

Prediction and Control by Dynamic Programming

CS60077: Reinforcement Learning

Abir Das

IIT Kharagpur

Sep 09, 10, 11 and 16, 2021



Resources

- § Reinforcement Learning by David Silver [[Link](#)]
- § Reinforcement Learning by Balaraman Ravindran [[Link](#)]
- § SB: Chapter 4

Evaluating a Random Policy in the Small Gridworld

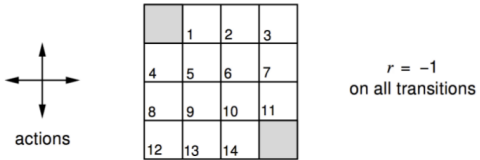


Figure credit: [SB] chapter 4

- § Undiscounted episodic MDP ($\lambda = 1$)
- § Non-terminal states are $\mathcal{S} = \{1, 2, \dots, 14\}$
- § Two terminal states (shown as shaded squares)
- § 4 possible actions in each state, $\mathcal{A} = \{up, down, right, left\}$
- § Deterministic state transitions
- § Actions leading out of the grid leave state unchanged
- § Reward is -1 until the terminal state is reached
- § Agent follows uniform random policy
 $\pi(n|..) = \pi(s|..) = \pi(e|..) = \pi(w|..)$

Improving a Policy: Policy Iteration

§ Given a policy π

- ▶ Evaluate the policy

$$v_\pi \doteq v^{(k+1)}(s) \leftarrow \sum_{a \in \mathcal{A}} \pi(a|s) \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) v^{(k)}(s') \right\}$$

- ▶ Improve the policy by acting greedily with respect to v_π

$$\pi' = \text{greedy}(v_\pi)$$

being greedy means choosing the action that will land the agent into best state *i.e.*, $\pi'(s) \doteq \arg \max_{a \in \mathcal{A}} q_\pi(s, a) =$

$$\arg \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) v_\pi(s') \right\}$$

§ In Small Gridworld improved policy was optimal $\pi' = \pi_*$

§ In general, need more iterations of improvement/evaluation

§ But this process of **policy iteration** always converges to π_*

Value Iteration

Algorithm 2: Value iteration

```

8 initialization:  $v \leftarrow v^0 \in \mathcal{V}$ , pick an  $\epsilon > 0$ ,  $n \leftarrow 0$ ;
9 while  $\|v^{n+1} - v^n\| > \epsilon \frac{1-\gamma}{2\gamma}$  do
10   foreach  $s \in \mathcal{S}$  do
11      $v^{n+1}(s) \leftarrow \max_a \left\{ r(s, a) + \gamma \sum_{s'} p(s'/s, a) v^n(s') \right\}$ 
12   end
13    $n \leftarrow n + 1$ ;
14 end
15 foreach  $s \in \mathcal{S}$  do
16   /* Note the use of  $\pi(s)$ . It means deterministic policy */
    $\pi(s) \leftarrow \arg \max_a \left\{ r(s, a) + \gamma \sum_{s'} p(s'/s, a) v^n(s') \right\}$ ; //  $n$  has
   already been incremented by 1
17 end

```

§ Take a note of the stopping criterion

Banach Fixed Point Theorem - Proof (2)

- § As m and $n \rightarrow \infty$ and as $\lambda < 1$, the norm of difference between v^{m+n} and v^n becomes less and less.
- § That means the sequence $\{v^n\}$ is Cauchy.
- § And since \mathcal{V} is a Banach space and since every Cauchy sequence converges to a point in that Banach space, therefore the Cauchy sequence $\{v^n\}$ also converges to a point in \mathcal{V} .
- § What we have proved till now is that the sequence $\{v^n\}$ will reach a converging point in the same space.
- § Lets say that the converging point is v^* .
- § What we will try to prove next is that v^* is a fixed point and then we will try to prove that v^* is an **unique** fixed point.

Banach Fixed Point Theorem - Proof (3)

§ Let us try to see what we get as the norm of the difference between v^* and Tv^* .

