

Обучение с подкреплением #2:

# Dynamic Programming

Policy Iteration, Value Iteration

Павел Темирчев

@cydoroga

# Напоминалочка

$s \sim \mathcal{S}; a \sim \mathcal{A}$  - пространства состояний \ действий

# Напоминалочка

$s \sim \mathcal{S}; a \sim \mathcal{A}$  - пространства состояний \ действий

$p(s_{t+1}|s_t, a_t)$  - динамика переходов в среде (марковская)

# Напоминалочка

$s \sim \mathcal{S}; a \sim \mathcal{A}$  - пространства состояний \ действий

$p(s_{t+1}|s_t, a_t)$  - динамика переходов в среде (марковская)

$r(s, a)$  - награда за действие  $a$  в состоянии  $s$

# Напоминалочка

$s \sim \mathcal{S}; a \sim \mathcal{A}$  - пространства состояний \ действий

$p(s_{t+1}|s_t, a_t)$  - динамика переходов в среде (марковская)

$r(s, a)$  - награда за действие  $a$  в состоянии  $s$

$\pi(a|s)$  - политика агента

# Напоминалочка

$s \sim \mathcal{S}; a \sim \mathcal{A}$  - пространства состояний \ действий

$p(s_{t+1}|s_t, a_t)$  - динамика переходов в среде (марковская)

$r(s, a)$  - награда за действие  $a$  в состоянии  $s$

$\pi(a|s)$  - политика агента

$p(\tau|\pi) = p(s_0) \prod_{t=0}^T \pi(a_t|s_t)p(s_{t+1}|a_t, s_t)$  - политика агента

где  $\tau = (s_0, a_0, s_1, a_1, \dots, s_T, a_T)$  - траектория агента

# Напоминалочка

$s \sim \mathcal{S}; a \sim \mathcal{A}$  - пространства состояний \ действия

$p(s_{t+1}|s_t, a_t)$  - динамика переходов в среде (марковская)

$r(s, a)$  - награда за действие  $a$  в состоянии  $s$

$\pi(a|s)$  - политика агента

$p(\tau|\pi) = p(s_0) \prod_{t=0}^T \pi(a_t|s_t)p(s_{t+1}|a_t, s_t)$  - политика агента

где  $\tau = (s_0, a_0, s_1, a_1, \dots, s_T, a_T)$  - траектория агента

$R_t = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \gamma^2 r(s_{t+2}, a_{t+2}) + \dots$  - reward to go

$$R_t = \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r(s_{t+\tau}, a_{t+\tau})$$

# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$

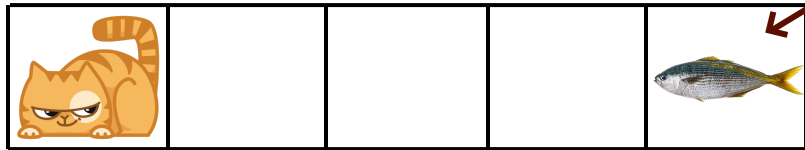


# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$

Политика  $\forall s$ :  $\longrightarrow$



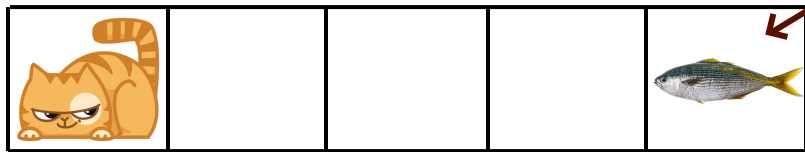
терминальное  
состояние

# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$

Политика  $\forall s$ :  $\longrightarrow$



терминальное  
состояние

$V$ -функция ценности:



$s_0$

$s_1$

$s_2$

$s_3$

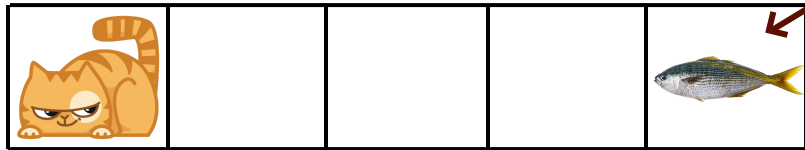
$s_4$

# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$

Политика  $\forall s$ :  $\longrightarrow$



терминальное  
состояние

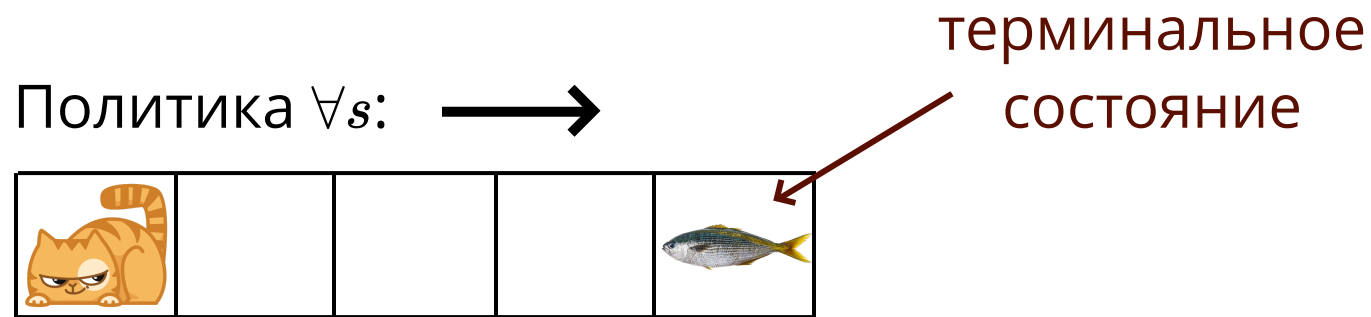
$V$ -функция ценности:

$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	1
$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$



$V$ -функция ценности:

$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	<b>1</b>
$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

