- Bias and Variance Approximation in Value Function Estimates

goal: provide confidence intervals (CI) over the future performance of a $\pi$ given existing data

could use simulation lemma but likely to be too wide to be informative

"parametric bias & variance"

how does uncertainty over model parameter estimates (due to limited data) contribute to uncertainty over the value function estimate for a policy $\pi$

1st consider given a policy $\pi$ and want to compute CI over $V^\pi$

assumptions

finite $|S|$ & $|A|$ space

prior set of data

$n(s_i, a)$ = # times tried action a in state $s_i$

$n(s_i, a, s_j)$ = # times tried action a in state $s_i$ & transitioned to state $s_j$

$c(s_i, a, s_j)$ = sum of rewards received in trying action a in state $s_i$ & transitioning to state $s_j$

assume $n(s_i, a) > 0 \quad \forall s_i \in S$ and $\forall a \in A$

if $n(s_i, a, s_j) = 0$ then $c(s_i, a, s_j) = 0$

$\pi(a|s_i)$ = prob of selecting action a in state $s_i$

define the approximate transition $\hat{T}$ & reward $\hat{R}$ models as

$$\hat{T}(s_i, a, s_j) = \frac{n(s_i, a, s_j)}{n(s_i, a)} \qquad \hat{R}(s_i, a, s_j) = \frac{c(s_i, a, s_j)}{n(s_i, a, s_j)} \text{ or } 0 \text{ if } n(s_i,a,s_j)=0$$

unbiased estimates of T & R

$$\hat{T}^\pi(s_i, s_j) = \sum_a \pi(a|s_i)\hat{T}(s_i, a, s_j)$$

$$\hat{R}^a(s_i) = \sum_j \hat{T}(s_i, a, s_j)\hat{R}(s_i, a, s_j) = \frac{\sum_j c(s_i, a, s_j)}{n(s_i, a)}$$

$$\hat{R}^\pi(s_i) = \sum_a \pi(a|s_i)\hat{R}^a(s_i)$$

performing direct policy evaluation using estimated models

$$\hat{V}^\pi = \underbrace{(I - \gamma \hat{T}^\pi)^{-1}}_{\text{nonlinear in } \hat{T}^\pi} \hat{R}^\pi$$

want to obtain estimate of bias $\quad V^\pi - E[\hat{V}^\pi]$

covariance of $\hat{V}^\pi \quad \text{cov}(\hat{V}^\pi) = E[\hat{V}^\pi(\hat{V}^\pi)^T] - E[\hat{V}^\pi]E[\hat{V}^\pi]^T$

use a $2^{nd}$ order approximation

Proposition. The expectation over $\hat{V}^\pi$ satisfies

$$E[\hat{V}^\pi] = V^\pi + \gamma^2 (I-\gamma T^\pi)^{-1} U V^\pi + \gamma (I-\gamma T^\pi)^{-1} B + L_{exp}$$

where $U_{ij} = cov^{(i)}_j (I-\gamma T^\pi)^{-1}_{:,i}$ row j

$$B_i = \sum_a \frac{\pi(a|s_i)^2}{n(s_i,a)} r(s_i,a) M_i^a (I-\gamma T^\pi)^{-1}_{:,i}$$ ← index certain column

$$M_i^a = \mathrm{diag}(T(s_i,a,\cdot)) - (T(s_i,a,\cdot)^T T(s_i,a,\cdot))$$

$$cov^{(i)} = \sum_a \frac{\pi(a|s_i)^2}{n(s_i,a)} M_i^a$$

$$L_{exp} = o\left(\frac{1}{\min_{(i,a)\ \pi(a|s_i)>0} n(s_i,a)}\right)$$ $o(\cdot)$ $\lim_{N\to\infty} o\left(\frac{1}{N}\right)\cdot N \to 0$

