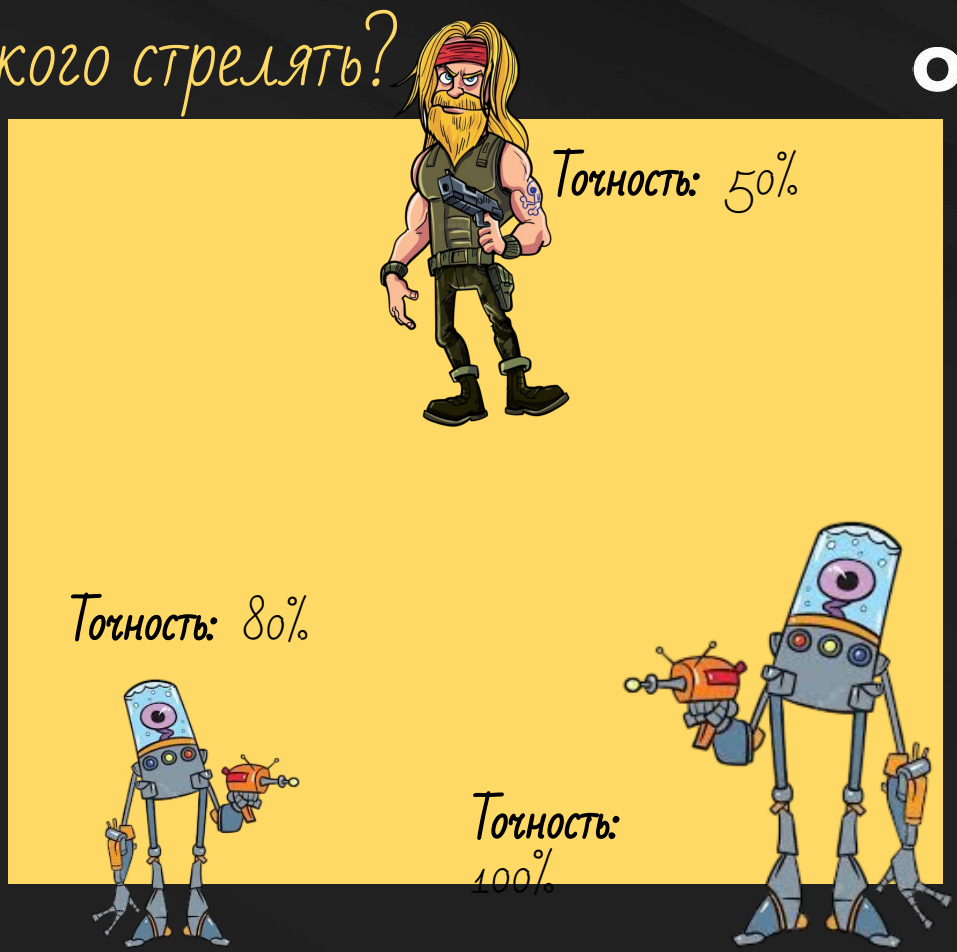


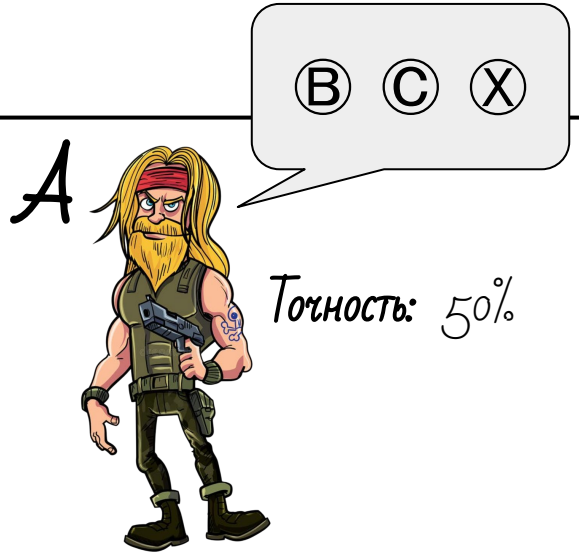
RL 1-03

Value Iteration

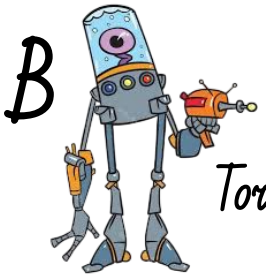
Начнем в 20:01

otus.ru

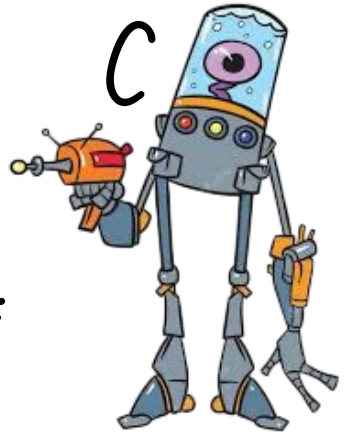
Что будем делать?



Точность: 50%



Точность: 80%



Точность:
100%

Дисклеймер

Занятие «Основные алгоритмы RL: Value based»

Цели занятия

познакомиться с основными алгоритмами, используемыми в обучении с подкреплением;
понять, что такое v-функция и q-функция;

Показать еще

Краткое содержание

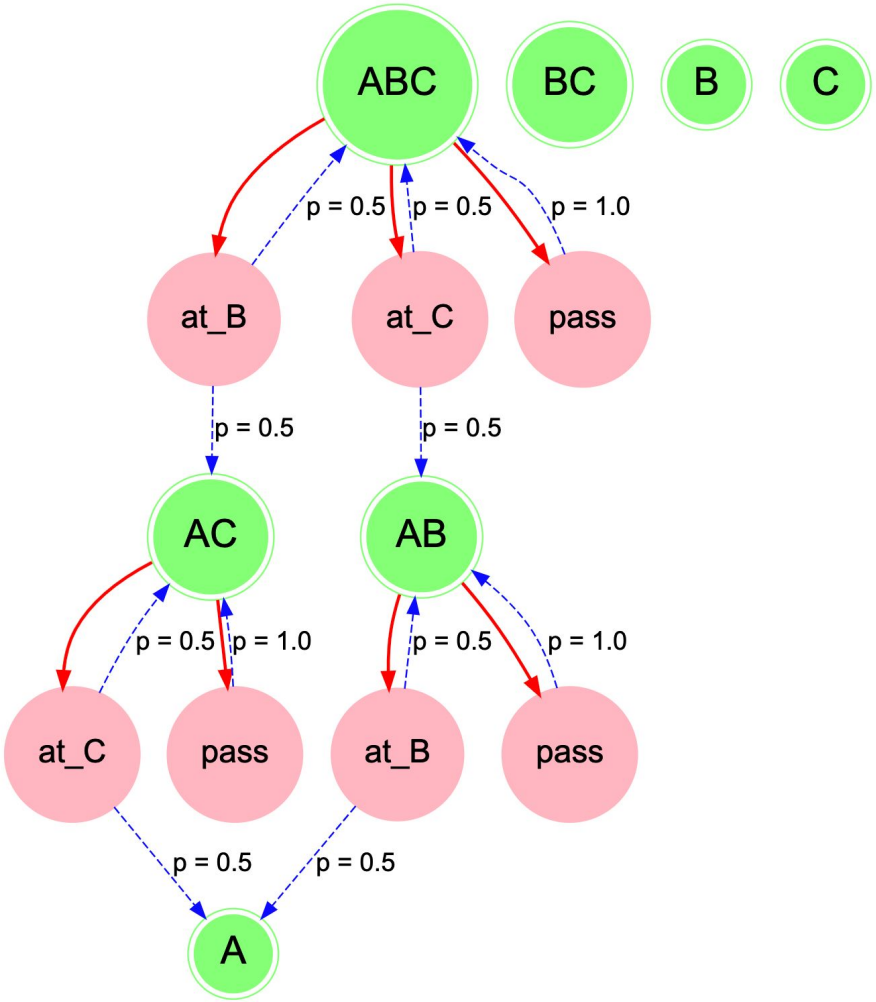
v-функция и q-функция;
уравнение Беллмана;
алгоритмы группы Value-based: SARSA, Q-learning.

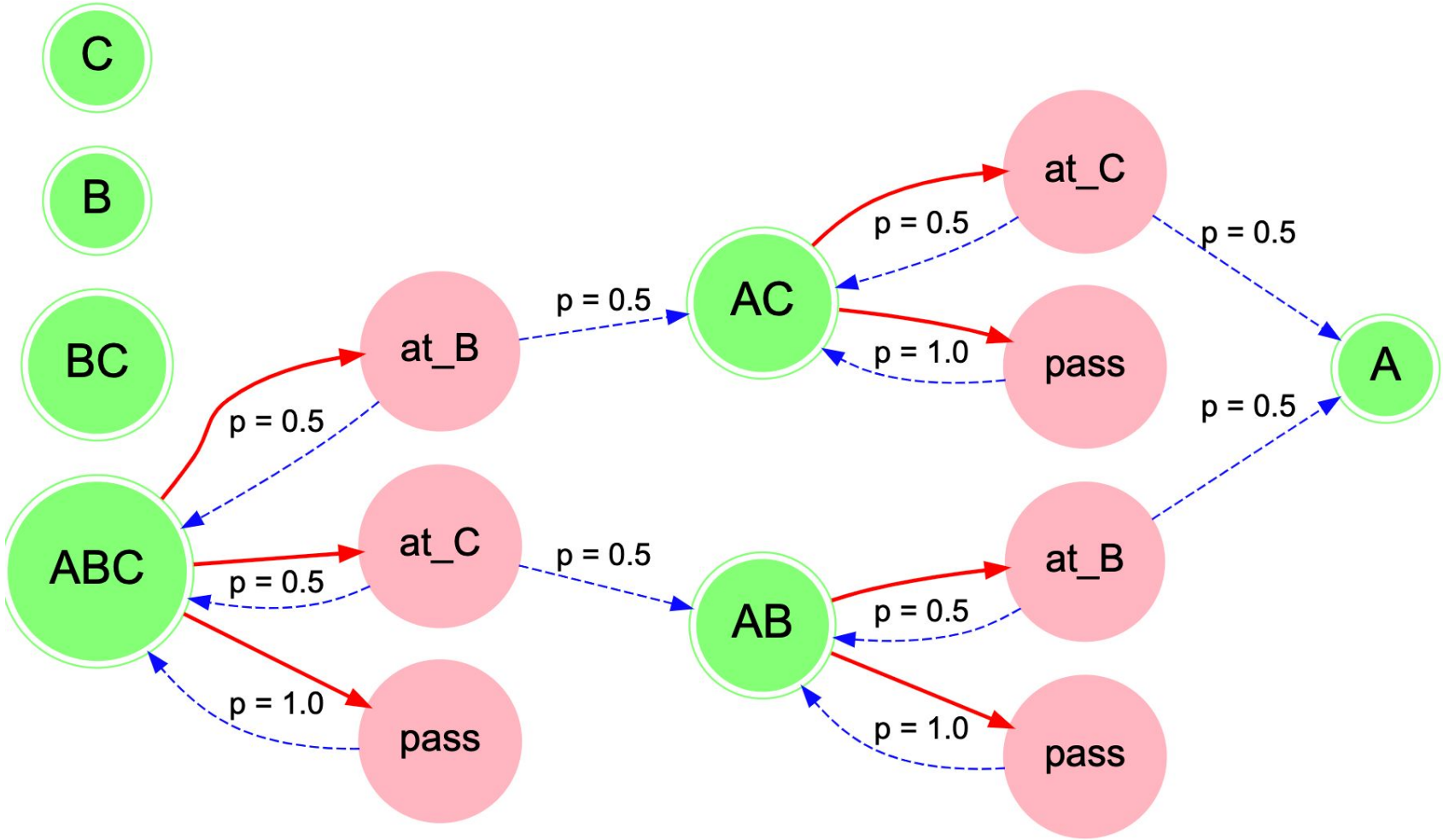
Свернуть

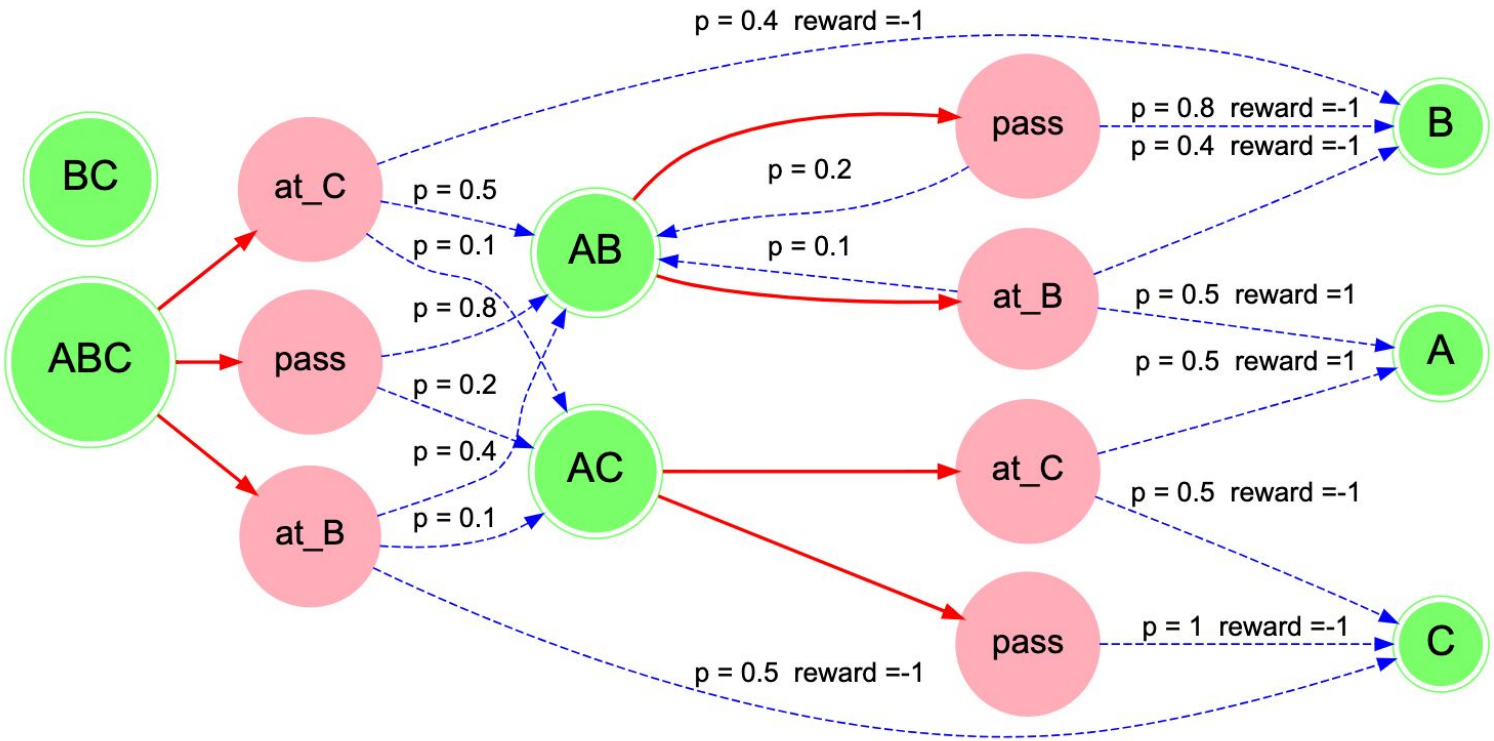
Результаты

конспект занятия;
ноутбук с кодом класса агента, реализующего алгоритм q-learning и мониторинг процесса обучения;
ответы на вопросы.

Свернуть

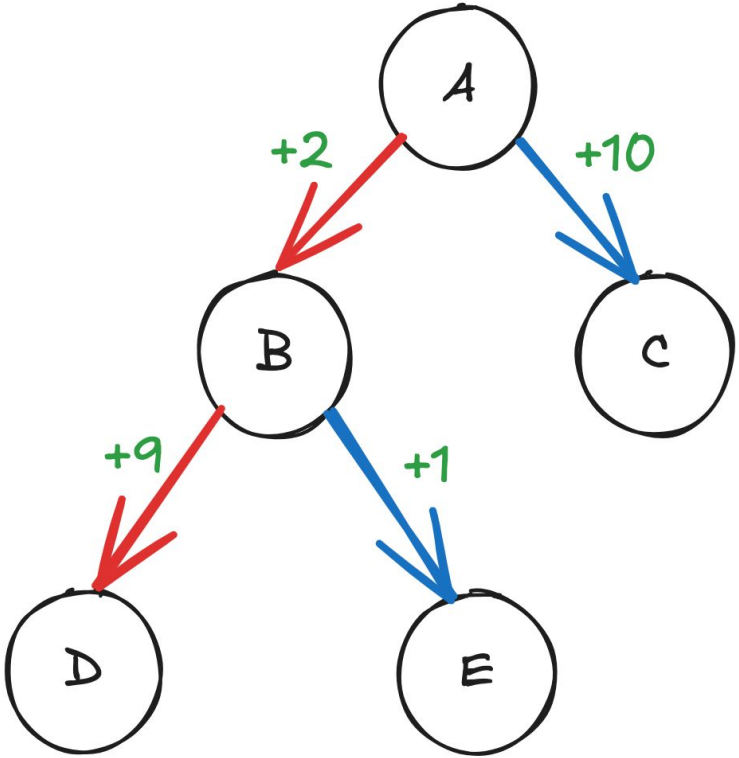

ABC
BC
B
C
p = 0.5
p = 0.5
p = 1.0
at_B
at_C
pass
p = 0.5
p = 0.5
AC
AB
p = 0.5
p = 1.0
p = 0.5
p = 1.0
at_C
pass
at_B
pass
p = 0.5
p = 0.5
A


C
B
BC
ABC
at_B
at_C
pass
p = 0.5
p = 0.5
p = 1.0
p = 0.5
AC
AB
at_C
at_B
pass
p = 0.5
p = 1.0
p = 0.5
p = 0.5
p = 1.0
A

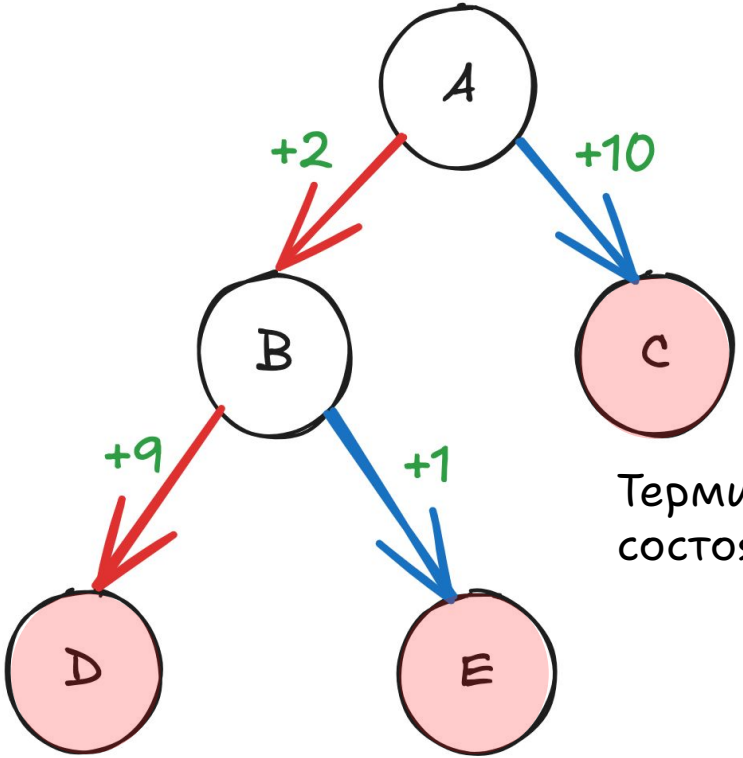

p = 0.4 reward =-1
BC
at_C
p = 0.5
p = 0.1
AB
pass
p = 0.2
p = 0.8 reward =-1
p = 0.4 reward =-1
B
p = 0.1
at_B
p = 0.5 reward =1
ABC
pass
p = 0.8
p = 0.2
A
p = 0.5 reward =1
p = 0.4
AC
at_C
p = 0.5 reward =-1
at_B
p = 0.1
pass
p = 1 reward =-1
C
p = 0.5 reward =-1

Value function

Детерминированная MDP

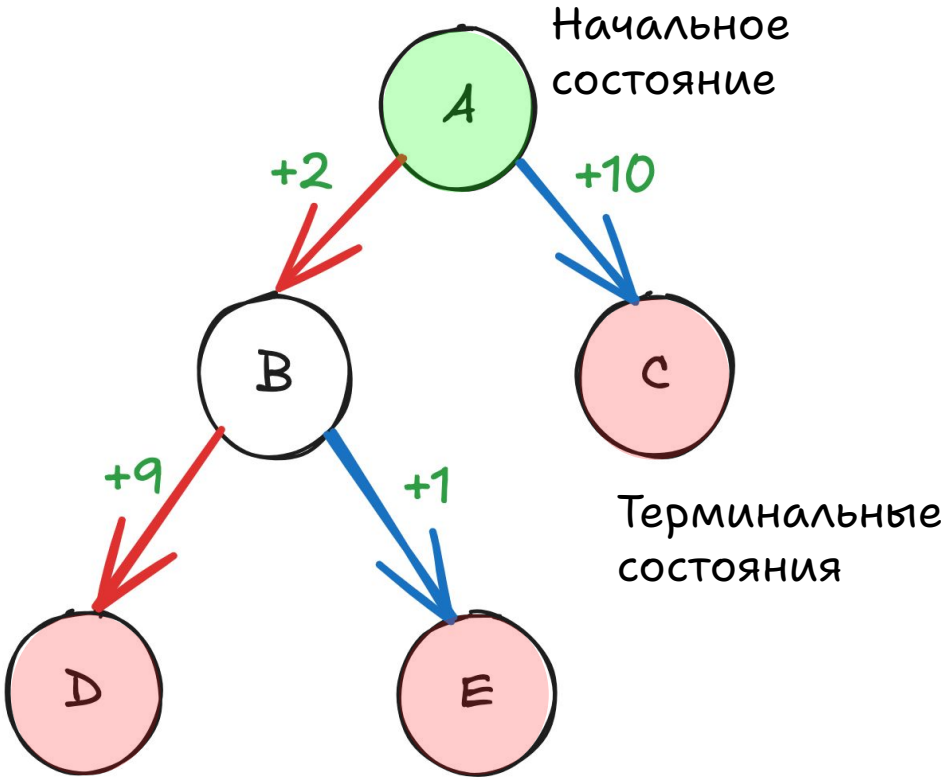


Детерминированная MDP

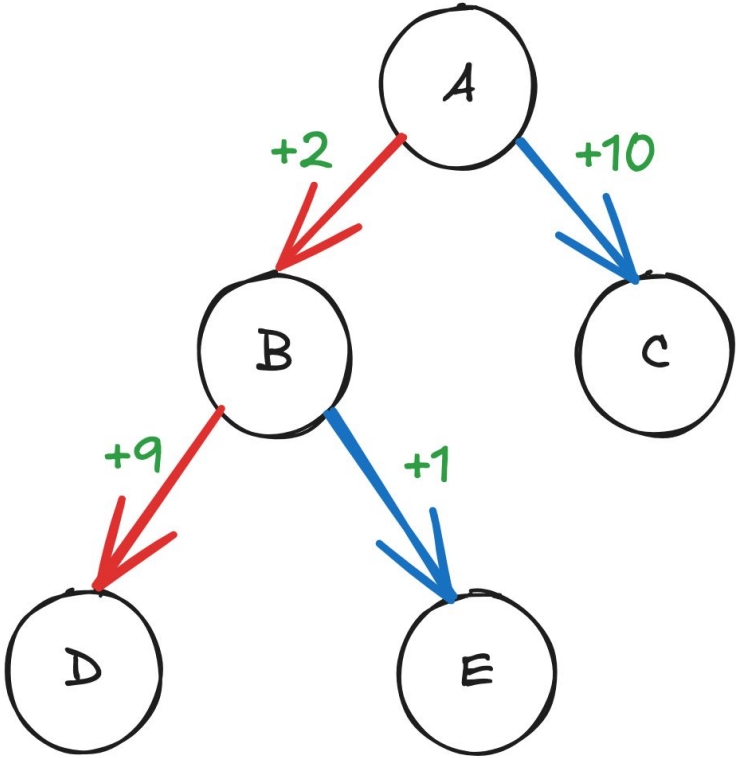


Терминальные состояния

Детерминированная MDP

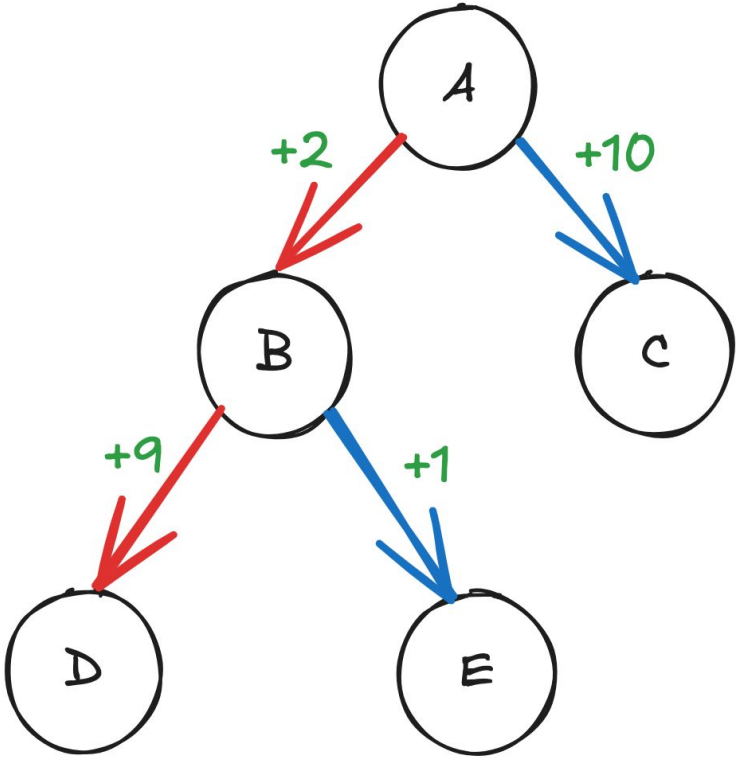


Детерминированная MDP



Какая цель?

