RL 1-04

Policy Iteration

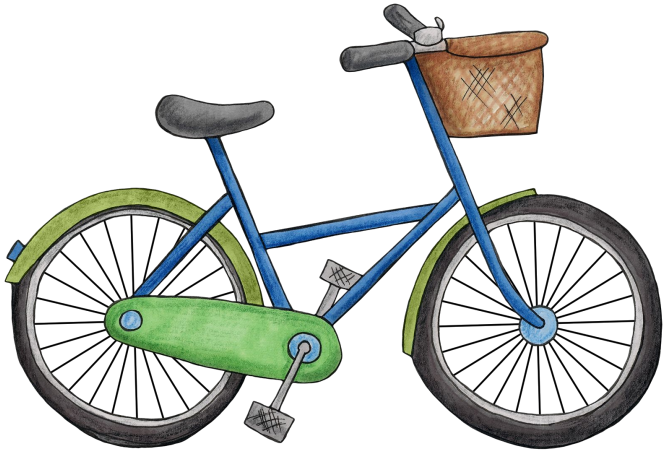
Начнем в 20:01

otus.ru

Можно ли **осознанно** разучиться ездить на велосипеде?

Warmup

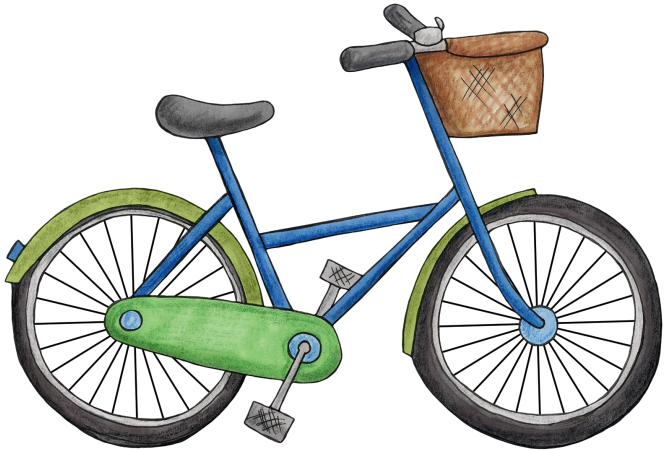
Можно ли **осознанно** разучиться ездить на велосипеде?



Model-free алгоритм – …

Warmup

Можно ли **осознанно** разучиться ездить на велосипеде?

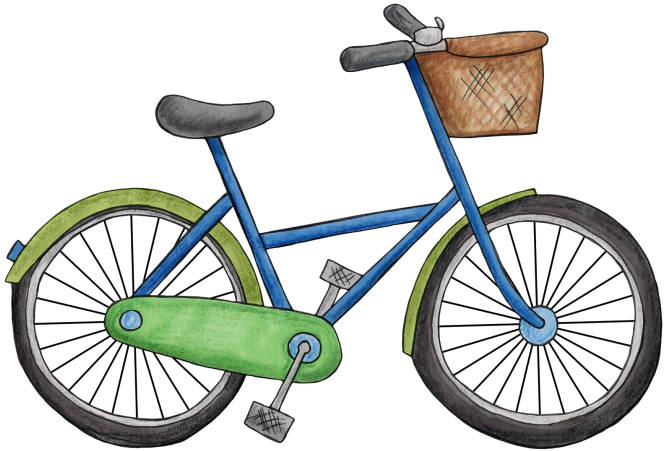


Model-free алгоритм не использует и не пытается выучить модель динамики среды $p(s' | s, a)$.

Warmup

“Обратный велосипед”

https://www.youtube.com/watch?v=MFzDaBzBlL0



Классификация

Классификация №1

1. Model free
2. Model based

Классификация №2

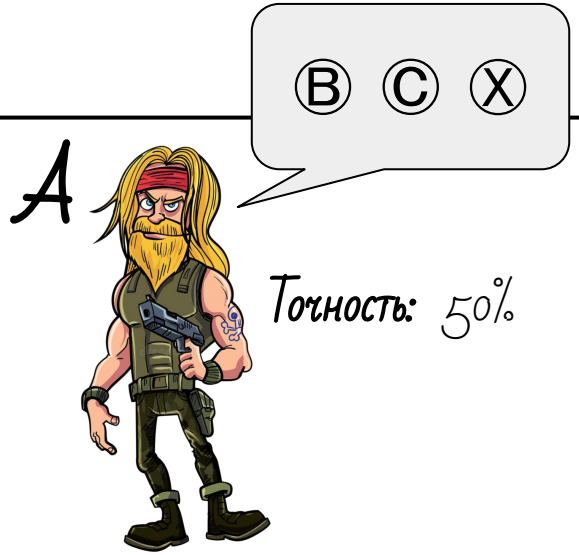
1. Value based
2. Policy based

Классификация №3

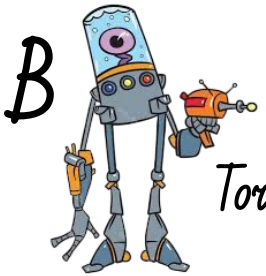
1. On-policy
2. Off-policy

Zagara

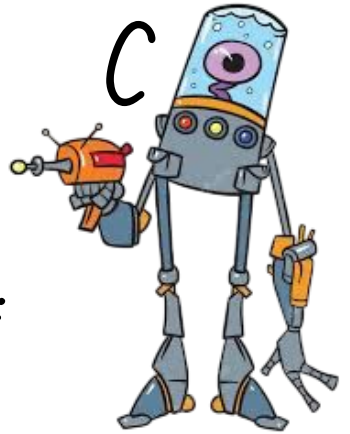
Что будем делать?



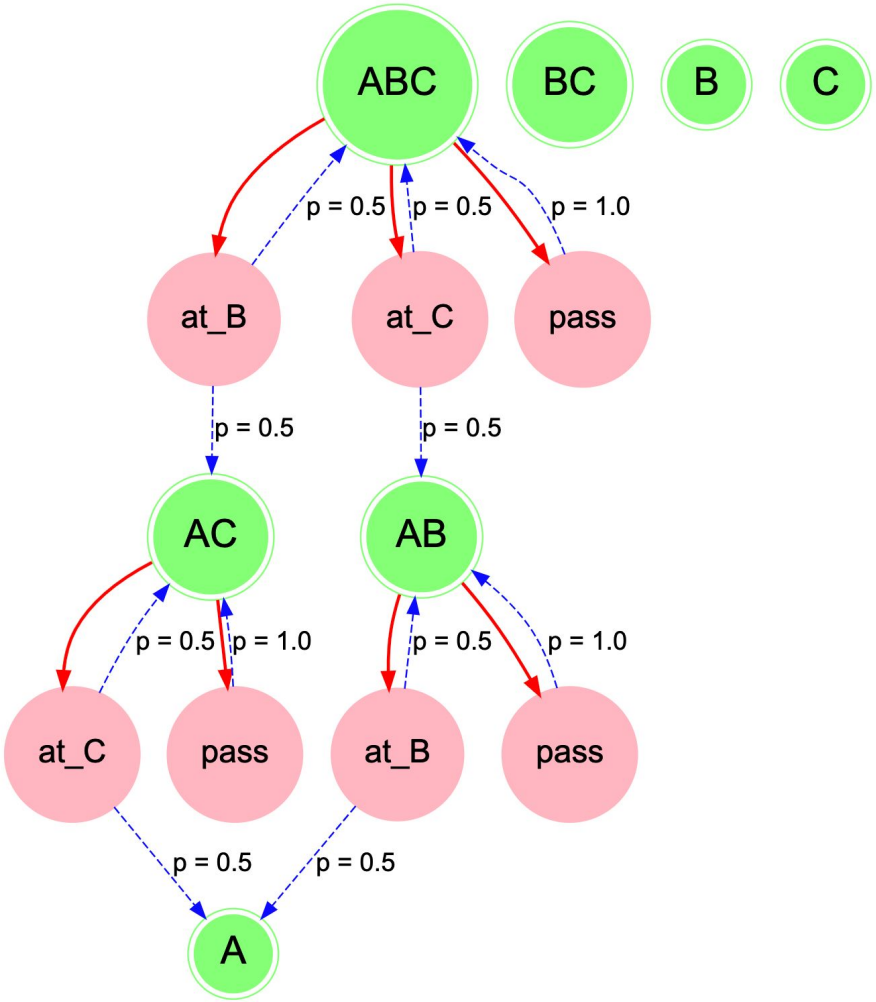
Точность: 50%

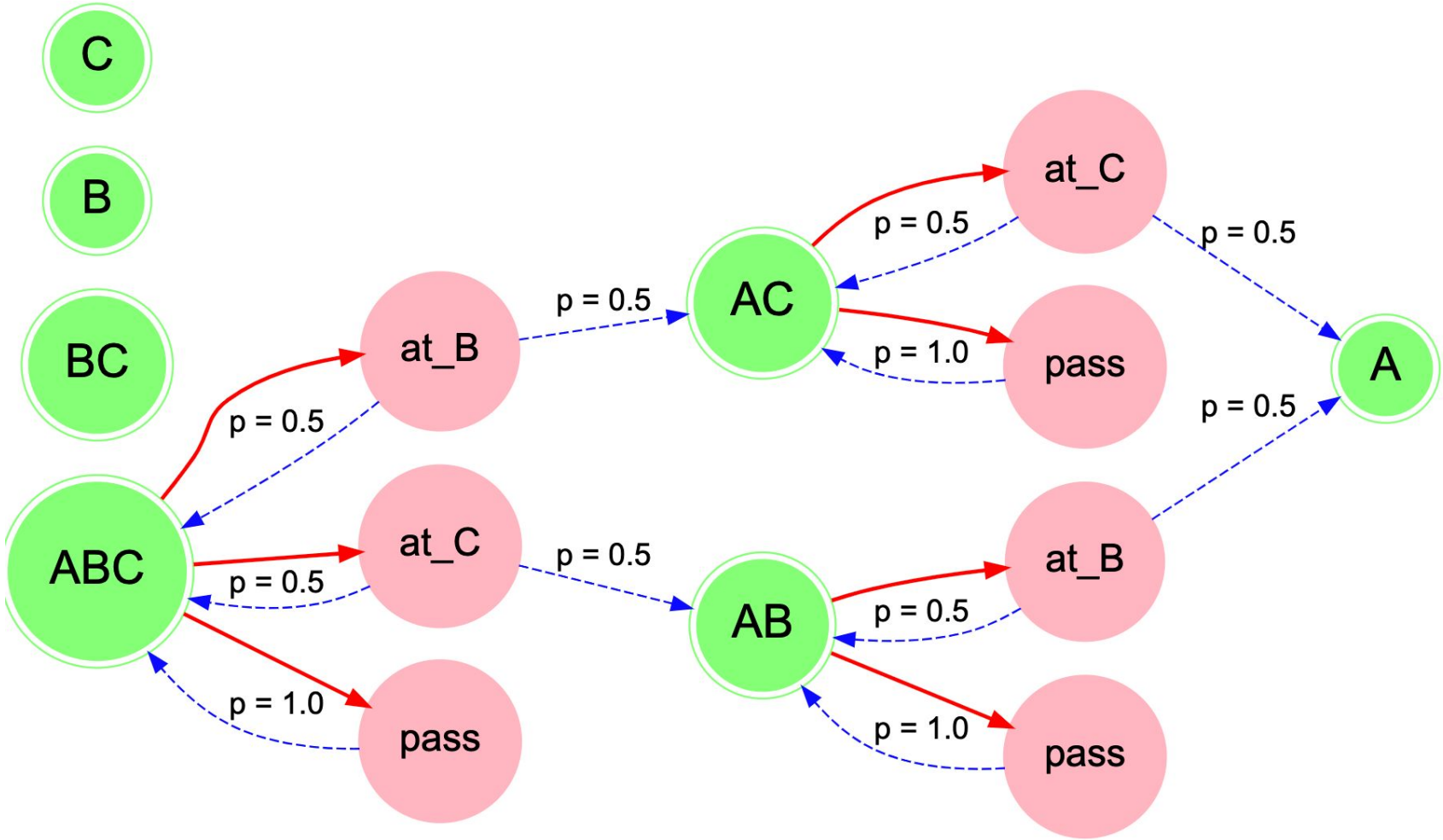


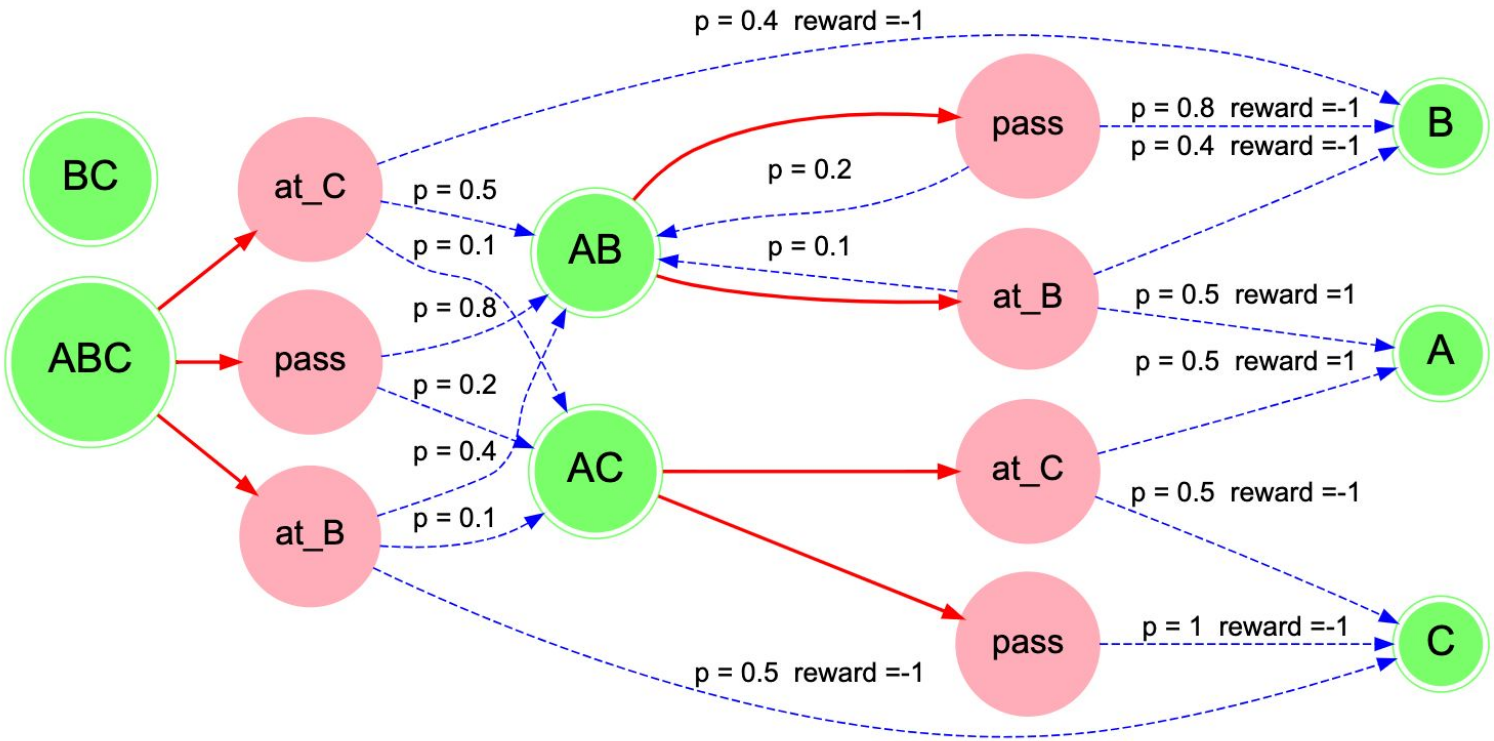
Точность: 80%



Точность:
100%


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

Практика

Модификация

Что мы делали?

- Меняли Q → V
- По Q определяли π

V-функция:

	ABC	AB	AC	BC	A	B	C
0	0.0	0.1	0.0	0.0	0.0	0.0	0.0

Q-функция:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	0.1	NaN	-0.8
ABC	-0.5	-0.4	0.0
AC	NaN	0.0	-1.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

Политика:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	1.0	NaN	0.0
ABC	0.0	0.0	1.0
AC	NaN	1.0	0.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

Модификация

Что мы делали?

- Меняли Q → V
- По Q определяли π

V-функция:

	ABC	AB	AC	BC	A	B	C
0	0.0	0.1	0.0	0.0	0.0	0.0	0.0

Q-функция:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	0.1	NaN	-0.8
ABC	-0.5	-0.4	0.0
AC	NaN	0.0	-1.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

Политика:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	1.0	NaN	0.0
ABC	0.0	0.0	1.0
AC	NaN	1.0	0.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

Модификация

Что мы делали?

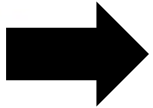
- Меняли Q → V
- По Q определяли π

V-функция:

	ABC	AB	AC	BC	A	B	C
0	0.0	0.1	0.0	0.0	0.0	0.0	0.0

Q-функция:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	0.1	NaN	-0.8
ABC	-0.5	-0.4	0.0
AC	NaN	0.0	-1.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN



Политика:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	1.0	NaN	0.0
ABC	0.0	0.0	1.0
AC	NaN	1.0	0.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

Модификация

Что мы делали?

- Меняли Q → V
- По Q определяли π

V-функция:

	ABC	AB	AC	BC	A	B	C
0	0.0	0.1	0.0	0.0	0.0	0.0	0.0

Q-функция:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	0.1	NaN	-0.8
ABC	-0.5	-0.4	0.0
AC	NaN	0.0	-1.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

Политика:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	1.0	NaN	0.0
ABC	0.0	0.0	1.0
AC	NaN	1.0	0.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

Модификация

Что мы делали?

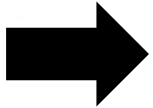
- Меняли Q → V
- По Q определяли π

V-функция:

	ABC	AB	AC	BC	A	B	C
0	0.0	0.1	0.0	0.0	0.0	0.0	0.0

Q-функция:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	0.1	NaN	-0.8
ABC	-0.5	-0.4	0.0
AC	NaN	0.0	-1.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN



Политика:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	1.0	NaN	0.0
ABC	0.0	0.0	1.0
AC	NaN	1.0	0.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

А так можем?

Что мы делали?

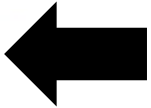
- Меняли Q → V
- По Q определяли π

V-функция:

	ABC	AB	AC	BC	A	B	C
0	0.0	0.1	0.0	0.0	0.0	0.0	0.0

Q-функция:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	0.1	NaN	-0.8
ABC	-0.5	-0.4	0.0
AC	NaN	0.0	-1.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN



Политика:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	1.0	NaN	0.0
ABC	0.0	0.0	1.0
AC	NaN	1.0	0.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

А так можем?

Другой вариант

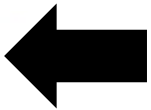
- Менять π
- По π определять Q → V

V-функция:

	ABC	AB	AC	BC	A	B	C
0	0.0	0.1	0.0	0.0	0.0	0.0	0.0

Q-функция:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	0.1	NaN	-0.8
ABC	-0.5	-0.4	0.0
AC	NaN	0.0	-1.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN



Политика:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	1.0	NaN	0.0
ABC	0.0	0.0	1.0
AC	NaN	1.0	0.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

А так можем?

