Planning in MDPs

Wen Sun

CS 6789: Foundations of Reinforcement Learning

Announcements

HW0 is due tmr

Recap: Value iteration

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

Recap: Value iteration

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

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

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

Recap: Value iteration

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

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

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

Q: when will π^t be the optimal policy?

Outline

1. Policy Iteration

2. Computation complexity of VI and PI

3. Linear Programming formulation

Policy Iteration Algorithm:

1. Initialization: $\pi^0 : S \mapsto A$

Policy Iteration Algorithm:

1. Initialization: $\pi^0 : S \mapsto A$

2. Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

Policy Iteration Algorithm:

1. Initialization: $\pi^0 : S \mapsto A$

2. Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

3. Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Policy Iteration Algorithm:

1. Initialization: $\pi^0 : S \mapsto A$

2. Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

3. Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Policy Iteration Algorithm:

Closed-form for PE
(see 1.1.3 in Monograph)

1. Initialization: $\pi^0 : S \mapsto A$

2. Policy Evaluation: $Q^{\pi^t}(s, a), \forall s, a$

3. Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s, a) \geq Q^{\pi^t}(s, a), \forall s, a$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s, a) \geq Q^{\pi^t}(s, a), \forall s, a$

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

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_{a} Q^{\pi^t}(s, a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s, a) \geq Q^{\pi^t}(s, a), \forall s, a$

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

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

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s, a$

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

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

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

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s, a) \geq Q^{\pi^t}(s, a), \forall s, a$

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

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

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

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Lemma: Monotonic improvement $Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \forall s, a$

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

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

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

$$V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s), \forall s$$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

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

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s, a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

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

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

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

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

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

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

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

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

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

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

$$\leq \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s') \right) \right)$$

Analysis of Policy Iteration

Recall: Policy Improvement $\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \forall s$

Theorem: Convergence $\|V^{\pi^{t+1}} - V^\star\|_\infty \le \gamma \|V^{\pi^t} - V^\star\|_\infty$

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

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

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

$$\le \max_a \left(r(s,a) + \gamma \mathbb{E}_{s'\sim P(s,a)} V^\star(s') - \left(r(s,a) + \gamma \mathbb{E}_{s'\sim P(s,a)} V^{\pi^t}(s') \right) \right)$$

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

Analysis of Policy Iteration

Q: what happens when π^{t+1} and π^t are exactly the same?

Show that π^t is an optimal policy $\pi^\star$

Q: does this imply that the algorithm will terminate?

Outline

1. Policy Iteration

2. Computation complexity of VI and PI

3. Linear Programming formulation

Computation complexity of VI and PI

Given an MDP $\mathscr{M} = (S, A, P, r, \gamma)$ can we exactly compute $Q^\star$ (or find $\pi^\star$) in time polynomial wrt $S, A, 1/(1-\gamma)$?

Computation complexity of VI and PI

Given an MDP $\mathscr{M} = (S, A, P, r, \gamma)$ can we exactly compute $Q^\star$ (or find $\pi^\star$) in time polynomial wrt $S, A, 1/(1-\gamma)$?

No for VI (i.e.,gap between second and best)

Computation complexity of VI and PI

Given an MDP $\mathscr{M} = (S, A, P, r, \gamma)$ can we exactly compute $Q^\star$ (or find $\pi^\star$) in time polynomial wrt $S, A, 1/(1-\gamma)$?

No for VI (i.e.,gap between second and best)

Yes for policy iteration:

$$(S^3 + S^2 A) \cdot \min \left\{ \frac{A^S}{S}, \frac{S^2 A \log \frac{S^2}{1-\gamma}}{1-\gamma} \right\}$$

Computation complexity of VI and PI

Given an MDP $\mathscr{M} = (S, A, P, r, \gamma)$ can we exactly compute $Q^\star$ (or find $\pi^\star$) in time polynomial wrt $S, A, 1/(1-\gamma)$?

No for VI (i.e.,gap between second and best)

Yes for policy iteration:

$$(S^3 + S^2A) \cdot \min\left\{ \frac{A^S}{S}, \frac{S^2 A \log \frac{S^2}{1-\gamma}}{1-\gamma} \right\}$$

What about poly(S, A) algs?

