

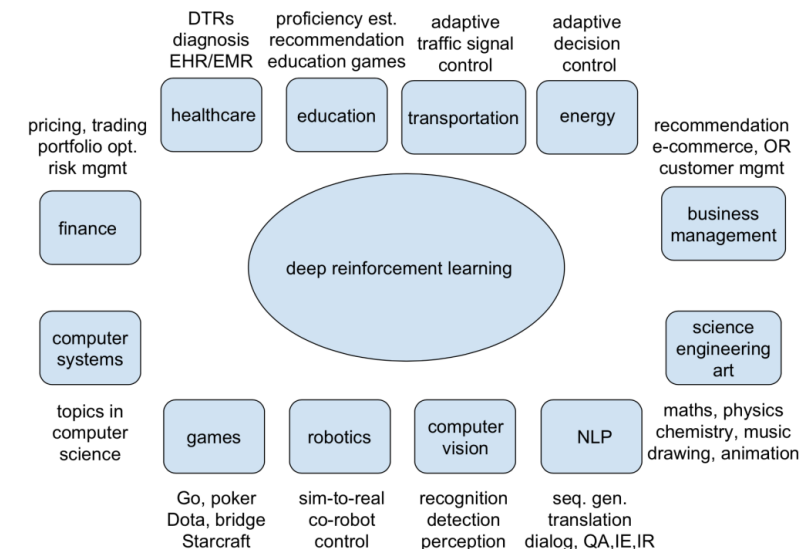
Осенняя Школа по информационным технологиям ОИЯИ  
7 – 11 октября 2023

# Глубокое обучение с подкреплением (Deep Reinforcement learning)

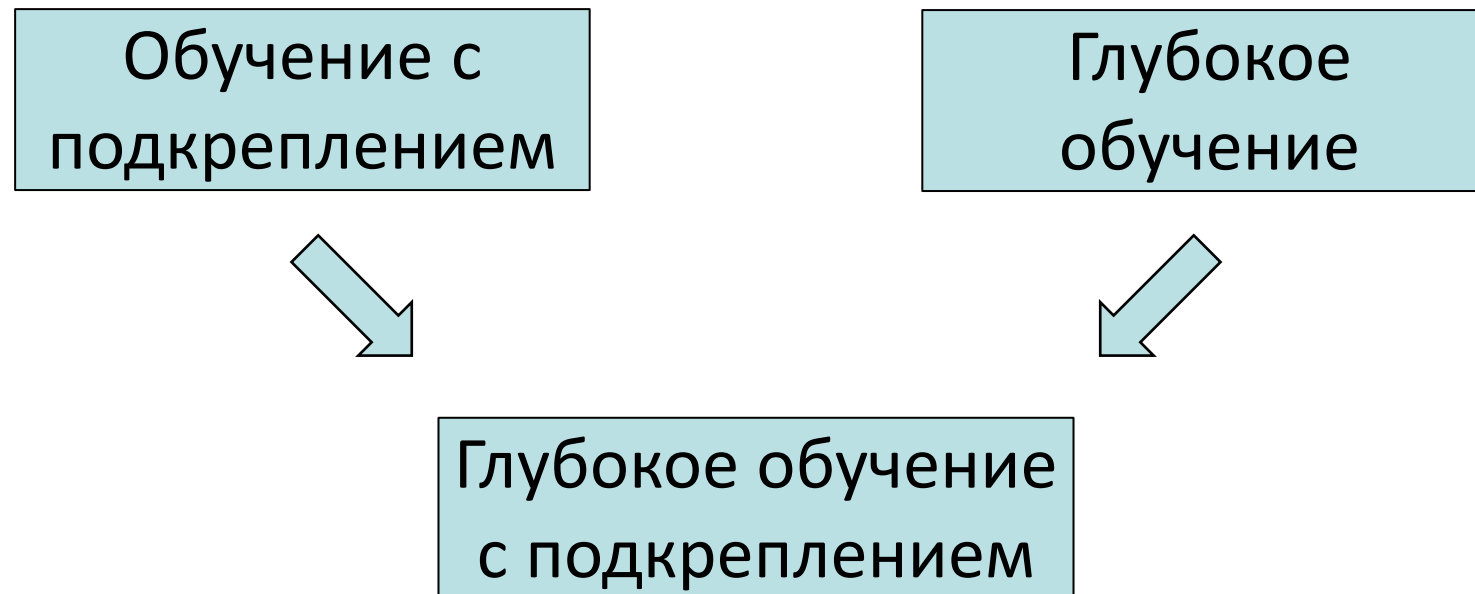
Соколинский Леонид Борисович  
доктор физ.-мат. наук, профессор

# Где применяется обучение с подкреплением

- Игры (шахматы, нарды, го и др.)
- Беспилотный автомобиль (self-driving car)
- Автоматические системы управления технологическими процессами (industry automation)
- Роботрейдинг: биржевые финансовые сделки (trading and finance)
- Обработка естественного языка (NLP - natural language processing)
- Рекомендательные сервисы
- Роботы-манипуляторы
- Маркетинг и реклама
- Здоровоохранение
- ...

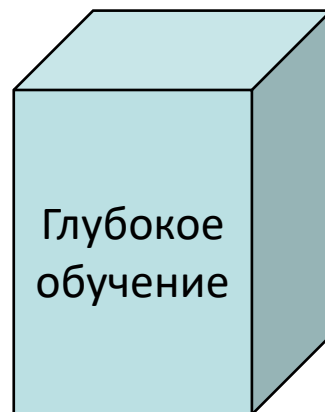
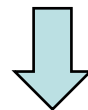


# ЭТИМОЛОГИЯ

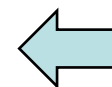


# Глубокое обучение

Необученная глубокая  
нейронная сеть

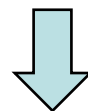


Глубокое  
обучение



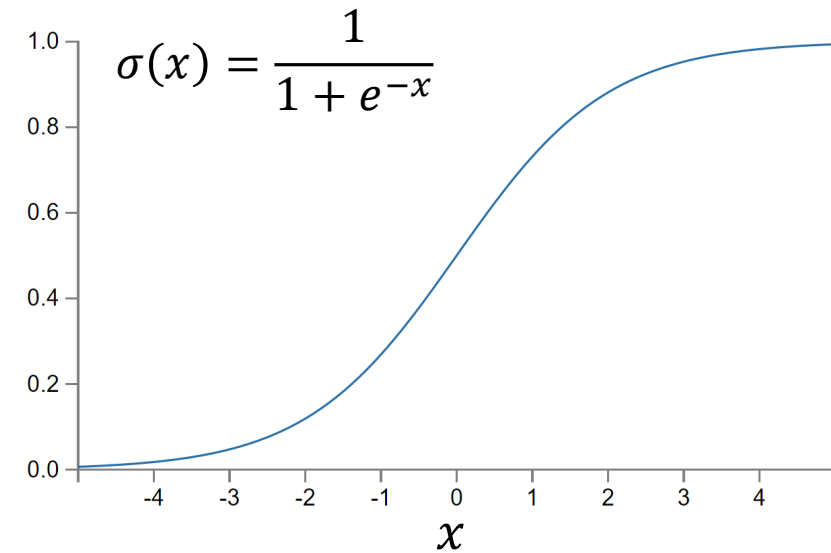
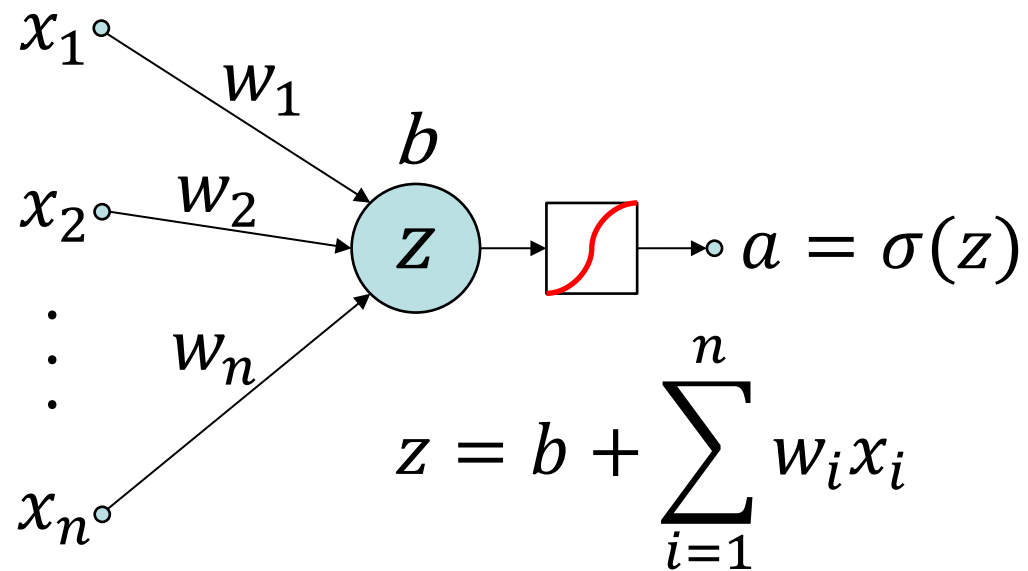
Множество  
прецедентов

Обученная глубокая  
нейронная сеть



Прецедент = (Данные задачи, Правильный ответ)

# Искусственный нейрон сигмоид



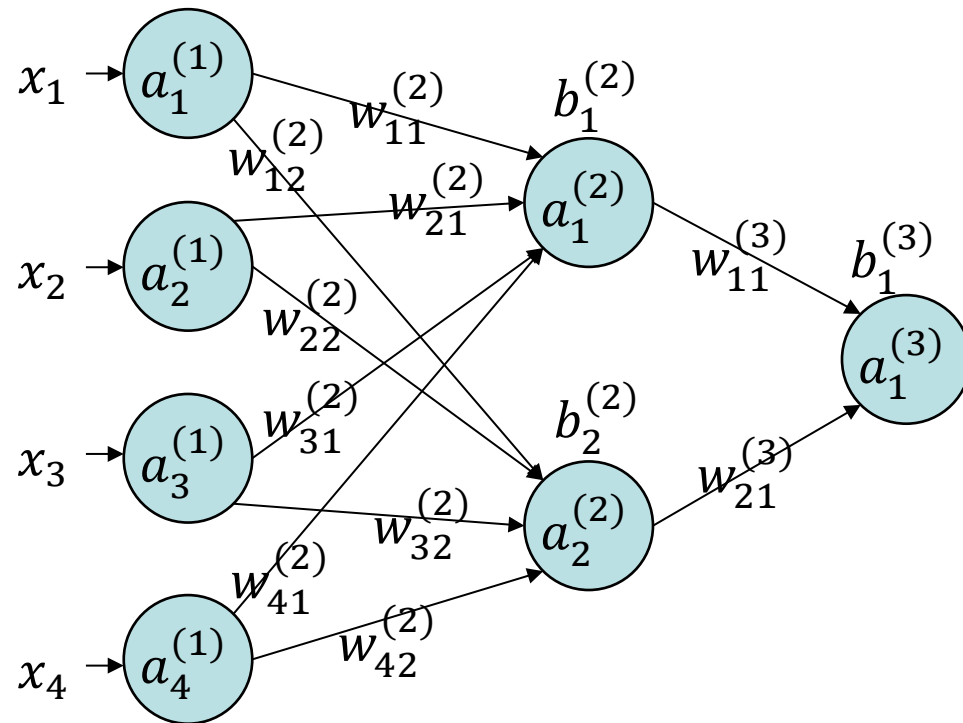
$\mathbf{x} = (x_1, x_2, \dots, x_n)$  – входные сигналы:  $0 \leq x_i \leq 1$

