



Faculty  
of  
**Computer**  
science  
Higher School of Economics

# Уравнения Беллмана Value iteration и Policy iteration

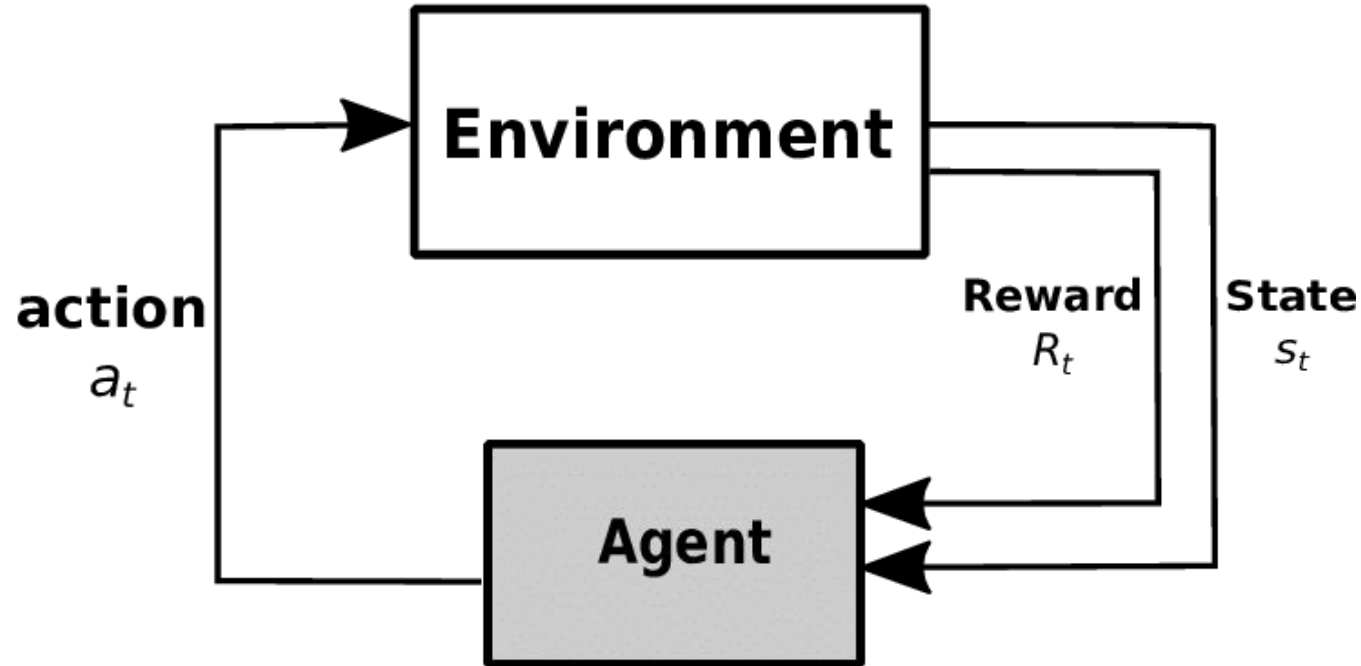
Подготовил:  
Казадаев Максим, БПМИ202

# План



1. Постановка задачи
2. Уравнения Беллмана.
3. Value iteration
4. Policy iteration
5. Сравнение двух методов
6. Теоретическое обоснование

# Постановка задачи



# Постановка задачи

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \sum_{j=0}^{\infty} \gamma^j r_{t+j+1}$$

Суммарный выигрыш

Фактор дисконтирования  
(discount factor)

Награды

# Уравнения Беллмана

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

Награда за действие

Выигрыш в следующем состоянии

# Уравнения Беллмана

$$\begin{aligned} V_{\pi}(s) &= \mathbb{E}_{\pi}[G_t | s_t = s] = \mathbb{E}_{\pi}[r_t + \gamma G_{t+1} | s_t = s] = \mathbb{E}_{s_t=s, s_{next} \sim \pi(s)}[r_t + \gamma V_{\pi}(s_{next})] = \\ &= \sum_a \pi(a|s) \cdot \sum_{s_{next}} P_{ss_{next}}^a (r(s, a) + \gamma V_{\pi}(s_{next})) \end{aligned}$$

Текущая политика

Переходные вероятности  
(из  $s$  в  $s_{next}$  с помощью действия  $a$ )

Текущая награда

Value-функция  
следующего состояния

# Уравнения Беллмана

$$V_{\pi}(s) = \sum_{s_{next}} P_{ss_{next}}^a (r(s, a) + \gamma V_{\pi}(s_{next})), \pi(s) = a$$

$$V_{\pi^*}(s) = \max_a \sum_{s_{next}} P_{ss_{next}}^a (r(s, a) + \gamma V_{\pi^*}(s_{next}))$$

Детерминированная  
политика

Оптимальная политика

# Value iteration

Инициализируем  $V(s)$  случайно

Повторяем, пока  $V(s)$  не сойдется:

Для всех состояний  $s$ :

Для всех действий  $a$ :

$$Q(s, a) = \mathbb{E}_{\pi}[G_t | S_t = s, A_t = a] = \sum_{s_{next}} P_{ss_{next}}^a \cdot (r(s, a) + \gamma V(s_{next}))$$

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

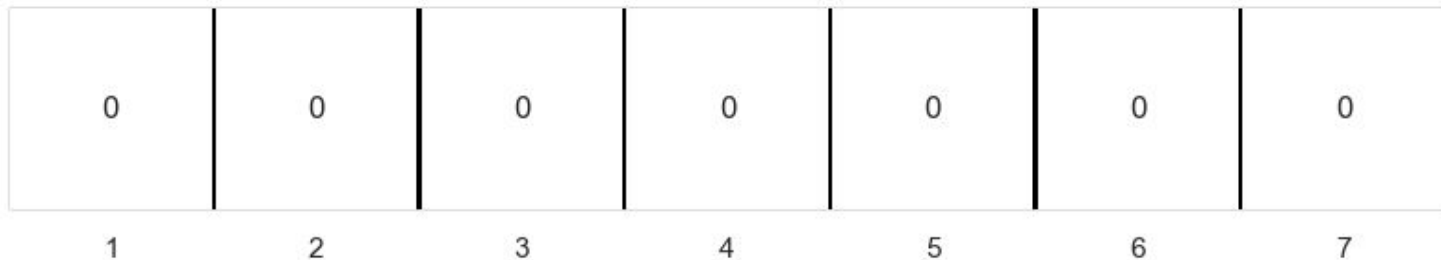
Итоговая политика:

$$\pi(s) = \arg \max_a Q(s, a)$$



# Value iteration: Frozen Lake

(gamma=1)