Детерминированная MDP

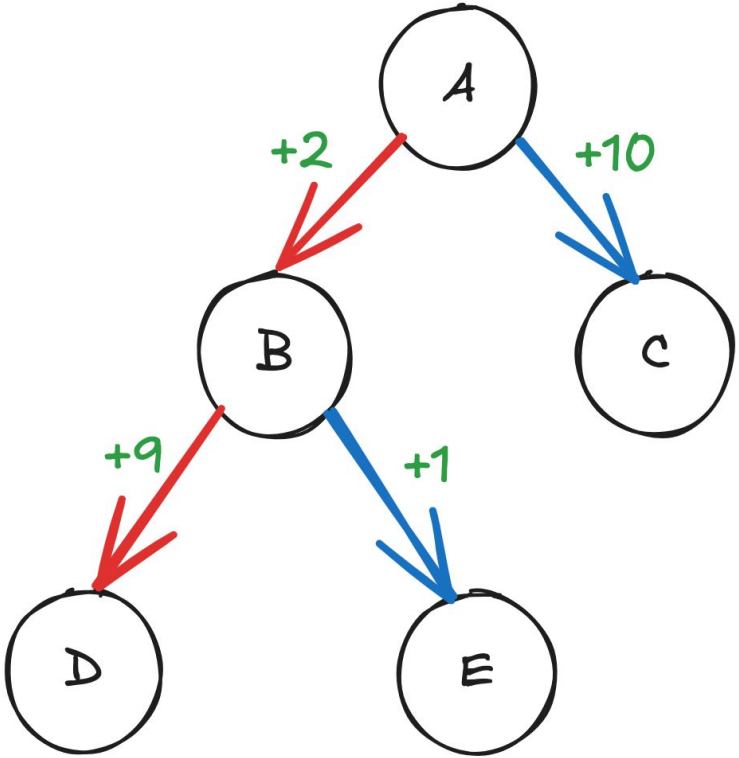


Какая цель?

Найти политику π, чтобы максимизировать сумму наград

$$\sum_{t \geq 0} r_t \to \max_{\pi},$$

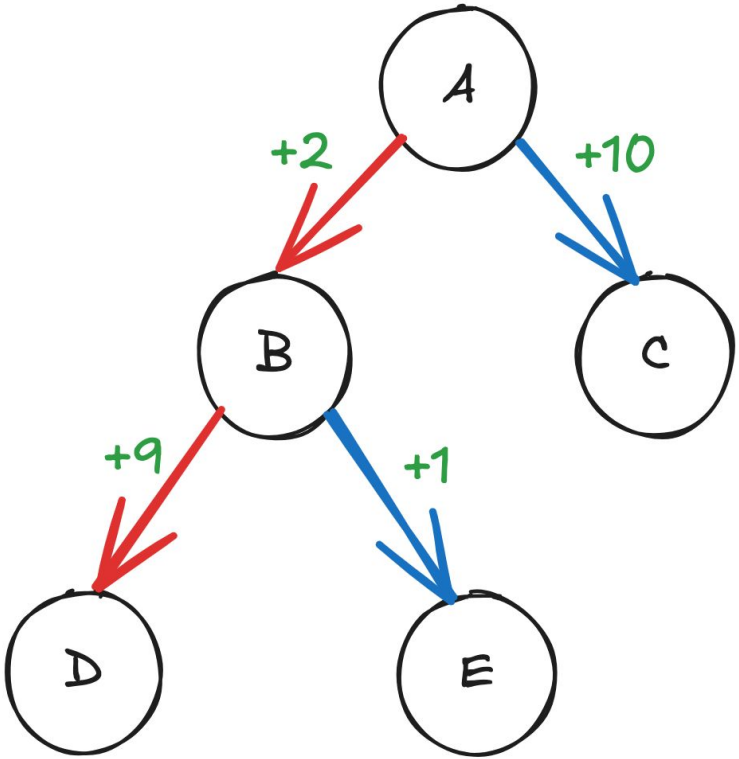
Детерминированная MDP



$$\mathcal{R}(\pi) = \sum_{t \geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

Детерминированная MDP

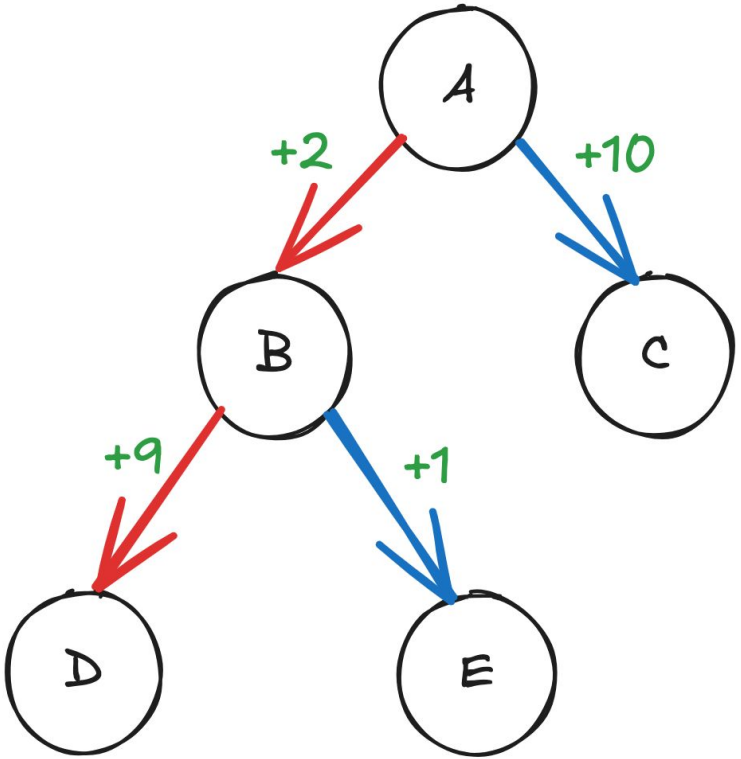


$$R(\pi) = \sum_{t \geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

P1 = {
A: red
B: red
}

Детерминированная MDP



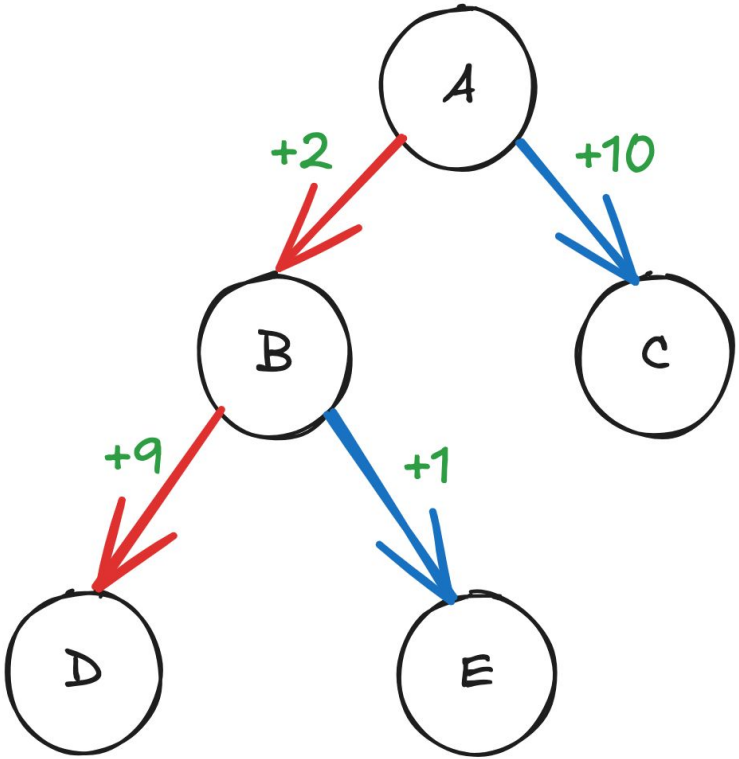
$$R(\pi) = \sum_{t \geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

P1 = {
A: red
B: red
}

P2 = {
A: red
B: blue
}

Детерминированная MDP



$$R(\pi) = \sum_{t \geq 0} r_t \to \max_{\pi},$$

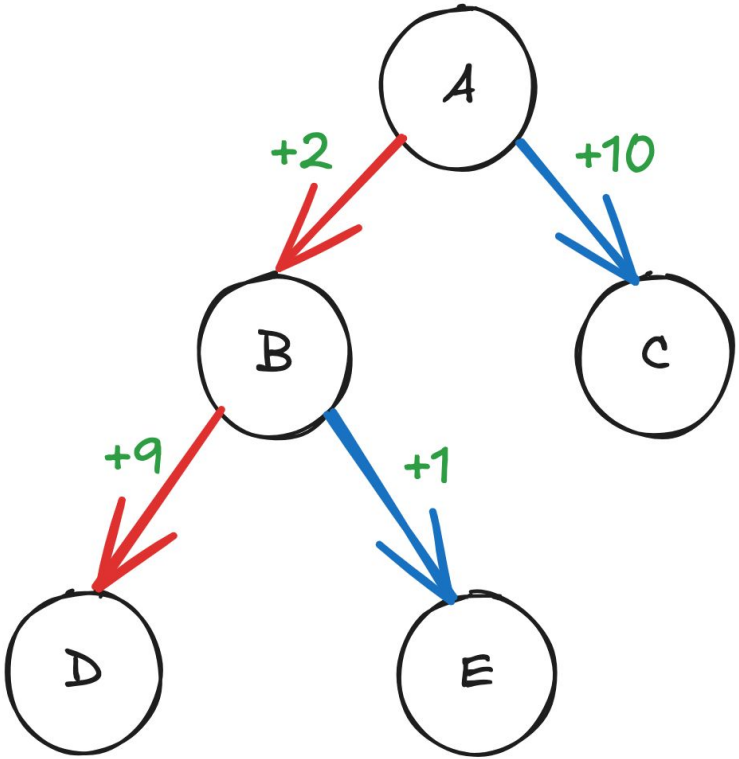
Какие политики есть?

P1 = {
A: red
B: red
}

P2 = {
A: red
B: blue
}

P3 = {
A: blue
}

Детерминированная MDP



$$R(\pi) = \sum_{t \geq 0} r_t \to \max_{\pi},$$

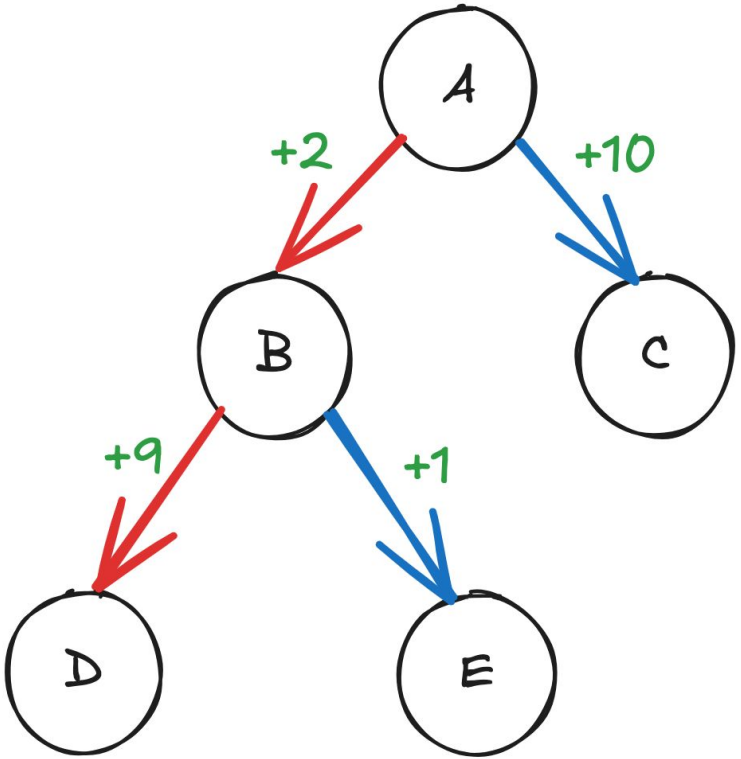
Какие политики есть?

```
P1 = {
    A: red
    B: red
}

P3 = {
    A: blue
}
```

```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

Детерминированная MDP



$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

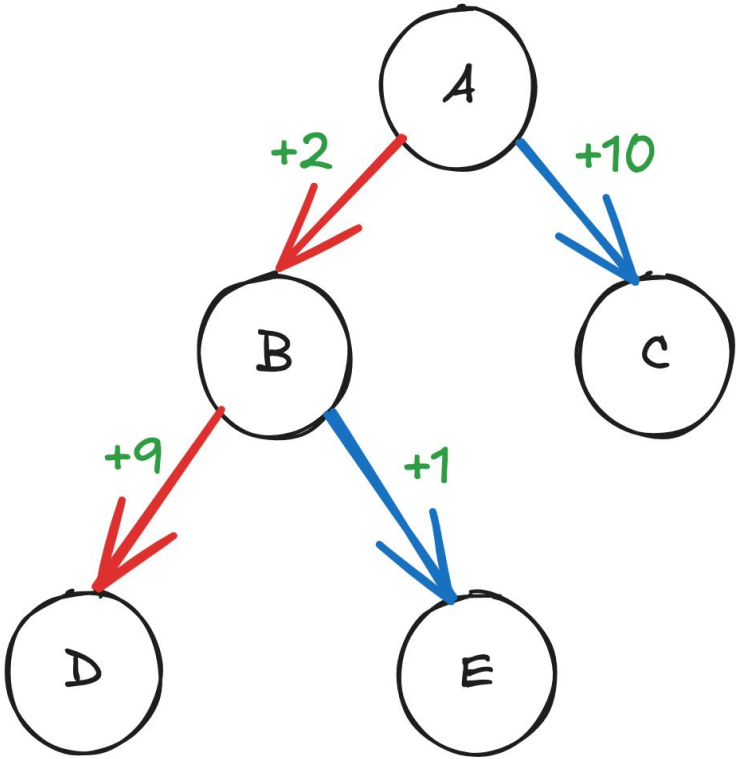
Какие политики есть?

```
P1 = {
      A: red
      B: red
}

P3 = {
      A: blue
}
```

```
P4 = {
      A: {
            blue: 0.2
            red: 0.8
      }
      B: {
            blue: 0.5
            red: 0.5
      }
}
...
```

Детерминированная MDP



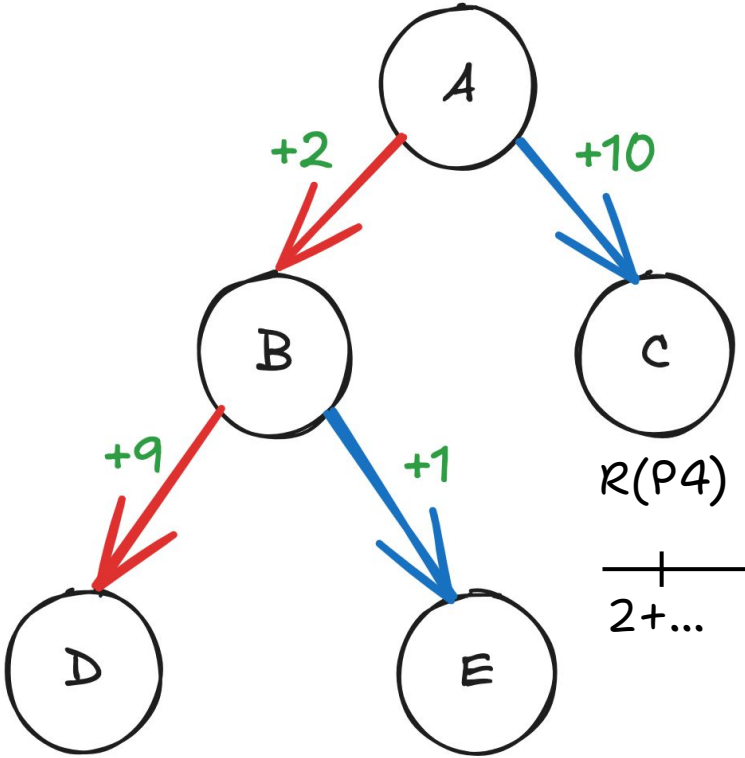
$$\mathcal{R}(\pi) = \mathbb{E}_{\mathcal{T}\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

$\mathcal{R}(\pi) = ?$

```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

Детерминированная MDP

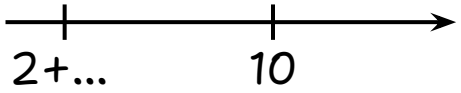


$$\mathcal{R}(\pi) = \mathbb{E}_{\mathcal{T}\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

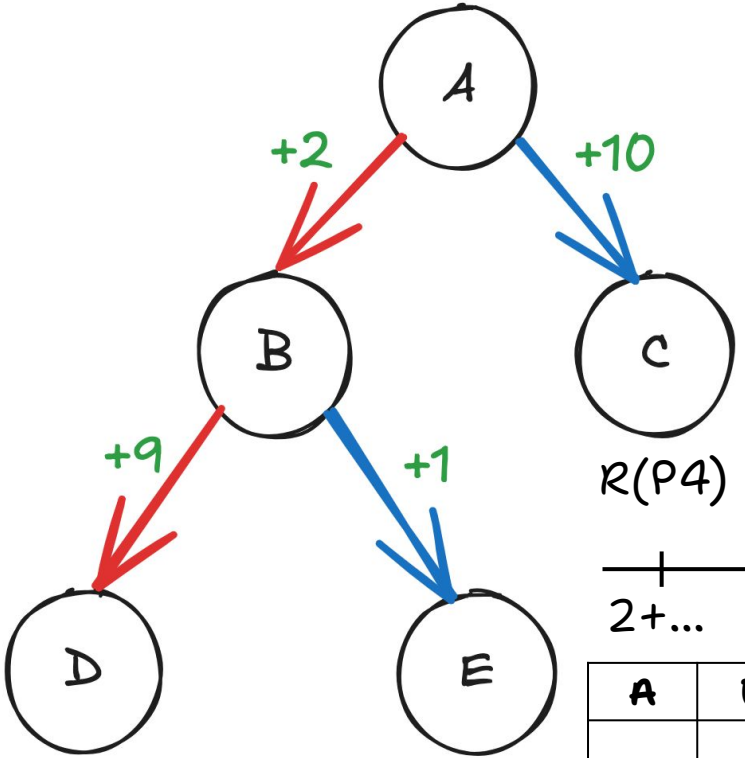
Какие политики есть?

```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

R(P4) = 0.2 * 10 + 0.8 * _



Детерминированная MDP

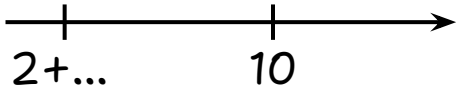


$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

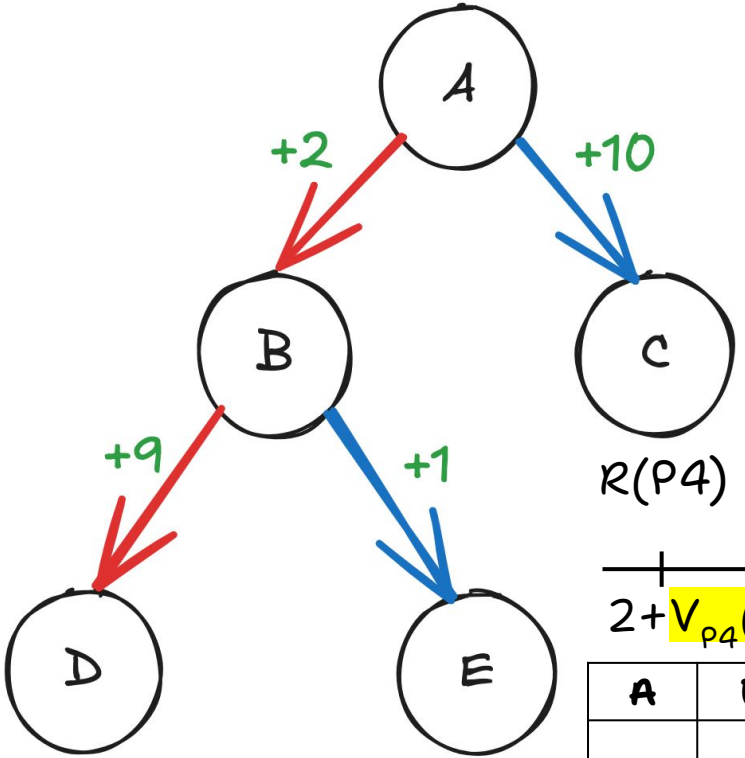
```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