Другой вариант

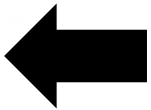
- Менять π
- По π определять Q → V
- По Q обновить π

V-функция:

	ABC	AB	AC	BC	A	B	C
0	0.0	0.1	0.0	0.0	0.0	0.0	0.0

Q-функция:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	0.1	NaN	-0.8
ABC	-0.5	-0.4	0.0
AC	NaN	0.0	-1.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN



Политика:

	at_B	at_C	pass
A	NaN	NaN	NaN
AB	1.0	NaN	0.0
ABC	0.0	0.0	1.0
AC	NaN	1.0	0.0
B	NaN	NaN	NaN
BC	NaN	NaN	NaN
C	NaN	NaN	NaN

№5. ДП: Policy-Iteration

Политика:

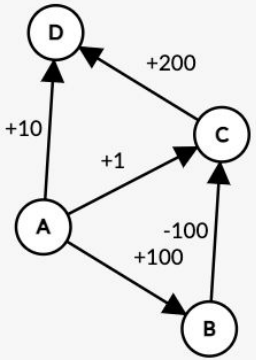
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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:

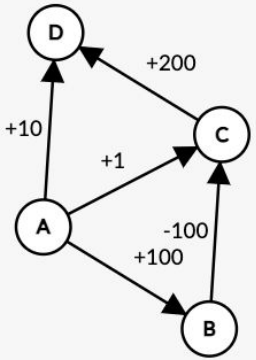
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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

Выход: $V_k(s)$

№5. ДП: Policy-Iteration

Оцениваем политику

Политика:

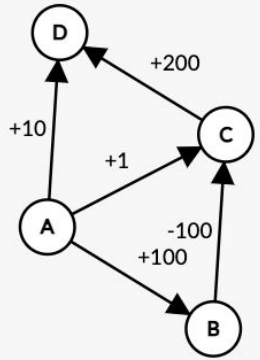
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

V:

A	B	C	D
0	0	0	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:

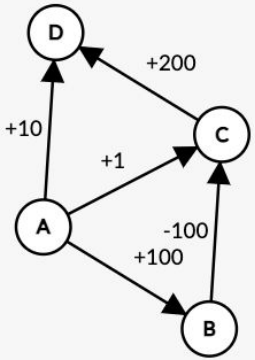
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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:

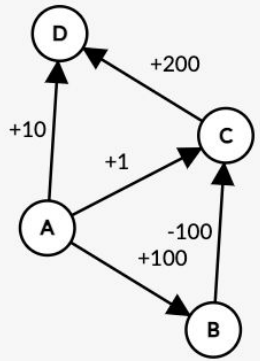
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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:

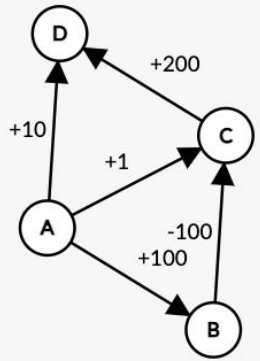
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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:

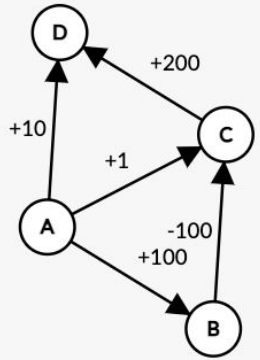
	A	B	C	D
A	✗	1.0	0.0	0.0
B	✗	✗	1.0	✗
C	✗	✗	✗	1.0
D	✗	✗	✗	✗

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:

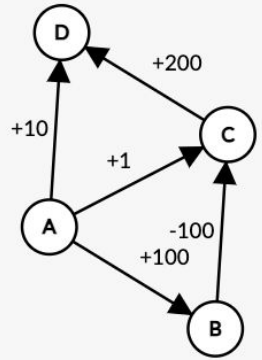
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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:

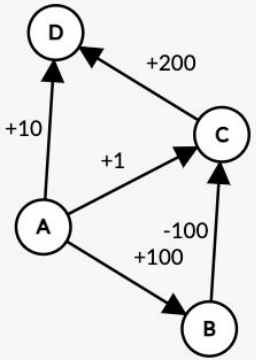
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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

Оцениваем политику

Политика:

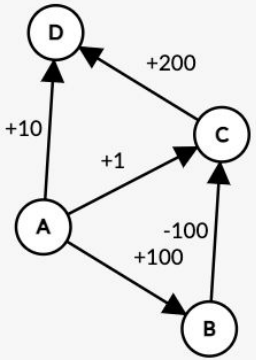
	A	B	C	D
A	✗	1.0	0.0	0.0
B	✗	✗	1.0	✗
C	✗	✗	✗	1.0
D	✗	✗	✗	✗

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:

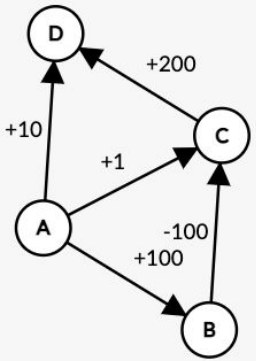
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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:

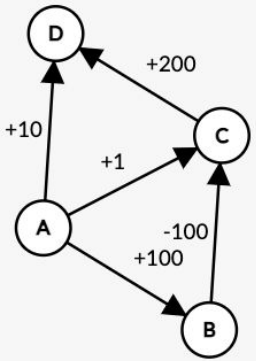
	A	B	C	D
A	❌	1.0	0.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
D	❌	❌	❌	❌

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	❌	1.0	1.0	0.0
B	❌	❌	1.0	❌
C	❌	❌	❌	1.0
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:** $\pi_k \equiv \pi_{k+1}$

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: $V_k(s)$

Model-free

Проблема

- А если мы не знаем **динамику среды**?
- … Или нам ее сложно описать

Не знаем меткость игроков

$$p(s' \mid s, a)$$

```
dynamics = {
    'ABC': {
        'at_B': {
            'C': 0.5, # Попали, убили B. C гарантированно убил нас.
            'AC': 0.5 * 0.2, # Не попали. B промахнулся. C убил B.
            'AB': 0.5 * 0.8 # Не попали. B попал.
        },
        'at_C': {
            'B': 0.5 * 0.8, # Мы попали. B стрелял в нас и попал
            'AB': 0.5 * 0.2 + 0.5 * 0.8, # Мы попали. B стрелял в нас и промахн
            'AC': 0.5 * 0.2, # Мы промахнулись. B стрелял в C и не попал. C уби
        },
        'pass': {
            'AB': 0.8, # B попал в C
            'AC': 0.2 # B  не попал в C
        }
    },
```

В каком месте проблема?

- А если мы не знаем **динамику среды**?
- … Или нам ее сложно описать

$$p(s' \mid s, a)$$

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

В каком месте проблема?

- А если мы не знаем **динамику среды**?
- … Или нам ее сложно описать

$$p(s' \mid s, a)$$

Алгоритм 7: Value Iteration

Вход: ε — критерий останова

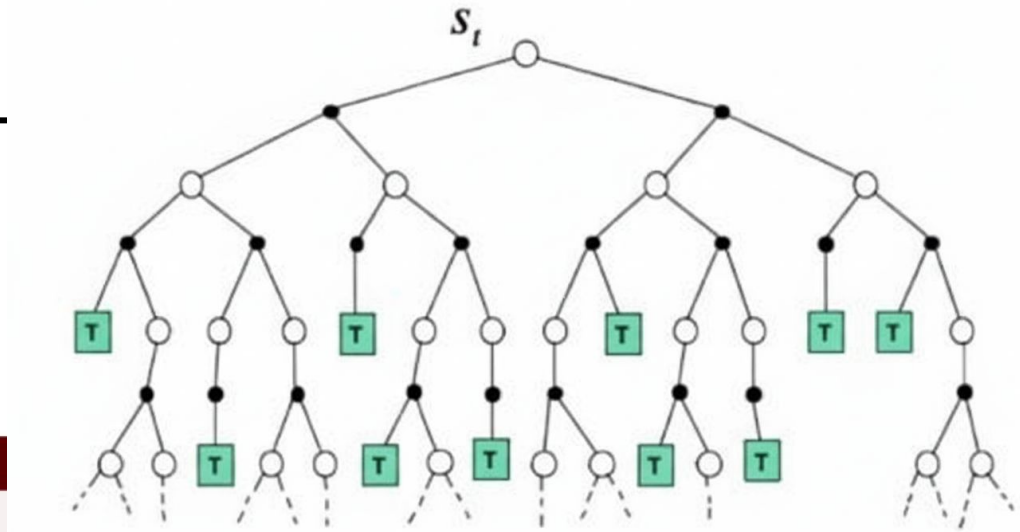
Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

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

Не сможем посчитать мат. ожидание по всем переходам

Сейчас:



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

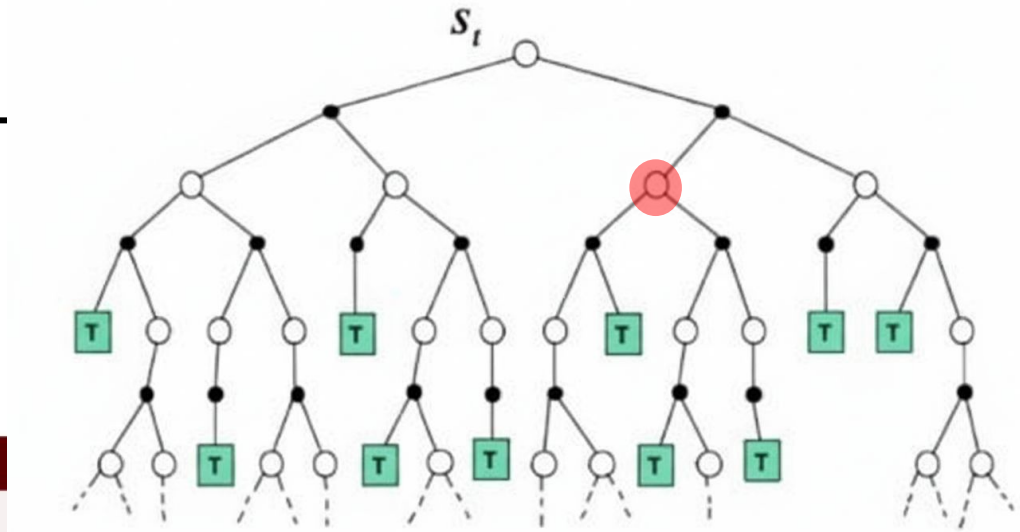
Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

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

Не сможем посчитать мат. ожидание по всем переходам

Сейчас:



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

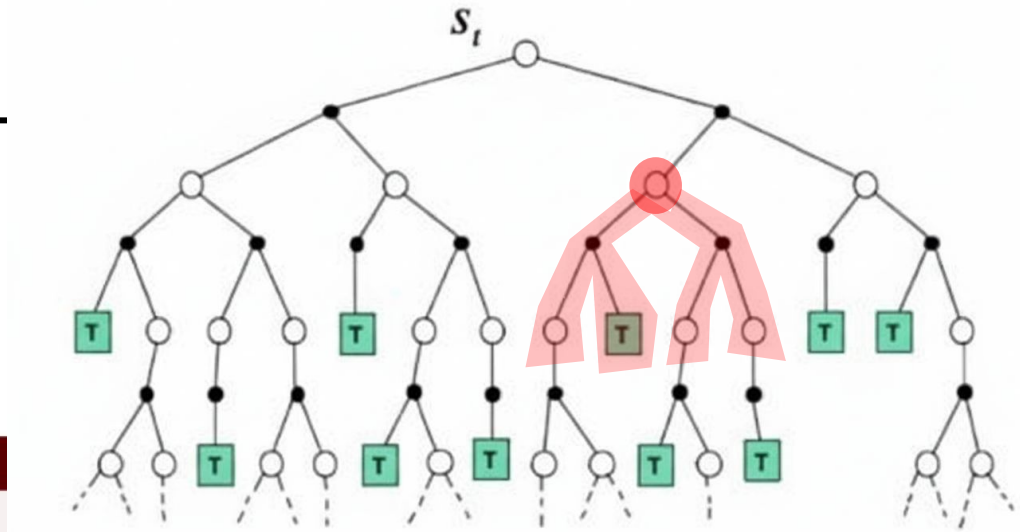
Инициализируем $V_0(s)$ произвольно для всех $s \in \mathcal{S}$
На k-ом шаге:

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

Не сможем посчитать мат. ожидание по всем переходам

Сейчас:



Алгоритм 7: Value Iteration

Вход: ε — критерий останова

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

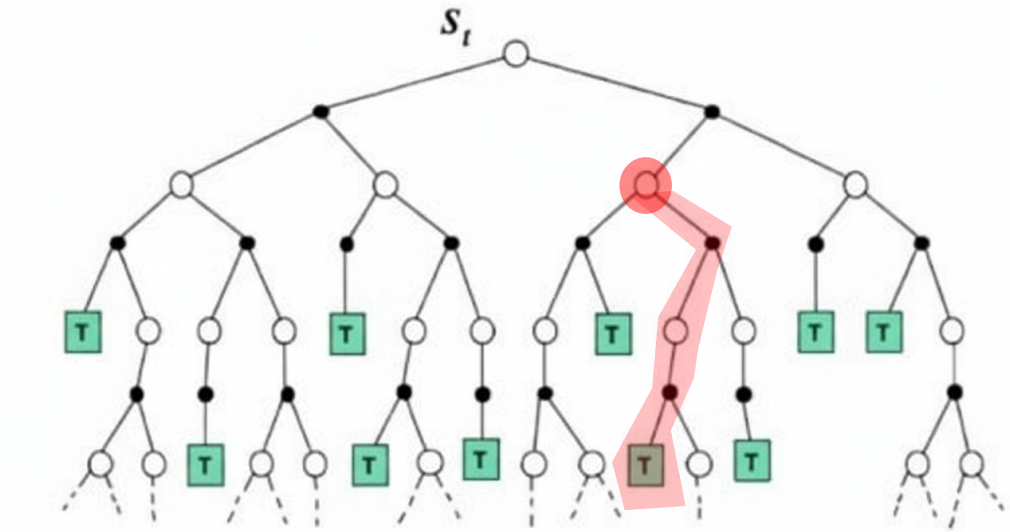
1. для всех s: $V_{k+1}(s) \coloneqq \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V_k(s')]$
2. **критерий останова:** $\max_s |V_{k+1}(s) - V_k(s)| < \varepsilon$

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

Не сможем посчитать мат. ожидание по всем переходам

Идея:

Играть до конца и усреднять!



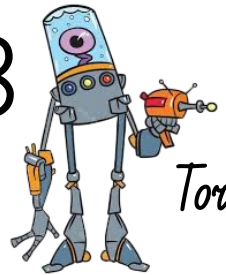
Монте-Карло

Метод Монте-Карло

A

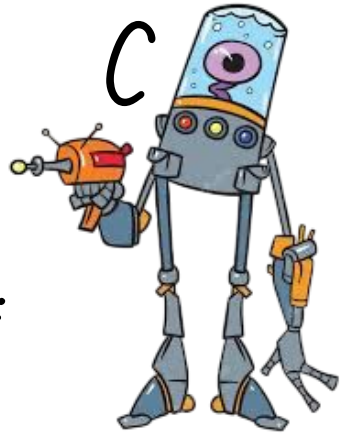
Точность: 50%

B



Точность: 80%

C



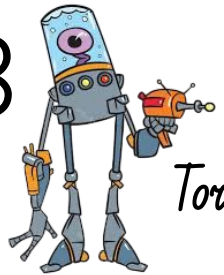
Точность: 100%

Шаг 1: Запасаемся агентами

A

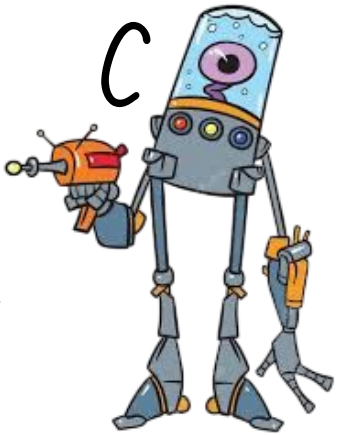
Точность: 50%

B



Точность: 80%

C

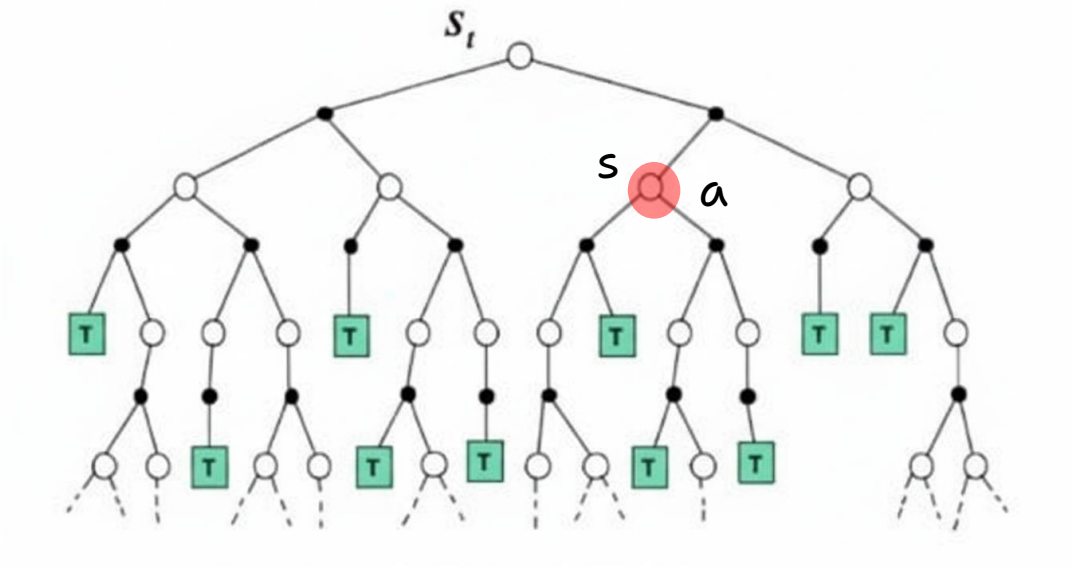


Точность:
100%

Шаг 2: Можем оценить Q средним

$$Q^{\pi_k}(s,a) \approx$$

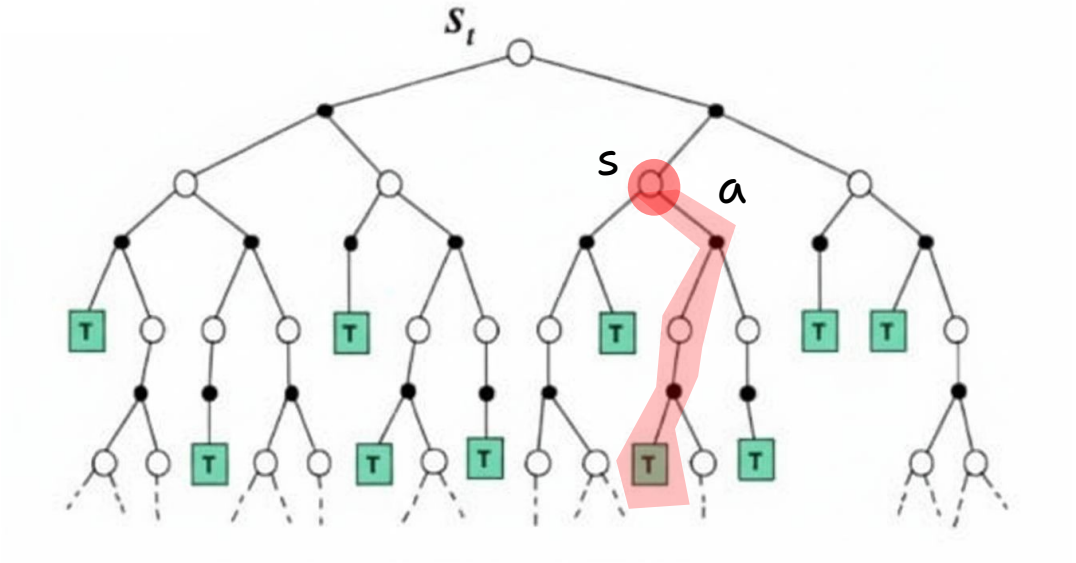
Для фиксированной политики



Шаг 2: Можем оценить Q средним

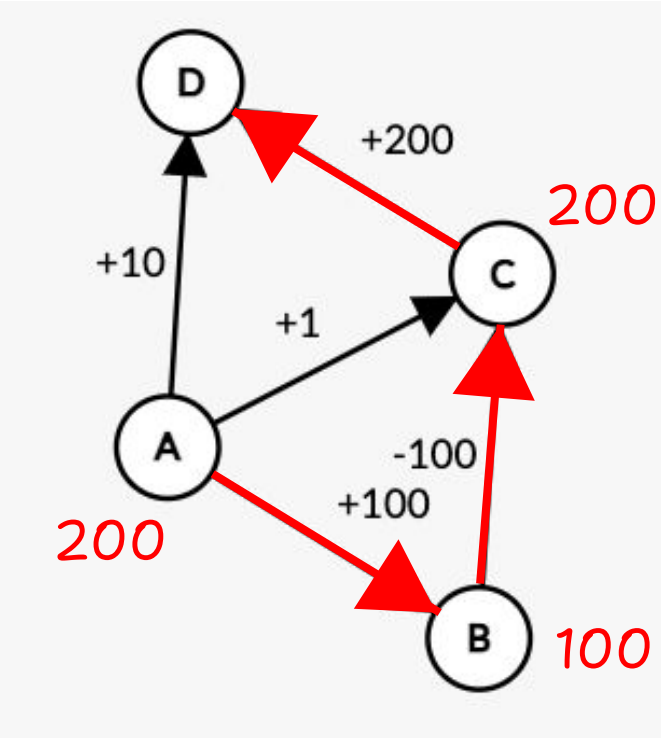
$$Q^{\pi_k}(s,a) \approx \frac{1}{N}\sum_{i=0}^{N} R(\mathcal{T}_i), \quad \mathcal{T}_i \sim \pi_k \mid s_0 = s, a_0 = a$$