Banach Fixed Point Theorem - Proof (3)

- § Let us try to see what we get as the norm of the difference between v^* and Tv^* .
- § In the first line below we apply triangle inequality where v^n is the value of v at the n^{th} iteration.

$$\begin{aligned}
 \|Tv^* - v^*\| &\leq \|Tv^* - v^n\| + \|v^n - v^*\| \\
 &= \|Tv^* - Tv^{n-1}\| + \|v^n - v^*\| \\
 &\leq \lambda \|v^* - v^{n-1}\| + \|v^n - v^*\| \text{ [Contraction property]}
 \end{aligned}
 \tag{4}$$

Banach Fixed Point Theorem - Proof (3)

- § Let us try to see what we get as the norm of the difference between v^* and Tv^* .
- § In the first line below we apply triangle inequality where v^n is the value of v at the n^{th} iteration.

$$\begin{aligned}
 \|Tv^* - v^*\| &\leq \|Tv^* - v^n\| + \|v^n - v^*\| \\
 &= \|Tv^* - Tv^{n-1}\| + \|v^n - v^*\| \\
 &\leq \lambda \|v^* - v^{n-1}\| + \|v^n - v^*\| \quad [\text{Contraction property}]
 \end{aligned}
 \tag{4}$$

- § Since $\{v^n\}$ is Cauchy and v^* is the convergence point, both the terms in the above equation will tend to 0 as $n \rightarrow \infty$.
- § So, as $n \rightarrow \infty$, $\|Tv^* - v^*\| \rightarrow 0$. That means in limit $Tv^* = v^*$. So, it is proved that v^* is a fixed point.

Banach Fixed Point Theorem - Proof (4)

§ Now we will show the uniqueness, *i.e.*, v^* is unique.

Banach Fixed Point Theorem - Proof (4)

- § Now we will show the uniqueness, *i.e.*, v^* is unique.
- § Let u^* and v^* be two fixed points of the space. From the contraction property, we can write $\|Tu^* - Tv^*\| \leq \lambda \|u^* - v^*\|$.
- § But, since u^* and v^* are fixed points, $Tu^* = u^*$ and $Tv^* = v^*$.

Banach Fixed Point Theorem - Proof (4)

- § Now we will show the uniqueness, *i.e.*, v^* is unique.
- § Let u^* and v^* be two fixed points of the space. From the contraction property, we can write $\|Tu^* - Tv^*\| \leq \lambda \|u^* - v^*\|$.
- § But, since u^* and v^* are fixed points, $Tu^* = u^*$ and $Tv^* = v^*$.
- § That means $\|u^* - v^*\| \leq \lambda \|u^* - v^*\|$ which can not be true for $\lambda < 1$ unless $v^* = u^*$.
- § So, it is proved that v^* is an unique fixed point.

Existence and Uniqueness of Bellman Equations

- § Now, we will start talking about the existence and uniqueness of the solution to Bellman expectation equations and the Bellman optimality equations.
- § In case of a finite MDP the value function v can be thought of as a vector in a $|\mathcal{S}|$ dimensional vector space \mathcal{V} .
- § Whenever, we will use norm $||\cdot||$ in this space we will mean the max norm, unless otherwise specified.

Existence and Uniqueness of Bellman Equations

§ Previously, we have seen

- ▶ $r_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s)r(s, a)$, one step expected reward for following policy π at state s .
- ▶ $p_\pi(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s)p(s'|s, a)$, one step transition probability under policy π .

§ Using these notations, the Bellman expectation equation becomes,

$$\begin{aligned}
 v_\pi(s) &= \sum_{a \in \mathcal{A}} \pi(a|s) \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v_\pi(s') \right\} \\
 &= r_\pi(s) + \gamma \sum_{s' \in \mathcal{S}} p_\pi(s'|s)v_\pi(s')
 \end{aligned}$$

Existence and Uniqueness of Bellman Equations

§
$$v_{\pi}(s) = r_{\pi}(s) + \gamma \sum_{s' \in \mathcal{S}} p_{\pi}(s'|s)v_{\pi}(s')$$

Existence and Uniqueness of Bellman Equations

$$\S v_{\pi}(s) = r_{\pi}(s) + \gamma \sum_{s' \in \mathcal{S}} p_{\pi}(s'|s)v_{\pi}(s')$$

§ Refresher from earlier lectures

$$\begin{bmatrix} v(s_1) \\ v(s_2) \\ \vdots \\ v(s_n) \end{bmatrix} = \begin{bmatrix} r(s_1) \\ r(s_2) \\ \vdots \\ r(s_n) \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \mathcal{P}_{12} & \cdots & \mathcal{P}_{1n} \\ \mathcal{P}_{21} & \mathcal{P}_{22} & \cdots & \mathcal{P}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{P}_{n1} & \mathcal{P}_{n2} & \cdots & \mathcal{P}_{nn} \end{bmatrix} \begin{bmatrix} v(s_1) \\ v(s_2) \\ \vdots \\ v(s_n) \end{bmatrix}$$

$$\S v_{\pi} = r_{\pi} + \gamma P_{\pi} v_{\pi}$$

§ r_{π} is a $|\mathcal{S}|$ dimensional vector while P_{π} is a $|\mathcal{S}| \times |\mathcal{S}|$ dimensional matrix.

