ТюмГУ

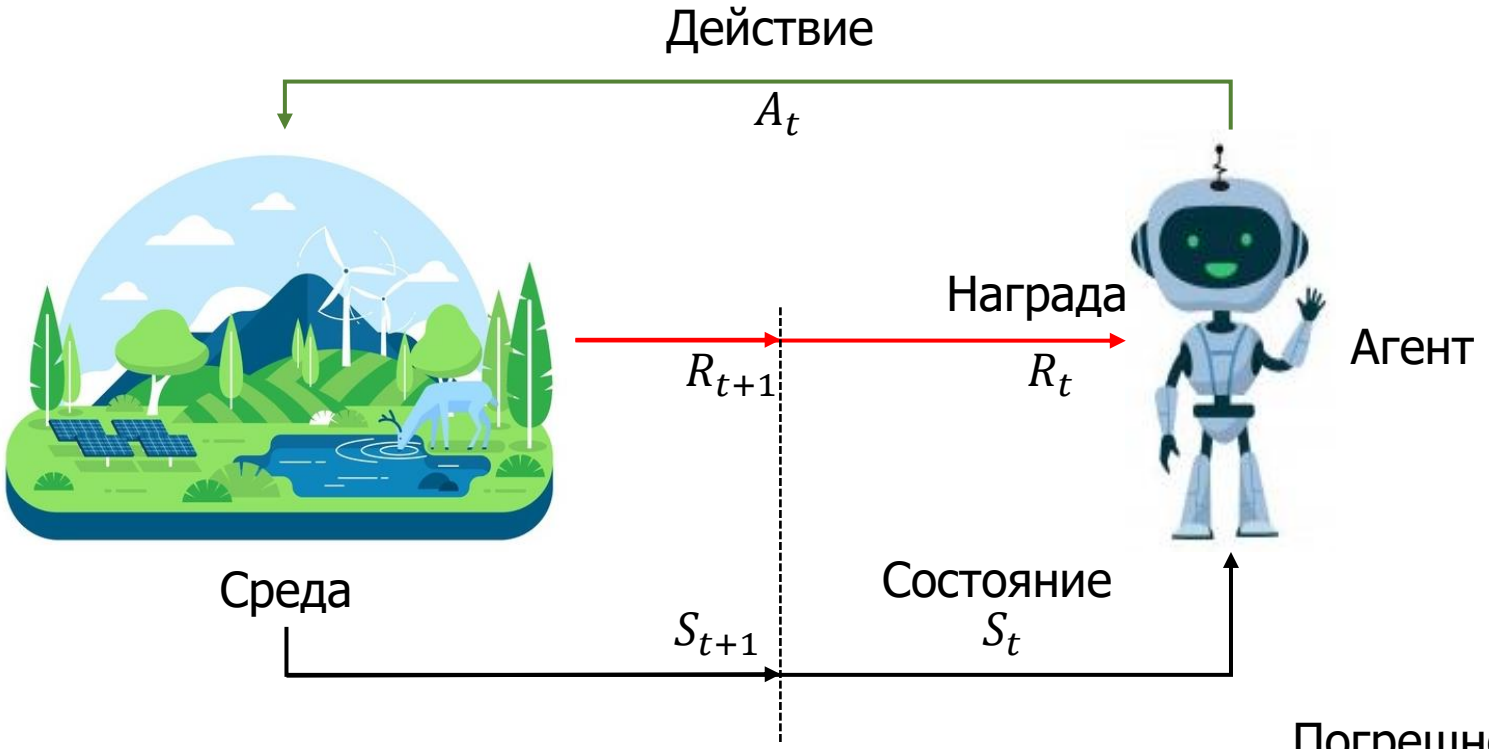
Обучение с подкреплением.

Ценностно-ориентированное глубокое обучение с подкреплением

Сергей Аксёнов,
Доцент отделения информационных технологий
Инженерной школы информационных технологий и робототехники
Томский политехнический университет

Тюмень-2024

Взаимодействие со средой

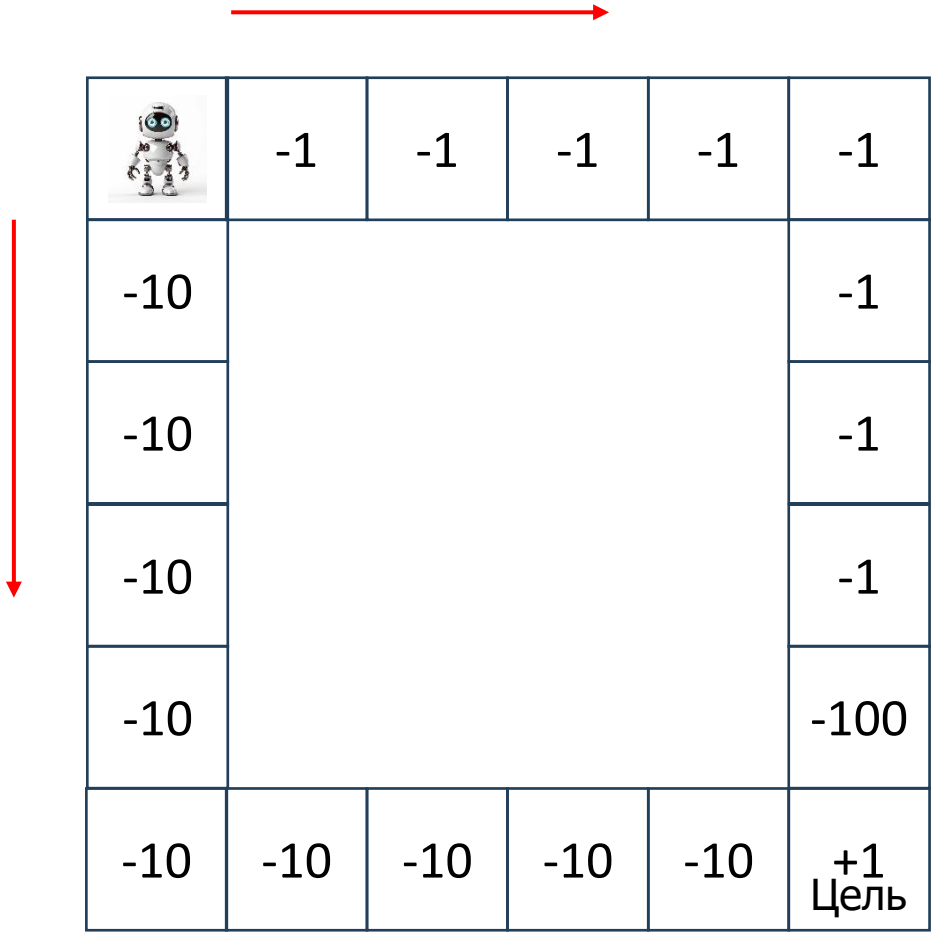


Погрешность Q-обучения

Уравнение Q-обучения: $$Q(S_t, A_t) = Q(S_t, A_t) + \alpha_t \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

