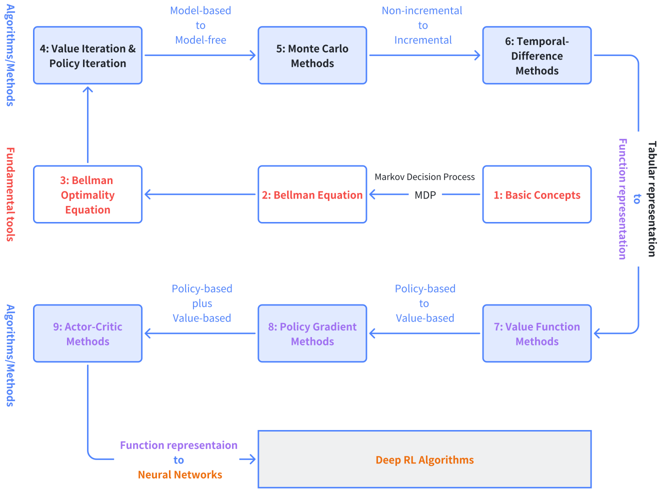
Lecture 2: State Values and the Bellman Equation

Dr. Wen Fuxi

Outline


Algorithms/Methods
4: Value Iteration & Policy Iteration
Model-based to Model-free
5: Monte Carlo Methods
Non-incremental to Incremental
6: Temporal-Difference Methods
Tabular representation to Function representation
Fundamental tools
3: Bellman Optimality Equation
2: Bellman Equation
Markov Decision Process
MDP
1: Basic Concepts
Algorithms/Methods
9: Actor-Critic Methods
Policy-based plus Value-based
8: Policy Gradient Methods
Policy-based to Value-based
7: Value Function Methods
Function representaion to Neural Networks
Deep RL Algorithms

In this lecture:

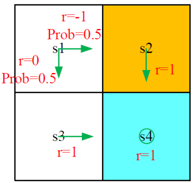
A core concept: state value

A fundamental tool: Bellman equation

Motivating example 1: Why return is important?

What is a return? The (discounted) sum of the rewards obtained along a trajectory.

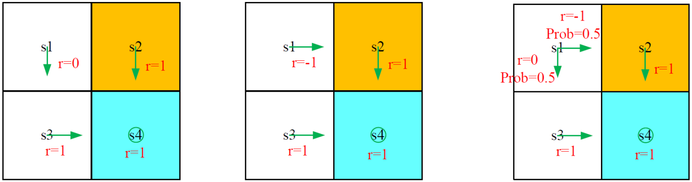
Why is return important? See the following examples.



Question: From the starting point s_1, which policy is the best (or worst)?

- Intuition: the first is the best and the second is the worst, because of the forbidden area.
- Math: Can we use mathematics to describe such intuition? Return could be used to evaluate policies. See the following.

Motivating example 1: Why return is important?



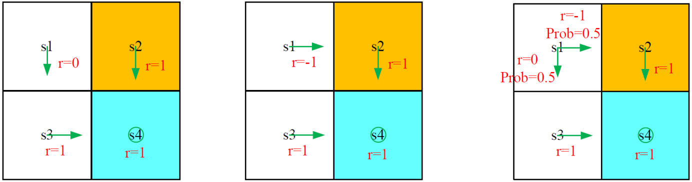
Left figure: $s_1 \xrightarrow{r=0} s_3 \xrightarrow{r=1} s_4 \xrightarrow{r=1} s_4 \xrightarrow{r=1} \cdots \xrightarrow{r=1} s_4$

Based on policy 1, starting from s_1, the discounted return is

$$\begin{aligned} \text{return}_1 &= 0 + \gamma 1 + \gamma^2 1 + \gamma^3 1 + \ldots \\ &= \gamma\left(1 + \gamma + \gamma^2 + \gamma^3 + \ldots\right) \\ &= \frac{\gamma}{1-\gamma} \end{aligned} \tag{1}$$

The geometric series formula: $\dfrac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \cdots, x \in (-1, 1)$

Motivating example 1: Why return is important?

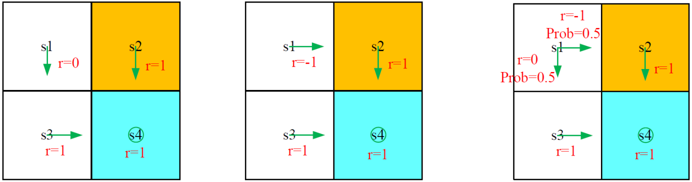


Exercise: Based on policy 2 (middle figure), starting from s_1, what is the discounted return?