Для фиксированной политики



Модификация 1

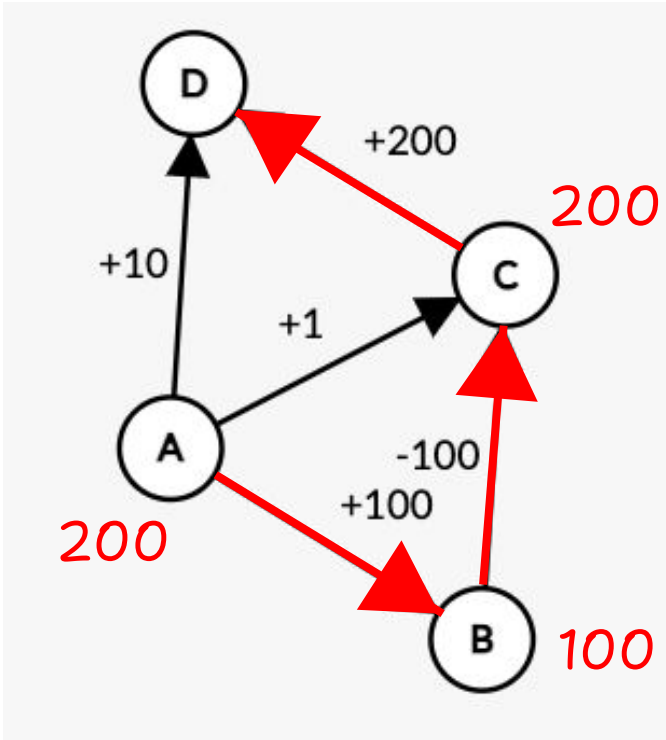
(s, a): [(r, s'), ...]



Модификация 1

(s, a): [(r, s'), ...]

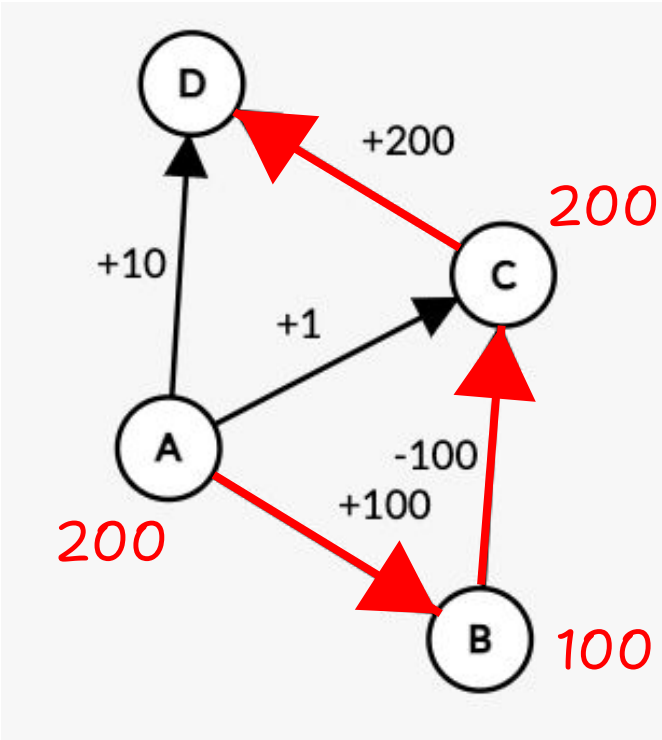
Первый **off-policy** алгоритм!



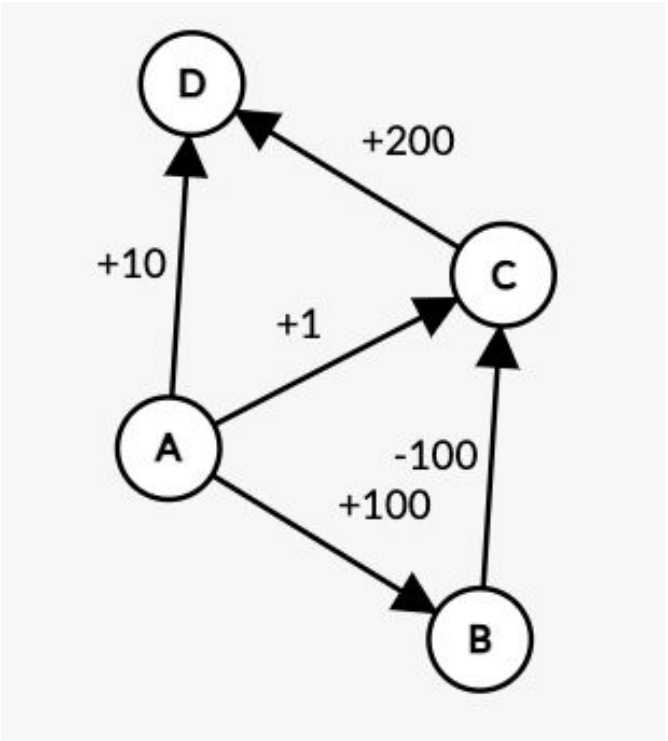
Off-policy

- **Target-policy** – стратегия, стремящаяся максимизировать суммарную награду
- **Behavior policy** – стратегия сбора данных

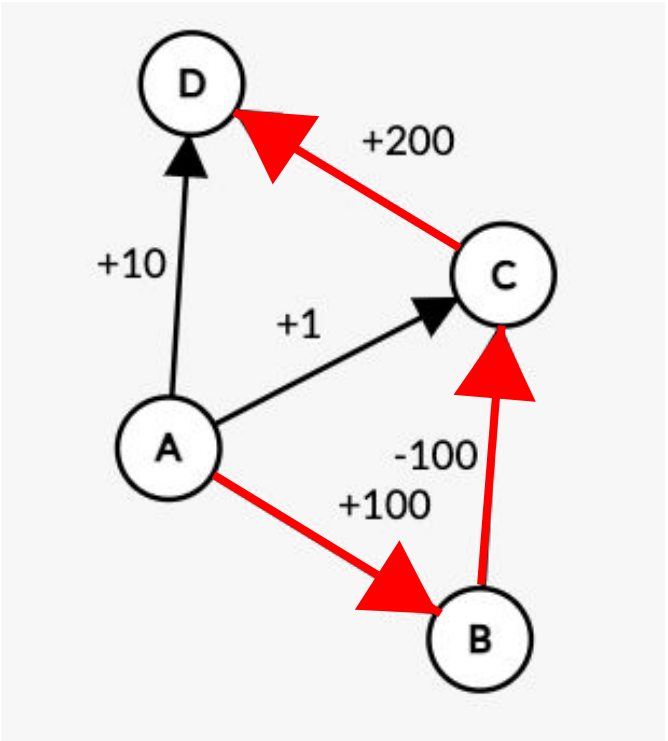
Off-policy алгоритм может обучаться на данных от произвольной behavior policy



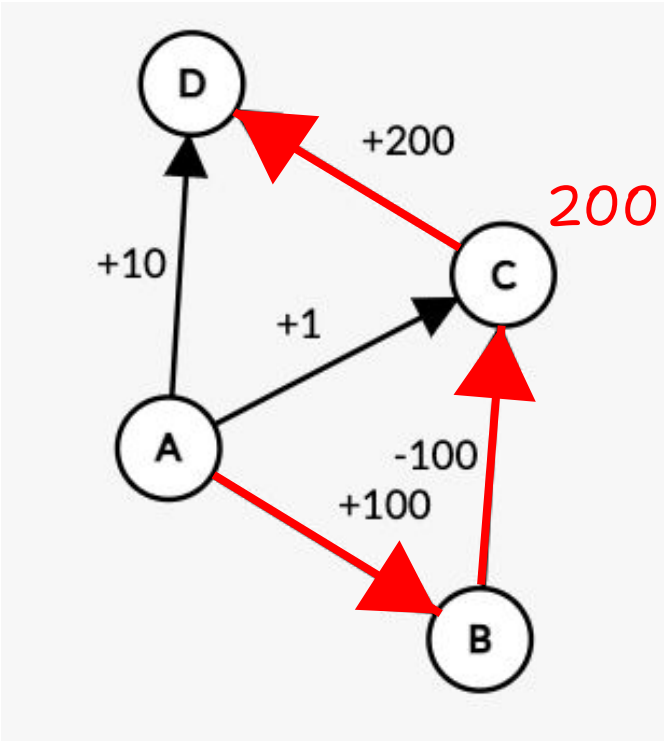
Пример



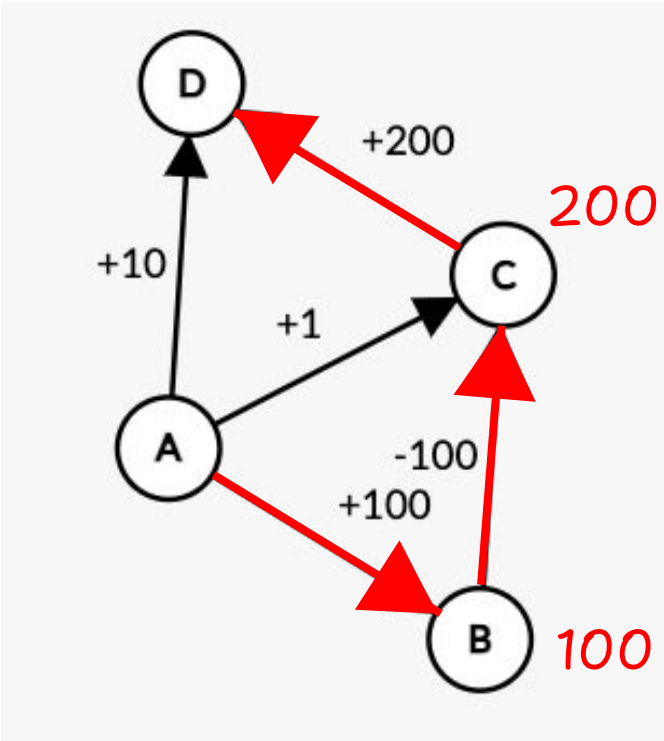
Пример



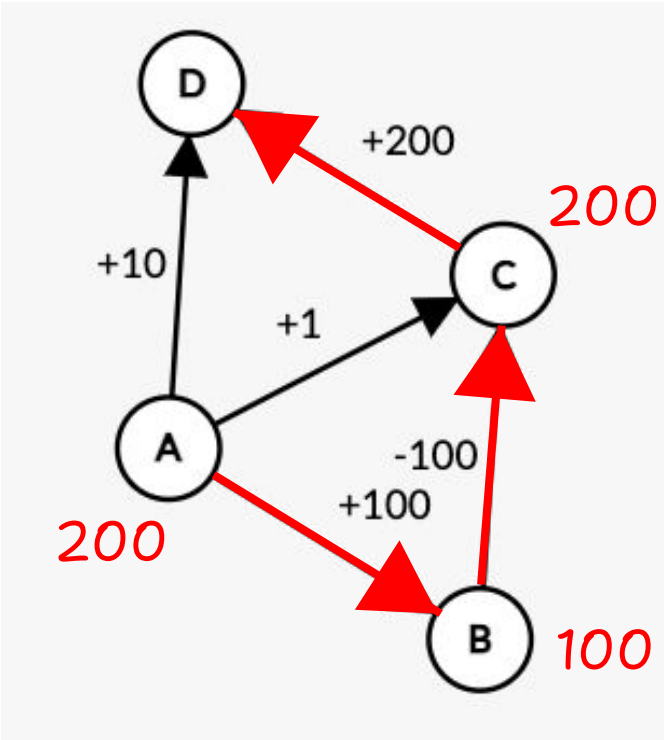
Пример



Пример

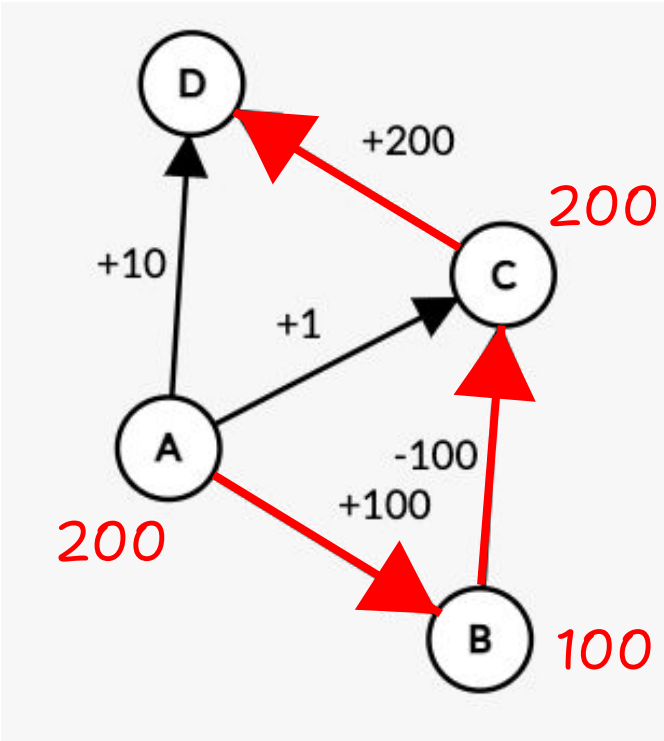


Пример



reward-to-go

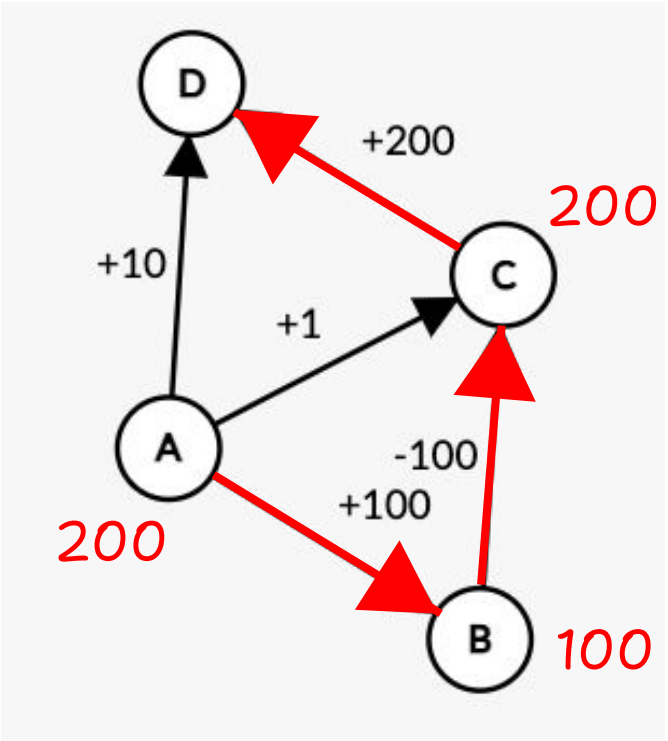
$$R_t := R(\mathcal{T}_{t:}) = \sum_{\hat{t} \geq t} \gamma^{\hat{t}-t} r_{\hat{t}}$$



reward-to-go

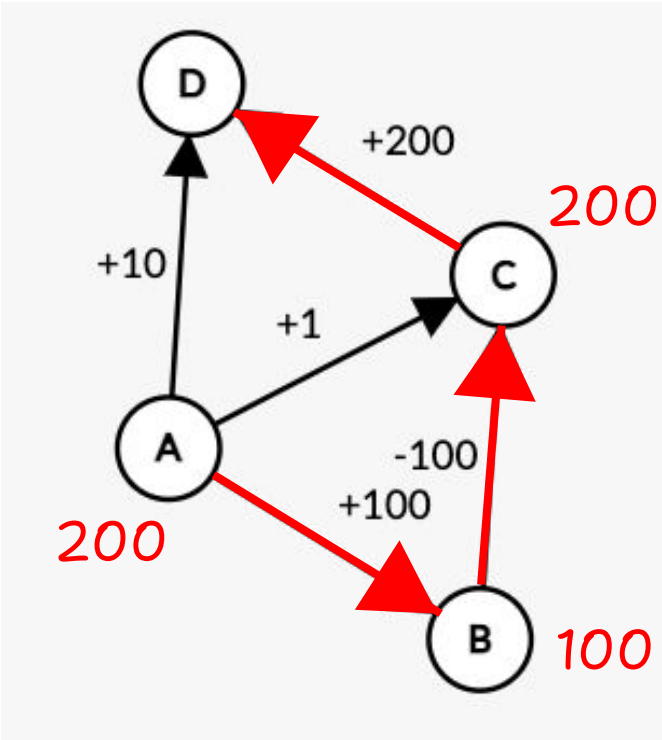
$$R_t := R(\mathcal{T}_{t:}) = \sum_{\hat{t} \geq t} \gamma^{\hat{t}-t} r_{\hat{t}}$$

$$Q^{\pi_k}(s,a) \approx \frac{1}{N} \sum_{i=0}^{N} R(\mathcal{T}_i), \quad \mathcal{T}_i \sim \pi_k \mid s_0 = s, a_0 = a$$