Цель Q-обучения

Последовательная обратная связь



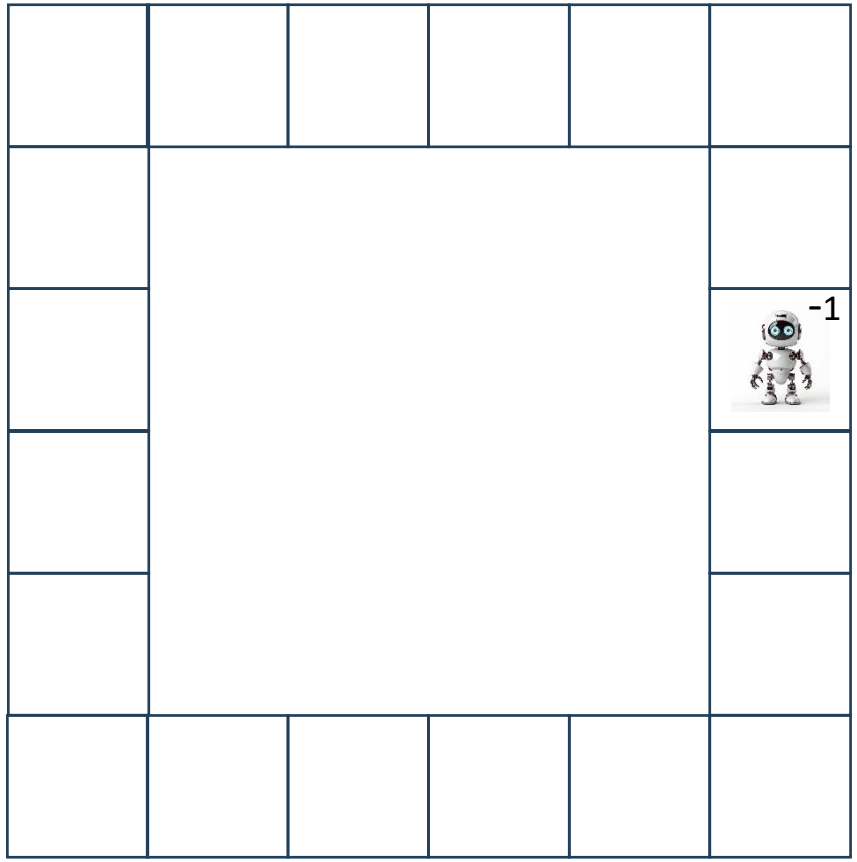
Последовательная связь:

- Награды получаются на каждом шаге: чтобы получить итоговую выгоду нужно закончить эпизод (отложенные последствия действий).

Если у задач нет последствий – одинарная (немедленная) связь:

- В классическом обучении с учителем имеется целевая переменная, связанная с входным вектором признаком.
- У многоруких бандитов результат немедленный после одного действия.

Оценочная обратная связь



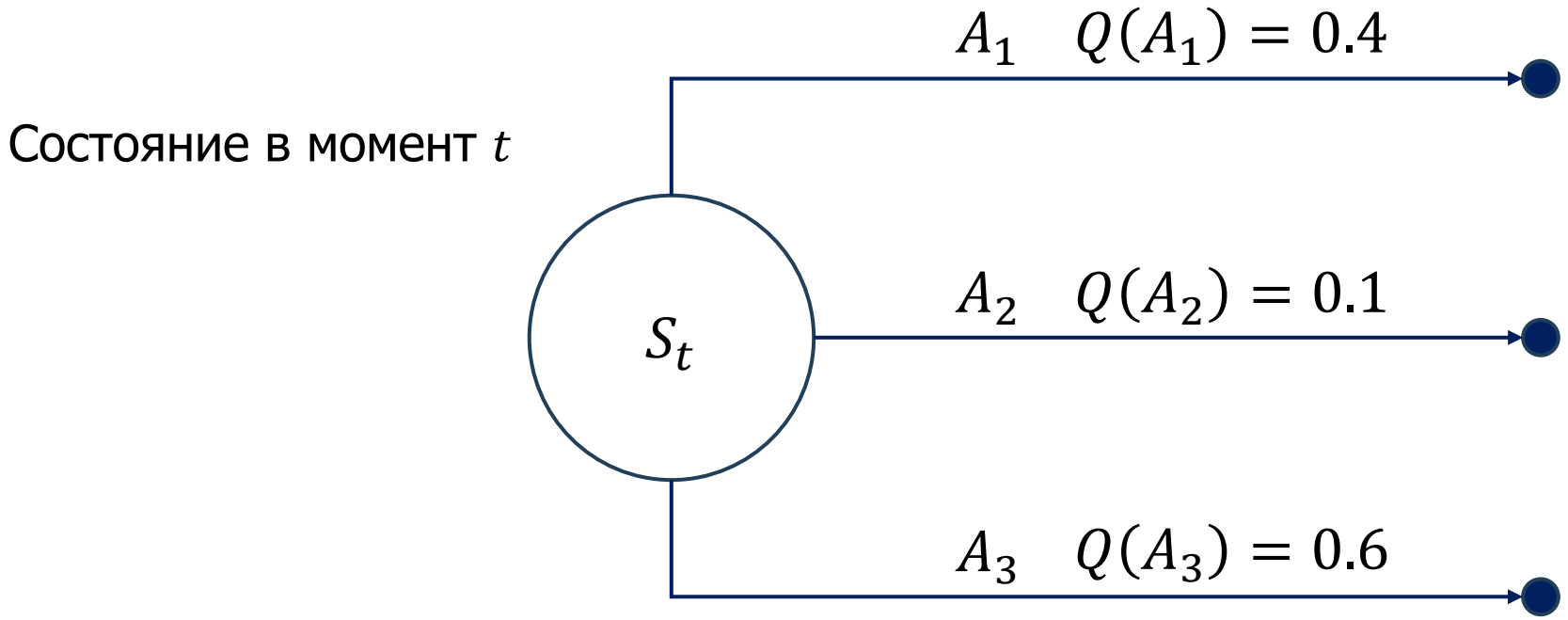
Оценочная связь:

- Неопределённость среды (не известны функции переходов и вознаграждения) приводит, к тому что размер наград не является постоянным при выполнении одного и того же действия.

Противоположность – контролируемая связь:

- В классическом обучении с учителем выборка содержит правильные значения целевой переменной.