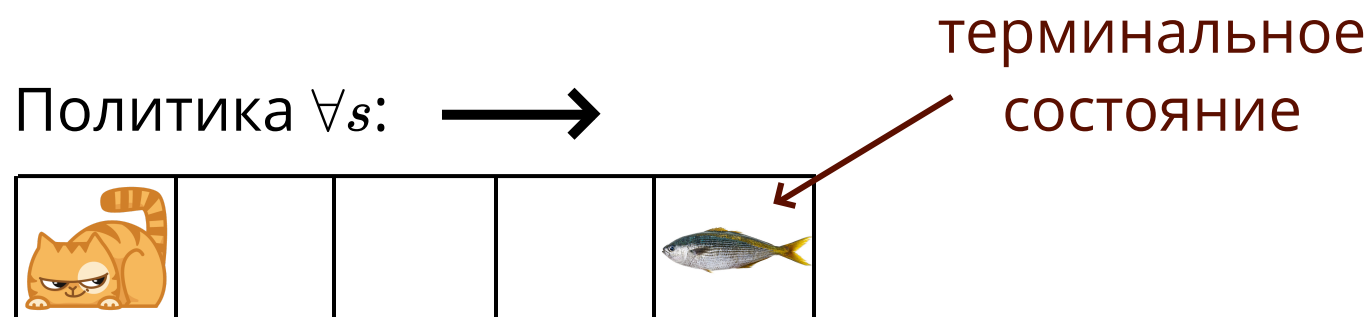
А что если в  $s$  "принудительно" выбрать действие  $a$ , а только затем идти по политике  $\pi$ ?

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s, a_t = a]$$

# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$

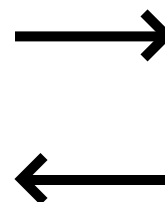


$V$ -функция ценности:

$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	1
$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

А что если в  $s$  "принудительно" выбрать действие  $a$ , а только затем идти по политике  $\pi$ ?

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s, a_t = a]$$



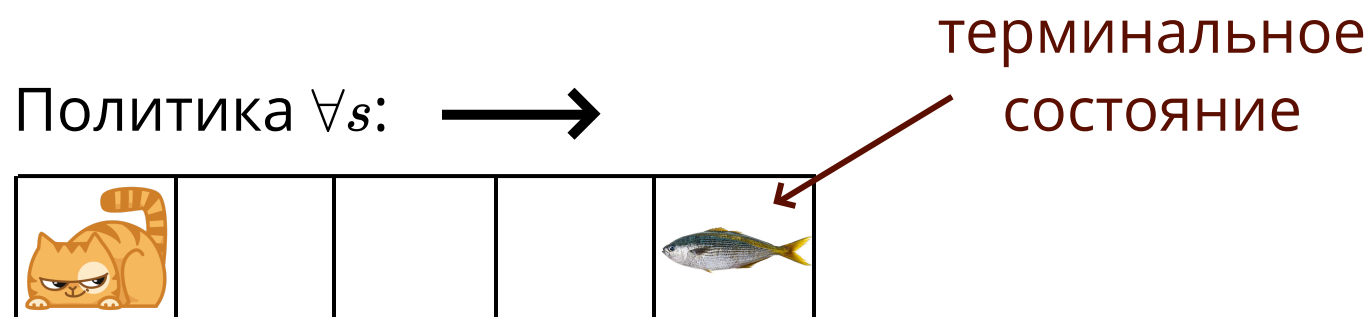
$Q$ -функция ценности:

$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$

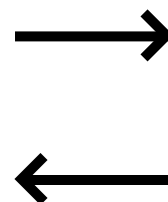


$V$ -функция ценности:

$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	1
$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

А что если в  $s$  "принудительно" выбрать действие  $a$ , а только затем идти по политике  $\pi$ ?

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s, a_t = a]$$

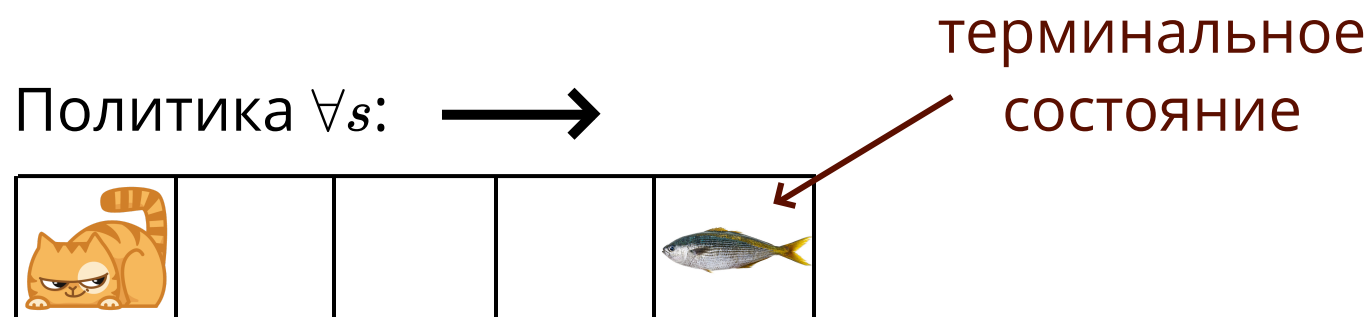


$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	1
$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$



$V$ -функция ценности:

$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	1
$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

А что если в  $s$  "принудительно" выбрать действие  $a$ , а только затем идти по политике  $\pi$ ?

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s, a_t = a]$$

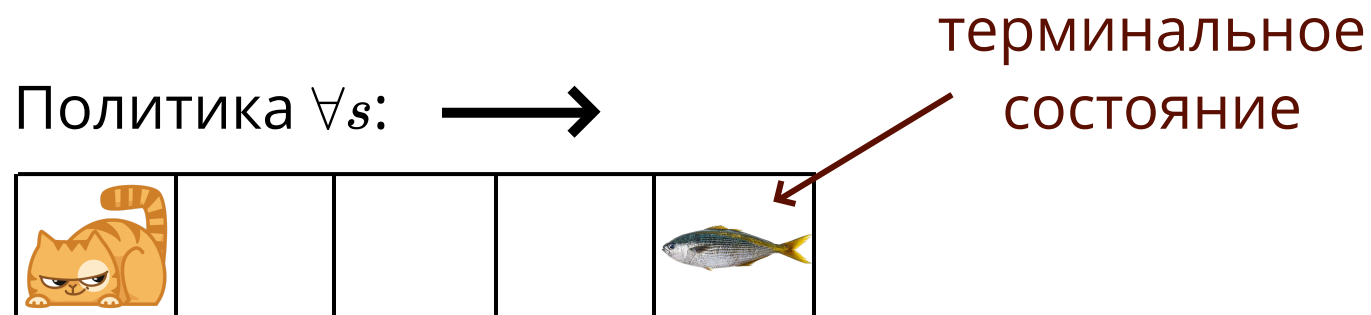
$Q$ -функция ценности:

$\longrightarrow$	$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	1
$\longleftarrow$	$\gamma^5$	$\gamma^5$	$\gamma^4$	$\gamma^3$	1
	$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

# Оценка политики

Насколько хороша политика  $\pi$ , если начать в состоянии  $s$ ?

