Analysis of the $R_{max,\gamma}$ algorithm:

Note that we only really need a π which tells us what to do for H steps - so it need not be a policy at all.

(ie) a stationary policy.

However what we show is that the optimal π^*_K obtained from M_K is good! So in fact we get a policy, a stationary policy.

- This π^*_K is what we use till K is updated.

- We could even work with an optimal H step policy in M_K. That too would work.

- But we work with the optimal policy in M_K.

- In fact with the "empirical" $\hat{M}_K$, and the optimal policy there since we are only estimating M_K.

LAST TIME:

Let M be an MDP, K - known states. M_K, the induced MDP. For any stationary policy π and state $s \in S$,

$$V^{\pi}_{M_K}(s) \geq V^{\pi}_{M}(s) \quad \&$$
$$V^{\pi}_{M}(s) \geq V^{\pi}_{M_K}(s) - \Pr\left[\text{escape}_K \text{ from} \,\middle|\, s_0 = s\right]$$

Cor:

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

Pf:

$$V_M^{\pi^*(M_K)}(s) \geq V_{M_K}^{\pi^*(M_K)}(s) - \mathbb{P}_{\sigma_M}^{\pi^*(M_K)}\left[\text{escape from}_K \mid s_0 = s\right]$$

$$\geq V_{M_K}^{\pi^*(M)}(s) - \mathbb{P}_{\sigma_M}^{\pi^*(M_K)}\left[\text{escape from } K \mid s_0 = s\right)$$

$$\geq V_M^{\pi^*(M)}(s) - \mathbb{P}_{\sigma_M}^{\pi^*(M_K)}\left[\text{escape from } K \mid s_0 = s\right]$$

- We will prove the main theorem.

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:

We should that the value function will be ε away from the optimal for at most

$$O\left(\frac{m H |S||A|}{\varepsilon}\log\left(\frac{|S||A|}{\delta}\right)\right)$$

rounds.

- We do not have the Markov Chain.

- We only sample from it!

So we cannot really compute M_k, only an approximation $\hat{M}_k$.

And we will simulate that!

We showed if:

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

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

for every policy π (stationary)

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

· We'll use this, assuming that $\underline{\varepsilon_1 = \varepsilon/2}$.

Then: $\forall t$,

$$\left| V^{\pi_t}_{M_K}(s) - V^{\pi_t}_{\hat{M}_K}(s) \right|_\infty \le \varepsilon/2$$

$\therefore$ $\forall s$,

π_t is optimal w.r.t. $\hat{M}_K$

$$\underbrace{V^{\pi_t}_{M_K}(s)} \ge V^{\pi_t}_{\hat{M}_K}(s) - \frac{\varepsilon}{2} \ge V^{\pi^*}_{\hat{M}_K}(s) - \frac{\varepsilon}{2} \ge \underbrace{V^{\pi^*}_{M_K}(s) - \varepsilon}$$

· So, $\forall s$,

$$V^{\pi_t}_{M}(s) \ge V^{\pi_t}_{M_K}(s) - \mathbb{P}^{\pi_t}_{M}\left[\text{escape from } K \mid s_0 = s\right]$$

$$\ge V^{\pi^*}_{M_K}(s) - \mathbb{P}^{\pi_t}_{M}\left[\text{escape from } K \mid s_0 = s\right] - \varepsilon$$

$$\ge V^{\pi^*}_{M}(s) - \varepsilon - \mathbb{P}^{\pi_t}_{M}\left[\text{escape from } K \mid s_0 = s\right]$$

$\therefore$ If $Pr_{\mu}^{\pi_t}$ [escape from $K \mid s_0 = s$] $\leq \varepsilon$, we are within 2ε of the optimal value.

O.W we have a good chance of escaping!

and then we know for all but

$$O\left(\frac{mH|S||A|}{\varepsilon}\log\frac{|S||A|}{\delta}\right)$$ rounds we

are within ε,

$\therefore$ we are within 2ε $\forall$ but rounds.

Now to find m, so M_K & $\hat{M}_K$ are close!