Состояния и действия



Возможные действия из состояния: A_i $Q(A_i)$ - ценность действия A_i

Затухающая ε-жадная стратегия

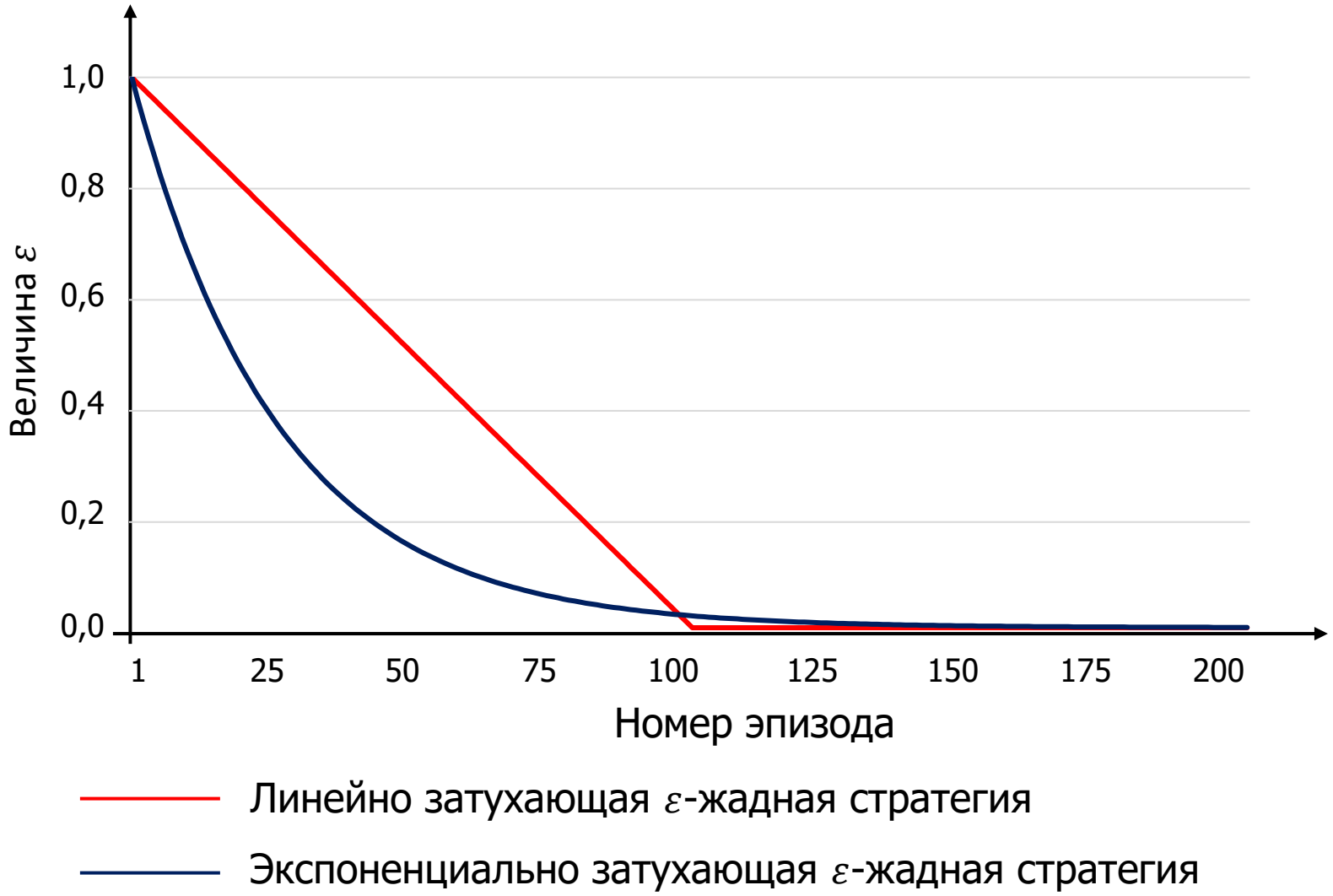
$$n\_episodes = 200$$

$$decay\_ratio(lin) = 0{,}5$$

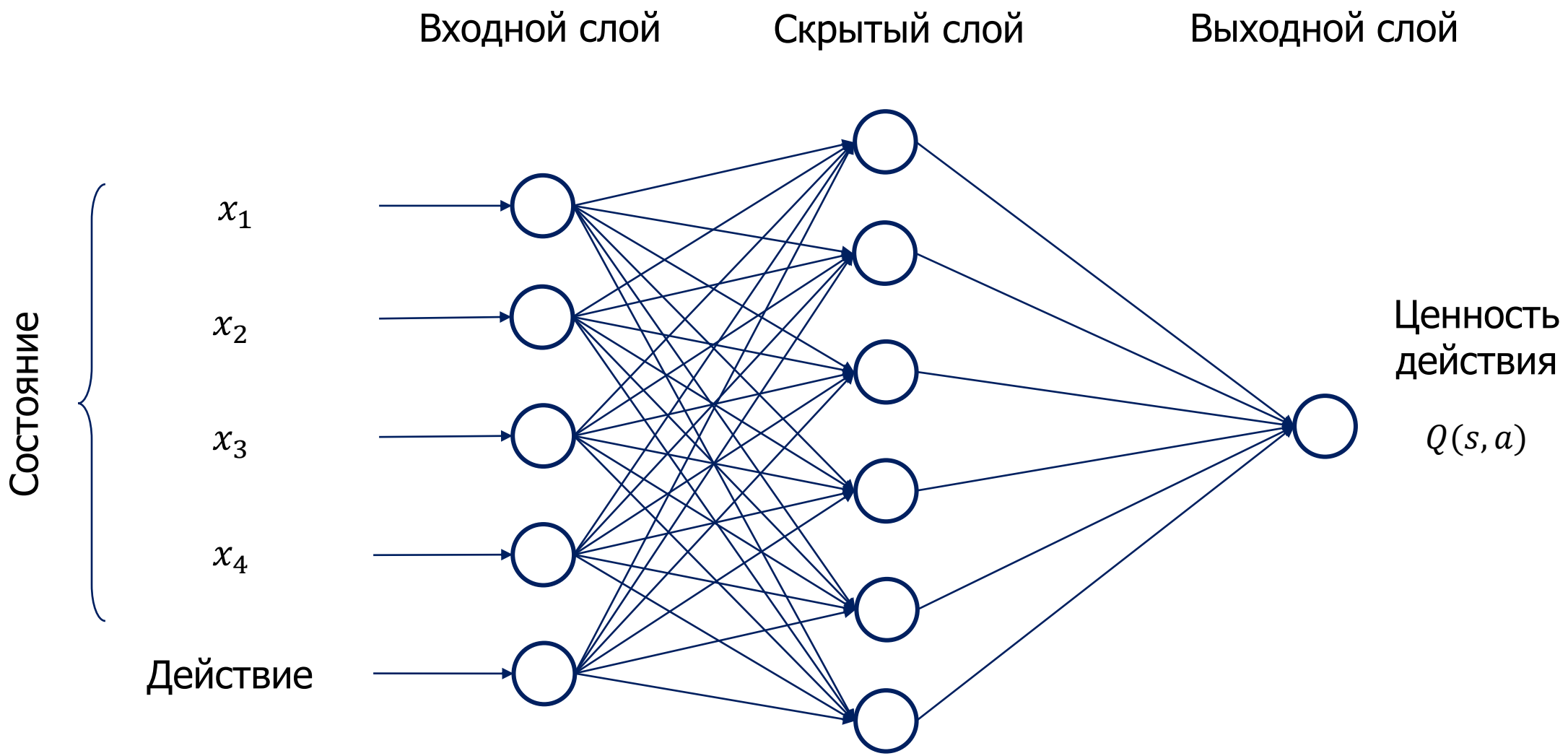
$$decay\_ratio(exp) = 0{,}6$$

$$\varepsilon_{max} = 1{,}0$$

$$\varepsilon_{min} = 0{,}01$$

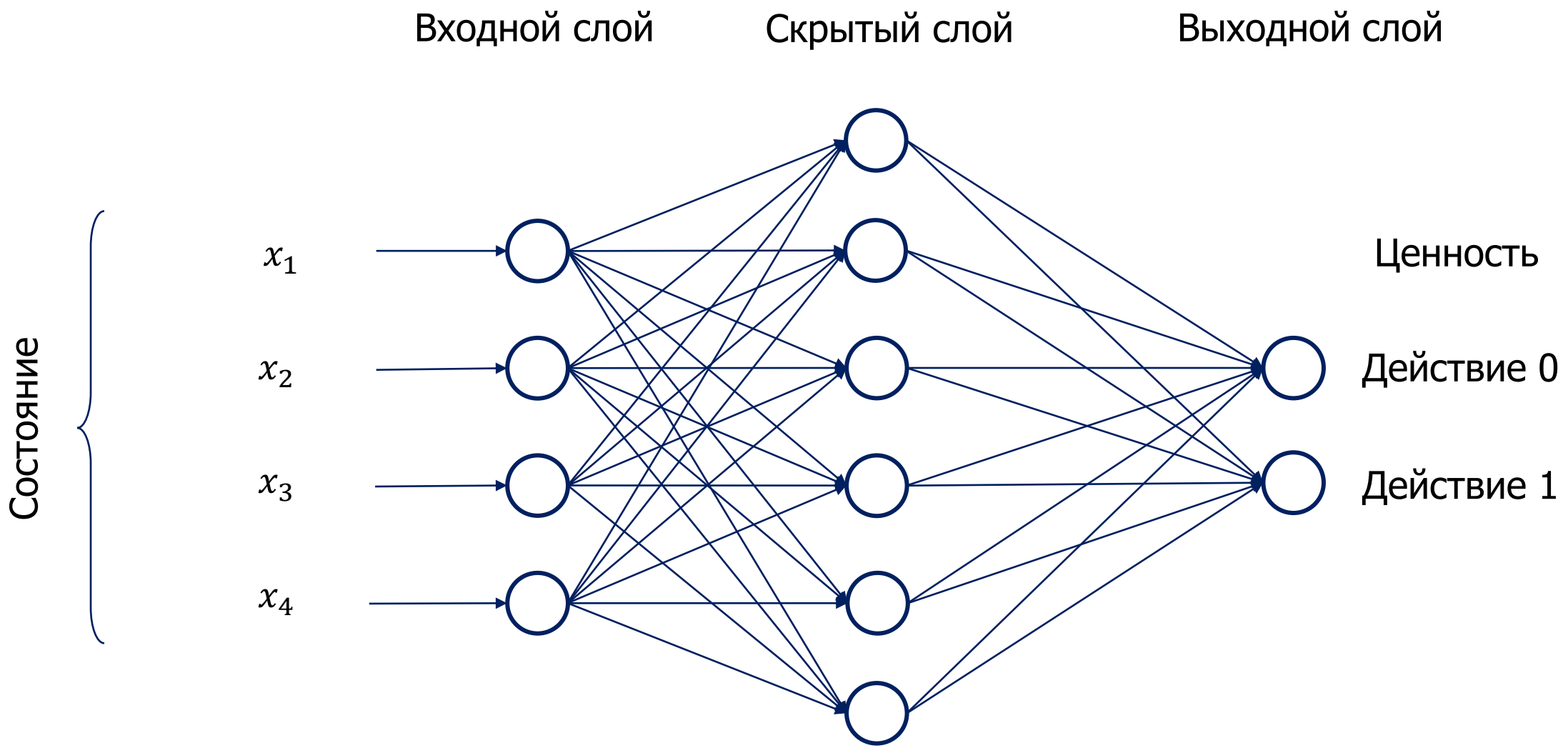


Пример архитектуры: State-action-in-value-out