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s]$$



$V$ -функция ценности:

$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	1
$s_0$	$s_1$	$s_2$	$s_3$	$s_4$

А что если в  $s$  "принудительно" выбрать действие  $a$ , а только затем идти по политике  $\pi$ ?

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} [R_t | s_t = s, a_t = a]$$

В сложных средах считать неудобно!

$Q$ -функция ценности:

$\gamma^4$	$\gamma^3$	$\gamma^2$	$\gamma$	1
$\gamma^5$	$\gamma^5$	$\gamma^4$	$\gamma^3$	1
$s_0$	$s_1$	$s_2$	$s_3$	$s_4$



# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

Получим рекурсивное соотношение для кумулятивной награды  $R_t$ :

$$R_t = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \gamma^2 r(s_{t+2}, a_{t+2}) + \dots$$

# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

Получим рекурсивное соотношение для кумулятивной награды  $R_t$ :

$$R_t = r(s_t, a_t) + \gamma (r(s_{t+1}, a_{t+1}) + \gamma r(s_{t+2}, a_{t+2}) + \dots)$$

# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

Получим рекурсивное соотношение для кумулятивной награды  $R_t$ :

$$R_t = r(s_t, a_t) + \gamma R_{t+1}$$

# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

Получим рекурсивное соотношение для кумулятивной награды  $R_t$ :

$$R_t = r(s_t, a_t) + \gamma R_{t+1}$$

Для  $V$ -функции:

$$V^\pi(s) = \mathbb{E}[R_t | s_t = s]$$

# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

Получим рекурсивное соотношение для кумулятивной награды  $R_t$ :

$$R_t = r(s_t, a_t) + \gamma R_{t+1}$$

Для  $V$ -функции:

$$V^\pi(s) = \mathbb{E} [r(s_t, a_t) + \gamma R_{t+1} | s_t = s]$$

# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

Получим рекурсивное соотношение для кумулятивной награды  $R_t$ :

$$R_t = r(s_t, a_t) + \gamma R_{t+1}$$

Для  $V$ -функции:

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} \mathbb{E}[R_{t+1} | s_{t+1} = s']]$$

# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

Получим рекурсивное соотношение для кумулятивной награды  $R_t$ :

$$R_t = r(s_t, a_t) + \gamma R_{t+1}$$

Для  $V$ -функции:

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V^\pi(s')]$$



# Динамическое программирование

Переформулирование *сложной* задачи в виде рекурсивной последовательности более *простых* задач.

Получим рекурсивное соотношение для кумулятивной награды  $R_t$ :

$$R_t = r(s_t, a_t) + \gamma R_{t+1}$$

Для  $V$ -функции:

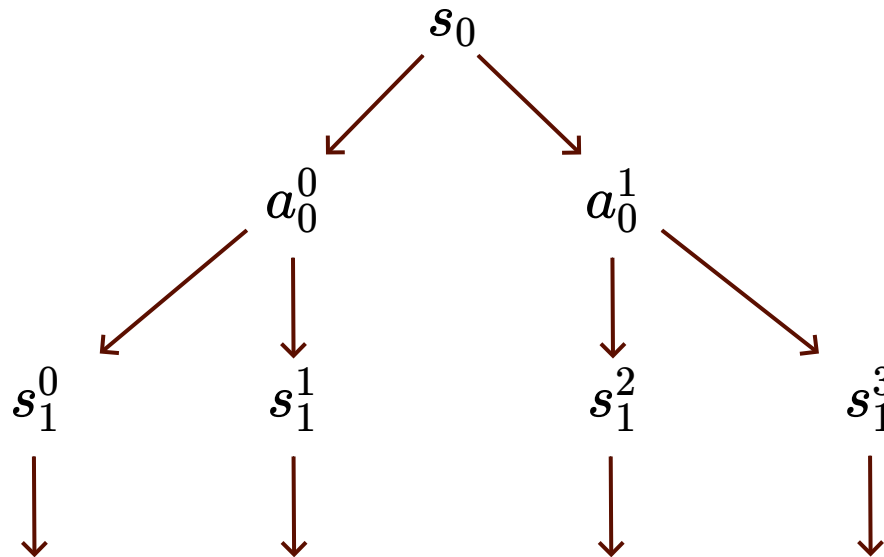
$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V^\pi(s')]$$

Для  $Q$ -функции:

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

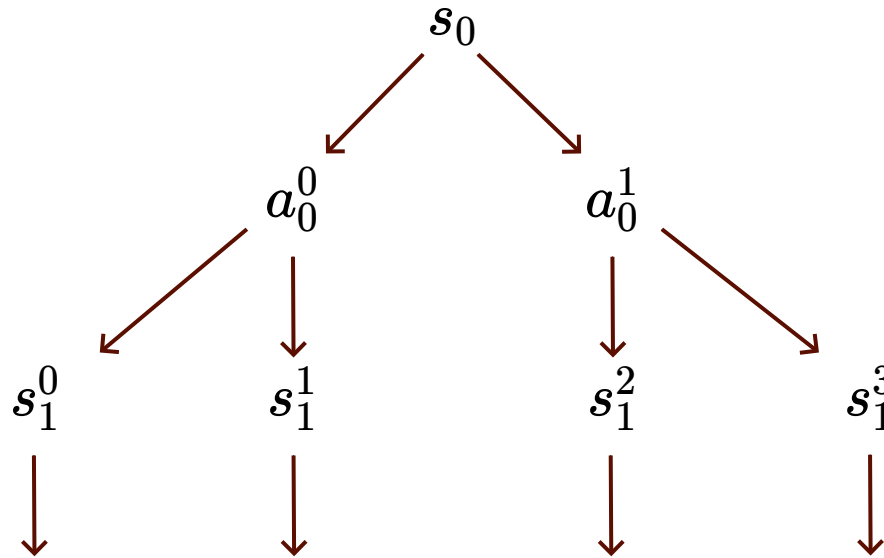
# Динамическое программирование

Если в среде состояния никогда не повторяются,  
граф этого MDP будет деревом



# Динамическое программирование

Если в среде состояния никогда не повторяются,  
граф этого MDP будет деревом



$$V^{\pi}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V^{\pi}(s')]$$

Уравнения Беллмана говорят, как посчитать ценность "задом наперед"