R(P4) = 0.2 * 10 + 0.8 * _



A	B	C	D	E

Value-функция

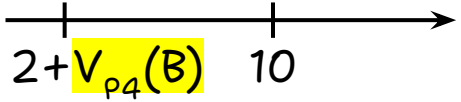


$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

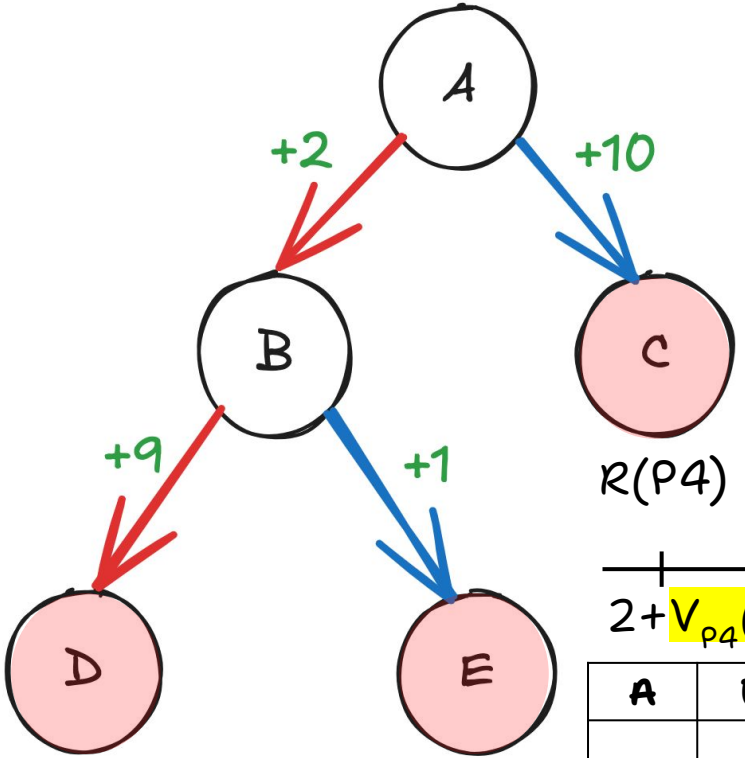
```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

$R(P4) = 0.2 * 10 + 0.8 * \_$



A	B	C	D	E

Value-функция

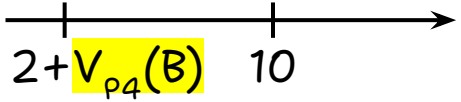


$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

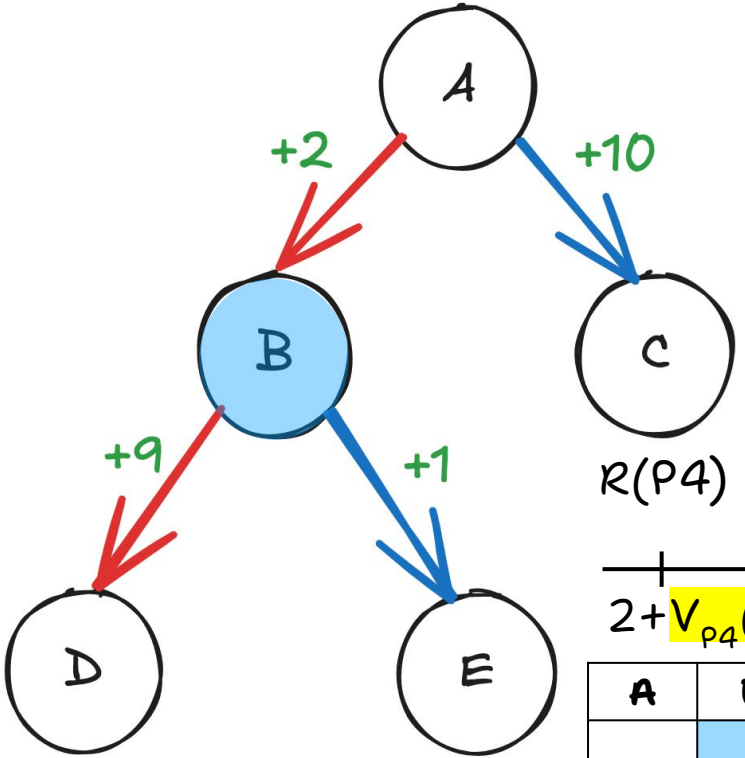
```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

R(P4) = 0.2 * 10 + 0.8 * _



A	B	C	D	E
		0	0	0

Value-функция

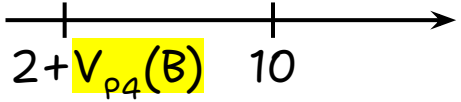


$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

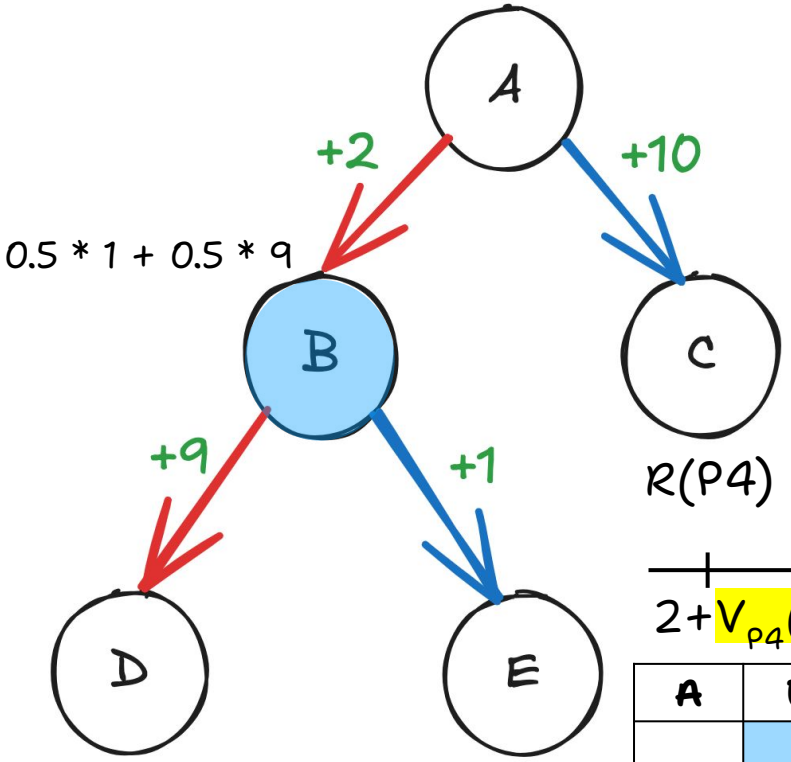
```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

$R(P4) = 0.2 * 10 + 0.8 * \_$



A	B	C	D	E
		0	0	0

Value-функция

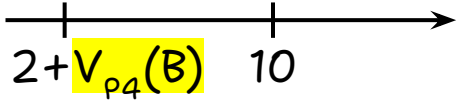


$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

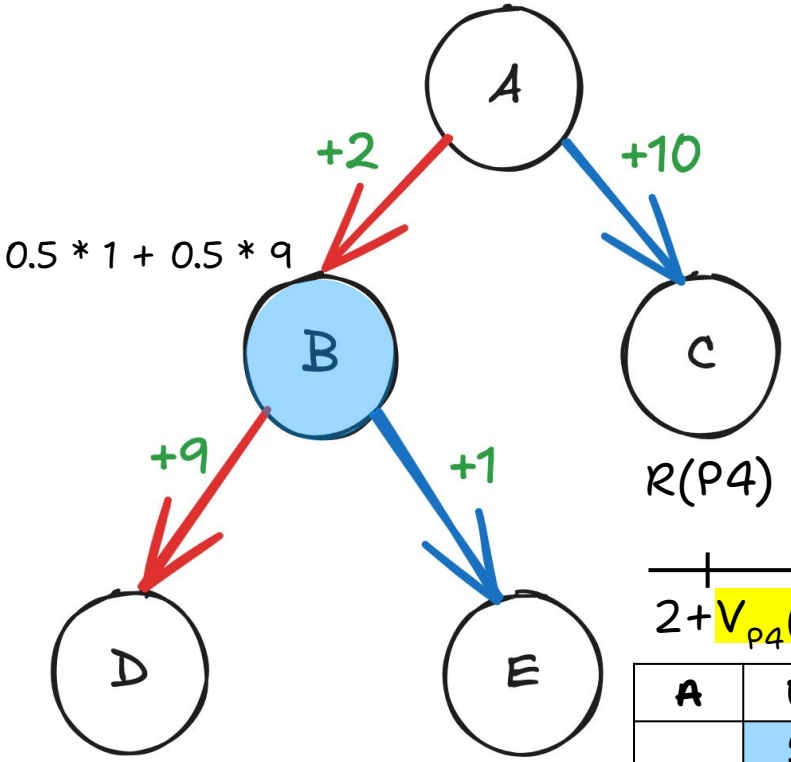
```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

$R(P4) = 0.2 * 10 + 0.8 * \_$



A	B	C	D	E
		0	0	0

Value-функция

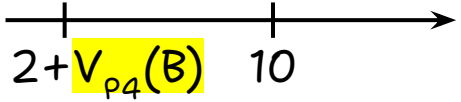


$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

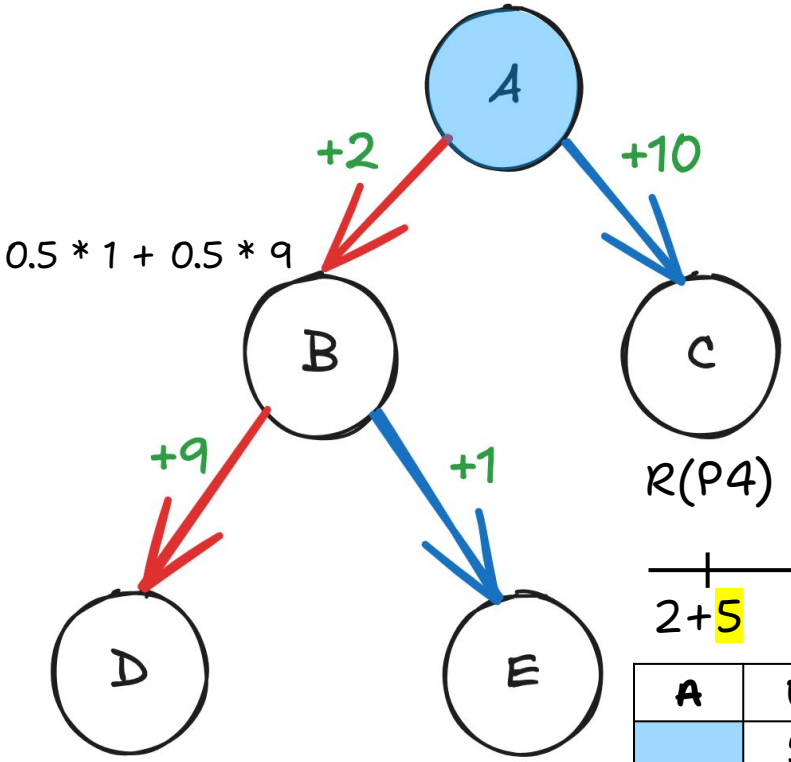
```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

$R(P4) = 0.2 * 10 + 0.8 * \_$



A	B	C	D	E
	5	0	0	0

Value-функция

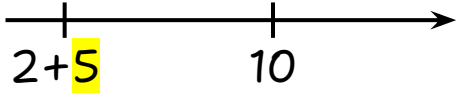


$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

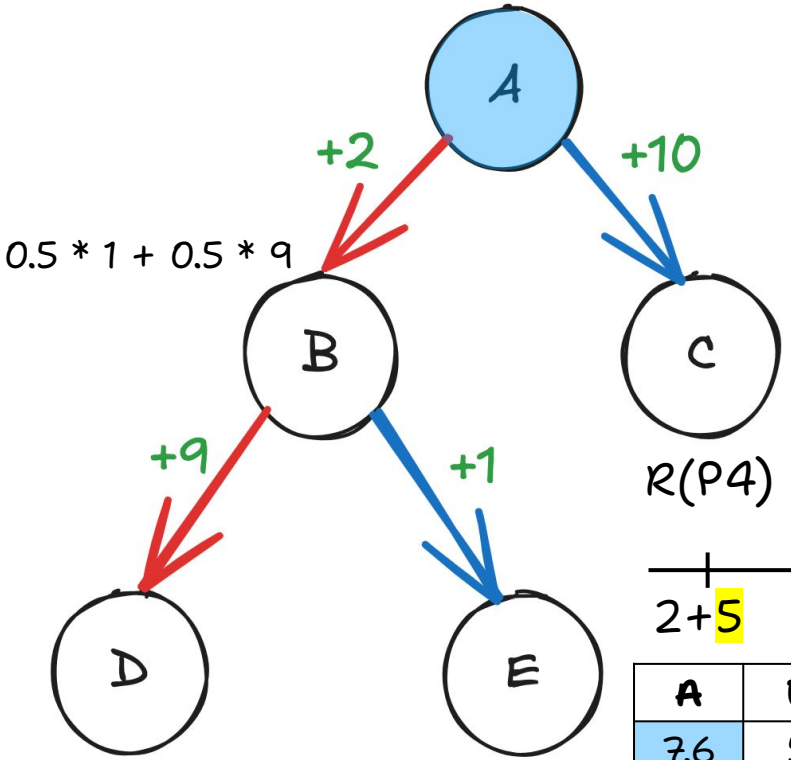
```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

R(P4) = 0.2 * 10 + 0.8 * 7



A	B	C	D	E
	5	0	0	0

Value-функция

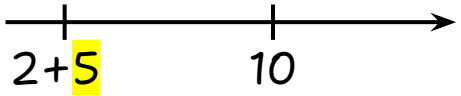


$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

Какие политики есть?

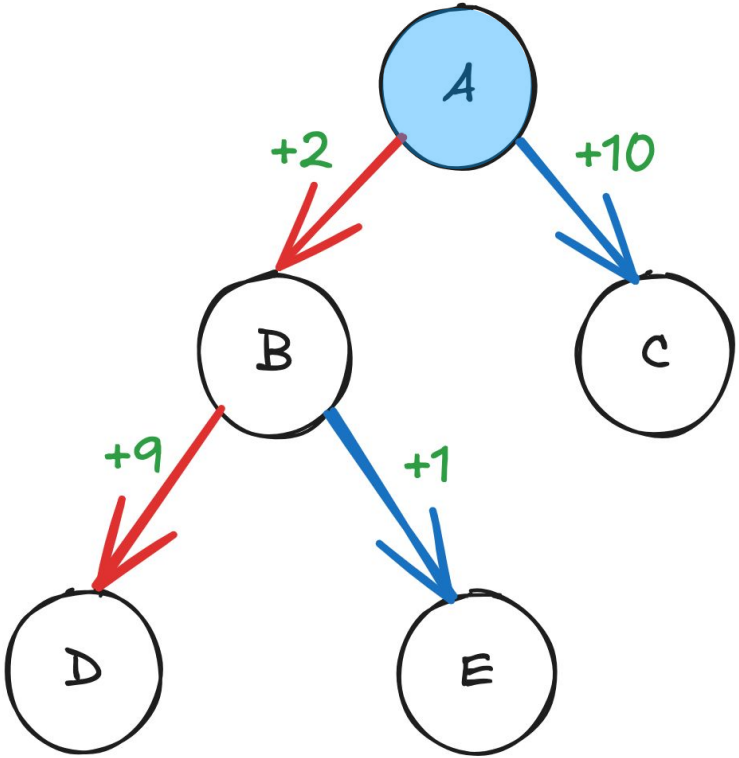
```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
...
```

R(P4) = 7.6



A	B	C	D	E
7.6	5	0	0	0

Какую политику назовем жадной?



$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

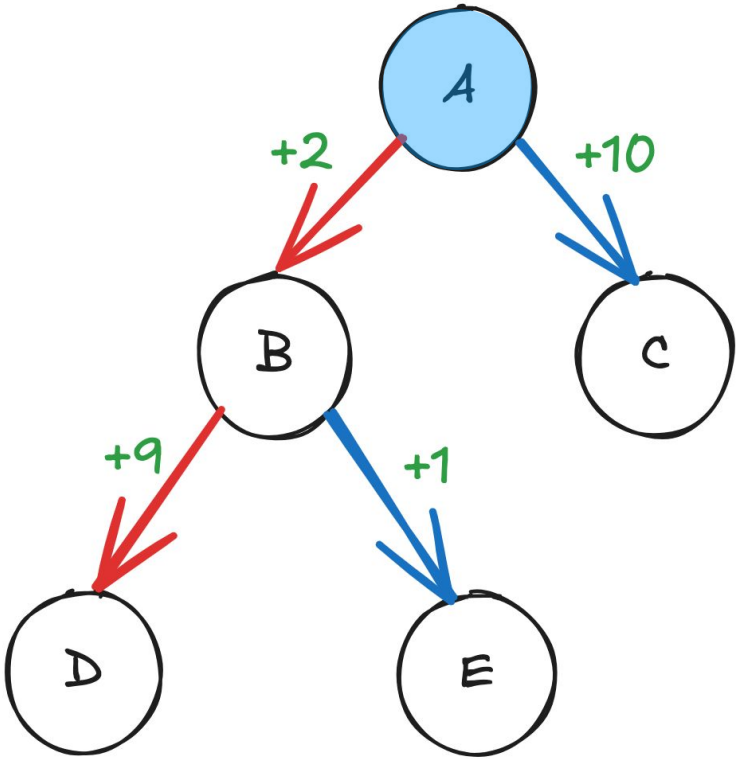
```
P1 = {
    A: red
    B: red
}
```

```
P3 = {
    A: blue
}
```

```
P2 = {
    A: red
    B: blue
}
```

```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
```

Какую политику назовем жадной?



$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

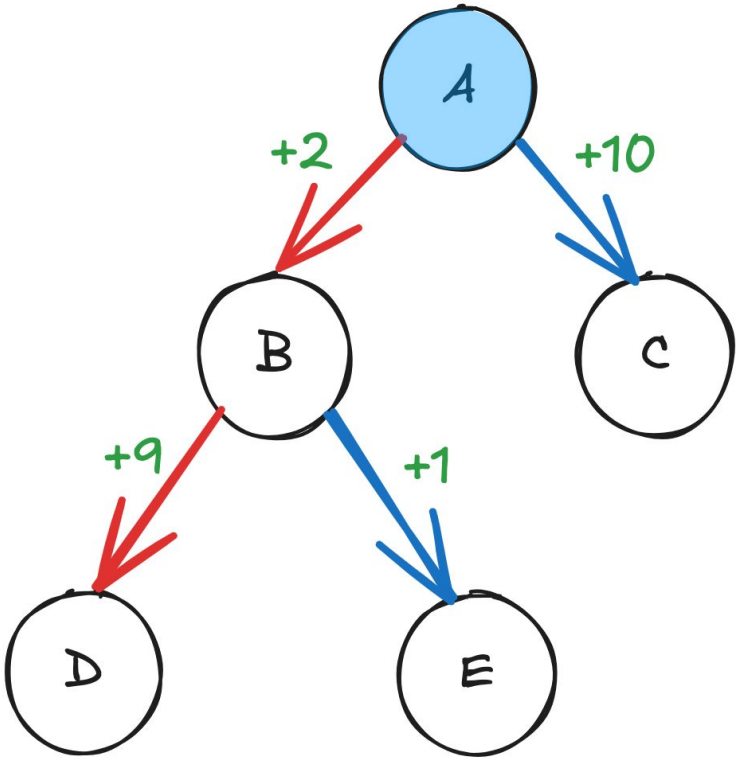
```
P1 = {
    A: red
    B: red
}
```

```
P3 = {
    A: blue
}
```

```
P2 = {
    A: red
    B: blue
}
```

```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
```

Какую политику назовем оптимальной?



$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

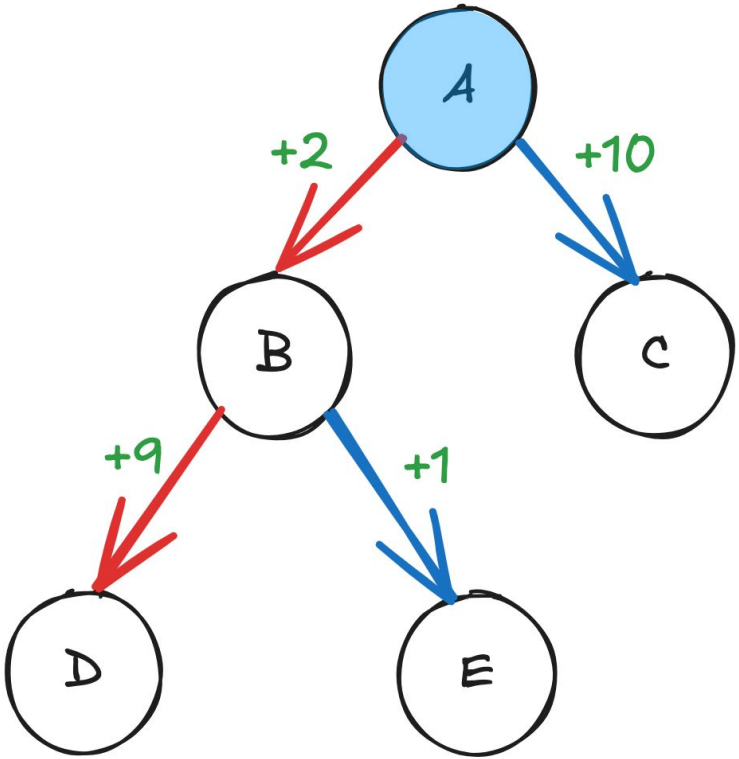
```
P1 = {
    A: red
    B: red
}
```

```
P3 = {
    A: blue
}
```

```
P2 = {
    A: red
    B: blue
}
```

```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
```

Какую политику назовем оптимальной?



