Evaluating Model Based Approaches to Batch RL Part 2

Previously discussed bounding estimation error when using a model to evaluate a π
- simulation lemma + model approx bounds
- bias & variance (Mannor, Tsitsiklis paper)

this approach can be used for a fixed model
what about when we have a set of possible models
and/or a set of possible features to define the model?

* Idea 1: Use feature selection methods discussed earlier to directly estimate V^π
now generally using value func based criteria for fit instead of model
Parr's method (ICML 2008, "An Analysis of Linear...") showed some model produce a V that is exactly equiv to LSTD soln
can use feature selection on those models
but criteria discussed so far has generally focussed on best approximating (using diff objectives) the value function estimated from training samples
not on generalization error
also the previously discussed approaches were for feature selection/generation to evaluate a single π
when evaluating many π, may want diff features/models
may impact empirical convergence

* Idea 2: Choose model that best fits the data process
let $\mathcal{M}$ be a MDP
let D be the batch data, consisting of N, H-length trajectories

$$\mathcal{M} = \arg\max_{M} P(D|M)$$

for ex for a single traj $(s_1, a_1, r_1, s_2, a_2, r_2, \ldots) = \tau$

$$P(\tau|M) = \prod_{i=1}^{H} \underbrace{p(s_{i+1}|a_i, s_i, M)}_{\text{prob of transition under model}} \underbrace{p(r_i|a_i, s_i, M)}_{\text{prob of reward under model}}$$

Is this a good idea? If have true model, is that sufficient for best π / V^π?

- 1st concern: overfitting
  finite data, might overfit
  soln: use AIC/BIC (to penalize complicated models)
  use cross validation

- bigger problem: will true model maximize likelihood?
  or satisfy min descrip length (MDL)
  $MDL(\mathcal{M}, D) = \underbrace{f(\mathcal{M}, D)}_{\text{some sublinear func}} - \log P(D|\mathcal{M})$ where $\mathcal{M}$ is some model

  e.g. if $\mathcal{M}$ has k params
  $AIC(\mathcal{M}, D) = \underbrace{2k}_{f(\mathcal{M}, D)} - 2\log P(D|\mathcal{M})$

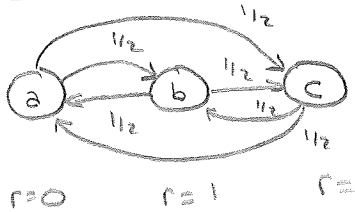
  $BIC(\mathcal{M}, D) = k \log(|D|) - 2\log P(D|\mathcal{M})$
  ($|D|$ = # data pts)

  MDL designed to help address overfitting

example 1

consider (from Hallak, Di-Castro & Mannor ICML 2013) 2 MDPs

MDP M_T



3 state MDP, 1 action
$P(s_{t+1} | s_t, a_t) = 1/2$ if $s_t \neq s_{t+1}$
true MDP

MDP M_L

alternate model

what is transition model?
reward " ?

r = 0 w/prob 1/3
r = 1 w/prob 2/3

consider single H-step traj
what is prob under M_T? $(1/2)^H$
M_L? $(1/3)^{H/3}(2/3)^{2H/3} \approx .53^H$

which will we choose under max likelihood? M_L
what if penalize # params? M_L
Is M_L a good model?
No, allows reward seq that are impossible, like 00

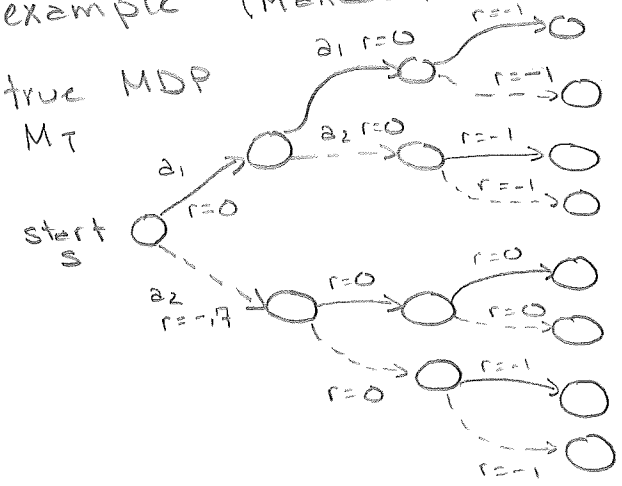
Why?
is new model Markov?
Can we know in advance if a model is Markov?