§ For all s' , $p_{\pi}(s'|s)$ is one row (s^{th} row) of the P_{π} matrix. Similarly, $v_{\pi}(s')$'s are the value functions for all states *i.e.*, in the vectorized notation, this is a vector v_{π} .

Existence and Uniqueness of Bellman Equations

§ Let u and v be in \mathcal{V} . So,

$$L_{\pi}u(s) = r_{\pi}(s) + \gamma \sum_{s'} p_{\pi}(s'|s)u(s')$$

$$L_{\pi}v(s) = r_{\pi}(s) + \gamma \sum_{s'} p_{\pi}(s'|s)v(s') \quad (7)$$

§ **One important note:** $L_{\pi}u(s)$ or $L_{\pi}v(s)$ does not mean L_{π} applied on $u(s)$ or $v(s)$. It means the s^{th} component of the vector $L_{\pi}u$ or $L_{\pi}v$

Existence and Uniqueness of Bellman Equations

§ Let us consider the case, $L_\pi v(s) > L_\pi u(s)$. Then,

$$\begin{aligned}
 0 &\leq L_\pi v(s) - L_\pi u(s) \\
 &= \gamma \sum_{s'} p_\pi(s'|s) \{v(s') - u(s')\} \\
 &\leq \gamma \|v - u\| \sum_{s'} p_\pi(s'|s) \\
 &\quad \text{[Why is this?]} \\
 &= \gamma \|v - u\| \quad \text{[Since } \sum_{s'} p_\pi(s'|s) = 1\text{]} \tag{8}
 \end{aligned}$$

§ Similarly, when $L_\pi u(s) > L_\pi v(s)$, we can show that,

$$0 \leq L_\pi u(s) - L_\pi v(s) \leq \gamma \|u - v\| = \gamma \|v - u\| \quad \text{[Since } \|u - v\| = \|v - u\|\text{]} \tag{9}$$

Existence and Uniqueness of Bellman Equations

§ Putting the two equations 8 and 9 together, we can get that

$$|L_\pi v(s) - L_\pi u(s)| \leq \gamma \|v - u\| \quad \forall s \in \mathcal{S} \tag{10}$$

§ Pointwise or componentwise the difference is being drawn closer by a factor of γ , so the maximum of the difference will also have come down.

$$\|L_\pi v - L_\pi u\| \leq \gamma \|v - u\| \tag{11}$$

§ So, that means that L_π is a contraction.

Existence and Uniqueness of Bellman Equations

§ Another proof of contraction property of Bellman expectation operator.

$$\begin{aligned}
 \|L_\pi v - L_\pi u\|_\infty &= \max_{s \in \mathcal{S}} |L_\pi v(s) - L_\pi u(s)| \\
 &= \max_{s \in \mathcal{S}} \left| r_\pi(s) + \gamma \sum_{s' \in \mathcal{S}} p_\pi(s'|s)v(s') - r_\pi(s) - \gamma \sum_{s' \in \mathcal{S}} p_\pi(s'|s)u(s') \right| \\
 &= \gamma \max_{s \in \mathcal{S}} \left| \sum_{s' \in \mathcal{S}} p_\pi(s'|s) \{v(s') - u(s')\} \right| \\
 &\leq \gamma \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} p_\pi(s'|s) |\{v(s') - u(s')\}| \\
 &\leq \gamma \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} p_\pi(s'|s) \| \{v - u\} \|_\infty
 \end{aligned}$$

[Absolute value each element \leq max norm of a vector]

$$= \gamma \| \{v - u\} \|_\infty \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} p_\pi(s'|s) = \gamma \| \{v - u\} \|_\infty$$

Existence and Uniqueness of Bellman Equations

§ Next we have to move on to the Bellman optimality equation's convergence proof.