$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

```
P1 = {
    A: red
    B: red
}
```

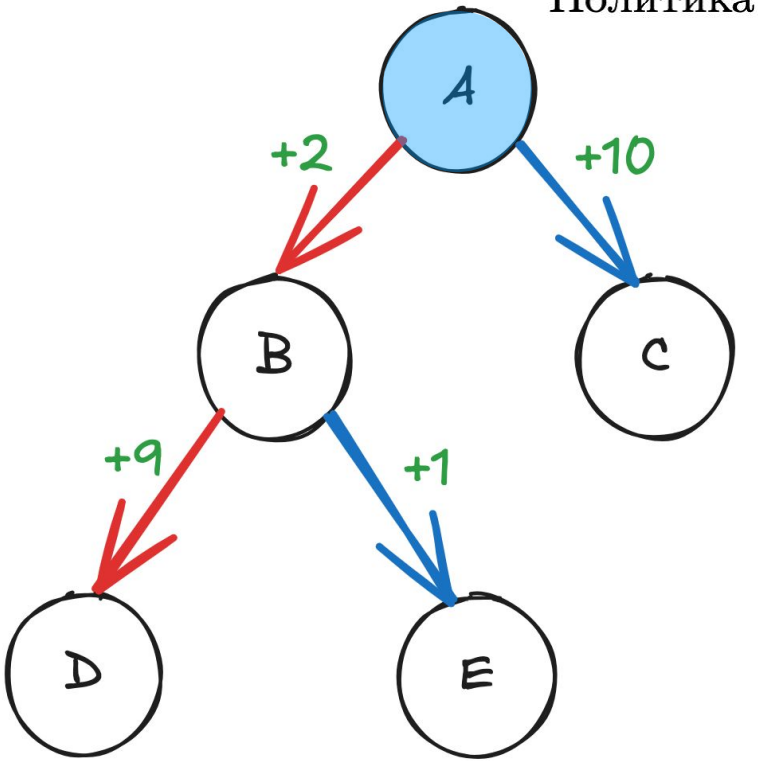
```
P3 = {
    A: blue
}
```

```
P2 = {
    A: red
    B: blue
}
```

```
P4 = {
    A: {
        blue: 0.2
        red: 0.8
    }
    B: {
        blue: 0.5
        red: 0.5
    }
}
```

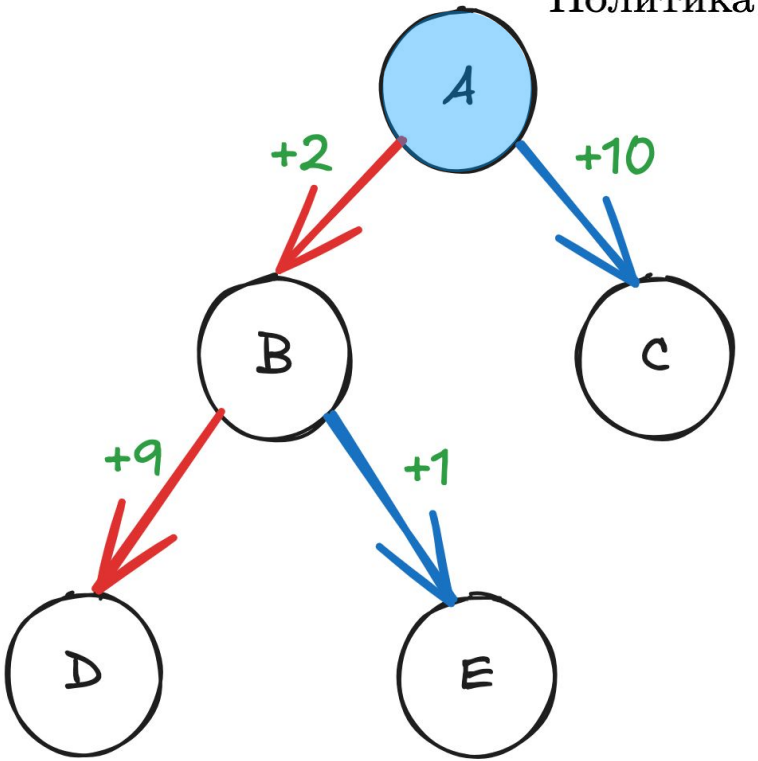
Оптимальная политика

Политика π^* ***оптимальна***, если $\forall \pi : V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$



P1 = {
A: red
B: red
}

Value-функция



Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$

```
P1 = {
      A: red
      B: red
}
```

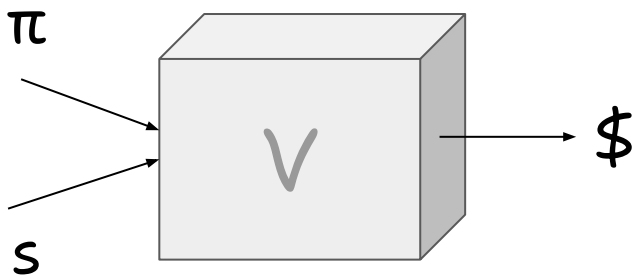
Сколько мы можем заработать, стартуя из данной вершины

$$V^{\pi}(s) := \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s} R(\mathcal{T})$$

Value-функция
(оценочна функция)

Value-функция

Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$




```
P1 = {
      A: red
      B: red
}
```

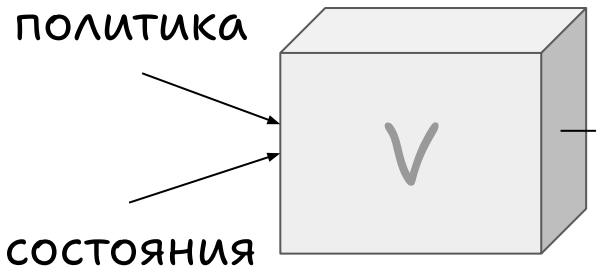
Сколько мы можем заработать, стартуя из данной вершины

$$V^{\pi}(s) := \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s} R(\mathcal{T})$$

Value-функция
(оценочна функция)

Value-функция

Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$



суммарная награда

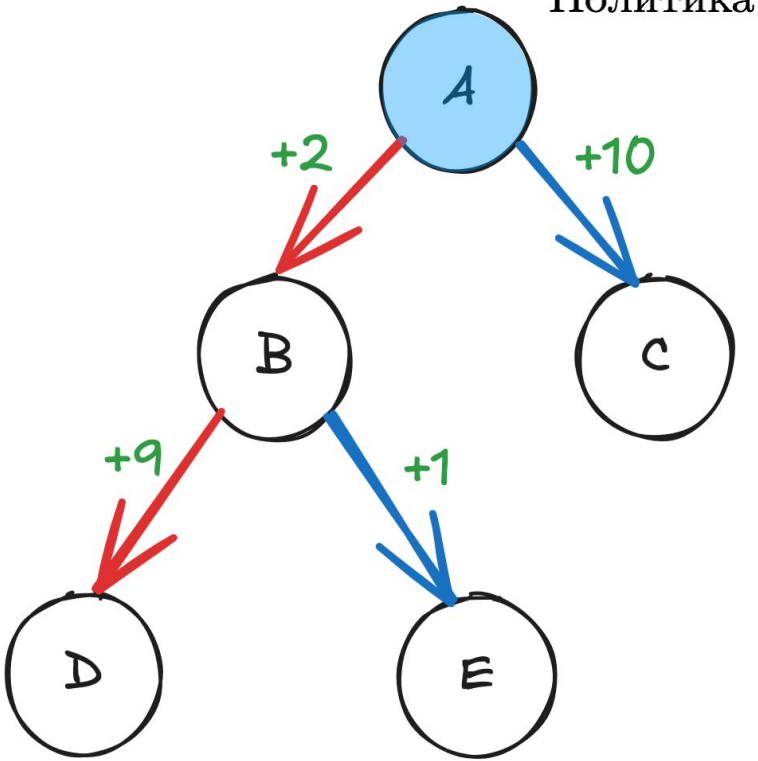
```
P1 = {
      A: red
      B: red
}
```

Сколько мы можем заработать, стартуя из данной вершины

$$V^{\pi}(s) := \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s} R(\mathcal{T})$$

Value-функция
(оценочна функция)

Value-функция



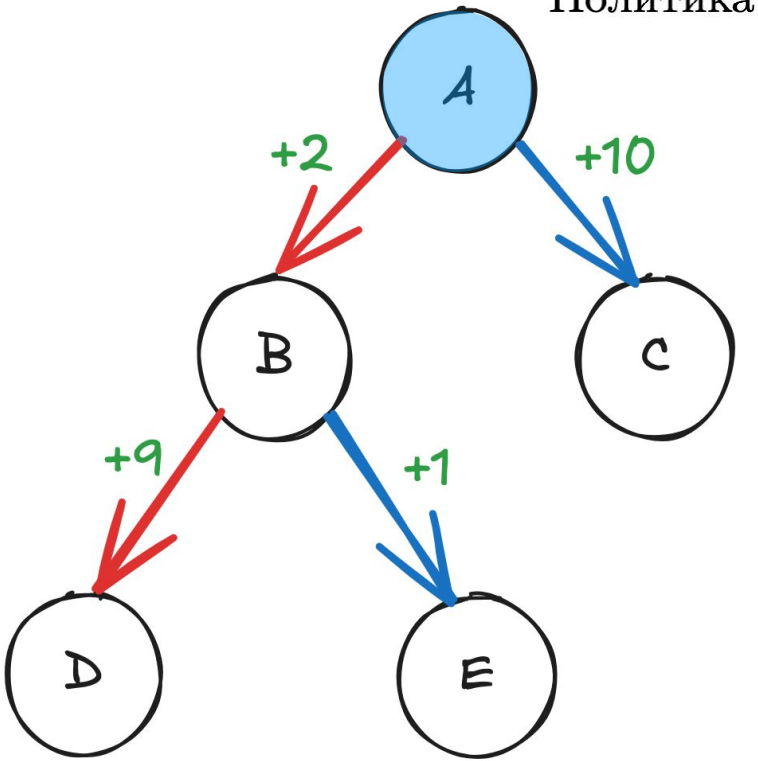
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$

```
P1 = {
      A: red
      B: red
}
```

Как докажем?

Оптимальная value-функция

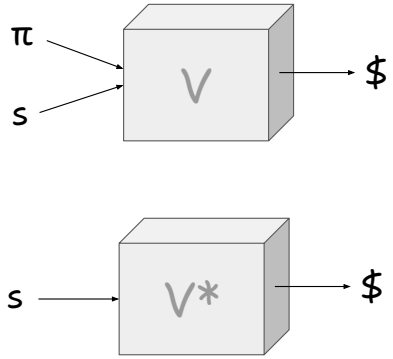
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$



P1 = {
A: red
B: red
}

Как докажем?

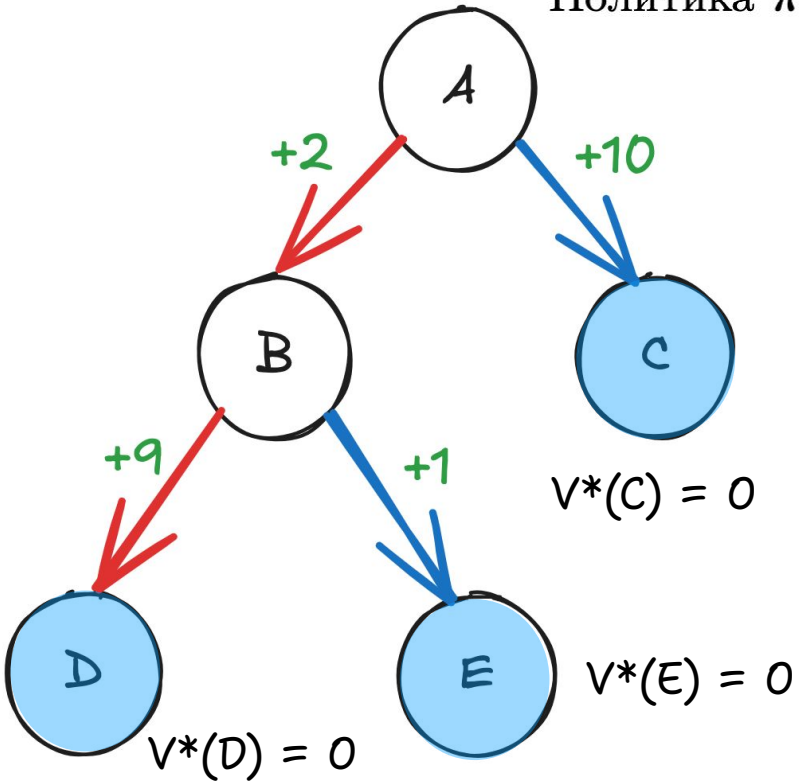
$V^* = V^{\pi^*}$



Оптимальная value-функция

A	B	C	D	E
		0	0	0

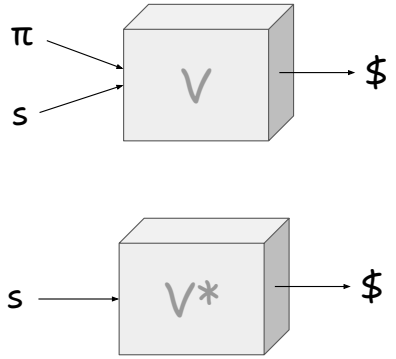
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$



P1 = {
A: red
B: red
}

Как докажем?

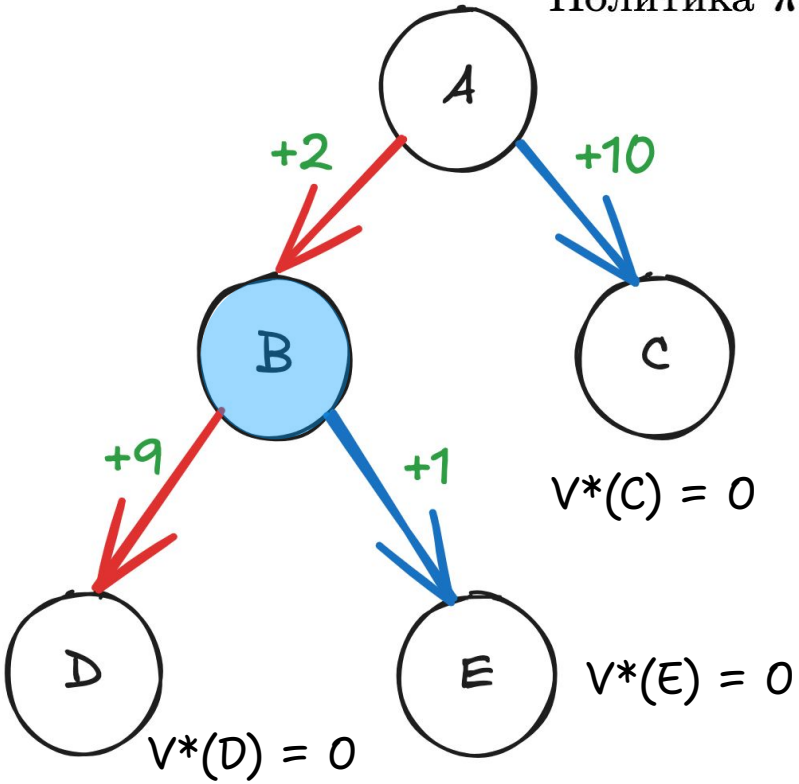
$V^* = V^{\pi^*}$



Оптимальная value-функция

A	B	C	D	E
		0	0	0

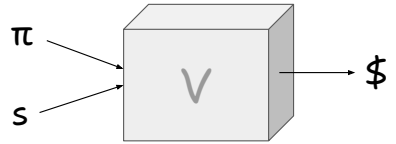
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$

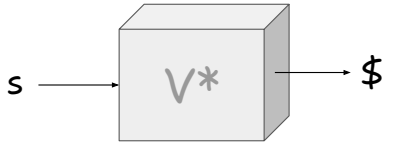


P1 = {
A: red
B: red
}

Как докажем?

$V^* = V^{\pi^*}$



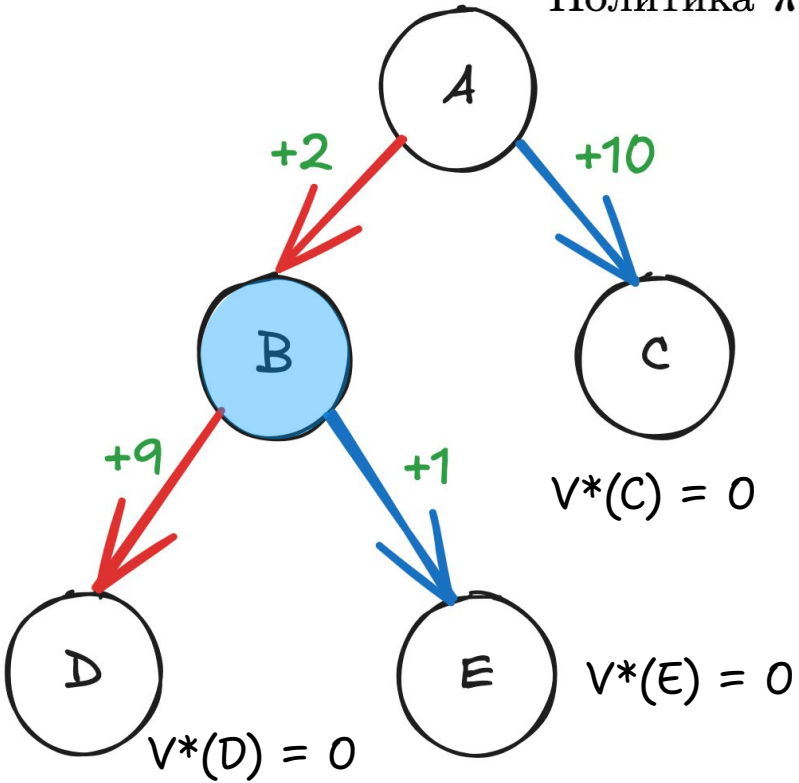


V*(B) =

Оптимальная value-функция

A	B	C	D	E
		0	0	0

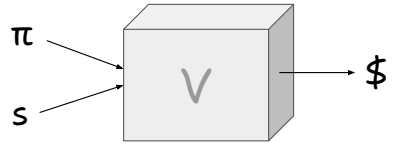
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$

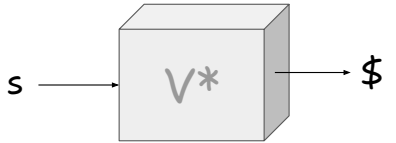


P1 = {
A: red
B: red
}

Как докажем?

$V^* = V^{\pi^*}$



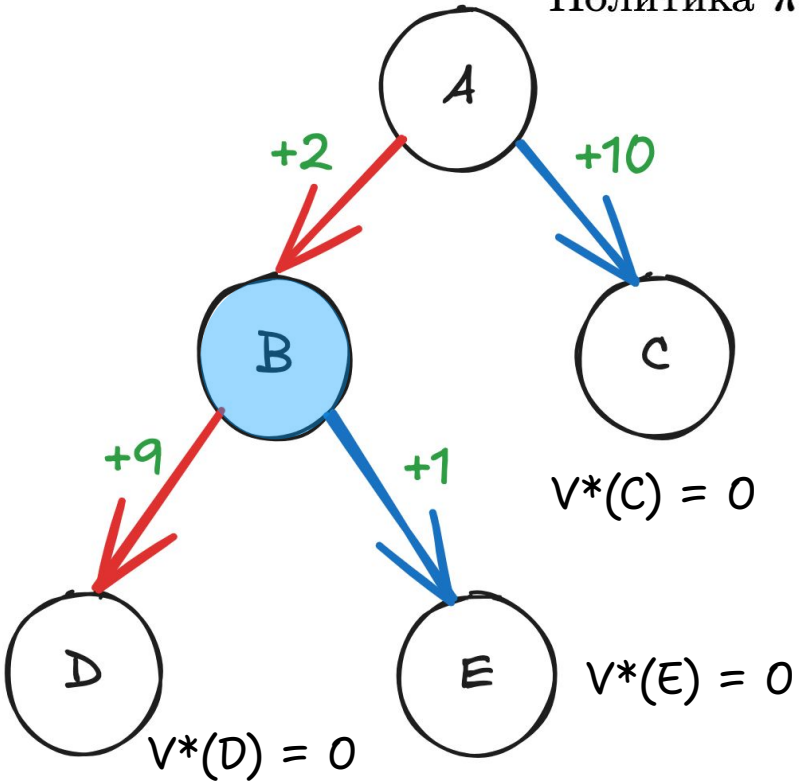


$V^*(B) = \max(9 + V^*(D), 1 + V^*(E))$

Оптимальная value-функция

A	B	C	D	E
	9	0	0	0

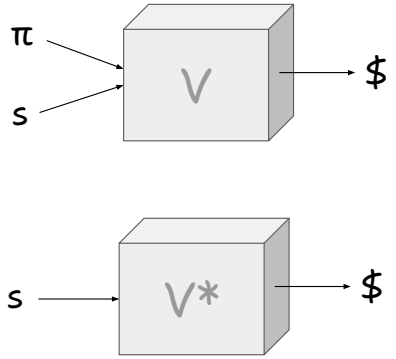
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$



P1 = {
A: red
B: red
}

Как докажем?

$V^* = V^{\pi^*}$

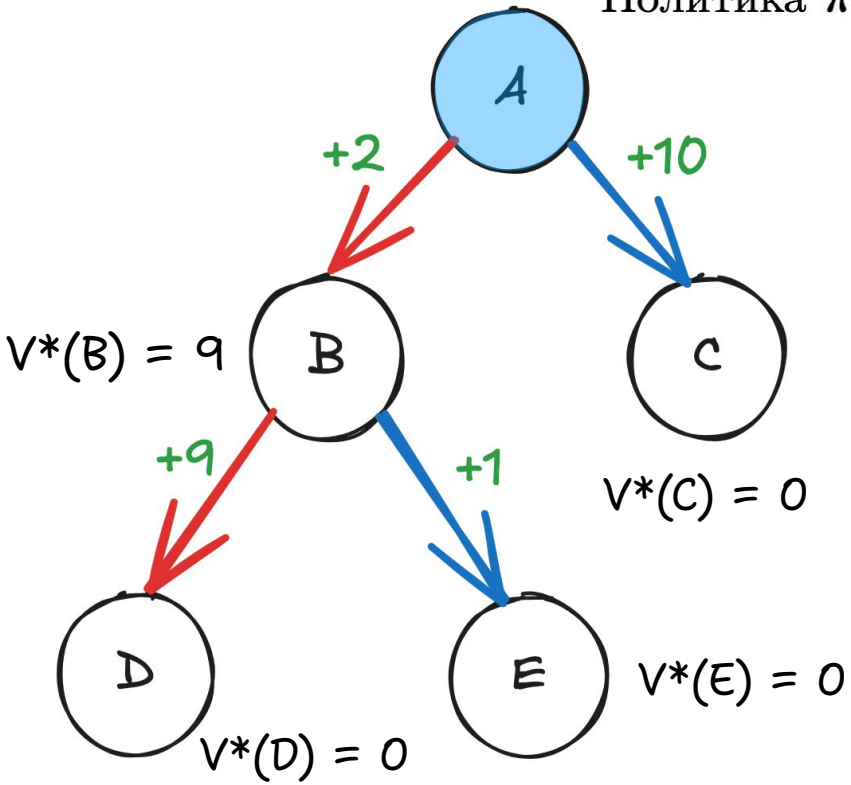


$V^*(B) = \max(9 + 0, 1 + 0)$

Оптимальная value-функция

A	B	C	D	E
	9	0	0	0

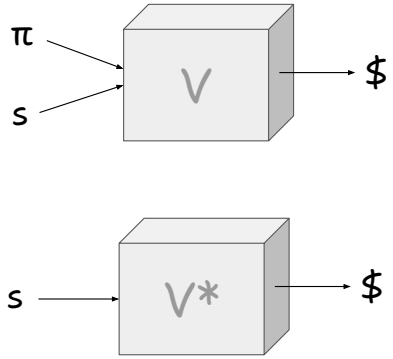
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$



P1 = {
A: red
B: red
}

Как докажем?