Архитектура сети определяется особенностями данных

Пример архитектуры: State-in-value-out



Архитектура сети определяется особенностями данных

Оптимизация

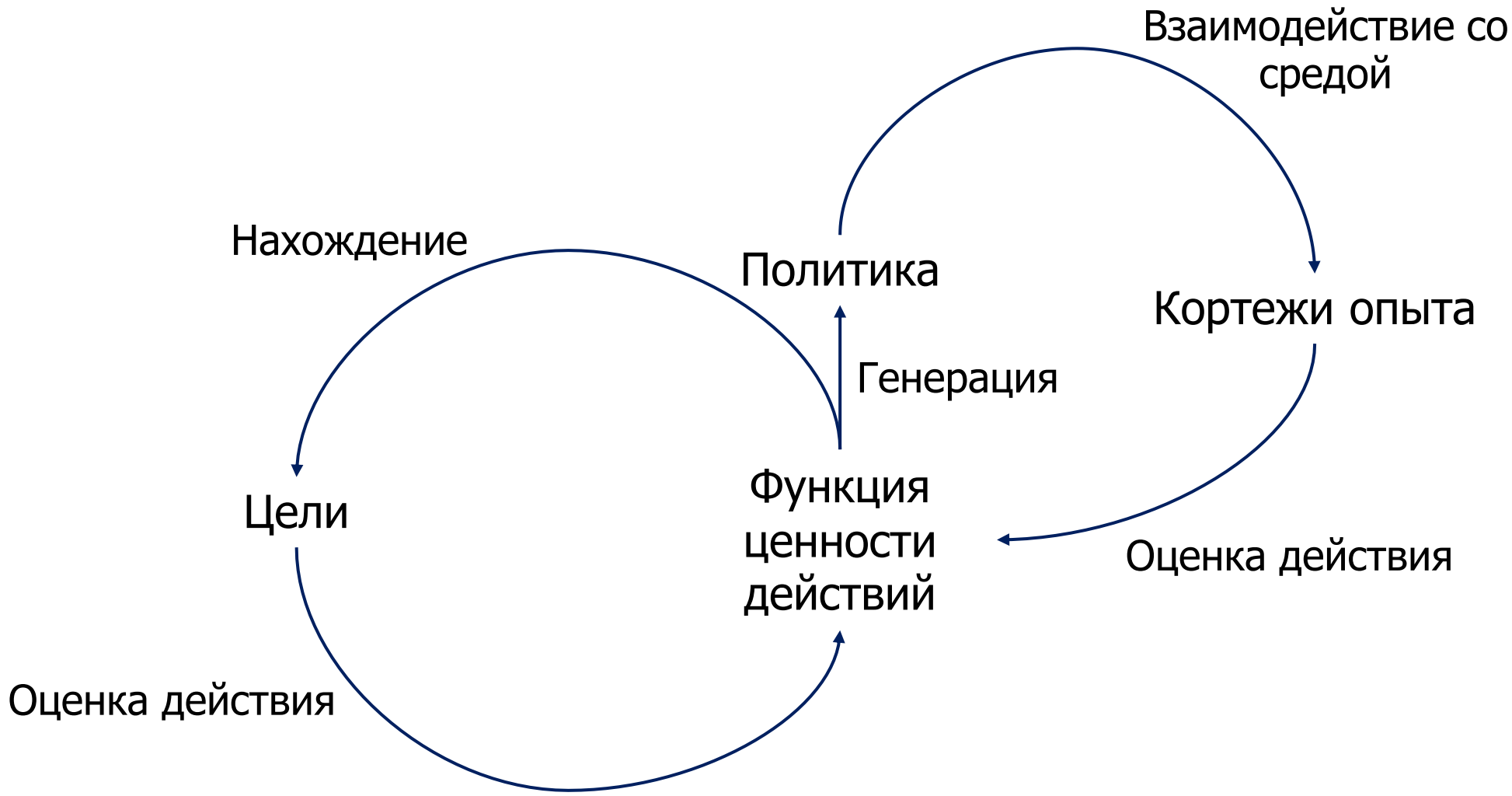
Идеальная цель

$$L_i(\theta_i) = \mathbb{E}_{s,a}\left[\left(q_*(s,a) - Q(s,a;\theta_i)\right)^2\right]$$

Оптимальная функция ценности действий

$$q_*(s,a) = \max_{\pi} \mathbb{E}_{\pi}[G_t | S_t = s, A_t = a], \forall s \in S, \forall a \in A(s)$$

