Policy iteration:

- Start with π_0 arbitrary;

for $k = 0 \ldots$

- Compute Q^{π_k} (Can be computed analytically $(1-\gamma)\left(I-\gamma P^{\pi}\right)^{-1} r$)

- Update $\pi_{k+1} = \pi_{Q^{\pi_k}}$

Recall $P^{\pi}_{(s,a)(s',a')} = P[s' \mid s,a]\,\pi(a' \mid s')$.

- $\therefore$ Policy evaluation followed by policy improvement.

Lemma: 1) $Q^{\pi_{k+1}} \geq TQ^{\pi_k} \geq Q^{\pi_k}$

2) $\|Q^{\pi_{k+1}} - Q^*\|_\infty \leq \|Q^{\pi_k} - Q^*\|_\infty$

Observe:

Our policies are deterministic.

$$\therefore V^{\pi_k}(s) = Q^{\pi_k}(s, \pi_k(s)) \quad \forall k, \forall s.$$

$$\therefore TQ^{\pi_k}(s,a) = (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q^{\pi_k}(s', a')\right]$$

$$\geq (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[Q^{\pi_k}(s', \pi_k(s'))\right]$$

i.e.

$$TQ^{\pi_k}(s,a) \geq Q^{\pi_k}(s,a).$$

Now:

$$Q^{\pi_{k+1}}(s,a) = (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[Q^{\pi_{k+1}}(s', \pi_{k+1}(s'))\right]$$

$$\geq (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[Q^{\pi_k}(s', \pi_{k+1}(s'))\right]$$

$$= (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q^{\pi_k}(s', a')\right]$$

$$= TQ^{\pi_k}(s,a).$$

Now:

$$\|Q^* - Q^{\pi_{k+1}}\|_\infty \le \|Q^* - TQ^{\pi_k}\|_\infty = \|TQ^* - TQ^{\pi_k}\|_\infty$$

$$\le \gamma \|Q^* - Q^{\pi_k}\|_\infty.$$

$$\le \gamma^k \; '' \; {}^{*} - Q_0\|_\infty \le \quad e^{-(1-\gamma)k}$$

$\therefore$ If $k \geq \dfrac{\log(1/\varepsilon)}{1-\gamma}$

$$\le \exp^{-\log(1/\varepsilon)}$$

$$\le \underline{\underline{\varepsilon}}.$$

Thm:

For any two policies π, π', we have:

Note – $Q_\pi(s,a)$

$$\left(\forall s \in S, \; \mathbb{E}_{a \sim \pi'(s)} \left[Q_\pi(s,a) \right] \geq \mathbb{E}_{a \sim \pi(s)} \left[Q_\pi(s,a) \right] \right) \Rightarrow$$

$$\left(\forall s, \; V_{\pi'}(s) \geq V_\pi(s) \right)$$

Proof:

$$V_\pi(s) = (1-\gamma) \mathop{\mathbb{E}}_{a\sim\pi(s)} Q_\pi(s,a)$$

Note: We can ignore $(1-\gamma)$ here, or ignore it here.

$$\leq (1-\gamma) \mathop{\mathbb{E}}_{a\sim\pi'(s)} Q_\pi(s,a)$$

$$= (1-\gamma) \left[\mathop{\mathbb{E}}_{a\sim\pi'(s)} \left[r(s,a) + \gamma V_\pi(s_1) \,\middle|\, s_0 = s \right] \right]$$

$$= (1-\gamma) \mathop{\mathbb{E}}_{a\sim\pi'(s)} \left[r(s,a) + \gamma \mathop{\mathbb{E}}_{a_1\sim\pi(s_1)} \left(Q_\pi(s_1,a_1) \right) \,\middle|\, s_0 = s \right]$$

$$\leq (1-\gamma) \mathop{\mathbb{E}}_{a\sim\pi'(s)} \left[r(s,a) + \gamma \mathop{\mathbb{E}}_{a_1\sim\pi'(s_1)} \left(Q_\pi(s_1,a_1) \right) \,\middle|\, s_0 = s \right]$$

$$= (1-\gamma) \mathop{\mathbb{E}}_{a\sim\pi'(s)} \left[r(s,a) + \gamma r(s_1,a_1) + \gamma^2 V_\pi(s_2) \,\middle|\, s_0 = s \right]$$

$$\vdots$$

$$\Rightarrow$$

$$V_\pi(s) \leq (1-\gamma) \mathop{\mathbb{E}}_{a\sim\pi'(s_t)} \left[\sum_{t=0}^{T} \gamma^t \mathbb{E}\left(r(s_t,a_t) \right) + \gamma^{T+1} V_\pi(s_{T+1}) \,\middle|\, s_0 = s \right]$$