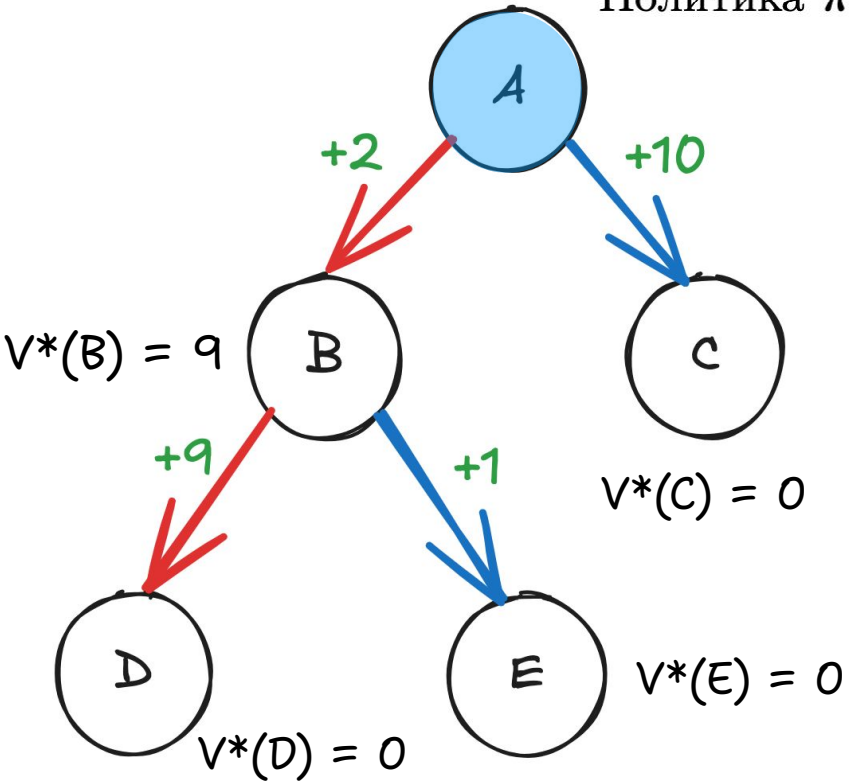
$V^* = V^{\pi^*}$



Оптимальная value-функция

A	B	C	D	E
	9	0	0	0

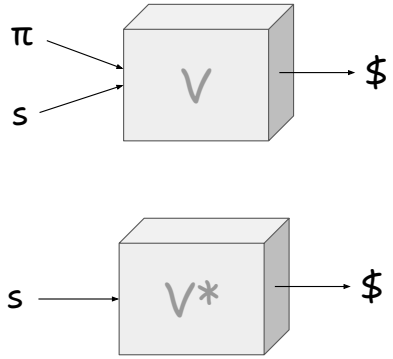
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$



P1 = {
A: red
B: red
}

Как докажем?

$V^* = V^{\pi^*}$

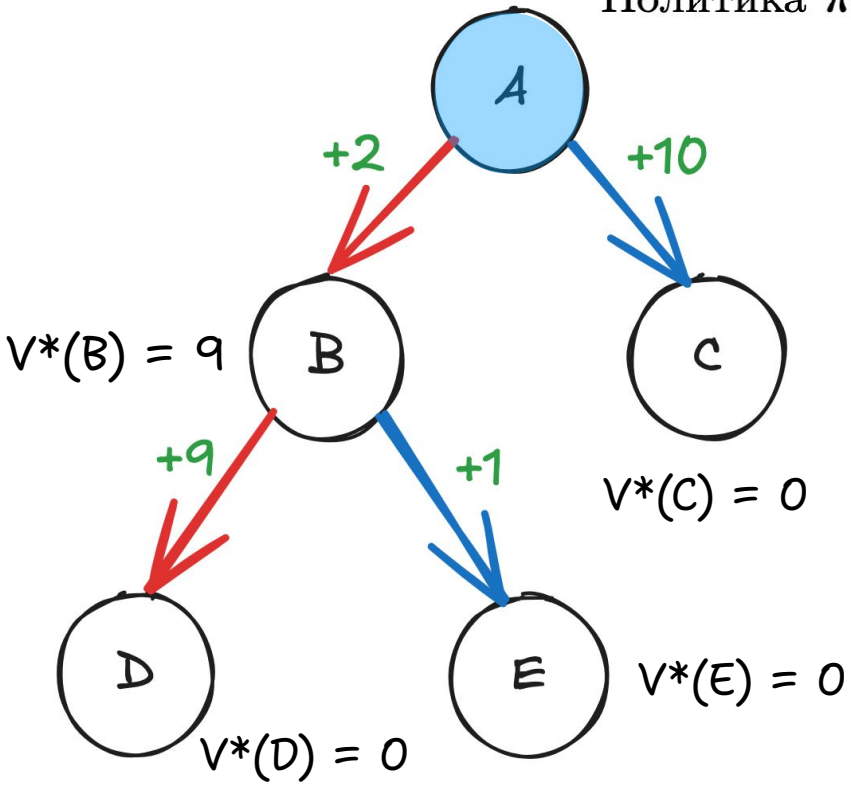


$V^*(A) =$

Оптимальная value-функция

A	B	C	D	E
	9	0	0	0

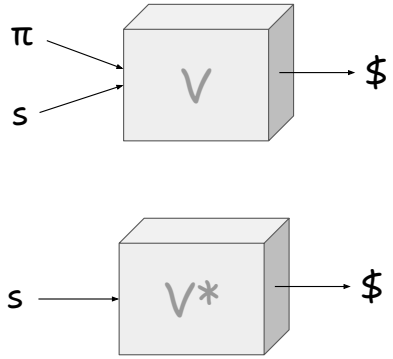
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$



P1 = {
A: red
B: red
}

Как докажем?

$V^* = V^{\pi^*}$

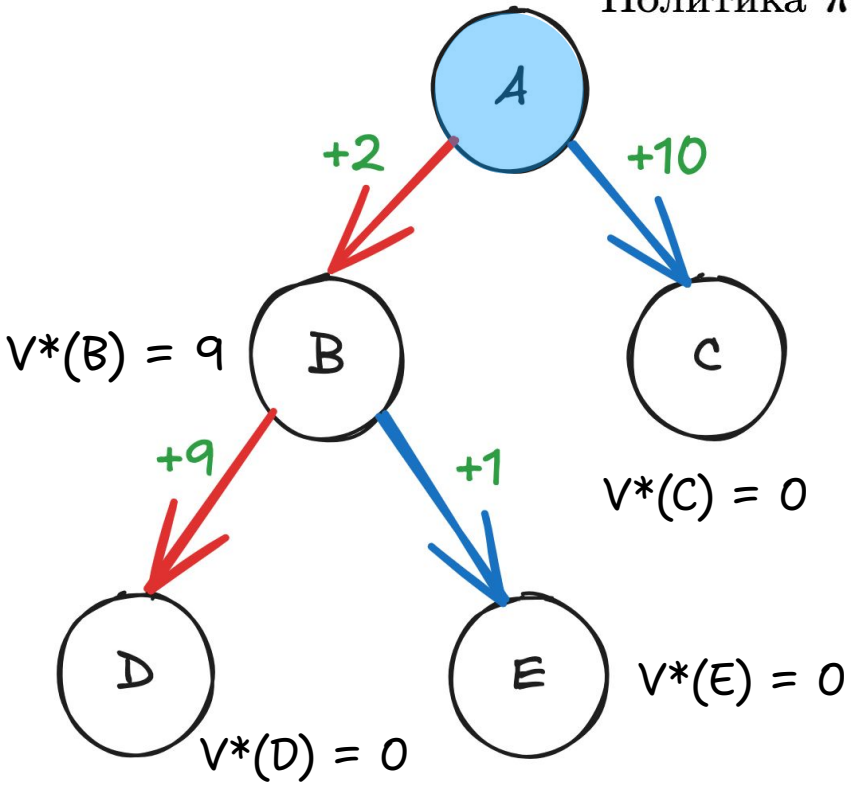


$V^*(A) = \max(2 + V^*(B), 10 + V^*(C))$

Оптимальная value-функция

A	B	C	D	E
	9	0	0	0

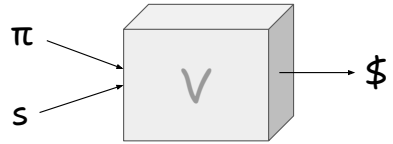
Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$

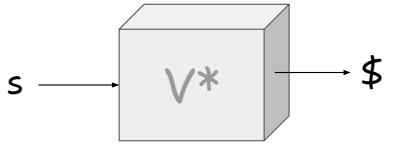


P1 = {
A: red
B: red
}

Как докажем?

$V^* = V^{\pi^*}$

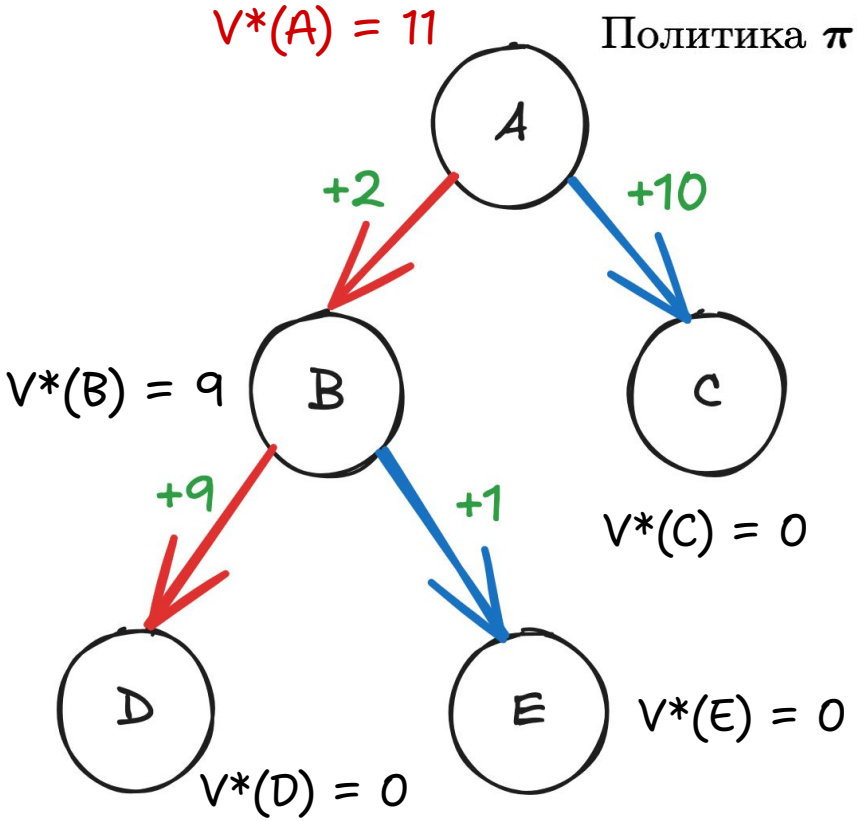




$V^*(A) = \max(2 + 9, 10 + 0)$

Оптимальная value-функция

A	B	C	D	E
11	9	0	0	0

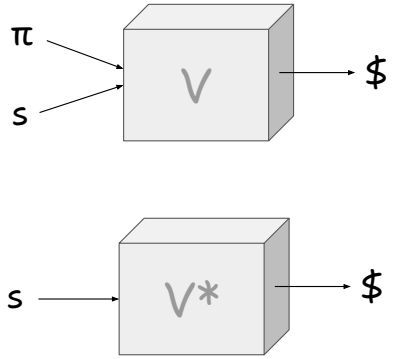


Политика π^* ***оптимальна***, если $\forall \pi: V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$

P1 = {
A: red
B: red
}

$V^* = V^{\pi^*}$

Как докажем?

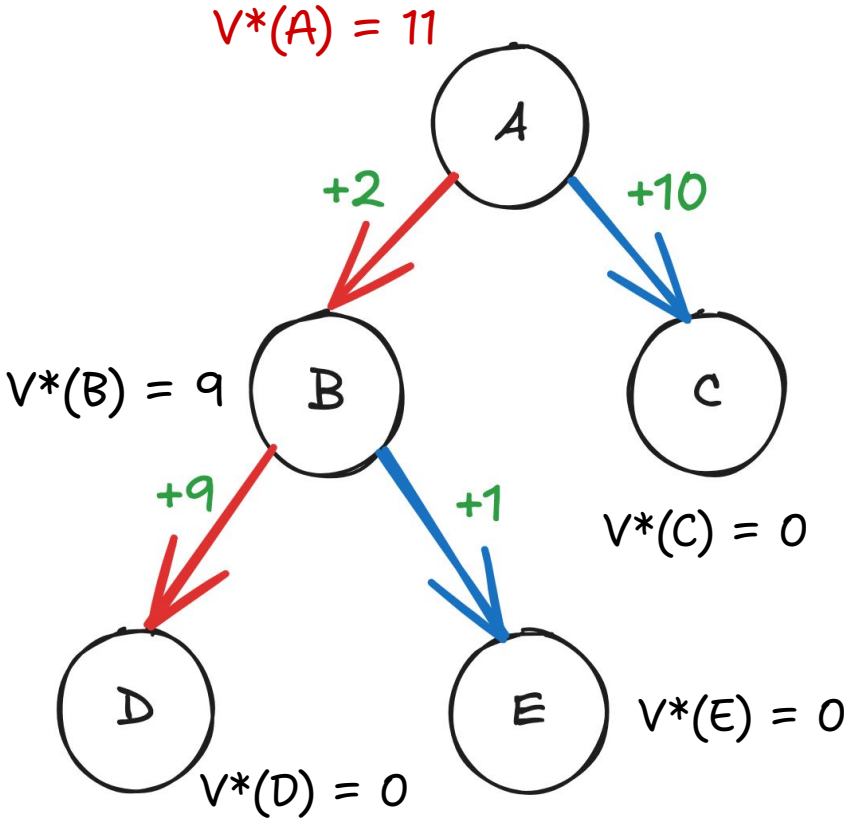


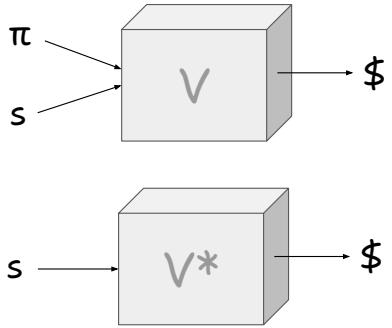
А как найти саму политику?

Метод обратного хода?

V*:

A	B	C	D	E
11	9	0	0	0

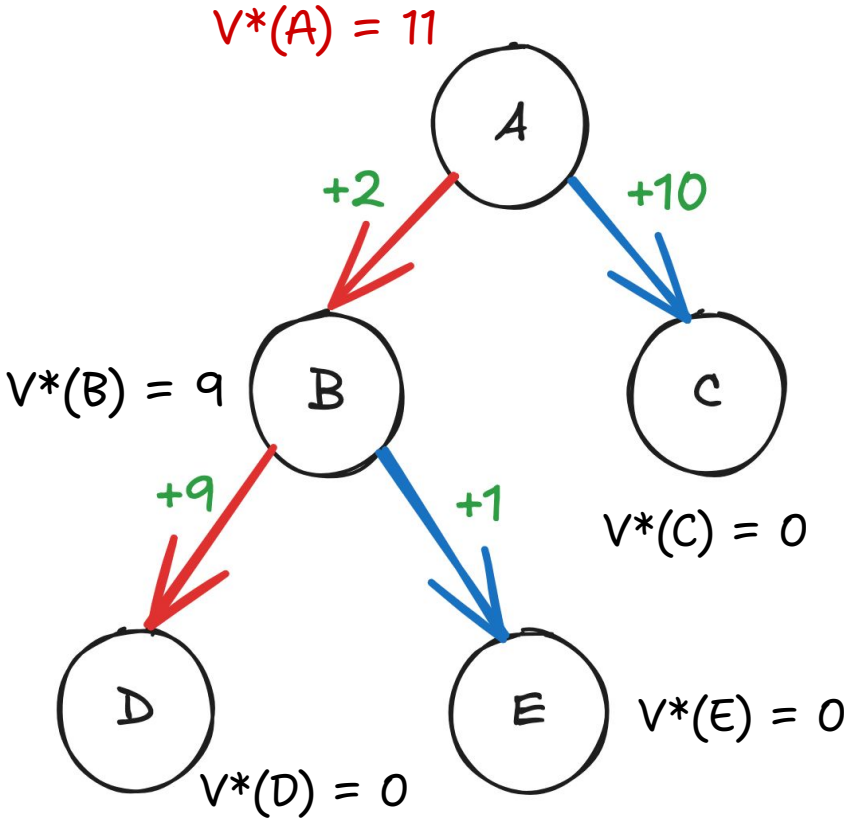


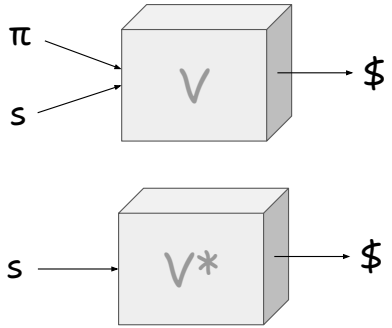


Можем сохранять сразу!

V*:

A	B	C	D	E
11	9	0	0	0

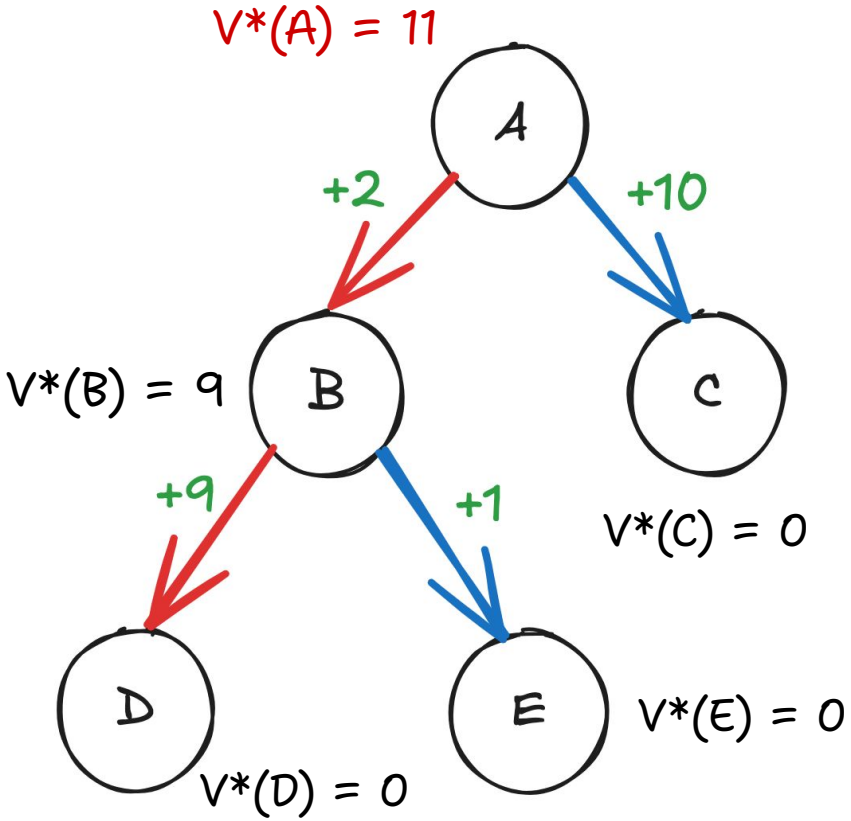


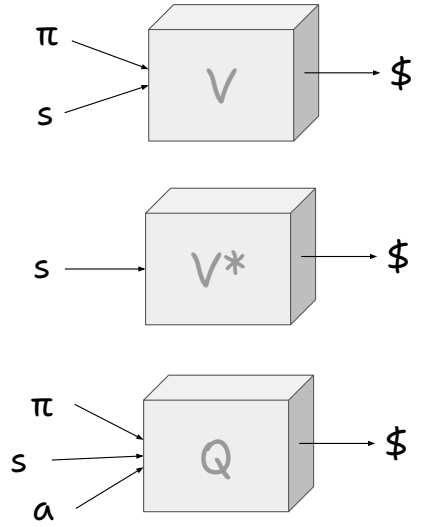


Quality function

V*:

A	B	C	D	E
11	9	0	0	0





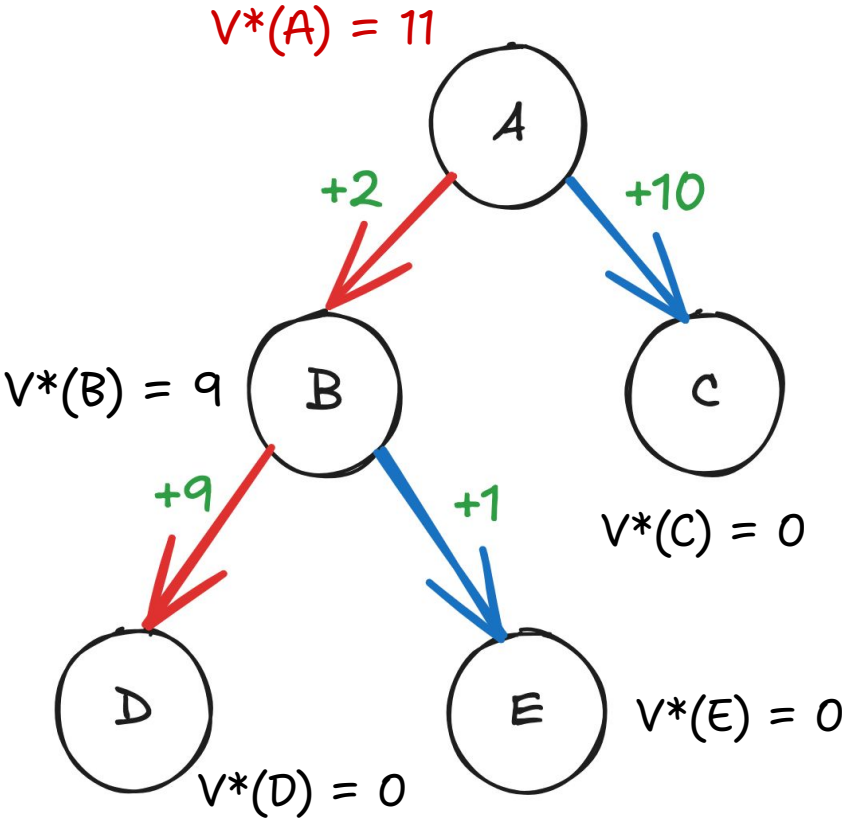
$$Q^{\pi}(s,a) := \mathbb{E}_{\mathcal{T}\sim\pi|s_0=s,a_0=a} \sum_{t\geq 0} r_t$$

