Lecture 3

Theorem (Policy Error Bound): Let $v : \mathcal{S} \to \mathbb{R}$ be arbitrary and π be the greedy policy w.r.t. v: $T_\pi v = Tv$. Then,

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

Singh & Yee, 1994

Theorem (Value Iteration): Consider an MDP with immediate rewards in the $[0, 1]$ interval. Pick an arbitrary positive number $\varepsilon > 0$. Let $v_0 = \mathbf{0}$ and set

$$v_{k+1} = Tv_k \quad \text{for } k = 0, 1, 2, \ldots$$

Then, for $k \geq \ln(1/(\varepsilon(1-\gamma))/\ln(1/\gamma)$, $\|v_k - v^*\|_\infty \leq \varepsilon$.

$$H_{\gamma,\varepsilon} := \frac{\ln(1/(\varepsilon(1-\gamma)))}{1-\gamma} \geq \frac{\ln(1/(\varepsilon(1-\gamma)))}{\ln(1/\gamma)}$$

Theorem (Runtime of Approximate Planning with Value Iteration): Fix a finite discounted MDP and a target accuracy $\delta > 0$. Then, after

$$O\left(\mathrm{S}^2\mathrm{A}H_{\gamma,\frac{\delta(1-\gamma)}{2\gamma}}\right) = \tilde{O}\left(\frac{\mathrm{S}^2\mathrm{A}}{1-\gamma}\ln\left(\frac{1}{\delta}\right)\right)$$

elementary arithmetic operations, value iteration produces a policy π that is δ-optimal: $v^\pi \geq v^* - \delta\mathbf{1}$, where the $\tilde{O}(\cdot)$ result holds when $\delta \leq 1/e$ is fixed and $\tilde{O}(\cdot)$ hides $\log(2/(1-\gamma))$.

$$H_{\gamma,\varepsilon} := \frac{\ln(1/(\varepsilon(1-\gamma)))}{1-\gamma}$$

Theorem (Computation Complexity of Planning in MDPs):

Let $0 \leq \delta < \gamma/(1-\gamma)$. Any algorithm that is guaranteed to produce δ-optimal policies in any finite MDP described with tables, with a fixed discount factor $0 \leq \gamma < 1$ and rewards in the $[0, 1]$ interval needs at least $\Omega(\mathrm{S}^2\mathrm{A})$ elementary arithmetic operations on some MDP with the above properties and whose state space is of size S and action space is of size A.

Policy Iteration

Policy iteration starts with an arbitrary deterministic (memoryless) policy π_0. Then, in step $k = 0, 1, 2, \ldots$, the following computations are done:

1. calculate v^{π_k}, and
2. obtain π_{k+1}, another deterministic memoryless policy, by "greedifying" w.r.t. v^{π_k}.

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

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)

$$
\begin{aligned}
v^{\pi'} - v^{\pi} &= (I - \gamma P_{\pi'})^{-1}[r_{\pi'} - (I - \gamma P_{\pi'})v^{\pi}] \\
&= (I - \gamma P_{\pi'})^{-1}[T_{\pi'}v^{\pi} - v^{\pi}] \,.
\end{aligned}
$$

Discussion

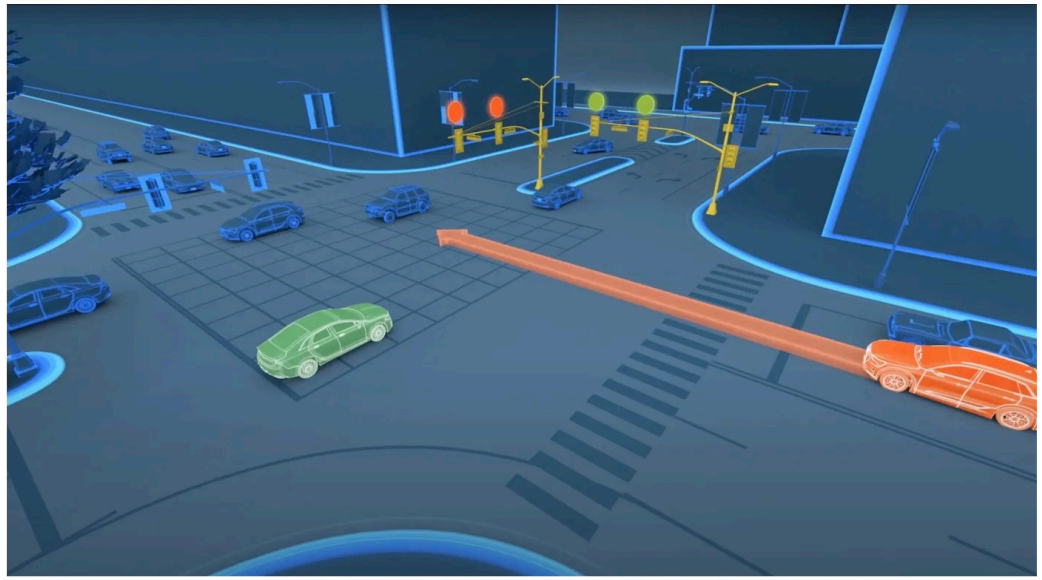
Simulate or not?

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?

Questions from slack

Kushagra Chandak

+5

I did not quite understand the argument for proving the lower bound of policy iteration. Could you maybe go over that?

Prabhat Nagarajan

+1

I’m a little confused about the bounds in the discounting section. We show that

H >= H*{gamma, eps}.

Then we say that H_{gamma, eps} is an upper bound on H*{gamma, eps}. This means

H_{gamma, eps} >= H*{gamma, eps}

From this how can we say that H >= H_{gamma, eps} ?

Here is a screenshot of the relevant part of the notes:

Prabhat Nagarajan

bound on the difference is below ε, we get that this bound on the difference holds as long as H satisfies

$$H \geq \underbrace{\frac{\ln\left(\frac{1}{\varepsilon(1-\gamma)}\right)}{\ln(1/\gamma)}}_{H^*_{\gamma,\varepsilon}}.$$

For the sake of simplicity, oftentimes this requirement is strengthened to $H \geq H_{\gamma,\varepsilon}$, where the latter quantity is defined as

$$H_{\gamma,\varepsilon} := \frac{\ln\left(\frac{1}{\varepsilon(1-\gamma)}\right)}{1-\gamma}.$$

Kushagra Chandak

+5

I did not quite understand the argument for proving the lower bound of policy iteration. Could you maybe go over that?

Kushagra Chandak

Also, while talking about the iteration complexity of policy iteration, you said there are SA-A free spaces in the table. Shouldn't it be SA-S, since there's one optimal action for *each* state?

True