$\mathbf{w} = (w_1, w_2, \dots, w_n)$  – синаптические веса:  $w_i \in \mathbb{R}$

$b$  – смещение:  $b \in \mathbb{R}$

$a$  – выходной сигнал:  $0 < a < 1$

# Вычисление выходного сигнала в нейронной сети



$$a_j^{(1)} = x_j \quad (j = 1, \dots, 4)$$

$$z_1^{(2)} = b_1^{(2)} + w_{11}^{(2)} a_1^{(1)} + w_{21}^{(2)} a_2^{(1)} + w_{31}^{(2)} a_3^{(1)} + w_{41}^{(2)} a_4^{(1)}$$

$$a_1^{(2)} = \sigma(z_1^{(2)})$$

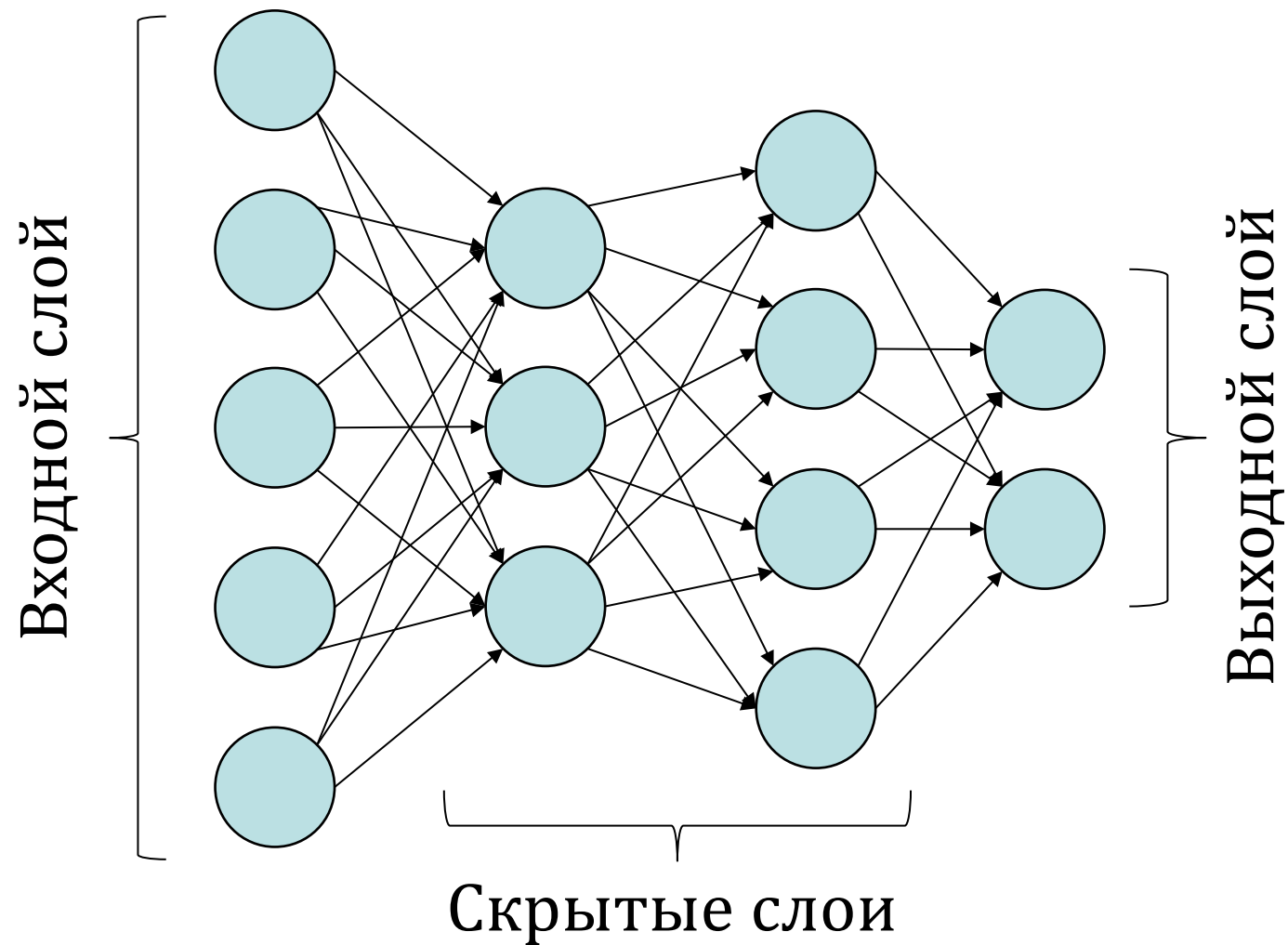
$$z_2^{(2)} = b_2^{(2)} + w_{21}^{(2)} a_1^{(1)} + w_{22}^{(2)} a_2^{(1)} + w_{23}^{(2)} a_3^{(1)} + w_{24}^{(2)} a_4^{(1)}$$

$$a_2^{(2)} = \sigma(z_2^{(2)})$$

$$z_1^{(3)} = b_1^{(3)} + w_{11}^{(3)} a_1^{(2)} + w_{12}^{(3)} a_2^{(2)}$$

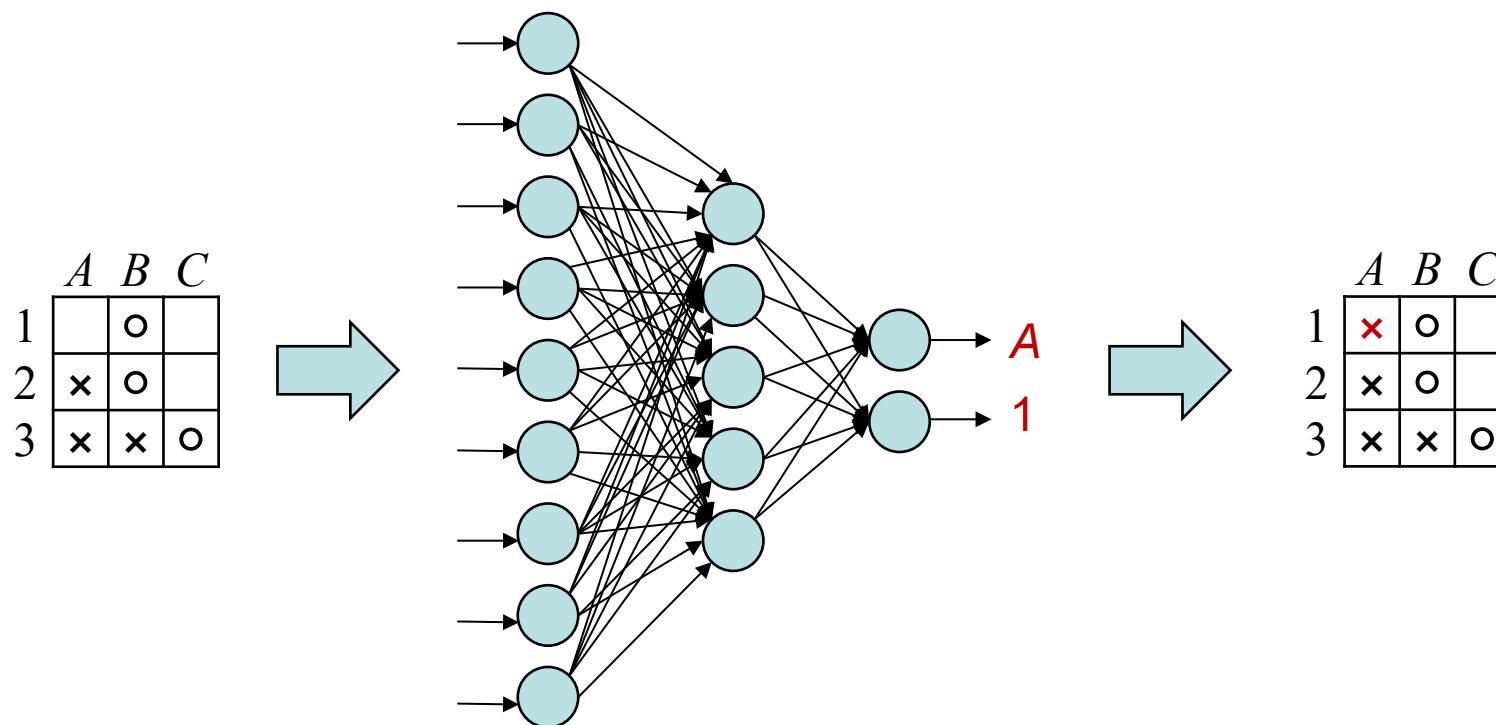
$$a_1^{(3)} = \sigma(z_1^{(3)})$$

# Глубокая нейронная сеть



# Пример с «крестиками-ноликами»

Наивный взгляд на глубокое обучение с подкреплением

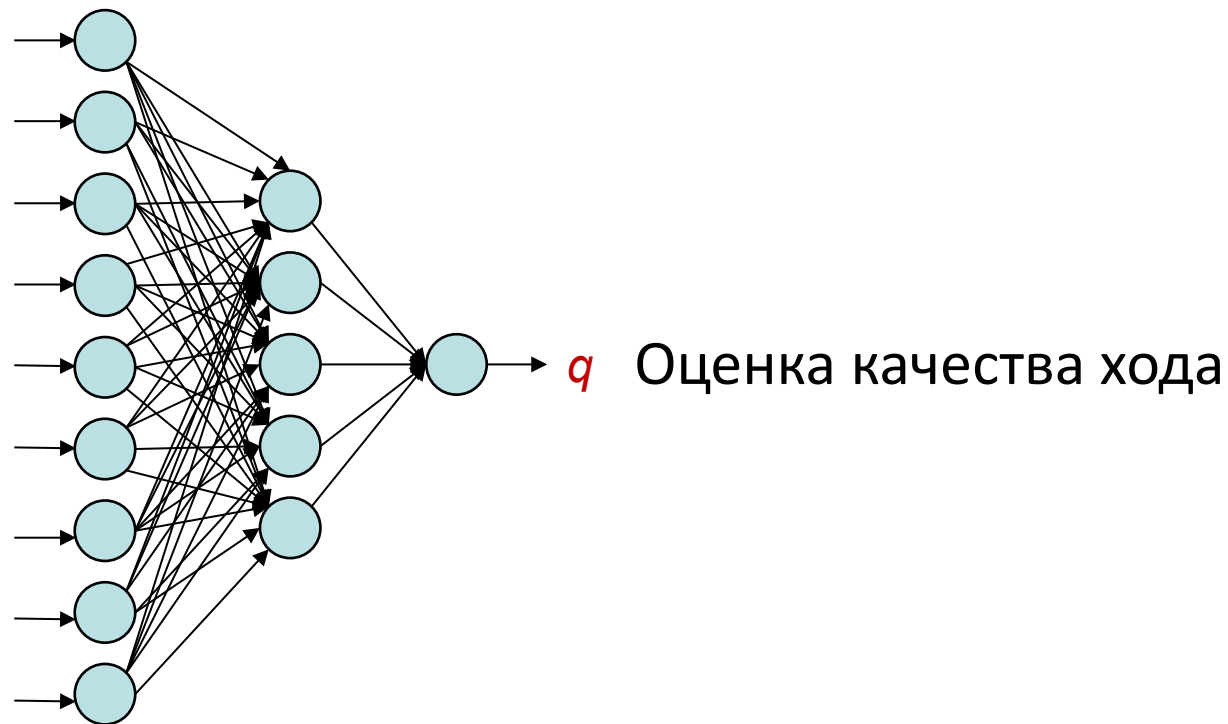
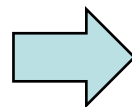




