Module 6
Value Iteration

CS 886 Sequential Decision Making and Reinforcement Learning

University of Waterloo

Markov Decision Process

- Definition
  - Set of states: S
  - Set of actions (i.e., decisions): A
  - Transition model: $\Pr(s_t | s_{t-1}, a_{t-1})$
  - Reward model (i.e., utility): $R(s_t, a_t)$
  - Discount factor: $0 \leq \gamma \leq 1$
  - Horizon (i.e., # of time steps): h

- Goal: find optimal policy π

Finite Horizon

- Policy evaluation

$$V_h^\pi(s) = \sum_{t=0}^{h} \gamma^t \Pr(S_t = s' | S_0 = s, \pi) R(s', \pi_t(s'))$$

- Recursive form (dynamic programming)

$$V_0^\pi(s) = R(s, \pi_0(s))$$

$$V_t^\pi(s) = R\big(s, \pi_t(s)\big) + \gamma \sum_{s'} \Pr(s' | s, \pi_t(s)) \, V_{t-1}^\pi(s')$$

Finite Horizon

- Optimal Policy π^*

$$V_h^{\pi^*}(s) \geq V_h^{\pi}(s) \;\; \forall \pi, s$$

- Optimal value function V^* (shorthand for V^{π^*})

$$V_0^*(s) = \max_a R(s,a)$$

$$\underbrace{V_t^*(s) = \max_a R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a) \, V_{t-1}^*(s')}_{\text{Bellman's equation}}$$

Value Iteration Algorithm

valueIteration(MDP)

$V_0^*(s) \leftarrow \max_a R(s,a) \;\; \forall s$

For $t = 1$ to h do

$V_t^*(s) \leftarrow \max_a R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a) \, V_{t-1}^*(s') \;\; \forall s$

Return V^*

Optimal policy π^*

$t = 0$: $\pi_0^*(s) \leftarrow \operatorname{argmax}_a R(s,a) \; \forall s$

$t > 0$: $\pi_t^*(s) \leftarrow \operatorname{argmax}_a R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a) \, V_{t-1}^*(s') \;\; \forall s$

NB: π^* is non stationary (i.e., time dependent)

Value Iteration

- Matrix form:

  R^a: $|S| \times 1$ column vector of rewards for a

  V_t^*: $|S| \times 1$ column vector of state values

  T^a: $|S| \times |S|$ matrix of transition prob. for a

valueIteration(MDP)

$V_0^* \leftarrow \max_a R^a$

For $t = 1$ to h do

$V_t^* \leftarrow \max_a R^a + \gamma T^a V_{t-1}^*$

Return V^*

Infinite Horizon

- Let $h \to \infty$
- Then $V_h^\pi \to V_\infty^\pi$ and $V_{h-1}^\pi \to V_\infty^\pi$

- Policy evaluation:

$$V_\infty^\pi(s) = R\big(s, \pi_\infty(s)\big) + \gamma \sum_{s'} \Pr(s'|s, \pi_\infty(s))\, V_\infty^\pi(s') \;\; \forall s$$

- Bellman's equation:

$$V_\infty^*(s) = \max_a R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V_\infty^*(s')$$

Policy evaluation

- Linear system of equations

$$V_\infty^\pi(s) = R\big(s, \pi_\infty(s)\big) + \gamma \sum_{s'} \Pr(s'|s, \pi_\infty(s))\, V_\infty^\pi(s') \;\; \forall s$$

- Matrix form:

  R: $|S| \times 1$ column vector of sate rewards for π

  V: $|S| \times 1$ column vector of state values for π

  T: $|S| \times |S|$ matrix of transition prob for π

$$V = R + \gamma T V$$

Solving linear equations

- Linear system: $V = R + \gamma TV$

- Gaussian elimination: $(I - \gamma T)V = R$
- Compute inverse: $V = (I - \gamma T)^{-1}R$
- Iterative methods
    - Value iteration (a.k.a. Richardson iteration)
    - Repeat $V \leftarrow R + \gamma TV$

Contraction

- Let $H(V) \stackrel{\text{def}}{=} R + \gamma TV$ be the policy eval operator
- Lemma 1: H is a contraction mapping.