Действия:

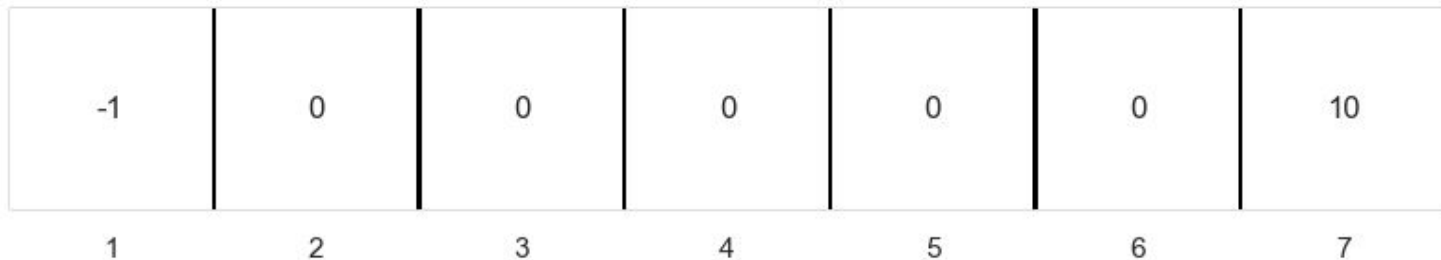
1. Остаться в клетке
2. Влево
3. Вправо

Награды:

1. +10 за пребывание в 7 квадрате
2. -1 за пребывание в 1 квадрате

# Value iteration: Frozen Lake

(gamma=1)



$$V(1) = -1; V(7) = 10$$

Действия:

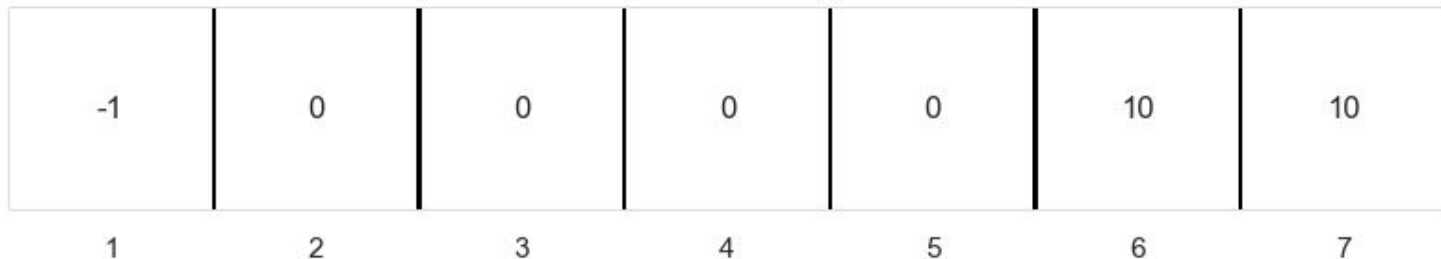
1. Остаться в клетке
2. Влево
3. Вправо

Награды:

1. +10 за пребывание в 7 квадрате
2. -1 за пребывание в 1 квадрате

# Value iteration: Frozen Lake

## (gamma=1)



*Делаем шаг.  $V(6) = 0 + \text{gamma} * \max(V(5), V(7)) = 10$*

Действия:

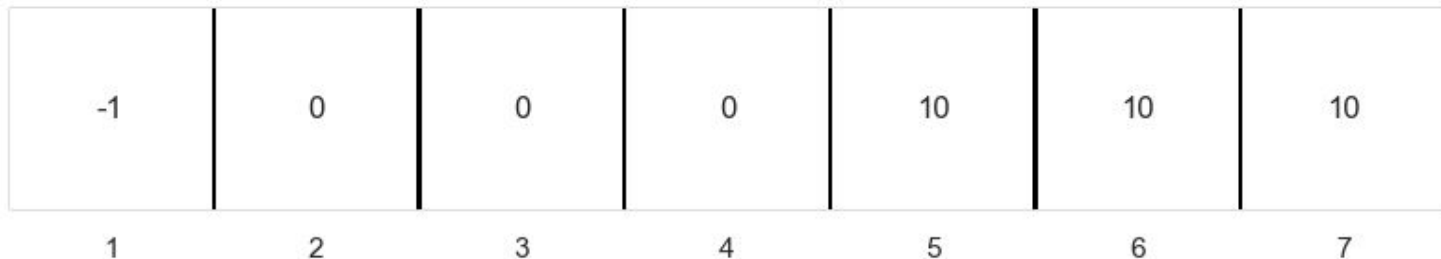
1. Остаться в клетке
2. Влево
3. Вправо

Награды:

1. +10 за пребывание в 7 квадрате
2. -1 за пребывание в 1 квадрате

# Value iteration: Frozen Lake

## (gamma=1)



**Еще шаг.  $V(5) = 0 + \text{gamma} * \max(V(4), V(6)) = 10$**

Действия:

