CS 4/5789 2 Feb 2022 Prof Sarah Dean

Lecture 4: Value & Policy Iteration, Dynamic Programming

1) Value Iteration

Recall algorithm, contraction proof from last lecture. Setting $Q^{t+1} \leftarrow \mathcal{J} Q^t$ results in approximately optimal Q function:

$$\|Q^t - Q^*\|_\infty \leq \gamma^t \|Q^0 - Q^*\|_\infty$$

From Q functions to policies

We know $\pi^*(s) = \arg\max_a Q^*(s,a)$

since $Q^t(s,a) \approx Q^*(s,a)$ during value iteration,

$\pi^t(s) = \arg\max_a Q^t(s,a)$

a good choice?

Theorem: The quality of π^t is bounded below:

$$V^{\pi^t}(s) \geq V^*(s) - \frac{2\gamma^t}{1-\gamma}\|Q^0 - Q^*\|_\infty \quad \forall s \in \mathcal{S}$$

Proof:

Assume the following claim is true:

$$V^{\pi^t}(s) - V^*(s) \geq \gamma \mathbb{E}_{s' \sim P(s,\pi^t(s))}\left[V^{\pi^t}(s') - V^*(s')\right] - 2\gamma^t \|Q^0 - Q^*\|_\infty$$

Then recursing k times,

$$V^{\pi^t}(s) - V^*(s) \geq \gamma^k \underset{s\sim P(s,\pi^t(s))}{\mathbb{E}}\left[V^{\pi^t}(s') - V^*(s)\right] - 2\sum_{\ell=0}^{k} \gamma^{\ell+t} \|Q^0 - Q^*\|_\infty$$

letting $k \to \infty$,

$$V^{\pi^t}(s) - V^*(s) \geq -2\gamma^t \sum_{l=0}^{\infty} \gamma^\ell \|Q^0 - Q^*\|_\infty$$

$$= \frac{-2\gamma^t}{1-\gamma} \|Q^0 - Q^*\|_\infty$$

Proof of claim:

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s)) \quad \text{(definition)}$$

$$\underbrace{- Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s))}_{} = 0$$

$$= \gamma \underset{s'\sim p(s,\pi^t(s))}{\mathbb{E}}\left[V^{\pi^t}(s') - V^*(s')\right] + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \underset{s'\sim p(s,\pi^t(s))}{\mathbb{E}}\left[V^{\pi^t}(s') - V^*(s')\right] + Q^*(s, \pi^t(s)) \overbrace{- Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s))}^{\geq 0 \text{ by } \pi^t \text{ optimality}} - Q^*(s, \pi^*(s))$$

$$\geq \gamma \underset{s'\sim p(s,\pi^t(s))}{\mathbb{E}}\left[V^{\pi^t}(s'), V^*(s')\right] - \|Q^t - Q^*\|_\infty - \|Q^t - Q^*\|_\infty$$

by definition of $\|\cdot\|_\infty$

$$\geq \gamma \underset{s'\sim p(s,\pi^t(s))}{\mathbb{E}}\left[V^{\pi^t}(s'), V^*(s')\right] - 2\gamma^t \|Q^0 - Q^*\|_\infty \quad \text{(convergence Lemma)}$$

□

Summary of Value Iteration (VI)

1) VI (fixed point)

$$Q^{t+1} \leftarrow \mathcal{T} Q^t$$

contraction

2) VI convergence

$$\|Q^t - Q^*\|_\infty \le \gamma^t \|Q^0 - Q^*\|_\infty$$

exponentially fast

"geometric rate"

$$\pi^t(s) = \arg\max_a Q^t(s,a)$$

3) policy performance

$$V^{\pi^t}(s) \ge V^*(s) - \frac{2\gamma^t}{1-\gamma}\|Q^0 - Q^*\|_\infty$$

Convergence argument is similar to Iterative Policy Eval (PE)

Bellman Eq: $V^\pi = R + \gamma P V^\pi$	fixed point	Bellman Optimality $Q^* = \mathcal{T} Q^*$
Iterative PE $V^{t+1} \leftarrow R + P V^t$	iteration	VI $Q^{t+1} \leftarrow \mathcal{T} Q^t$
by contraction, $\|V^t - V^\pi\|_\infty \le \gamma^t \|V^0 - V^\pi\|_\infty$	converges	$\|Q^t - Q^*\|_\infty \le \gamma^t \|Q^0 - Q^*\|_\infty$

2) Policy Iteration

Another iterative Algorithm for approximating the optimal policy π^*. While value iteration updates Q-function at each timestep (and then at the very end we transform Q^t into π^t), policy iteration updates both a policy and a Q function at each timestep.

```
Algorithm: Policy Iteration
Initialize $\pi^0: \mathcal{S} \to \Delta(\mathcal{A})$
for $t = 0, 1, \ldots$
        Policy Evaluation: $Q^{\pi^t}(s, a)\ \forall s, a)$
        Policy Improvement: $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a)\ \forall s$
```

In each iteration, we first use policy evaluation to compute the Q function associated with the current policy. Then, we "argmax" that Q function to generate a new policy, aka, policy improvement.

