

Lecture 2a: Convergence Properties

CS885 Reinforcement Learning

2022-09-12

Complementary readings: [SutBar] Sec. 4.1, 4.4, [Sze] Sec. 2.2, 2.3, [Put] Sec. 6.1-6.3, [SigBuf] Chap. 1

Pascal Poupart
David R. Cheriton School of Computer Science



Outline

- Convergence properties of
 - Policy evaluation
 - Value iteration

Recap: Value Iteration

- Matrix form:

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

V_n^* : $|S| \times 1$ 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

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

Return V^*

Infinite Horizon

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

- **Policy evaluation:**

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

- **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

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

- Matrix form:

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

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

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

$$V = R + \gamma TV$$

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 evaluation operator
- **Lemma 1:** H is a **contraction mapping**.

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

- Proof $\|H(\tilde{V}) - H(V)\|_{\infty} = \|R + \gamma T\tilde{V} - R - \gamma TV\|_{\infty}$ (by definition)
 $= \|\gamma T(\tilde{V} - V)\|_{\infty}$ (simplification)
 $\leq \gamma \|T\|_{\infty} \|\tilde{V} - V\|_{\infty}$ (since $\|AB\| \leq \|A\| \|B\|$)
 $= \gamma \|\tilde{V} - V\|_{\infty}$ (since $\max_s \sum_{s'} T(s, s') = 1$)

Convergence

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

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

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

Approximate Policy Evaluation

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

Approximate Policy Evaluation

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

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

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

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

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

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

Optimal Value Function

- Non-linear system of equations

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

- 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 4:** H^* is a contraction mapping. Then $\|H^*(\tilde{V}) - H^*(V)\|_\infty \leq \gamma \|\tilde{V} - V\|_\infty$
- Proof: without loss of generality, let $H^*(\tilde{V})(s) \geq H^*(V)(s)$ and
let $a_s^* = \operatorname{argmax}_a R(s, a) + \gamma \sum_{s'} \Pr(s'|s, a) V(s')$
 $\tilde{a}_s^* = \operatorname{argmax}_a R(s, a) + \gamma \sum_{s'} \Pr(s'|s, a) \tilde{V}(s')$

Contraction (proof continued)

- Then $0 \leq H^*(\tilde{V})(s) - H^*(V)(s)$ (by assumption)
= $R(s, \tilde{a}_s^*) + \gamma \sum_{s'} \Pr(s'|s, \tilde{a}_s^*) \tilde{V}(s') - R(s, a_s^*) - \gamma \sum_{s'} \Pr(s'|s, a_s^*) V(s')$ (by def.)
 $\leq R(s, \tilde{a}_s^*) + \gamma \sum_{s'} \Pr(s'|s, \tilde{a}_s^*) \tilde{V}(s') - R(s, \tilde{a}_s^*) - \gamma \sum_{s'} \Pr(s'|s, \tilde{a}_s^*) V(s')$
(since \tilde{a}_s^* suboptimal for V)
= $\gamma \sum_{s'} \Pr(s'|s, \tilde{a}_s^*) [\tilde{V}(s') - V(s')]$
 $\leq \gamma \sum_{s'} \Pr(s'|s, \tilde{a}_s^*) \|\tilde{V} - V\|_\infty$ (maxnorm upper bound)
= $\gamma \|\tilde{V} - V\|_\infty$ (since $\sum_{s'} \Pr(s'|s, \tilde{a}_s^*) = 1$)
- Repeat the same argument for $H^*(V)(s) \geq H^*(\tilde{V})(s)$ and for each s

Convergence

- **Theorem 5:** Value iteration converges to V^* for any initial estimate V

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

- Proof

- By definition $V^* = H^{*(\infty)}(0)$, but value iteration computes $H^{*(\infty)}(V)$ for some initial V
- By Lemma 4, $\left\| H^{*(n)}(V) - H^{*(n)}(\tilde{V}) \right\|_{\infty} \leq \gamma^n \left\| V - \tilde{V} \right\|_{\infty}$
- Hence, when $n \rightarrow \infty$, then $\left\| H^{*(n)}(V) - H^{*(n)}(0) \right\|_{\infty} \rightarrow 0$ and $H^{*(\infty)}(V) = V^* \quad \forall V$

Value Iteration

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

valueIteration(MDP)

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

Repeat

$$n \leftarrow n + 1$$

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

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

Return V_n

Induced Policy

- Since $\|V_n - V_{n-1}\|_\infty \leq \epsilon$, by Theorem 5:
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 ?

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

- How far is V^{π_n} from V^* ?

Induced Policy

- **Theorem 6:** $\|V^{\pi_n} - V^*\|_{\infty} \leq \frac{2\epsilon}{1-\gamma}$
- Proof

$$\begin{aligned}\|V^{\pi_n} - V^*\|_{\infty} &= \|V^{\pi_n} - V_n + V_n - V^*\|_{\infty} \\ &\leq \|V^{\pi_n} - V_n\|_{\infty} + \|V_n - V^*\|_{\infty} \quad (\|A + B\| \leq \|A\| + \|B\|) \\ &= \left\| H^{\pi_n^{(\infty)}}(V_n) - V_n \right\|_{\infty} + \left\| V_n - H^{*(\infty)}(V_n) \right\|_{\infty} \\ &\leq \frac{\epsilon}{1-\gamma} + \frac{\epsilon}{1-\gamma} = \frac{2\epsilon}{1-\gamma} \quad (\text{by Theorems 2 and 5})\end{aligned}$$

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