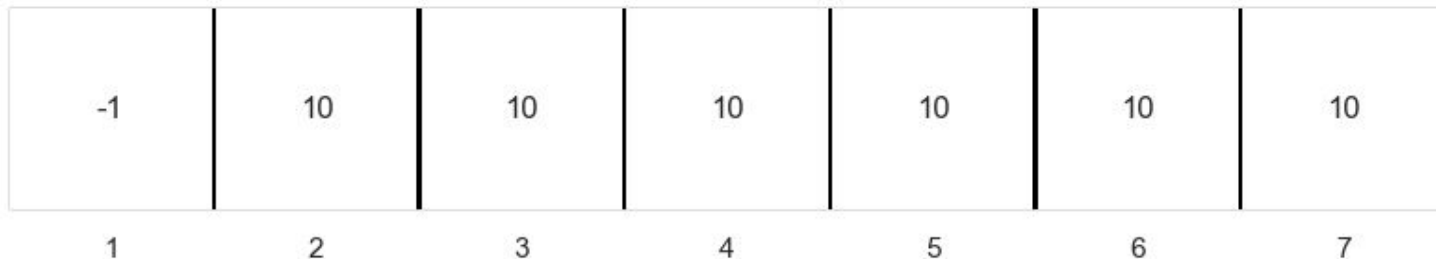
1. Остаться в клетке
2. Влево
3. Вправо

Награды:

1. +10 за пребывание в 7 квадрате
2. -1 за пребывание в 1 квадрате

# Value iteration: Frozen Lake

(gamma=1)



*Продолжаем, пока Value-функция не перестанет меняться*

Действия:

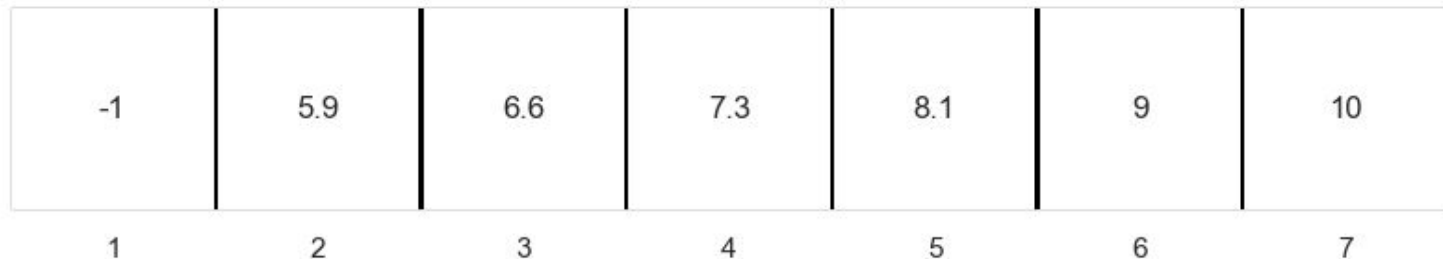
1. Остаться в клетке
2. Влево
3. Вправо

Награды:

1. +10 за пребывание в 7 квадрате
2. -1 за пребывание в 1 квадрате

# Value iteration: Frozen Lake

(gamma=0.9)



**Итоговые значения Value-функции, если  $\gamma=0.9$**

Действия:

1. Остаться в клетке
2. Влево
3. Вправо

Награды:

1. +10 за пребывание в 7 квадрате
2. -1 за пребывание в 1 квадрате

# Policy iteration

Инициализируем  $\pi(s)$  и  $V(s)$  случайно

1. **Policy Evaluation (считаем value-функцию для текущей политики)**

Повторяем, пока  $V(s)$  не сойдется:

Для всех состояний  $s$ :

$$V(s) = \sum_{s_{next}} P_{ss_{next}}^{a=\pi(s)} (r(s, a = \pi(s)) + \gamma V(s_{next}))$$

2. **Policy Improvement (обновляем политику, чтобы максимизировать value-функцию)**

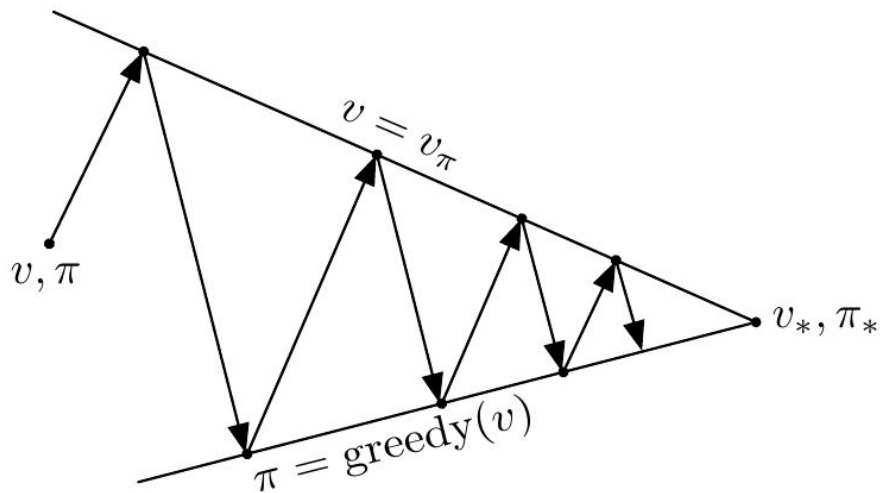
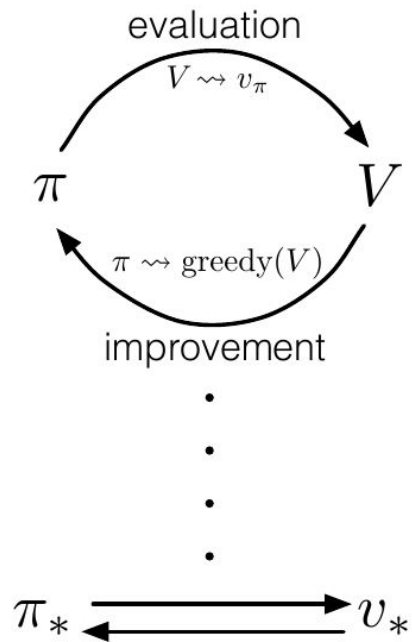
Обновляем политику:

$$Q(s, a) = \sum_{s_{next}} P_{ss_{next}}^a (r(s, a) + \gamma V(s_{next}))$$
























































































$$\pi(s) = \arg \max_a Q(s, a)$$

3. **Переходим к шагу 1, пока политика не сойдется**

# Policy iteration





0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 
0.00 					0.00 				0.00 
0.00 	0.00 	0.00 	0.00 		0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 