# Связь Q и V функций

Выражение V через Q:

$$V^{\pi}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} Q^{\pi}(s, a)$$

Выражение Q через V:

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

# Как решить Беллмана?

Как СЛАУ:

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V^\pi(s')]$$

# Как решить Беллмана?

Как СЛАУ:

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

# Как решить Беллмана?

Как СЛАУ:

$$V^\pi(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^\pi(s')$$

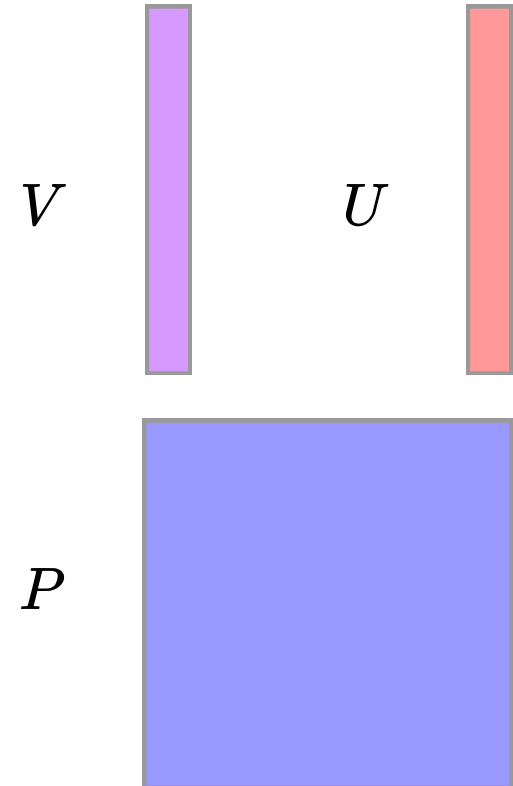
# Как решить Беллмана?

Как СЛАУ:

$$V^\pi(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^\pi(s')$$

Относительно  $V$  все линейно

$$V = U + \gamma PV$$





# Как решить Беллмана?

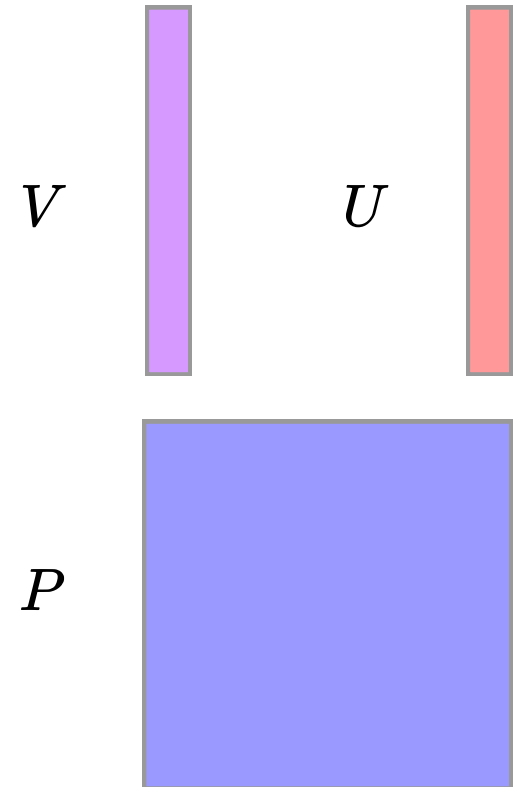
Как СЛАУ:

$$V^\pi(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^\pi(s')$$

Относительно  $V$  все линейно

$$V = U + \gamma P V$$

$$(I - \gamma P)V = U$$



# Как решить Беллмана?

Как СЛАУ:

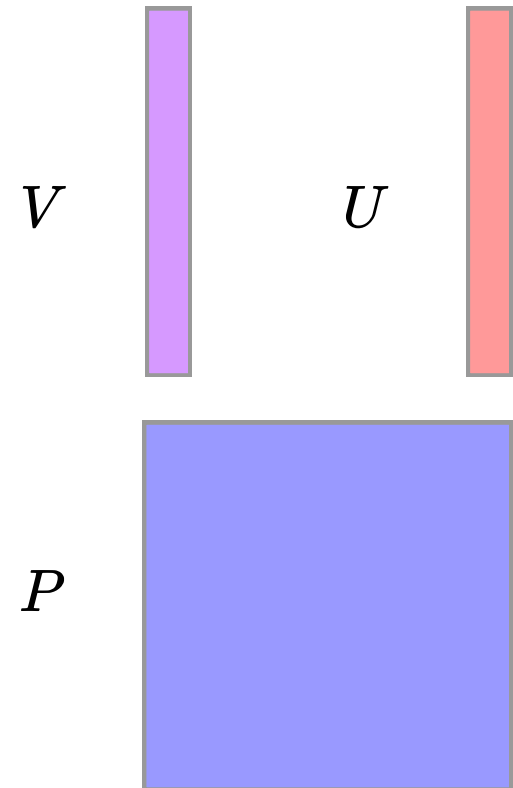
$$V^\pi(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^\pi(s')$$