$$\left|\left|H(\tilde{V}) - H(V)\right|\right|_\infty \leq \gamma \left|\left|\tilde{V} - V\right|\right|_\infty$$

- Proof $\left|\left|H(\tilde{V}) - H(V)\right|\right|_\infty$

$= \left|\left|R + \gamma T\tilde{V} - R - \gamma TV\right|\right|_\infty$ (by definition)

$= \left|\left|\gamma T(\tilde{V} - V)\right|\right|_\infty$ (simplification)

$\leq \gamma \left|\left|T\right|\right|_\infty \left|\left|\tilde{V} - V\right|\right|_\infty$ (since $\left|\left|AB\right|\right| \leq \left|\left|A\right|\right| \left|\left|B\right|\right|$)

$= \gamma \left|\left|\tilde{V} - V\right|\right|_\infty$ (since $\max_s \sum_{s'} T(s, s') = 1$)

Convergence

- Theorem 2: Policy evaluation converges to V^{π} for any initial estimate V

$$\lim_{n\to\infty} H^{(n)}(V) = V^{\pi} \ \ \forall V$$

- Proof
  - By definition $\mathrm{V}^{\pi} = H^{(\infty)}(0)$, but policy evaluation computes $H^{(\infty)}(V)$ for any initial V
  - By lemma 1, $\left|\left|H^{(n)}(V) - H^{(n)}(\tilde{V})\right|\right|_{\infty} \leq \gamma^{n} \left|\left|V - \tilde{V}\right|\right|_{\infty}$
  - Hence, when $n \to \infty$, then $\left|\left|H^{(n)}(V) - H^{(n)}(0)\right|\right|_{\infty} \to 0$ and $H^{(\infty)}(V) = V^{\pi} \ \ \forall V$

Approximate Policy Evaluation

- In practice, we can't perform an infinite number of iterations.
- Suppose that we perform value iteration for k steps and $\left\|H^{(k)}(V) - H^{(k-1)}(V)\right\|_\infty = \epsilon$, how far is $H^{(k)}(V)$ from V^π?

Approximate Policy Evaluation

- Theorem 3: If $\left|\left|H^{(k)}(V) - H^{(k-1)}(V)\right|\right|_\infty \leq \epsilon$ then

$$\left|\left|V^\pi - H^{(k)}(V)\right|\right|_\infty \leq \frac{\epsilon}{1-\gamma}$$

- Proof $\left|\left|V^\pi - H^{(k)}(V)\right|\right|_\infty$

$= \left|\left|H^{(\infty)}(V) - H^{(k)}(V)\right|\right|_\infty$ (by Theorem 2)

$= \left|\left|\sum_{t=1}^{\infty} H^{(t+k)}(V) - H^{(t+k-1)}(V)\right|\right|_\infty$

$\leq \sum_{t=1}^{\infty} \left|\left|H^{(t+k)}(V) - H^{(t+k-1)}(V)\right|\right|_\infty$ $(||A + B|| \leq ||A|| + ||B||)$

$= \sum_{t=1}^{\infty} \gamma^t \epsilon = \frac{\epsilon}{1-\gamma}$ (by Lemma 1)

Optimal Value Function

- Non-linear system of equations

$$V_\infty^*(s) = \max_a R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V_\infty^*(s') \;\; \forall s$$

- Matrix form:

  R^a: $|S| \times 1$ column vector of rewards for a

  V^*: $|S| \times 1$ column vector of optimal values

  T^a: $|S| \times |S|$ matrix of transition prob for a

$$V^* = \max_a R^a + \gamma T^a V^*$$

Contraction

- Let $H^*(V) \stackrel{\text{def}}{=} \max_a R^a + \gamma T^a V$ be the operator in value iteration

- Lemma 3: H^* is a contraction mapping.

$$\left\|H^*(\tilde{V}) - H^*(V)\right\|_\infty \leq \gamma \left\|\tilde{V} - V\right\|_\infty$$

- Proof: without loss of generality,

  let $H^*(\tilde{V})(s) \geq H^*(V)(s)$ and

  let $a_s^* = \operatorname*{argmax}_a R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V(s')$