But $V_\pi(s_{T+1})$ is bounded $\therefore \gamma^{T+1} V_\pi(s_{T+1}) \to 0$ as $T \to \infty$

$\therefore$ Taking limit as $T \to \infty$,

$$V_\pi(s) \le (1-\gamma)\left[\mathop{\mathbb{E}}_{a_t \sim \pi'(s_t)}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[r(s_t, a_t)\right] \middle| s_0 = s\right]\right]$$

$$= V_{\pi'}(s);$$

---x---

We get.

THM: (Bellman's optimality condition)

A policy π is optimal iff for any pair $(s,a) \in S \times A$ with $\pi(s)(a) > 0$ (recall $\pi(s)$ is a distribution on actions) the following holds

$$a \in \operatorname*{argmax}_{a' \in A} Q_\pi(s, a').$$

Proof: Clear.

If this condition does not hold for some (s,a) then π is not optimal.

Consider π' s.t. $\pi'(s') = \pi(s)$ for $s' \neq s$ and $\pi'(s)$ is concentrated on $\underset{a' \in A}{\operatorname{argmax}} \, Q_\pi(s, a')$.

For π', $V_{\pi'}(s) \gneq V_\pi(s) \; \forall s$.

$\Longleftarrow$ Conversely

If π' is not optimal, $\exists \, \pi$ with $V_\pi(s) > V_{\pi'}(s)$ some s;

But then $\exists \, s$ with
$$\mathbb{E}_{a \sim \pi'(s)} \left[Q_\pi(s,a) \right] < \mathbb{E}_{a \sim \pi(s)} \left[Q_\pi(s,a) \right].$$

THM: $\exists$ a optimal det policy.

SAMPLE COMPLEXITY with a GENERATIVE MODEL.

- What is the sample complexity of estimating Q^*?

- Will assume that the reward function is known & deterministic.

- Assume we have access to a generative model

- Given s, a provides a sample $s' \sim \mathbb{P}(\cdot \mid s, a)$.

- Invoke our simulator N times $\forall$ $S \times A$ pair.

Define:

$$\hat{P}(s' \mid s, a) = \frac{\#(s', s, a)}{N}$$

times our simulator transits to s' from s on a.

Let $\hat{M}$ = empirical MDP.

Uses $\hat{P}$ instead of P.

· First Hoeffding's inequality:

· Recall Markov's inequality:

For any nonnegative random variable X

$$\Pr[X > t] \leq \frac{E(X)}{t}.$$

In fact:

if ϕ is a strictly monotonically $\uparrow$ non-negative-valued function then for any random variable X and $t \in \mathbb{R}$,

$$\Pr[X > t] = \Pr[\phi(X) > \phi(t)] \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)}.$$

=

Using $\underline{\phi(a) = a^2}$:

$$\Pr[|X - \mathbb{E}(X)| > t] = \Pr[(X - \mathbb{E}(X))^2 > t^2]$$

$$\leq \frac{\mathbb{E}[(X - E(X))^2]}{t^2}$$

$$= \frac{Var(X)}{t^2}$$ ... Chebyshev.

- Taking $\phi(x) = e^{sx}$, s arbitrary +ve number

$$\Pr[X > t] = \Pr[e^{sX} > e^{st}] \leq \frac{\mathbb{E}[e^{sX}]}{e^{st}}.$$

find $s > 0$ which makes the RHS small!

CHERNOFF.

