

Lecture 6

Local/online planning, part 2

Value iteration


$$\pi_k(s) = \operatorname{argmax}_a q_{k+1}(s, a)$$

$$q_k = \tilde{T}^k \mathbf{0}$$

$$\tilde{T}q = r + \gamma PMq$$

$$k \geq H_{\gamma, \delta(1-\gamma)/(2\gamma)}$$

```
1. define q(k,s):  
2.   if k = 0 return 0 # base case  
3.   return [ r(s,a) + gamma * sum( [P(s,a,s') * max(q(k-1,s')) for s' in S] ) for a in A ]  
4. end
```



Cost: $O((SA)^k)$

Deterministic systems

Next state: $g(s, a)$

```
1. define q(k,s):  
2.   if k = 0 return 0 # base case  
3.   return [ r(s,a) + gamma * max(q(k-1,g(s,a))) for a in A ]  
4. end
```

Cost: $O(A^k)$ – independent of S

Stochastic systems

```
1. define q(k,s):  
2.   if k = 0 return 0 # base case  
3.   return [ r(s,a) + gamma/m * sum( [max(q(k-1,s')) for s' in C(s,a)] ) for a in A ]  
4. end
```

$O((mA)^k)$ runtime

```
1. define C(s,a):  
2.   if (s,a) in _C: return _C[(s,a)]  
3.   _C[(s,a)] = [ sim.nextstate(s,a) for i in range(m) ]  
4.   return _C[(s,a)]  
5. end
```

```
1. define C(s,a):  
2.   return [ sim.nextstate(s,a) for i in range(m) ]  
3. end
```

(Kearns, Mansour and Ng, 2002)
“sparse lookahead trees”

we go with this..

$$\boxed{\pi: S \rightarrow \mathcal{M}_1(\mathcal{A})}$$

$$\pi(s) \in \mathcal{M}_1(\mathcal{A})$$

$$\pi(s)(a) = \pi(a|s)$$

$$\forall s \in S$$

$$\sum_a \pi(a|s) \mathbb{I}(q^*(s,a) \geq v^*(s) - \epsilon) \geq 1 - \zeta$$

$$A \sim \pi(\cdot|s) : \mathbb{P}(q^*(s,A) \geq v^*(s) - \epsilon) \geq 1 - \zeta$$

Lemma (Policy error bound II): Let $\zeta \in [0, 1]$, π be a memoryless policy that selects ϵ -optimizing actions with probability at least $1 - \zeta$ in each state. Then,

$$v^\pi \geq v^* - \left[\frac{\epsilon + 2\zeta \|q^*\|_\infty}{1 - \gamma} \right] \mathbf{1}.$$

$$H \approx \frac{1}{1 - \gamma}$$

Lemma (Policy error bound - I.): Let π be a memoryless policy and choose a function $q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and $\epsilon \geq 0$. Then, the following hold:

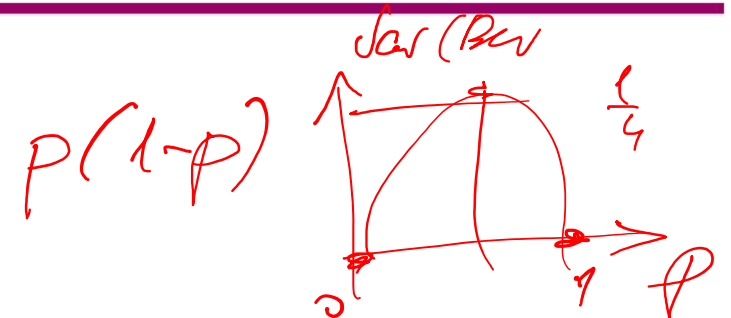
- 1 If π is ϵ -**optimizing** in the sense that $\sum_a \pi(a|s)q^*(s, a) \geq v^*(s) - \epsilon$ holds for every state $s \in \mathcal{S}$ then π is $\epsilon/(1 - \gamma)$ suboptimal: $v^\pi \geq v^* - \frac{\epsilon}{1-\gamma} \mathbf{1}$.
- 2 If π is greedy with respect to q then π is 2ϵ -optimizing with $\epsilon = \|q - q^*\|_\infty$ and thus

$$v^\pi \geq v^* - \frac{2\|q - q^*\|_\infty}{1 - \gamma} \mathbf{1}.$$

Lemma (Hoeffding's Inequality): Given m independent, identically distributed (i.i.d.) random variables that take values in the $[0, 1]$ interval, for any $0 \leq \zeta < 1$, with probability at least $1 - \zeta$ it holds that

$$\left| \frac{1}{m} \sum_{i=1}^m X_i - \mathbb{E}[X_1] \right| \leq \sqrt{\frac{\log \frac{2}{\zeta}}{2m}}.$$