Проблемы

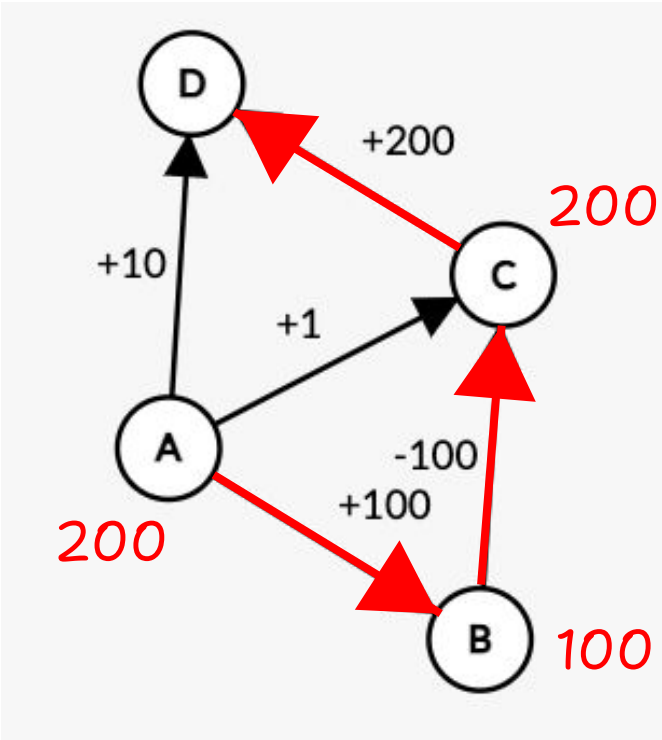
1. На этом мы закончили …



Проблемы

1. На этом мы закончили …

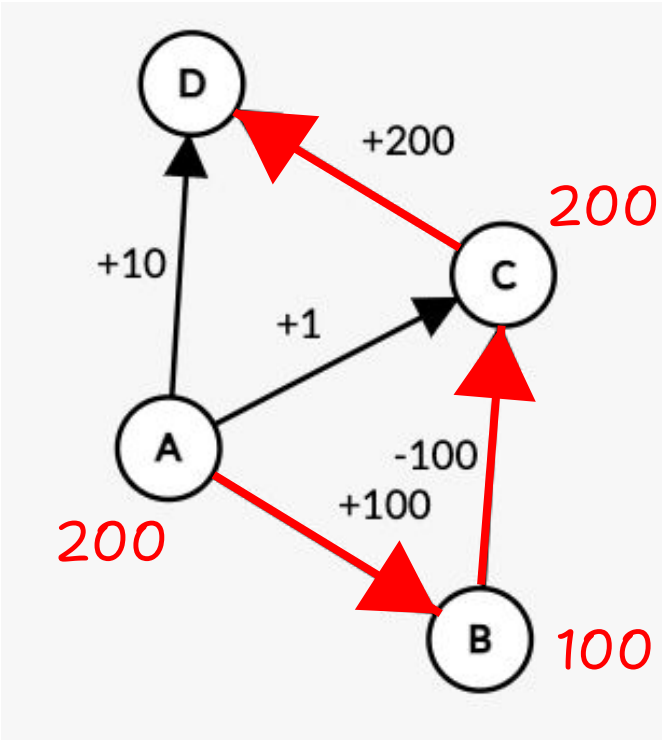
   Наша политика не включала ребро AC, поэтому мы не скорректировали политику



Модификация 1

epsilon-жадная стратегия

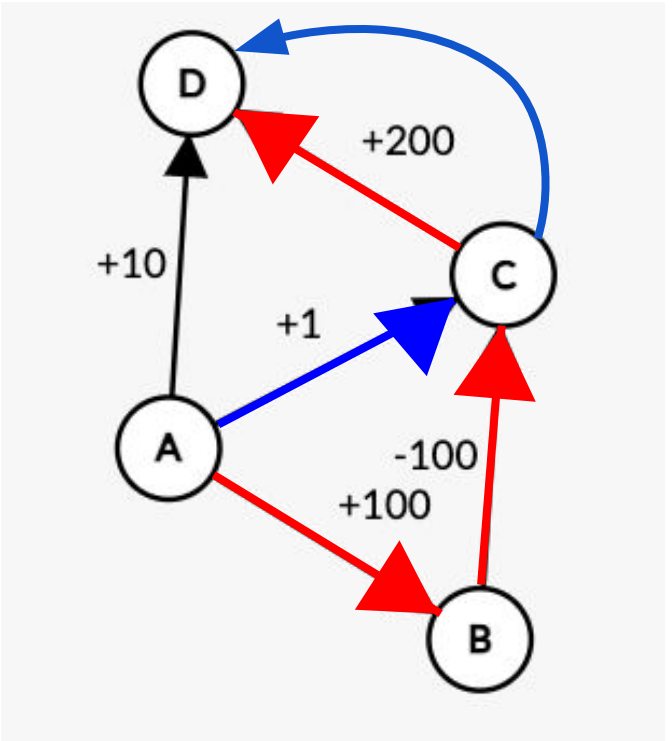
с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$



Модификация 1

epsilon-жадная стратегия

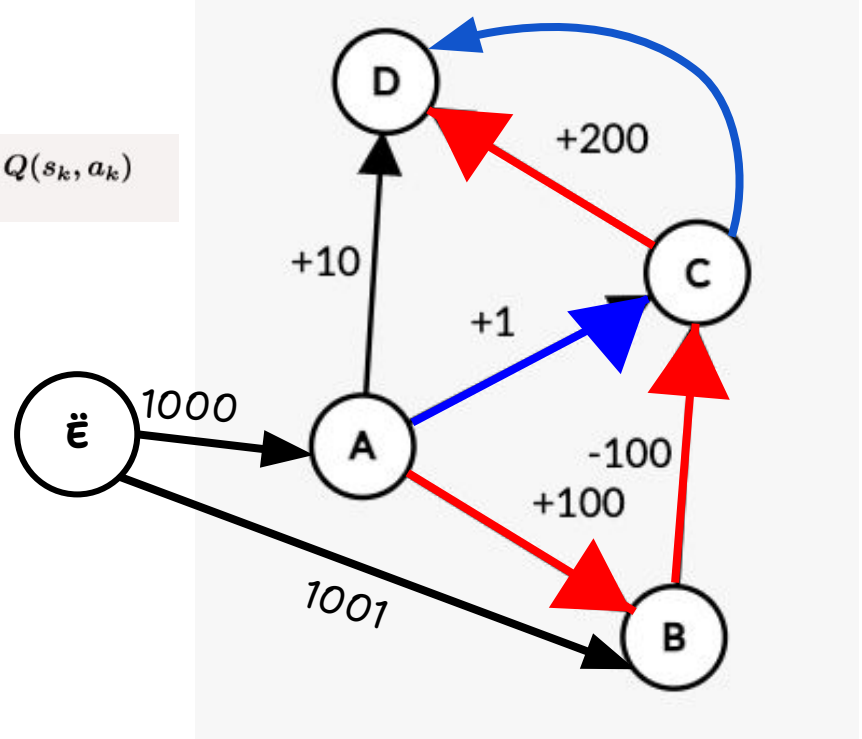
с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$



Модификация 1

epsilon-жадная стратегия

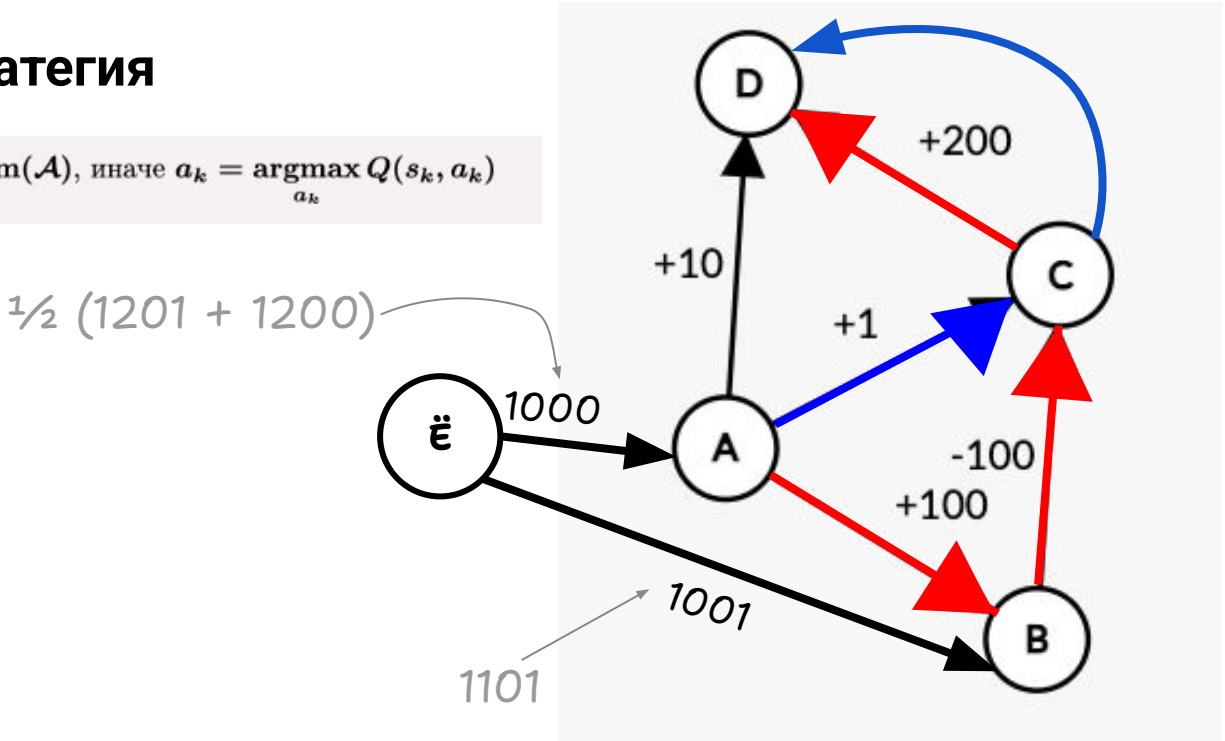
с вероятностью ϵ играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$



Модификация 1

epsilon-жадная стратегия

с вероятностью ϵ играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$

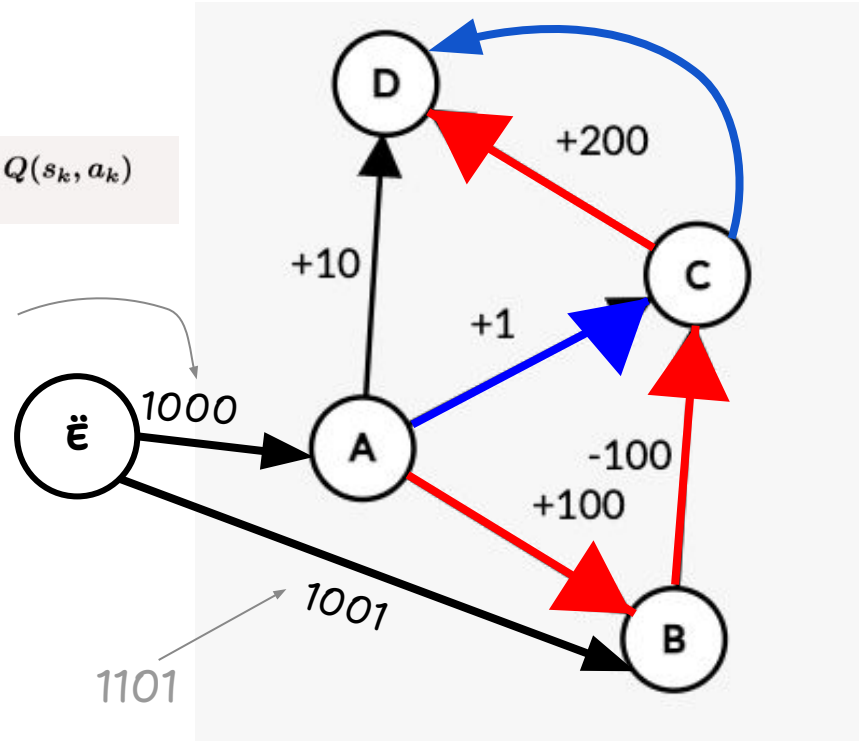


Модификация 1

epsilon-жадная стратегия

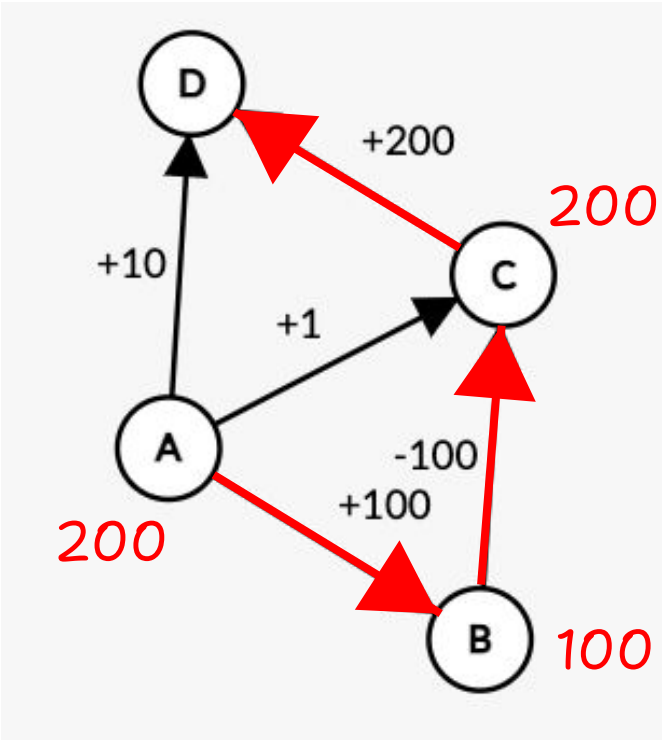
с вероятностью ϵ играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$

1/9 * 1201 + 8/9 * 1200



Проблемы

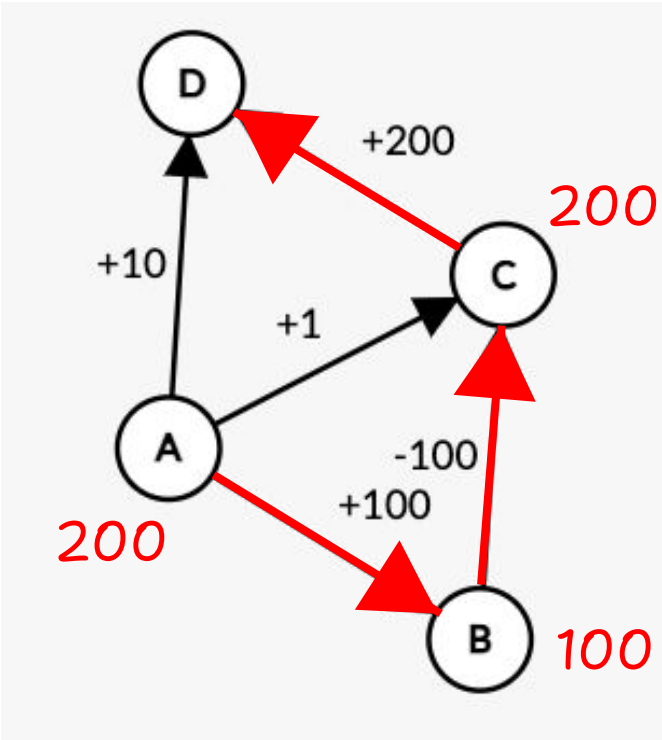
1. Если политика не включала ребро, то нет шансов включить его в **target policy**.



Проблемы

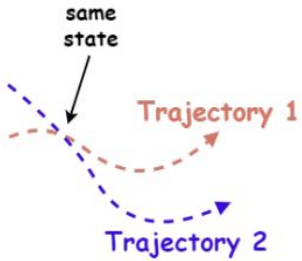
1. Если политика не включала ребро, то нет шансов включить его в **target policy**.
2. Нужно доигрывать эпизоды до конца

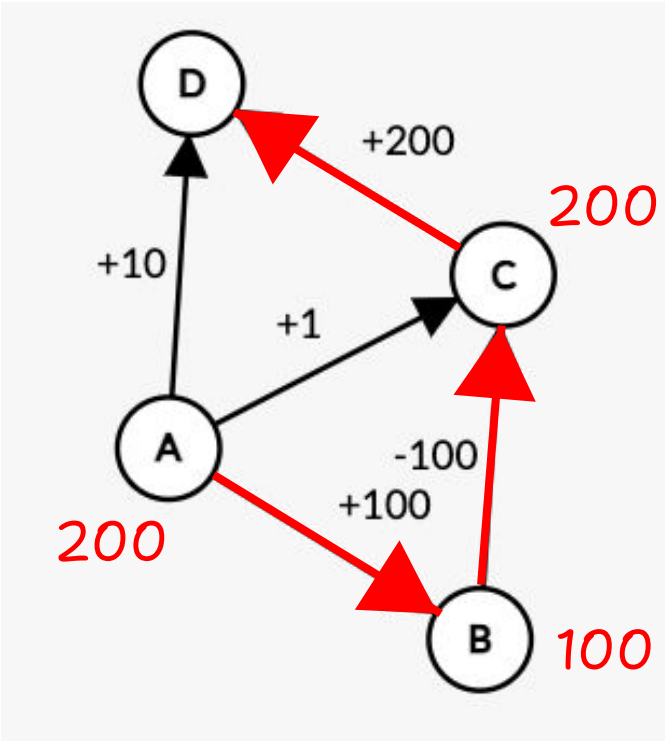
   Когда обновлять политику?



Проблемы

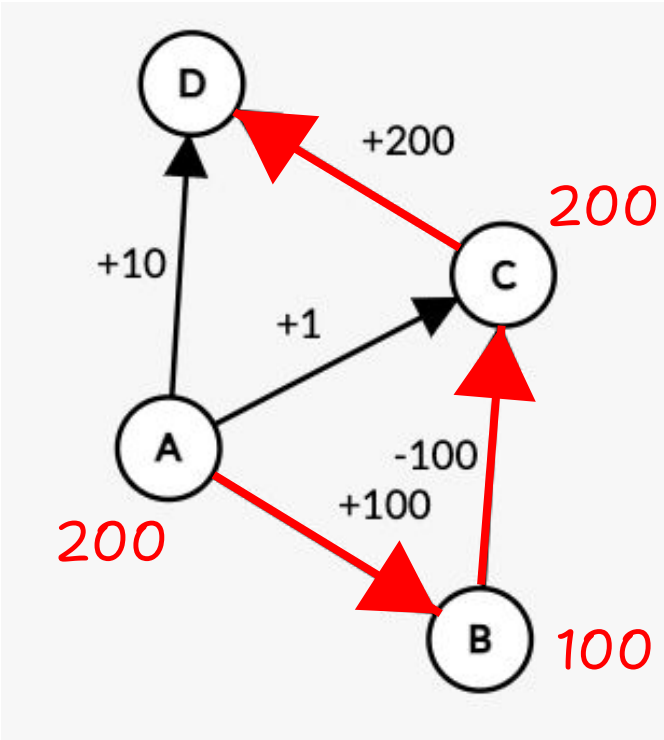
1. Если политика не включала ребро, то нет шансов включить его в **target policy**.
2. Нужно доигрывать эпизоды до конца
3. Теряем кучу информации на пересечениях





Модификация 2

Давайте сразу менять Q!



Модификация 2

Давайте сразу менять Q!

1. Устанавливаем количество эпизодов N.
2. Внутри эпизода начинаем проходить по шагам траектории t и на каждом шаге:
    - Выполняем выбранное действие и получаем следующее состояние, вознаграждение и флаг завершения
    - Получаем следующее действие с использованием epsilon-жадной стратегии
    - Обновляем Q-функцию с использованием следующего состояния и следующего действия **согласно политики**
    - Добавляем полученное вознаграждение к общему вознаграждению текущего эпизода
3. Завершаем эпизод, обновляем ϵ и переходим к следующему эпизоду.

SARSA

Давайте сразу менять Q!

1. Устанавливаем количество эпизодов N.
2. Внутри эпизода начинаем проходить по шагам траектории t и на каждом шаге:
    - Выполняем выбранное действие и получаем следующее состояние, вознаграждение и флаг завершения
    - Получаем следующее действие с использованием epsilon-жадной стратегии
    - Обновляем Q-функцию с использованием следующего состояния и следующего действия **согласно политики**
    - Добавляем полученное вознаграждение к общему вознаграждению текущего эпизода
3. Завершаем эпизод, обновляем ϵ и переходим к следующему эпизоду.

Q-learning

Давайте сразу менять Q!

1. Устанавливаем количество эпизодов N.
2. Внутри эпизода начинаем проходить по шагам траектории t и на каждом шаге:
   - Выполняем выбранное действие и получаем следующее состояние, вознаграждение и флаг завершения
   - Получаем следующее действие с использованием epsilon-жадной стратегии
   - Обновляем Q-функцию с использованием следующего состояния и следующего действия ~~согласно политики~~
   - Добавляем полученное вознаграждение к общему вознаграждению текущего эпизода
3. Завершаем эпизод, обновляем ϵ и переходим к следующему эпизоду.

жадно

Алгоритм Q-learning

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

Задание 7. ε = 0 γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	?	?	?

s1	a1	r1	s2
?	?	?	?

