finite data $\xrightarrow{①}$ error in models $\rightarrow$ error in resulting ② value function

① Finite data yields error in model parameter estimates
if have limited amount of samples sampled from a stochastic model, always possible that if estimate model parameters from that data, that model parameters will be different than estimates

but under some assumptions can bound difference
- 1st consider estimating reward model $r(s)$ for a particular state s where the rewards are drawn randomly iid given the state with some unknown mean $\mu(s)$.
- assume all rewards are bounded in $[0, R_{max}]$
- given n samples
- can use Hoeffding inequality to compare empirical average $\frac{1}{n}\sum_{i=1}^{n} r_i(s)$ where $r_i^{(s)}$ is the reward received on the i^{th} sample of state s

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} r_i(s) - \mu(s)\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-2n^2\epsilon^2}{nR_{max}^2}\right)$$

set right hand side to δ (implies larger dev happens with prob $\leq \delta$)

solve for ϵ

$$2\exp\left(-2n^2\epsilon^2/(nR_{max}^2)\right) = \delta$$

$$2n\epsilon^2/R_{max}^2 = \log(2/\delta)$$

$$\epsilon = R_{max}\sqrt{\frac{\log(2/\delta)}{2n}}$$

implies for a given number of samples n, with prob at least $1-\delta$, the estimate of the mean reward for state s will be within $R_{max}\sqrt{\frac{\log(2/\delta)}{2n}}$ of the true mean.

- Similar expression holds for discrete state-action transition models (Weissman et al. 2003)

define $\hat{p}(s'|s,a) = \frac{\#\ (s,a,s')\ \text{tuples}}{\#\ (s,a)\ \text{tuples}}$

then with prob at least $1-\delta$

$$\underbrace{\|\hat{p}(\cdot|s,a) - p(\cdot|s,a)\|_1}_{L_1\ \text{distance between two } |s|\text{-dim vectors}} \leq \sqrt{\frac{2\left(\ln\left(2^{|s|}-2\right) - \ln(\delta)\right)}{\#\ s,a\ \text{tuples}}}$$

- Simulation lemma [1st by Kearns & Singh 1999/2002, many similar later uses & slight variations]

let MDPs $M_1 = (S, A, T_1, R_1, \gamma)$ and $M_2 = (S, A, T_2, R_2, \gamma)$ be two MDPs with the same state & action space and whose rewards are bounded between 0 & R_{max}.

If $|R_1(s,a) - R_2(s,a)| \leq \alpha \quad \forall s,a$ and

$\|T_1(s,a,\cdot) - T_2(s,a,\cdot)\| \leq \beta \quad \forall s,a$

Then the following condition holds for all s,a & stationary, deterministic policies π

$$|Q_1^\pi(s,a) - Q_2^\pi(s,a)| \leq \frac{\alpha + \gamma \frac{R_{max}}{1-\gamma} \cdot \beta}{(1-\gamma)}$$

proof

$$|Q_1^\pi(s,a) - Q_2^\pi(s,a)| = |R_1(s,a) + \gamma \textstyle\sum_{s'} T_1(s,a,s') V_1^\pi(s') - R_2(s,a) - \gamma \sum_{s'} T_2(s,a,s') V_2^\pi(s')|$$

(triangle inequality)

$$\leq |R_1(s,a) - R_2(s,a)| + \gamma |\textstyle\sum_{s'} T_1(s,a,s') V_1^\pi(s') - T_2(s,a,s') V_2^\pi(s')|$$

$$= |R_1(s,a) - R_2(s,a)| + \gamma |\textstyle\sum_{s'} T_1(s,a,s') V_1^\pi(s') - T_2(s,a,s') V_1^\pi(s') + T_2(s,a,s') V_1^\pi(s') - T_2(s,a,s') V_2^\pi(s')|$$

$$\leq |R_1(s,a) - R_2(s,a)| + \gamma |\textstyle\sum_{s'} (T_1(s,a,s') - T_2(s,a,s')) V_1^\pi(s')| + \gamma |\textstyle\sum_{s'} T_2(s,a,s') (V_1^\pi(s') - V_2^\pi(s'))|$$