Handwritten notes: The interval $[0, 1]$ is underlined. The term $\frac{1}{m} \sum_{i=1}^m X_i$ is circled. The term $\mathbb{E}[X_1]$ has a wavy line underneath. The term $\sqrt{\frac{\log \frac{2}{\zeta}}{2m}}$ is circled, and the expression $\log \frac{2}{\zeta}$ is also circled. A red arrow points from the text "subgaussian" to the circled term.



$\|T - \hat{T}\| \approx 0$

$$(\hat{T}q)(s, a) = r_a(s) + \frac{\gamma}{m} \sum_{s' \in C(s, a)} \max_{a' \in A} q(s', a')$$

sampled version of T

$$A = \arg \max_{a \in A} (\hat{T}^H \mathbf{0})(s_0, a)$$

$Q_H(s_0, a)$



Model-Predictive Control

$$\delta_H = \|Q_H(s_0, \cdot) - q^*(s_0, \cdot)\|_\infty \leq \epsilon$$

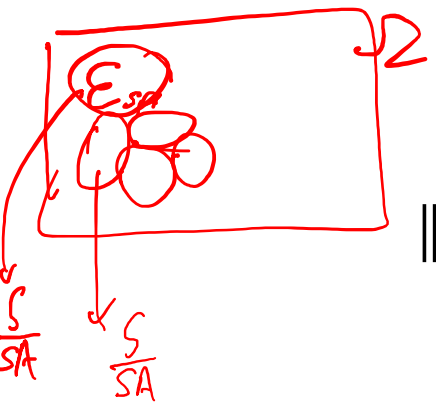
← we want this..

$$\delta_h \leq \gamma \delta_{h-1} + \|\hat{T}q^* - q^*\|_{S_{H-h}} \leq \gamma \delta_{h-1} + \underbrace{\|\hat{T}q^* - q^*\|_{S_{H-1}}}_{=:\epsilon'/(1-\gamma)}$$

$$S_h = \{s \in S \mid \text{dist}(s_0, s) \leq h\}$$

$$\delta_H \leq \frac{\gamma^H + \epsilon'(1 + \gamma + \dots + \gamma^{H-1})}{1 - \gamma} \leq \left(\gamma^H + \frac{\epsilon'}{1 - \gamma} \right) \frac{1}{1 - \gamma}$$

$$\delta_0 = \|q^*\|_{S_H} \leq \|q^*\|_\infty \leq \frac{1}{1 - \gamma}$$



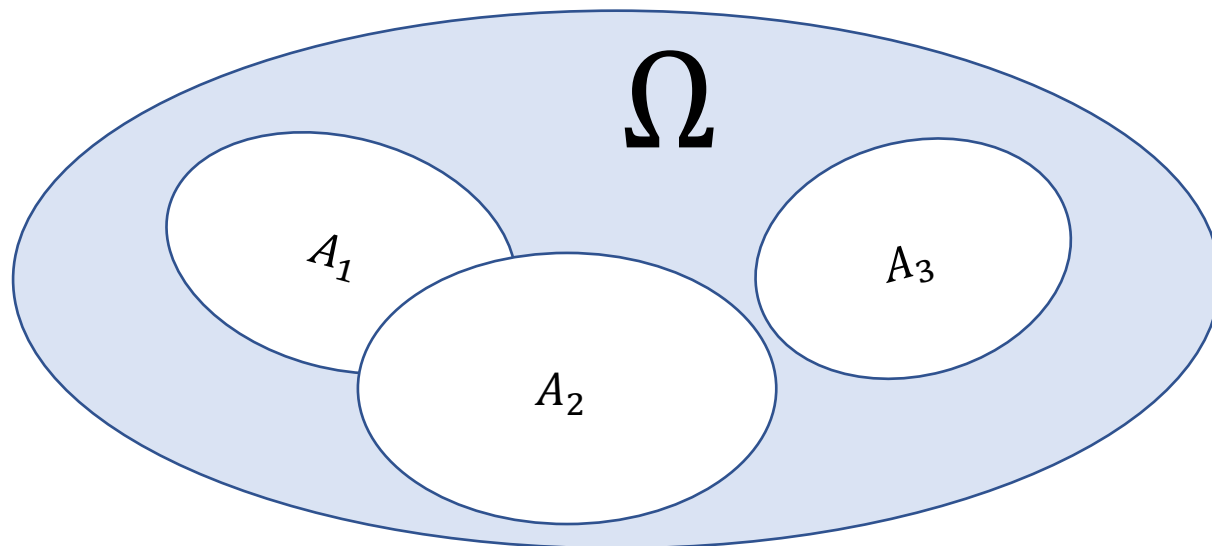
$$\|\hat{T}q^* - q^*\|_{S_{H-1}} \leq \|\hat{T}q^* - q^*\|_\infty = \max_{s, a} |\hat{T}q^*(s, a) - Tq^*(s, a)| \leq \frac{\gamma}{1 - \gamma} \sqrt{\frac{\log(2SA)}{2m}}$$