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

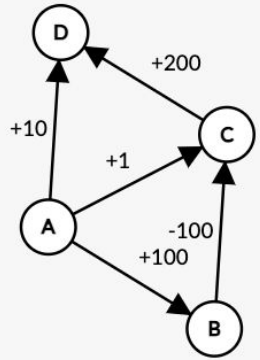
	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ε = 0 γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	?	?	?

s1	a1	r1	s2
?	?	?	?

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

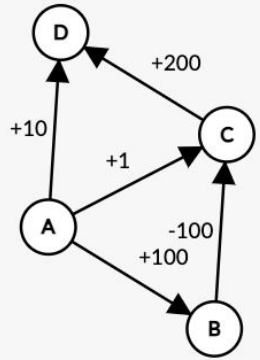
	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ε = 0 γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	?	?

s1	a1	r1	s2
?	?	?	?

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

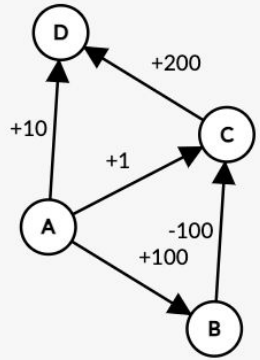
	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ε = 0 γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
?	?	?	?

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

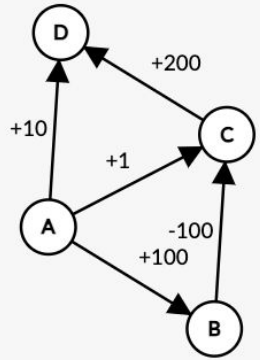
	A	B	C	D
A	?	?	?	?
B	?	?	?	?
C	?	?	?	?
D	?	?	?	?

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ε = 0 γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
D	?	?	?

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

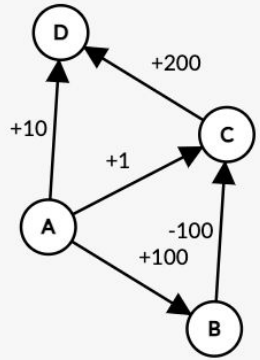
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ε = 0- γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	?	?	?

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

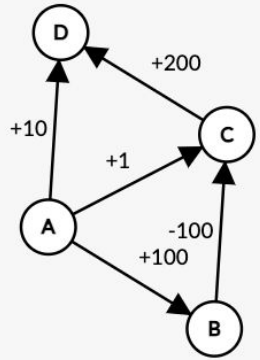
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	?	?	?

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

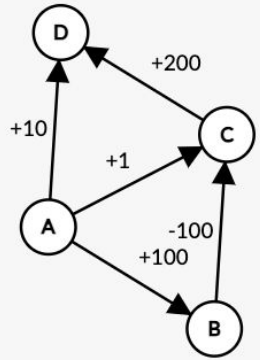
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	?	?

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

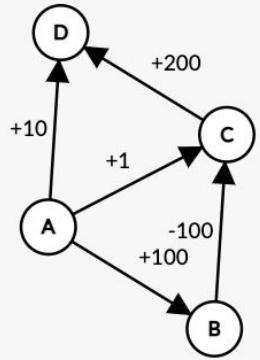
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

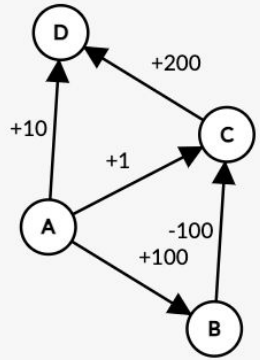
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

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

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
?	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

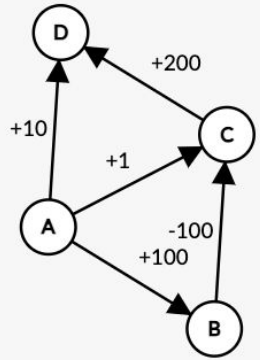
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	?	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

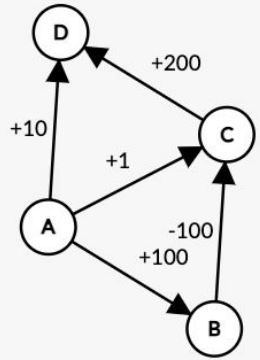
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	?	?

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

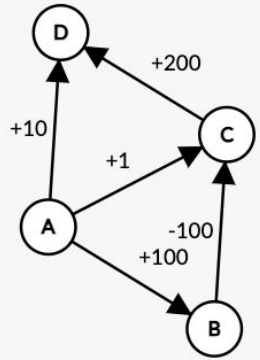
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

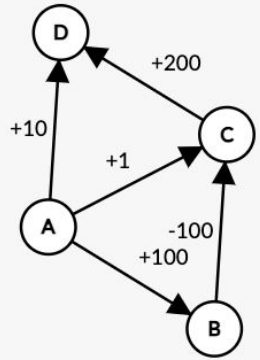
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

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



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
?	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

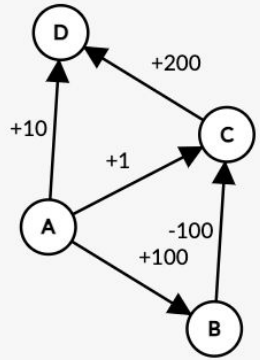
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	?	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

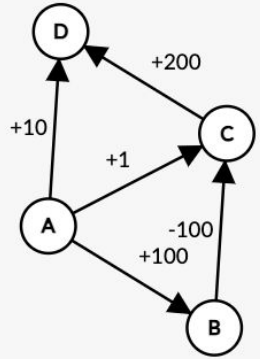
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	D	?	?

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

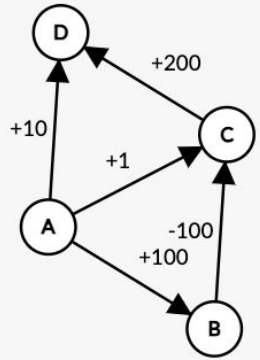
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	D	200	D

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

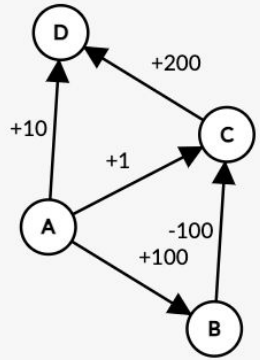
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	D	200	D

s4	a4	r4	s5
?	?	?	?

Q0:

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

Q1:

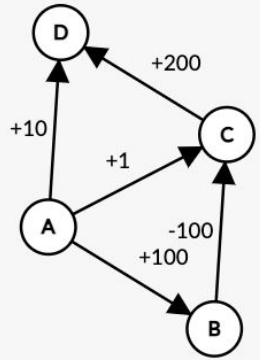
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	200
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	D	200	D

s4	a4	r4	s5
D	?	?	?

Q0:

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

Q1:

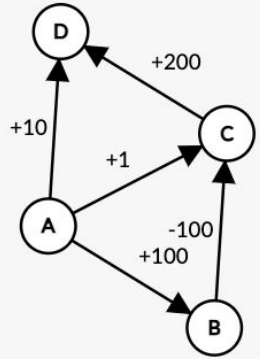
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	200
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	D	200	D

s4	a4	r4	s5
A	?	?	?

Q0:

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

Q1:

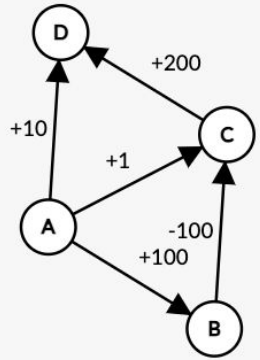
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	200
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	D	200	D

s4	a4	r4	s5
A	B	?	?

Q0:

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

Q1:

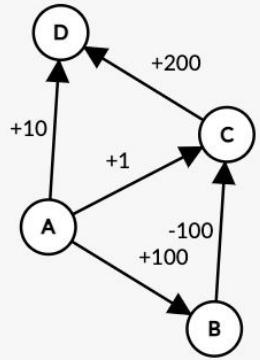
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	200
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	D	200	D

s4	a4	r4	s5
A	B	-100	B

Q0:

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

Q1:

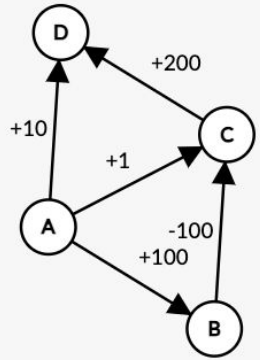
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q3-...:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	200
D	✗	✗	✗	✗



Остальное будем менять прямо тут

Задание 7. ~~ε = 0~~ γ = 1.0 α = 1

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

s0	a0	r0	s1
A	AD	10	D

s1	a1	r1	s2
A	B	100	B

s2	a2	r2	s3
B	C	-100	C

s3	a3	r3	s4
C	D	200	D

s4	a4	r4	s5
A	B	-100	B

Q0:

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

Q1:

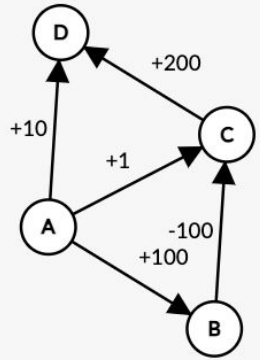
	A	B	C	D
A	✗	0	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

Q2:

	A	B	C	D
A	✗	100	0	10
B	✗	✗	0	✗
C	✗	✗	✗	0
D	✗	✗	✗	✗

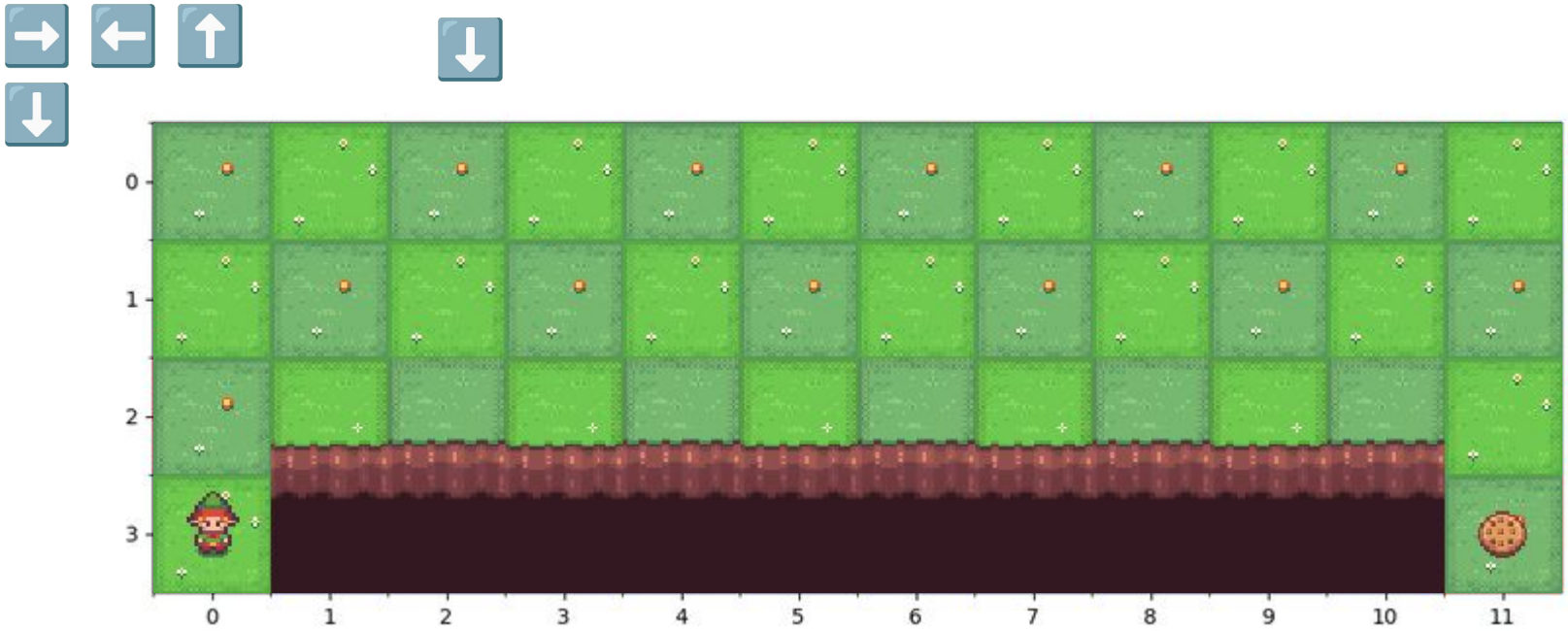
Q3-...:

	A	B	C	D
A	✗	0	0	10
B	✗	✗	-100	✗
C	✗	✗	✗	200
D	✗	✗	✗	✗

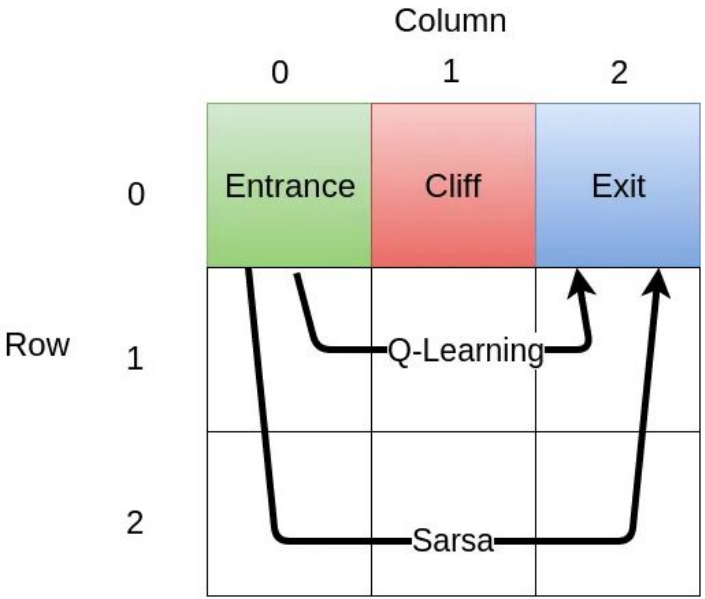


Остальное будем менять прямо тут

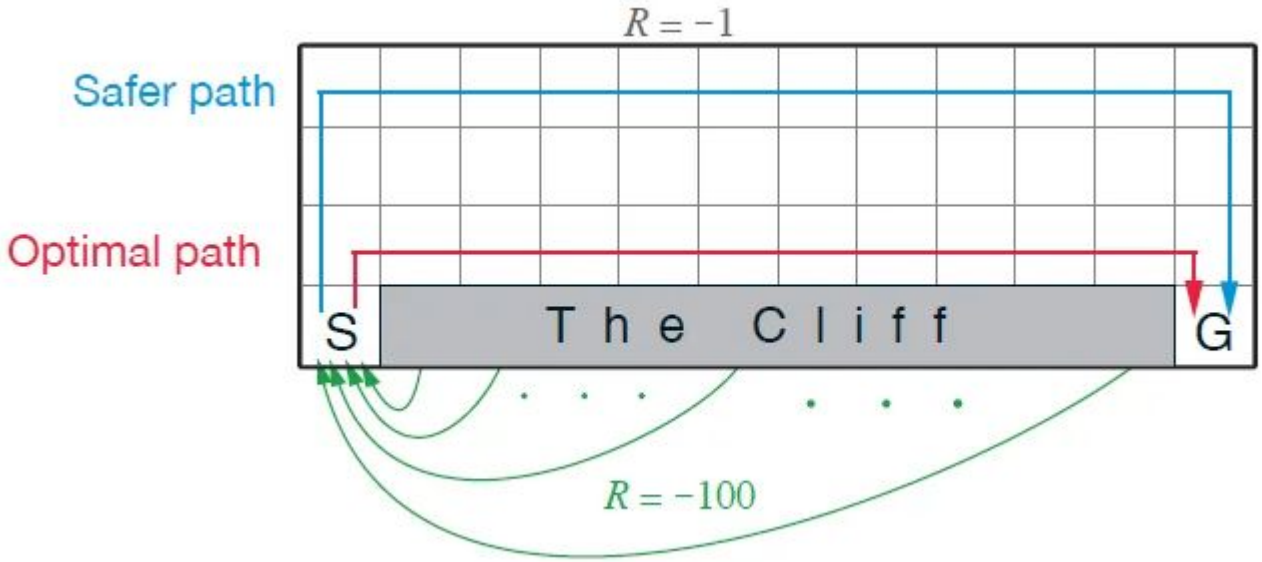
Cliff World



Cliff World



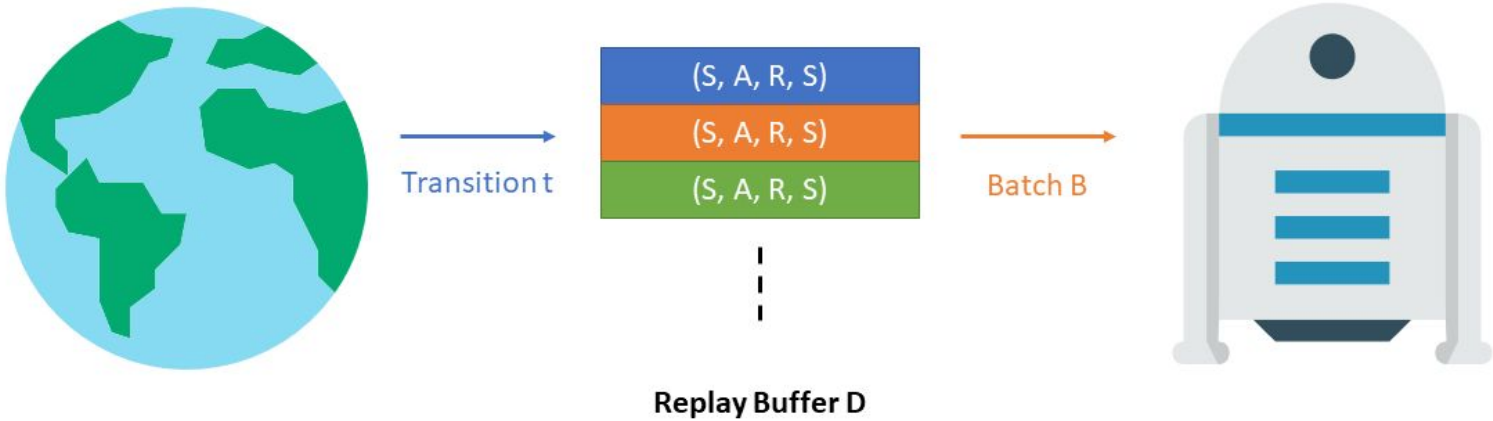
Cliff World



https://awjuliani.github.io/web-rl-playground

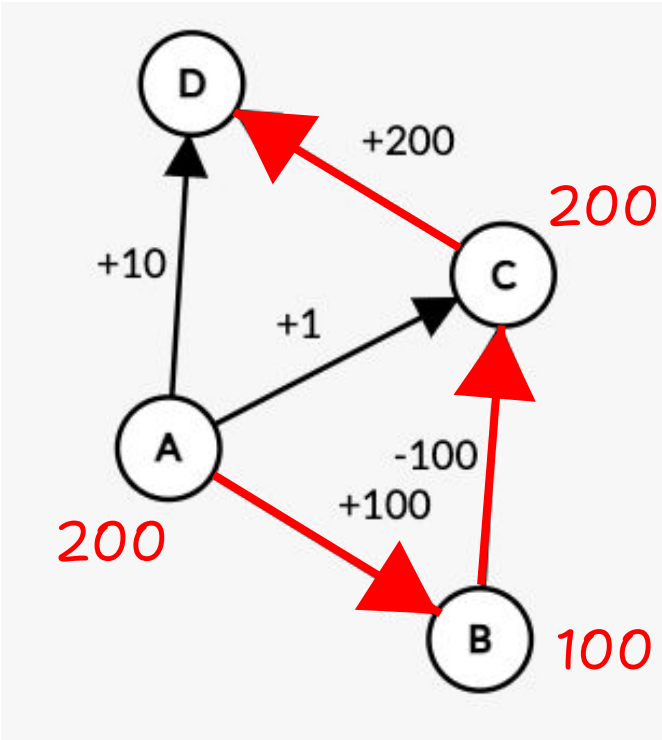
Реплей буфер

Replay buffer, experience replay — это память со всеми собранными агентом переходами (*s, a, r, s0, done*)

