Lecture 4
Policy iteration & local planning

$$k^* := \lceil H_{\gamma,1} \rceil + 1$$

Theorem (Runtime Bound for Policy Iteration): Consider a finite, discounted MDP with rewards in $[0, 1]$. Let k^* be as in the progress lemma, $\{\pi_k\}_{k \geq 0}$ the sequence of policies obtained by policy iteration starting from an arbitrary initial policy π_0. Then, after at most $k = k^*(\mathrm{SA} - \mathrm{S}) = \tilde{O}\left(\frac{\mathrm{SA}-\mathrm{S}}{1-\gamma}\right)$ iterations, the policy π_k produced by policy iteration is optimal: $v^{\pi_k} = v^*$. In particular, policy iteration computes an optimal policy with at most $\tilde{O}\left(\frac{\mathrm{S}^4\mathrm{A}+\mathrm{S}^3\mathrm{A}^2}{1-\gamma}\right)$ arithmetic and logic operations.

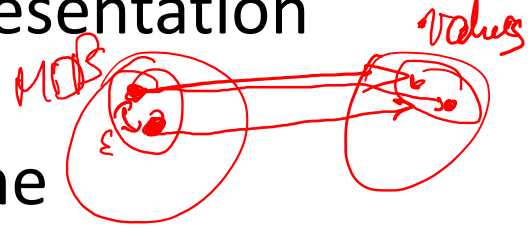
Yinyu Ye (2011). Bruno Scherrer (2016)

Questions from slack

$p_1 + p_2 + \dots + p_k = 1, \; p_i \geq 0 \qquad 1 \leq I \leq k : \mathbb{P}(I = i) = p_i$

- **Shuai Liu** I'm a bit confused about the "simulation model" and "table representation of MDPs".

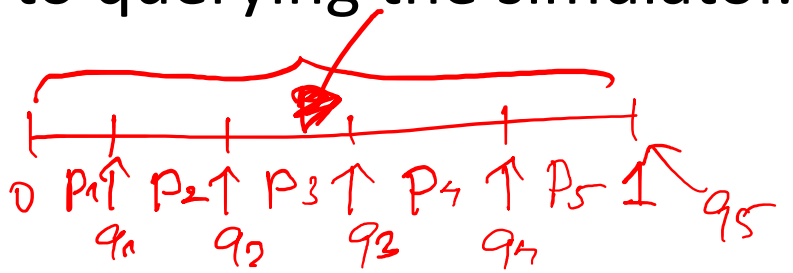
From the endnote of lecture 3, I infer the difference lies in we cannot perform vector calculations directly in simulation models as we did in table representation of MDPs. But can we still query the simulator multiple times and construct a table on which we can perform VI/PI? Or is there any other way to apply the algorithms we developed in table representation to simulation model? If we do so, is there any price we need to pay for transferring methods on table representation to simulation model?



Following the above question, I'm further confused about the difference between queries in table and simulation model as queries in table is just looking up the table (I guess so), which in some sense is similar to querying the simulator.

+9

$q_i = p_1 + \dots + p_i$



- **Jiamin He** 13 hours ago
  In lecture note 5, we define local planning with an oracle simulator. However, what if the simulator is not accurate, can we still get meaningful results on these local planning algorithms? If yes, how are we going to formulate the inaccuracy of the simulator?
  (I am asking this question because I am thinking of model-based reinforcement learning in which the model is learned and inaccurate. But let's put aside the learning of the model and discuss a simulator with a static error.)
- **+2**

- **Ehsan Imani** 4 hours ago
  In the policy iteration algorithm in the lecture the value of each policy is computed from scratch. Is there a kind of structure among the sequence of policies in any class of problems so that we can reuse the value of the previous policies $\pi_1, \dots, \pi_{k-1}$ to learn the value of the new policy π_k with less computation?
- +1

Discussion

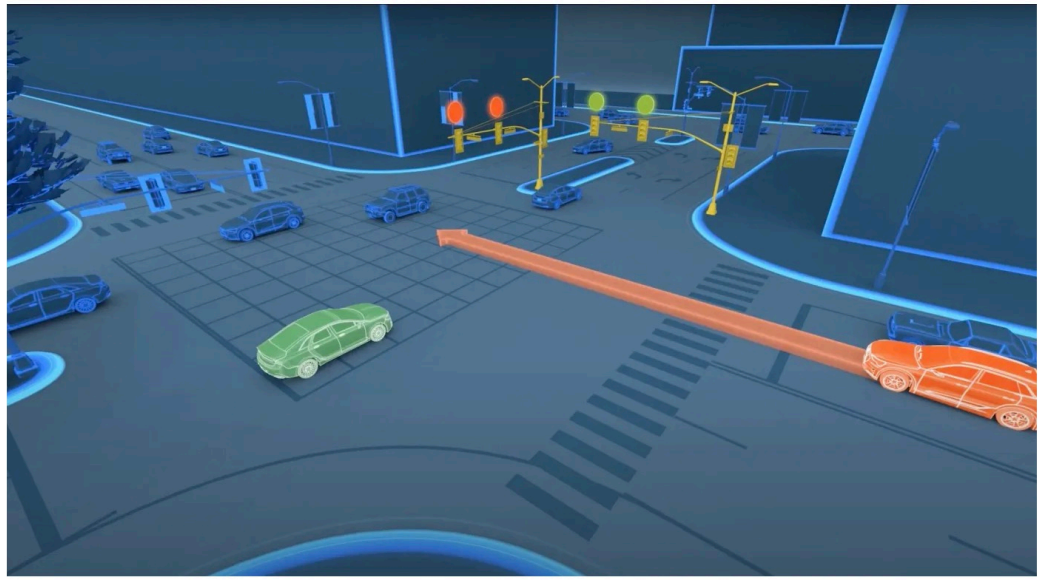
Simulate or not?

https://bdtechtalks.com/2022/01/06/real-world-reinforcement-learning/

“Basically, it comes down to this question: is it easier to create a brain, or is it easier to create the universe? I think it’s easier to create a brain, because it is part of the universe,”

- Sergey Levine

No more simulations



One of the great benefits of offline and self-supervised RL is learning from real-world data instead of simulated environments.

Simulate or not? Breakout rooms!

- Planning folks

Argue for the advantages of simulation

Argue against learning from batch of data/interaction

- Non-planning folks

Argue ~~for~~ against the advantages of simulation

Argue ~~against~~ for learning from batch of data/interaction

- Discussion for 5 mins in the rooms
- After rejoining the main room, one person from each group should summarize the arguments of the group
- It does not have to be perfect (time is short)

Computational complexity

- What is computation?
- How do we account for compute cost?