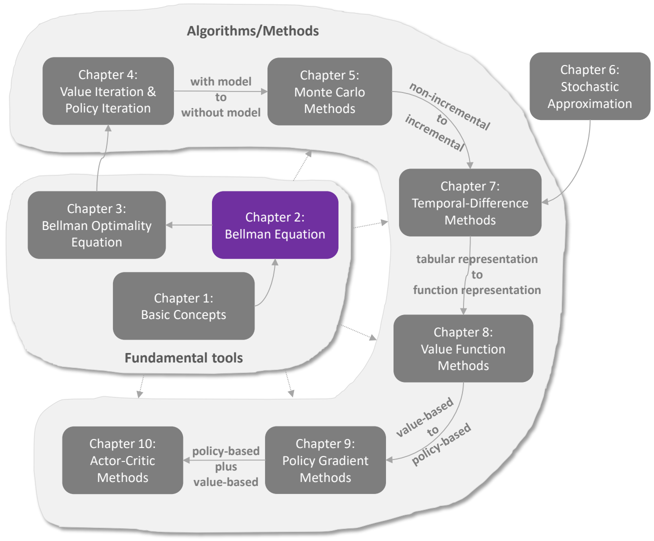
Lecture 2: State Value and Bellman Equation

Shiyu Zhao

Department of Artificial Intelligence

Westlake University

Outline



Algorithms/Methods
Chapter 4: Value Iteration & Policy Iteration
with model to without model
Chapter 5: Monte Carlo Methods
non-incremental to incremental
Chapter 6: Stochastic Approximation
Chapter 3: Bellman Optimality Equation
Chapter 2: Bellman Equation
Chapter 7: Temporal-Difference Methods
Chapter 1: Basic Concepts
tabular representation to function representation
Fundamental tools
Chapter 8: Value Function Methods
value-based to policy-based
Chapter 10: Actor-Critic Methods
policy-based plus value-based
Chapter 9: Policy Gradient Methods

Outline

In this lecture:

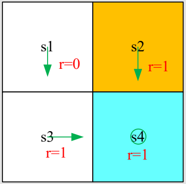
- A core concept: state value
- A fundamental tool: Bellman equation

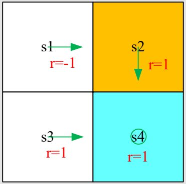
Outline

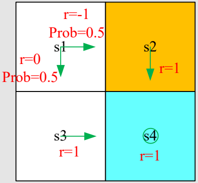
Outline

Motivating example 1: Why return is important?

- What is return? The (discounted) sum of the rewards obtained along a trajectory.
- Why is return important? See the following examples.



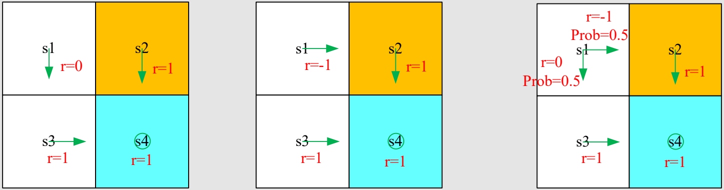




- Question: From the starting point s_1, which policy is the "best"? Which is the "worst"?
  - Intuition: the first is the best and the second is the worst, because of the forbidden area.
  - Math: can we use mathematics to describe such intuition?

  Return could be used to evaluate policies. See the following.

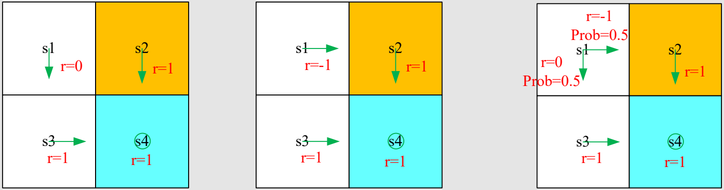
Motivating example 1: Why return is important?



Based on policy 1 (left figure), starting from s_1, the discounted return is

$$\begin{aligned}\text{return}_1 &= 0 + \gamma 1 + \gamma^2 1 + \dots \\ &= \gamma(1 + \gamma + \gamma^2 + \dots) \\ &= \frac{\gamma}{1-\gamma}\end{aligned}$$

Motivating example 1: Why return is important?

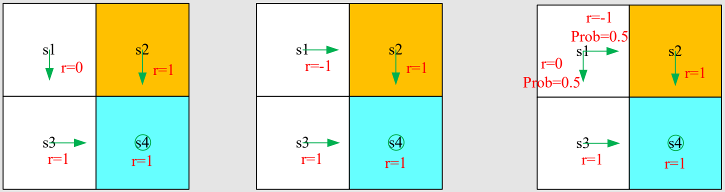


Exercise: Based on policy 2 (middle figure), starting from s_1, what is the discounted return?