· Sums of independent random variables:

$S_n = X_1 + \cdots + X_n$, X_i independent real-valued random variables.

$$\Pr\left[\,|S_n - \mathbb{E}(S_n)| > t\,\right] \leq \frac{\mathrm{Var}(S_n)}{t^2}$$

$$= \frac{\sum \mathrm{Var}(X_i)}{t^2}$$

Writing $\sigma^2 = \frac{1}{n} \sum \mathrm{Var}(X_i)$ we get

Then $$\Pr\left[\,\left|\frac{S_n}{n} - \mathbb{E}(X_i)\right| > \varepsilon\,\right] \leq \frac{\sigma^2}{n\varepsilon^2}$$

- Not very satisfactory.

∵ Central limit thm:

$$\Pr\left[\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum x_i - \mathbb{E}(x_i)\right) \geq y\right] \longrightarrow 1-\phi(y) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-y^2/2}}{y}$$

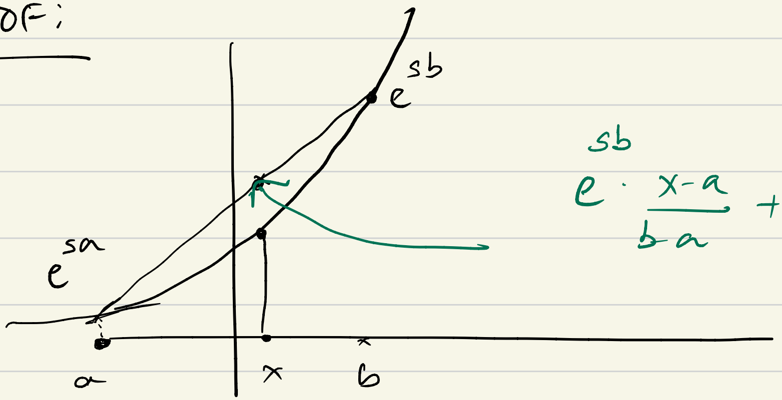
Expect: $\Pr\left[\frac{1}{n} \sum x_i - \mathbb{E}(x_i) \geq \varepsilon\right] \simeq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$

Much smaller than what Chebyshev gives!

HOEFFDING: Let x be a r.v with $\mathbb{E}(x)=0$, $a \leq x \leq b$, $\forall s > 0$,

$$\mathbb{E}\left[e^{sx}\right] \leq e^{s^2(b-a)^2/8}$$

PROOF:



$$e^{sb} \cdot \frac{x-a}{b-a} + e^{sa} \cdot \frac{b-x}{b-a}$$

$$\therefore \quad e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}$$

$$\therefore \quad \mathbb{E}\left[e^{sx}\right] \leq \frac{\mathbb{E}(X-a)}{b-a} e^{sb} + \frac{\mathbb{E}(b-X)}{b-a} e^{sa}$$

$$= \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}$$

Setting $p \triangleq -\dfrac{a}{b-a} \quad \therefore \quad 1-p = 1 + \dfrac{a}{b-a} = \dfrac{b}{b-a}$

$$= (1-p)\, e^{-sp(b-a)} + p \cdot e^{s(b-a)(1-p)}$$

$$= \left[1-p+pe^{s(b-a)}\right] e^{-sp(b-a)}$$

Set

$$u = s(b-a)$$

$$\mathbb{E}\left[e^{sX}\right] \le e^{\phi(u)},$$

$$\phi(u) = -pu + \log\left(1-p+pe^{u}\right).$$

$$\phi'(u) = -p + \frac{pe^{u}}{1-p+pe^{u}}$$

$$= -p + \frac{p}{p+(1-p)e^{-u}}$$

Now $\phi(0) = 0$; $\phi'(0) = 0$

$$\phi''(u) = \frac{p(1-p)e^{-u}}{\left(p+(1-p)e^{-u}\right)^2} \le \frac{1}{4}$$

$\theta \in [0, u]$

$$\therefore \phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta)$$

$$\le \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$$

$$\therefore \mathbb{P}\left[S_n - \mathbb{E}(S_n) \geq t \right]$$

$$\le e^{-st} \, e^{\sum_{i=1}^{n} s(x_i - \mathbb{E}(x_i))}$$

$$= e^{-st} \, e^{sx_1} \cdot e^{sx_2} \cdots e^{sx_n}$$

$$\le e^{-st} \prod e^{\frac{s^2(b_i-a_i)^2}{8}}$$

$$= \underbrace{e^{-st} \, e^{s^2} \cdot e^{\sum_{i=1}^{n} \frac{(b_i-a_i)^2}{8}}}$$

Choose $s = \dfrac{4t}{\sum_{i=1}^{n}(b_i-a_i)^2}$

$$\le e^{-2t^2 / \sum_{i=1}^{n}(b_i-a_i)^2}$$

Thm: HOEFFDING's tail inequality.