Contraction

- Proof continued:

- Then $0 \leq H^*(\tilde{V})(s) - H^*(V)(s)$ (by assumption)

  $\leq R(s, a_s^*) + \gamma \sum_{s'} \Pr(s'|s, a_s^*)\, \tilde{V}(s')$ (by definition)

  $-R(s, a_s^*) - \gamma \sum_{s'} \Pr(s'|s, a_s^*)\, V(s')$

  $= \gamma \sum_{s'} \Pr(s'|s, a_s^*)\, [\tilde{V}(s') - V(s')]$

  $\leq \gamma \sum_{s'} \Pr(s'|s, a_s^*)\, \left|\left|\tilde{V} - V\right|\right|_\infty$ (maxnorm upper bound)

  $= \gamma \left|\left|\tilde{V} - V\right|\right|_\infty$ (since $\sum_{s'} \Pr(s'|s, a_s^*) = 1$)

- Repeat the same argument for $H^*(V)(s) \geq H^*(\tilde{V})(s)$ and for each s

Convergence

- Theorem 4: Value iteration converges to V^* for any initial estimate V

$$\lim_{n\to\infty} H^{*(n)}(V) = V^* \quad \forall V$$

- Proof
  - By definition $\mathrm{V}^* = H^{*(\infty)}(0)$, but value iteration computes $H^{*(\infty)}(V)$ for some initial V
  - By lemma 3, $\left\|H^{*(n)}(V) - H^{*(n)}(\tilde{V})\right\|_\infty \leq \gamma^n \left\|V - \tilde{V}\right\|_\infty$
  - Hence, when $n \to \infty$, then $\left\|H^{*(n)}(V) - H^{*(n)}(0)\right\|_\infty \to 0$ and $H^{*(\infty)}(V) = V^* \;\; \forall V$

Value Iteration

- Even when horizon is infinite, perform finitely many iterations
- Stop when $\left|\left|V_n - V_{n-1}\right|\right| \leq \epsilon$

valueIteration(MDP)

$V_0^* \leftarrow \max_a R^a$; $\quad n \leftarrow 0$

Repeat

$\quad n \leftarrow n + 1$

$\quad V_n \leftarrow \max_a R^a + \gamma T^a V_{n-1}$

Until $\left|\left|V_n - V_{n-1}\right|\right|_\infty \leq \epsilon$

Return V_n

Induced Policy

- Since $||V_n - V_{n-1}||_\infty \leq \epsilon$, by Theorem 4: we know that $||V_n - V^*||_\infty \leq \frac{\epsilon}{1-\gamma}$

- But, how good is the stationary policy $\pi_n(s)$ extracted based on V_n?

$$\pi_n(s) = \underset{a}{\operatorname{argmax}}\, R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V_n(s')$$

- How far is V^{π_n} from V^*?

Induced Policy

- Theorem 5: $\left|\left|V^{\pi_n} - V^*\right|\right|_\infty \leq \frac{2\epsilon}{1-\gamma}$
- Proof

$$\left|\left|V^{\pi_n} - V^*\right|\right|_\infty = \left|\left|V^{\pi_n} - V_n + V_n - V^*\right|\right|_\infty$$

$$\leq \left|\left|V^{\pi_n} - V_n\right|\right|_\infty + \left|\left|V_n - V^*\right|\right|_\infty \quad (\left|\left|A+B\right|\right| \leq \left|\left|A\right|\right| + \left|\left|B\right|\right|)$$

$$= \left|\left|H^{\pi_n(\infty)}(V_n) - V_n\right|\right|_\infty + \left|\left|V_n - H^{*(\infty)}(V_n)\right|\right|_\infty$$

$$\leq \frac{\epsilon}{1-\gamma} + \frac{\epsilon}{1-\gamma} \qquad \text{(by Theorems 2 and 4)}$$

$$= \frac{2\epsilon}{1-\gamma}$$

Summary

- Value iteration
  - Simple dynamic programming algorithm
  - Complexity: $O(n|A||S|^2)$
    - Here n is the number of iterations

- Can we optimize the policy directly instead of optimizing the value function and then inducing a policy?
  - Yes: by policy iteration