Answer:

$$\begin{aligned}\text{return}_2 &= -1 + \gamma 1 + \gamma^2 1 + \gamma^3 1 + \dots \\ &= -1 + \gamma\left(1 + \gamma + \gamma^2 + \gamma^3 + \dots\right) \\ &= -1 + \frac{\gamma}{1-\gamma}\end{aligned} \tag{2}$$

Motivating example 1: Why return is important?

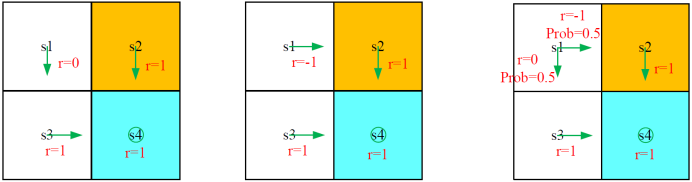


Exercise: Based on policy 3 (right figure), starting from s_1, what is the discounted return? (Policy 3 is stochastic!)

Answer:

$$\begin{aligned}\text{return}_3 &= 0.5\left(-1+\frac{\gamma}{1-\gamma}\right)+0.5\left(\frac{\gamma}{1-\gamma}\right)\\ &= -0.5+\frac{\gamma}{1-\gamma}\end{aligned} \tag{3}$$

Motivating example 1: Why return is important?



In summary, starting from s_1,

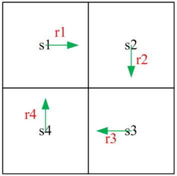
$$\text{return}_1 > \text{return}_3 > \text{return}_2$$

The above inequality suggests that the first policy is the best and the second policy is the worst, which aligns with our intuition.

Calculating the return is important for evaluating a policy.

Motivating example 2: How to calculate return?

While return is important, how do we calculate it?



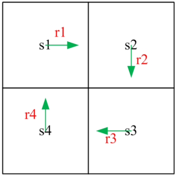
Method 1: by definition

Let v_i denote the return obtained starting from s_i $(i = 1, 2, 3, 4)$

$$\begin{aligned} v_1 &= r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \dots \\ v_2 &= r_2 + \gamma r_3 + \gamma^2 r_4 + \gamma^3 r_1 + \dots \\ v_3 &= r_3 + \gamma r_4 + \gamma^2 r_1 + \gamma^3 r_2 + \dots \\ v_4 &= r_4 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \dots \end{aligned} \tag{4}$$

Motivating example 2: How to calculate return?

While return is important, how to calculate it?



Method 2: Let v_i denote the return obtained starting from s_i $(i = 1, 2, 3, 4)$

$$v_1 = r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \ldots = r_1 + \gamma(\underline{r_2 + \gamma r_3 + \gamma^2 r_4 + \ldots}) = r_1 + \gamma v_2$$
$$v_2 = \underline{r_2 + \gamma r_3 + \gamma^2 r_4 + \gamma^3 r_1 + \ldots} = r_2 + \gamma(r_3 + \gamma r_4 + \gamma^2 r_1 + \ldots) = r_2 + \gamma v_3$$
$$v_3 = r_3 + \gamma r_4 + \gamma^2 r_1 + \gamma^3 r_2 + \ldots = r_3 + \gamma(r_4 + \gamma r_1 + \gamma^2 r_2 + \ldots) = r_3 + \gamma v_4$$
$$v_4 = r_4 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \ldots = r_4 + \gamma(r_1 + \gamma r_2 + \gamma^2 r_3 + \ldots) = r_4 + \gamma v_1$$

The returns rely on each other. Bootstrapping!

Motivating example 2: How to calculate return?

How to solve these equations? Write in the following matrix-vector form:

$$\underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{\mathbf{v}} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \begin{bmatrix} \gamma v_2 \\ \gamma v_3 \\ \gamma v_4 \\ \gamma v_1 \end{bmatrix} = \underbrace{\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}}_{\mathbf{r}} + \gamma \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{\mathbf{v}} \tag{5}$$

which can be rewritten as

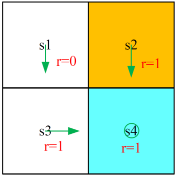
$$\mathbf{v} = \mathbf{r} + \gamma \mathbf{P} \mathbf{v} \tag{6}$$

This is the Bellman equation (for this specific deterministic problem)!!