# Policy iteration: пример



## Value iteration

0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖					0.00 ↖				0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	-1.00 ↖ R-1.0		0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖		-1.00 ↖ R-1.0	-1.00 ↖ R-1.0	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖		1.00 ↖ R 1.0	-1.00 ↖ R-1.0	0.00 ↖	-1.00 ↖ R-1.0	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖		0.00 ↖	0.00 ↖	0.00 ↖	-1.00 ↖ R-1.0	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	-1.00 ↖ R-1.0		-1.00 ↖ R-1.0	-1.00 ↖ R-1.0	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖

## 1. Policy Evaluation

## Policy iteration

-0.06 ↖	-0.05 ↖	-0.04 ↖	-0.05 ↖	-0.08 ↖	-0.11 ↖	-0.09 ↖	-0.08 ↖	-0.10 ↖	-0.13 ↖
-0.09 ↖	-0.05 ↖	-0.04 ↖	-0.06 ↖	-0.10 ↖	-0.20 ↖	-0.11 ↖	-0.09 ↖	-0.11 ↖	-0.20 ↖
-0.19 ↖					-0.58 ↖				-0.42 ↖
-0.34 ↖	-0.61 ↖	-1.26 ↖	-2.88 ↖ R-1.0		-1.22 ↖	-1.48 ↖	-1.19 ↖	-1.00 ↖	-0.78 ↖
-0.35 ↖	-0.50 ↖	-0.84 ↖	-1.33 ↖		-2.15 ↖ R-1.0	-2.67 ↖ R-1.0	-1.62 ↖	-1.47 ↖	-1.18 ↖
-0.33 ↖	-0.42 ↖	-0.63 ↖	-0.88 ↖		0.94 ↖ R 1.0	-2.20 ↖ R-1.0	-1.87 ↖	-2.74 ↖ R-1.0	-1.67 ↖
-0.32 ↖	-0.42 ↖	-0.67 ↖	-1.07 ↖		-1.02 ↖	-1.71 ↖	-1.76 ↖	-2.71 ↖ R-1.0	-1.67 ↖
-0.32 ↖	-0.45 ↖	-0.87 ↖	-2.13 ↖ R-1.0		-2.73 ↖ R-1.0	-2.64 ↖ R-1.0	-1.54 ↖	-1.42 ↖	-1.18 ↖
-0.29 ↖	-0.38 ↖	-0.60 ↖	-0.95 ↖	-0.86 ↖	-1.30 ↖	-1.32 ↖	-1.02 ↖	-0.90 ↖	-0.83 ↖
-0.29 ↖	-0.34 ↖	-0.48 ↖	-0.64 ↖	-0.71 ↖	-0.87 ↖	-0.89 ↖	-0.79 ↖	-0.72 ↖	-0.70 ↖

# Policy iteration: пример

## Value iteration

0.00 ↙	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↘	0.00 ↗	0.00 ↗	0.00 ↗	0.00 ↗	0.00 ↗	0.00 ↗	0.00 ↗	0.00 ↗	0.00 ↗
0.00 ↘					0.00 ↗				0.00 ↗
0.00 ↘	0.00 ↗	0.00 ↗	-1.00 ↙ R -1.0		0.00 ↗	0.00 ↗	0.00 ↗	0.00 ↗	0.00 ↗
0.00 ↘	0.00 ↗	0.00 ↗	0.00 ↖		-1.00 ↓ R -1.0	-1.00 ↙ R -1.0	0.00 ↗	0.00 ↗	0.00 ↗
0.00 ↘	0.00 ↗	0.00 ↗	0.00 ↗		1.00 ↗ R 1.0	-1.00 ↖ R -1.0	0.00 ↓	-1.00 ↙ R -1.0	0.00 ↓
0.00 ↘	0.00 ↗	0.00 ↗	0.00 ↗		0.00 ↗	0.00 ↔	0.00 ↘	-1.00 ↙ R -1.0	0.00 ↓
0.00 ↘	0.00 ↗	0.00 ↗	-1.00 ↙ R -1.0		-1.00 ↓ R -1.0	-1.00 ↙ R -1.0	0.00 ↗	0.00 ↗	0.00 ↗
0.00 ↘	0.00 ↗	0.00 ↗	0.00 ↖	0.00 ↗	0.00 ↖	0.00 ↖	0.00 ↗	0.00 ↗	0.00 ↗
0.00 ↙	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖

## 2. Policy Improvement

## Policy iteration

-0.06 →	-0.05 →	-0.04 ↓	-0.05 ←	-0.08 ←	-0.11 ←	-0.09 →	-0.09 ↓	-0.10 ←	-0.14 ←
-0.09 →	-0.05 →	-0.05 ←	-0.06 ←	-0.10 ←	-0.20 ←	-0.11 →	-0.09 ↓	-0.11 ←	-0.20 ←
-0.19 ↑					-0.18 ↑				-0.12 ↑
-0.14 ↑	-0.61 ←	-1.26 ←	-2.88 ← R -1.0		-1.12 ↑	-1.48 →	-1.19 →	-1.00 →	-0.18 ↑
-0.35 ↓	-0.50 ←	-0.84 ←	-1.33 ←		-2.15 ↓ R -1.0	-2.07 ↑ R -1.0	-1.62 ↑	-1.47 ↑	-1.18 ↑
-0.33 ↓	-0.42 ←	-0.63 ←	-0.88 ←		0.94 ↗ R 1.0	-2.20 ↑ R -1.0	-1.87 ↑	-2.74 ↑ R -1.0	-1.67 ↑
-0.32 ↓	-0.42 ←	-0.67 ←	-1.07 ←		-1.02 ↑	-1.71 ←	-1.76 ↓	-2.71 ↑ R -1.0	-1.67 ↓
-0.32 ↓	-0.45 ←	-0.87 ←	-2.13 ← R -1.0		-2.13 ↑ R -1.0	-2.64 ↓ R -1.0	-1.54 ↓	-1.42 ↓	-1.18 ↓
-0.30 ↓	-0.38 ←	-0.60 ←	-0.95 ←	-0.86 ↓	-1.30 ←	-1.32 ↓	-1.02 ↓	-0.90 ↓	-0.83 ↓
-0.19 ↑	-0.34 ←	-0.48 ←	-0.64 ←	-0.71 ←	-0.87 ←	-0.89 →	-0.79 →	-0.72 →	-0.70 ←