Относительно  $V$  все линейно

$$V = U + \gamma P V$$

$$(I - \gamma P)V = U$$

$$V = (I - \gamma P)^{-1} U$$



# Как решить Беллмана?

Как СЛАУ:

$$V^\pi(s) = u(s) + \gamma \mathbb{E}_{s' \sim p(s'|s)} V^\pi(s')$$

Относительно  $V$  все линейно

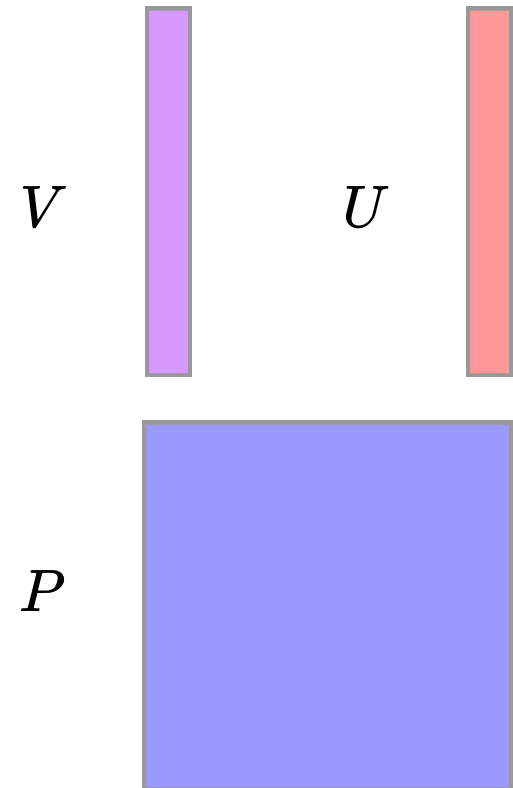
$$V = U + \gamma P V$$

$$(I - \gamma P)V = U$$

$$V = (I - \gamma P)^{-1} U$$

Дороговато будет!

Без учета  $|\mathcal{A}|$  - уже  $O(|\mathcal{S}|^3)$



# Как решить Беллмана?

Методом простой итерации:

$$V^{new} = F(V^{old})$$

$$F(V)_s = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V_s]$$

# Как решить Беллмана?

Методом простой итерации:

$$V^{new} = F(V^{old})$$

$$F(V)_s = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V_{s'}]$$

Будет ли алгоритм сходиться?

# Как решить Беллмана?

Методом простой итерации:

$$V^{new} = F(V^{old})$$

$$F(V)_s = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V_{s'}]$$

Будет ли алгоритм сходиться?      Будет, если отображение  $F$  сжимающее

# Как решить Беллмана?

Методом простой итерации:

$$V^{new} = F(V^{old})$$

$$F(V)_s = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V_{s'}]$$

Будет ли алгоритм сходиться?      Будет, если отображение  $F$  сжимающее



# Как решить Беллмана?

Методом простой итерации:

$$V^{new} = F(V^{old})$$

$$F(V)_s = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V_{s'}]$$

Будет ли алгоритм сходиться?      Будет, если отображение  $F$  сжимающее





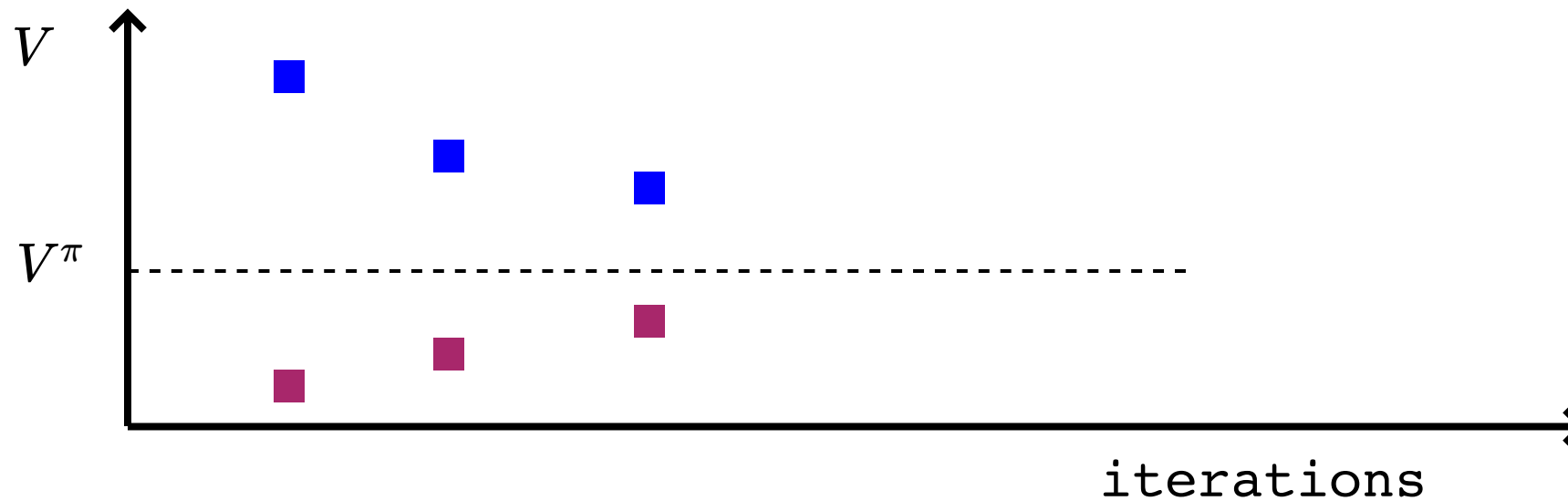
# Как решить Беллмана?

Методом простой итерации:

$$V^{new} = F(V^{old})$$

$$F(V)_s = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(s'|s, a)} V_{s'}]$$

Будет ли алгоритм сходиться? Будет, если отображение  $F$  сжимающее



# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\|F(V) - F(W)\|_{\infty} =$$

# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\|F(V) - F(W)\|_{\infty} = \|U + \gamma PV - U - \gamma PW\|_{\infty} =$$

# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\begin{aligned} \|F(V) - F(W)\|_\infty &= \|U + \gamma PV - U - \gamma PW\|_\infty = \\ &= \|\gamma P(V - W)\|_\infty \leq \end{aligned}$$

# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\begin{aligned} \|F(V) - F(W)\|_\infty &= \|U + \gamma PV - U - \gamma PW\|_\infty = \\ &= \|\gamma P(V - W)\|_\infty \leq \gamma \|P\|_\infty \|V - W\|_\infty \end{aligned}$$

# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\begin{aligned} \|F(V) - F(W)\|_\infty &= \|U + \gamma PV - U - \gamma PW\|_\infty = \\ &= \|\gamma P(V - W)\|_\infty \leq \gamma \|P\|_\infty \|V - W\|_\infty \end{aligned}$$

где матричная норма:

$$\|P\|_\infty = \max_{x: \|x\|_\infty=1} \|Px\|_\infty =$$

# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\begin{aligned} \|F(V) - F(W)\|_\infty &= \|U + \gamma PV - U - \gamma PW\|_\infty = \\ &= \|\gamma P(V - W)\|_\infty \leq \gamma \|P\|_\infty \|V - W\|_\infty \end{aligned}$$

где матричная норма:

$$\|P\|_\infty = \max_{x: \|x\|_\infty=1} \|Px\|_\infty = \max_{x: \|x\|_\infty=1} \max_i \left| \sum_j P_{ij} x_j \right|$$

# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\begin{aligned} \|F(V) - F(W)\|_\infty &= \|U + \gamma PV - U - \gamma PW\|_\infty = \\ &= \|\gamma P(V - W)\|_\infty \leq \gamma \|P\|_\infty \|V - W\|_\infty \end{aligned}$$

где матричная норма:

$$\|P\|_\infty = \max_{x: \|x\|_\infty=1} \|Px\|_\infty = \max_{x: \|x\|_\infty=1} \max_i \left| \sum_j P_{ij} x_j \right|$$


$$x_j = \text{sign}(P_{ij})$$



# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\begin{aligned} \|F(V) - F(W)\|_\infty &= \|U + \gamma PV - U - \gamma PW\|_\infty = \\ &= \|\gamma P(V - W)\|_\infty \leq \gamma \|P\|_\infty \|V - W\|_\infty \end{aligned}$$

где матричная норма:

$$\begin{aligned} \|P\|_\infty &= \max_{x: \|x\|_\infty=1} \|Px\|_\infty = \max_{x: \|x\|_\infty=1} \max_i \left| \sum_j P_{ij} x_j \right| \\ &= \max_i \left| \sum_j P_{ij} \right| = 1 \end{aligned}$$


$$x_j = \text{sign}(P_{ij})$$

# Является ли оператор Беллмана сжатием?

По норме-бесконечность:

$$\begin{aligned} \|F(V) - F(W)\|_\infty &= \|U + \gamma PV - U - \gamma PW\|_\infty = \\ &= \|\gamma P(V - W)\|_\infty \leq \gamma \|P\|_\infty \|V - W\|_\infty \end{aligned}$$

где матричная норма:

$$\begin{aligned} \|P\|_\infty &= \max_{x: \|x\|_\infty=1} \|Px\|_\infty = \max_{x: \|x\|_\infty=1} \max_i \left| \sum_j P_{ij} x_j \right| \\ &= \max_i \left| \sum_j P_{ij} \right| = 1 \end{aligned}$$


$$x_j = \text{sign}(P_{ij})$$

$$\text{ЧТД: } \|F(V) - F(W)\|_\infty < \|V - W\|_\infty$$

# Алгоритм Policy Evaluation

- инициализируем  $V(s) \quad \forall s$
- повторять:
  - $\Delta = 0$
  - для всех  $s$ :
    - $v = V(s)$
    - $V(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [r(s, a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s, a)} V(s')]$
    - $\Delta = \max(\Delta, |v - V(s)|)$
- пока  $\Delta > \epsilon$

# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

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

# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

$$V^*(s) = \max_{\pi_0, \pi_1, \dots} (\mathbb{E}_a [r(s, a) + \gamma \mathbb{E}_{s'} V^{\pi_1, \dots}(s')])$$

# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

$$V^*(s) = \max_{\pi_0} (\mathbb{E}_a [r(s, a) + \gamma \mathbb{E}_{s'} \max_{\pi_1, \dots} V^{\pi_1, \dots}(s')])$$

# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

$$V^*(s) = \max_{\pi_0} (\mathbb{E}_a [r(s, a) + \gamma \mathbb{E}_{s'} \max_{\pi_1, \dots} V^{\pi_1, \dots}(s')])$$

Относительно  $\pi_0$  решается задача вида:

$$\mathbb{E}_{a \sim \pi(\cdot|s)} y(s, a) \rightarrow \max_{\pi}$$



# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

$$V^*(s) = \max_{\pi_0} (\mathbb{E}_a [r(s, a) + \gamma \mathbb{E}_{s'} \max_{\pi_1, \dots} V^{\pi_1, \dots}(s')])$$

Относительно  $\pi_0$  решается задача вида:

$$\left\{ \begin{array}{l} \sum_i \pi_i y(s, a_i) \rightarrow \max_{\pi} \\ \pi_i \geq 0 \\ \sum_i \pi_i = 1 \end{array} \right.$$

# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

$$V^*(s) = \max_{\pi_0} (\mathbb{E}_a [r(s, a) + \gamma \mathbb{E}_{s'} \max_{\pi_1, \dots} V^{\pi_1, \dots}(s')])$$

Относительно  $\pi_0$  решается задача вида:

$$\left\{ \begin{array}{l} \sum_i \pi_i y(s, a_i) \rightarrow \max_{\pi} \\ \pi_i \geq 0 \\ \sum_i \pi_i = 1 \end{array} \right. \quad \text{Задача ЛП}$$

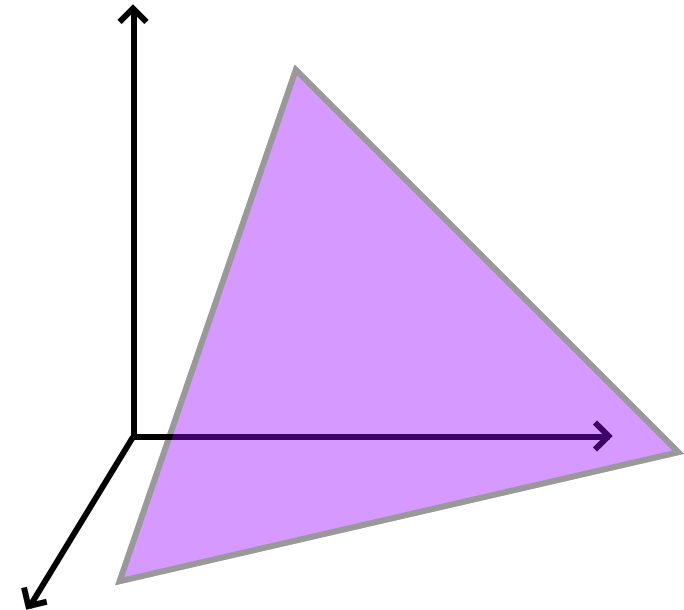
# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

$$V^*(s) = \max_{\pi_0} (\mathbb{E}_a [r(s, a) + \gamma \mathbb{E}_{s'} \max_{\pi_1, \dots} V^{\pi_1, \dots}(s')])$$