§ Bellman optimality equation is given by

$$v_*(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v_*(s') \right\} \quad (12)$$

§ Let us define the Bellman optimality operator,

$L : \mathcal{V} \rightarrow \mathcal{V}$ such that

$$(Lv)(s) \equiv \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v(s') \right\} \quad \forall v \in \mathcal{V} \quad (13)$$

§ To declutter notation, we will use $Lv(s)$ to denote $(Lv)(s)$.

§ Then Bellman optimality equation becomes

$$v_* = Lv_* \quad \backslash \backslash \text{Componentwise} \quad (14)$$

Existence and Uniqueness of Bellman Equations

- § Now we will prove that L is contraction by taking the same route as we took for L_π .
- § Let u and v be in \mathcal{V} . Let us also assume, first, that $Lv(s) \geq Lu(s)$. Then we can write,

$$\begin{aligned}
 0 &\leq Lv(s) - Lu(s) \\
 &= \left\{ r(s, a_s^*) + \gamma \sum_{s'} p(s'/s, a_s^*) v(s') \right\} - \left\{ r(s, (a'_s)^*) + \gamma \sum_{s'} p(s'/s, (a'_s)^*) u(s') \right\} \\
 &\leq \left\{ r(s, a_s^*) + \gamma \sum_{s'} p(s'/s, a_s^*) v(s') \right\} - \left\{ r(s, a_s^*) + \gamma \sum_{s'} p(s'/s, a_s^*) u(s') \right\}
 \end{aligned}$$

[why?? Note what has changed!] (15)

Existence and Uniqueness of Bellman Equations

- § Now we will prove that L is contraction by taking the same route as we took for L_π .
- § Let u and v be in \mathcal{V} . Let us also assume, first, that $Lv(s) \geq Lu(s)$. Then we can write,

$$\begin{aligned}
 0 &\leq Lv(s) - Lu(s) \\
 &= \left\{ r(s, a_s^*) + \gamma \sum_{s'} p(s'/s, a_s^*) v(s') \right\} - \left\{ r(s, (a'_s)^*) + \gamma \sum_{s'} p(s'/s, (a'_s)^*) u(s') \right\} \\
 &\leq \left\{ r(s, a_s^*) + \gamma \sum_{s'} p(s'/s, a_s^*) v(s') \right\} - \left\{ r(s, a_s^*) + \gamma \sum_{s'} p(s'/s, a_s^*) u(s') \right\}
 \end{aligned}$$

[why?? Note what has changed!] (15)

- § The two actions a_s^* and $(a'_s)^*$ maximize the value functions v and u respectively at state s . So replacing $(a'_s)^*$ with a_s^* , in the second part reduces the value of the second part.

Existence and Uniqueness of Bellman Equations

$$\begin{aligned}0 &\leq Lv(s) - Lu(s) \\ &\leq \left\{ r(s, a_s^*) + \gamma \sum_{s'} p(s'/s, a_s^*) v(s') \right\} - \left\{ r(s, a_s^*) + \gamma \sum_{s'} p(s'/s, a_s^*) u(s') \right\} \\ &= \gamma \sum_{s'} p(s'/s, a_s^*) [v(s') - u(s')] \\ &\leq \gamma \sum_{s'} p(s'/s, a_s^*) \|v - u\| \quad [\text{Use of max norm similar to } L_\pi] \\ &= \gamma \|v - u\| \quad [\text{Since } \sum_{s'} p(s'/s, a_s^*) = 1]\end{aligned}\tag{16}$$

Similarly, for the second case $Lu(s) \geq Lv(s)$, we can write,

$$0 \leq Lu(s) - Lv(s) \leq \gamma \|v - u\|\tag{17}$$

Combining eqns. (16) and (17), $|Lv(s) - Lu(s)| \leq \gamma \|v - u\| \forall s \in \mathcal{S}$ which again from definition of max norm leads to $\|Lv - Lu\| \leq \gamma \|v - u\|$

Value Iteration Theorem

Theorem (Value Iteration Theorem(ref. S P. Singh and R C. Yee, 1993))

Let $v^0 \in \mathcal{V}, \epsilon > 0$, sequence $\{v^n\}$ is obtained from $v^{n+1} = Lv^n$, Then

- i. v^n converges in norm to v^* .
- ii. \exists a finite N at which the condition $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ is met $\forall n > N$.
- iii. $\pi(s)$ (obtained by $\arg \max_a \left\{ r(s, a) + \gamma \sum_{s'} p(s'/s, a) v^{n+1}(s') \right\} \forall s \in \mathcal{S}$) is ϵ optimal.
- iv. $\|v^{n+1} - v^*\| \leq \frac{\epsilon}{2}$ when the condition $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ holds.