- Though simple, it demonstrates the core idea: the value of one state relies on the values of other states.
- A matrix-vector form is more clear to see how to solve the state values.

Motivating example 2: How to calculate return?

Exercise: Consider the policy shown in the figure. Please write out the relation among the returns (that is to write out the Bellman equation)



Answer:

$$\begin{aligned} v_1 &= 0 + \gamma v_3 \\ v_2 &= 1 + \gamma v_4 \\ v_3 &= 1 + \gamma v_4 \\ v_4 &= 1 + \gamma v_4 \end{aligned} \tag{7}$$

Exercise: How to solve them? We can first calculate v_4, and then v_3, v_2, v_1.

Some notations

Consider the following single-step process (S_t, A_t, R_{t+1} are all random variables):

$$S_t \xrightarrow{A_t} R_{t+1}, S_{t+1}$$

- $t, t+1$: discrete time instances
- S_t : state at time t
- A_t : the action taken in state S_t
- R_{t+1} : the reward obtained after taking A_t
- S_{t+1} : the state transited to after taking A_t

The following probability distributions govern this step:

- $S_t \rightarrow A_t$ is governed by $\pi(A_t = a \mid S_t = s)$
- $S_t, A_t \rightarrow R_{t+1}$ is governed by $p(R_{t+1} = r \mid S_t = s, A_t = a)$
- $S_t, A_t \rightarrow S_{t+1}$ is governed by $p(S_{t+1} = s' \mid S_t = s, A_t = a)$

At this moment, we assume we know the model (i.e., the probability distributions)!

Some notations

Consider the following multi-step trajectory:

$$S_t \xrightarrow{A_t} R_{t+1}, S_{t+1} \xrightarrow{A_{t+1}} R_{t+2}, S_{t+2} \xrightarrow{A_{t+2}} R_{t+3}, \ldots \tag{8}$$

The discounted return is

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \tag{9}$$

- $\gamma \in (0,1)$ is a discount rate.
- G_t is also a random variable since $R_{t+1}, R_{t+2}, \ldots$ are random variables.

State value

The expectation (or called expected value or mean) of G_t starting from state s under policy π is defined as the state-value function or simply state value:

$$
\begin{aligned}
v^{\pi}(s) &\stackrel{\text{def}}{=} \mathbb{E}_{\pi}\left\{G_t \mid S_t = s\right\} = \mathbb{E}_{\pi}\left\{\sum_{i=0}^{+\infty} \gamma^{i} r_{t+i} \mid S_t = s\right\} \\
&= \mathbb{E}_{a_t, s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \cdots}\left\{\sum_{i=0}^{+\infty} \gamma^{i} r_{t+i} \mid S_t = s\right\}
\end{aligned}
$$

Remarks:

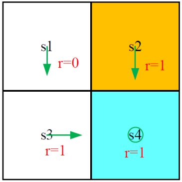
- It is a function of s. It is a conditional expectation with the condition that the state starts from s.
- It is based on the policy π. For a different policy, the state value may be different.

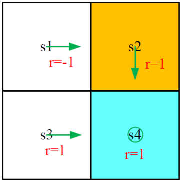
Q: What is the relationship between return and state value?

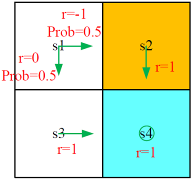
A: The state value is the mean of all possible returns that can be obtained starting from a state. If everything - $\pi(a \mid s), p(r \mid s, a), p\left(s' \mid s, a\right)$ - is deterministic, then the state value is the same as the return.

State value

Example: Which policy is good, and which is bad?







Recall the returns obtained from s_1 for the three examples:

$$v_{\pi_1}(s_1) = 0 + \gamma 1 + \gamma^2 1 + \cdots = \gamma\left(1 + \gamma + \gamma^2 + \dots\right) = \frac{\gamma}{1-\gamma}$$

$$v_{\pi_2}(s_1) = -1 + \gamma 1 + \gamma^2 1 + \cdots = -1 + \gamma\left(1 + \gamma + \gamma^2 + \dots\right) = -1 + \frac{\gamma}{1-\gamma}$$

$$v_{\pi_3}(s_1) = 0.5\left(-1 + \frac{\gamma}{1-\gamma}\right) + 0.5\left(\frac{\gamma}{1-\gamma}\right) = -0.5 + \frac{\gamma}{1-\gamma}$$

Bellman equation

- While state value is important, how to calculate? The answer lies in the Bellman equation.
- In a word, the Bellman equation describes the relationship among the values of all states.
- Next, we derive the Bellman equation.
  - There is some math.
  - We already have the intuition.

