Reinforcement Learning for Business, Economics, and Social Sciences

Unit 2-4: Value Iteration: Technicalities

Davud Rostam-Afschar (Uni Mannheim)

Solving for state-value functions in a system of linear equations

Value Iteration

- Idea: Optimize value function and then induce a policy
- Convergence properties of
  - Policy evaluation
  - Value iteration

Readings: Value Iteration
Sutton and Barto (2018, sections 4.1, 4.4)
Szepesvári (2022, sections 2.2, 2.3)
Puterman (2014, sections 6.1-6.3)
Sigaud and Buffet (2013, chapter 1)

Value Iteration Algorithm

valueIteration(MDP)

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

For $t = 1$ to h do

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

Return V^*

Value Iteration Algorithm

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

For $t = 1$ to h do
$V_t^*(s) \leftarrow \max_a R(s,a) + \gamma \sum_{S'} \Pr(s' \mid s,a)\, V_{t-1}^*(s')\ \forall s$

Return V^*

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

Value Iteration Algorithm

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

For $t = 1$ to h do
$\quad V_t^*(s) \leftarrow \max_a R(s, a) + \gamma \sum_{S'} \Pr(s' \mid s, a)\, V_{t-1}^*(s')\ \forall s$

Return V^*

Optimal policy π^*
$t = 0 : \pi_0^*(s) \leftarrow \underset{a}{\operatorname{argmax}}\ R(s, a)\ \forall s$
$t > 0 : \pi_t^*(s) \leftarrow \underset{a}{\operatorname{argmax}}\ R(s, a) + \gamma \sum_{s'} \Pr(s' \mid s, a)\, V_{t-1}^*(s')\ \forall s$

Value Iteration Algorithm

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

For $t = 1$ to h do
$$V_t^*(s) \leftarrow \max_a R(s,a) + \gamma \sum_{S'} \Pr\left(s' \mid s,a\right) V_{t-1}^*\left(s'\right)\ \forall s$$

Return V^*

Optimal policy π^*
$t = 0 : \pi_0^*(s) \leftarrow \underset{a}{\operatorname{argmax}}\ R(s,a)\ \forall s$
$t > 0 : \pi_t^*(s) \leftarrow \underset{a}{\operatorname{argmax}}\ R(s,a) + \gamma \sum_{s'} \Pr\left(s' \mid s,a\right) V_{t-1}^*\left(s'\right)\ \forall s$
NB: t indicates the # of time steps to go (till end of process)
π^* is non stationary (i.e., time dependent)

- 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

  Two-state, two-action Markov Decision Process

$$T^{a_1} = \begin{array}{ccc} & s_1' & s_2' \\ s_1 & 0.3 & 0.7 \\ s_2 & 0.8 & 0.2 \end{array} \qquad T^{a_2} = \begin{array}{ccc} & s_1' & s_2' \\ s_1 & 0.7 & 0.3 \\ s_2 & 0.2 & 0.8 \end{array}$$

$$R^{a_1} = \begin{array}{cc} s_1 & 0 \\ s_2 & 10 \end{array} \qquad R^{a_2} = \begin{array}{cc} s_1 & -5 \\ s_2 & 5 \end{array}$$

Value Iteration Example

- 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

$$\max R^a + \gamma T^a V_{t-1}^*$$

$$\max \left\{ \begin{pmatrix} 0 \\ 10 \end{pmatrix} + 0.9 \begin{pmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{pmatrix} \begin{pmatrix} V^*(s_1) \\ V^*(s_2) \end{pmatrix}, \right.$$

$$\left. \begin{pmatrix} -5 \\ 5 \end{pmatrix} + 0.9 \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix} \begin{pmatrix} V^*(s_1) \\ V^*(s_2) \end{pmatrix} \right\}$$

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:**

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

- **Bellman's equation:**

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

Policy Evaluation

- Linear system of equations

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

- 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 prob for π

(Non-optimal) policy $\pi(s_1) = a_1; \pi(s_2) = a_2$

$$T^\pi = \begin{array}{c|cc} & s'_1 & s'_2 \\ s_1 & 0.3 & 0.7 \\ s_2 & 0.2 & 0.8 \end{array} \qquad R^\pi = \begin{array}{c|c} s_1 & 0 \\ s_2 & 5 \end{array}$$

Policy Evaluation

- Linear system of equations

$$V^{\pi}_{\infty}(s) = R\left(s, \pi_{\infty}(s)\right) + \gamma \sum_{s\prime} \Pr\left(s' \mid s, \pi_{\infty}(s)\right) V^{\pi}_{\infty}\left(s'\right) \forall s$$