Value Iteration Theorem

Theorem (Value Iteration Theorem(ref. S P. Singh and R C. Yee, 1993))

Let $v^0 \in \mathcal{V}, \epsilon > 0$, sequence $\{v^n\}$ is obtained from $v^{n+1} = Lv^n$, Then

- i. v^n converges in norm to v^* .
- ii. \exists a finite N at which the condition $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ is met $\forall n > N$.
- iii. $\pi(s)$ (obtained by $\arg \max_a \left\{ r(s, a) + \gamma \sum_{s'} p(s'/s, a) v^{n+1}(s') \right\} \forall s \in \mathcal{S}$) is ϵ optimal.
- iv. $\|v^{n+1} - v^*\| \leq \frac{\epsilon}{2}$ when the condition $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ holds.

§ statement III means $\|v_\pi - v^*\| \leq \epsilon$. And statement IV tells that $\|v^{n+1} - v^*\| \leq \frac{\epsilon}{2}$. Are they redundant?

Value Iteration Theorem

Theorem (Value Iteration Theorem(ref. S P. Singh and R C. Yee, 1993))

Let $v^0 \in \mathcal{V}, \epsilon > 0$, sequence $\{v^n\}$ is obtained from $v^{n+1} = Lv^n$, Then

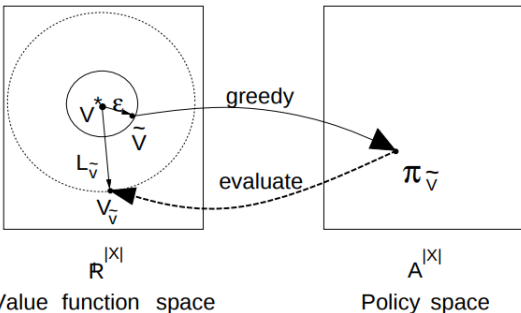
- I. v^n converges in norm to v^* .
- II. \exists a finite N at which the condition $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ is met $\forall n > N$.
- III. $\pi(s)$ (obtained by $\arg \max_a \left\{ r(s, a) + \gamma \sum_{s'} p(s'/s, a) v^{n+1}(s') \right\} \forall s \in \mathcal{S}$) is ϵ optimal.
- IV. $\|v^{n+1} - v^*\| \leq \frac{\epsilon}{2}$ when the condition $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ holds.

§ statement III means $\|v_\pi - v^*\| \leq \epsilon$. And statement IV tells that $\|v^{n+1} - v^*\| \leq \frac{\epsilon}{2}$. Are they redundant?

§ No! Think about what is v_π and what is v^{n+1} .

Value Iteration Theorem

§ Though the figure is related to policy iteration, remember the figure in slide (17).



§ Figure credit: [Singh and Yee, 1993]

§ Equality occurs if and only if value function given by the value iteration algorithm is equal to the optimal policy.

§ What V is telling is that v_π is ϵ optimal and what V is telling is that v^{n+1} is $\frac{\epsilon}{2}$ optimal given condition in V .

Proof

§ Proof: Suppose, for some n , *II* is met i.e., $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ and $\pi(s)$ obtained by *III*. Now, by triangle inequality,

$$\|v_\pi - v^*\| \leq \|v_\pi - v^{n+1}\| + \|v^{n+1} - v^*\| \tag{18}$$

Proof

§ **Proof:** Suppose, for some n , *II* is met i.e., $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ and $\pi(s)$ obtained by *III*. Now, by triangle inequality,

$$\|v_\pi - v^*\| \leq \|v_\pi - v^{n+1}\| + \|v^{n+1} - v^*\| \quad (18)$$

§ Now we have seen L_π to be such that

$$\begin{aligned} L_\pi v(s) &= r_\pi(s) + \gamma \sum_{s'} p_\pi(s'|s)v(s') \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v(s') \right\} \end{aligned} \quad (19)$$

Proof

§ **Proof:** Suppose, for some n , II is met i.e., $\|v^{n+1} - v^n\| < \epsilon \frac{1-\gamma}{2\gamma}$ and $\pi(s)$ obtained by III . Now, by triangle inequality,

$$\|v_\pi - v^*\| \leq \|v_\pi - v^{n+1}\| + \|v^{n+1} - v^*\| \tag{18}$$

§ Now we have seen L_π to be such that

$$\begin{aligned} L_\pi v(s) &= r_\pi(s) + \gamma \sum_{s'} p_\pi(s'|s)v(s') \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v(s') \right\} \end{aligned} \tag{19}$$