Answer:

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

Motivating example 1: Why return is important?



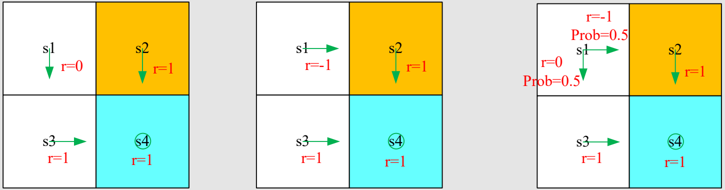
Policy 3 is stochastic!

Exercise: Based on policy 3 (right figure), starting from s_1, the discounted return is

Answer:

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

Motivating example 1: Why return is important?



In summary, starting from s_1,

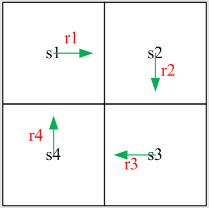
$$\text{return}_1 > \text{return}_3 > \text{return}_2$$

The above inequality suggests that the first policy is the best and the second policy is the worst, which is exactly the same as our intuition.

Calculating return is important to evaluate a policy.

Motivating example 2: How to calculate return?

While return is important, how to calculate it?



Method 1: by definition

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

$$v_1 = r_1 + \gamma r_2 + \gamma^2 r_3 + \dots$$

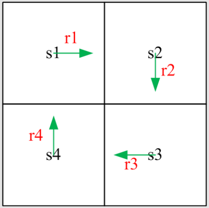
$$v_2 = r_2 + \gamma r_3 + \gamma^2 r_4 + \dots$$

$$v_3 = r_3 + \gamma r_4 + \gamma^2 r_1 + \dots$$

$$v_4 = r_4 + \gamma r_1 + \gamma^2 r_2 + \dots$$

Motivating example 2: How to calculate return?

While return is important, how to calculate it?



Method 2:

$$v_1 = r_1 + \gamma(r_2 + \gamma r_3 + \dots) = r_1 + \gamma v_2$$
$$v_2 = r_2 + \gamma(r_3 + \gamma r_4 + \dots) = r_2 + \gamma v_3$$
$$v_3 = r_3 + \gamma(r_4 + \gamma r_1 + \dots) = r_3 + \gamma v_4$$
$$v_4 = r_4 + \gamma(r_1 + \gamma r_2 + \dots) = r_4 + \gamma v_1$$

- 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:

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

which can be rewritten as

只有v未知

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

$(I - \gamma P) v = r$

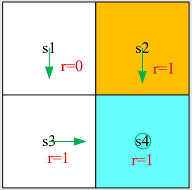
$v = (I - \gamma P)^{-1} r$

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)



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

Answer:

$$v_1 = 0 + \gamma v_3$$

$$v_2 = 1 + \gamma v_4$$

$$v_3 = 1 + \gamma v_4$$

$$v_4 = 1 + \gamma v_4$$

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

Outline

Consider the following single-step process:

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

- $t, t+1$: discrete time instances t是当前时刻，t+1 下一时刻
- 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

Note that S_t, A_t, R_{t+1} are all *random variables*.

This step is governed by the following probability distributions:

由…决定

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

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

Some notations

Consider the following multi-step trajectory:

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

The discounted return is

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

- $\gamma \in (0, 1)$ 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 is defined as the *state-value function* or simply *state value*:

$$v_\pi(s) = \mathbb{E}[G_t|S_t = s]$$

Remarks:

- It is a function of s. It is a conditional expectation with the condition that the state starts from s. 从 S 出发的所有 trajectory 的 return 的期望值
- It is based on the policy π. For a different policy, the state value may be different.

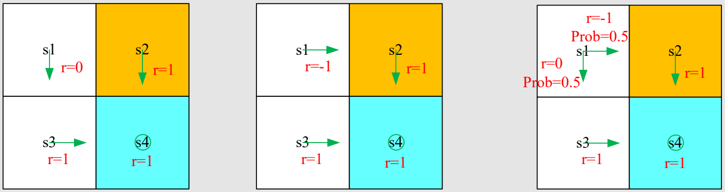
return 是对单个

state value 是对多个的平均值

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 - $\pi(a|s)$, $p(r|s,a)$, $p(s'|s,a)$ - is deterministic, then state value is the same as return.

Example: which policy is good, which is bad?



Recall the returns obtained from s_1 for the three examples: 3个 state value.

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

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

$$v_{\pi_3}(s_1) = 0.5\left(-1 + \frac{\gamma}{1-\gamma}\right) + 0.5\left(\frac{\gamma}{1-\gamma}\right) = -0.5 + \frac{\gamma}{1-\gamma}$$ 两个return的期望

Outline

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

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

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

在策略π已知的情况下：

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

Next, calculate the two terms, respectively.

