Q-learning

Владимир Морозов

Агент

Агент

Окружающая среда

Действие

Агент

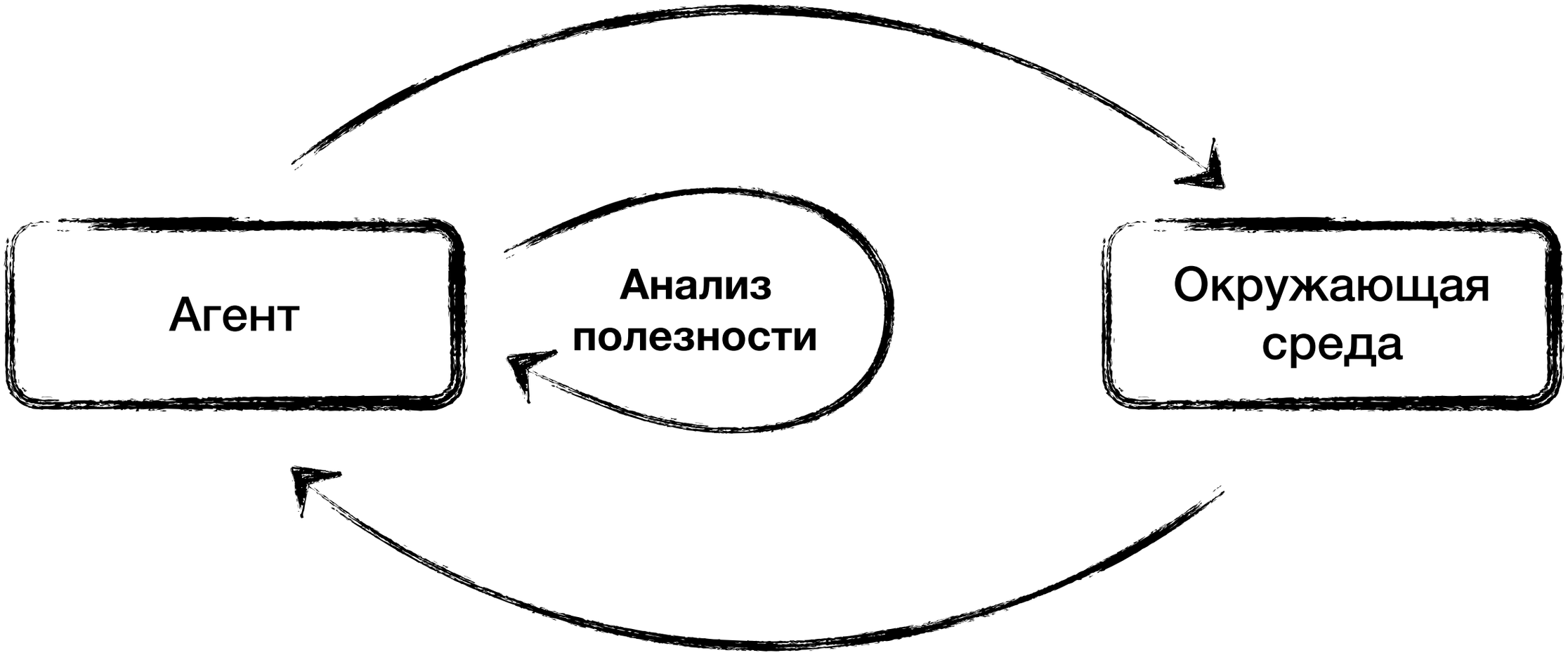
Окружающая среда

Действие

Агент

Окружающая среда

Вознаграждение


Действие
Агент
Анализ полезности
Окружающая среда
Вознаграждение

Алгоритм

1.Инициализация

2.Наблюдение

3.Обновление

4.Выбор действия

Инициализация

Инициализируем функцию полезности Q нулевыми или случайными значениями:

$$\forall\ i, j: Q[s_i, a_j] = 0$$
$$\forall\ i, j: Q[s_i, a_j] = RND$$

Пока модель ничего не знает о вознаграждениях окружающей среды.