Aside: How do we get Q^{π^t} from policy evaluation?

$$V^{\pi^t} = (I - \gamma \underset{\substack{\text{entries}\\ P(s'|s,\pi^t(s))}}{P})^{-1} \underset{\substack{\text{entries}\\ r(s,\pi^t(s))}}{R}$$

Then $$Q^{\pi^t}(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)}\left[V^{\pi^t}(s')\right] \quad \forall s, a$$

We will prove two key properties of Policy iteration.

1) Monotonic Improvement
$$Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a) \quad \forall s, a$$

2) convergence
$$\|V^* - V^{\pi^t}\|_\infty \leq \gamma^t \|V^* - V^{\pi^0}\|_\infty$$

Lemma (Monotonic Improvement):

For policy iteration, $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a) \;\forall s,a.$

Proof:

$$Q^{\pi^{t+1}}(s,a) - Q^{\pi^t}(s,a) = \cancel{r(s,a)} + \gamma \mathop{\mathbb{E}}_{s'\sim P(s,a)}\left[V^{\pi^{t+1}}(s')\right] - \cancel{r(s,a)} - \gamma \mathop{\mathbb{E}}_{s'\sim P(s,a)}\left[V^{\pi^t}(s')\right]$$

(definition of Q fn)

$$= \gamma \mathop{\mathbb{E}}_{s'\sim P(s,a)}\left[Q^{\pi^{t+1}}(s',\pi^{t+1}(s')) - Q^{\pi^t}(s',\pi^t(s'))\right]$$

(linearity of expectation and definition of value fn.)

$$= \gamma \mathop{\mathbb{E}}_{s'\sim P(s,a)}\Big[Q^{\pi^{t+1}}(s',\pi^{t+1}(s')) - Q^{\pi^t}(s',\pi^{t+1}(s')) + Q^{\pi^t}(s',\pi^{t+1}(s')) - Q^{\pi^t}(s',\pi^t(s'))\Big] \geq 0$$

$\pi^{t+1}(s')$ is defined as argmax for Q^{π^t}

$$\geq \gamma \mathop{\mathbb{E}}_{s'\sim P(s,a)}\left[Q^{\pi^{t+1}}(s',\pi^{t+1}(s')) - Q^{\pi^t}(s',\pi^{t+1}(s'))\right]$$

(iterate)

$$\geq \gamma^2 \mathop{\mathbb{E}}_{\substack{s'\sim P(s,a)\\ s''\sim P(s',\pi^{t+1}(s'))}}\left[Q^{\pi^{t+1}}(s'',\pi^{t+1}(s'')) - Q^{\pi^t}(s'',\pi^{t+1}(s''))\right]$$

(iterate k times)

$$\ldots \geq \gamma^k \mathop{\mathbb{E}}_{s_1,s_2,\ldots s_k}\left[Q^{\pi^{t+1}}(s_k,\pi^{t+1}(s_k)) - Q^{\pi^t}(s_k,\pi^{t+1}(s_k))\right]$$

$$\rightarrow 0 \text{ as } k \rightarrow \infty. \qquad \square$$

Does this imediately imply that $V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s)$?

(see proof below)

Theorem (convergence):

For policy iteration, $\|V^{\pi t} - V^*\|_\infty \le \gamma^t \|V^{\pi_0} - V^*\|_\infty$

Proof:

$$V^*(s) - V^{\pi^{t+1}}(s) = \max_a \Big[r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(s,a)} V^*(s')\Big] - \Big[r(s,\pi^{t+1}(s)) + \gamma \mathop{\mathbb{E}}_{s'\sim P(s,\pi^{t+1}(s))} V^{\pi^{t+1}}(s')\Big]$$

(Bellman optimality & definition)

Aside: by Lemma, $Q^{\pi^{t+1}}(s,a) \ge Q^{\pi^t}(s,a)\quad \forall s, a$

setting $a = \pi^{t+1}(s)$,

$$Q^{\pi^{t+1}}(s, \pi^{t+1}(s)) \ge Q^{\pi^t}(s, \pi^{t+1}(s)) \ \forall s$$

Recall that $\pi^{t+1}(s)$ is defined to maximize $Q^{\pi^t}(s, \cdot)$. Therefore

definition $\quad V^{\pi^{t+1}}(s) \ge Q^{\pi^t}(s, a) \quad \forall s, a$

choosing $a = \pi^t(s)$,

$$V^{\pi^{t+1}}(s) \ge V^{\pi^t}(s) \quad \forall s.$$

$$V^*(s) - V^{\pi^{t+1}}(s) \le \max_a \Big[r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(s,a)} V^*(s')\Big] - \Big[r(s,\pi^{t+1}(s)) + \gamma \mathop{\mathbb{E}}_{s'\sim P(s,\pi^{t+1}(s))} V^{\pi^t}(s')\Big]$$