Deriving the Bellman equation

在策略π中，s状态下有很多的a

所以 R_{t+1} 的期望为

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

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

在s下，选a的概率

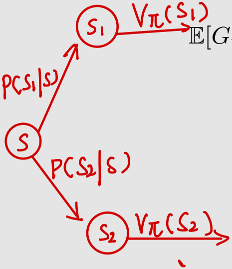
在s下，采取a时，得到reward r的概率

Note that

- This is the mean of *immediate rewards*

因为马尔可夫的 memoryless 的性质，$S_t = s$ 是没有意义的

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



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

$\sum_{s'}$ 是所有可能的后继 state.

Note that

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

Therefore, we have

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

Highlights:

- The above equation is called the *Bellman equation*, which characterizes the relationship among 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!!!

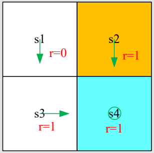
Therefore, we have

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

Highlights: symbols in this equation

- $v_\pi(s)$ and $v_\pi(s')$ are state values to be calculated. Bootstrapping!
- $\pi(a|s)$ is a given policy. Solving the equation is called policy evaluation.
- $p(r|s,a)$ and $p(s'|s,a)$ 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:

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

This example is simple because the policy is deterministic.

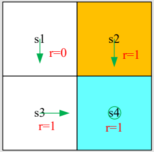
First, consider the state value of s_1:

- $\pi(a = a_3|s_1) = 1$ and $\pi(a \neq a_3|s_1) = 0$.
- $p(s' = s_3|s_1, a_3) = 1$ and $p(s' \neq s_3|s_1, a_3) = 0$.
- $p(r = 0|s_1, a_3) = 1$ and $p(r \neq 0|s_1, a_3) = 0$.

Substituting them into the Bellman equation gives

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

An illustrative example



Write out the Bellman equation according to the general expression:

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

Similarly, it can be obtained that

每个 state 都有一个 Bellman 公式
联立可求解.

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

An illustrative example

How to solve them?

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

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

$$
\begin{aligned}
v_\pi(s_4) &= \frac{1}{1-\gamma}, \\
v_\pi(s_3) &= \frac{1}{1-\gamma}, \\
v_\pi(s_2) &= \frac{1}{1-\gamma}, \\
v_\pi(s_1) &= \frac{\gamma}{1-\gamma}.
\end{aligned}
$$

An illustrative example

γ越大，越注重未来

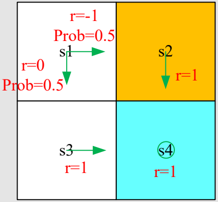
If $\gamma = 0.9$, then

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

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

Exercise

$$V_\pi(s_1) = 0.5 \times \left(-1 + \gamma V_\pi(s_2)\right) + 0.5 \times \left(0 + \gamma V_\pi(s_3)\right) = 0.5\gamma\left(V_\pi(s_2) + V_\pi(s_3)\right)$$



Exercise:

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

- 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:

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

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

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

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.

Outline

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?

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

- 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 very elegant and important.

Matrix-vector form of the Bellman equation

Recall that:

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

Rewrite the Bellman equation as

相当于又简化回去了

$$v_\pi(s) = \underline{r_\pi(s)} + \gamma \sum_{s'} p_\pi(s'|s) v_\pi(s') \qquad (1)$$

where 从当前S出发，所能得到的immediate reward的一个期望值

$$r_\pi(s) \triangleq \sum_a \pi(a|s) \sum_r p(r|s,a)r, \qquad p_\pi(s'|s) \triangleq \sum_a \pi(a|s) p(s'|s,a)$$

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

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

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

因为P_π是个特殊的matrix

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

左右两个V_π是同一个向量

where

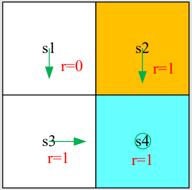
- $v_\pi = [v_\pi(s_1), \ldots, v_\pi(s_n)]^T \in \mathbb{R}^n$
- $r_\pi = [r_\pi(s_1), \ldots, r_\pi(s_n)]^T \in \mathbb{R}^n$
- $P_\pi \in \mathbb{R}^{n \times n}$, where $[P_\pi]_{ij} = p_\pi(s_j|s_i)$, is the *state transition matrix*

P_π是一个$n \times n$矩阵，i行j列的元素为$P_\pi(S_j|S_i)$

Illustrative examples

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

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

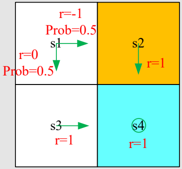


For this specific example:

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

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

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



For this specific example:

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

Outline

给出一个 Policy ⟶ 求 Bellman Equation ⟶ 解得 state value of each state .