(triangle inequality)

$$\leq \alpha + \gamma V_{max} |\textstyle\sum_{s'} T_1(s,a,s') - T_2(s,a,s')| + \gamma |\textstyle\sum_{s'} T_2(s,a,s') (V_1^\pi(s') - V_2^\pi(s'))|$$

(upper bound 1st 2 terms)

$$\leq \alpha + \gamma V_{max} \beta + \gamma \max_{s,a} |Q_1^\pi(s,a) - Q_2^\pi(s,a)| |\textstyle\sum_{s'} T_2(s,a,s')|$$

2nd term by assumption
3rd term upper bounded by max Q diff

this expression holds for any arbitrary (s,a) pair so must also hold for pair that yields max diff

$$\max_{s,a} |Q_1^\pi(s,a) - Q_2^\pi(s,a)| \leq \alpha + \gamma V_{max} \beta + \gamma \max_{s,a} |Q_1^\pi(s,a) - Q_2^\pi(s,a)|$$

(since T_2 must $\sum$ to 1)

$$(1-\gamma) \max_{s,a} |Q_1^\pi(s,a) - Q_2^\pi(s,a)| \leq \alpha + \gamma V_{max} \beta$$

$$\max_{s,a} |Q_1^\pi(s,a) - Q_2^\pi(s,a)| \leq \frac{\alpha + \gamma V_{max} \beta}{(1-\gamma)}$$

$$\leq \frac{\alpha}{1-\gamma} + \gamma \frac{R_{max} \beta}{(1-\gamma)^2} \quad \text{using } V_{max} \leq \frac{R_{max}}{1-\gamma}$$

□

Note the above holds for all π, including π_2^* for M_2 so if compute the optimal π_2^* for M_2 and then execute that π_2^* in MDP M_1, the resulting state-action values will be a bounded distance from the values predicted in M_2

putting the two ideas together

assume have $n(s,a)$ samples for a state action pair (s,a)

let M_1 be the true MDP

M_2 be the MDP whose parameters are estimated from data

Let π_2^* be the optimal π for MDP M_2

$$|Q_1^{\pi_2^*}(s,a) - Q_2^{\pi_2^*}(s,a)| \leq \min_{n(s,a)} R_{max} \sqrt{\frac{\log(2|S||A|/\delta)}{2n(s,a)}} + \frac{\gamma R_{max} \sqrt{\frac{2(\ln(2^{|S|}-2) - \ln(\delta/|S||A|)}{n(s,a)}}}{(1-\gamma)^2}$$

exercise

let $R_{max} = 1$ $\quad |S| = 2$ $\quad |A| = 2$

$\gamma = 0.9$ $\quad \delta = .2$

how many samples (min n) to make bound not trivial? (e.g. $\geq V_{max}$)

$$\frac{\sqrt{\frac{\log(2|S||A|/\delta)}{2n(s,a)}}}{(1-\gamma)} + \frac{\gamma \sqrt{\frac{2(\ln(2^{|S|}-2) - \ln(\delta/|S||A|)}{n(s,a)}}}{(1-\gamma)^2} \leq \frac{1}{1-\gamma} = V_{max}$$

$$\sqrt{\frac{\log(2|S||A|/\delta)}{2n(s,a)}} + \frac{\gamma}{1-\gamma} \sqrt{\frac{2(\ln(2^{|S|}-2) - \ln(\delta/|S||A|)}{n(s,a)}} \leq 1$$

2nd term dominates generally so

$$\frac{\gamma^2}{(1-\gamma)^2} \cdot 2(\ln(2^{|S|}-2) + \ln(|S||A|/\delta) \leq n(s,a)$$

$$260 \leq n(s,a)$$ ← till not vacuous!

but if actually have 260 samples for a (s,a) pair, likely have a great model

bounds looser than experiments suggest is needed