* Idea 3: Compare V_M^π for different M (or even diff π)
where estimate V_M^π from model M w/parameters
estimated from data

- What is the potential problem w/this?
  - overfitting (but could use one data set to compute π & another to estimate M's params and perform V_M^π, ala "Bias and Variance..." paper)
- Still can be a problem

example (Mandel, Liu, Levine, Brunskill & Popovic AAMAS 2014)

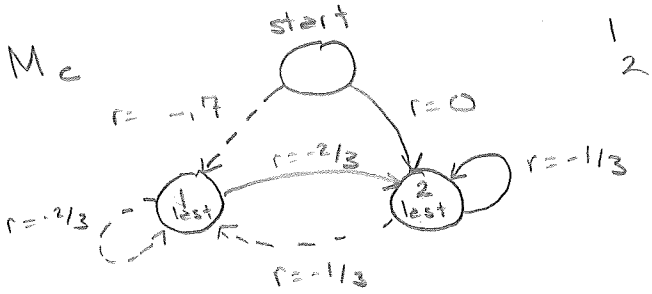
example 2



— = a_1 --- = a_2

3 step horizon
2 actions

optimal π? 2 1 1/2
value = -.7 (γ=1



1 last means took a_1
2 " " " a_2

assume have ∞ data

optimal 3-step π? 1 1 1
Value? -2/3

would chose M_c as $V_{M_c}^{\pi^*_{M_c}} > V_{M_T}^{\pi^*_{M_T}}$

but in M_T π: [1 1 1] would yield $V_{M_T}^{\pi^*_{M_c}} = -1$

why? M_c not Markov
state aliasing

- what if we want to compare different models (& corresponding π), some of which may be partially observable MDPs; fairly in terms of predicted performance?

importance sampling!

but compare $\pi_e(a_t \mid \underbrace{s_t, a_{t-1}, s_{t-1} \dots s_0}_{\text{full history } h_{t-1}})$ to $\pi_b(a_t \mid s_t, a_{t-1} \dots)$

(π_e: evaluation π; π_b: behavior / sampling π)

$$V_{IS}^{\pi_e} = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi_e(a_{1i}, a_{2i} \dots a_{Hi} \mid \tau_i)}{\pi_b(a_{1i}, a_{2i} \dots a_{Hi} \mid \tau_i)} \sum_{t=1}^{H} \gamma^{t-1} r_{ti}$$

$$= \sum_{t=1}^{H} \frac{1}{N} \sum_{i=1}^{N} \frac{\pi_e(a_{1i} \dots a_{ti} \mid h_{t-1})}{\pi_b(a_{1i} \dots a_{ti} \mid h_{t-1})} \gamma^{t-1} r_{t,i}$$

(since rewards only depend on prior steps)

still have problem phil mentioned, can be very high var in long horizon problems

but in short horizon with a broad π_b, decent

→ show examples in slides

* Another idea: reward based criteria (Hallak, Di-Castro, & Mannor)

define a traj $\tau = (o_t, a_t, r_t)_{t=1}^{H}$ where o_t is an observation

MDP M will define a state space

d-delayed reward error is

$$RE_d(\tau) = \frac{1}{H} \sum_{t=1}^{H-d} \left(r_{t+d} - \underbrace{E[R_{t+d} \mid s_t, a_t]}_{\text{empirical estim of rewards obtain from } s_t, a_t \text{ after } d \text{ steps}} \right)^2 + \underbrace{|S|}_{\#\text{states in MDP}} \frac{f(H)}{H}$$

($f(H)$: sublinear)

$$\lim_{H \to \infty} \frac{f(H)}{H} \to 0$$

exercise: show for ∞ data, RE_0 will choose correct model for example 1

thm: If $\forall s_1, s_2 \in S \; E[R_{t+d} \mid s_t = s_1] \neq E[R_{t+d} \mid s_t = s_2]$ then RE_d is weakly consistent (weak consistency means that given finite set of models incl. true model, given ∞ data, prob of selecting non true model $\to 0$)

show plots

not always clear what d in RE_d is needed

can use criteria to choose a model even if noise are perfect