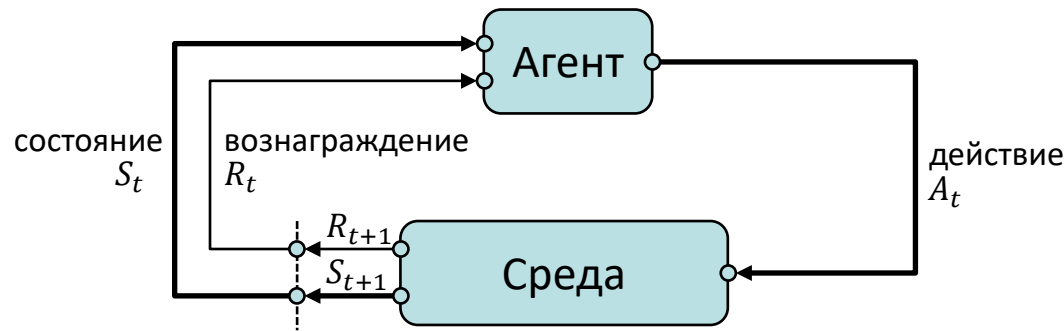
# Пример с «крестиками-ноликами»

## Реальный взгляд на глубокое обучение с подкреплением

	<i>A</i>	<i>B</i>	<i>C</i>
1	×	○	
2	×	○	
3	×	×	○



# Марковский процесс



© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

## Андрей Андреевич Марков



- А. А. Марков был сыном чиновника Андрея Григорьевича Маркова, служившего в Лесном департаменте в чине коллежского советника, а затем вышедшего в отставку и служившего в Санкт-Петербурге частным поверенным.
- Андрей Марков страдал туберкулёзом коленного сустава и до 10 лет ходил на костылях. После операции, проведённой известным хирургом Кады, он получил возможность ходить нормально.
- В 1866 году его отдали в 5-ю Петербургскую гимназию. Это классическое учебное заведение с преподаванием древних языков (латинского и греческого) пришлось ему не по вкусу; по большинству предметов он учился плохо, исключение составлял только один предмет — математика.
- В 1874 году А. А. Марков окончил гимназию и поступил в Санкт-Петербургский университет. Там он слушал лекции профессоров А. Н. Коркина и Е. И. Золотарёва, а также Пафнутия Львовича Чебышёва, оказавшего определяющее влияние на выбор научной деятельности Андрея Маркова. 31 мая 1878 года он окончил Петербургский университет по математическому разряду физико-математического факультета со степенью кандидата.
- С 13 декабря 1886 года, по предложению Чебышёва, он был избран адъюнктом физико-математического отделения (чистая математика); с 3 марта 1890 года — экстраординарный академик, а с 2 марта 1896 года — ординарный академик Императорской Санкт-Петербургской академии наук. С 1880 года — приват-доцент, с 1886 года — профессор физико-математического факультета Санкт-Петербургского университета. С 1898 года — действительный статский советник.
- Умер в Петрограде в 1922 году. Похоронен на Митрофаньевском кладбище Санкт-Петербурга. В 1954 году перезахоронен на Литераторских мостках, Волжское кладбище.

Андрей Андреевич Марков (2 июня 1856 — 20 июля 1922) — русский математик, академик, внесший большой вклад в теорию вероятностей, математический анализ и теорию чисел.

Осенняя Школа по информационным технологиям ОИЯИ

- *Агент (agent)* — сторона, которая обучается и принимает решения
- *Среда (environment)* — сторона, с которой агент взаимодействует (игра + противник + судья)
- *Вознаграждение (reward)* — числовое значение, генерируемое судьей в зависимости от успешности действия агента (в простейшем случае: 1 — выигрыш, 0 — проигрыш)
- Агент и среда взаимодействуют на каждом шаге дискретной последовательности временных шагов:  $t = 0, 1, 2, 3, \dots$
- На каждом шаге  $t$  агент получает *состояние (state) среды* (ответный ход противника)  $S_t \in \mathcal{S}$ , на основе которого выполняет *действие (action)*  $A_t \in \mathcal{A}$
- На следующем шаге  $t + 1$  агент получает *вознаграждение*  $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$  и переходит в новое состояние  $S_{t+1} \in \mathcal{S}$

$$(S_0, A_0, 0) \rightarrow (S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow (S_3, A_3, R_3) \rightarrow \dots$$

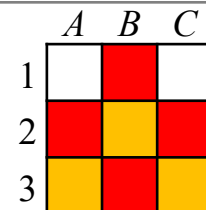
# «Цветные крестики-нолики»

- Агент начинает игру и ставит «крестики»
- Противник (среда) ставит «нолики»
- Можно использовать только белые клетки
- Судья (среда) после каждого хода освобождает одну желтую клетку
- Побеждает тот, кто поставит три своих знака по диагонали, вертикали или горизонтали
- За правильные ходы начисляются очки (вознаграждение)
- За победу назначается вознаграждение **+20**

	A	B	C
1		■	
2	■	■	■
3	■	■	■

# Дерево игры

↓ Начальная позиция

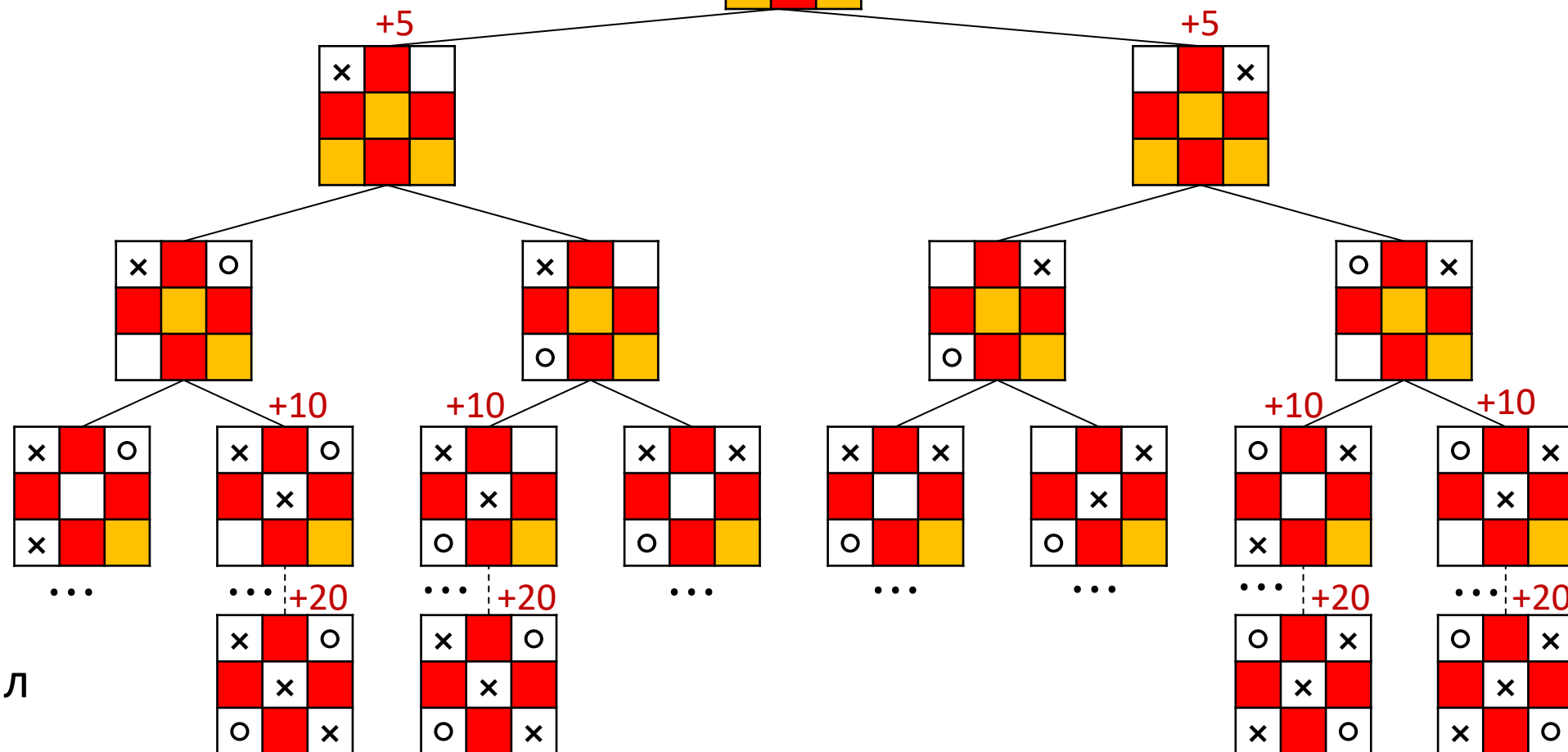


↓ Ход противника

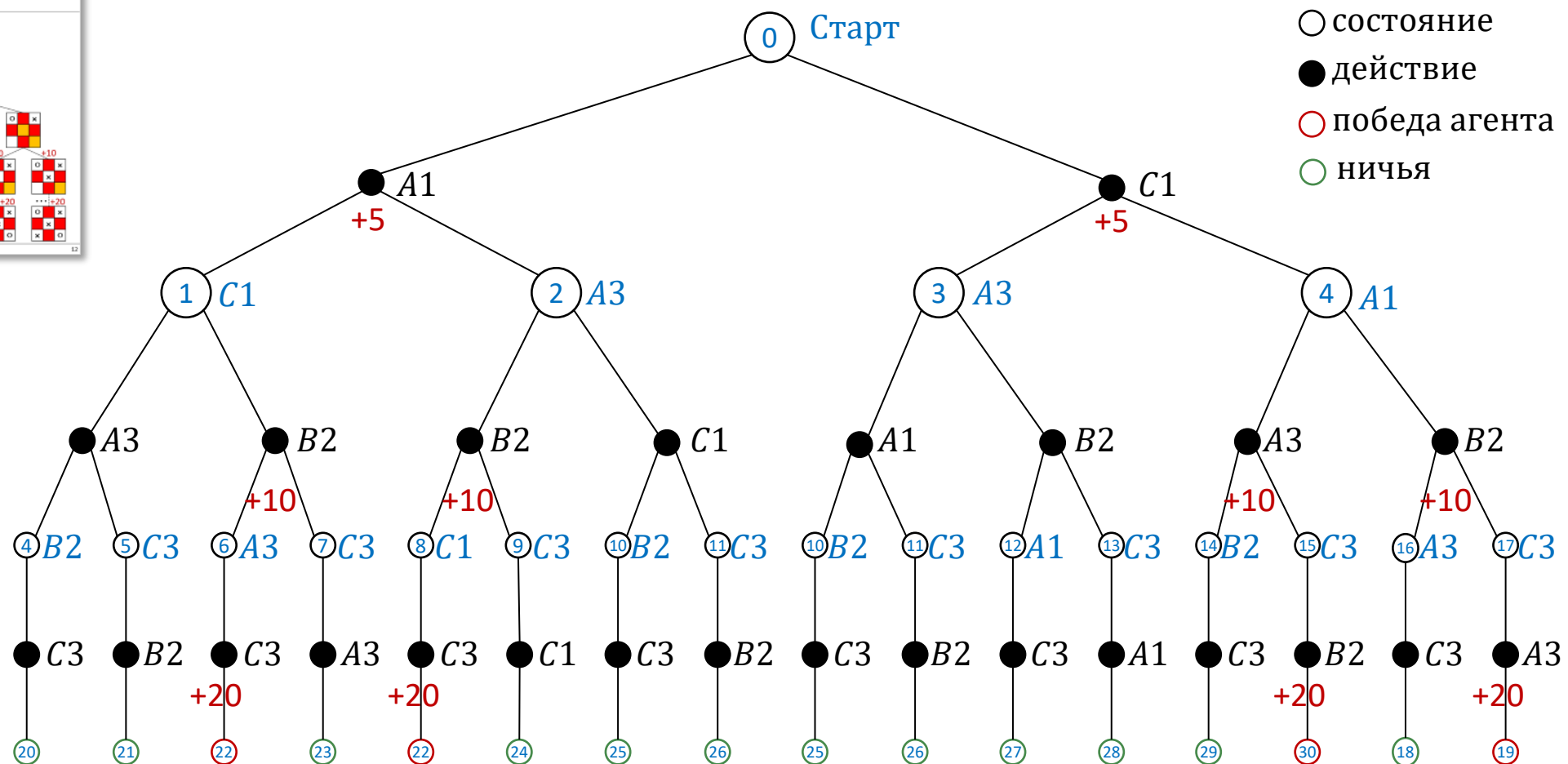
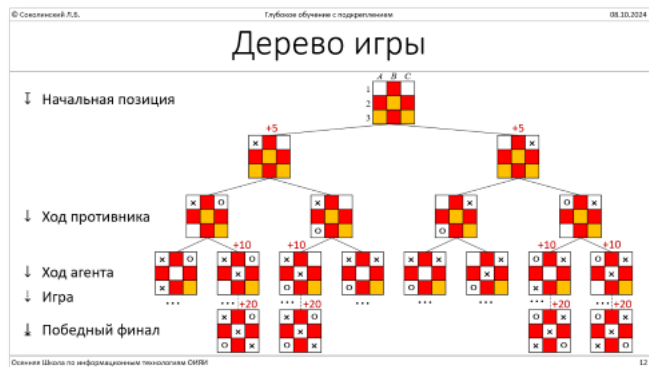
↓ Ход агента