Lemma:

Assume m samples are obtained from a distribution p, whose support is of size N. $\hat{p}$, empirical distribution.

If $m = O\left(\frac{N^2}{\varepsilon^2} \log\left(\frac{N}{\delta}\right)\right)$, with prob $1-\delta$,

$$\sum_i |\hat{p}(i) - p(i)| \leq \varepsilon.$$

* Can improve this. will be in H.W.

We select m, samples; And for each i we calculate,

$$\frac{\# \text{ i's seen}}{m};$$

If we ensure $\Pr\left[|\hat{p}(i) - p(i)| \leq \frac{\varepsilon}{N}\right] \geq 1 - \frac{\delta}{N}$

then with prob $(1-\delta)$, $\forall i$, $\Pr\left[|\hat{p}(i) - p(i)| \leq \frac{\varepsilon}{N}\right]$

and so with prob $1-\delta$,

$$\sum_i \left| \hat{p}(i) - p(i) \right| \le \frac{\varepsilon}{N} \cdot N \le \varepsilon .$$

For one i, use Hoeffding;

$$\Pr\left[\left| \sum_k X_k^i - \sum_k \mathbb{E}(X_k^i) \right| \ge t \right] \le \exp\left(\frac{-2t^2}{\sum_{k=1}^{m} (b_k^i - a_k^i)^2} \right)$$

Here $b_i - a_i = 1$.

<u>So</u>: $$\Pr\left[\left| \frac{\sum X_k^i}{m} - p_i \right| \ge \frac{t}{m} \right] \le \exp\left(\frac{-2t^2}{m} \right)$$

$$\frac{t}{m} = \frac{\varepsilon}{N} \quad \therefore \; t = \frac{\varepsilon m}{N} ;$$

$$\le \exp\left(- \frac{2 \cdot 4 \cdot \varepsilon^2 m^2}{N^2 \cdot m} \right)$$

$$\le \frac{1}{\exp\left(\frac{8 \varepsilon^2 m}{N^2} \right)} \le \frac{\delta}{N} \quad \text{if}$$

$$m \ge \frac{N^2}{8 \varepsilon^2} \log\left(\frac{N}{\delta} \right)$$

Want:

$$\sum_{s'} \left| \mathbb{P}_{M_K}(s' \mid s,a) - \mathbb{P}_{\hat{M}_K}(s' \mid s,a) \right| \le \varepsilon_1$$

then: $$\left\| V^{\pi_t}_{M_K}(s) - V^{\pi_t}_{\hat{M}_K}(s) \right\| \le \underbrace{\frac{\gamma}{1-\gamma}\varepsilon_1}$$

Want: $\le \varepsilon/2$.

What about ε_1?

We want for H consecutive transitions that the two Markov chains agree to ε_1 of their transition probabilities

But error accumulates to $\varepsilon_1 H$ in H rounds.
So we can only guarantee $\frac{\gamma}{1-\gamma}\varepsilon_1 H$ closeness!

$\therefore$ Set $\frac{\gamma}{1-\gamma}\varepsilon_1 H = \frac{\varepsilon}{2}$

$$\therefore \boxed{\varepsilon_1 = \frac{1-\gamma}{\gamma} \frac{\varepsilon}{2H}}$$

$\boxed{\delta}$

. Want this confidence for all (s,a)

$\therefore$ We can afford an $\frac{\delta}{|S||A|}$ error for one (s,a), and this is what we plug into

$$m \geq \frac{N^2}{8\varepsilon_0^2} \log\left(\frac{N}{\delta_0}\right)$$

Finally: $\varepsilon_0 = \frac{1-\gamma}{\gamma} \frac{\varepsilon}{2H}$; $\delta_0 = \frac{\delta}{|S||A|}$

$$\therefore m \geq \frac{N^2 \cdot 4H^2}{8\varepsilon^2} \left(\frac{\gamma^2}{(1-\gamma)^2}\right) \log\left[\frac{N|S||A|}{\delta}\right]$$

Here $N = |S||A|$.

$$\therefore O\left(\frac{S^2A^2H^2}{\varepsilon^2} \log\left[\frac{|S|^2|A|^2}{\delta}\right]\right)$$

Improves to:

$$O\left(\frac{|S||A| H^L}{\varepsilon^2} \log \frac{|S||A|}{\delta}\right).$$

- Next chapter is policy gradient.
- We look at more Monte Carlo methods. See Barto & Sutton.