- 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 prob for π

(Non-optimal) policy $\pi\left(s_1\right) = a_1; \pi\left(s_2\right) = a_2$

$$V = R + \gamma TV$$

Solving Linear Equations

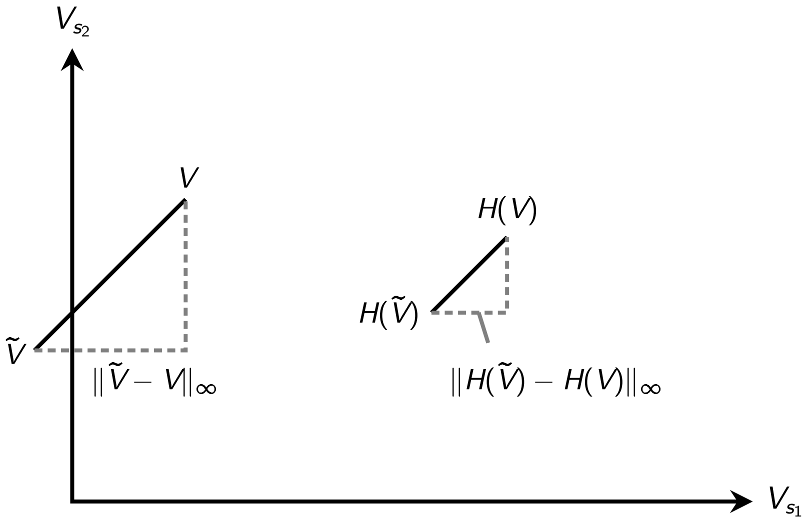
- Linear system: $V = R + \gamma TV$
- Gaussian elimination: $(I - \gamma T)V = R$
- Compute inverse: $V = (I - \gamma T)^{-1}R$
- Iterative methods
  - Value iteration (a.k.a. Richardson iteration)
  - Repeat $V \leftarrow R + \gamma TV$

With whatever estimate of the value function we start,

...

we shrink the distance with the discount factor

Contraction: Transform with H to Shrink the Maxnorm Distance



Contraction

- Let $H(V) \equiv R + \gamma TV$ be the policy evaluation operator
- **Lemma 1**: H is a contraction mapping.

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

Contraction

- Let $H(V) \equiv R + \gamma TV$ 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 $\|H(\tilde{V}) - H(V)\|_\infty$

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

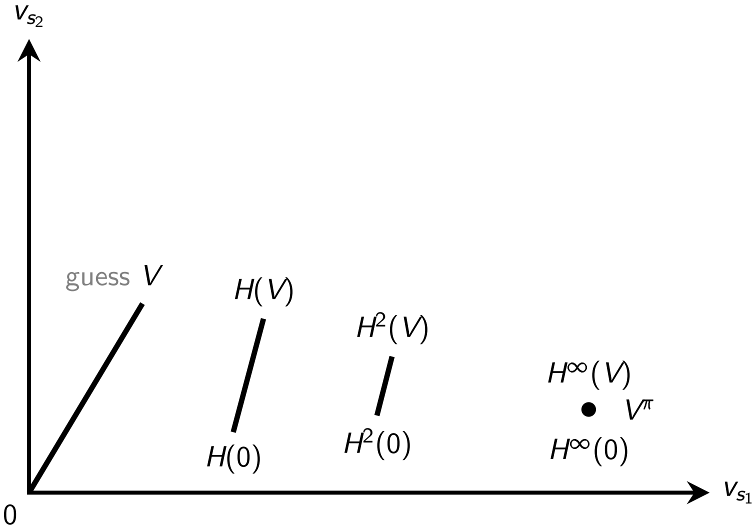
$$= \|\gamma T(\tilde{V} - V)\|_\infty \qquad \text{(simplification)}$$

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

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

Wherever we start, we contract to the optimal value

Contraction: Whatever Initial Guess Gets the True Point



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

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

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

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

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

When we stop early, how far are we from the optimal value?