⋮ Игра

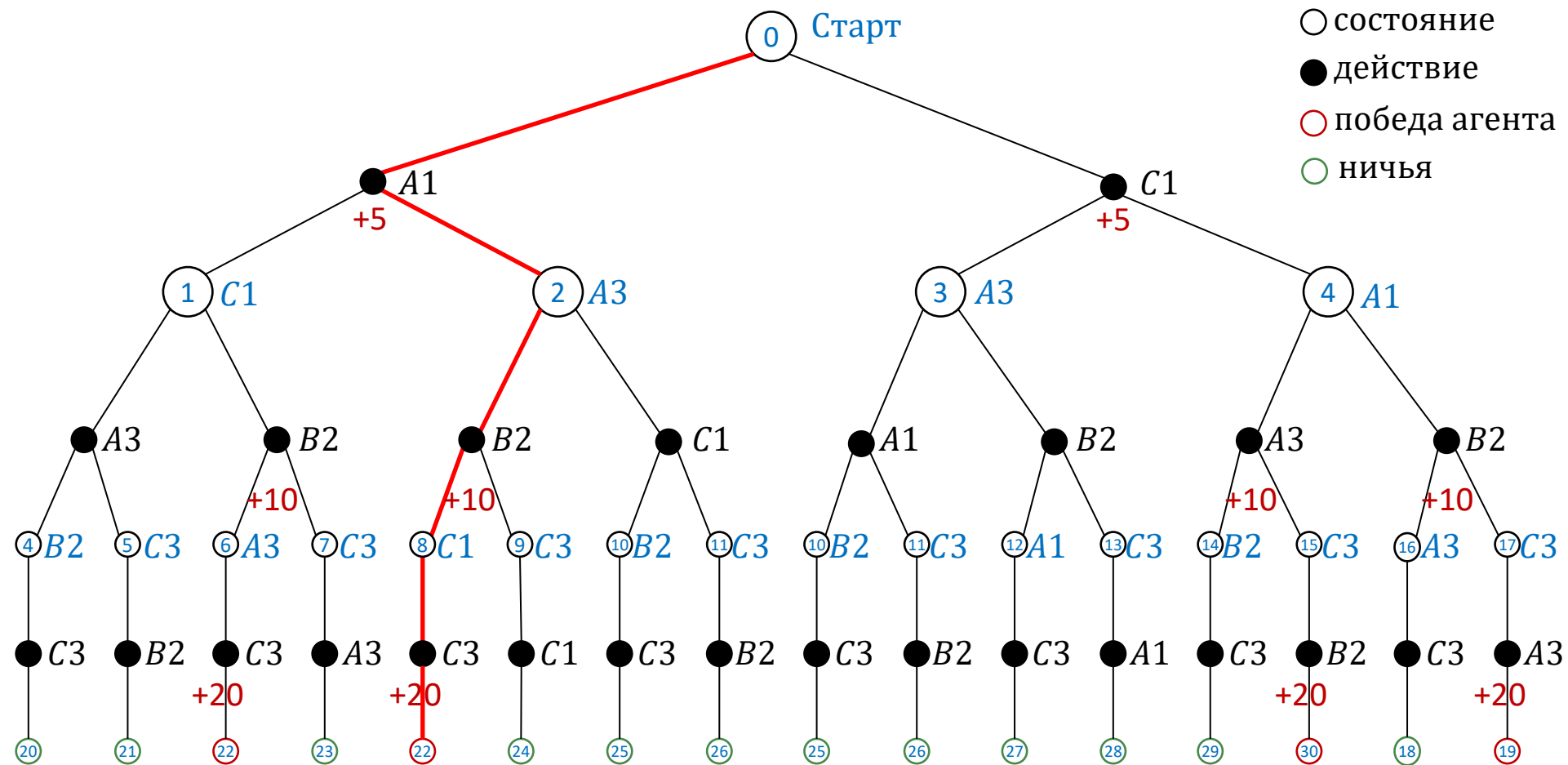
↓ Победный финал



# Дерево состояний



# Пример марковского процесса



$(0; A1; 0) \rightarrow (2; B2; 5) \rightarrow (8; C3; 10) \rightarrow (22; \blacksquare; 20)$

# Эпизоды и доходы

*Эпизод (episode)* – серия взаимодействий агента со средой (партия в игре)

Эпизод представляется в виде конечного марковского процесса:

$$(S_0, A_0, 0) \rightarrow (S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow \dots \rightarrow (S_T, A_T, R_T)$$

*Доход (return)* с момента времени  $t$  до конца эпизода:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{T-t-1} R_T = \sum_{k=t+1}^T \gamma^{k-t-1} R_k$$

$0 \leq \gamma \leq 1$  – коэффициент обесценивания (*discount rate*)

# Пример вычисления дохода

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

## Эпизоды и доходы

Эпизод (*episode*) – серия взаимодействий агента со средой (партия в игре)

Эпизод представляется в виде конечного марковского процесса:

$$(S_0, A_0, 0) \rightarrow (S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow \dots \rightarrow (S_T, A_T, R_T)$$

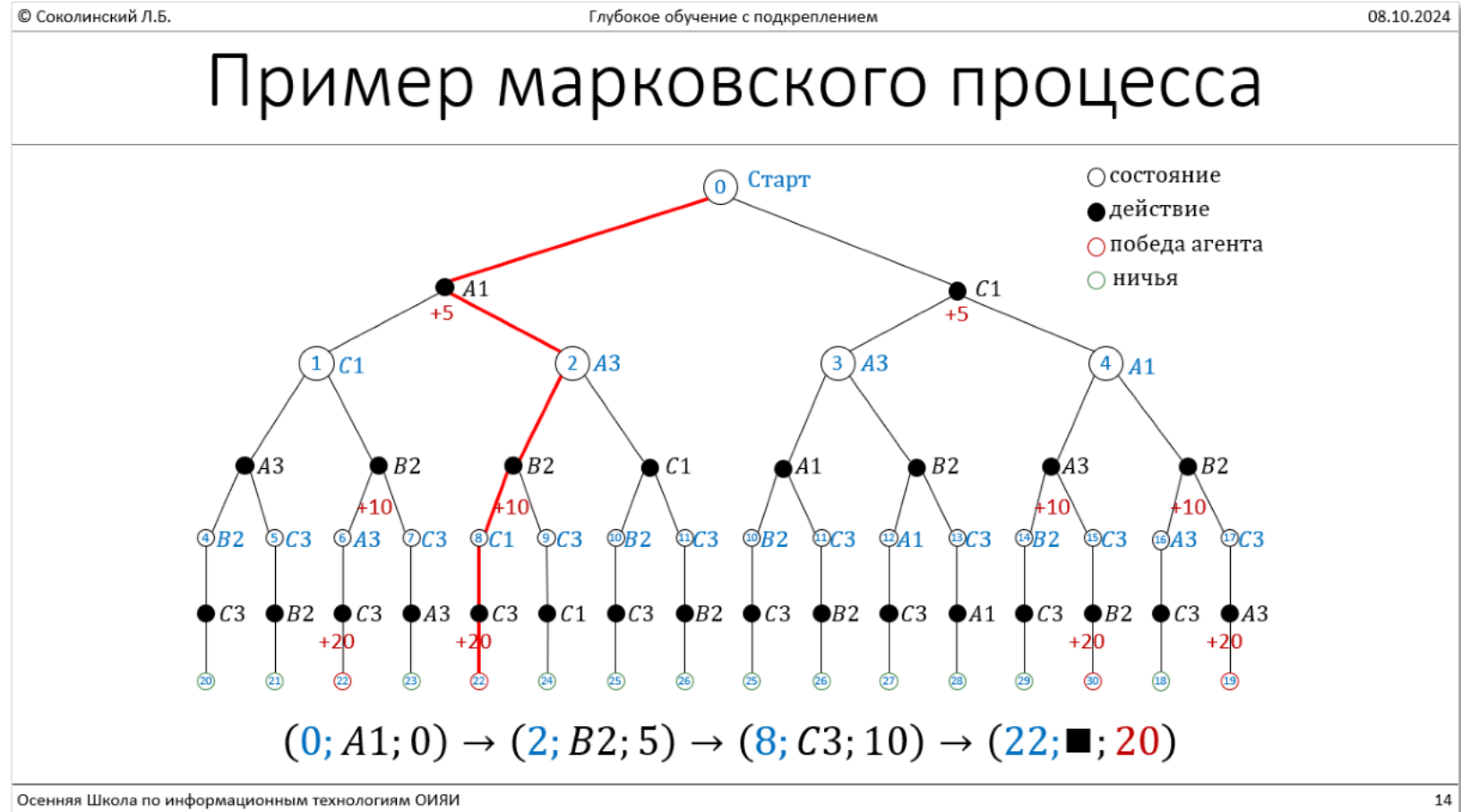
Доход (*return*) с момента времени  $t$  до конца эпизода:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{T-t-1} R_T = \sum_{k=t+1}^T \gamma^{k-t-1} R_k$$

$0 \leq \gamma \leq 1$  – коэффициент обесценивания (*discount rate*)

Осенняя Школа по информационным технологиям ОИЯИ 15

Коэффициент  
обесценивания  $\gamma = \frac{1}{2}$



$$\text{Доход } G_0 = 5 + \frac{1}{2}10 + \frac{1}{4}20 = 15$$



# Стратегия (Policy)

$$\pi: \mathcal{A} \times \mathcal{S} \rightarrow [0; 1]$$

Стратегия определяет для каждого состояния  $s \in \mathcal{S}$  с какой вероятностью следует предпринять то или иное действие из всех ВОЗМОЖНЫХ:

$$\forall s \in \mathcal{S}: \sum_{a \in \mathcal{A}_s} \pi(a|s) = 1$$

$\mathcal{A}_s$  - все возможные действия в состоянии  $s$

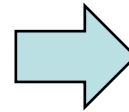
# Стратегия в классических «крестиках-ноликах»

- Стратегия противника
  1. Поставить ○ в центральную клетку (если она свободна)
  2. Поставить ○ в свободную угловую клетку
  3. Занять диагональ и выиграть
- Какой стратегии должен придерживаться агент, чтобы выиграть игру?