$r_a(s) \in [0, 1]$

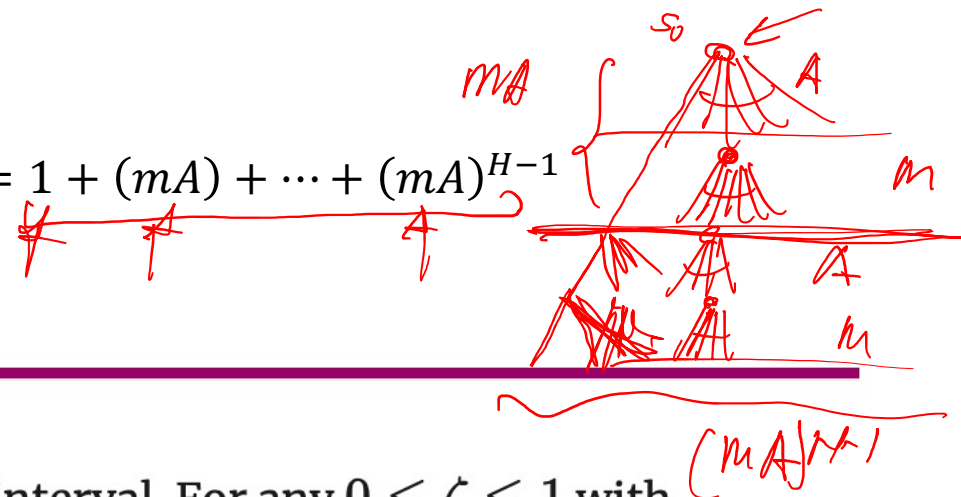
wp. 1-5

Lemma (Union Bound): For any probability measure \mathbb{P} and any countable sequence of events A_1, A_2, \dots of the underlying measurable space,

$$\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i).$$



S_i : the state that q is called for the i^{th} time, $i = 0, 1, \dots, n := 1 + (mA) + \dots + (mA)^{H-1}$



Lemma: Assume that the immediate rewards belong to the $[0, 1]$ interval. For any $0 \leq \zeta \leq 1$ with probability $1 - An\zeta$, for any $1 \leq i \leq n$,

$$\|\hat{T}q^*(S_i, \cdot) - q^*(S_i, \cdot)\|_{\infty} \leq \Delta(\zeta, m),$$

where Δ is given by (3).

$$\frac{\gamma}{1 - \gamma} \sqrt{\frac{\log \frac{2}{\zeta}}{2m}} =: \Delta(\zeta, m),$$

Theorem: Assume that the immediate rewards belong to the $[0, 1]$ interval. There is a local planner such that for any $\delta \geq 0$, in any discounted MDP with discount factor γ , the planner induces a δ -optimal policy and uses at most $O((m^* A)^H)$ elementary arithmetic and logic operations per its calls, where $m^*(\delta, A)$ is given by (8) and $H = \lceil H_{\gamma, (1-\gamma)\delta/3} \rceil$.