☆ → U & B decrease like $(1/\min_{(i,a)} n(s_i,a))$

corollary 4.1

$$cov(\hat{V}^\pi) = (I-\gamma T^\pi)^{-1} W ((I-\gamma T^\pi)^{-1})^T + o\left(\frac{1}{\min_{n(i,a)\ \pi(a|s_i)>0} n(s_i,a)}\right)$$

$$W_{ii} = \sum_a \frac{\pi(a|s_i)^2}{n(s_i,a)} \left[(\gamma (V^\pi)^T + r(s_i,a,\cdot)) \cdot M_i^a (\gamma V^\pi + r(s_i,a,\cdot)^T) + V^{a\,T}_{i,:} P^a_{i,:}\right]$$

$$V_{ij}^a = Var(r(s_i,a,s_j))$$

note

1) as $n(s_i,a) \to \infty$ $\forall s_i,a$

- $cov^i \to 0$
- $U \to 0$
- $B \to 0$
- $W \to 0$
- bias $\to 0$
- variance $\to 0$

2) expressions depend on true unknown parameters

use estimates instead

☆ 3) stdev goes to zero at a rate of $1/\sqrt{\min_{i,a} n(s_i,a)}$

use asymptotic normal assumption to get CI

bias goes to 0 at a rate of $1/\min_{i,a} n(s_i,a)$

stdev " " " " " " $1/\sqrt{\min_{i,a} n(s_i,a)}$

} so expect variance to dominate uncertainty in estimate

how does this compare to result from simulation lemma & bounds on uncertainty of model?

rate of error convergence is similar $1/\sqrt{\min_{s,a} n(s,a)}$

but constants matter. Prior bound had a $\left(\frac{1}{1-\gamma}\right)^2$ term in front...

but cov term also will yield something similar due to $I-\gamma T^\pi$ terms

The problem w/control

Typically also want to use data to estimate a good $\pi$ (instead of just policy eval)

one idea: use data to compute model parameters
compute $\hat{\pi}^*$ for those parameters
then use above procedure to estimate $\hat{V}^{\hat{\pi}^*}$ and CI around it

are $\hat{T}^{\pi}$ & $\hat{R}^{\pi}$ still unbiased estimates of $T^{\pi}$ & $R^{\pi}$?

no

example

MDP with 1 state 2 actions

$R(a_1) = R(a_2) = \mathcal{N}(0, \sigma^2)$

$V^{a_1} = V^{a_2} = 0$

Jensen's inequality: $f(E[X]) <= E[f(X)]$ $f$ convex $X$ $\mathbb{R}^n$ vector

$f(x_1 \ldots x_n) = \max(x_1 \ldots x_n)$ convex

$E[\hat{V}^{\pi}] = E[\max(\hat{R}^{a_1}, \hat{R}^{a_2})] \geq \max[E[\hat{R}^{a_1}], E[\hat{R}^{a_2}]]$

$= \max[0, 0]$

$= 0$

↳ see ex. of why max is convex on next pg

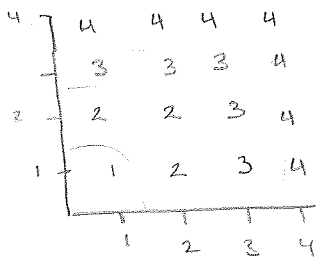
solution

split into train & test
train $\pi$ on part of data
use other part to evaluate

raises important implementation issue: how to split data?
- if have larger training set to compute policy, expect to get better policy
- if have larger test set expect to get better estimate of $\pi$'s performance (bias & variance estimate scale like $1/\min_{i,a} n(s_i,a_i)$ & $1/\sqrt{\min_{i,a} n(s_i,a_i)}$ respectively)

* note: prior approach using simulation lemma did not directly address this issue. It provided a policy evaluation estimate & bounds on its accuracy. If one uses same data to compute $\pi$ & evaluate it, likely same issue as above

max is convex




max function

$f(\vec{X}) = \max_i (X_1 \dots X_n)$

$f(rx + (1-r)y) \le r f(x) + (1-r) f(y)$

$x = [1\ \ 3]$ $y = [5\ \ 2]$

$r = .5$ $rx + (1-r)y = [3\ \ 2.5]$ $\max([3\ \ 2.5]) = 3$

$.5 \cdot 3 + .5 \cdot 5 = 4$

$f(E[x]) \le E[f(x)]$ convex

$E[\hat{V}^{\pi}] = E[\max(\hat{R}^0, \hat{R}^1)] \ge \max[E[\hat{R}^0], E[\hat{R}^1]] = 0$