# Policy iteration: пример



## Value iteration

0.00 ↙	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘
0.00 ↘					0.00 ↘				0.00 ↘
0.00 ↘	0.00 ↘	0.00 ↘	-1.00 ↘ R -1.0		0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘
0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘		-0.10 ↓ R -1	-1.00 ↘ R -1.0	0.00 ↘	0.00 ↘	0.00 ↘
0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘		1.00 ↘ R 1.0	-0.10 ↖ R -1.0	0.00 ↓	-1.00 ↖ R -1.0	0.00 ↓
0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘		0.90 ↘ R 1.0	0.00 ↔	0.00 ↘	-1.00 ↖ R -1.0	0.00 ↓
0.00 ↘	0.00 ↘	0.00 ↘	-1.00 ↘ R -1.0		-1.00 ↓ R -1.0	-1.00 ↘ R -1.0	0.00 ↘	0.00 ↘	0.00 ↘
0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘	0.00 ↘
0.00 ↙	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖

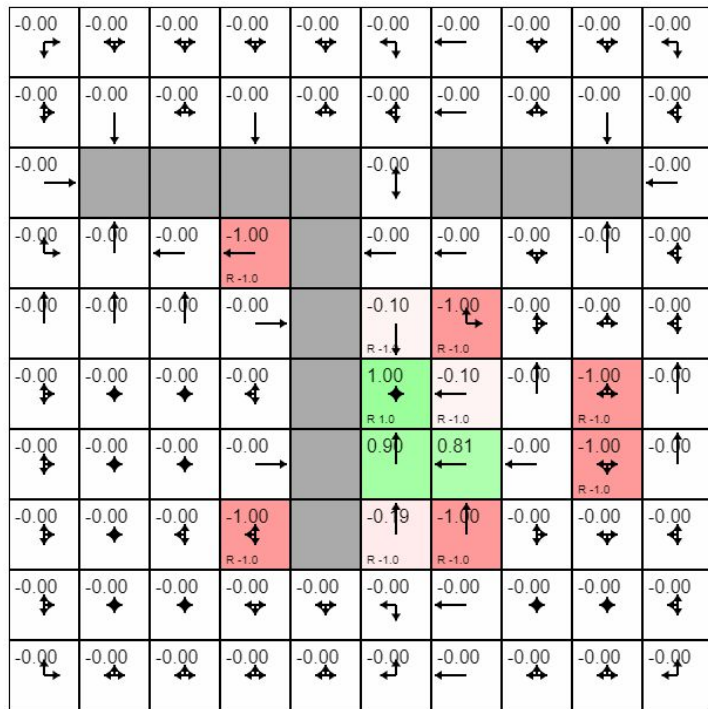
## 3. Policy Evaluation

## Policy iteration

























































































-0.00 →	-0.00 →	-0.00 ↓	-0.00 ←	-0.00 ←	-0.00 ←	-0.00 →	-0.00 ↓	-0.00 ←	-0.00 ←
-0.00 →	-0.00 →	-0.00 ↓	-0.00 ←	-0.00 ←	-0.00 ←	-0.00 →	-0.00 ↓	-0.00 ←	-0.00 ←
-0.00 ↓					-0.00 ↓				-0.00 ↓
-0.00 ↓	-0.00 ←	-0.00 ←	-1.00 ← R -1.0		-0.00 ↓	-0.00 →	-0.00 →	-0.00 →	-0.00 ↓
-0.00 ↓	-0.00 ←	-0.00 ←	-0.00 ←		-0.10 ↓ R -1	-1.00 ↖ R -1.0	-0.00 ↖	-0.00 ↖	-0.00 ↖
-0.00 ↓	-0.00 ←	-0.00 ←	-0.00 ←		1.00 ↘ R 1.0	-0.10 ↖ R -1.0	-0.00 ↖	-1.00 ↖ R -1.0	-0.00 ↖
-0.00 ↓	-0.00 ←	-0.00 ←	-0.00 ←		0.90 ↘ R 1.0	0.81 ←	-0.00 ↓	-1.00 ↖ R -1.0	-0.00 ↓
-0.00 ↓	-0.00 ←	-0.00 ←	-1.00 ← R -1.0		-0.10 ↖ R -1.0	-1.00 ↓ R -1.0	-0.00 ↓	-0.00 ↓	-0.00 ↓
-0.00 ↓	-0.00 ←	-0.00 ←	-0.00 ←	-0.00 ↓	-0.00 ←	-0.00 ↓	-0.00 ↓	-0.00 ↓	-0.00 ↓
-0.00 ↓	-0.00 ←	-0.00 ←	-0.00 ←	-0.00 ←	-0.00 →	-0.00 →	-0.00 →	-0.00 →	-0.00 ←

## 4. Policy Improvement











































## Policy iteration



## 5. Policy Evaluation

0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 
0.00 					0.00 				0.00 
0.00 	0.00 	0.00 	-1.00  R-1.0		0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		-0.10  R-1.0	-1.00  R-1.0	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		1.00  R 1.0	-0.10  R-1.0	0.00 	-1.00  R-1.0	0.00 
0.00 	0.00 	0.00 	0.00 		0.90  R 1.0	0.81  R-1.0	0.00 	-1.00  R-1.0	0.00 
0.00 	0.00 	0.00 	-1.00  R-1.0		-0.10  R-1.0	-1.00  R-1.0	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 