$$m^*(\delta, A) = 2c_\delta \left[H \log(c_\delta H) + \log \left(\frac{12}{(1-\gamma)^2 \delta} \right) + (H+1) \log(A) \right]$$

$$c_\delta = \frac{18}{\delta^2 (1-\gamma)^6} = \frac{18}{\delta^2} H^6$$

Bellman's “curse of dimensionality”

Using Randomization to Break the Curse of Dimensionality

John Rust, *Econometrica* Vol. 65, No. 3 (May, 1997), pp. 487-516

There is an important practical limitation to one's ability to solve continuous MDPs arbitrarily accurately, Bellman's curse of dimensionality. This is the well-known exponential rise in the time and space required to compute an approximate solution to an MDP problem as the dimension (i.e. the number of state and control variables) increases.

<https://www.jstor.org/stable/2171751>

$$s \in [0, 1]^d$$

$$H = 3$$

$$(m(A+H)A)^{H^3}$$

Csaba's addendum. This should be: “..exponential rise in the time required to compute an approximate solution to an MDP problem as **both** the planning horizon and the dimension increases.”

Extra reading:

O' Curse of Dimensionality, Where is Thy Sting?, K. L. Judd, 2008

https://kenjudd.org/wp-content/uploads/2017/02/Curse_in_Dallas.pdf

“Monte-Carlo propaganda” or truth??

Questions from slack

Ehsan Imani 3 days ago

We removed dependence on the size of the state space by sampling.
Can we do something similar for actions if the action space is too large?
When it comes to actions we're dealing with a max operator instead of expected value. So is there any hope in getting a (provably) good estimate without trying all actions, at least in some problems?

+9

w/o structure: No.

$$\mathbb{E}(A^H)$$

$$= \mathbb{E}(2^{H \log A})$$

w structure: concavity.



$$q^*(s, \cdot) \rightarrow \text{concave} \uparrow$$
$$q^{\pi}(s, \cdot)$$

[Yilin Wang](#) 11 hours ago

A question about the statement on the runtime in lecture 6. In the below proof of "runtime independence on size of the state space", It seems to be a default that the sampling size m is set independently from the state space size S . But how can we formally prove that " m can be set independently of S while meeting our target for the suboptimal other of the induced policy"? Am I missing some preconditions or preliminaries here?

The total runtime of this function is now $O((mA)^{k+1})$. What is important is that this will give us a compute time independent of the size of the state space as long as we can show that m can be set independently of S while meeting our target for the suboptimality of the induced policy.

+3

Homayoon Farrahi 7 hours ago

What prevents us from taking the sampling technique of local planning for stochastic MDPs and applying it to value iteration for tabular MDPs? Perform lookups for and average over some limited number of entries in the table instead of all entries. Could that allow us to drop a factor of S from its computational complexity?

+3

Homayoon Farrahi 7 hours ago

What prevents us from taking the sampling technique of local planning for stochastic MDPs and applying it to value iteration for tabular MDPs? Perform lookups for and average over some limited number of entries in the table instead of all entries. Could that allow us to drop a factor of S from its computational complexity?

+3

[Alireza Bakhtiari](#) [3 hours ago](#)

In this lemma we proved a lower bound for a planner which samples the next state from its simulator, and we also defined failure events which are the time steps that the planner's action value is noticeably different from the correct action value (maybe due to the bad luck in sampling from simulator?). What I expected was finding a lower bound that holds with some probability, since a planner could be unlucky and get bad samples from its simulator which results in failure events. So isn't there any probability that a planner doesn't achieve a good policy in these cases?

Lemma (Policy error bound II): Let $\zeta \in [0, 1]$, π be a memoryless policy that selects ϵ -optimizing actions with probability at least $1 - \zeta$ in each state. Then,

$$v^\pi \geq v^* - \frac{\epsilon + 2\zeta \|q^*\|_\infty}{1 - \gamma} \mathbf{1}.$$

Discussion

Computational complexity

- How do we account for compute cost?
- What is computation?
 - Turing model/bit model
 - RAM model/computation over the reals
 - Random bits?
 - Biological computation? Liquid computers? ??
 - Other models? What do we expect of a model of computation?
 - Implications of choices
 - Input size depends on model
 - Cost depends on model
 - Which model is a better fit to “reality”?