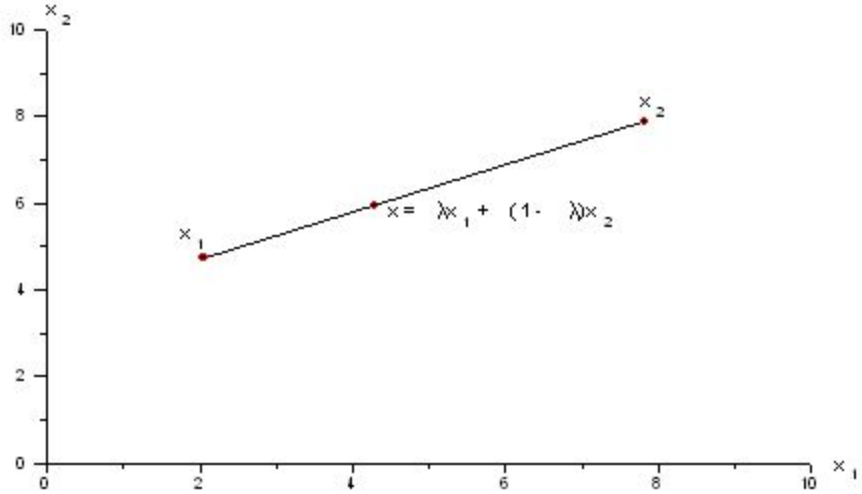


Проблемы

1. Если политика не включала ребро, то нет шансов включить его в **target policy**.
2. Нужно доигрывать эпизоды до конца
3. Теряем кучу информации на пересечениях
4. В текущем значении Q уже содержится много информации. Жалко ее терять



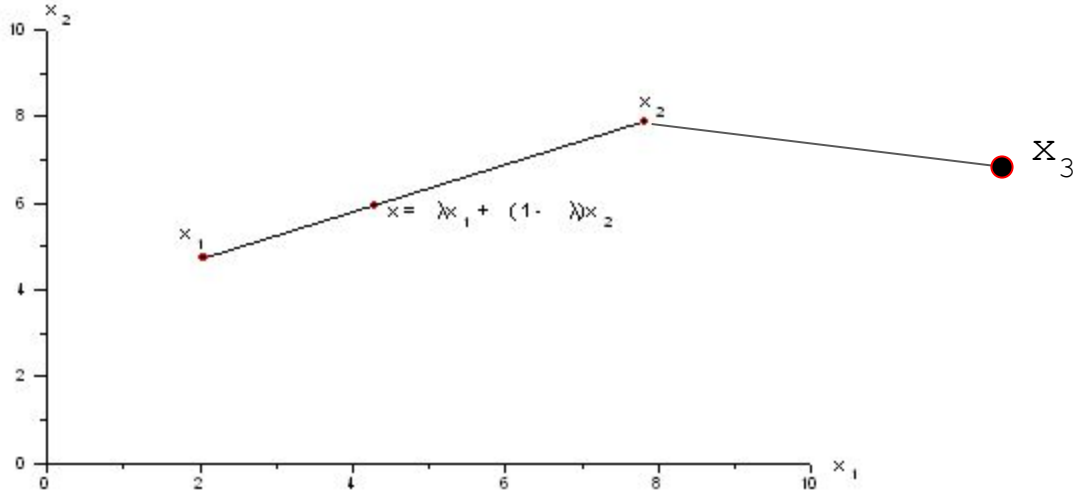
Выпуклая комбинация



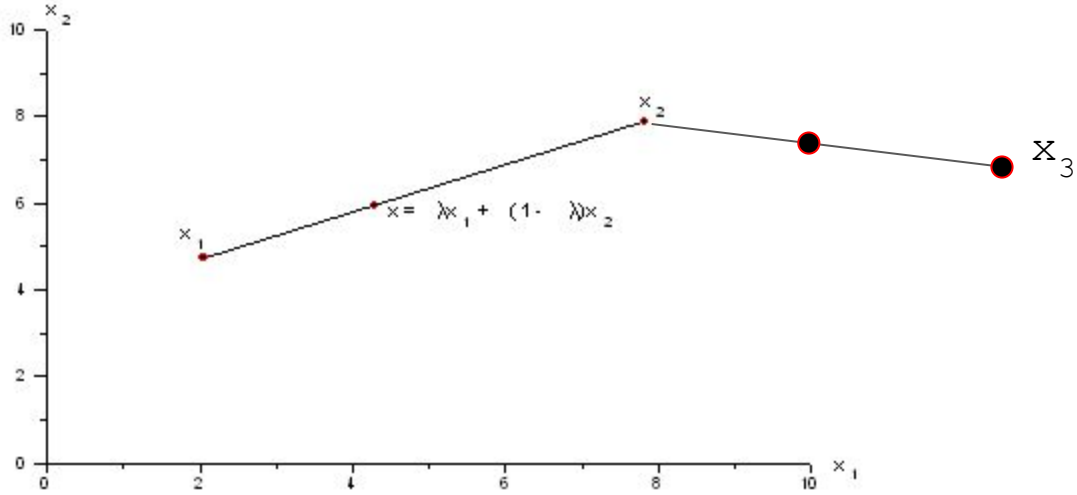
https://www.geogebra.org/m/WBU57uED

https://www.geogebra.org/m/rtrqmkyq

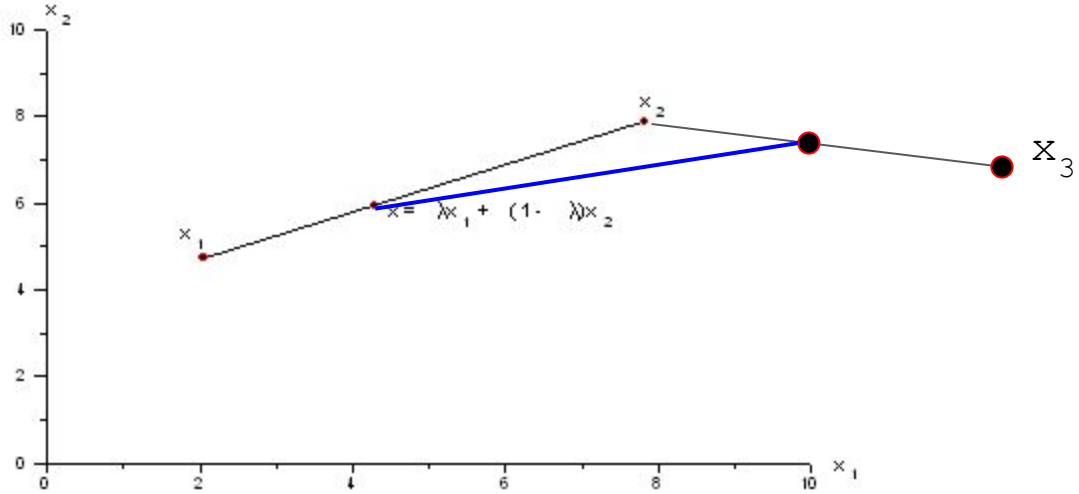
Выпуклая комбинация



Выпуклая комбинация

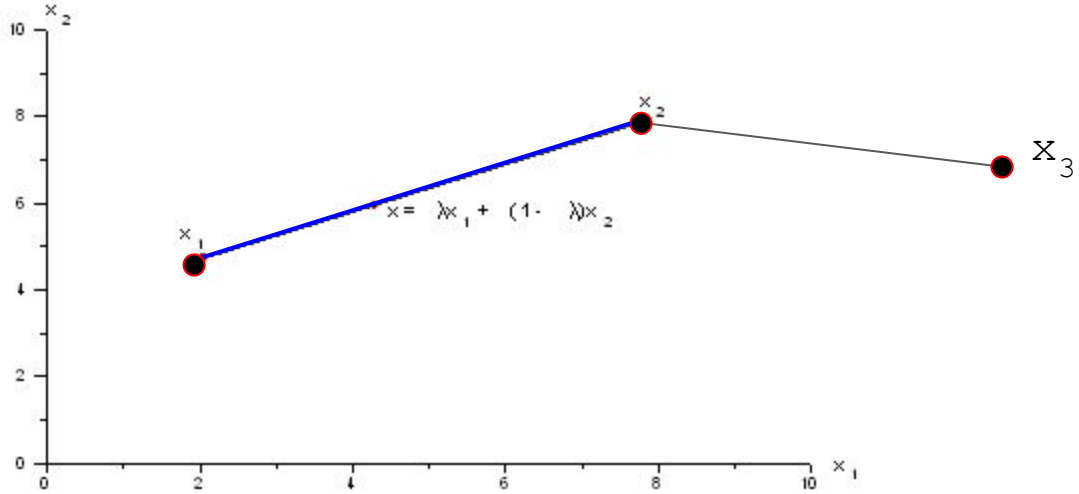


Выпуклая комбинация

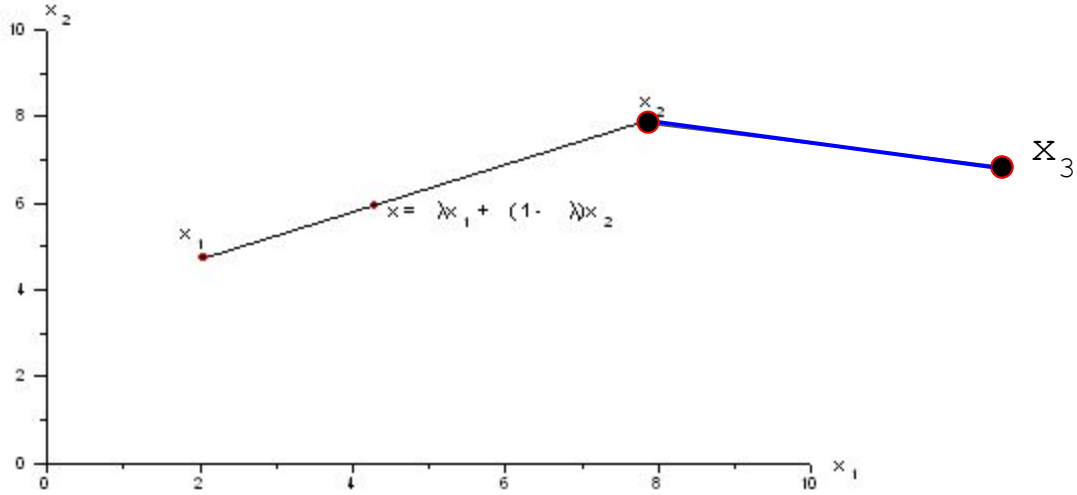


Выпуклая комбинация

$\lambda = 1$

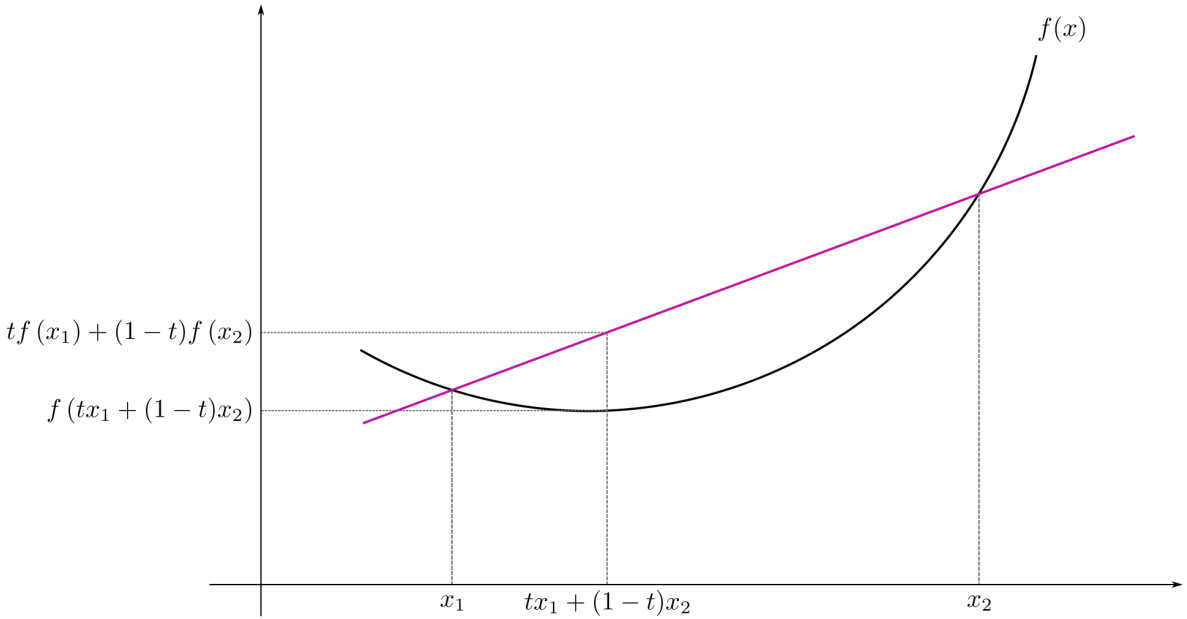


Выпуклая комбинация



λ = 0

Выпуклая комбинация



Экспоненциальное сглаживание

в

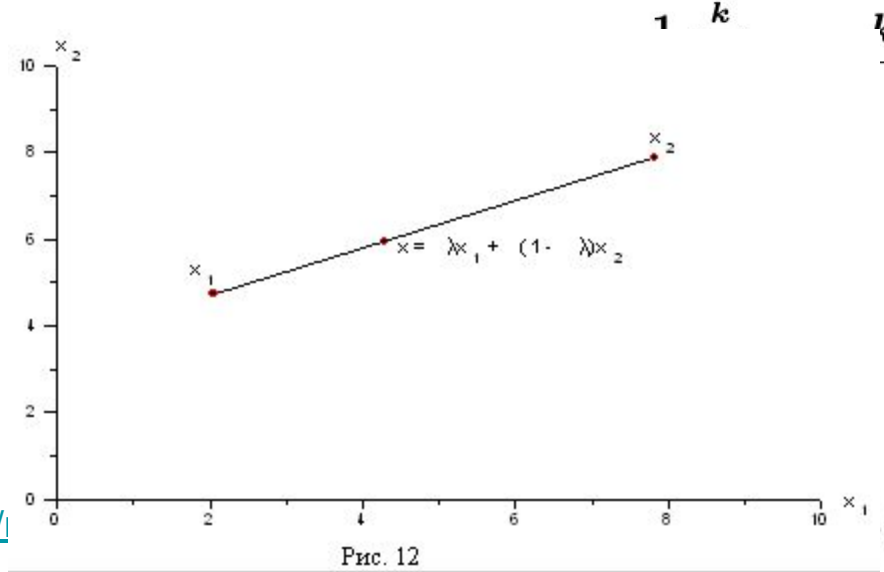


Рис. 12

$$\frac{k-1}{k} m_{k-1} + \frac{1}{k} x_k$$

$$\alpha_k := \frac{1}{k}$$

$$m_{k-1} + \alpha_k x_k$$

https://www.geogebra.org/

Q-learning

$$Q_{k+1}(s,a) \leftarrow Q_k(s,a) + \alpha_k \underbrace{\left(\overbrace{r + \gamma Q_k(s',a')}^{\text{таргет}} - Q_k(s,a) \right)}_{\text{временная разность}}$$

Q-learning

$$Q_{k+1}(s,a) \leftarrow Q_k(s,a) + \alpha_k \underbrace{\left(\overbrace{r + \gamma Q_k(s',a')}^{\text{таргет}} - Q_k(s,a) \right)}_{\text{временная разность}}$$

temporal difference

learning rate

Q-learning

Алгоритм 10: Q-learning

Гиперпараметры: α — параметр экспоненциального сглаживания, ε — параметр исследований

Инициализируем $Q(s, a)$ произвольно для всех $s \in \mathcal{S}, a \in \mathcal{A}$
Наблюдаем s_0
На k-ом шаге:

1. с вероятностью ε играем $a_k \sim \text{Uniform}(\mathcal{A})$, иначе $a_k = \underset{a_k}{\text{argmax}}\, Q(s_k, a_k)$
2. наблюдаем r_k, s_{k+1}
3. обновляем $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \left(r_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) \right)$

$$Q_{k+1}(s,a) \leftarrow Q_k(s,a) + \alpha_k \underbrace{\left(\overbrace{r + \gamma Q_k(s', a')}^{\text{target}} - Q_k(s,a) \right)}_{\text{временная разность}}$$

Резюме

Классификация

Классификация №1

1. Model free
2. Model based

Классификация №2

1. Value based
2. Policy based

Классификация №3

1. On-policy
2. Off-policy

Способ 1

Вообще в этой игре существует выигрышная стратегия для первого игрока. Т.к. вне зависимости от ходов второго, первый может выиграть.

Способ 1

Достаточно первым ходом взять одну палочку, а дальше дополнять ход соперника до трёх палочек.

Что это значит?

Способ 1

Алгоритм

Если соперник взял 1 палочку – бери ___

Если соперник взял 2 палочки – бери ___

Способ 2

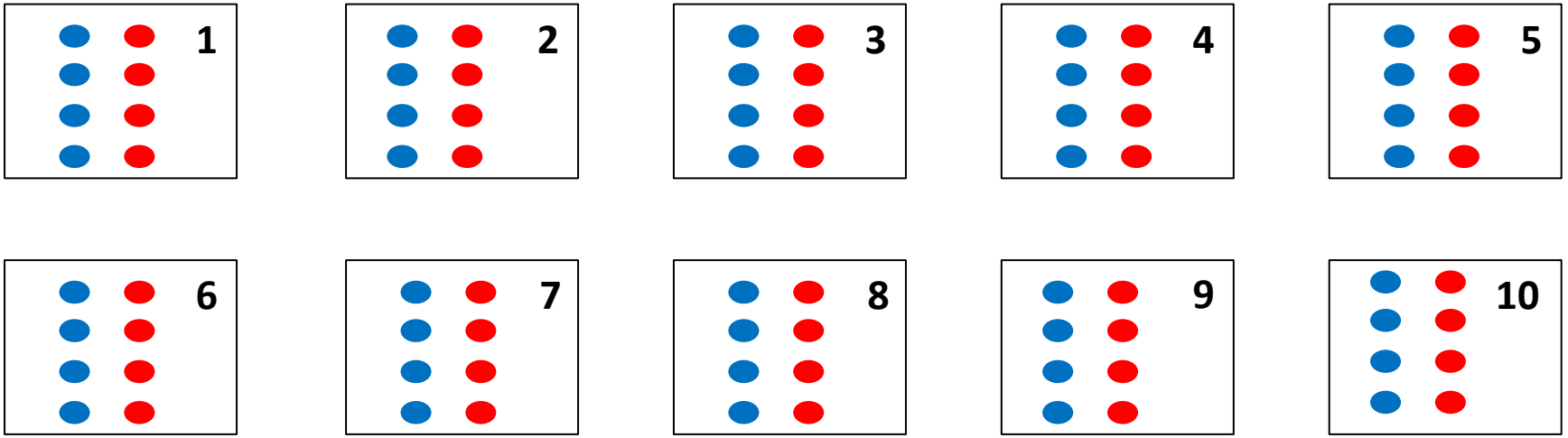
А теперь попробуем создать «искусственный интеллект», который будет сам обучаться выигрышной стратегии без наших подсказок.