Зависимости функции ценности действий



Выборка для настройки нейросети

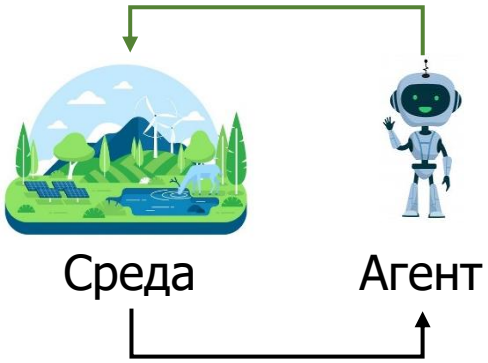
Состояние

Ценность действий

x_1	x_2	x_3	x_4	$Q(A_1)$	$Q(A_2)$

Алгоритм NFQ

Сеть определяет ценность состояний

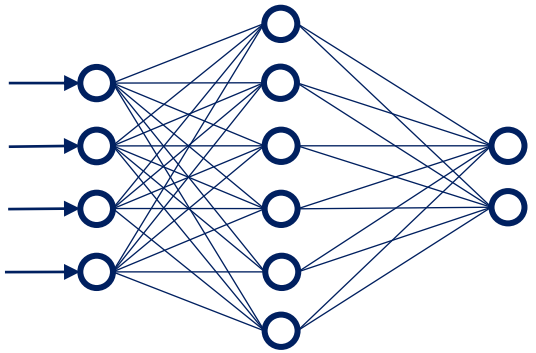


Среда Агент

Взаимодействие
Получение значений Q.
Q-learning

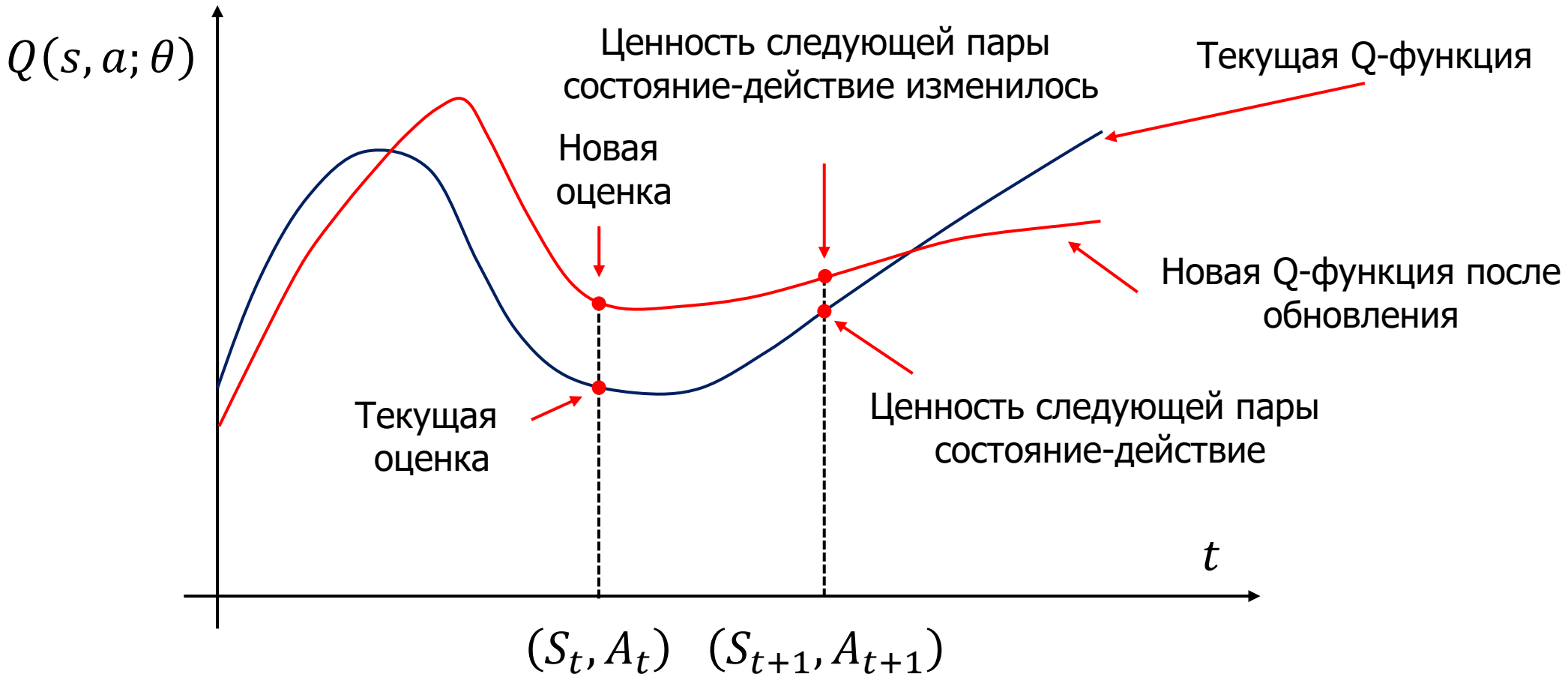
x_1	x_2	x_3	x_4	$Q(A_1)$	$Q(A_2)$

Выборка



Настройка нейросети

Проблемы обучения: Нестационарная цель



Проблемы обучения: Данные

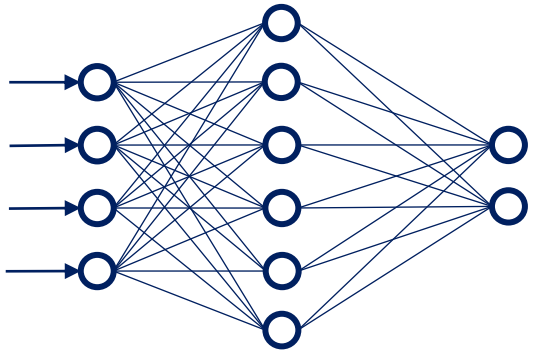
Для хорошего глубокого обучения данные должны быть независимы и одинаково распределены.

Что происходит у нас:

- Данные берутся из эпизода (состояния/ действия зависят друг от друга)
- Данные не распределены одинаково, т.к. зависят от политики, генерирующей действия

Глубокая Q-сеть: DQN

Сеть определяет ценность состояний



Целевая сеть

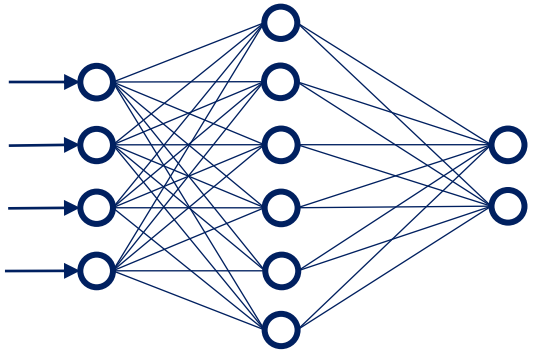
Обновляем веса целевой сети через N шагов