## Policy iteration

-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 
-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 
-0.00 					-0.00 				-0.00 
-0.00 	-0.00 	-0.00 	-1.00  R-1.0		-0.00 	-0.00 	-0.00 	-0.00 	-0.00 
-0.00 	-0.00 	-0.00 	-0.00 		-0.10  R-1.0	-1.00  R-1.0	-0.00 	-0.00 	-0.00 
-0.00 	-0.00 	-0.00 	-0.00 		1.00  R 1.0	-0.10  R-1.0	-0.00 	-1.00  R-1.0	-0.00 
-0.00 	-0.00 	-0.00 	-0.00 		0.90  R 1.0	0.81  R-1.0	0.73  R-1.0	-0.75  R-1.0	-0.00 
-0.00 	-0.00 	-0.00 	-1.00  R-1.0		-0.19  R-1.0	-0.47  R-1.0	0.28 	0.12 	0.05 
-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	0.09 	0.07 	0.05 
-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	0.04 	0.04 	0.04 



# Policy iteration: пример

## Value iteration

0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00					0.00				0.00
0.00	0.00	0.00	-1.00		0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		-0.10	-1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		1.00	-0.10	0.00	-1.00	0.00
0.00	0.00	0.00	0.00		0.90	0.81	0.00	-1.00	0.00
0.00	0.00	0.00	-1.00		-0.10	-1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

## 6. Policy Improvement

## Policy iteration

-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00					-0.00				-0.00
-0.00	-0.00	-0.00	-1.00		-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00		-0.10	-1.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00		1.00	-0.10	-0.00	-1.00	-0.00
-0.00	-0.00	-0.00	-0.00		0.90	0.81	0.73	-0.75	-0.00
-0.00	-0.00	-0.00	-1.00		-0.10	-0.10	0.28	0.12	0.05
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.05
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00

# Policy iteration: пример



## Value iteration

0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00					0.00				0.00
0.00	0.00	0.00	-1.00		0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		-0.10	-1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		1.00	-0.10	0.00	-1.00	0.00
0.00	0.00	0.00	0.00		0.90	0.81	0.73	-1.00	0.00
0.00	0.00	0.00	-1.00		-0.19	-0.27	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

## 7. Policy Evaluation

## Policy iteration

-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00					-0.00				-0.00
-0.00	-0.00	-0.00	-1.00		-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00		-0.10	-1.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00		1.00	-0.10	0.66	-0.41	-0.00
-0.00	-0.00	-0.00	-0.00		0.90	0.81	0.73	-0.34	0.48
-0.00	-0.00	-0.00	-1.00		-0.19	-0.27	0.60	0.59	0.53
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.53	0.50	0.48
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.48	0.50	0.40



# Policy iteration: пример



## Value iteration

0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00					0.00				0.00
0.00	0.00	0.00	-1.00		0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		-0.10	-1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		1.00	-0.10	0.00	-1.00	0.00
0.00	0.00	0.00	0.00		0.90	0.81	0.73	-1.00	0.00
0.00	0.00	0.00	-1.00		-0.19	-0.27	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

## 8. Policy Improvement

## Policy iteration

-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00					-0.00				-0.00
-0.00	-0.00	-0.00	-1.00		-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00		-0.10	-1.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00		1.00	-0.10	0.66	-0.41	-0.00
-0.00	-0.00	-0.00	-0.00		0.90	0.81	0.73	-0.34	0.48
-0.00	-0.00	-0.00	-1.00		-0.19	-0.27	0.60	0.59	0.53
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.53	0.50	0.53	0.48
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.48	0.50	0.48	0.43

# Policy iteration: пример



## Value iteration

## 9. Policy Evaluation

## Policy iteration

0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00					0.00				0.00
0.00	0.00	0.00	-1.00		0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		-0.10	-1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		1.00	-0.10	0.66	-1.00	0.00
0.00	0.00	0.00	0.00		0.90	0.81	0.73	-0.34	0.00
0.00	0.00	0.00	-1.00		-0.19	-0.27	0.66	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00					-0.00				-0.00
-0.00	-0.00	-0.00	-1.00		0.17	0.20	0.31	0.27	0.26
-0.00	-0.00	-0.00	-0.00		-0.10	-0.64	0.59	0.35	0.31
-0.00	-0.00	-0.00	-0.00		1.00	-0.10	0.66	-0.41	0.43
-0.00	-0.00	-0.00	-0.00		0.90	0.81	0.73	-0.34	0.48
-0.00	-0.00	-0.00	-0.96		-0.19	-0.27	0.66	0.59	0.53
0.00	-0.10	0.04	0.12	0.24	0.48	0.53	0.59	0.53	0.48
0.01	0.02	0.06	0.13	0.24	0.43	0.48	0.53	0.48	0.43

# Policy iteration: пример



## Value iteration

0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00					0.00				0.00
0.00	0.00	0.00	-1.00		0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		-0.10	-1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		1.00	-0.10	0.66	-1.00	0.00
0.00	0.00	0.00	0.00		0.90	0.81	0.73	-0.34	0.00
0.00	0.00	0.00	-1.00		-0.19	-0.27	0.66	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

## 10. Policy Improvement

## Policy iteration

-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00					-0.00				-0.00
-0.00	-0.00	-0.00	-1.00		0.17	0.20	0.31	0.27	0.26
-0.00	-0.00	-0.00	-0.00		-0.10	-0.64	0.59	0.35	0.31
-0.00	-0.00	-0.00	-0.00		1.00	-0.10	0.66	-0.41	0.43
-0.00	-0.00	-0.00	-0.00		0.90	0.81	0.73	-0.34	0.48
-0.00	-0.00	-0.00	-0.96		-0.19	-0.27	0.66	0.59	0.53
0.00	-0.00	0.04	0.12	0.24	0.48	0.53	0.59	0.53	0.48
0.01	0.02	0.06	0.13	0.24	0.43	0.48	0.53	0.48	0.43

# Policy iteration: пример



## Value iteration

## 11. Policy Evaluation







































## Policy iteration

0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00					0.00				0.00
0.00	0.00	0.00	-1.00		0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		-0.10	-1.00	0.59	0.00	0.00
0.00	0.00	0.00	0.00		1.00	-0.10	0.66	-0.41	0.00
0.00	0.00	0.00	0.00		0.90	0.81	0.73	-0.34	0.00
0.00	0.00	0.00	-1.00		-0.19	-0.17	0.66	0.59	0.00
0.00	0.00	0.00	0.00		0.00	0.00	0.59	0.00	0.00
0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00