Способ 2

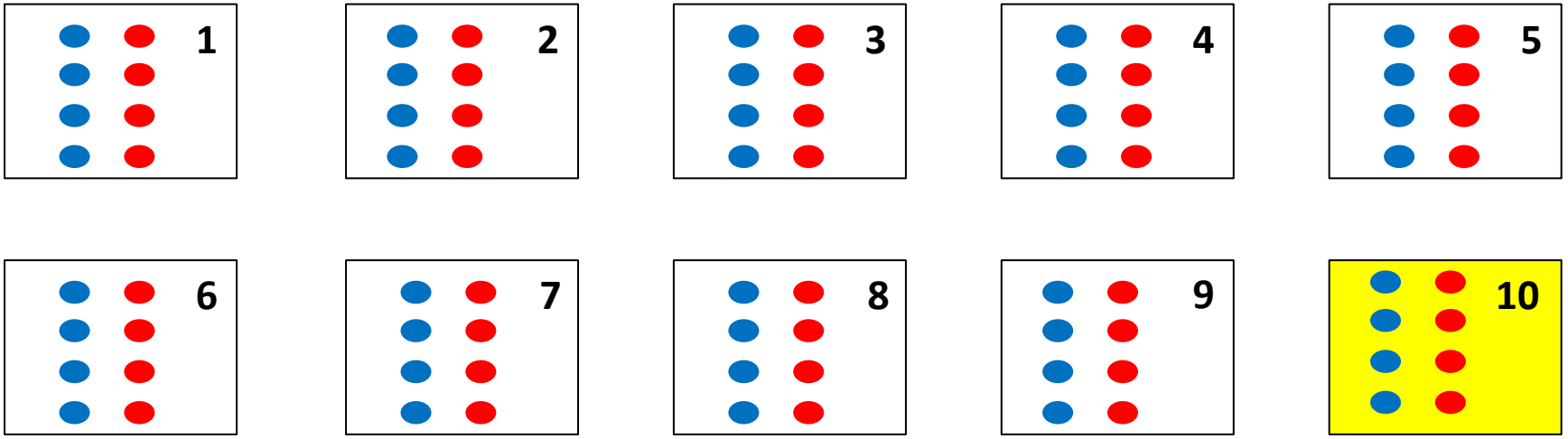
Для этого даже не понадобится компьютер.

Способ 2

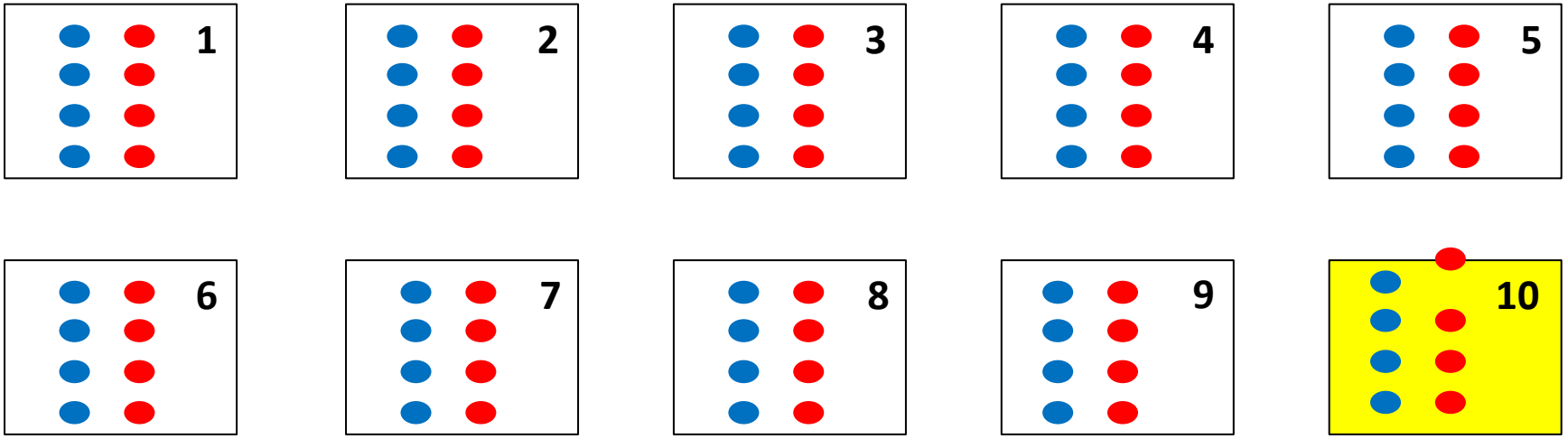
- Всего перед ходом игрока может оказаться 10 вариантов количества палочек на столе: 1, 2, 3, …, 10.
- Поэтому возьмем 10 коробочек и промаркируем их числами от 1 до 10.
- В каждую из них положим 4 красных и 4 синих пуговицы.



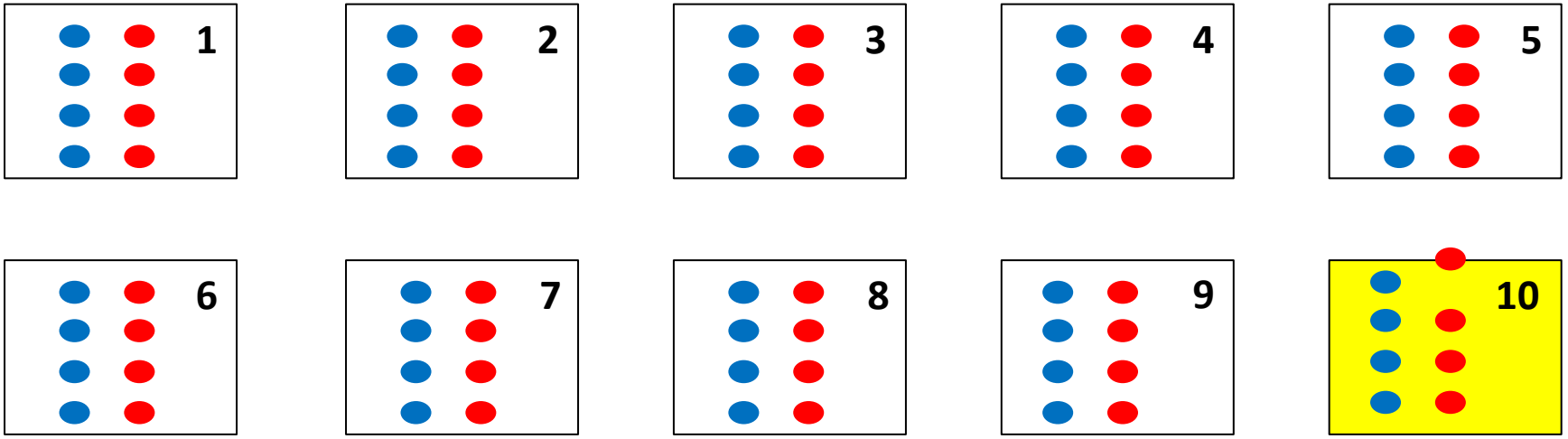
- Перед нами 10 палочек, поэтому возьмем 10-ую коробку и вытащим не глядя одну пуговицу. Если она будет синей – возьмем 1 палочку, если красной, то



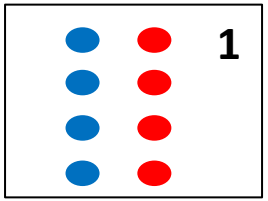
- Перед нами 10 палочек, поэтому возьмем 10-ую коробку и вытащим не глядя одну пуговицу. Если она будет синей – возьмем 1 палочку, если красной, то

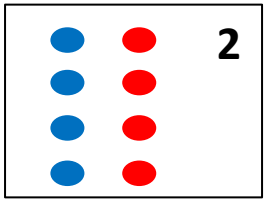


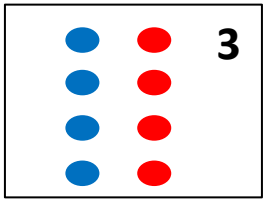
- Перед нами 10 палочек, поэтому возьмем 10-ую коробку и вытащим не глядя одну пуговицу. Если она будет синей – возьмем 1 палочку, если красной, то

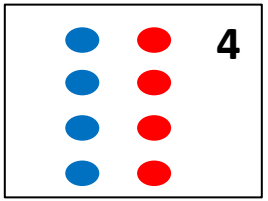


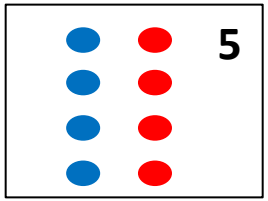
- Теперь ходит первый игрок. Допусти он возьмёт одну палочку.

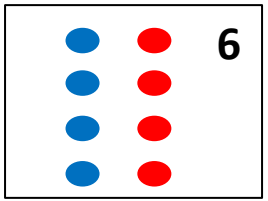


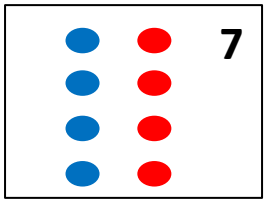


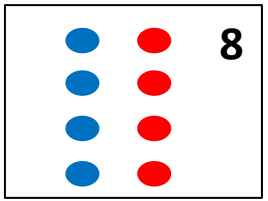


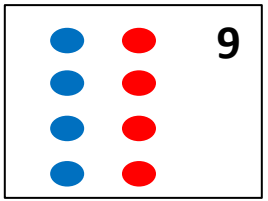


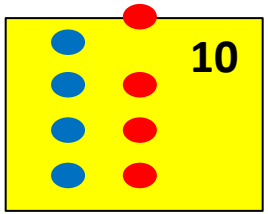




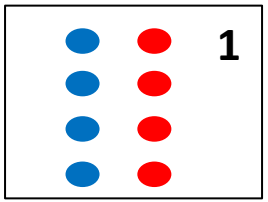


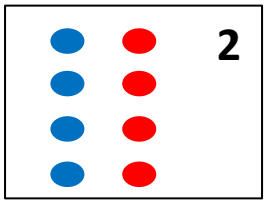


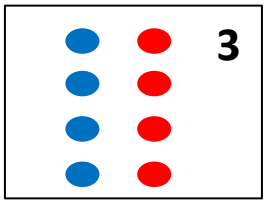


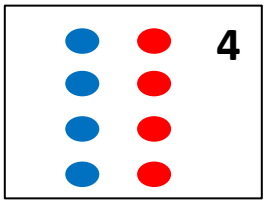


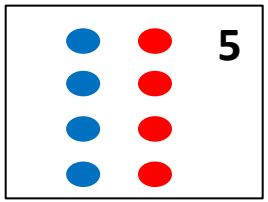
- Теперь ходит первый игрок. Допусти он возьмёт одну палочку.

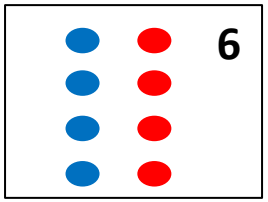


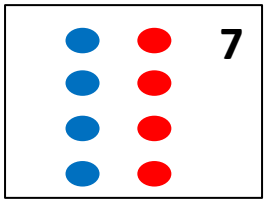


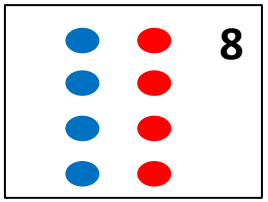


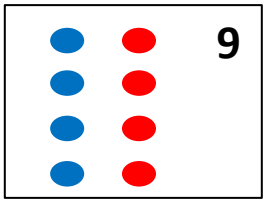


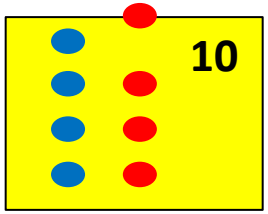




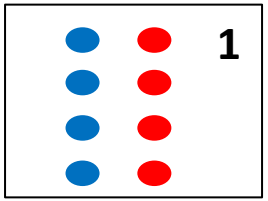


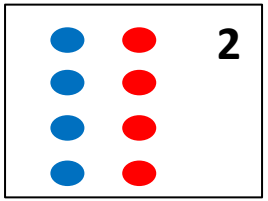


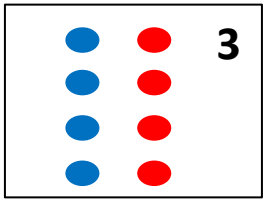


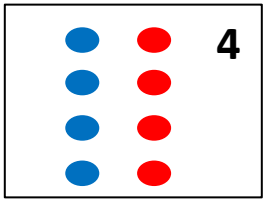


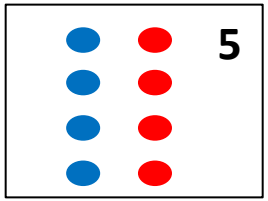
- Перед нами 7 палочек. Поэтому возьмём седьмую коробку.

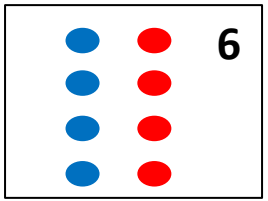


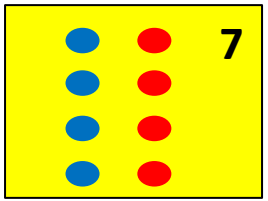


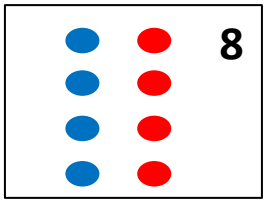


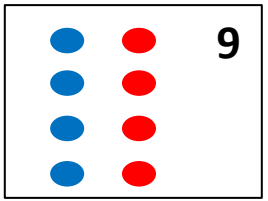


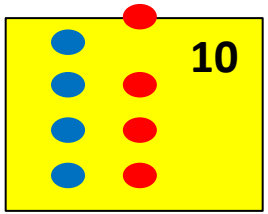




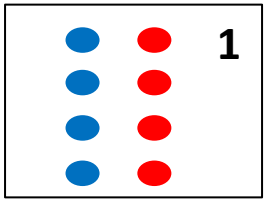


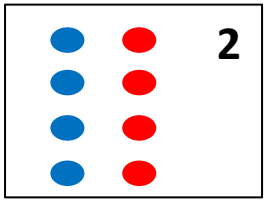


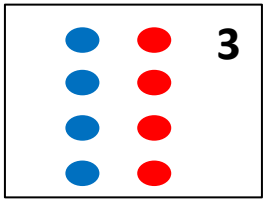


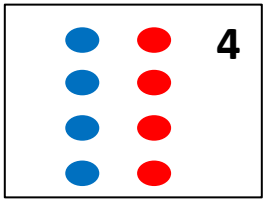


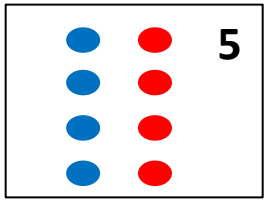
- Вытащим наугад пуговицу

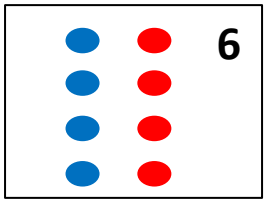


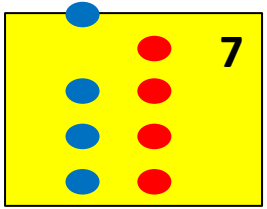


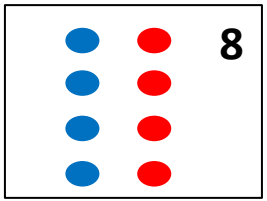


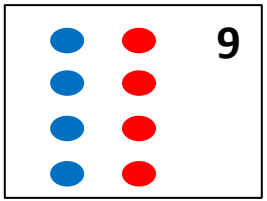


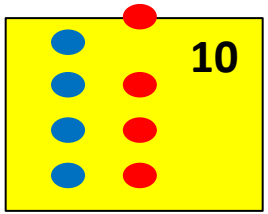




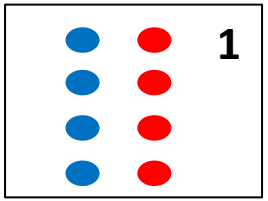


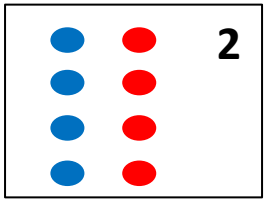


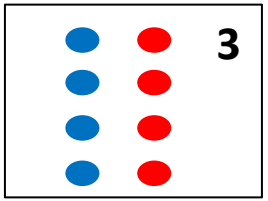


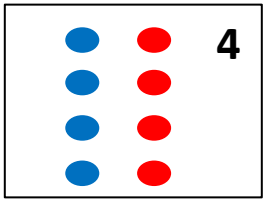


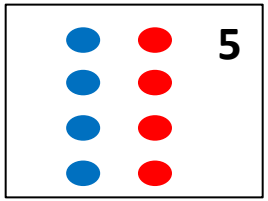
- Сходим

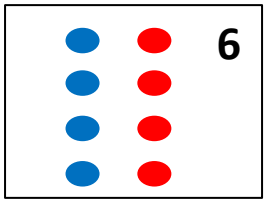


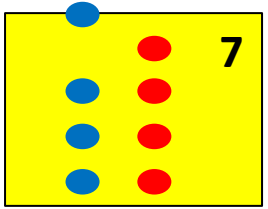


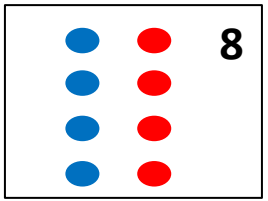


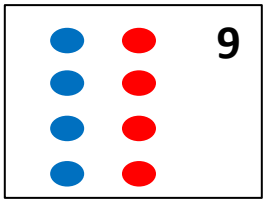


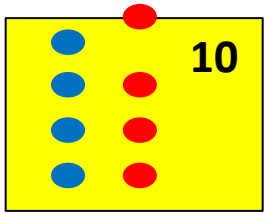




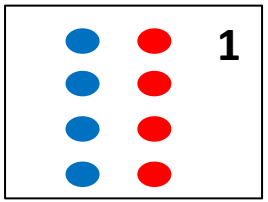


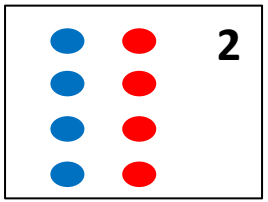


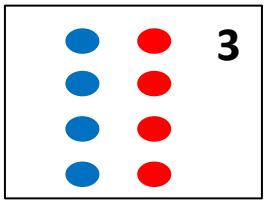


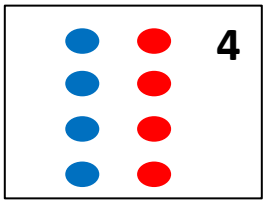


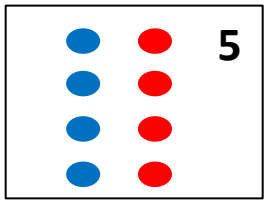
- Допустим игрок возьмёт 2 палочки.

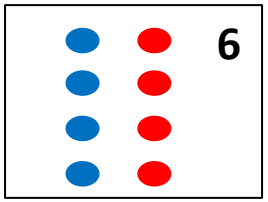


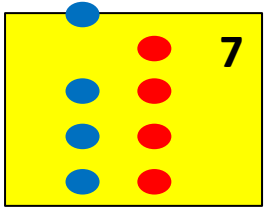


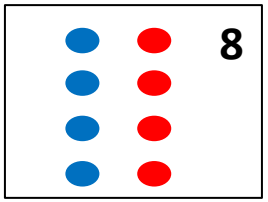


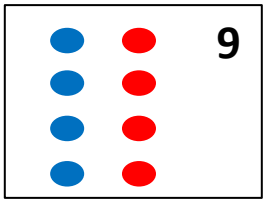


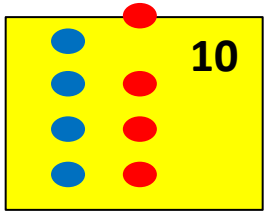




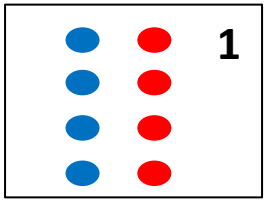


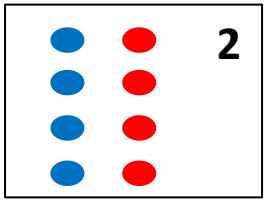


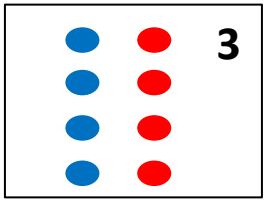


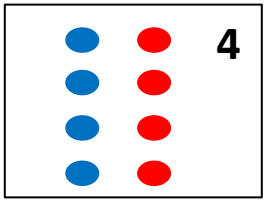


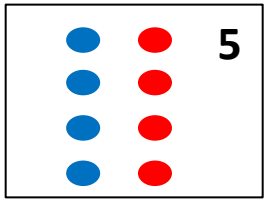
- Допустим игрок возьмёт 2 палочки.