# Стратегия агента

$s_0$	A	B	C
1	$\frac{1}{4}$	0	$\frac{1}{4}$
2	0	0	0
3	$\frac{1}{4}$	0	$\frac{1}{4}$

$$\sum_{a \in \mathcal{A}} \pi(a|s_0) = 1$$

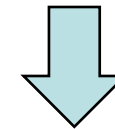


$s_1$	A	B	C
1	×	0	0
2	0	○	0
3	0	0	1

$$\sum_{a \in \mathcal{A}} \pi(a|s_1) = 1$$

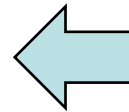
$$\begin{aligned} \pi(A1|s_0) &= \pi(A3|s_0) = \pi(C1|s_0) = \pi(C3|s_0) = \frac{1}{4}; \\ \pi(B1|s_0) &= \pi(B2|s_0) = \pi(B3|s_0) = \pi(A2|s_0) = \\ &= \pi(C2|s_0) = 0 \end{aligned}$$

$$\begin{aligned} \pi(C3|s_1) &= 1; \\ \pi(A2|s_1) &= \pi(A3|s_1) = \pi(B1|s_1) = \\ &= \pi(B3|s_1) = \pi(C1|s_1) = \pi(C2|s_1) = 0 \end{aligned}$$



$s_3$	A	B	C
1	×	0	○
2	○	○	0
3	×	1	×

$$\sum_{a \in \mathcal{A}} \pi(a|s_3) = 1$$



$s_2$	A	B	C
1	×	0	○
2	0	○	0
3	1	0	×

$$\sum_{a \in \mathcal{A}} \pi(a|s_2) = 1$$

$$\begin{aligned} \pi(B3|s_2) &= 1; \\ \pi(B1|s_2) &= \pi(C2|s_2) = 0 \end{aligned}$$

$$\begin{aligned} \pi(A3|s_2) &= 1; \\ \pi(A2|s_2) &= \pi(B1|s_2) = \\ &= \pi(B3|s_2) = \pi(C2|s_2) = 0 \end{aligned}$$

# Функция качества действия (Q-функция)

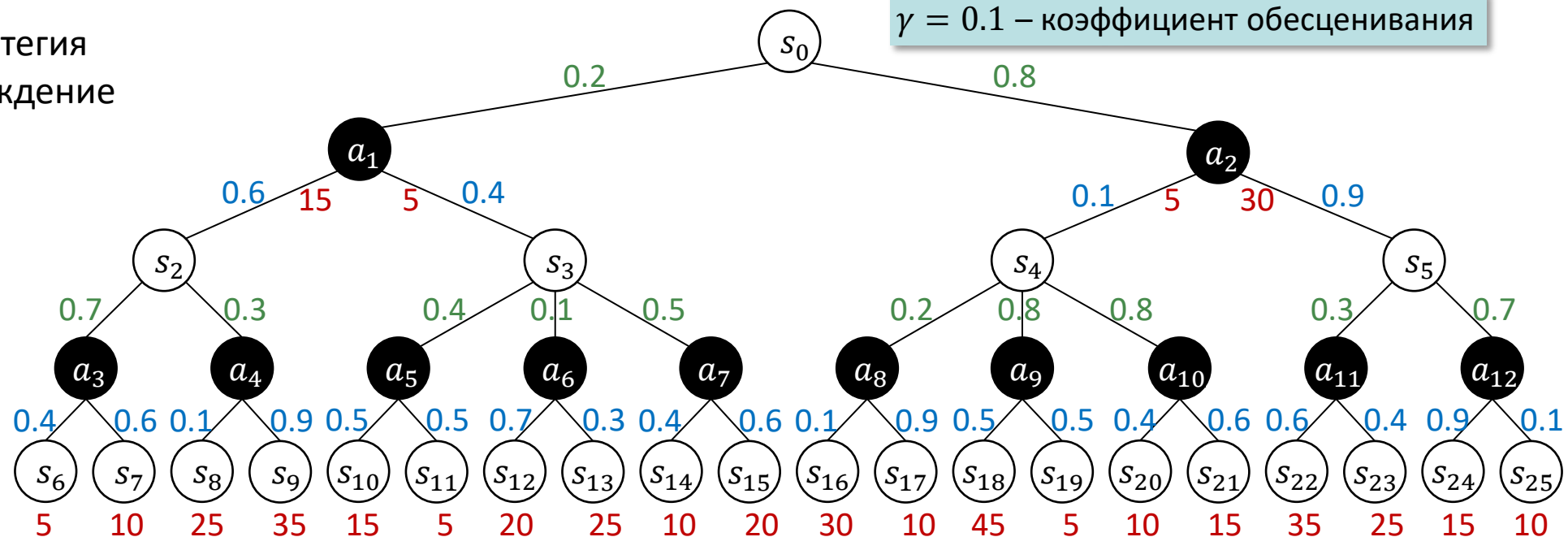
Функция  $q_{\pi}(s, a)$  качества действия  $a$  в состоянии  $s$  при стратегии  $\pi$  – это ожидаемый доход, когда агент предпринимает действие  $a$  в состоянии  $s$  и в дальнейшем следует стратегии  $\pi$

# Вычисление качества действия

$p(\acute{s}, \acute{r}|s, a)$  - вероятность перехода из состояния  $s$  в состояние  $\acute{s}$  и получение вознаграждения  $\acute{r}$  при выполнении действия  $a$

$\pi(a|s)$  - стратегия

$\acute{r}$  - вознаграждение



$$v_{\pi}(s_2) = 15.8$$

$$v_{\pi}(s_3) = 14.15$$

$$\begin{aligned} q_{\pi}(s_0, a_1) &= p(s_2, 15|s_0, a_1) \cdot (15 + \gamma v_{\pi}(s_2)) + p(s_3, 5|s_0, a_1) \cdot (5 + \gamma v_{\pi}(s_3)) \\ &= 0.6 \cdot (15 + 0.1 \cdot 15.8) + 0.4 \cdot (5 + 0.1 \cdot 14.15) = 12.514 \end{aligned}$$

# Отношение квазипорядка на множестве стратегий

$$\pi \succcurlyeq \pi' \iff \forall s \in \mathcal{S}: q_{\pi}(s, a) \geq q_{\pi'}(s, a)$$

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

## Отношение квазипорядка

- Рефлексивность  
 $a \preceq a$
- Транзитивность  
 $a \preceq b \wedge b \preceq c \Rightarrow a \preceq c$

Отношение (частичного) порядка

- Рефлексивность  $a \preceq a$
- Транзитивность  $a \preceq b \wedge b \preceq c \Rightarrow a \preceq c$
- Антисимметричность  $a \preceq b \wedge b \preceq a \Rightarrow a = b$

Осенняя Школа по информационным технологиям ОИЯИ 49

# Оптимальная стратегия $\pi_*$

$$\forall \pi: \pi_* \succeq \pi$$

Если дерево состояний конечно, то существует как минимум одна оптимальная стратегия  $\pi_*$

# Оптимальная функция качества действия $q_*(s, a)$

Все оптимальные стратегии имеют одинаковую функцию качества действия, называемую *оптимальной*:

$$q_*(s, a)$$

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

## Оптимальная стратегия $\pi_*$

$$\forall \pi: \pi_* \succcurlyeq \pi$$

Если дерево состояний конечно, то существует как минимум одна оптимальная стратегия  $\pi_*$

Осенняя Школа по информационным технологиям ОИЯИ 23

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

## Отношение квазипорядка на множестве стратегий

$$\pi \succcurlyeq \pi' \Leftrightarrow \forall s \in \mathcal{S}: q_\pi(s, a) \geq q_{\pi'}(s, a)$$

### Отношение квазипорядка

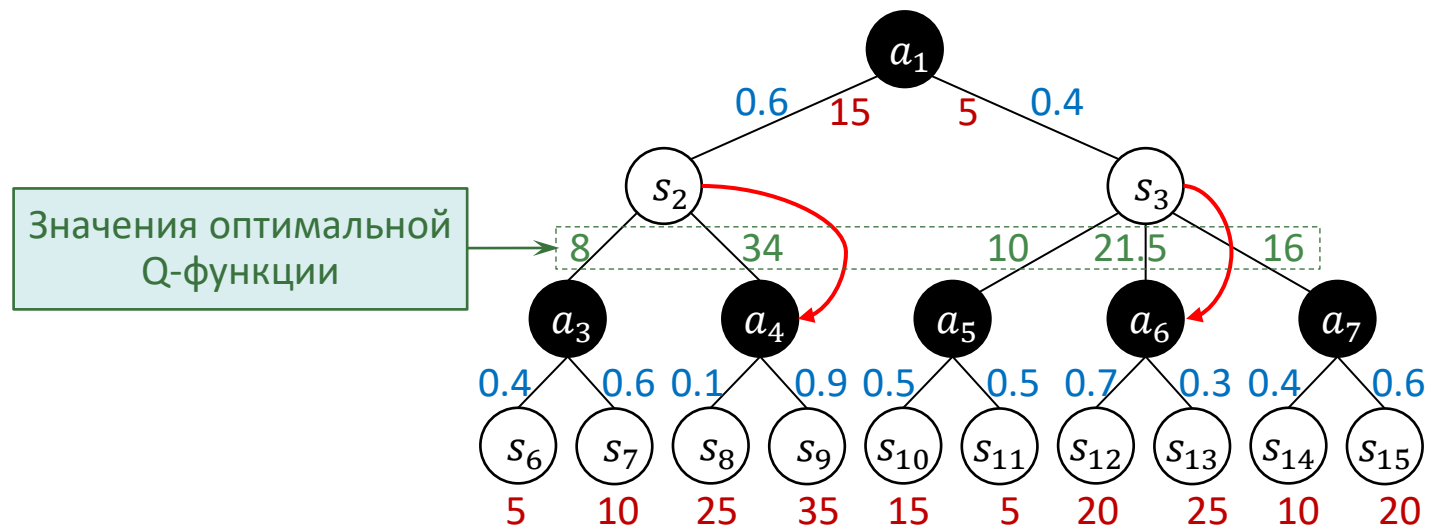
- Рефлексивность  $a \preccurlyeq a$
- Транзитивность  $a \preccurlyeq b \wedge b \preccurlyeq c \Rightarrow a \preccurlyeq c$