Относительно  $\pi_0$  решается задача вида:

$$\left\{ \begin{array}{l} \sum_i \pi_i y(s, a_i) \rightarrow \max_{\pi} \\ \pi_i \geq 0 \\ \sum_i \pi_i = 1 \end{array} \right. \quad \text{Задача ЛП}$$



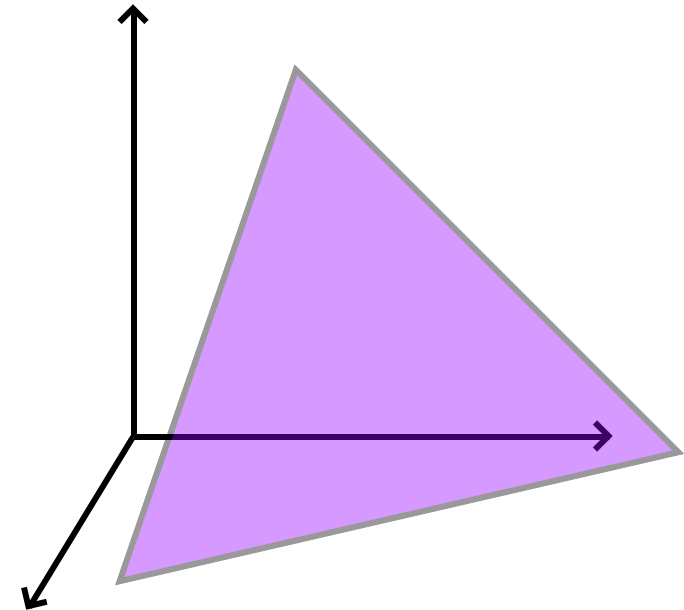
# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

$$V^*(s) = \max_{\pi_0} (\mathbb{E}_a [r(s, a) + \gamma \mathbb{E}_{s'} \max_{\pi_1, \dots} V^{\pi_1, \dots}(s')])$$

Относительно  $\pi_0$  решается задача вида:

$$\left\{ \begin{array}{l} \sum_i \pi_i y(s, a_i) \rightarrow \max_{\pi} \\ \pi_i \geq 0 \\ \sum_i \pi_i = 1 \end{array} \right. \quad \text{Задача ЛП}$$



Среди оптимальных политик всегда есть детерминированная (жадная)

# Оптимальные уравнения Беллмана

V-функция для оптимальной политики:

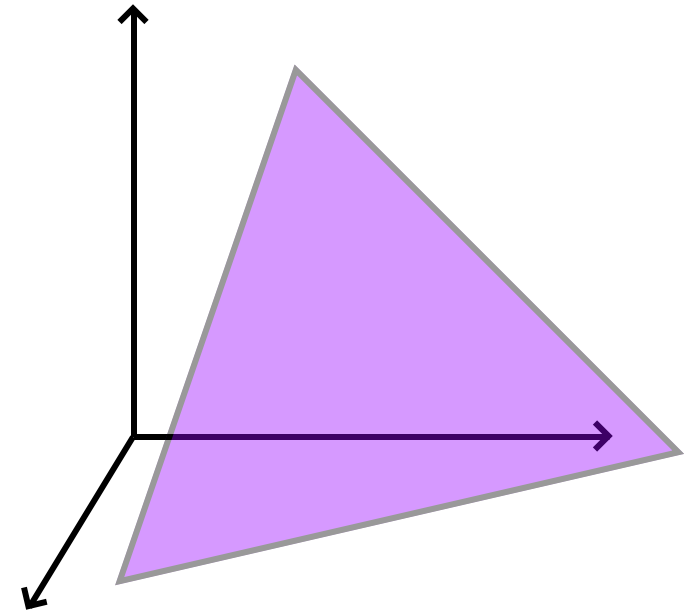
$$V^*(s) = \max_{\pi_0} (\mathbb{E}_a [r(s, a) + \gamma \mathbb{E}_{s'} \max_{\pi_1, \dots} V^{\pi_1, \dots}(s')])$$

Относительно  $\pi_0$  решается задача вида:

$$\left\{ \begin{array}{l} \sum_i \pi_i y(s, a_i) \rightarrow \max_{\pi} \\ \pi_i \geq 0 \\ \sum_i \pi_i = 1 \end{array} \right. \quad \text{Задача ЛП}$$

Оптимальное уравнение Беллмана:

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



Среди оптимальных политик всегда есть детерминированная (жадная)

# Оптимальные уравнения Беллмана

Оптимальное уравнение Беллмана для V-функции:

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

# Оптимальные уравнения Беллмана

Оптимальное уравнение Беллмана для V-функции:

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

Оптимальное уравнение Беллмана для Q-функции:

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

# Оптимальные уравнения Беллмана

Оптимальное уравнение Беллмана для V-функции:

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

Оптимальное уравнение Беллмана для Q-функции:

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

Выражение  $V^*$  через  $Q^*$ :

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



# Оптимальные уравнения Беллмана

Оптимальное уравнение Беллмана для V-функции:

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

Оптимальное уравнение Беллмана для Q-функции:

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

Выражение  $V^*$  через  $Q^*$ :

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

Выражение  $Q^*$  через  $V^*$ :

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

# Улучшение политики

**Опр.**  $\pi' \geq \pi$  если

$$V^{\pi'}(s) \geq V^{\pi}(s) \quad \forall s$$

# Улучшение политики

**Опр.**  $\pi' \geq \pi$  если

$$V^{\pi'}(s) \geq V^{\pi}(s) \quad \forall s$$

Наша стратегия обновления политики:

- пусть  $\exists s$  такой, что:  
 $\exists a : Q^{\pi}(s, a) > V^{\pi}(s)$

# Улучшение политики

**Опр.**  $\pi' \geq \pi$  если

$$V^{\pi'}(s) \geq V^{\pi}(s) \quad \forall s$$

Наша стратегия обновления политики:

- пусть  $\exists s$  такой, что:  
 $\exists a : Q^{\pi}(s, a) > V^{\pi}(s)$
- тогда  $\pi'(s) := a$ ,  
а во всех  $\tilde{s} \neq s$  ПОЛОЖИМ  $\pi'(\tilde{s}) = \pi(\tilde{s})$

# Улучшение политики

**Опр.**  $\pi' \geq \pi$  если

$$V^{\pi'}(s) \geq V^{\pi}(s) \quad \forall s$$

Наша стратегия обновления политики:

- пусть  $\exists s$  такой, что:  
 $\exists a : Q^{\pi}(s, a) > V^{\pi}(s)$
- тогда  $\pi'(s) := a$ ,  
а во всех  $\tilde{s} \neq s$  положим  $\pi'(\tilde{s}) = \pi(\tilde{s})$

В таком случае,  $\pi' \geq \pi$     **ПРОВЕРИМ**

# Улучшение политики

Наша стратегия обновления политики:

- пусть  $\exists s$  такой, что:  
 $\exists a : Q^\pi(s, a) > V^\pi(s)$
- тогда  $\pi'(s) := a$ ,  
а во всех  $\tilde{s} \neq s$  положим  $\pi'(\tilde{s}) = \pi(\tilde{s})$

# Улучшение политики

Наша стратегия обновления политики:

- пусть  $\exists s$  такой, что:  
 $\exists a : Q^\pi(s, a) > V^\pi(s)$
- тогда  $\pi'(s) := a$ ,  
а во всех  $\tilde{s} \neq s$  положим  $\pi'(\tilde{s}) = \pi(\tilde{s})$

$$V^\pi(s) \leq Q^\pi(s, \pi'(s))$$

# Улучшение политики

Наша стратегия обновления политики:

- пусть  $\exists s$  такой, что:  
 $\exists a : Q^\pi(s, a) > V^\pi(s)$
- тогда  $\pi'(s) := a$ ,  
а во всех  $\tilde{s} \neq s$  положим  $\pi'(\tilde{s}) = \pi(\tilde{s})$

$$V^\pi(s) \leq Q^\pi(s, \pi'(s))$$

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



# Улучшение политики

Наша стратегия обновления политики:

- пусть  $\exists s$  такой, что:  
 $\exists a : Q^\pi(s, a) > V^\pi(s)$
- тогда  $\pi'(s) := a$ ,  
а во всех  $\tilde{s} \neq s$  положим  $\pi'(\tilde{s}) = \pi(\tilde{s})$

$$V^\pi(s) \leq Q^\pi(s, \pi'(s))$$

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

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

# Улучшение политики

Наша стратегия обновления политики:

- пусть  $\exists s$  такой, что:  
 $\exists a : Q^\pi(s, a) > V^\pi(s)$
- тогда  $\pi'(s) := a$ ,  
а во всех  $\tilde{s} \neq s$  положим  $\pi'(\tilde{s}) = \pi(\tilde{s})$

$$V^\pi(s) \leq Q^\pi(s, \pi'(s))$$

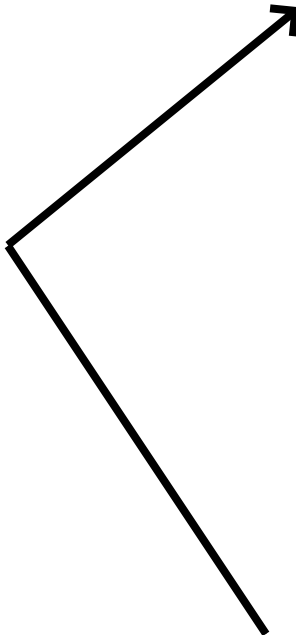
$$= r(s, \pi'(s)) + \mathbb{E}_{s'} V^\pi(s') \leq$$

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

$$\leq \dots \leq V^{\pi'}$$

ЧТД

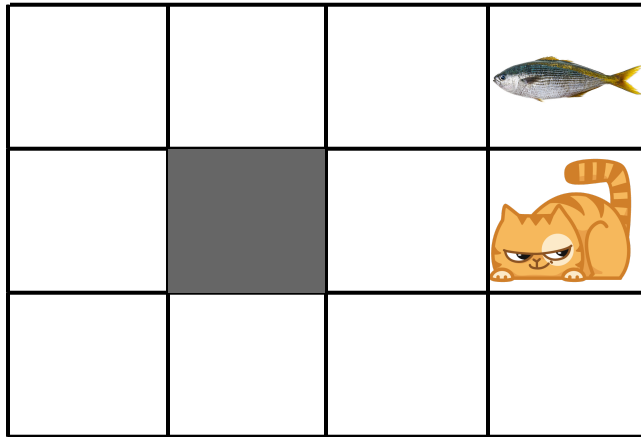
# Алгоритм Policy Iteration

- 
- инициализируем  $V(s)$ ,  $\pi(s) \quad \forall s$
  - оценить  $V$  для политики  $\pi$  методом PE
  - $stop = True$
  - для всех  $s$ :
    - $a = \pi(s)$
    - $\pi(s) = \arg \max_a [r(s, a) + \mathbb{E}_{s'} V(s')]$
    - если  $a \neq \pi(s)$ :
      - $stop = False$
  - если не stop

# Алгоритм Value Iteration

- инициализируем  $V(s) \quad \forall s$
- повторять:
  - $\Delta = 0$
  - для всех  $s$ :
    - $v = V(s)$
    - $V(s) = \max_a [r(s, a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} V(s')]$
    - $\Delta = \max(\Delta, |v - V(s)|)$
- пока  $\Delta > \epsilon$

# Пример Value Iteration



дружественная рыбка + 1

вражеский кот - 1