Deriving the Bellman equation

Consider a random trajectory:

$$S_t \xrightarrow{A_t} R_{t+1}, S_{t+1} \xrightarrow{A_{t+1}} R_{t+2}, S_{t+2} \xrightarrow{A_{t+2}} R_{t+3}, \ldots$$

The return G_t can be written as

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots, \\ &= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \ldots \right), \\ &= R_{t+1} + \gamma G_{t+1}, \end{aligned}$$

Then, it follows from the definition of the state value that

$$\begin{aligned} v_\pi(s) &= \mathbb{E}\left[G_t \mid S_t = s \right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s \right] \\ &= \mathbb{E}\left[R_{t+1} \mid S_t = s \right] + \gamma \mathbb{E}\left[G_{t+1} \mid S_t = s \right] \end{aligned}$$

Next, calculate the two terms, respectively.

Deriving the Bellman equation

First, calculate the first term $\mathbb{E}\left[R_{t+1} \mid S_t = s\right]$:

$$\begin{aligned}\mathbb{E}\left[R_{t+1} \mid S_t = s\right] &= \sum_a \pi(a \mid s)\mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right] \\ &= \sum_a \pi(a \mid s) \sum_r p(r \mid s, a) r\end{aligned}$$

Note that

- This is the mean of immediate rewards

Deriving the Bellman equation

Second, calculate the second term $\mathbb{E}\left[G_{t+1} \mid S_t = s\right]$:

$$
\begin{aligned}
\mathbb{E}\left[G_{t+1} \mid S_t = s\right] &= \sum_{s'} \mathbb{E}\left[G_{t+1} \mid S_t = s, S_{t+1} = s'\right] p\left(s' \mid s\right) \\
&= \sum_{s'} \mathbb{E}\left[G_{t+1} \mid S_{t+1} = s'\right] p\left(s' \mid s\right) \\
&= \sum_{s'} v_\pi\left(s'\right) p\left(s' \mid s\right) \\
&= \sum_{s'} v_\pi\left(s'\right) \sum_{a} p\left(s' \mid s, a\right) \pi(a \mid s)
\end{aligned}
$$

Note that

- This is the mean of future rewards
- $\mathbb{E}\left[G_{t+1} \mid S_t = s, S_{t+1} = s'\right] = \mathbb{E}\left[G_{t+1} \mid S_{t+1} = s'\right]$ due to the memoryless Markov property.

Deriving the Bellman equation

Therefore, we have

$$\begin{aligned} v_\pi(s) &= \mathbb{E}\left[R_{t+1} \mid S_t = s\right] + \gamma \mathbb{E}\left[G_{t+1} \mid S_t = s\right], \\ &= \underbrace{\sum_a \pi(a \mid s) \sum_r p(r \mid s, a) r}_{\text{mean of immediate rewards}} + \underbrace{\gamma \sum_a \pi(a \mid s) \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right)}_{\text{mean of future rewards}}, \\ &= \sum_a \pi(a \mid s)\left[\sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right)\right], \quad \forall s \in \mathcal{S}. \end{aligned}$$

Highlights:

- The above equation is known as the Bellman equation, which characterizes the relationship between the state-value functions of different states.
- It consists of two terms: the immediate reward term and the future reward term.
- A set of equations: every state has an equation like this.

Deriving the Bellman equation

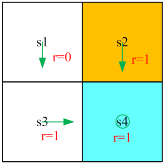
Therefore, we have

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}\left[R_{t+1} \mid S_t = s\right] + \gamma \mathbb{E}\left[G_{t+1} \mid S_t = s\right], \\
&= \underbrace{\sum_a \pi(a \mid s) \sum_r p(r \mid s, a) r}_{\text{mean of immediate rewards}} + \underbrace{\gamma \sum_a \pi(a \mid s) \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right)}_{\text{mean of future rewards}}, \\
&= \sum_a \pi(a \mid s)\left[\sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right)\right], \quad \forall s \in \mathcal{S}.
\end{aligned}
$$

Highlights: symbols in this equation

- $v_\pi(s)$ and $v_\pi\left(s'\right)$ are state values to be calculated. Bootstrapping!
- $\pi(a \mid s)$ is a given policy. Solving the equation is referred to as policy evaluation.
- $p(r \mid s, a)$ and $p\left(s' \mid s, a\right)$ represent the dynamic model. What if the model is known or unknown?