Осенняя Школа по информационным технологиям ОИЯИ 22



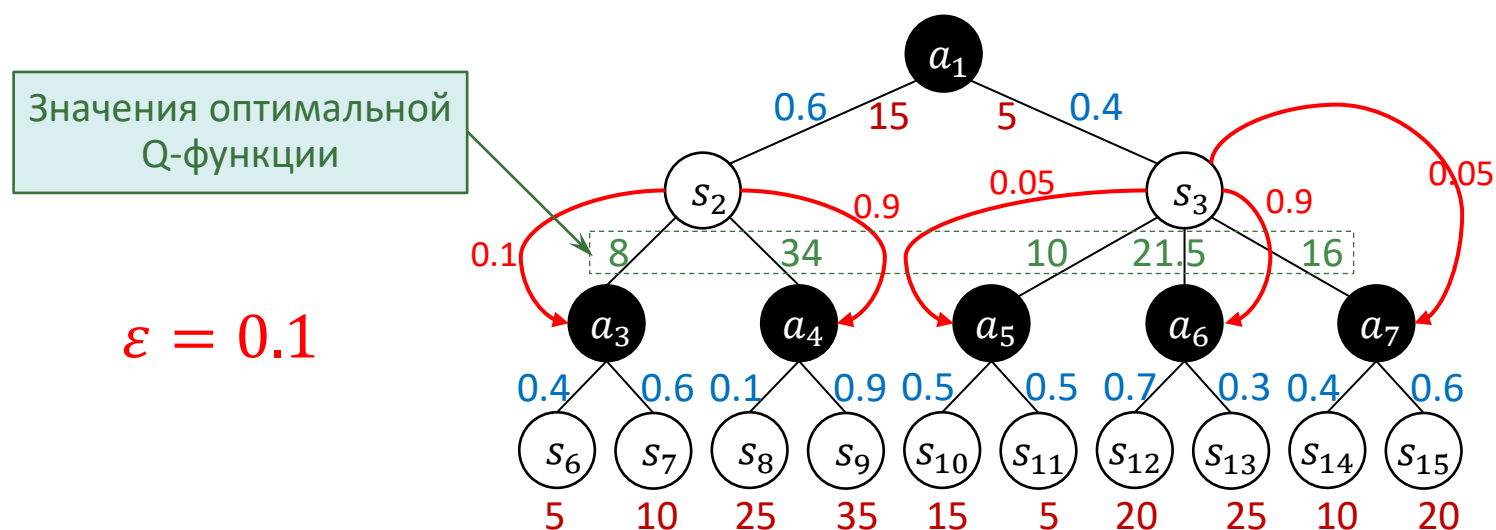
# Жадная оптимальная стратегия для известной оптимальной Q-функции

В состоянии  $s$  выбираем действие  $a$ , которое имеет максимальное качество



# $\epsilon$ -жадная оптимальная стратегия для известной оптимальной Q-функции

- В состоянии  $s$  с вероятностью  $(1 - \epsilon)$  выбираем действие  $a$ , которое имеет максимальное качество
- Остальным возможным  $k$  действиям назначаем вероятность выбора  $\frac{\epsilon}{k}$
- $\epsilon$  – малое положительное число
- $\epsilon$ -жадные оптимальные стратегии используются в обучении с подкреплением



# Q-таблица

		Действия		
		$a_1$	$a_2$	$a_3$
СОСТОЯНИЯ	$s_1$	$q_*(s_1, a_1)$	$q_*(s_1, a_2)$	$q_*(s_1, a_3)$
	$s_2$	$q_*(s_2, a_1)$	$q_*(s_2, a_2)$	$q_*(s_2, a_3)$
	$s_3$	$q_*(s_3, a_1)$	$q_*(s_3, a_2)$	$q_*(s_3, a_3)$
	$s_4$	$q_*(s_4, a_1)$	$q_*(s_4, a_2)$	$q_*(s_4, a_3)$

Q-таблица

- Q-таблица содержит значения оптимальной  $q$ -функции
- Q-таблица однозначно определяет жадную оптимальную стратегию



Q-таблица вычисляется  
с помощью Q-обучения

# Q-обучение

Для каждого эпизода

$$(S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow \dots \rightarrow (S_T, A_T, R_T)$$

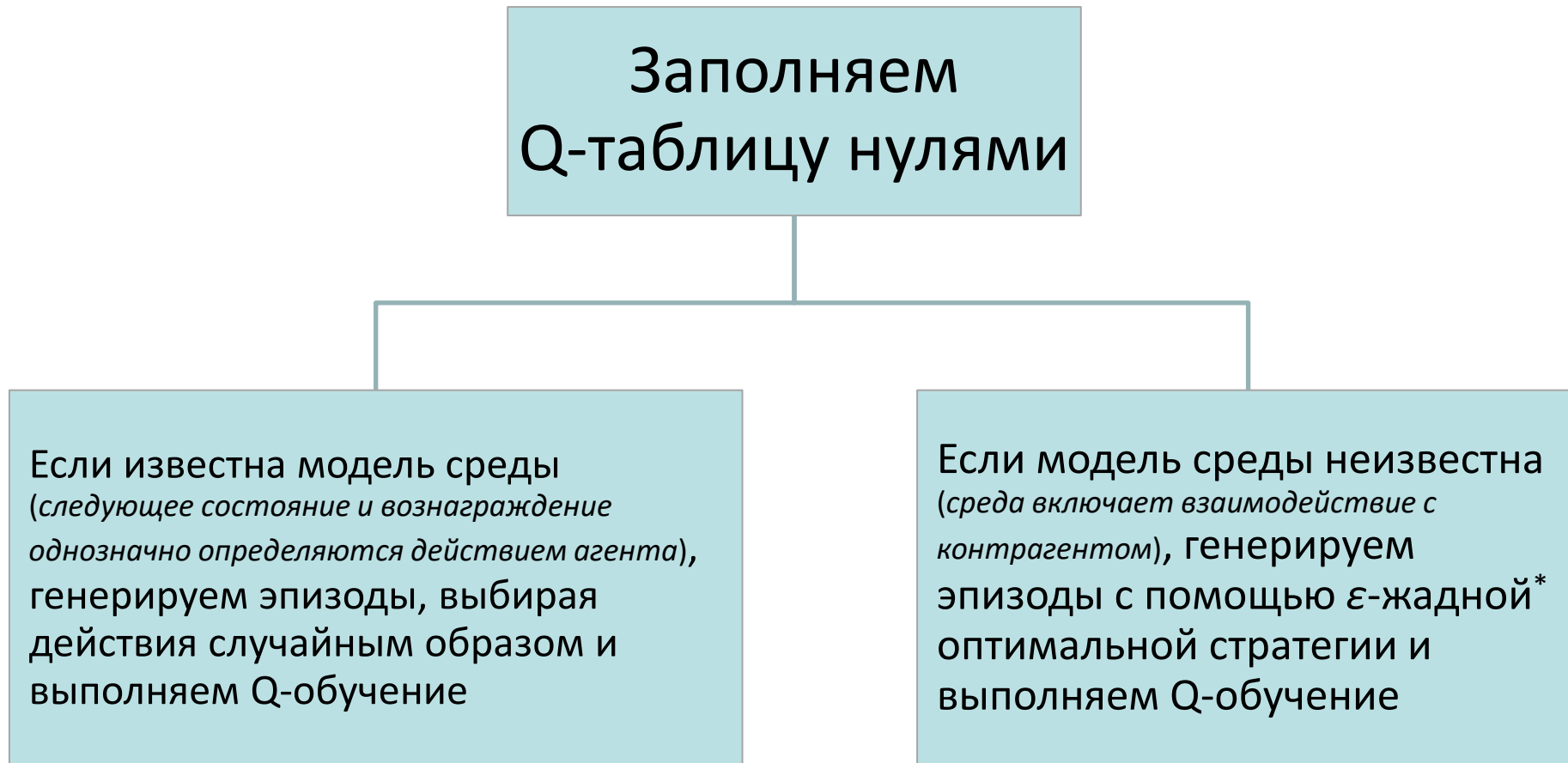
обновляем соответствующие элементы Q-таблицы по формуле

$$Q_{S_t, A_t} := Q_{S_t, A_t} + \eta \left( R_{t+1} + \gamma \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

$\eta$  – скорость обучения

$\gamma$  – коэффициент обесценивания

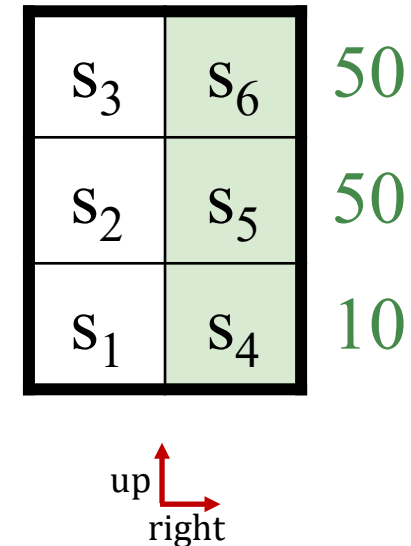
# Алгоритм Q-обучения



\*В этом случае нельзя использовать жадную стратегию, так как она не гарантирует обход всего дерева состояний

# Задача «Микрошашки» (пример Q-обучения)

- Имеется одна шашка, установленная в клетке  $s_1$
- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)
- **Выполнить цикл Q-обучения на основе последовательности эпизодов** (начальные значения элементов Q-таблицы положить равными нулю, использовать  $\eta = 0.1$ ,  $\gamma = 1$ )

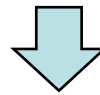


# Эпизод 1

$(s_1, \text{up}, 0) \rightarrow (s_2, \text{right}, 50) \rightarrow (s_5, \blacksquare, 0)$


- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	0	0	0
right	0	0	0



	$s_1$	$s_2$	$s_3$
up	-0.1 <small>= 0.1(-1)</small>	0	0
right	0	5 <small>= 0.1 · 50</small>	0

$s_3$	$s_6$	50
$s_2$	$s_5$	50
$s_1$	$s_4$	10

up   
right

$$Q_{S_t, A_t} := Q_{S_t, A_t} + 0.1 \left( R_{t+1} + \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$



# Эпизод 2

$(s_1, \text{right}, 10) \rightarrow (s_4, \blacksquare, 0)$

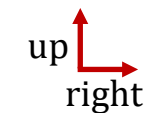
- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	-0.1	0	0
right	0	5	0



	$s_1$	$s_2$	$s_3$
up	-0.1	0	0
right	1 <small>= 0.1 · 10</small>	5	0

$s_3$	$s_6$	50
$s_2$	$s_5$	50
$s_1$	$s_4$	10



$$Q_{S_t, A_t} := Q_{S_t, A_t} + 0.1 \left( R_{t+1} + \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

# Эпизод 3

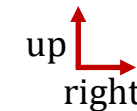
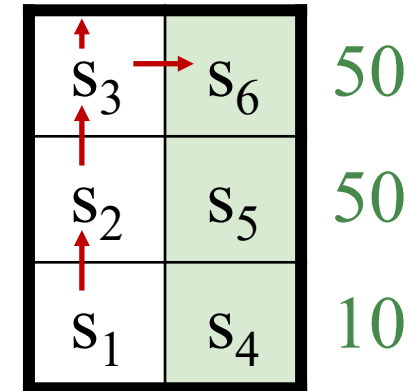
$(s_1, \text{up}, 0) \rightarrow (s_2, \text{up}, -1) \rightarrow (s_3, \text{up}, -10) \rightarrow (s_3, \text{right}, 50) \rightarrow (s_6, \blacksquare, 0)$

- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	-0.1	0	0
right	1	5	0



	$s_1$	$s_2$	$s_3$
up	0.31 <small><math>= -0.1 + 0.1(-1 + 5 + 0.1)</math></small>	-0.1 <small><math>= 0.1(-1)</math></small>	-1 <small><math>= 0.1(-10)</math></small>
right	1	5	5 <small><math>= 0.1 \cdot 50</math></small>



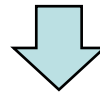
$$Q_{S_t, A_t} := Q_{S_t, A_t} + 0.1 \left( R_{t+1} + \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