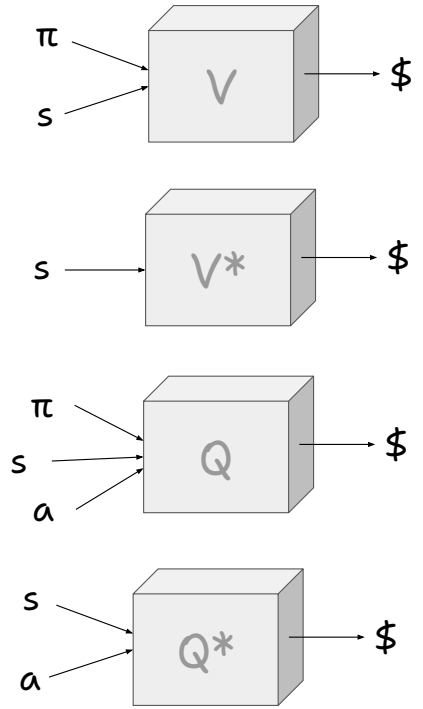
Выполняем **a**, а потом по политике **π**

Quality function

V*:

A	B	C	D	E
11	9	0	0	0



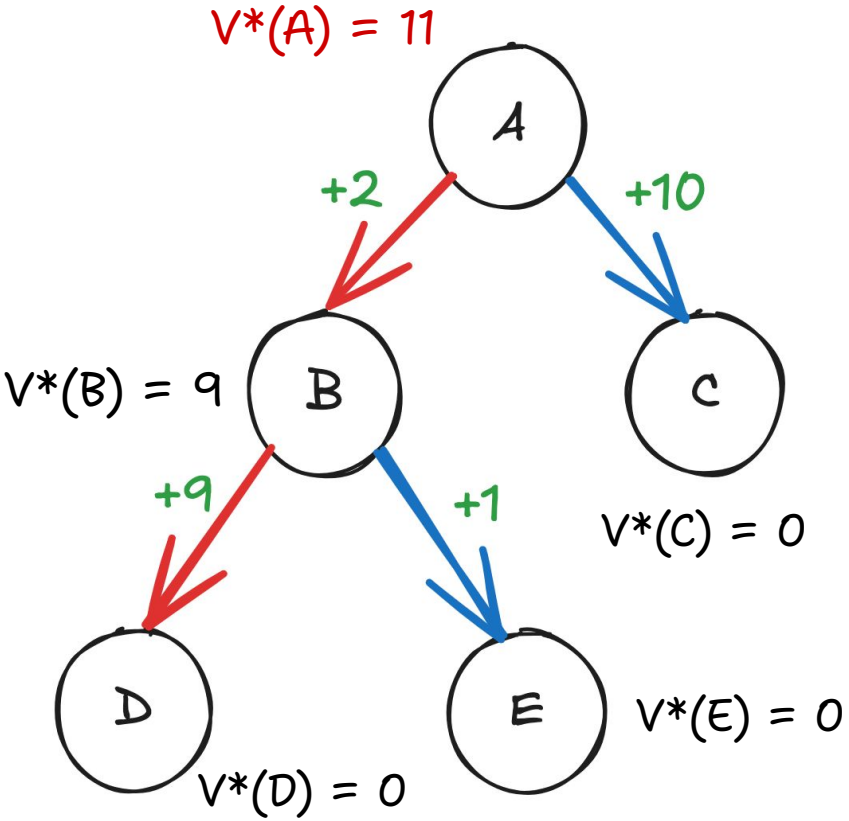


Выполняем **a**, а потом по политике **π**

Quality function

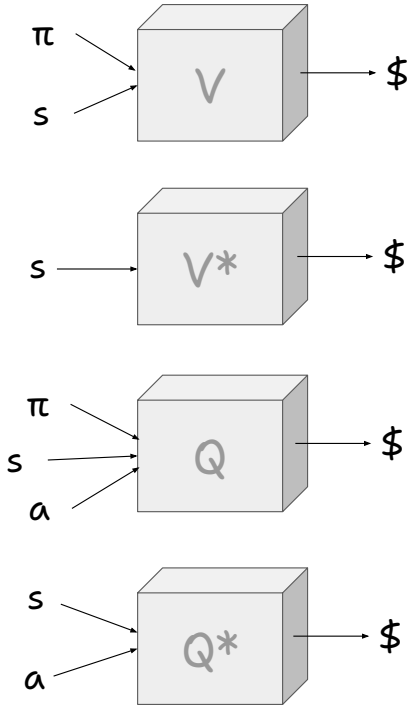
V*:

A	B	C	D	E
11	9	0	0	0



Q*:

A		
B		
C		
D		
E		

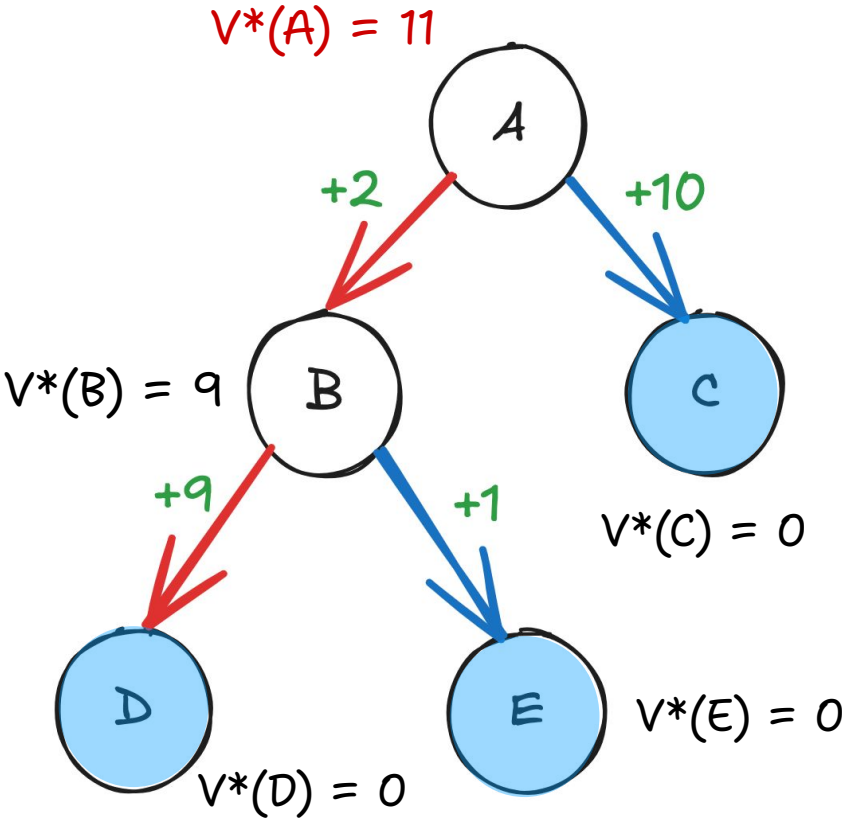


Выполняем **a**, а потом по политике **π**

Quality function

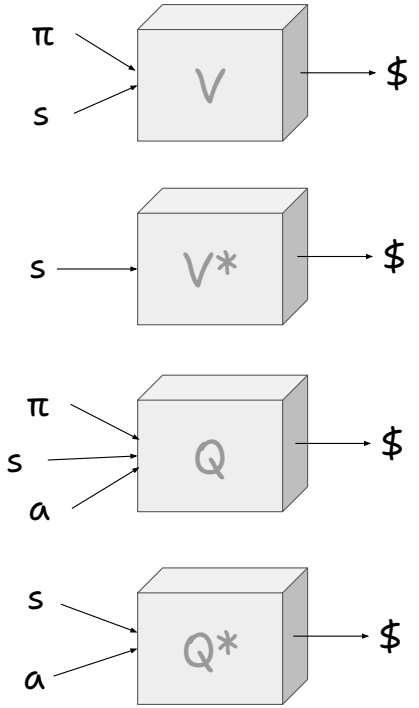
V*:

A	B	C	D	E
		0	0	0



Q*:

A		
B		
C	✗	✗
D	✗	✗
E	✗	✗

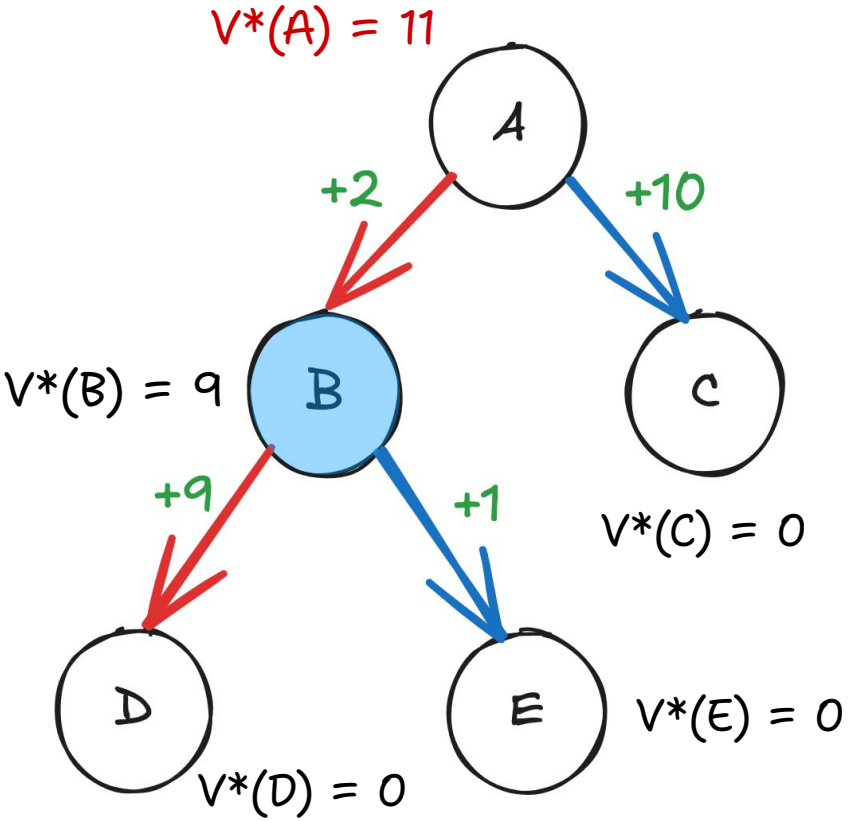


Выполняем **a**, а потом по политике **π**

Quality function

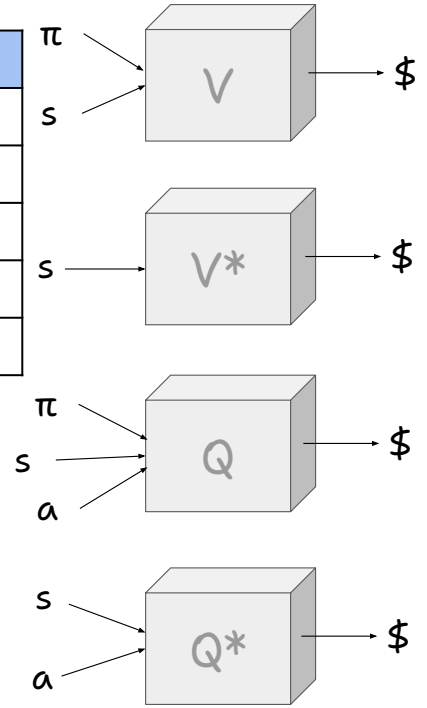
V*:

A	B	C	D	E
		0	0	0



Q*:

A		
B	9+V(D)	1+V(E)
C	✗	✗
D	✗	✗
E	✗	✗

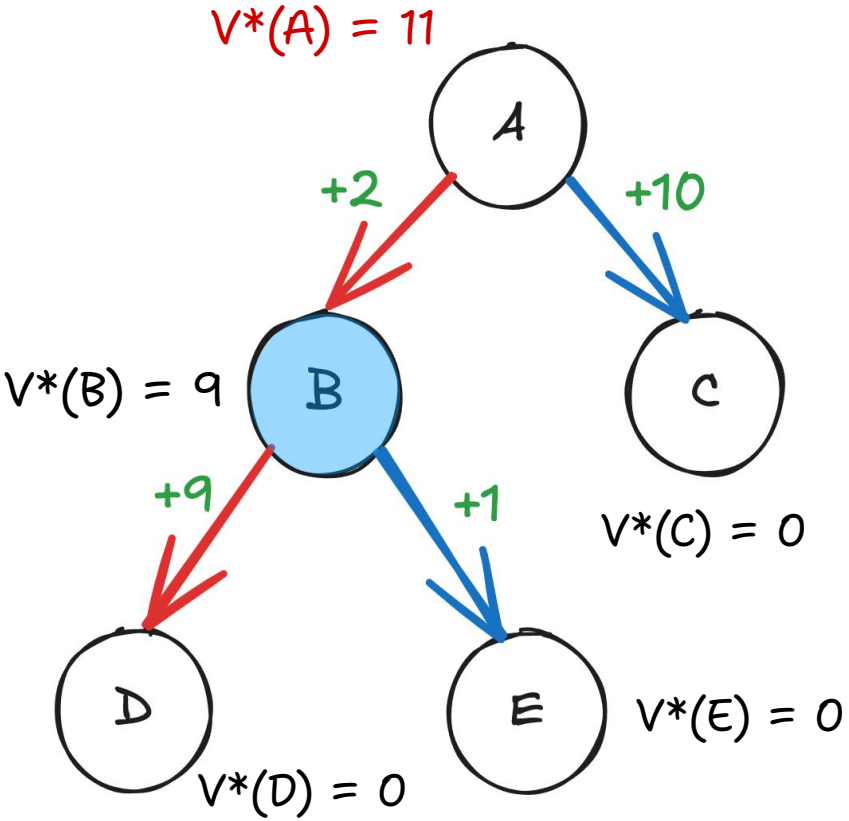


Выполняем **a**, а потом по политике **π**

Quality function

V*:

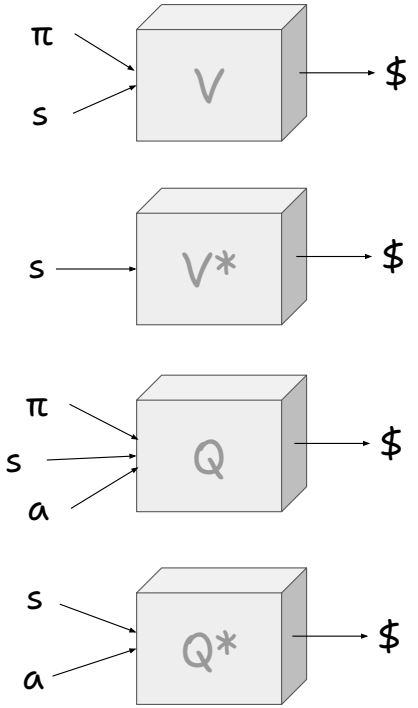
A	B	C	D	E
		0	0	0



Q*:

A		
B	9	1
C	✗	✗
D	✗	✗
E	✗	✗

$$V^*(s) = \max_a Q^*(s, a)$$

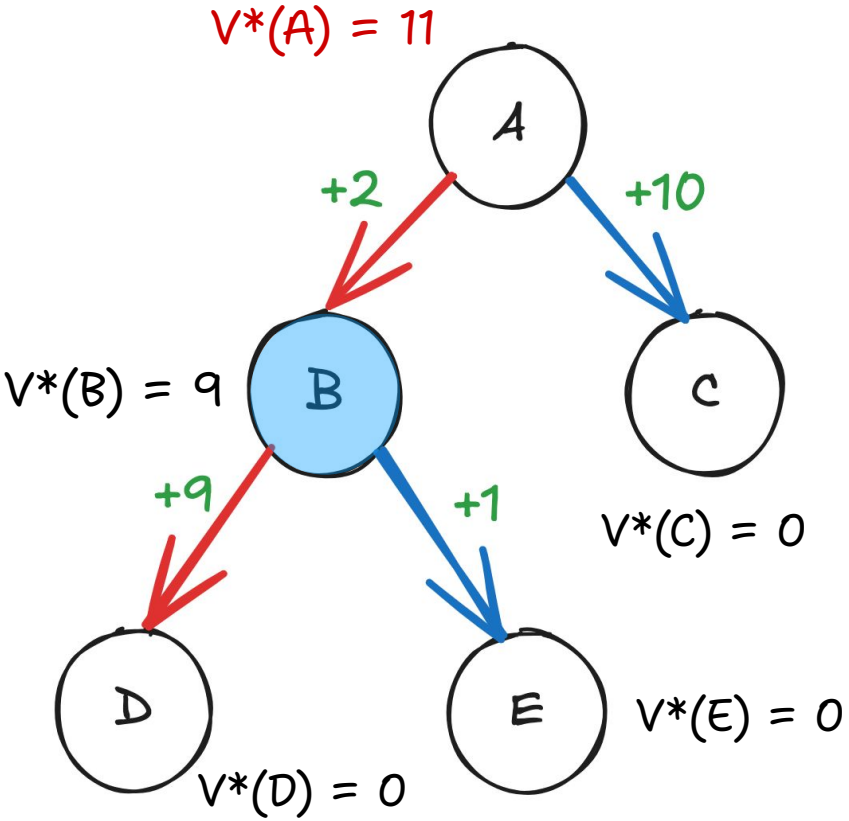


Выполняем **a**, а потом по политике **π**

Quality function

V*:

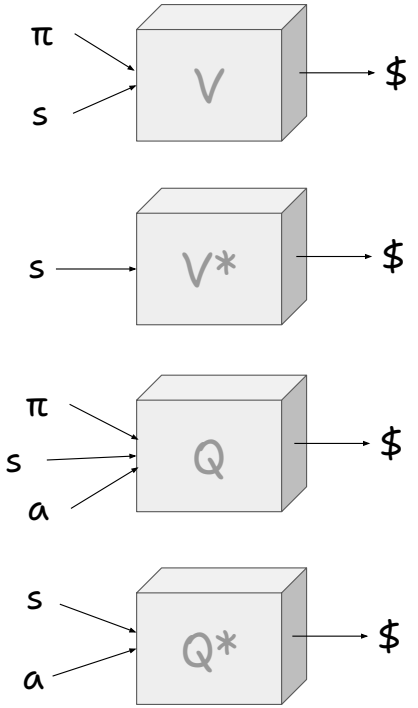
A	B	C	D	E
	9	0	0	0



Q*:

A		
B	9	1
C	✗	✗
D	✗	✗
E	✗	✗

$$V^*(s) = \max_a Q^*(s, a)$$

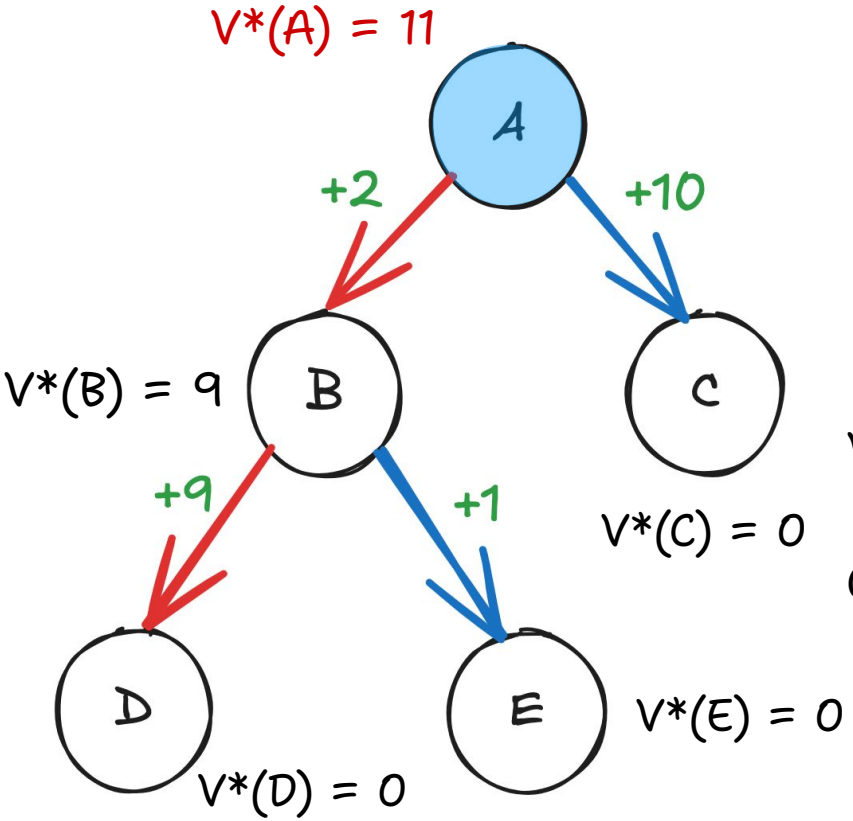


Выполняем **a**, а потом по политике **π**

Quality function

V*:

A	B	C	D	E
	9	0	0	0

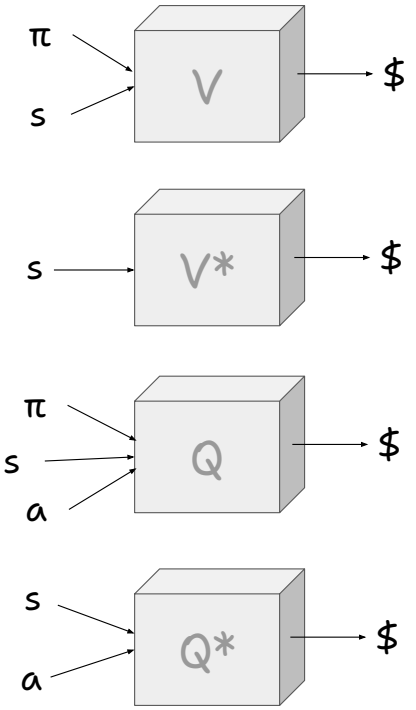


Q*:

A		
B	9	1
C	✗	✗
D	✗	✗
E	✗	✗

$$V^*(s) = \max_a Q^*(s, a)$$

$$Q^*(s, a) = r(s,a) + V^*(s, a)$$

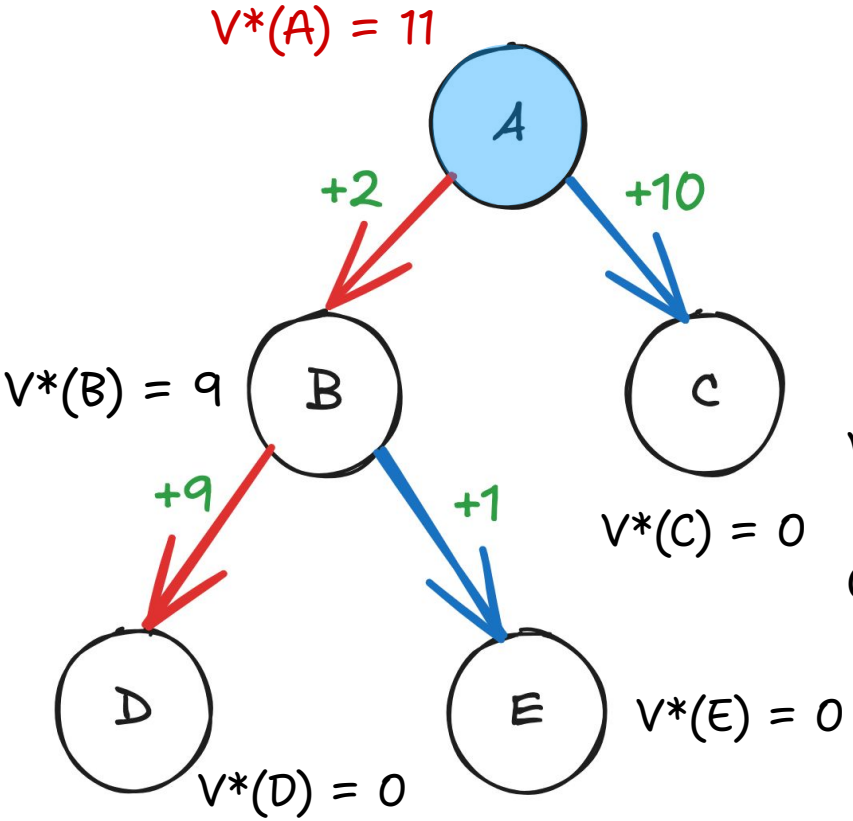


Выполняем **a**, а потом по политике **π**

Quality function

V*:

A	B	C	D	E
	9	0	0	0

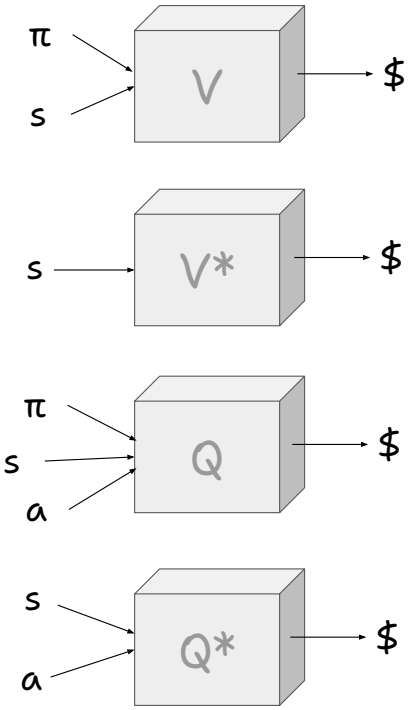


Q*:

A	11	10
B	9	1
C	✗	✗
D	✗	✗
E	✗	✗

$$V^*(s) = \max_a Q^*(s, a)$$

$$Q^*(s, a) = r(s,a) + V^*(s, a)$$

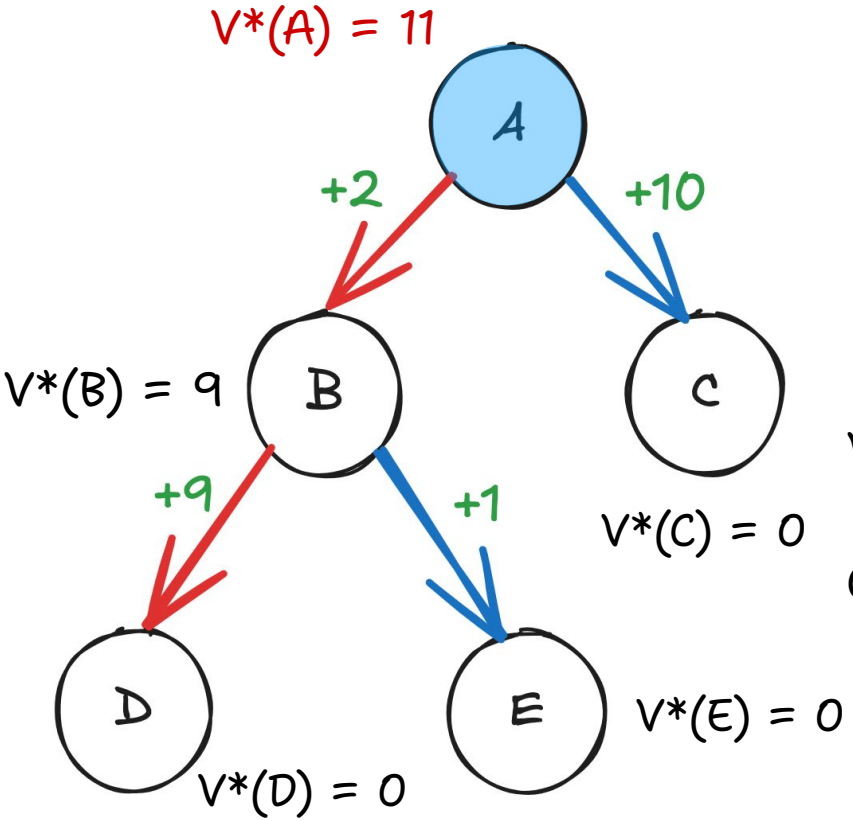


Выполняем **a**, а потом по политике **π**

Quality function

V*:

A	B	C	D	E
11	9	0	0	0

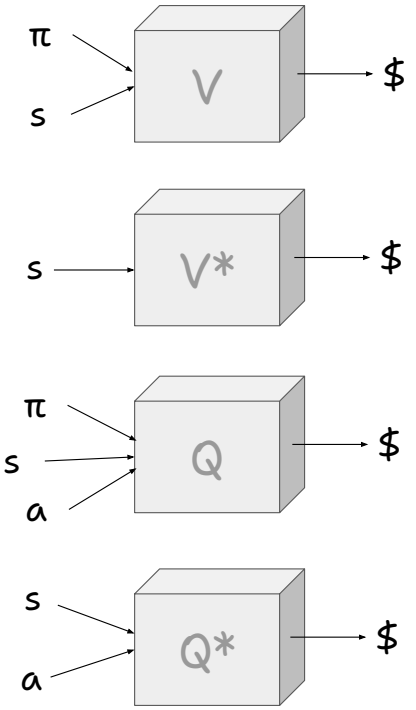


Q*:

A	11	10
B	9	1
C	✗	✗
D	✗	✗
E	✗	✗

$$V^*(s) = \max_a Q^*(s, a)$$

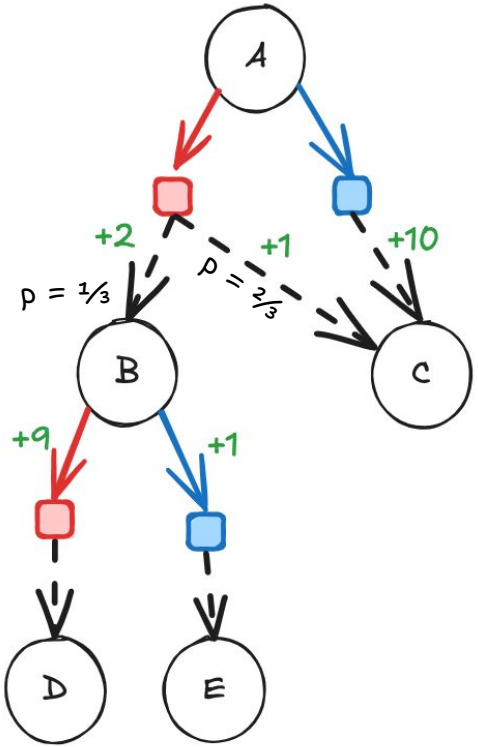
$$Q^*(s, a) = r(s,a) + V^*(s, a)$$



Выполняем **a**, а потом по политике **π**

Стохастический случай

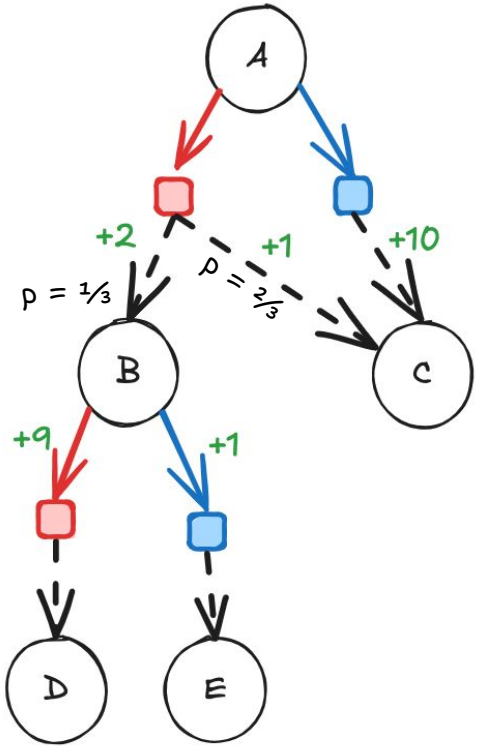
Стохастический случай



$$V^*(s) = \max_a Q^*(s, a)$$

$$Q^*(s, a) = r(s, a) + V^*(s, a)$$

Стохастический случай



$$V^*(s) = \max_a Q^*(s, a)$$

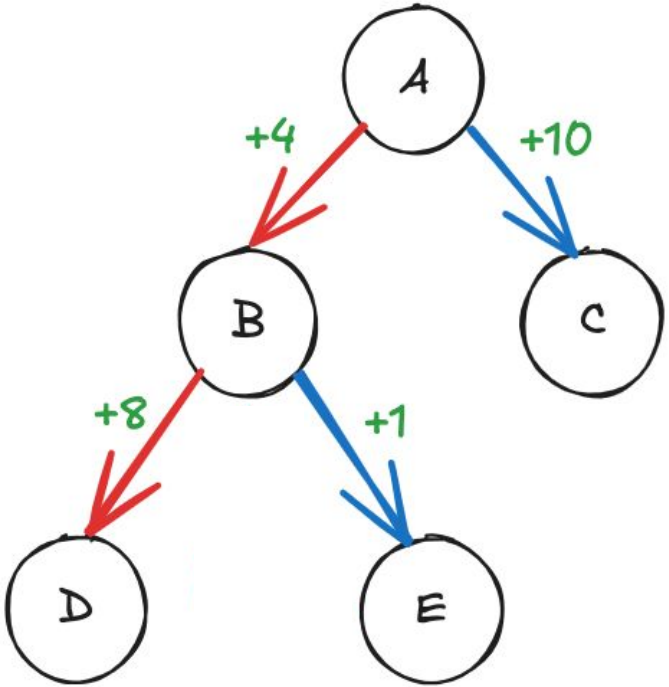
$$Q^*(s, a) = \mathbf{E}(r(s, a) + V^*(s, a))$$

Более реалистично...

Более реалистично?

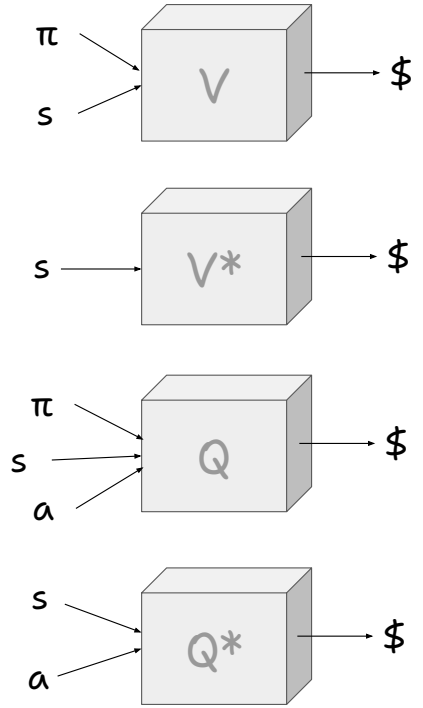
V*:

A	B	C	D	E



Q*:

A		
B		
C	✗	✗
D	✗	✗
E	✗	✗



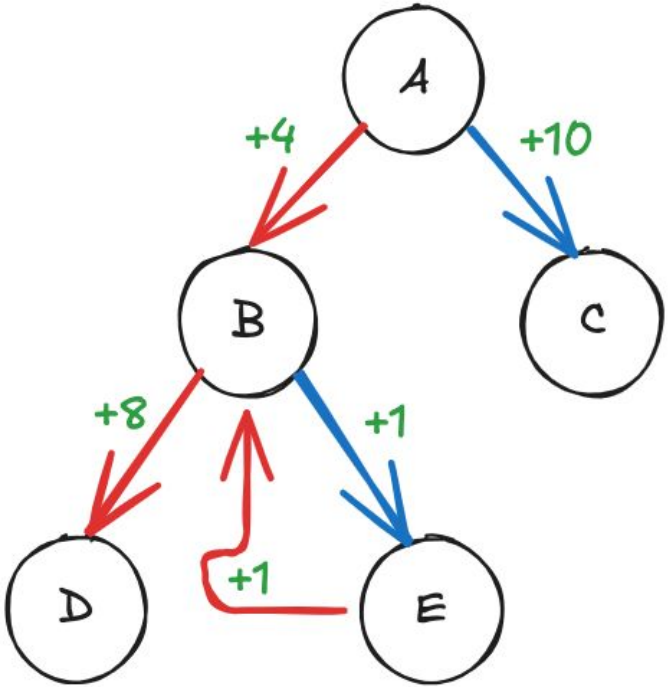
Выполняем **a**, а потом по политике **π**

Перерыв до 20:57

Циклы

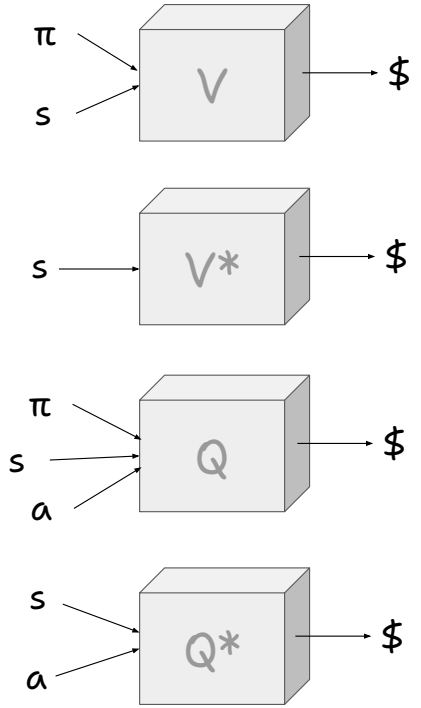
V*:

A	B	C	D	E



Q*:

A		
B		
C	✗	✗
D	✗	✗
E		✗

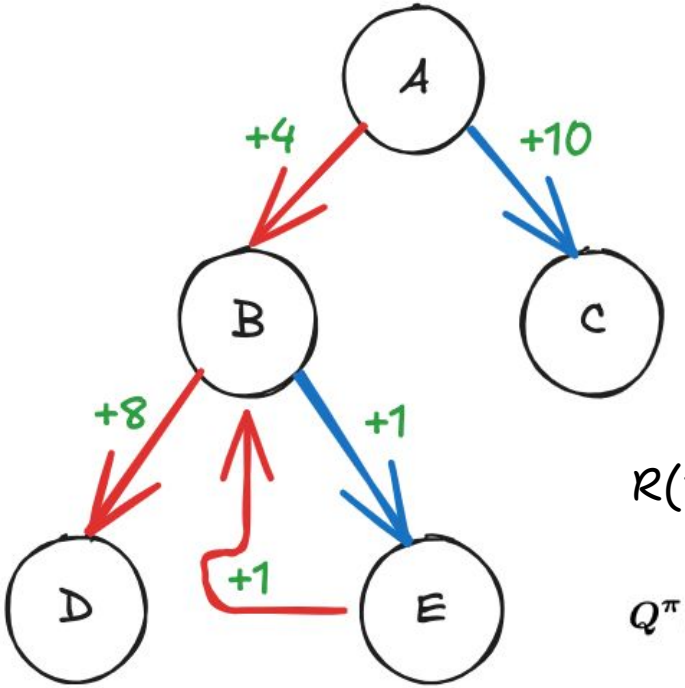


Выполняем **a**, а потом по политике **π**

В чем проблема?

V*:

A	B	C	D	E

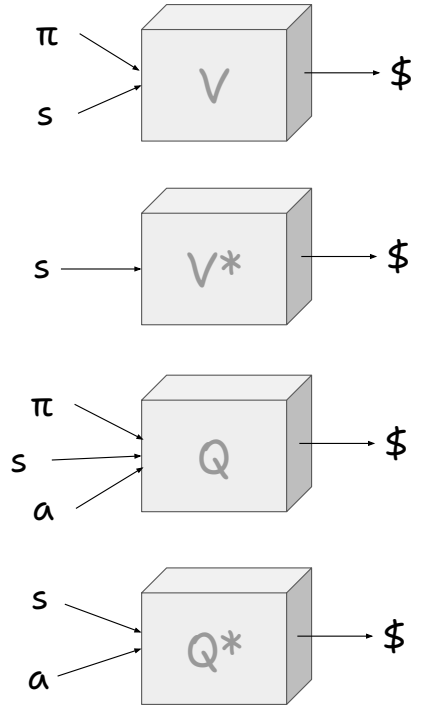


Q*:

A		
B		
C	✗	✗
D	✗	✗
E		✗

$$R(\pi) = \mathbb{E}_{T\sim\pi} \sum_{t\geq 0} r_t \to \max_{\pi},$$

$$Q^{\pi}(s,a) := \mathbb{E}_{T\sim\pi|s_0=s,a_0=a} \sum_{t\geq 0} r_t$$

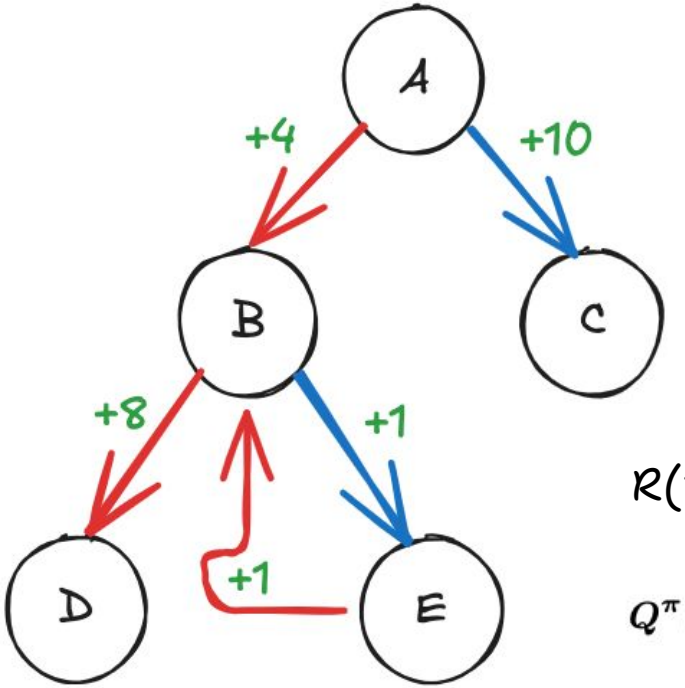


Выполняем **a**, а потом по политике **π**

Коэффициент дисконтирования

V*:

A	B	C	D	E

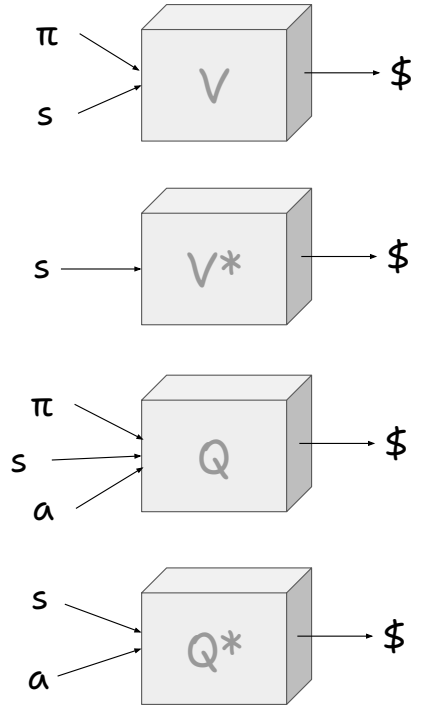


Q*:

A		
B		
C	✗	✗
D	✗	✗
E		✗

$$\mathcal{R}(\pi) = \mathbb{E}_{\mathcal{T}\sim\pi} \sum_{t\geq 0} \gamma^t r_t \to \max_{\pi},$$

$$Q^{\pi}(s,a) := \mathbb{E}_{\mathcal{T}\sim\pi|s_0=s,a_0=a} \sum_{t\geq 0} \gamma^t r_t$$

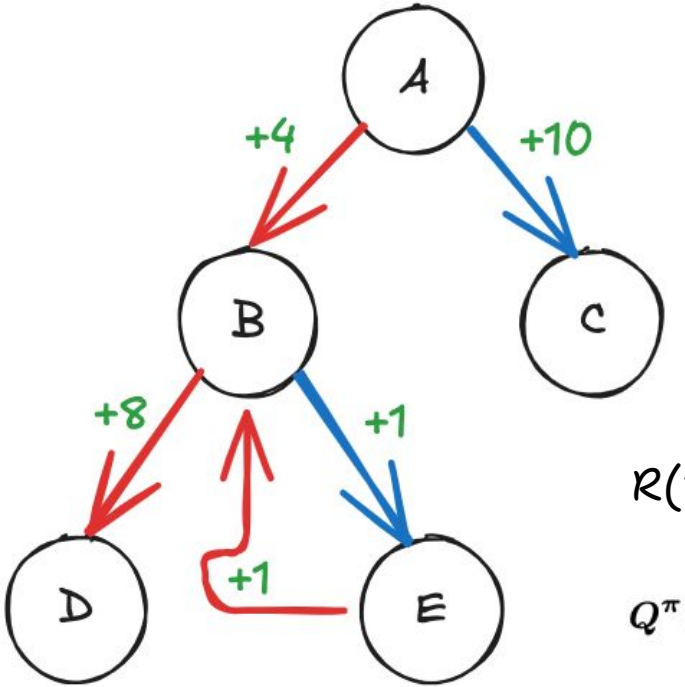


Выполняем **a**, а потом по политике **π**

Коэффициент дисконтирования

V*:

A	B	C	D	E



Q*:

A		
B		
C	✗	✗
D	✗	✗
E		✗

$$R(\pi) = \mathbb{E}_{\tau\sim\pi} \sum_{t\geq 0} \gamma^t r_t \to \max_{\pi},$$

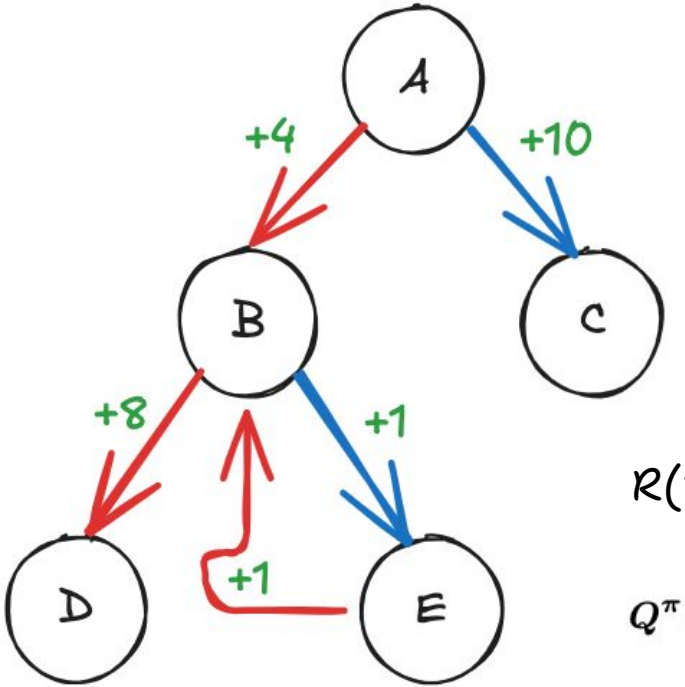
$$Q^{\pi}(s,a) := \mathbb{E}_{\tau\sim\pi|s_0=s,a_0=a} \sum_{t\geq 0} \gamma^t r_t$$