An illustrative example



Write out the Bellman equation according to the general expression:

$$v_\pi(s)=\sum_a \pi(a \mid s)\left[\sum_r p(r \mid s, a) r+\gamma \sum_{s^{\prime}} p\left(s^{\prime} \mid s, a\right) v_\pi\left(s^{\prime}\right)\right]$$

This example is simple because the policy is deterministic.

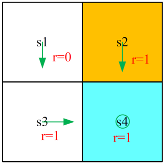
First, consider the state value of s_1 :

- $\pi\left(a=a_3 \mid s_1\right)=1$ and $\pi\left(a \neq a_3 \mid s_1\right)=0$.
- $p\left(s^{\prime}=s_3 \mid s_1, a_3\right)=1$ and $p\left(s^{\prime} \neq s_3 \mid s_1, a_3\right)=0$.
- $p\left(r=0 \mid s_1, a_3\right)=1$ and $p\left(r \neq 0 \mid s_1, a_3\right)=0$.

Substituting them into the Bellman equation gives

$$v_\pi\left(s_1\right)=0+\gamma v_\pi\left(s_3\right)$$

An illustrative example



Write out the Bellman equation according to the general expression:

$$v_\pi(s) = \sum_a \pi(a \mid s)\left[\sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right)\right]$$

Similarly, it can be obtained that

$$v_\pi\left(s_1\right) = 0 + \gamma v_\pi\left(s_3\right)$$

$$v_\pi\left(s_2\right) = 1 + \gamma v_\pi\left(s_4\right)$$

$$v_\pi\left(s_3\right) = 1 + \gamma v_\pi\left(s_4\right)$$

$$v_\pi\left(s_4\right) = 1 + \gamma v_\pi\left(s_4\right)$$

An illustrative example

How to solve them?

$$v_\pi(s_1) = 0 + \gamma v_\pi(s_3),$$
$$v_\pi(s_2) = 1 + \gamma v_\pi(s_4),$$
$$v_\pi(s_3) = 1 + \gamma v_\pi(s_4),$$
$$v_\pi(s_4) = 1 + \gamma v_\pi(s_4).$$

Solve the above equations one by one from the last to the first:

$$v_\pi(s_4) = \frac{1}{1-\gamma}$$
$$v_\pi(s_3) = \frac{1}{1-\gamma}$$
$$v_\pi(s_2) = \frac{1}{1-\gamma}$$
$$v_\pi(s_1) = \frac{\gamma}{1-\gamma}$$

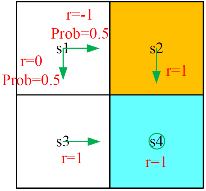
An illustrative example

If $\gamma = 0.9$, then

$$v_\pi(s_4) = \frac{1}{1-0.9} = 10$$
$$v_\pi(s_3) = \frac{1}{1-0.9} = 10$$
$$v_\pi(s_2) = \frac{1}{1-0.9} = 10$$
$$v_\pi(s_1) = \frac{0.9}{1-0.9} = 9$$

What to do after we have calculated state values? Be patient (calculating action value and improve policy)

Exercise



$$v_\pi(s)=\sum_a \pi(a \mid s)\left[\sum_r p(r \mid s, a) r+\gamma \sum_{s^{\prime}} p\left(s^{\prime} \mid s, a\right) v_\pi\left(s^{\prime}\right)\right]$$

- Write out the Bellman equations for each state.
- Solve the state values from the Bellman equations.
- Compare with the policy in the last example.

Exercise

Answer:

$$
\begin{aligned}
v_\pi(s_1) &= 0.5\left[0+\gamma v_\pi(s_3)\right]+0.5\left[-1+\gamma v_\pi(s_2)\right] \\
v_\pi(s_2) &= 1+\gamma v_\pi(s_4) \\
v_\pi(s_3) &= 1+\gamma v_\pi(s_4) \\
v_\pi(s_4) &= 1+\gamma v_\pi(s_4)
\end{aligned}
$$

Solve the above equations one by one from the last to the first.

$$
\begin{aligned}
v_\pi(s_4) &= \frac{1}{1-\gamma}, \quad v_\pi(s_3) = \frac{1}{1-\gamma}, \quad v_\pi(s_2) = \frac{1}{1-\gamma} \\
v_\pi(s_1) &= 0.5\left[0+\gamma v_\pi(s_3)\right]+0.5\left[-1+\gamma v_\pi(s_2)\right] \\
&= -0.5+\frac{\gamma}{1-\gamma}
\end{aligned}
$$

Substituting $\gamma = 0.9$ yields

$$
v_\pi(s_4) = 10, \quad v_\pi(s_3) = 10, \quad v_\pi(s_2) = 10, \quad v_\pi(s_1) = -0.5+9 = 8.5
$$

Compare with the previous policy. This one is worse.