(definition of π^{t+1})

$$= \max_a \Big[r(s,a) + \gamma \mathbb{E} V^*(s')\Big] - \max_a \Big[r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(s,a)} V^{\pi^t}(s')\Big]$$

$(\max_x f(x) - \max_x g(x) \le \max_x f(x) - g(x))$

$$\le \max_a \Big[\cancel{r(s,a)} + \gamma \mathop{\mathbb{E}}_{s'} V^*(s') - \big(\cancel{r(s,a)} + \gamma \mathop{\mathbb{E}}_{s'} V^{\pi^t}(s')\big)\Big]$$

$(\mathop{\mathbb{E}}_{x\sim D(x)} f(x) \le \max_x f(x))$

$$\le \max_{a, s'} \gamma \big(V^*(s') - V^{\pi^t}(s')\big) = \gamma \|V^* - V^{\pi^t}\|_\infty$$

Thus $\|V^{t+1} - V^*\|_\infty \le \gamma \|V^t - V^*\|_\infty$

$$\Rightarrow \|V^t - V^*\|_\infty \le \gamma^t \|V^0 - V^*\|_\infty$$

□

Both value iteration and policy iteration have geometric / exponential convergence

Value It.

$\|V^{\pi^t} - V^*\|_\infty \le \frac{2\gamma^t}{1-\gamma} \|Q^0 - Q^*\|_\infty$

Policy It.

$\|V^{\pi^t} - V^*\|_\infty \le \gamma^t \|V^0 - V^*\|_\infty$

while this is a very fast convergence rate, for any finite t, its not equal to 0, i.e. it is not exact.

In HW1, you will see that in fact Policy iteration is guaranteed to exactly converge to the optimal policy in a finite number of steps. (The same is not true for value Iteration)

3) Finite Horizon MDP

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, H, \mu_0\}$$

states $\mathcal{S}$, actions $\mathcal{A}$, transitions P, rewards r as before

Horizon $H \in \mathbb{N}^+$ (length of time)

Initial state distribution $\mu_0 \in \Delta(\mathcal{S})$

$s_0 \sim \mu_0$

The task starts from an initial distribution and lasts for H steps (common in robotics)

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H-1} r(a_t, s_t) \middle| s_{t+1} \sim P(s_t, a_t),\ s_0 \sim \mu_0,\ a_t = \pi_t(s_t) \right]$$

(deterministic reward & policy)

In general, we consider time-vary policies

$$\pi = (\pi_0, \pi_1, \dots, \pi_{H-1})$$

The value and Q function are

$$V_t^\pi(s) = \mathbb{E}\left[\sum_{k=t}^{H-1} r(s_k, a_k) \,\middle|\, \begin{array}{l} s_t = s, \; a_k = \pi_k(s_k), \\ s_{k+1} \sim P(s_k, a_k) \end{array}\right]$$

$$Q_t^\pi(s,a) = \mathbb{E}\left[\sum_{k=t}^{H-1} r(s_k, a_k) \,\middle|\, \begin{array}{l} (s_t, a_t) = (s, a), \\ a_k = \pi_k(s_k), \\ s_{k+1} \sim P(s_k, a_k) \end{array}\right]$$

time-varying!

Bellman Equation:

$$Q_t^\pi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)}\left[V_{t+1}^\pi(s')\right]$$

next state & next timestep

Because the horizon is finite, the recursion implied by the Bellman equation is finite and we can compute the optimal policy backwards through time.

4) Dynamic Programming

To find $\pi^* = (\pi_0^*, \dots, \pi_{H-1}^*)$

Start with H-1 (note $V_H(s) = 0$ since H is past horizon)

$$Q^*_{H-1}(s,a) = r(s,a) \qquad \pi^*_{H-1}(s) = \arg\max_a Q^*_{H-1}(s,a)$$

$$V^*_{H-1}(s) = \max_a Q^*_{H-1}(s,a) = Q^*_{H-1}(s, \pi^*_{H-1}(s))$$

Bellman optimality

Then if we have computed $V_{t+1}^*(s)$, $(t \leq H-2)$

$$\left.\begin{aligned} Q_t^*(s,a) &= r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V_{t+1}^*(s') \\ \pi_t^*(s) &= \underset{a}{\operatorname{argmax}}\, Q_t^*(s,a) \end{aligned}\right\} \text{finite time version of Value/Policy iteration.}$$

I.e., if we know how to act optimally at time t+1, we can figure out how to act optimally at time t.

Dynamic Programming will terminate in H steps (t from H-1 to 0)

Result in <u>exact</u> π^*, no discounting.