Let $X_1, \dots, X_n$ be independent r.v s.t

$X_i \in [a_i, b_i]$ with prob. 1. Then $\forall t > 0$

$$\Pr\left[S_n - \mathbb{E}(S_n) \geq t\right] \leq e^{-\frac{2t^2}{\sum (b_i - a_i)^2}}$$

$$\Pr\left[S_n - \mathbb{E}(S_n) \leq -t\right] \leq e^{-\frac{2t^2}{\sum (b_i - a_i)^2}}$$

Coming back to the empirical model:

$$\hat{P}(s' \mid s, a) = \frac{m(s', s, a)}{n}$$

n = # calls to the model per (s, a) pair.

$\therefore$ # calls $n|S||A|$.

Based on $\hat{P}$ we get

$$\hat{T} Q(z) = r(z) + \gamma(\hat{P} V)(z), \quad V(x) = \max_{a \in A} Q(x, a)$$

Can also define $\hat{T}^{\pi}$ - operator on Q as

$$\hat{T}^{\pi} Q(z) = r(z) + \gamma \hat{P}^{\pi} Q(z).$$

- Can define the above operators for V as well.

<u>Algorithm</u>: Model-based Q value iteration.

Input: Q_0; n - samples per $S \times A$ pair;
$k = \#$ iterations.

$\hat{P}$ = ESTIMATE MODEL (n)

for $j = 0, \ldots, k-1$ do
  for each $s \in S$ do
    $\pi_j(s) = \arg\max_{a \in A} Q_j(s, a)$
    for each $a \in A$
      $\hat{T} Q(s,a) = r(s,a) + \gamma \left(\hat{P}^{\pi_j} Q_j\right)(s,a)$
    end
      $Q_{j+1}(s,a) = \hat{T} Q_j(s,a)$
  end
end
Return Q_K

Model based policy iteration:

Input: reward, γ, π_0, n, k.

$\hat{P} = \text{ESTIMATEMODEL}(n)$

$Q_0 = \left(I - \gamma \hat{P}^{\pi_0}\right)^{-1} r;$

for $j = 0, 1, \ldots k-1$

for each $s \in S$ do

$\pi_j(s) = \arg\max_{a \in A} Q_j(s, a)$

end

$\hat{Q}^{\pi_j} = \left(I - \gamma \hat{P}^{\pi_j}\right)^{-1} r$

$Q_{j+1} = \hat{Q}^{\pi_j}$

end

return Q_K;

denote $S \times A$ by $\mathcal{Z}$;

MODELESTIMATE (n):

$\forall (s, z) \in S \times \mathcal{Z}$ set $m(s, z) = 0$

for each $z \in \mathcal{Z}$

for $i = 1, \ldots, n$ do

$s \sim P(\cdot \mid z)$

$m(s, z) = m(s, z) + 1$

end.

$\forall s \in S$, $\hat{P}(s \mid z) = \dfrac{m(s, z)}{n}$

end

return $\hat{P}$

ASSUMPTIONS: $Z = S \times A$ is finite; Assume $r(s,a) \in [0,1]$;

THEOREM: $\exists$ constants c, c_0, d & d_0 s.t $\forall \varepsilon \in (0,1), \forall \delta \in (0,1)$, a total sampling budget

$$T = \left\lceil \frac{c\,|S|\,|A|}{(1-\gamma)^3 \cdot \varepsilon^2} \log\left(\frac{c_0\,|S|\,|A|}{\delta}\right) \right\rceil$$

suffices for $\|Q^* - Q_k\|_\infty \le \varepsilon$ with probability $1-\delta$, after $k = \left\lceil d \log\left(\frac{d_0}{\varepsilon(1-\gamma)}\right) \Big/ \log\left(\frac{1}{\gamma}\right) \right\rceil$ iterations of QVI or PI

The above theorem holds for finding a near optimal policy

The bounds are almost tight;