-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
-0.00					0.39				0.31
-0.00	-0.00	-0.00	-1.00		0.43	0.48	0.53	0.48	0.35
-0.00	-0.00	-0.00	-0.00		-0.10	-0.47	0.59	0.53	0.39
-0.00	-0.00	-0.00	-0.00		1.00	-0.10	0.66	-0.41	0.43
-0.00	-0.00	-0.00	-0.00		0.90	0.81	0.73	-0.34	0.48
0.21	-0.00	0.31	-0.65		-0.19	-0.17	0.66	0.59	0.53
0.23	0.31	0.35	0.39	0.43	0.48	0.53	0.59	0.53	0.48
0.25	0.28	0.31	0.35	0.39	0.43	0.48	0.53	0.48	0.43

## 12. Policy Improvement

## Policy iteration

0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 
0.00 					0.00 				0.00 
0.00 	0.00 	0.00 	-1.00  R-1.0		0.00 	0.00 	0.00 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		-0.10  R-1	-1.00  R-1.0	0.59 	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 		1.00  R 1.0	-0.10  R-1.0	0.66 	-0.41  R-1.0	0.00 
0.00 	0.00 	0.00 	0.00 		0.90  R 1.0	0.81  R-1.0	0.73  R-1.0	-0.34  R-1.0	0.00 
0.00 	0.00 	0.00 	-1.00  R-1.0		-0.19  R-1.0	-0.27  R-1.0	0.66  R-1.0	0.59  R-1.0	0.00 
0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.59  R-1.0	0.00 	0.00 
0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 	0.00 

-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 
-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 	-0.00 
-0.00 					0.39 				0.31 
-0.00 	-0.00 	-0.00 	-1.00  R-1.0		0.43 	0.48 	0.53 	0.48 	0.35 
-0.00 	-0.00 	-0.00 	-0.00 		-0.10  R-1.0	-0.47  R-1.0	0.59 	0.53 	0.39 
-0.00 	-0.00 	-0.00 	-0.00 		1.00  R-1.0	-0.10  R-1.0	0.66 	-0.41  R-1.0	0.43 
-0.00 	-0.00 	-0.00 	-0.00 		0.90  R-1.0	0.81  R-1.0	0.73  R-1.0	-0.34  R-1.0	0.48 
0.21 	-0.00 	0.31 	-0.65  R-1.0		-0.19  R-1.0	-0.47  R-1.0	0.66  R-1.0	0.59  R-1.0	0.53 
0.23 	0.31 	0.35 	0.39 	0.43 	0.48 	0.53 	0.59 	0.53 	0.48 
0.25 	0.28 	0.31 	0.35 	0.39 	0.43 	0.48 	0.53 	0.48 	0.43 

# Policy iteration: пример



## Value iteration

0.22	0.25	0.27	0.31	0.34	0.38	0.34	0.31	0.34	0.38
0.25	0.27	0.31	0.34	0.38	0.42	0.38	0.34	0.38	0.42
0.27					0.46				0.46
0.20	0.22	0.25	-0.78		0.52	0.57	0.64	0.57	0.52
0.22	0.25	0.27	0.25		0.08	-0.36	0.71	0.64	0.57
0.25	0.27	0.31	0.27		1.20	0.08	0.79	-0.29	0.52
0.27	0.31	0.34	0.31		1.08	0.97	0.87	-0.21	0.57
0.31	0.34	0.38	-0.58		-0.03	-0.13	0.79	0.71	0.64
0.34	0.38	0.42	0.46	0.52	0.57	0.64	0.71	0.64	0.57
0.31	0.34	0.38	0.42	0.46	0.52	0.57	0.64	0.57	0.52

## Optimal

## Policy iteration

0.22	0.25	0.27	0.31	0.34	0.38	0.34	0.31	0.34	0.38
0.25	0.27	0.31	0.34	0.38	0.42	0.38	0.34	0.38	0.42
0.27					0.46				0.46
0.20	0.22	0.25	-0.78		0.52	0.57	0.64	0.57	0.52
0.22	0.25	0.27	0.25		0.08	-0.36	0.71	0.64	0.57
0.25	0.27	0.31	0.27		1.20	0.08	0.79	-0.29	0.52
0.27	0.31	0.34	0.31		1.08	0.97	0.87	-0.21	0.57
0.31	0.34	0.38	-0.58		-0.03	-0.13	0.79	0.71	0.64
0.34	0.38	0.42	0.46	0.52	0.57	0.64	0.71	0.64	0.57
0.31	0.34	0.38	0.42	0.46	0.52	0.57	0.64	0.57	0.52

# Policy iteration vs Value iteration

## Value iteration

$$V^0 \rightarrow \pi_0 \rightarrow V^1 \rightarrow \pi_1 \rightarrow \dots \rightarrow \pi^* \rightarrow V^*$$

↗  
Не истинная Value-функция ни для какой политики

Упрощенная версия Policy iteration, где мы для текущей политики делаем только один шаг, чтобы посчитать Value-функцию

## Policy iteration

$$\pi_0 \rightarrow V^{\pi_0} \rightarrow \pi_1 \rightarrow V^{\pi_1} \rightarrow \pi_2 \rightarrow \dots \rightarrow \pi^* \rightarrow V^*$$

↖  
Value-функция для политики  $\pi_0$

Оценивает Value-функцию по текущей политике точно

Обычно **сходится быстрее**, чем Value iteration, поэтому лучше использовать этот метод



# Value iteration: доказательство

1. Оператор Беллмана:  
Value-функция  $\rightarrow$  Value-функция

$$V \in \mathbb{R}^{|S|}$$

$$\mathcal{T} : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|S|}$$

$$(\mathcal{T}V)(s) = \max_a [r(s, a) + \gamma \mathbb{E}_{s'} [V(s') | a_t = a]]$$