Matrix-vector form of the Bellman equation

Why consider the matrix-vector form? Because we need to solve the state values from it!

- One unknown relies on another unknown. How to solve the unknowns?

$$v_\pi(s)=\sum_a \pi(a \mid s)\left[\sum_r p(r \mid s, a) r+\gamma \sum_{s^{\prime}} p\left(s^{\prime} \mid s, a\right) v_\pi\left(s^{\prime}\right)\right]$$

- Elementwise form: The above elementwise form is valid for every state $s \in \mathcal{S}$. That means there are $|\mathcal{S}|$ equations like this!
- Matrix-vector form: If we put all the equations together, we have a set of linear equations, which can be concisely written in a matrix-vector form. The matrix-vector form is exquisite and important.

Matrix-vector form of the Bellman equation

Recall that:

$$v_\pi(s) = \sum_a \pi(a \mid s)\left[\sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right)\right]$$

Rewrite the Bellman equation as

$$v_\pi(s) = r_\pi(s) + \gamma \sum_{s'} p_\pi\left(s' \mid s\right) v_\pi\left(s'\right) \tag{1}$$

where

$$r_\pi(s) \triangleq \sum_a \pi(a \mid s) \sum_r p(r \mid s, a) r, \quad p_\pi\left(s' \mid s\right) \triangleq \sum_a \pi(a \mid s) p\left(s' \mid s, a\right)$$

Matrix-vector form of the Bellman equation

Suppose the states could be indexed as $s_i (i = 1, \ldots, n)$.
For state s_i, the Bellman equation is

$$v_\pi (s_i) = r_\pi (s_i) + \gamma \sum_{s_j} p_\pi (s_j \mid s_i) v_\pi (s_j)$$

Put all these equations for all the states together and rewrite to a matrix-vector form

$$v_\pi = r_\pi + \gamma P_\pi v_\pi$$

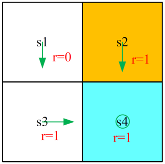
where

- $v_\pi = [v_\pi (s_1), \ldots, v_\pi (s_n)]^T \in \mathbb{R}^n$
- $r_\pi = [r_\pi (s_1), \ldots, r_\pi (s_n)]^T \in \mathbb{R}^n$
- $P_\pi \in \mathbb{R}^{n \times n}$, where $[P_\pi]_{ij} = p_\pi (s_j \mid s_i)$, is the state transition matrix

Illustrative examples

If there are four states, $v_\pi = r_\pi + \gamma P_\pi v_\pi$ can be written out as

$$\underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} = \underbrace{\begin{bmatrix} r_\pi(s_1) \\ r_\pi(s_2) \\ r_\pi(s_3) \\ r_\pi(s_4) \end{bmatrix}}_{r_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(s_1 \mid s_1) & p_\pi(s_2 \mid s_1) & p_\pi(s_3 \mid s_1) & p_\pi(s_4 \mid s_1) \\ p_\pi(s_1 \mid s_2) & p_\pi(s_2 \mid s_2) & p_\pi(s_3 \mid s_2) & p_\pi(s_4 \mid s_2) \\ p_\pi(s_1 \mid s_3) & p_\pi(s_2 \mid s_3) & p_\pi(s_3 \mid s_3) & p_\pi(s_4 \mid s_3) \\ p_\pi(s_1 \mid s_4) & p_\pi(s_2 \mid s_4) & p_\pi(s_3 \mid s_4) & p_\pi(s_4 \mid s_4) \end{bmatrix}}_{P_\pi} \underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi}$$