Среда Агент

Взаимодействие
Получение значений Q.
Q-learning

x_1	x_2	x_3	x_4	$Q(A_1)$	$Q(A_2)$

Выборка



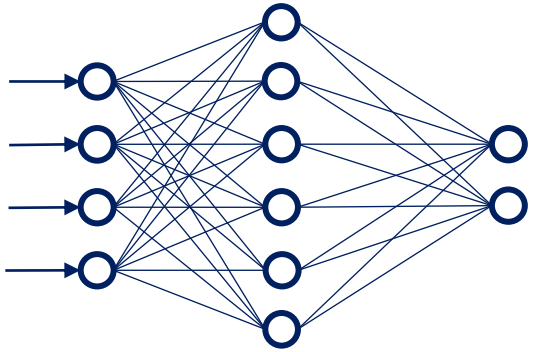
Динамическая сеть

Воспроизведение опыта

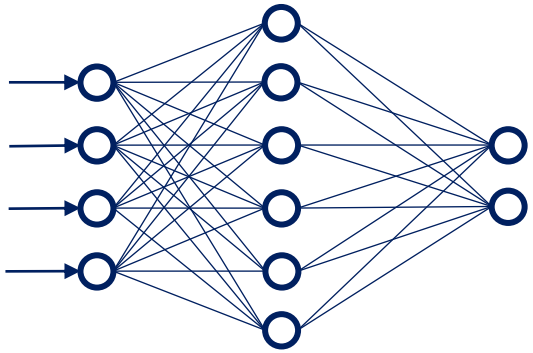
Использование буфера воспроизведения, из которого берутся данные для настройки модели

- Хранение данных из предыдущих выборок
- Данные берутся из разных траекторий и политик
- Повышение стабильности работы процедуры оптимизации
- При достижении максимального размера буфера удаляются более старые записи