Наблюдение

S_{prev} = S	Сохранить предыдущее состояние системы
A_{prev} = A	Сохранить предыдущее действие
S = OBSERVE	Получить текущее состояние (например, с датчиков)
R = FROM_ENV	Получить награду за предыдущее действие

Обновление

Фактор обучения
(**L**earning **F**actor)

Наибольшая полезность действия
в строке текущего состояния S

$$Q[S_{prev}, A_{prev}] = Q[S_{prev}, A_{prev}] + LF * (R + DF * MAX(Q,S) - Q[S_{prev}, A_{prev}])$$

Предыдущее состояние

Фактор дисконтирования
(**D**iscount **F**actor)

Вознаграждение
за предыдущее действие

Действие

Дейстаие с наибольшей полезностью
в строке текущего состояния S

A = ARGMAX(Q, S)

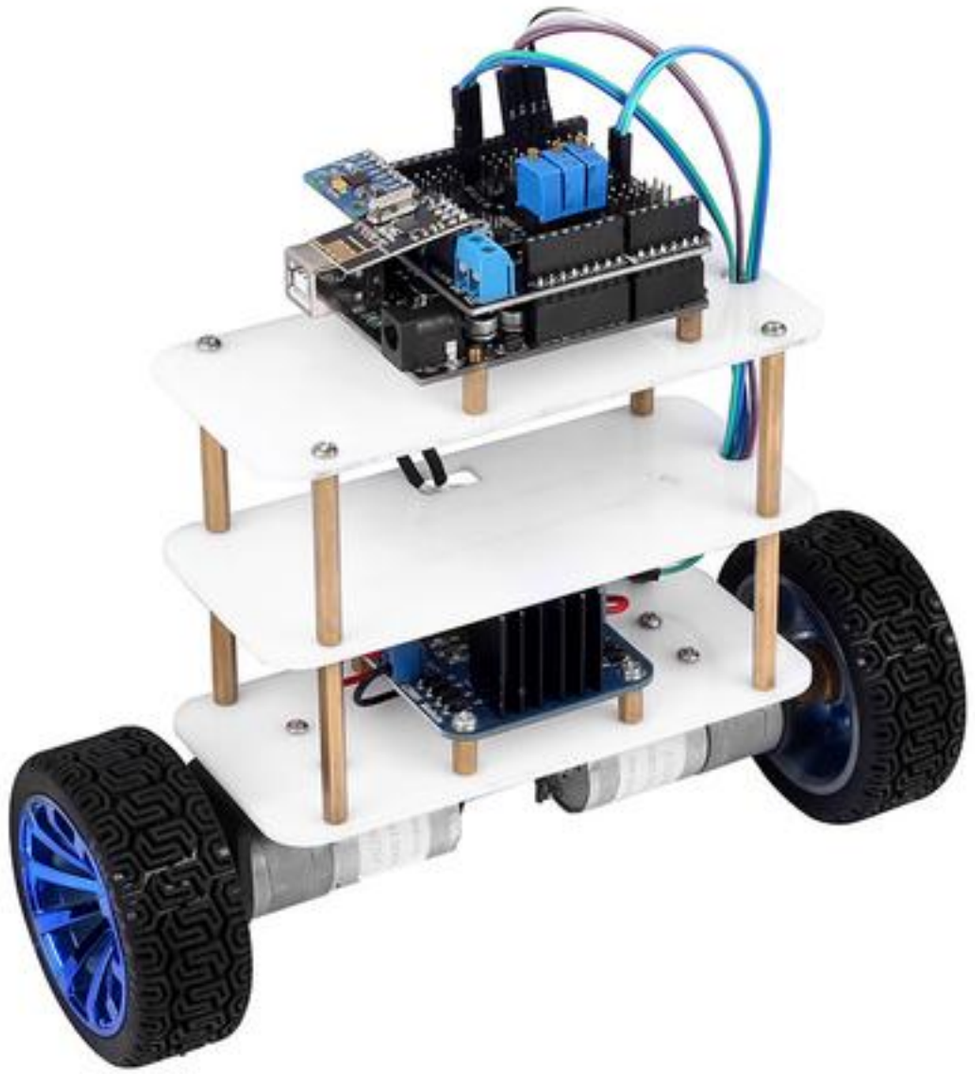
Текущее действие

Текущее состояние

Представление Q

	Действие 1	Действие 2	...	Действие N
Состояние 1	Q_{11}	Q_{12}	...	Q_{1N}
Состояние 2	Q_{21}	Q_{22}	...	Q_{2N}
...	...	...	...	...
Состояние M	Q_{M1}	Q_{M2}	...	Q_{MN}