- ESTIMATING STATE-VALUE FUNCTION for a policy π (* π is fixed)

Start from s $\to$

Generate a history $(S_0 = s, A_0, R_0, S_1, a_1, r_1, \ldots)$

Compute $\sum_{k=0}^{\infty} \gamma^k R_k$

We are constructing an "unbiased" estimator of $v^\pi(s)$

Alg:

Input π :

$v \leftarrow 0$; $R(s) \leftarrow$ null $\forall s$;

for $i = 1, \dots k$

Generate an episode $s_{0,i}, a_{0i}, r_{0i}, s_{1i}, a_{1i}, r_{1i} \dots\dots$

for each state s appearing in episode,

$t \leftarrow$ time of first occurrence of s.

$G_{i,s} \leftarrow \sum \gamma^k R_{t+k}$

Append G_i to $R(s)$

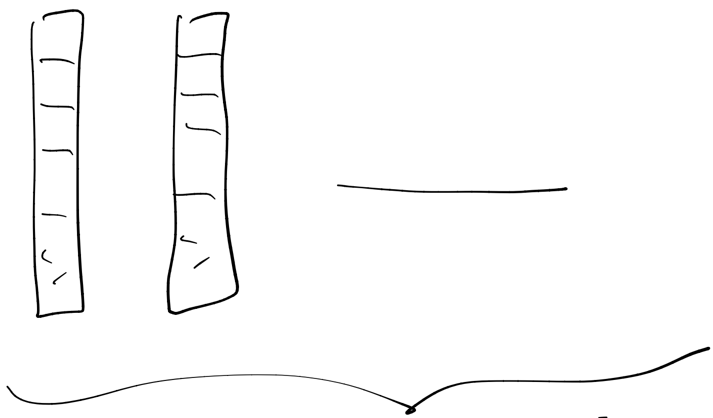
end

$$v(s) = \frac{1}{k} \sum G_i.$$

Q: Can we compute this using last occurrence of s?

- $\mathbb{E}(G_i, s) = v^\pi(s)$;

- i.i.d samples $\therefore$ $\mathbb{E}\left(\frac{\sum G_i}{n}\right) = v^\pi(s)$;



finite states $\therefore$ finite vectors;

Works if every state is visited infinitely often

A variant:

Every visit Monte Carlo:

- for each state s appearing in the episode and each time t it occurs,

- find the discounted reward from that time
- append all such rewards to the list (s);
- Average of the rewards is $\hat{v}_\pi(s)$;

Gradient based Monte Carlo:

- Start with $v \in \mathbb{R}^{|S|}$; At time t

$$v(S_t) \leftarrow v(S_t) + \alpha \left[G_t - V(S_t) \right]$$

- Mean square error $(v) = \frac{1}{2} \mathop{\mathbb{E}}_{S} \left[\left(v^\pi(s) - v(s) \right)^2 \right]$

Gradient descent:

$$v \leftarrow v - \alpha \frac{\partial MSE(v)}{\partial v}$$

$$= v - \alpha \, \mathbb{E} \left(\left(v^\pi(s) - v(s) \right)(-1) \frac{\partial v(s)}{\partial v} \right)$$

$$= v + \alpha\, \mathbb{E}\left[\left(v^{\pi}(s) - v(s) \right) \frac{\partial v(s)}{\partial v} \right]$$

$$\begin{array}{cc} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} & \leftarrow s \end{array}$$

$$\boxed{V_{new}(s_t) = V_{old}(s_t) + \alpha \left(G_t - V_{old}(s_t) \right)}$$

$$\mathbb{E}(G_t) = v^{\pi}(s_t)$$