# Эпизод 4

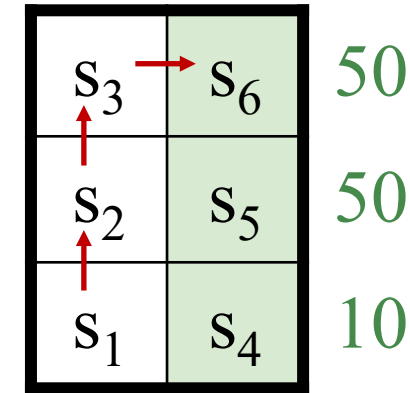
$(s_1, \text{up}, 0) \rightarrow (s_2, \text{up}, -1) \rightarrow (s_3, \text{right}, 50) \rightarrow (s_6, \blacksquare, 0)$

- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	0.31	-0.1	-1
right	1	5	5



	$s_1$	$s_2$	$s_3$
up	0.679 <small><math>= 0.31 + 0.1(-1 + 5 - 0.31)</math></small>	0.31 <small><math>= -0.1 + 0.1(-1 + 5 + 0.1)</math></small>	-1
right	1	5	9.5 <small><math>= 5 + 0.1(50 - 5)</math></small>



$$Q_{S_t, A_t} := Q_{S_t, A_t} + 0.1 \left( R_{t+1} + \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

# Эпизод 5

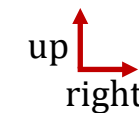
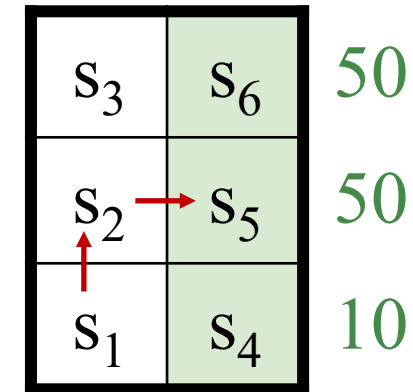
$(s_1, \text{up}, 0) \rightarrow (s_2, \text{right}, 50) \rightarrow (s_5, \blacksquare, 0)$

- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	0.679	0.31	-1
right	1	5	9.5



	$s_1$	$s_2$	$s_3$
up	1.0111 <small><math>= 0.679 + 0.1(-1 + 5 - 0.679)</math></small>	0.31	-1
right	1	9.5 <small><math>= 5 + 0.1(50 - 5)</math></small>	9.5



$$Q_{S_t, A_t} := Q_{S_t, A_t} + 0.1 \left( R_{t+1} + \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

# Эпизод 6

$(s_1, \text{right}, 10) \rightarrow (s_4, \blacksquare, 0)$

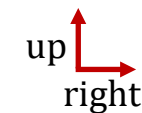
- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	1.0111	0.31	-1
right	1	9.5	9.5



	$s_1$	$s_2$	$s_3$
up	1.0111	0.31	-1
right	1.9 <small><math>= 1 + 0.1(10 - 1)</math></small>	9.5	9.5

$s_3$	$s_6$	50
$s_2$	$s_5$	50
$s_1$	$s_4$	10



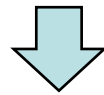
$$Q_{S_t, A_t} := Q_{S_t, A_t} + 0.1 \left( R_{t+1} + \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

# Эпизод 7

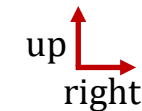
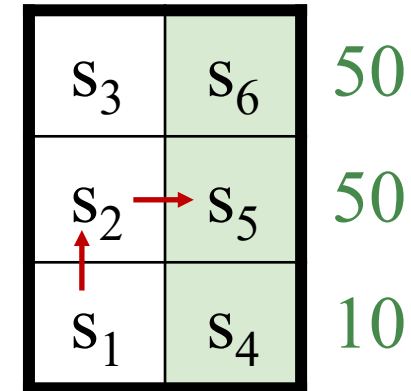
$(s_1, \text{up}, 0) \rightarrow (s_2, \text{right}, 50) \rightarrow (s_5, \blacksquare, 0)$

- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	1.0111	0.31	-1
right	1.9	9.5	9.5



	$s_1$	$s_2$	$s_3$
up	1.76 <small><math>\approx 1.0111 + 0.1(-1 + 9.5 - 1.0111)</math></small>	0.31	-1
right	1.9	13.55 <small><math>= 9.5 + 0.1(50 - 9.5)</math></small>	9.5



$$Q_{S_t, A_t} := Q_{S_t, A_t} + 0.1 \left( R_{t+1} + \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

# Эпизод 8

$(s_1, \text{right}, 10) \rightarrow (s_4, \blacksquare, 0)$


- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	1.76	0.31	-1
right	1.9	13.55	9.5



	$s_1$	$s_2$	$s_3$
up	1.76	0.31	-1
right	2.71 <small><math>= 1.9 + 0.1(10 - 1.9)</math></small>	13.55	9.5

$s_3$	$s_6$	50
$s_2$	$s_5$	50
$s_1$	$s_4$	10

up   
right

$$Q_{S_t, A_t} := Q_{S_t, A_t} + 0.1 \left( R_{t+1} + \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

# Эпизод 9

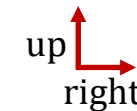
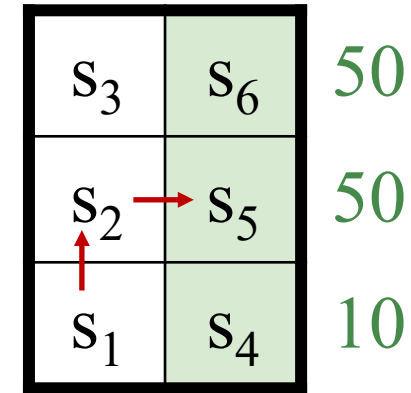
$(s_1, \text{up}, 0) \rightarrow (s_2, \text{right}, 50) \rightarrow (s_5, \blacksquare, 0)$

- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение в соседнюю белую клетку начисляется вознаграждение (-1)

	$s_1$	$s_2$	$s_3$
up	1.76	0.31	-1
right	2.71	13.55	9.5



	$s_1$	$s_2$	$s_3$
up	2.84 <small><math>\approx 1.76 + 0.1(-1 + 13.55 - 1.76)</math></small>	0.31	-1
right	2.71	17.2 <small><math>\approx 13.55 + 0.1(50 - 13.55)</math></small>	9.5



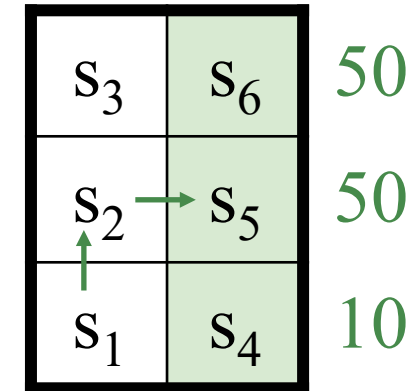
$$Q_{s_t, a_t} := Q_{s_t, a_t} + 0.1 \left( R_{t+1} + \max_a Q_{s_{t+1}, a} - Q_{s_t, a_t} \right)$$



# После Q-обучения жадная оптимальная стратегия обеспечивает максимальный доход

$(s_1, \text{up}, 0) \rightarrow (s_2, \text{right}, 50) \rightarrow (s_5, \blacksquare, 0)$

	$s_1$	$s_2$	$s_3$
up	2.84	0.31	-1
right	2.71	17.2	9.5



$$G_1 = -1 + 50 = 49$$

Эпизоды и доходы

Эпизод (*episode*) – серия взаимодействий агента со средой (партия в игре)

Эпизод представляется в виде конечного марковского процесса:

$$(S_0, A_0, 0) \rightarrow (S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow \dots \rightarrow (S_T, A_T, R_T)$$

Доход (*return*) с момента времени  $t$  до конца эпизода:

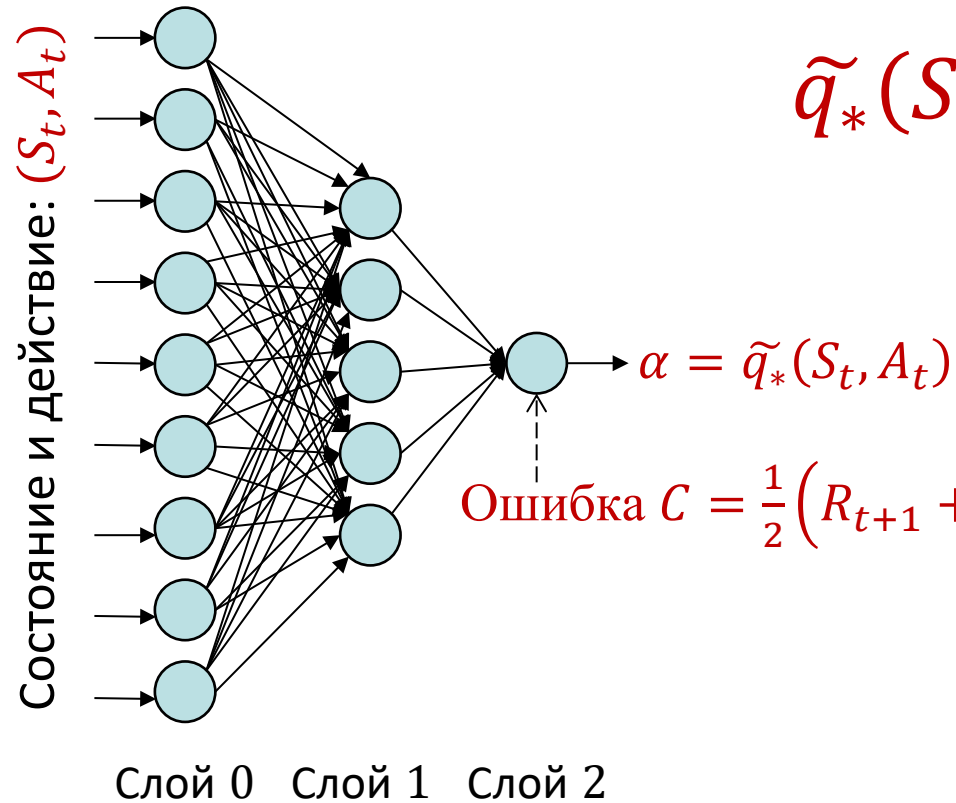
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{T-t-1} R_T = \sum_{k=t+1}^T \gamma^{k-t-1} R_k$$

$0 \leq \gamma \leq 1$  – коэффициент обесценивания (*discount rate*)

# Проблема комбинаторного взрыва

- Во многих задачах, к которым мы хотели бы применять обучение с подкреплением, пространство состояний комбинаторное, а его размер огромен
- Например, количество неповторяющихся шахматных партий многократно превышает количество атомов в наблюдаемой Вселенной
- В таких случаях невозможно найти оптимальную стратегию в результате Q-обучения
- Для сложных задач можно найти стратегию, приближающуюся к оптимальной, с помощью *глубокого обучения с подкреплением*

# Для аппроксимация оптимальной Q-функции используем нейронную сеть



$$\tilde{q}_*(S_t, A_t) \approx q_*(S_t, A_t)$$

Ошибка  $C = \frac{1}{2} \left( R_{t+1} + \gamma \max_a \tilde{q}_*(S_{t+1}, a) - \alpha \right)^2$

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

## Q-обучение

Для каждого эпизода

$$(S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow \dots \rightarrow (S_T, A_T, R_T)$$

