching } s'' \notin K \text{ in 2 steps} \right]$

· (ic) Prob of escaping from K in 2 steps $* \gamma^2$.

· As we expand this:

$$\leq \mathbb{1}(s \notin K) + \mathbb{1}(s \in K) \cdot \gamma P_{M}(s' \notin K \mid s, \pi(s))$$

$$+ \mathbb{1}(s \in K)\, \mathbb{1}(s' \in K)\, \gamma^2 P_{M}(s'' \notin K \mid s, s' \in K)$$

$$+ \mathbb{1}(s \in K)\, \mathbb{1}(s' \in K)\, \mathbb{1}(s'' \in K)\, \gamma^3 P_{M}(s''' \notin K \mid \underset{\in K}{s, s', s''})$$

$\Uparrow$

exactly $\mathbb{P}^{\pi}_{M}[\text{escape from } K \mid s_0 = s]$

* Jump

– lemma

- With sufficiently many visits to states with a large escape probability, all states become known. } whp
- With high probability the # rounds following a visit where the policy's value function is significantly suboptimal is at most H.

COR:

Implicit Explore - Exploit:

$$V_M^{\pi^*(M_K)}(s) \geq V_M^*(s) - P_M^{\pi^*(M_K)}[\text{escape for } K \mid s_0 = s]$$

From lemma:

$$V_M^{\pi^*(M_K)}(s) \geq V_{M_K}^{\pi^*(M_K)}(s) - \mathbb{P}_M^{\pi^*(M_K)}[\text{escape for } K \mid s_0 = s]$$

Why $\geq V_{M_K}^{\pi^*(M)}(s) - \mathbb{P}_M^{\pi^*(M_K)}[\text{escape for } K \mid s_0 = s]$

Why. $\geq V_M^{\pi^*(M)}(s) - \mathbb{P}_M^{\pi^*(M_K)}[\text{escape for } K \mid s_0 = s]$

Policy computed in each episode is near optimal

Lemma:

with probability at least $1-\delta$, the # rounds t with $V_M^{\pi_t}(s_t) \leq V_M^*(s_t) - \varepsilon$ is at most $\mathcal{O}\left(\frac{mHSA}{\varepsilon}\log(1/\delta)\right)$

Proof:

2 steps.

1) If in round t, the prob. of escape from K starting from s_t is large, then we have a large escape probability in the next H steps.

2) In fact, with high probability, the algorithm encounters a unknown state in the next H steps.

1) $s = s_t$, π_t; Suppose $\mathbb{P}_M^{\pi_t}[\text{escape from } K \mid s_0 = s] \geq \varepsilon$

Define: $$p_H = \mathbb{1}(s \notin K) + \sum_{t=1}^{H} \underbrace{P_M^{\pi_t}(s_t \notin K \mid s_0, \ldots, s_{t-1} \in K)}_{\text{Prob of escaping from } K \text{ in } t \text{ steps}}$$

- Let s_t = state encountered at round t & π_t be the policy.

Suppose $\mathbb{P}^{\pi}_{M}\left[\text{escape from } K \mid s_0 = s\right] \geq \varepsilon$.

- Set

$$\underbrace{p_H = \mathbb{1}(s \notin K) + \sum_{t=1}^{H} \mathbb{P}^{\pi}_{M}\left(s_t \notin K \mid s_0, \ldots s_{t-1} \in K\right)}$$

Probability of escaping from K in H steps!

No discounted probability!!

$$\varepsilon \leq \mathbb{1}(s \notin K) + \sum_{t=1}^{\infty} \gamma^t \, \mathbb{P}^{\pi}_{M}\left(s_t \notin K \mid s_0, \ldots s_{t-1} \in K\right)$$

$$\leq p_H + \sum_{t=H+1}^{\infty} \gamma^t \leq p_H + \frac{\gamma^{H+1}}{1-\gamma}$$

for H as chosen $\dfrac{\gamma^{H+1}}{1-\gamma} \leq \varepsilon/2$

for this H $\boxed{p_H \geq \varepsilon/2}$

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

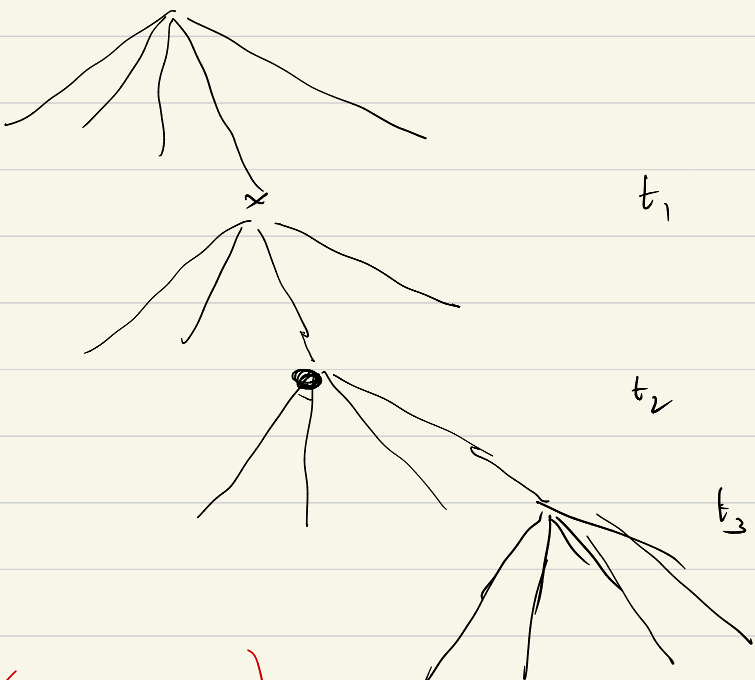
⇒ By defin
$$\mathbb{E}\left(X_i \mid s_{t_i}\right) \geq \varepsilon.$$

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 $\mathcal{F}_i$



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}.

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

Now

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

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

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

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

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

$X_t \leq R$ almost surely $\forall t$;

$\forall \delta \in (0,1)$ & $\lambda \in [0, 1/R]$ with prob at least $1-\delta$,

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

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

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

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

Applying Friedman's inequality:

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

$$= \sum_{i=1}^{n} \mathbb{E}\left[X_i \middle| \mathcal{F}_{i-1}\right] - X_i$$

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

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

$$\le \frac{1}{2}\sum \mathbb{E}\left[X_i \middle| \mathcal{F}_{i-1}\right] + 3\ln(1/\delta)$$

$$\therefore \quad \sum X_i \ge \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[X_i \middle| \mathcal{F}_{i-1}\right] - 3\ln(1/\delta)$$

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

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

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

$$n \geq \frac{4}{\varepsilon}\left[mSA + 3\ln(1/\delta)\right].$$

- 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.