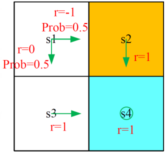
For this specific example:

$$\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}$$

Illustrative examples

If there are four states, $v_\pi = r_\pi + \gamma P_\pi v_\pi$ can be written out as

$$\underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} = \underbrace{\begin{bmatrix} r_\pi(s_1) \\ r_\pi(s_2) \\ r_\pi(s_3) \\ r_\pi(s_4) \end{bmatrix}}_{r_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(s_1 \mid s_1) & p_\pi(s_2 \mid s_1) & p_\pi(s_3 \mid s_1) & p_\pi(s_4 \mid s_1) \\ p_\pi(s_1 \mid s_2) & p_\pi(s_2 \mid s_2) & p_\pi(s_3 \mid s_2) & p_\pi(s_4 \mid s_2) \\ p_\pi(s_1 \mid s_3) & p_\pi(s_2 \mid s_3) & p_\pi(s_3 \mid s_3) & p_\pi(s_4 \mid s_3) \\ p_\pi(s_1 \mid s_4) & p_\pi(s_2 \mid s_4) & p_\pi(s_3 \mid s_4) & p_\pi(s_4 \mid s_4) \end{bmatrix}}_{P_\pi} \underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi}$$



For this specific example:

$$\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix} = \begin{bmatrix} 0.5(0) + 0.5(-1) \\ 1 \\ 1 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}$$

Solve state values

Why to solve state values?

- Given a policy, finding out the corresponding state values is called *policy evaluation*!
- It is a fundamental problem in RL. It is the foundation to find better policies.
- Therefore, it is important to understand how to solve the Bellman equation.

Solve state values

The Bellman equation in matrix-vector form is

$$v_\pi = r_\pi + \gamma P_\pi v_\pi$$

- The closed-form solution is:

$$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

- The matrix $I - \gamma P_\pi$ is inevitable.
- We still need to use numerical algorithms to calculate the matrix inverse.
- Can we avoid the matrix inverse operation? Yes, as shown below.
- An iterative solution is:

$$v_{k+1} = r_\pi + \gamma P_\pi v_k$$

This algorithm leads to a sequence $\{v_0, v_1, v_2, \ldots\}$. We can show that

$$v_k \to v_\pi = (I - \gamma P_\pi)^{-1} r_\pi, \quad k \to \infty$$

Solve state values

Proof: Define the error as $\delta_k = v_k - v_\pi$. We only need to show $\delta_k \to 0$. Substituting $v_{k+1} = \delta_{k+1} + v_\pi$ and $v_k = \delta_k + v_\pi$ into $v_{k+1} = r_\pi + \gamma P_\pi v_k$ gives

$$\delta_{k+1} + v_\pi = r_\pi + \gamma P_\pi \left(\delta_k + v_\pi\right)$$

which can be rewritten as

$$\delta_{k+1} = -v_\pi + r_\pi + \gamma P_\pi \delta_k + \gamma P_\pi v_\pi = \gamma P_\pi \delta_k$$

As a result,

$$\delta_{k+1} = \gamma P_\pi \delta_k = \gamma^2 P_\pi^2 \delta_{k-1} = \cdots = \gamma^{k+1} P_\pi^{k+1} \delta_0$$

Note that $0 \leq P_\pi^k \leq 1$, which means every entry of P_π^k is no greater than 1 for any $k = 0, 1, 2, \ldots$ That is because $P_\pi^k \mathbf{1} = \mathbf{1}$, where $\mathbf{1} = [1, \ldots, 1]^T$. On the other hand, since $\gamma < 1$, we know $\gamma^k \to 0$ and hence $\delta_{k+1} = \gamma^{k+1} P_\pi^{k+1} \delta_0 \to 0$ as $k \to \infty$.

Solve state values

Examples: $r_{\text{boundary}} = r_{\text{forbidden}} = -1, r_{\text{target}} = +1, \gamma = 0.9$

- The following are two "good" policies and the state values. The two policies are different for the top two states in the forth column.

	1	2	3	4	5
1	→	→	→	↓	↓
2	↑	↑	→	↓	↓
3	↑	←	↓	→	↓
4	↑	→	○	←	↓
5	↑	→	↑	←	←

	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

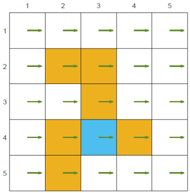
	1	2	3	4	5
1	→	→	→	→	↓
2	↑	↑	→	→	↓
3	↑	←	↓	→	↓
4	↑	→	○	←	↓
5	↑	→	↑	←	←

	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

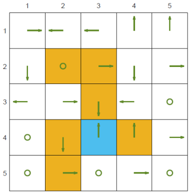
Solve state values

Examples: $r_{\text{boundary}} = r_{\text{forbidden}} = -1, r_{\text{target}} = +1, \gamma = 0.9$

- The following are two "bad" policies and the state values. The state values are less than those of the good policies.