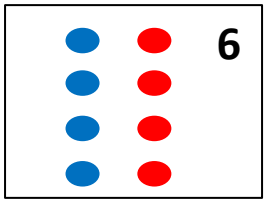


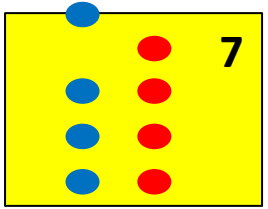


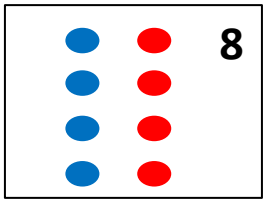


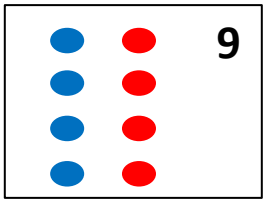


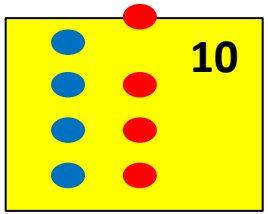




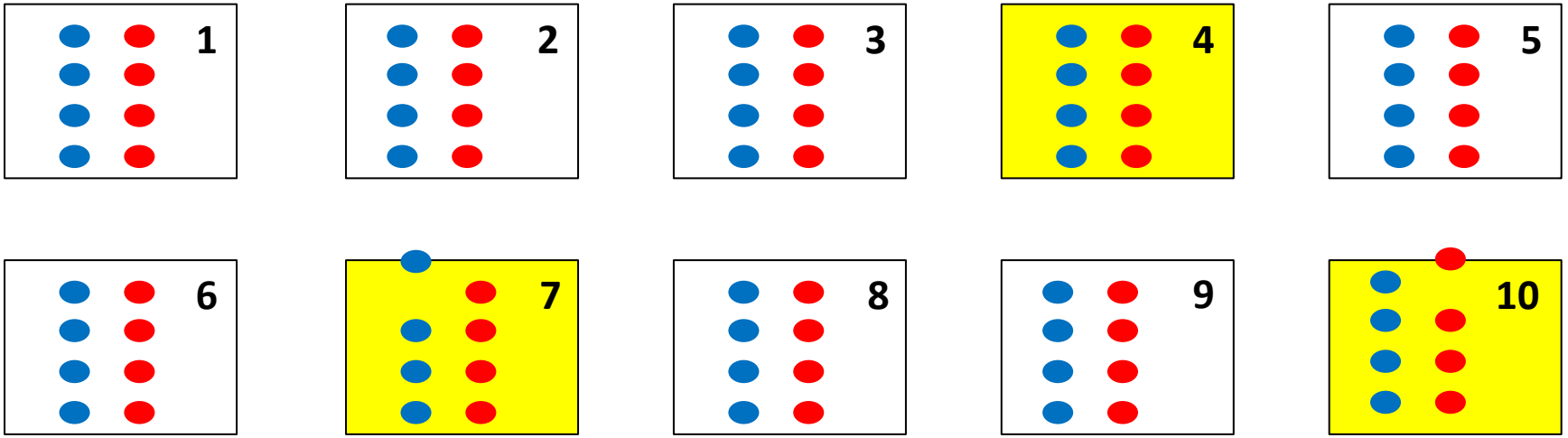




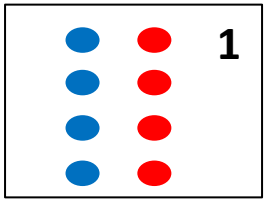


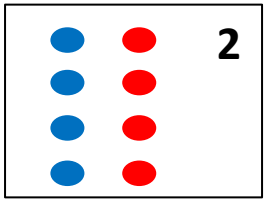


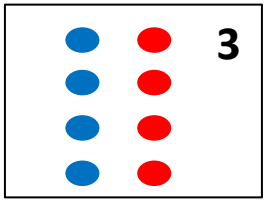
- Перед нами 4 палочки. Берём коробку №4

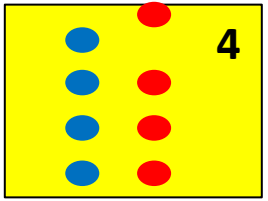


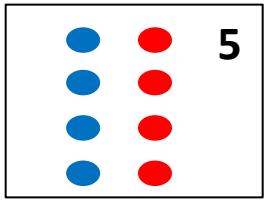
- Тянем пуговицу.

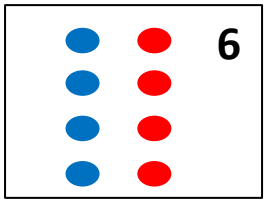


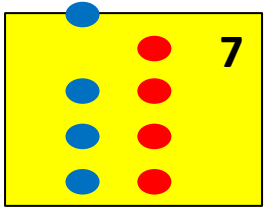


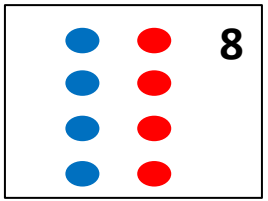


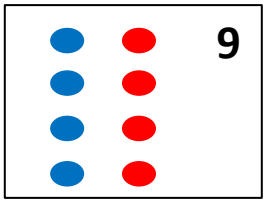


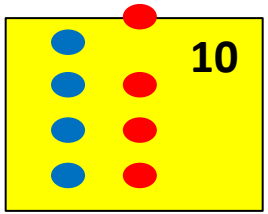




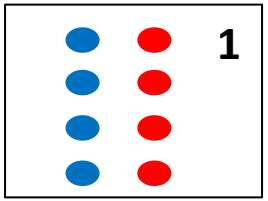


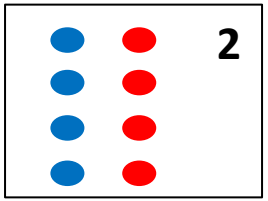


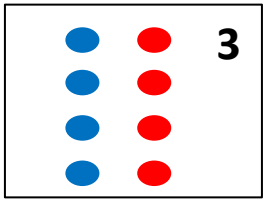


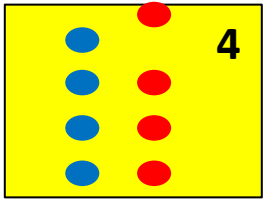


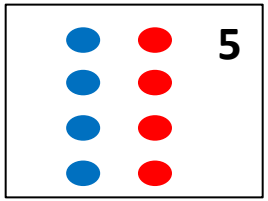
- Делаем ход.

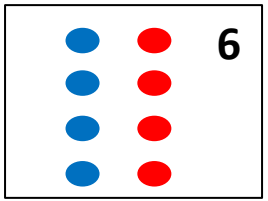


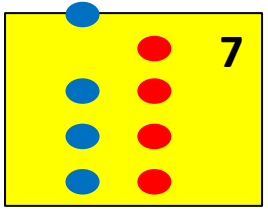


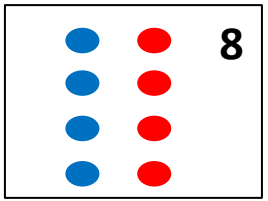


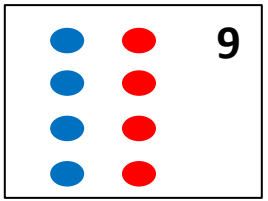


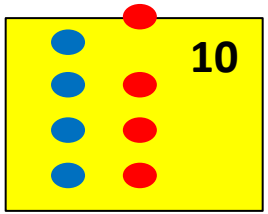




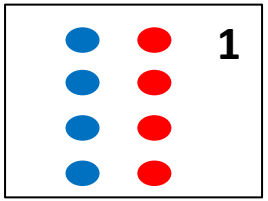


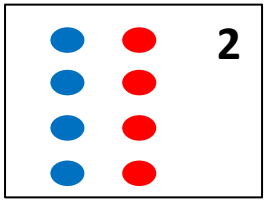


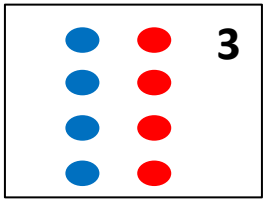


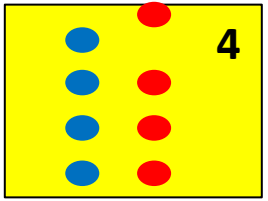


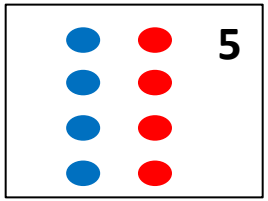
- Первый игрок выигрывает! А наш искусственный интеллект проигрывает

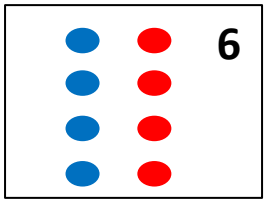


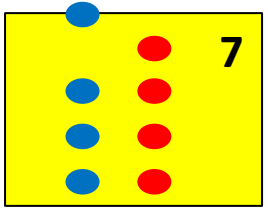


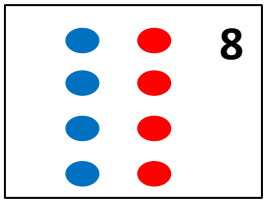


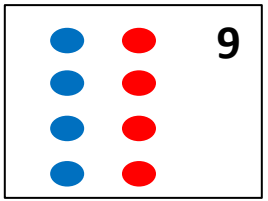


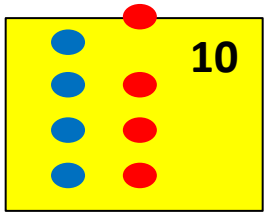




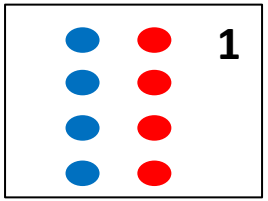


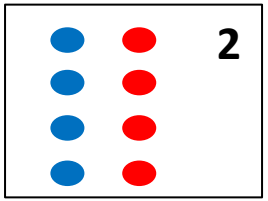


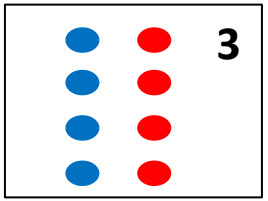


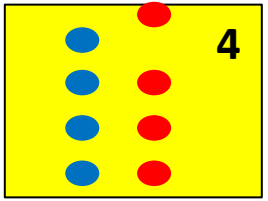


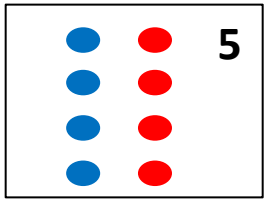
- Поэтому накажем его. И из всех коробочек, с которыми мы работали заберем пуговицы, которые привели нас к проигрышу.

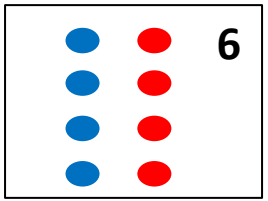


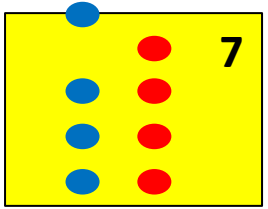


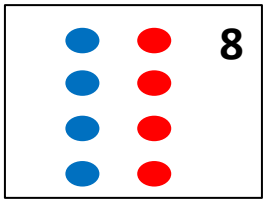


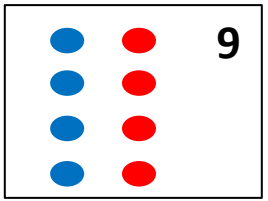


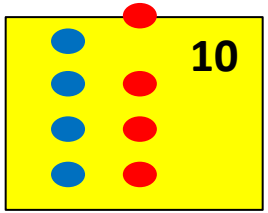




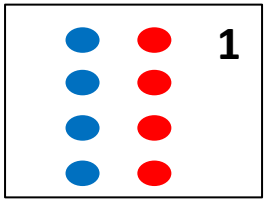


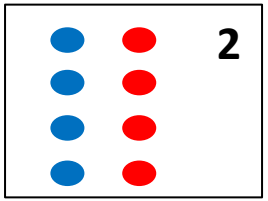


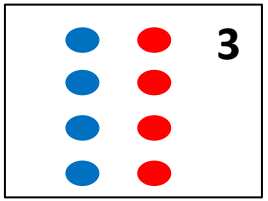


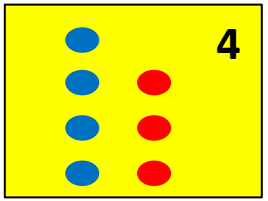


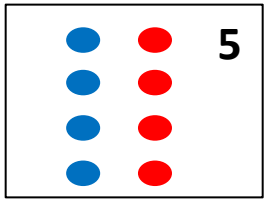
- Поэтому накажем его. И из всех коробочек, с которыми мы работали заберем пуговицы, которые привели нас к проигрышу.

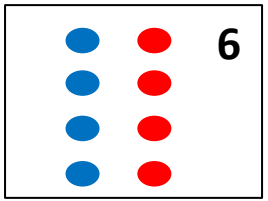


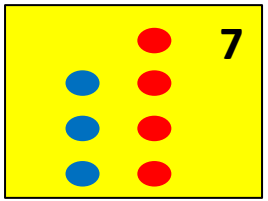


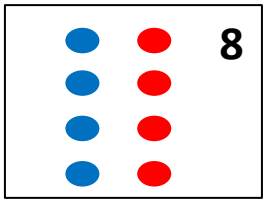


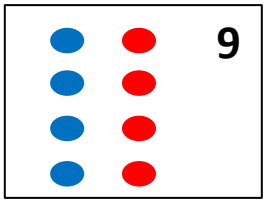


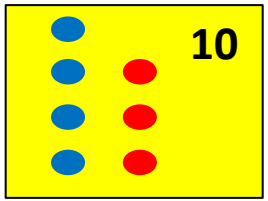




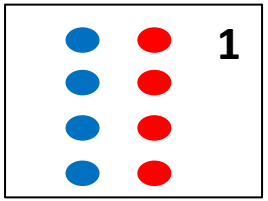


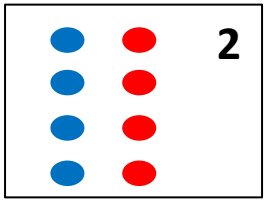


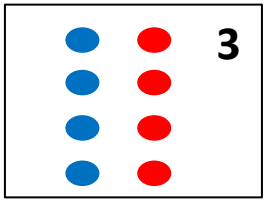


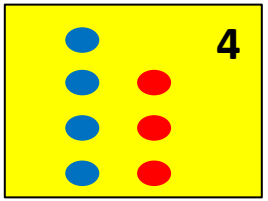


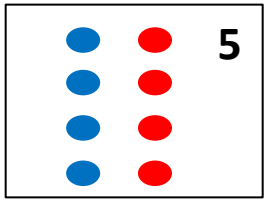
- Если бы мы выиграли – мы бы наоборот добавили по пуговице того цвета, которая привела нас к победе.

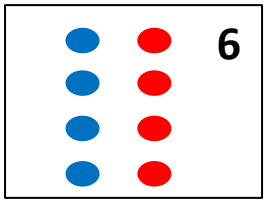


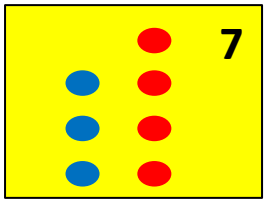


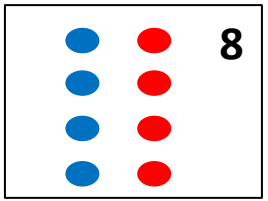


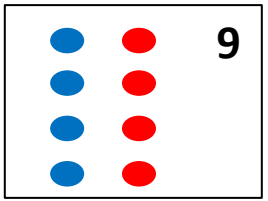


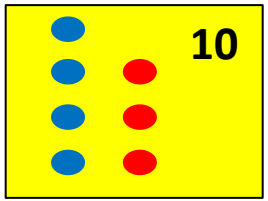






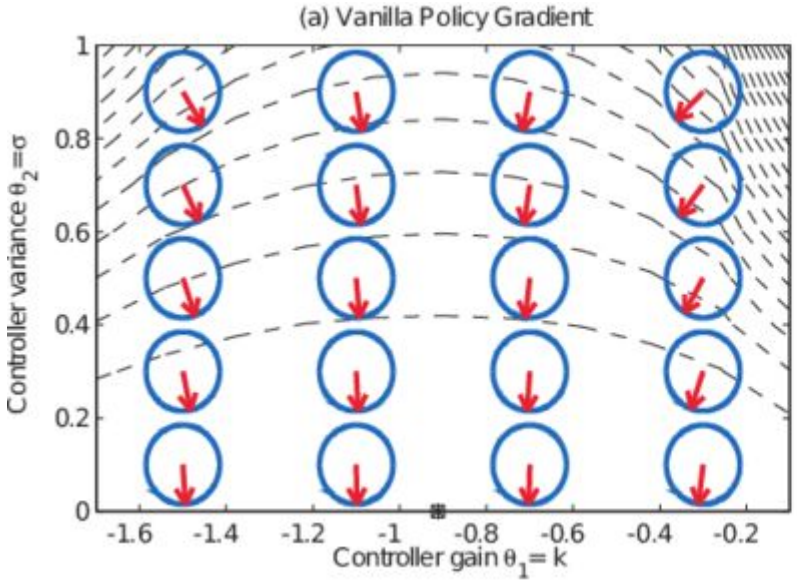


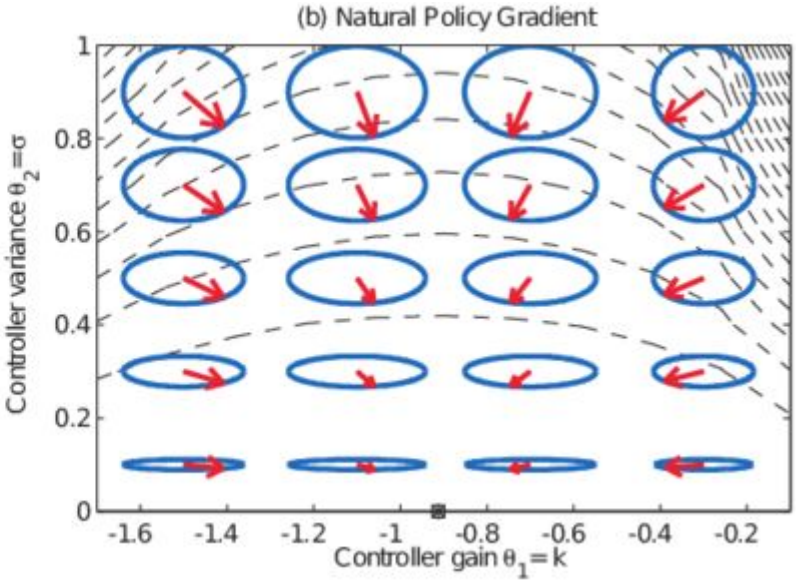




Policy gradient vs. Natural policy gradient

Адаптивные шаги!





Классификация

Классификация №1

1. Model free
2. Model based

Классификация №2

1. Value based
2. Policy based

Классификация №3

1. On-policy
2. Off-policy

План

- Buffer replay
- TD = temporal difference
- E-greedy
- Q-learning
- SARSA

Опрос в конце: https://otus.ru/polls/141248/

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

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