Outline

1. Policy Iteration

2. Computation complexity of VI and PI

3. Linear Programming formulation

The primal linear programming

Recall the Bellman consistency:

$$V(s) = \max_a \left\{ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V(s') \right] \right\}, \forall s$$

The primal linear programming

Recall the Bellman consistency:

$$V(s) = \max_a \left\{ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V(s') \right] \right\}, \forall s$$

We can re-write this as a linear program

The primal linear programming

Recall the Bellman consistency:

$$V(s) = \max_a \left\{ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V(s') \right] \right\}, \forall s$$

We can re-write this as a linear program

$$\min \sum_s \mu(s) V(s)$$

s.t. $V(s) \geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \quad \forall s, a \in S \times A$

The primal linear programming

Recall the Bellman consistency:

$$V(s) = \max_a \left\{ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V(s') \right] \right\}, \forall s$$

We can re-write this as a linear program

$$\min \sum_s \mu(s) V(s)$$

s.t. $V(s) \geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \quad \forall s, a \in S \times A$

(Proof in HW1)

LP Runtime

[Ye, '05]: there is an interior point algorithm (CIPA)
which is ("nearly") strongly polynomial, i.e., no poly dependence on $1/(1-\gamma)$

$$S^4 A^4 \ln\left(\frac{S}{1-\gamma}\right)$$

What about the Dual LP?

What about the Dual LP?

- Let us now consider the dual LP.
  - It is also very helpful conceptually.
  - In some cases, it also provides a reasonable algorithmic approach

What about the Dual LP?

- Let us now consider the dual LP.
  - It is also very helpful conceptually.
  - In some cases, it also provides a reasonable algorithmic approach
- Let us start by understanding the dual variables

State action occupancy measure

$\mathbb{P}_h(s, a; \mu, \pi)$: probability of π visiting (s, a) at time step $h \in \mathbb{N}$, starting at $s_0 \sim \mu$

State action occupancy measure

$\mathbb{P}_h(s, a; \mu, \pi)$: probability of π visiting (s, a) at time step $h \in \mathbb{N}$, starting at $s_0 \sim \mu$

$$d_\mu^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h(s, a; \mu, \pi)$$

State action occupancy measure

$\mathbb{P}_h(s, a; \mu, \pi)$: probability of π visiting (s, a) at time step $h \in \mathbb{N}$, starting at $s_0 \sim \mu$

$$d_\mu^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h(s, a; \mu, \pi)$$

$$\mathbb{E}_{s_0 \sim \mu} V^\pi(s_0) = \frac{1}{1 - \gamma} \sum_{s,a} d_\mu^\pi(s, a) r(s, a)$$

A Bellman equation like property for $d_\mu^\pi(s, a)$

$$\sum_a d_\mu^\pi(s, a) = (1 - \gamma)\mu(s) + \gamma \sum_{\bar{s},\bar{a}} P(s \mid \bar{s}, \bar{a}) d_\mu^\pi(\bar{s}, \bar{a})$$

Proof:

The "State-Action" Polytope

- Let us define the state-action polytope K as follows:

$$K_\mu := \left\{ d \,|\, d \geq 0 \quad \text{and} \quad \sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s \,|\, s', a') d(s', a') \right\}$$

The "State-Action" Polytope

- Let us define the state-action polytope K as follows:

$$K_\mu := \left\{ d \,|\, d \geq 0 \;\; \text{and} \;\; \sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s \,|\, s', a') d(s', a') \right\}$$

- This set precisely characterizes all state-action visitation distributions:

The "State-Action" Polytope

- Let us define the state-action polytope K as follows:

$$K_\mu := \left\{ d \,|\, d \geq 0 \;\; \text{and} \right.$$

$$\left. \sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s\,|\,s',a')d(s',a') \right\}$$

- This set precisely characterizes all state-action visitation distributions:

Lemma: $d \in K_\mu$ if and only if there exists a (possibly randomized) policy π s.t. $d_\mu^\pi = d$

The "State-Action" Polytope

- Let us define the state-action polytope K as follows:

$$K_\mu := \left\{ d \,|\, d \geq 0 \ \text{ and } \right.$$

$$\left. \sum_a d(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s\,|\,s',a')d(s',a') \right\}$$

- This set precisely characterizes all state-action visitation distributions:

Lemma: $d \in K_\mu$ if and only if there exists a (possibly randomized) policy π s.t. $d_\mu^\pi = d$

(Proof in HW1)

The Dual LP

$$\max \sum_{s,a} d(s,a) r(s,a)$$

$$\text{s.t.} \quad d \in K_\mu$$

- One can verify that this is the dual of the primal LP.

Summary

Notations: Value / Q functions, state-action occupant measures, Bellman equation / optimality

Planning algorithms: VI, PI, LP (primal and dual)