Why to solve state values?

- Given a policy, finding out the corresponding state values is called *policy evaluation*! 用于评价一个 Policy 的好坏
- 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.

The Bellman equation in matrix-vector form is

$(I-\gamma P_\pi)v_\pi = r_\pi$

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

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

- The *closed-form solution* is:

在实际操作中不用逆矩阵，因为计算量太大 $O(n^3)$

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

  - The matrix $I - \gamma P_\pi$ is inevitable. See details in my book.
  - We still need to use numerical algorithms to calculate the matrix inverse.
  - Can we avoid the matrix inverse operation? Yes, as shown below.

通过不断迭代来求近似值　从任意初始向量v_0出发：

$v_0 \to v_1$

$v_1 \to v_2$

$v_2 \to v_3$

......

- *An iterative solution is:*

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

发现 v_{k+1} 的值在收敛到一个数 这个值就是 v_π

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

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

Solve state values (optional)

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

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

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,

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

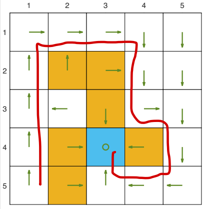
Note that $0 \leq P_\pi^k \leq 1$, which means every entry of P_π^k is no greater than 1 for any $k = 0, 1, 2, \ldots$. That is because $P_\pi^k \mathbf{1} = \mathbf{1}$, where $\mathbf{1} = [1, \ldots, 1]^T$. On the other hand, since $\gamma < 1$, we know $\gamma^k \to 0$ and hence $\delta_{k+1} = \gamma^{k+1}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

Solve state values

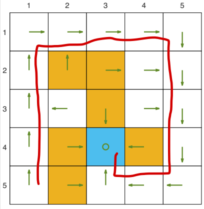
Examples: $r_{\text{boundary}} = r_{\text{forbidden}} = -1$, $r_{\text{target}} = +1$, $\gamma = 0.9$

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

靠近 target area的 state value都比较大.



	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

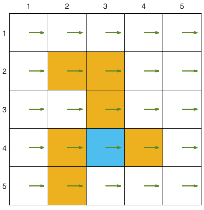


	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

Solve state values

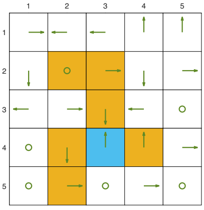
Examples: $r_{\text{boundary}} = r_{\text{forbidden}} = -1$, $r_{\text{target}} = +1$, $\gamma = 0.9$

- 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

由Bad policy生成的state value都为负数，所以不好



	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

Outline

Action value

state value $\longrightarrow$ policy的好坏

action value $\longrightarrow$ action的好坏

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 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.

已知一个Policy π，在 s 下，采取动作 a，所有 return value 的期望值

Definition:

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

- $q_\pi(s,a)$ 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

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

state value = $\sum_a$ (一个动作的 action value × 该动作的概率)

Hence,

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

action value 和 state value 互相转化.

Action value

Recall that the state value is given by

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

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

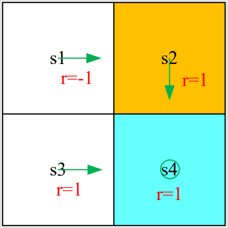
$$q_\pi(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \qquad (4)$$

$$s \xrightarrow{a,r} s' \to \cdots$$

(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.

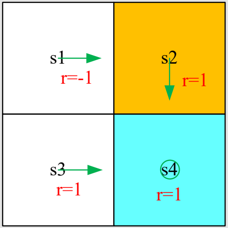
$$q_\pi(s_1, a_2) = -1 + \gamma v_\pi(s_2),$$

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!

即使Policy没有a_1, a_3, a_4, a_5，但其action values都是存在的

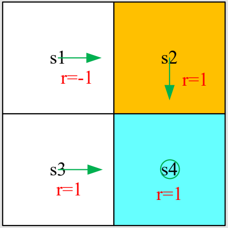
Illustrative example for action value



For the other actions:

$$
\begin{aligned}
q_\pi(s_1, a_1) &= -1 + \gamma v_\pi(s_1), \\
q_\pi(s_1, a_3) &= 0 + \gamma v_\pi(s_3), \\
q_\pi(s_1, a_4) &= -1 + \gamma v_\pi(s_1), \\
q_\pi(s_1, a_5) &= 0 + \gamma v_\pi(s_1).
\end{aligned}
$$

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.

Outline

Summary

Key concepts and results:

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

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

$$= \sum_a \pi(a|s) q_\pi(s,a)$$

- The Bellman equation (matrix-vector form):

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

- How to solve the Bellman equation: closed-form solution, <u>iterative solution</u>