Двойная глубокая Q-сеть: DDQN



Целевая нейросеть



Динамическая нейросеть

Оценка лучшего действия

Выбор лучшего действия

Среда

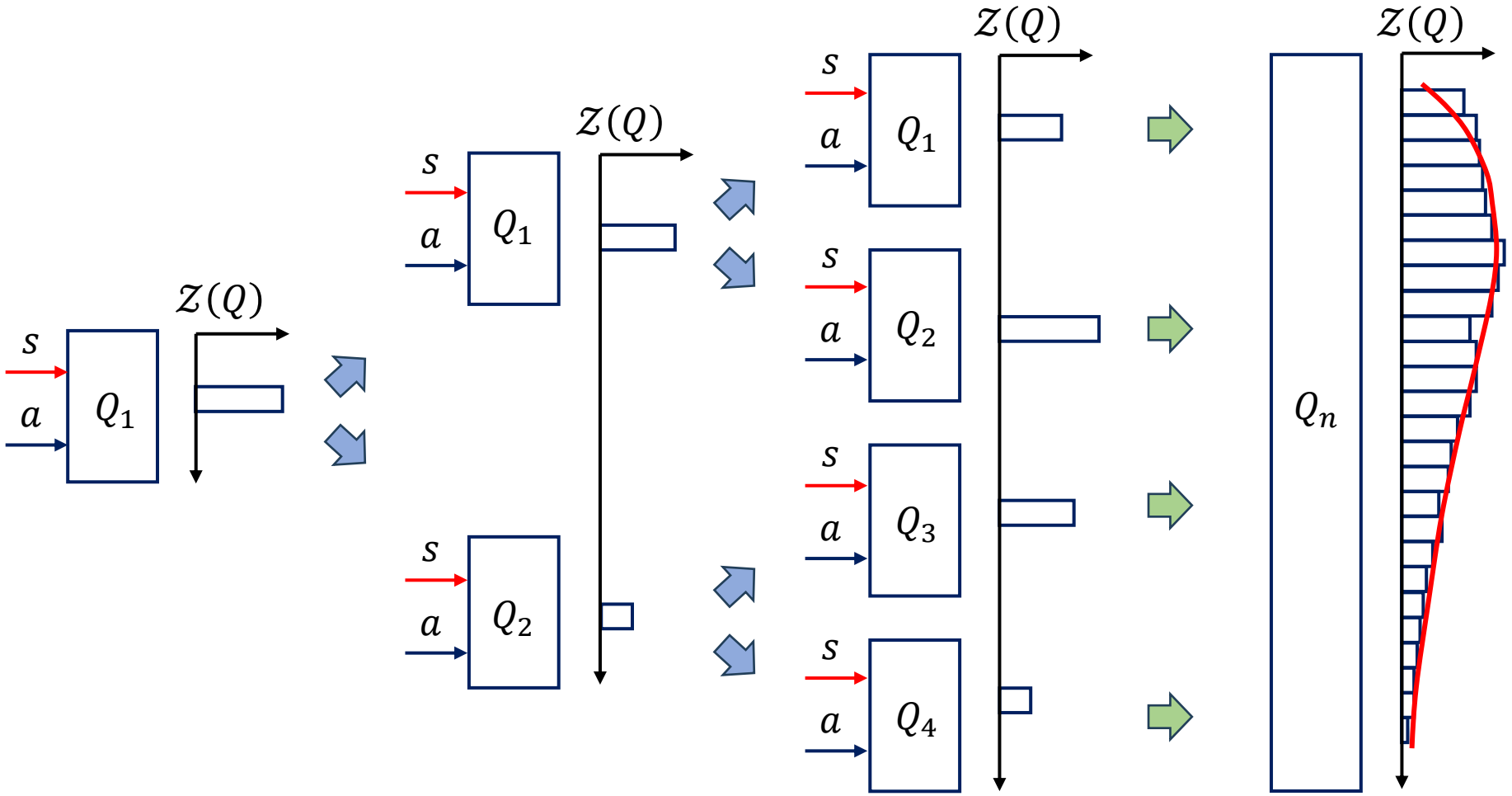
Агент

Взаимодействие
Получение значений Q.
Q-learning

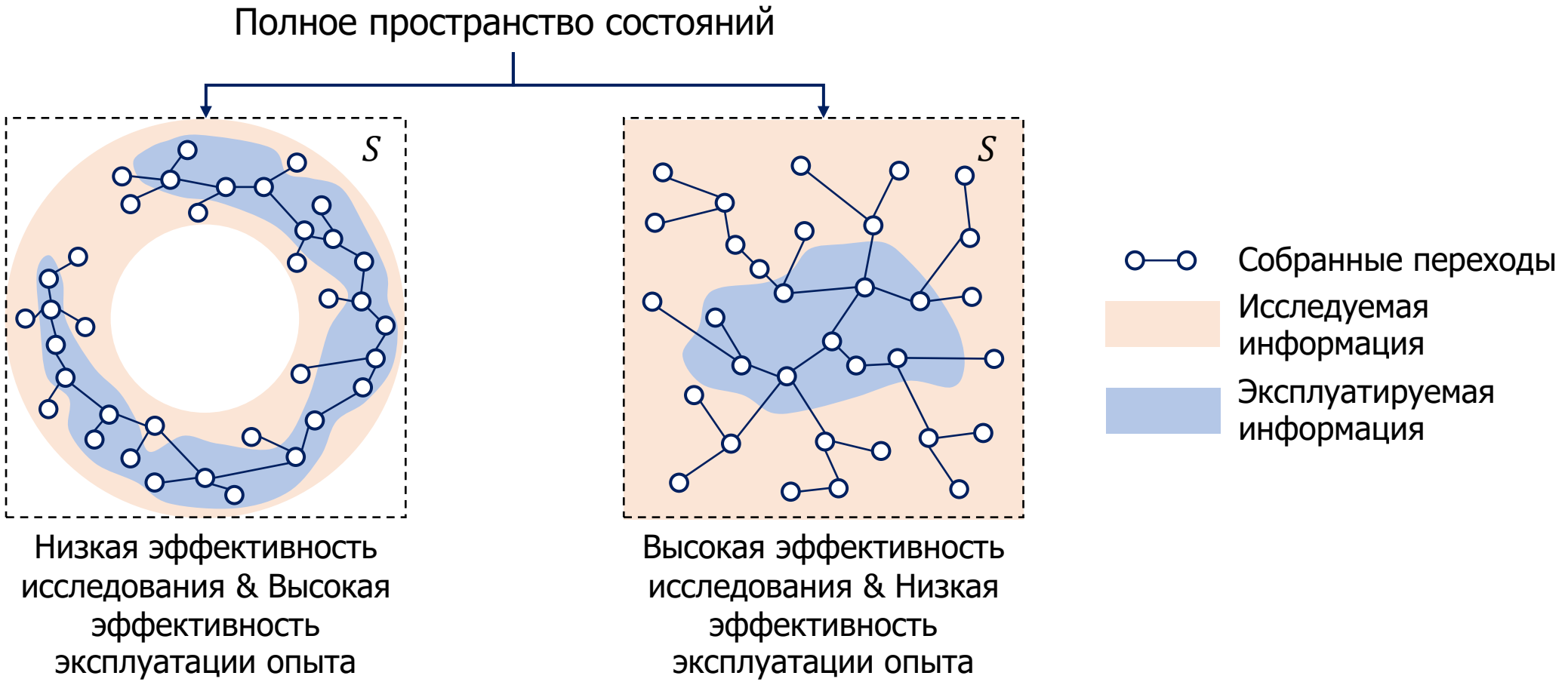
x_1	x_2	x_3	x_4	$Q(A_1)$	$Q(A_2)$

Выборка

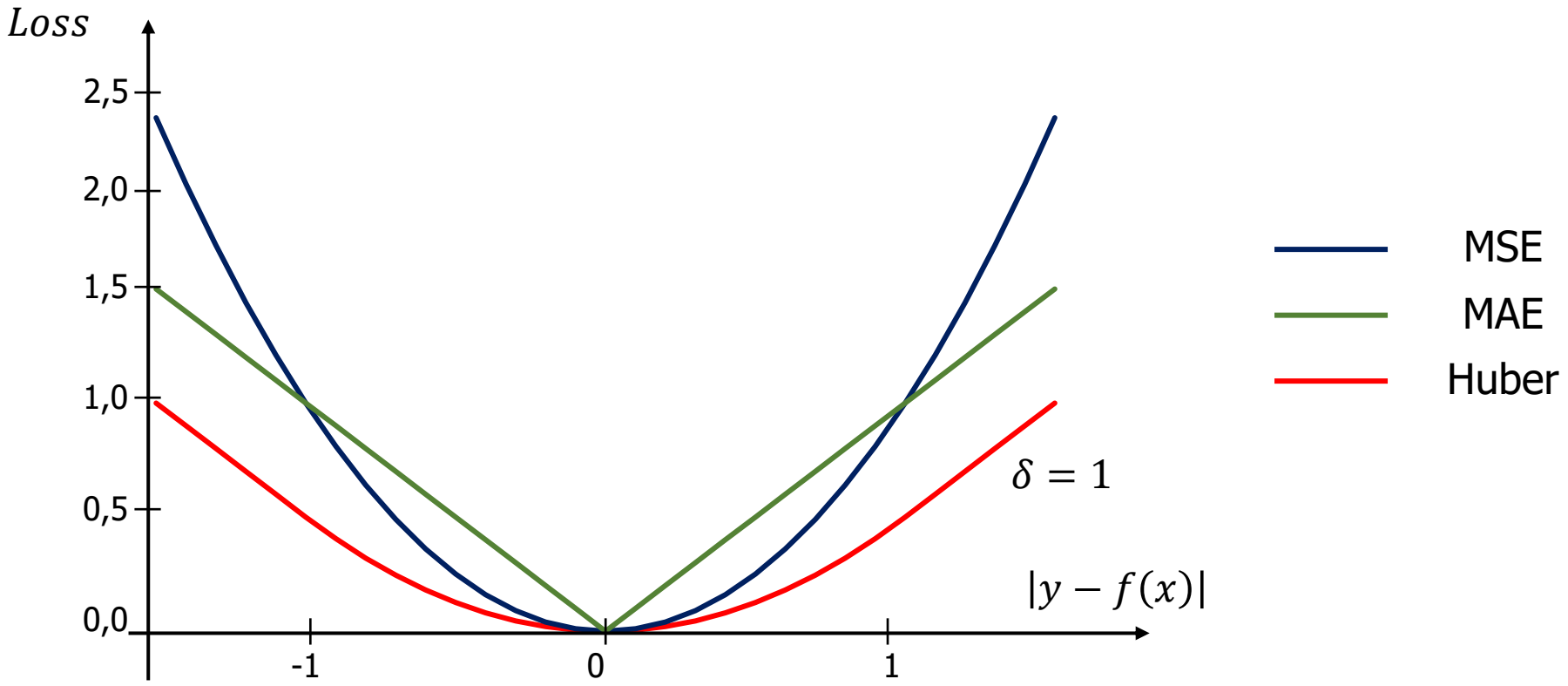
Идея распределенной функции выгоды



Неэффективность кортежей опыта



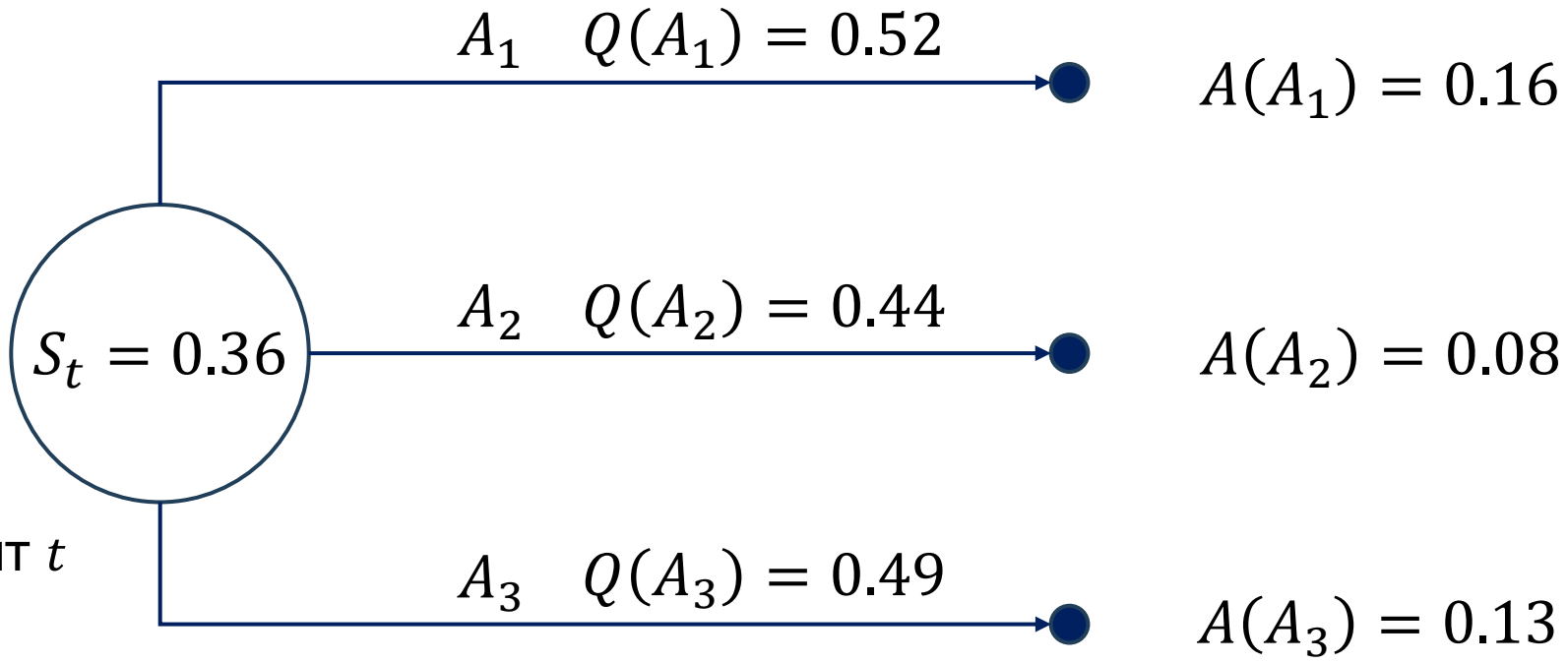
Функция потерь Хьюбера



$$L_\delta\big(y, f(x)\big) = \begin{cases} \frac{1}{2}\big(y - f(x)\big)^2, |y - f(x)| \le \delta \\ \delta \cdot \left(|y - f(x)| - \frac{1}{2}\delta\right), |y - f(x)| > \delta \end{cases}$$