	1	2	3	4	5
1	-6.6	-7.3	-8.1	-9.0	-10.0
2	-8.5	-8.3	-8.1	-9.0	-10.0
3	-7.5	-8.3	-8.1	-9.0	-10.0
4	-7.5	-7.2	-9.1	-9.0	-10.0
5	-7.6	-7.3	-8.1	-9.0	-10.0



	1	2	3	4	5
1	0.0	0.0	0.0	-10.0	-10.0
2	-9.0	-10.0	-0.4	-0.5	-10.0
3	-10.0	-0.5	0.5	-0.5	0.0
4	0.0	-1.0	-0.5	-0.5	-10.0
5	0.0	0.0	0.0	0.0	0.0

Action value

From state value to action value:

- State value: the average return the agent can get starting from a state.
- Action value: the average return the agent can get starting from a state and taking an action.

Why do we care about action value? Because we want to know which action is better. This point will be clearer in the following lectures.

We will frequently use action values.

Action value

Definition:

$$q_\pi(s, a) = \mathbb{E}[G_t \mid S_t = s, A_t = a]$$

- $q_\pi(s, a)$ is a function of the state-action pair (s, a)
- $q_\pi(s, a)$ depends on π

It follows from the properties of conditional expectation that

$$\underbrace{\mathbb{E}[G_t \mid S_t = s]}_{v_\pi(s)} = \sum_a \underbrace{\mathbb{E}[G_t \mid S_t = s, A_t = a]}_{q_\pi(s,a)} \pi(a \mid s)$$

Hence,

$$v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s, a) \tag{2}$$

Action value

Recall that the state value is given by

$$v_\pi(s) = \sum_a \pi(a \mid s) \underbrace{\Big[\sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right)\Big]}_{q_\pi(s,a)} \tag{3}$$

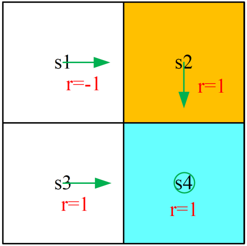
By comparing (2) and (3), we have the action-value function as

$$q_\pi(s, a) = \sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right) \tag{4}$$

(2) and (4) are the two sides of the same coin:

- (2) shows how to obtain state values from action values.
- (4) shows how to obtain action values from state values.

Illustrative example for action value



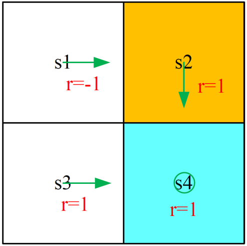
Write out the action values for state s_1.

$$q_\pi(s_1, a_2) = -1 + \gamma v_\pi(s_2),$$

Questions:

- $q_\pi(s_1, a_1), q_\pi(s_1, a_3), q_\pi(s_1, a_4), q_\pi(s_1, a_5) = ?$ Be careful!

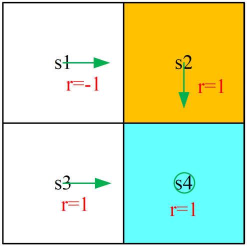
Illustrative example for action value



For the other actions:

$$q_\pi(s_1, a_1) = -1 + \gamma v_\pi(s_1),$$
$$q_\pi(s_1, a_3) = 0 + \gamma v_\pi(s_3),$$
$$q_\pi(s_1, a_4) = -1 + \gamma v_\pi(s_1),$$
$$q_\pi(s_1, a_5) = 0 + \gamma v_\pi(s_1).$$

Illustrative example for action value



Highlights:

- Action value is important since we care about which action to take.
- We can first calculate all the state values and then calculate the action values.
- We can also directly calculate the action values with or without models.

Summary

Key concepts and results:

- State value: $v_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$
- Action value: $q_\pi(s, a) = \mathbb{E}[G_t \mid S_t = s, A_t = a]$
- The Bellman equation (elementwise form):

$$
\begin{aligned}
v_\pi(s) &= \sum_a \pi(a \mid s) \underbrace{\left[\sum_r p(r \mid s, a) r + \gamma \sum_{s'} p\left(s' \mid s, a\right) v_\pi\left(s'\right)\right]}_{q_\pi(s, a)} \\
&= \sum_a \pi(a \mid s) q_\pi(s, a)
\end{aligned}
$$

- The Bellman equation (matrix-vector form):

$$
v_\pi = r_\pi + \gamma P_\pi v_\pi
$$

- How to solve the Bellman equation: closed-form solution, iterative solution