обновляем соответствующие элементы Q-таблицы по формуле

$$Q_{S_t, A_t} := Q_{S_t, A_t} + \eta \left( R_{t+1} + \gamma \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t} \right)$$

$\eta$  – скорость обучения

$\gamma$  – коэффициент обесценивания

Осенняя Школа по информационным технологиям ОИЯИ 29

# Для построения эпизода используем $\varepsilon$ -приближенную жадную оптимальную стратегию

- Находясь в состоянии  $S_t$  выполняем действие  $a_t = \arg \max_a \tilde{q}_*(s, a)$
- Для вычисления  $\tilde{q}_*(s, a)$  используем нейронную сеть

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

Для аппроксимация оптимальной Q-функции используем нейронную сеть

Состояние и действие:  $(S_t, A_t)$

Слой 0 Слой 1 Слой 2

$\tilde{q}_*(S_t, A_t) \approx q_*(S_t, A_t)$

$\alpha = \tilde{q}_*(S_t, A_t)$

Ошибка  $C = \frac{1}{2} (R_{t+1} + \gamma \max_a \tilde{q}_*(S_{t+1}, a) - \alpha)^2$

**Q-обучение**

Для каждого эпизода  
 $(S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow \dots \rightarrow (S_T, A_T, R_T)$

обновляем соответствующие элементы Q-таблицы по формуле

$$Q_{S_t, A_t} := Q_{S_t, A_t} + \eta (R_{t+1} + \gamma \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t})$$

$\eta$  – скорость обучения  
 $\gamma$  – коэффициент обесценивания

Осенняя Школа по информационным технологиям ОИЯИ 43

# Глубокое обучение с подкреплением

1. Инициализируем веса  $w$  каким-либо образом
2. С помощью  $\varepsilon$ -жадной оптимальной стратегии генерируем эпизод, вычисляя качество возможных действий с помощью нейронной сети:

$$(S_0, A_0, 0) \rightarrow (S_1, A_1, R_1) \rightarrow \dots \rightarrow (S_T, \blacksquare, R_T)$$

3. На каждом шаге  $t = 0, \dots, T - 1$  корректируем веса:

$$w := w - \eta \nabla_w C$$

- Ошибка вычисляется по формуле

$$C = \frac{1}{2} \left( R_{t+1} + \gamma \max_a \tilde{q}_*(S_{t+1}, a) - \tilde{q}_*(S_t, A_t) \right)^2$$

4. Повторяем шаги 2-3 много раз

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

Для аппроксимации оптимальной Q-функции используем нейронную сеть

$\tilde{q}_*(S_t, A_t) \approx q_*(S_t, A_t)$

$\alpha = \tilde{q}_*(S_t, A_t)$

Ошибка  $C = \frac{1}{2} (R_{t+1} + \gamma \max_a \tilde{q}_*(S_{t+1}, a) - \alpha)^2$

Состояние и действие:  $(S_t, A_t)$

Слой 0 Слой 1 Слой 2

**Q-обучение**

Для каждого эпизода  $(S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow \dots \rightarrow (S_T, A_T, R_T)$

обновляем соответствующие элементы Q-таблицы по формуле

$$Q_{S_t, A_t} := Q_{S_t, A_t} + \eta (R_{t+1} + \gamma \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t})$$

$\eta$  – скорость обучения

$\gamma$  – коэффициент обесценивания

Осенняя Школа по информационным технологиям ОИЯИ 43

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

## Q-обучение

Для каждого эпизода

$$(S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow \dots \rightarrow (S_T, A_T, R_T)$$

обновляем соответствующие элементы Q-таблицы по формуле

$$Q_{S_t, A_t} := Q_{S_t, A_t} + \eta (R_{t+1} + \gamma \max_a Q_{S_{t+1}, a} - Q_{S_t, A_t})$$

$\eta$  – скорость обучения

$\gamma$  – коэффициент обесценивания

Осенняя Школа по информационным технологиям ОИЯИ 29

**Спасибо за внимание!**

# Вспомогательные слайды

# Андрей Андреевич Марков



**Андрей Андреевич Марков**  
(2 июня 1856 — 20 июля 1922)  
русский математик, академик,  
внесший большой вклад в  
теорию вероятностей,  
математический анализ и  
теорию чисел.

- А. А. Марков был сыном чиновника Андрея Григорьевича Маркова, служившего в Лесном департаменте в чине коллежского советника, а затем вышедшего в отставку и служившего в Санкт-Петербурге частным поверенным.
- Андрей Марков страдал туберкулёзом коленного сустава и до 10 лет ходил на костылях. После операции, проведённой известным хирургом Кадэ, он получил возможность ходить нормально.
- В 1866 году его отдали в 5-ю Петербургскую гимназию. Это классическое учебное заведение с преподаванием древних языков (латинского и греческого) пришлось ему не по вкусу; по большинству предметов он учился плохо, исключение составлял только один предмет — математика.
- В 1874 году А. А. Марков окончил гимназию и поступил в Санкт-Петербургский университет. Там он слушал лекции профессоров А. Н. Коркина и Е. И. Золотарёва, а также Пафнутия Львовича Чебышёва, оказавшего определяющее влияние на выбор научной деятельности Андрея Маркова. 31 мая 1878 года он окончил Петербургский университет по математическому разряду физико-математического факультета со степенью кандидата.
- С 13 декабря 1886 года, по предложению Чебышёва, он был избран адъюнктом физико-математического отделения (чистая математика); с 3 марта 1890 года — экстраординарный академик, а с 2 марта 1896 года — ординарный академик Императорской Санкт-Петербургской академии наук. С 1880 года — приват-доцент, с 1886 года — профессор физико-математического факультета Санкт-Петербургского университета. С 1898 года — действительный статский советник.
- Умер в Петрограде в 1922 году. Похоронен на Митрофаниевском кладбище Санкт-Петербурга. В 1954 году перезахоронен на Литераторских мостках, Волковское кладбище.



# Отношение квазипорядка

- Рефлексивность

$$a \preceq a$$

- Транзитивность

$$a \preceq b \wedge b \preceq c \Rightarrow a \preceq c$$

© Соколинский Л.Б. Глубокое обучение с подкреплением 08.10.2024

Отношение (частичного) порядка

- Рефлексивность  $a \preceq a$
- Транзитивность  $a \preceq b \wedge b \preceq c \Rightarrow a \preceq c$
- Антисимметричность  $a \preceq b \wedge b \preceq a \Rightarrow a = b$

Осенняя Школа по информационным технологиям ОИЯИ 50

# Отношение (частичного) порядка

- Рефлексивность

$$a \preceq a$$

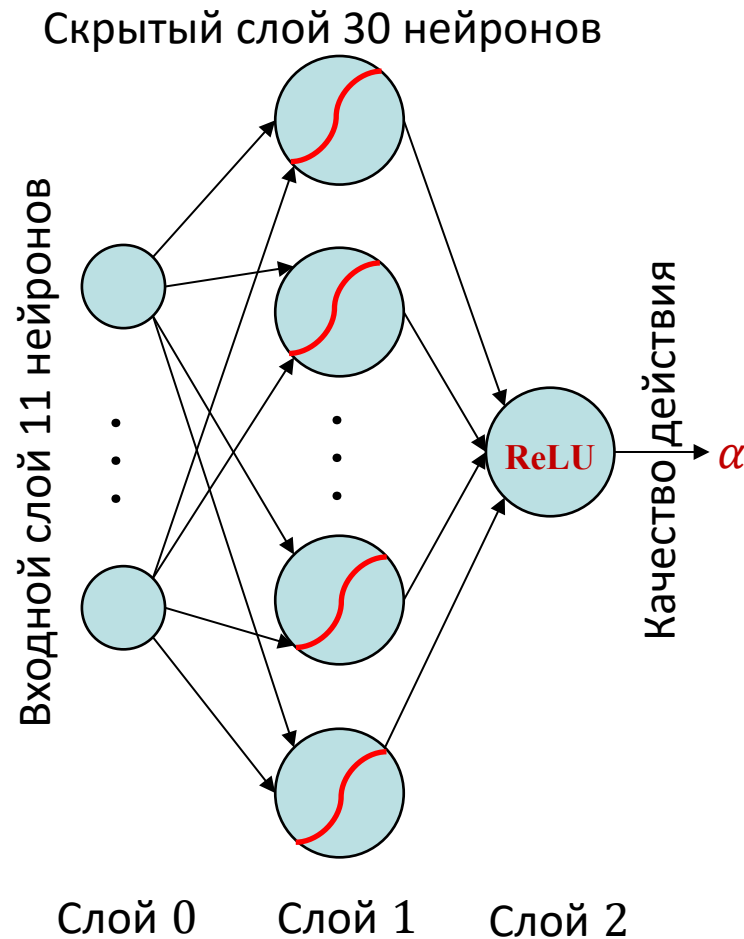
- Транзитивность

$$a \preceq b \wedge b \preceq c \Rightarrow a \preceq c$$

- Антисимметричность

$$a \preceq b \wedge b \preceq a \Rightarrow a = b$$

# Глубокое обучение с подкреплением в игре «крестики-нолики»



- Векторизуем игровое поле (9 нейронов входного слоя)

0 – пустая клетка

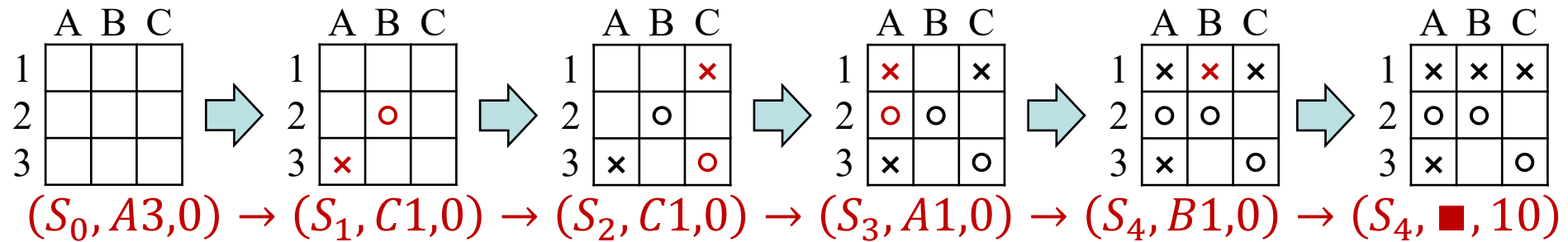
1 – крестик

2 – нолик

	1	2	3
1	×	o	
2	×	o	
3	×	×	o

- Добавляем действие: координаты клетки, куда ставим очередной крестик (2 нейрона входного слоя)
- Используем сигмоидные нейроны для скрытого слоя
- Всем состояниям, кроме финальных назначаем вознаграждение 0
- Проигрышу соответствует вознаграждение 0
- Ничьей соответствует вознаграждение 10
- Выигрышу соответствует вознаграждение 100
- Полагаем  $\gamma = 1, \eta = 0.1$

# Генерация эпизодов и обучение



- Методы генерации эпизодов

- Контрагент, делает случайные ходы (примитивная модель среды)
- Агент играет сам с собой (предпочтительный метод при начальном обучении)
- Агент играет с человеком (применяется в завершающей стадии обучения)

- Тактики обучения

- Непосредственное обучение в процессе игры
- Отложенное обучение:
  - Перед началом игры делается копия вектора весов:  $\hat{w} := w$
  - На протяжении одной игры ходы делаются на основе  $w$ , а корректируется  $\hat{w}$
  - По завершению игры выполняется присваивание:  $w := \hat{w}$
- Рекомендуется постепенно уменьшать скорость обучения  $\eta$