Функции ценности состояния и преимущества действий

Функция преимущества действий: $A_\pi(s,a) \stackrel{\text{def}}{=} Q_\pi(s,a) - V_\pi(s)$

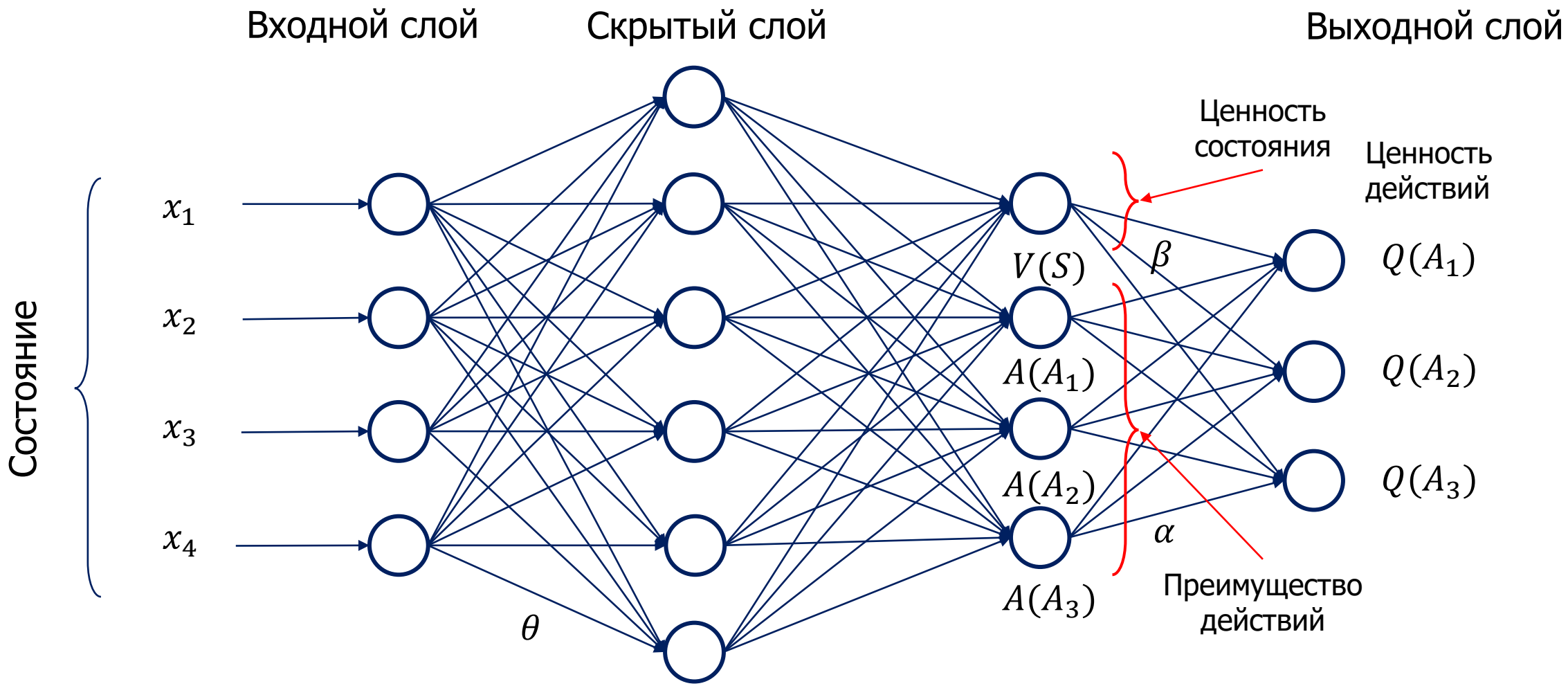


Состояние в момент t

Возможные действия из состояния: A_i $Q(A_i)$ - ценность действия A_i

$A(A_i)$ - преимущество действия A_i

Дуэльная сеть



Архитектура сети определяется особенностями данных

Агрегация дуэльной архитектуры

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + A(s,a;\theta,\alpha)$$

Аппроксимация Q:

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s,a';\theta,\alpha) \right)$$

α - Веса потока преимущества действий

β - Веса потока ценности состояний

θ - Веса разделяемых слоёв

Непрерывное обновление целевой сети

Синхронизация целевой сети с динамической

Усреднение Поляка:

$$\theta_i^- = \tau\theta_i + (1-\tau)\,\theta_i^-$$

$$\alpha_i^- = \tau\alpha_i + (1-\tau)\,\alpha_i^-$$

$$\beta_i^- = \tau\beta_i + (1-\tau)\,\beta_i^-$$

Приоритетное воспроизведение полезного опыта (PER)

Абсолютная погрешность TD – приоритет:

$$|\delta_i| = |\underbrace{\underbrace{r + \gamma Q(s', \underset{a'}{\operatorname{argmax}}\, Q(s', a', \theta_i, \alpha_i, \beta_i); \theta^-, \alpha^-, \beta^-)}_{\text{Цель дуэльной сети}} - Q(s, a; \theta_i, \alpha_i, \beta_i)}_{\text{Погрешность дуэльной сети}}|$$

Абсолютная погрешность дуэльной сети

Назначение приоритетов:

- Жадная приоритизация
- Стохастическая выборка приоритетного опыта
- Пропорциональная приоритизация
- На основе ранжирования

Пропорциональная приоритизация

Приоритет выборки: $p_i = |\delta_i| + \varepsilon$

Вероятность извлечения выборки: $P(i) = \frac{p_k^\alpha}{\sum_k p_k^\alpha}$

$\alpha = 0$ Равные приоритеты

$\alpha = 1$ Вероятность ~ Абсолютная погрешность TD

Приоритизация на основе ранжирования

Приоритет выборки: $$p_i = \frac{1}{rank(i)}$$

$rank(i)$ - Величина порядкового номера выборки после ранжирования по абсолютной погрешности TD

Вероятность извлечения выборки: $$P(i) = \frac{p_k^\alpha}{\sum_k p_k^\alpha}$$

Взвешивание выборки по значимости

Вес – значимость выборки: $w_i = \left(N \times P(i)\right)^{-\beta}$

N	Число выборок в буфере воспроизведения
β	Управление важностью взвешивания
$\beta = 0$	Коррекция не выполняется
$\beta = 1$	Полная коррекция смещения

Нормализация весов: $w_i = \dfrac{w_i}{\max\limits_j \left(w_j\right)}$