Пример



Самобалансирующийся двухколёсный робот

Implementation of Q Learning and Deep Q Network For Controlling a Self Balancing Robot Model

MD Muhaimin Rahman[1,a], SM Hasanur Rashid[2,a] and M.M Hossain [b]

Матрица Q

Континуум состояний [-10, 10]	-200 рад/с	-100 рад/с	-50 рад/с	...	+200 рад/с
	0	0	0	...	0
	0	0	0	...	0
	...	...	...	...	...
	0	0	0	...	0

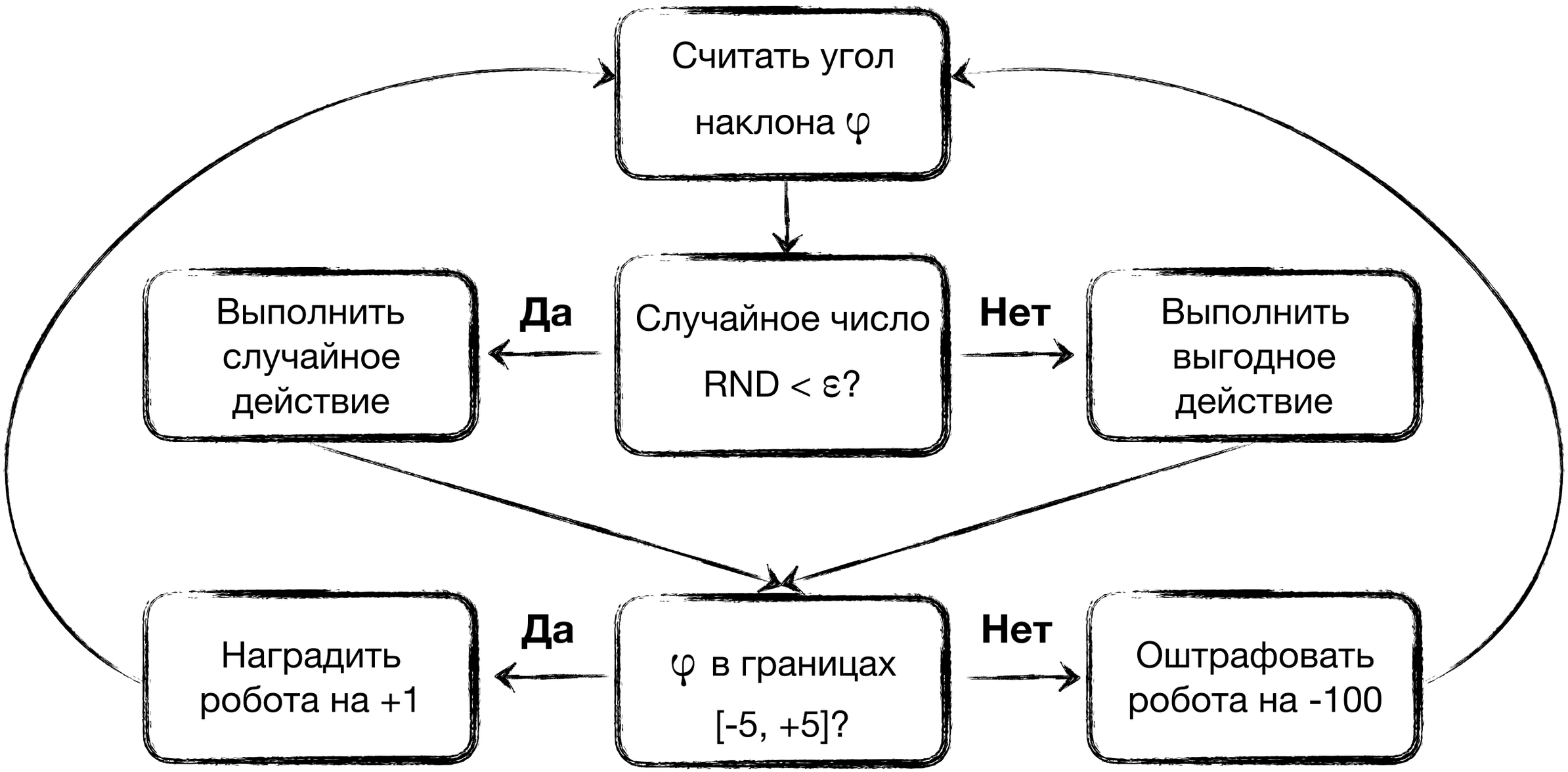
Матрица Q

	-200 рад/с	-100 рад/с	-50 рад/с	...	+200 рад/с
-10°	0	0	0	...	0
-9°	0	0	0	...	0
...	...	...	...	...	...
+10°	0	0	0	...	0

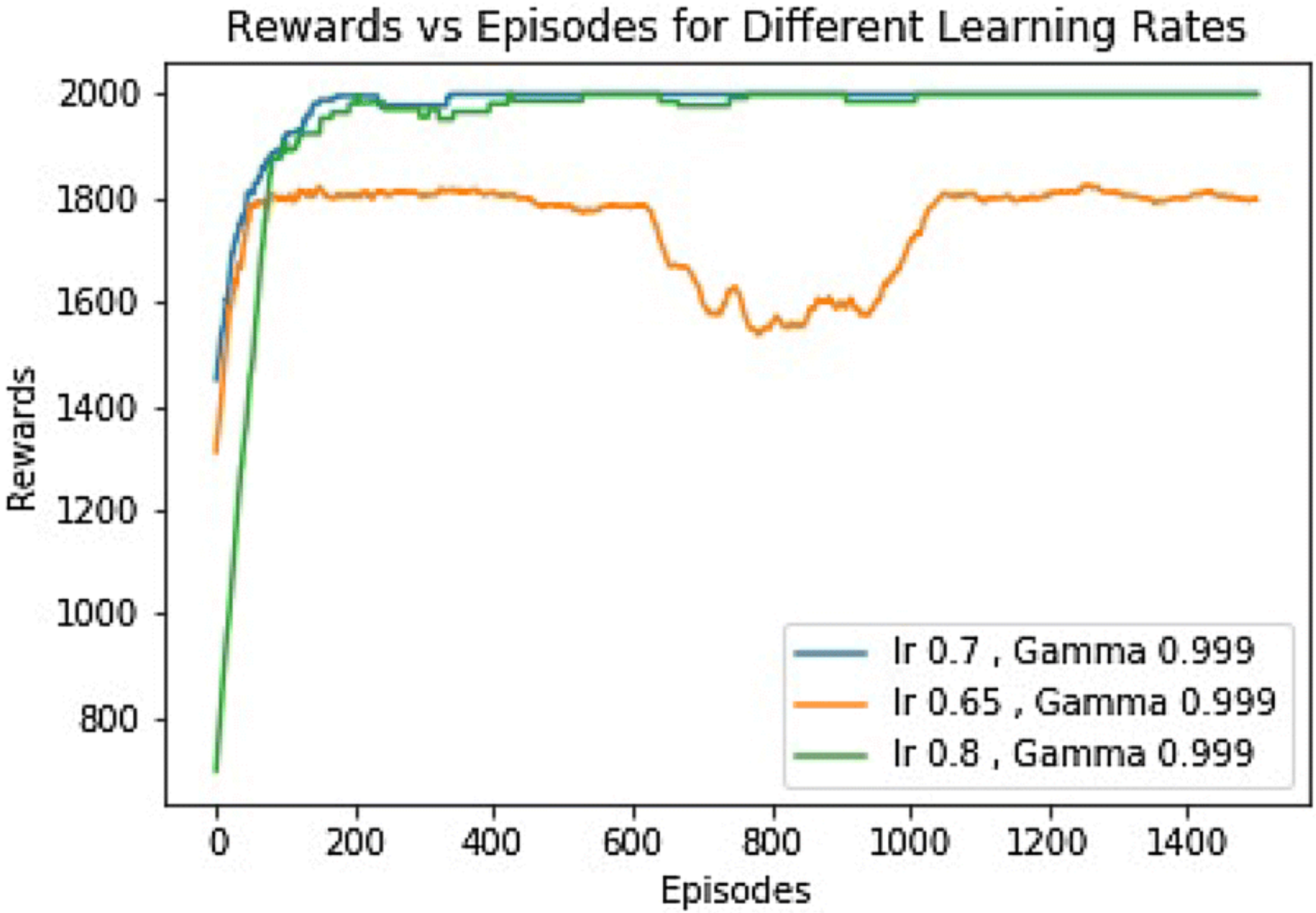
Обучение

Вознаграждения:

- Выход за границы [-5, +5] : *-100*
- Нахождение в интервале [-5, +5] : *+1*



Результаты


Rewards vs Episodes for Different Learning Rates
Rewards
2000
1800
1600
1400
1200
1000
800
0
200
400
600
800
1000
1200
1400
Episodes
lr 0.7 , Gamma 0.999
lr 0.65 , Gamma 0.999
lr 0.8 , Gamma 0.999

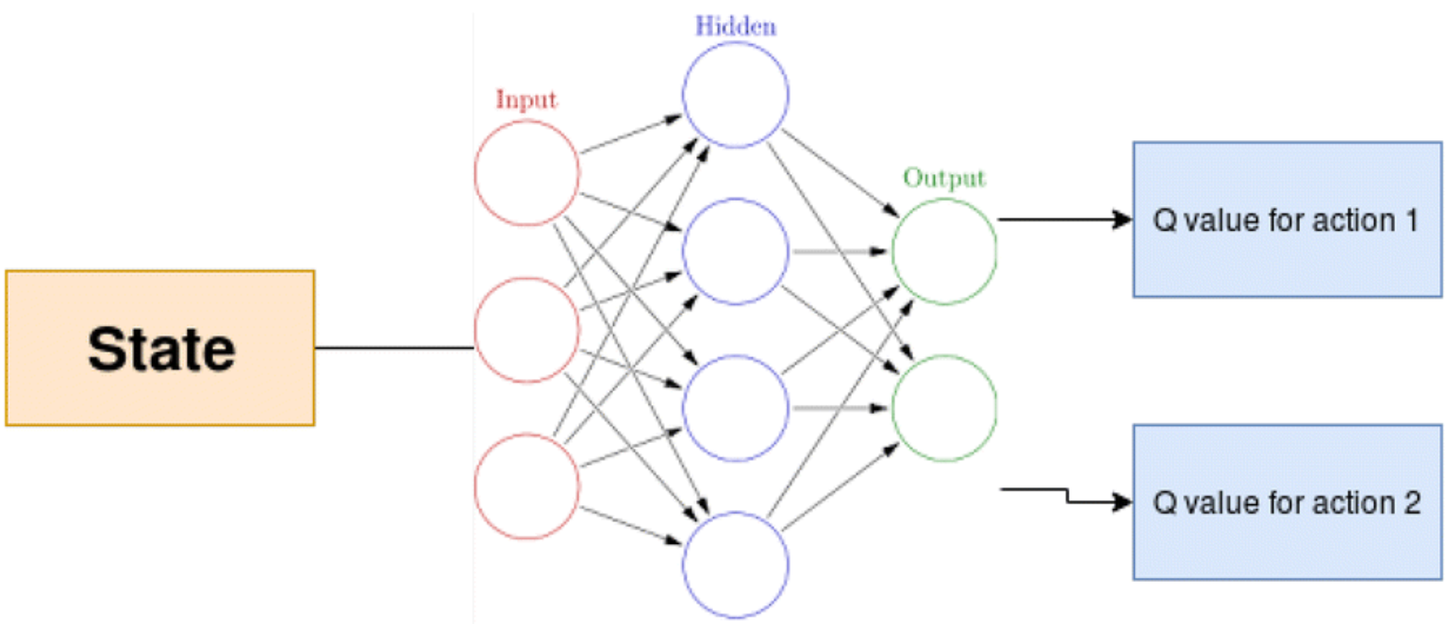
Модификации

"Experience replay"

Сохранение на каждом шаге текущих состояния, награды, действия. При новой инициализации данные не заполняются нулями, а берутся случайным образом из сохранённых

Предсказание значений Q

Вместо постоянного пересчёта значений качества их можно предсказывать по текущему состоянию.



Спасибо за внимание!