§ Let us apply L_π on v^{n+1} and remember that π is deterministic policy. So,

$$L_\pi v^{n+1}(s) = r(s, \pi(s)) + \gamma \sum_{s'} p(s'/s, \pi(s))v^{n+1}(s') \tag{20}$$

Proof

§ Now we have seen L to be such that

$$Lv(s) \equiv \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v(s') \right\} \quad \forall v \in \mathcal{V} \quad (21)$$

Proof

§ Now we have seen L to be such that

$$Lv(s) \equiv \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v(s') \right\} \quad \forall v \in \mathcal{V} \quad (21)$$

§ So, similarly, let us apply L on v^{n+1} . So,

$$Lv^{n+1}(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v^{n+1}(s') \right\} \quad (22)$$

Proof

§ Repeating eqn. (20) and (22)

$$L_{\pi}v^{n+1}(s) = r(s, \pi(s)) + \gamma \sum_{s'} p(s'/s, \pi(s))v^{n+1}(s') \quad (23)$$

§

$$Lv^{n+1}(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v^{n+1}(s') \right\} \quad (24)$$

Proof

§ Repeating eqn. (20) and (22)

$$L_{\pi}v^{n+1}(s) = r(s, \pi(s)) + \gamma \sum_{s'} p(s'/s, \pi(s))v^{n+1}(s') \quad (23)$$

§

$$Lv^{n+1}(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v^{n+1}(s') \right\} \quad (24)$$

§ Now, because π was chosen such that π maximizes the argument inside the $\max \{.\}$ operator, so whether we apply L_{π} on v^{n+1} or L on v^{n+1} , they are the same, i.e., $Lv^{n+1} = L_{\pi}v^{n+1}$.

Proof

§ Now let us take the first term in eqn. (18) and proceed.

$$\begin{aligned}
 \|v_\pi - v^{n+1}\| &= \|L_\pi v_\pi - v^{n+1}\| \text{ [By eqn. 6 - fixed point]} \\
 &\leq \|L_\pi v_\pi - Lv^{n+1}\| + \|Lv^{n+1} - v^{n+1}\| \text{ [Triangle inequality]} \\
 &= \|L_\pi v_\pi - L_\pi v^{n+1}\| + \|Lv^{n+1} - Lv^n\| \\
 &\quad \text{[1. Using previous slide 2. } v^{n+1} = Lv^n\text{]} \\
 &\leq \gamma \|v_\pi - v^{n+1}\| + \gamma \|v^{n+1} - v^n\| \text{ [Contraction mappings]} \\
 \implies \|v_\pi - v^{n+1}\| &\leq \frac{\gamma}{1 - \gamma} \|v^{n+1} - v^n\| \\
 &\leq \frac{\gamma}{1 - \gamma} \epsilon \frac{1 - \gamma}{2\gamma} \text{ [By statement // of the theorem]} \\
 &= \frac{\epsilon}{2} \tag{25}
 \end{aligned}$$

Proof

§ Now let us take the second term in eqn. (18) and proceed.

$$\begin{aligned}
 \|v^{n+1} - v^*\| &\leq \sum_{k=0}^{\infty} \|v^{n+k+2} - v^{n+k+1}\| \text{ [Triangle inequality repeatedly]} \\
 &= \sum_{k=0}^{\infty} \|L^{k+1}v^{n+1} - L^{k+1}v^n\| \text{ [From iterative application of } L\text{]} \\
 &\leq \sum_{k=0}^{\infty} \gamma^{k+1} \|v^{n+1} - v^n\| \text{ [} L \text{ is a contraction mapping]} \\
 &= \frac{\gamma}{1-\gamma} \|v^{n+1} - v^n\| \text{ [G.P. sum]} \\
 &\leq \frac{\gamma}{1-\gamma} \epsilon \frac{1-\gamma}{2\gamma} \text{ [By statement II of the theorem]} \\
 &= \frac{\epsilon}{2}
 \end{aligned}
 \tag{26}$$

§ this is also the proof of statement IV of the theorem

Proof

Now putting eqn. 25 and eqn. 26 in eqn. 18, we get,

$$\|v_\pi - v^*\| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \tag{27}$$

So, statement III is proved.