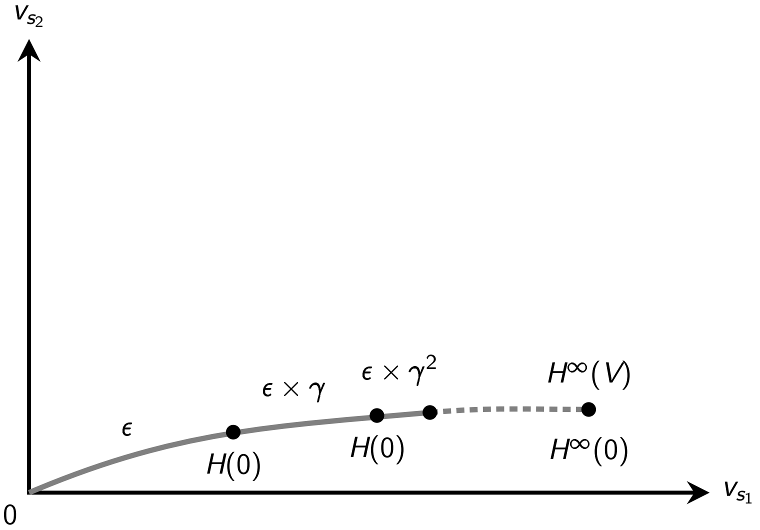
Approximate Policy Evaluation

- 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?

Contraction



Approximate Policy Evaluation

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

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

Approximate Policy Evaluation

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

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

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

$$
\begin{aligned}
&= \left\| H^{(\infty)}(V) - H^{(n)}(V) \right\|_\infty \quad \text{(by Theorem 2)} \\
&= \left\| \sum_{t=1}^{\infty} H^{(t+n)}(V) - H^{(t+n-1)}(V) \right\|_\infty \\
&\le \sum_{t=1}^{\infty} \left\| H^{(t+n)}(V) - H^{(t+n-1)}(V) \right\|_\infty (\|A+B\| \le \|A\| + \|B\|) \\
&= \sum_{t=1}^{\infty} \gamma^t \epsilon = \frac{\epsilon}{1-\gamma} \quad \text{(by Lemma 1)}
\end{aligned}
$$

How to find the best policy?

Optimal Value Function

- Non-linear system of equations

$$V^*_\infty(s) = \max_a R(s,a) + \gamma \sum_{s'} \Pr\left(s' \mid s, a\right) V^*_\infty\left(s'\right) \forall s$$

- 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 with max

- Even with $\max_a$ we get a contraction mapping
- Let $H^*(V) \equiv \max_a R^a + \gamma T^a V$ be the operator in value iteration
- **Lemma 4:** H^* is a contraction mapping.

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

Contraction with max

- Even with $\max_a$ we get a contraction mapping
- Let $H^*(V) \equiv \max_a R^a + \gamma T^a V$ be the operator in value iteration
- **Lemma 4:** H^* is a contraction mapping.

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

- Proof: without loss of generality,
  - let $H^*(\tilde{V})(s) \geq H^*(V)(s)$ and
  - let $a_s^* = \text{argmax } R(s,a) + \gamma \sum_{s'} \Pr(s' \mid s,a) V(s')$

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

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

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

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

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

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

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

Convergence with max

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

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

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

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

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

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

valueIteration(MDP)
$V_0^*(s) \leftarrow \max_a R^a; \quad n \leftarrow 0$
Repeat
  $n \leftarrow n + 1$
  $V_n \leftarrow \max_a R^a + \gamma T^a V_{n-1}$
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?
- $\pi_n(s) = \underset{a}{\operatorname{argmax}}\ R(s,a) + \gamma \sum_{s'} \Pr(s' \mid s, a)\, V_n(s')$
- How far is V^{π_n} from V^*?

Induced Policy

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

Induced Policy

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

- Proof

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

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

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

$$\leq \frac{\epsilon}{1-\gamma} + \frac{\epsilon}{1-\gamma} \quad \text{(by Theorems 2 and 5)}$$

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

Summary Value Iteration Algorithm

- Value iteration
  - Simple dynamic programming algorithm
  - Complexity: $\mathcal{O}\left(n|A||S|^2\right)$
    - Here n is the number of iterations,
    - A number of actions,
    - S number of states

PUTERMAN, M. L. (2014): *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

SIGAUD, O., AND O. BUFFET (2013): *Markov decision processes in artificial intelligence*. John Wiley & Sons, Available at https://zodml.org/sites/default/files/Markov_Decision_Processes_and_Artificial_Intelligence.pdf.

SUTTON, R. S., AND A. G. BARTO (2018): "Reinforcement learning: An introduction," *A Bradford Book*, Available at http://incompleteideas.net/book/the-book-2nd.html.

SZEPESVÁRI, C. (2022): *Algorithms for reinforcement learning*. Springer nature, Available at https://sites.ualberta.ca/~szepesva/RLBook.html.

Takeaways

How Does the Value Iteration Algorithm Work?

- Repeatedly applies the Bellman optimality update to converge to V^*
- Approximate solutions in infinite-horizon settings:
  Can stop early (threshold on update size)
- Policy error decreases each iteration