Вероятность того, что эпизод сейчас завершится

Коэффициент дисконтирования

V*:

A	B	C	D	E



Q*:

A		
B		
C	✗	✗
D	✗	✗
E		✗

$$R(\pi) = \mathbb{E}_{\tau\sim\pi} \sum_{t\geq 0} \gamma^t r_t \to \max_{\pi},$$

$$Q^{\pi}(s,a) := \mathbb{E}_{\tau\sim\pi|s_0=s,a_0=a} \sum_{t\geq 0} \gamma^t r_t$$

Вероятность того, что эпизод сейчас завершится

Уравнения Беллмана

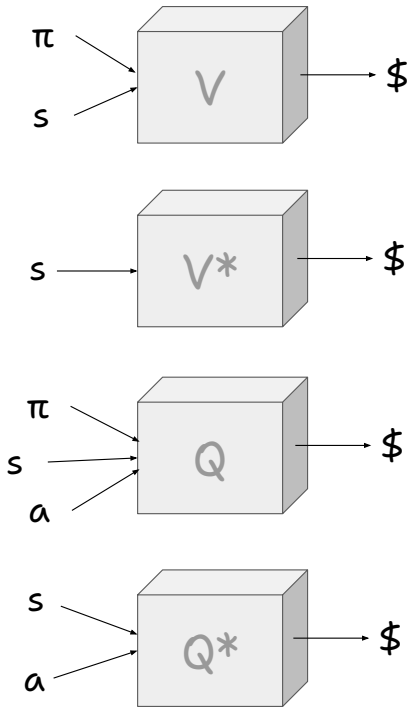
Уравнения Беллмана

$$V^{\pi}(s) = \mathbb{E}_a \left[r(s,a) + \gamma \mathbb{E}_{s'} V^{\pi}(s') \right]$$

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} \mathbb{E}_{a'} Q^{\pi}(s',a')$$

$$Q^{*}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} \max_{a'} Q^{*}(s',a')$$

$$V^{*}(s) = \max_{a} \left[r(s,a) + \gamma \mathbb{E}_{s'} V^{*}(s') \right]$$



V-функция (state value function) ~ сколько, в среднем набирает агент из данного состояния, действую по заданной политике.

$$V^{\pi}(s) := \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s} R(\mathcal{T})$$

Q-функция (**action** state value function) ~ сколько, в среднем набирает агент из данного состояния,

- Сначала выбирая действие **a**
- А потом действую по заданной политике **π**

$$Q^{\pi}(s, a) := \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s, a_0 = a} \sum_{t \geq 0} \gamma^t r_t$$

V-функция (state value function) ~ сколько, в среднем набирает агент из данного состояния, действую по заданной политике.

$$V^{\pi}(s) := \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s} R(\mathcal{T})$$

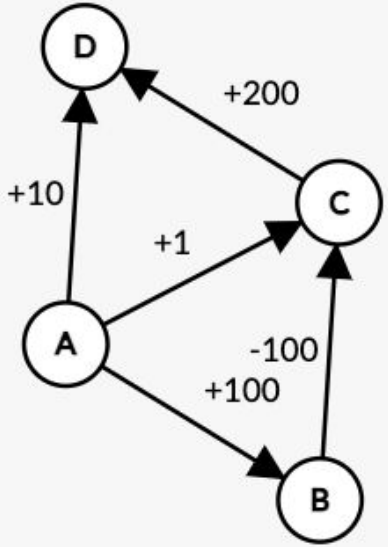
Q-функция (**action** state value function) ~ сколько, в среднем набирает агент из данного состояния,

- Сначала выбирая действие **a**
- А потом действую по заданной политике **π**

$$Q^{\pi}(s, a) = r(s, a) + \gamma \mathbb{E}_{s'} V^{\pi}(s')$$

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:				
Шаг 2. V:				
Шаг 3. V:				
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

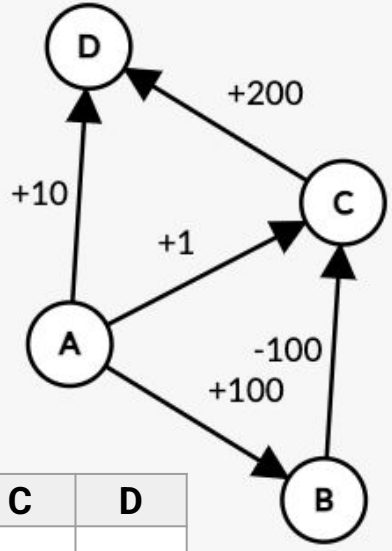
Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. для всех s: $V_{k+1}(s) \coloneqq \max\limits_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max\limits_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}\limits_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:				
Шаг 2. V:				
Шаг 3. V:				
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

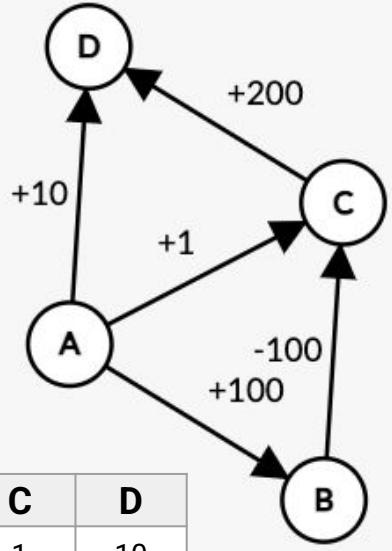
1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

	A	B	C	D
A	❌			
B	❌	❌		❌
C	❌	❌	❌	
D	❌	❌	❌	❌

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:				
Шаг 2. V:				
Шаг 3. V:				
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

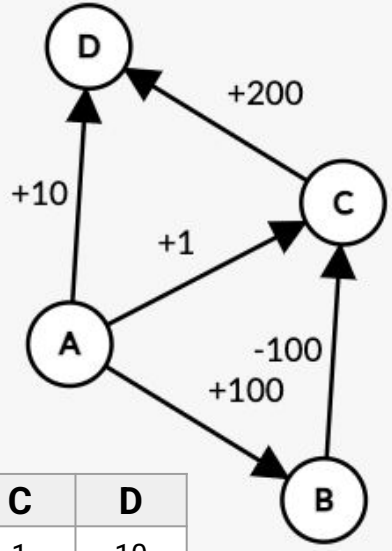
1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

	A	B	C	D
A	❌	100	1	10
B	❌	❌	-100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:	100	-100	200	0
Шаг 2. V:				
Шаг 3. V:				
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

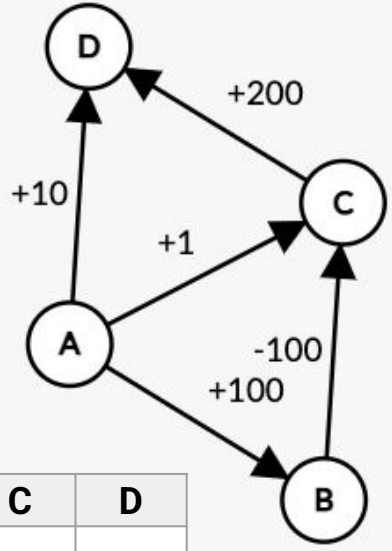
1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

	A	B	C	D
A	❌	100	1	10
B	❌	❌	-100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:	100	-100	200	0
Шаг 2. V:				
Шаг 3. V:				
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

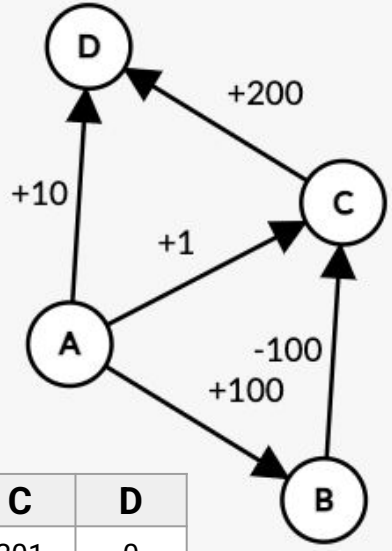
1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

	A	B	C	D
A	❌			
B	❌	❌		❌
C	❌	❌	❌	
D	❌	❌	❌	❌

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:	100	-100	200	0
Шаг 2. V:				
Шаг 3. V:				
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

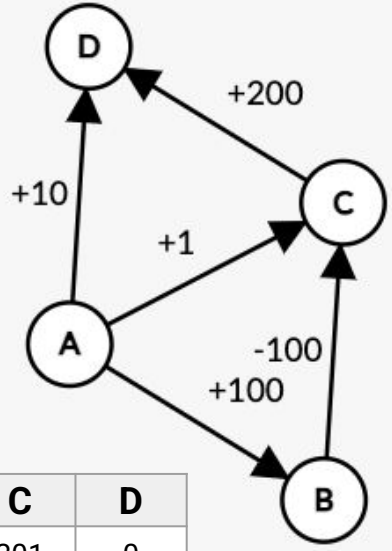
1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

	A	B	C	D
A	❌	0	201	0
B	❌	❌	100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:	100	-100	200	0
Шаг 2. V:	201	100	200	0
Шаг 3. V:				
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

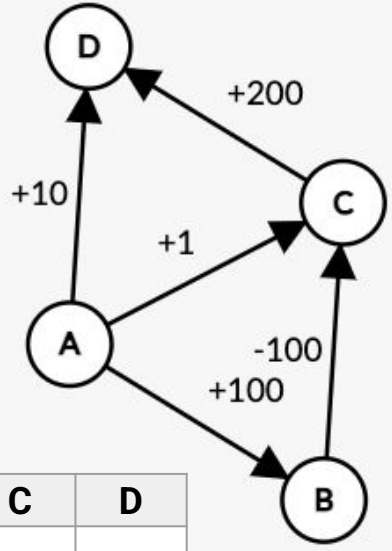
1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

	A	B	C	D
A	❌	0	201	0
B	❌	❌	100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:	100	-100	200	0
Шаг 2. V:	201	100	200	0
Шаг 3. V:				
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

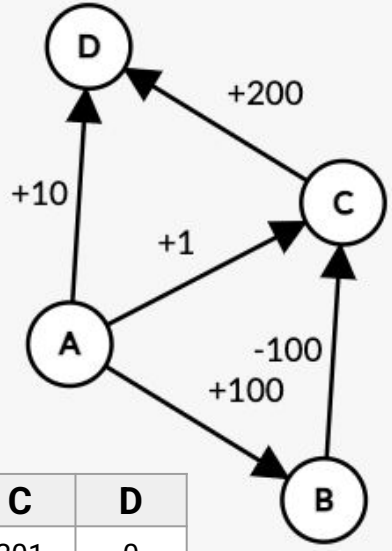
1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

	A	B	C	D
A	❌			
B	❌	❌		❌
C	❌	❌	❌	
D	❌	❌	❌	❌

№4. ДП: Value-Iteration

	A	B	C	D
Шаг 0. V:	0	0	0	0
Шаг 1. V:	100	-100	200	0
Шаг 2. V:	201	100	200	0
Шаг 3. V:	201	100	200	0
Шаг 4. V:				



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

Выход: $\pi(s) \coloneqq \operatorname{argmax}_a [r(s,a) + \gamma \mathbb{E}_{s'} V(s')]$

	A	B	C	D
A	❌	0	201	0
B	❌	❌	100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌

Policy Iteration

Алгоритм Policy Iteration

Алгоритм 8: Policy Iteration

Гиперпараметры: ε — критерий останова для процедуры **PolicyEvaluation**

Инициализируем $\pi_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $V^{\pi_k} := \text{PolicyEvaluation}(\pi_k, \varepsilon)$
2. $Q^{\pi_k}(s,a) := r(s,a) + \gamma \mathbb{E}_{s'} V^{\pi_k}(s')$
3. $\pi_{k+1}(s) := \underset{a}{\text{argmax}}\, Q^{\pi_k}(s,a)$
4. **критерий останова:** $\pi_k \equiv \pi_{k+1}$

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s$: $V_{k+1}(s) := \mathbb{E}_a \left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s') \right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Политика:

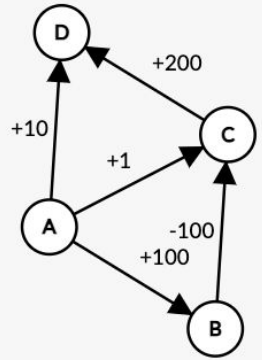
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
?	?	?	?

Q:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B	C	D
?	?	?	?

Q:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

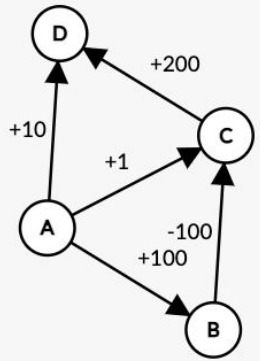
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
0	0	0	0

Q:

	A	B	C	D
A	✗	0	0	0
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s\colon V_{k+1}(s) \coloneqq \mathbb{E}_a\left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')\right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

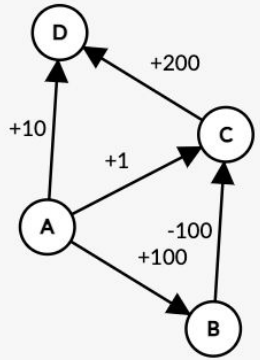
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
100	-100	200	0

Q:

	A	B	C	D
A	❌	100	200	10
B	❌	❌	-100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s\colon V_{k+1}(s) \coloneqq \mathbb{E}_a\left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')\right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

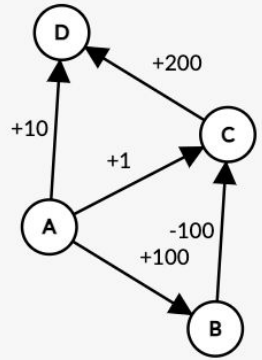
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
100	-100	200	0

Q:

	A	B	C	D
A	✗			
B	✗	✗		✗
C	✗	✗	✗	
D	✗	✗	✗	✗



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s\colon V_{k+1}(s) \coloneqq \mathbb{E}_a\left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')\right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

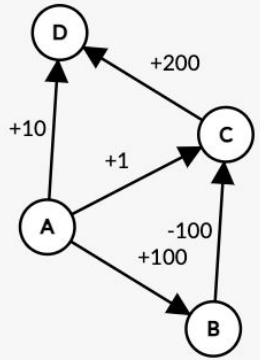
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
100	-100	200	0

Q:

	A	B	C	D
A	❌	0	201	10
B	❌	❌	100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s\colon V_{k+1}(s) \coloneqq \mathbb{E}_a\left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')\right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

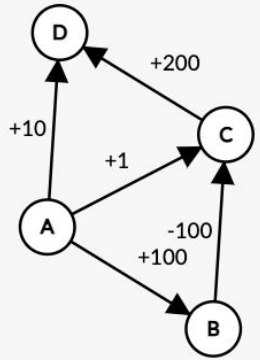
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
201	100	200	0

Q:

	A	B	C	D
A	❌	0	201	10
B	❌	❌	100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s\colon V_{k+1}(s) \coloneqq \mathbb{E}_a\left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')\right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

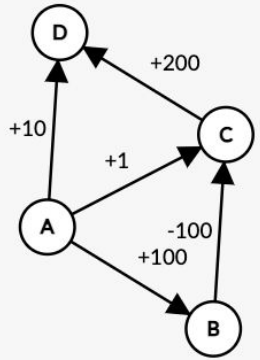
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
201	100	200	0

Q:

	A	B	C	D
A	✗			
B	✗	✗		✗
C	✗	✗	✗	
D	✗	✗	✗	✗



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s\colon V_{k+1}(s) \coloneqq \mathbb{E}_a\left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')\right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

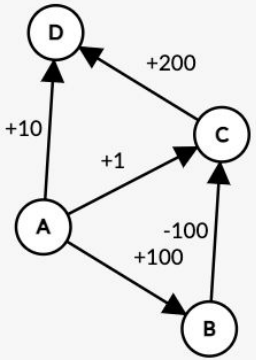
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
201	100	200	0

Q:

	A	B	C	D
A	❌	200	201	10
B	❌	❌	100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s$: $V_{k+1}(s) \coloneqq \mathbb{E}_a \left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s') \right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

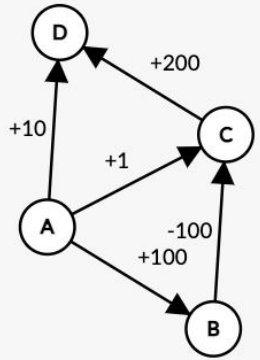
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
201	100	200	0

Q:

	A	B	C	D
A	❌	200	201	10
B	❌	❌	100	❌
C	❌	❌	❌	200
D	❌	❌	❌	❌



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 6: Policy Evaluation

Вход: $\pi(a \mid s)$ — стратегия
Гиперпараметры: ε — критерий останова

Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $\forall s\colon V_{k+1}(s) \coloneqq \mathbb{E}_a\left[r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')\right]$
2. **критерий останова:** $\max_s |V_k(s) - V_{k+1}(s)| < \varepsilon$

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Улучшаем политику

Политика:

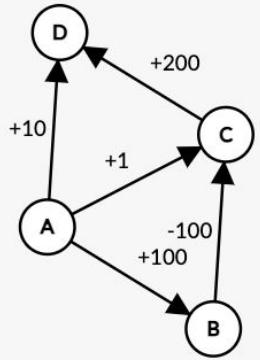
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
201	100	200	0

Q:

	A	B	C	D
A	✗	200	201	10
B	✗	✗	100	✗
C	✗	✗	✗	200
D	✗	✗	✗	✗



Политика:

	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

V:

A	B
?	?

Q:

Алгоритм 8: Policy Iteration

Гиперпараметры: ε — критерий останова для процедуры **PolicyEvaluation**

Инициализируем $\pi_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $V^{\pi_k} := \text{PolicyEvaluation}(\pi_k, \varepsilon)$
2. $Q^{\pi_k}(s,a) := r(s,a) + \gamma \mathbb{E}_{s'} V^{\pi_k}(s')$
3. $\pi_{k+1}(s) := \underset{a}{\text{argmax}}\, Q^{\pi_k}(s,a)$
4. **критерий останова:** $\pi_k \equiv \pi_{k+1}$

№5. ДП: Policy-Iteration

Улучшаем политику

Политика:

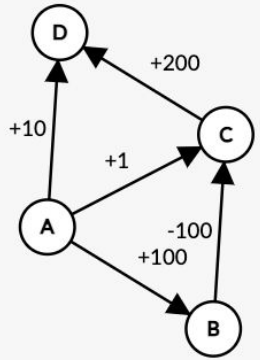
	A	B	C	D
A	0.0	1.0	0.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B	C	D
201	100	200	0

Q:

	A	B	C	D
A	✗	200	201	10
B	✗	✗	100	✗
C	✗	✗	✗	200
D	✗	✗	✗	✗



Политика:

	A	B	C	D
A	0.0	0.0	1.0	0.0
B	0.0	0.0	1.0	0.0
C	0.0	0.0	0.0	1.0
D	0.0	0.0	0.0	0.0

V:

A	B
?	?

Q:

Алгоритм 8: Policy Iteration

Гиперпараметры: ε — критерий останова для процедуры **PolicyEvaluation**

Инициализируем $\pi_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

1. $V^{\pi_k} := \textbf{PolicyEvaluation}(\pi_k, \varepsilon)$
2. $Q^{\pi_k}(s,a) := r(s,a) + \gamma \mathbb{E}_{s'} V^{\pi_k}(s')$
3. $\pi_{k+1}(s) := \underset{a}{\operatorname{argmax}}\, Q^{\pi_k}(s,a)$
4. **критерий останова:** $\pi_k \equiv \pi_{k+1}$

Резюме

Итого

1. **Agent**
2. **Environment:**
   State, Actions, Dynamics

Дано

MDP

3. Rewards
4. Алгоритм

Найти

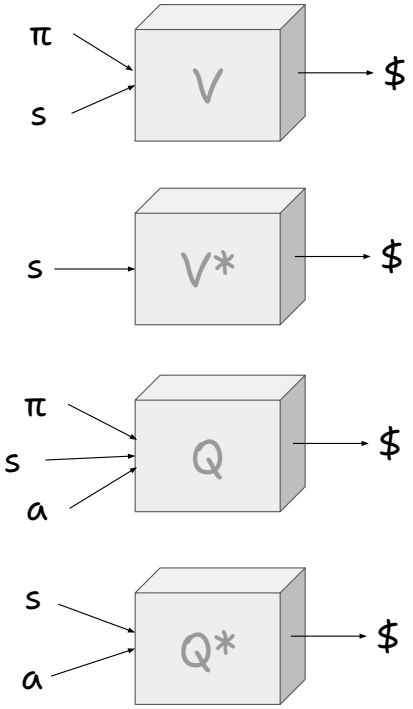
5. Политика

Результат

Опрос в конце: https://otus.ru/polls/141246/

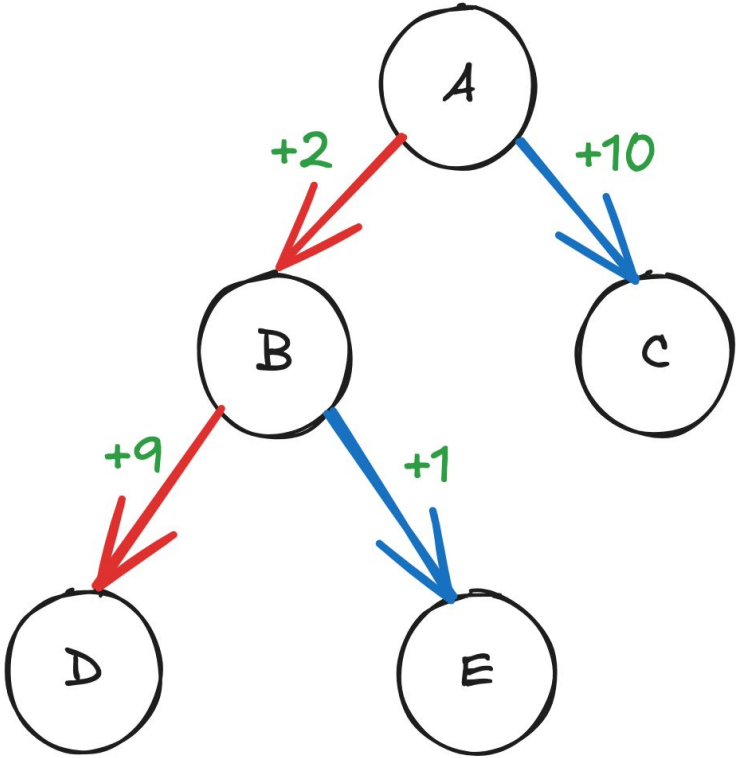
План

- Терминальные состояния
- Оптимальная политика
- Value-функция
- Q-функция
- Дисконтирование
- **Value iteration**
- Policy iteration



Опрос в конце: https://otus.ru/polls/141247/

Вопрос



Какая цель?

$$\mathbb{E}_{\tau \sim \pi} \sum_{t \geq 0} r_t \to \max_{\pi},$$

Откуда может взяться случайность?

Дисклеймер: В презентации использованы личные материалы **@dmi3eva**.

Образовательная площадка **Otus** не несет за них ответственность.