2. Value iteration:

$$V_{k+1} = \mathcal{T}V_k$$

3. Нам нужна неподвижная точка.  
Value-функция сходится к ней  
путем применения оператора  
Беллмана

$$V^* = \mathcal{T}V^*$$

$$\|\mathcal{T}U - \mathcal{T}V\|_\infty \leq \gamma \|U - V\|_\infty, \forall U, V \in \mathbb{R}^{|S|}$$

$$\Rightarrow \lim_{k \rightarrow \infty} \mathcal{T}^k V_0 = V^*$$



# Policy iteration: доказательство

1. Policy improvement:

$$Q^{\pi_k}(s, a) = \sum_{s_{next}} P_{ss_{next}}^a (r(s, a) + \gamma V^{\pi_k}(s_{next}))$$
$$\pi_{k+1}(s) = \arg \max_a Q^{\pi_k}(s, a)$$

2. Policy evaluation:

Value-функция улучшается  
на всех состояниях

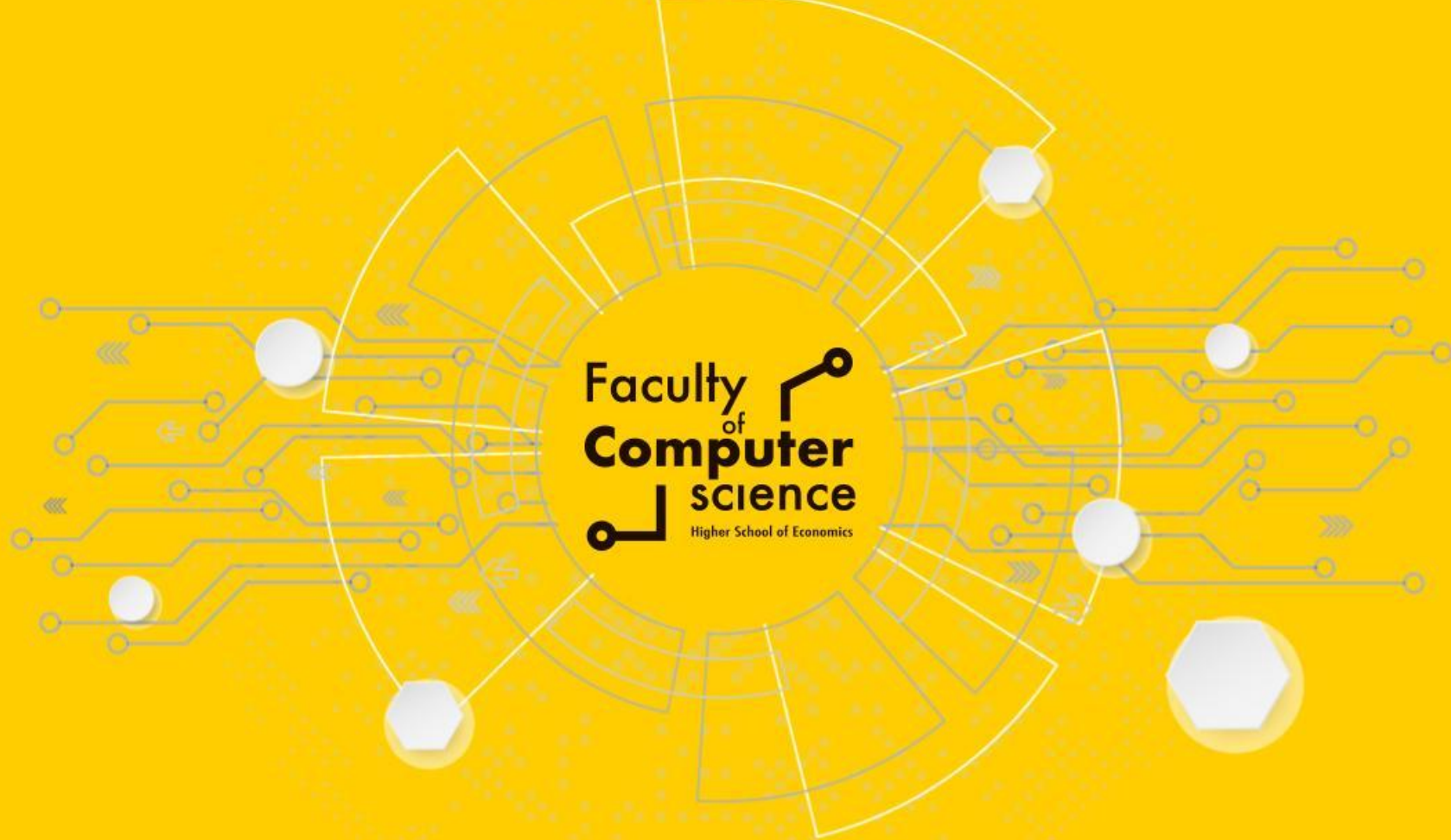
$$V^{\pi_{k+1}}(s) = \mathbb{E}_{a \sim \pi_{k+1}(s)} [Q^{\pi_{k+1}}(s, a)] \geq V^{\pi_k}(s)$$

3. В оптимуме Value-функция  
не изменяется

$$V^{\pi_{k+1}}(s) = V^{\pi_k}(s) \Rightarrow \pi_k = \pi^*$$

# Материалы

1. Value iteration:  
<http://incompleteideas.net/book/ebook/node44.html>
2. Policy iteration:  
<http://incompleteideas.net/book/ebook/node43.html>
3. Математическое обоснование:  
[https://yuanz.web.illinois.edu/teaching/IE498fa19/lec\\_16.pdf](https://yuanz.web.illinois.edu/teaching/IE498fa19/lec_16.pdf)
4. Игровой пример Value Iteration и Policy Iteration:  
[https://cs.stanford.edu/people/karpathy/reinforcejs/gridworld\\_dp.html](https://cs.stanford.edu/people/karpathy/reinforcejs/gridworld_dp.html)



[mskazadaev@edu.hse.ru](mailto:mskazadaev@edu.hse.ru)