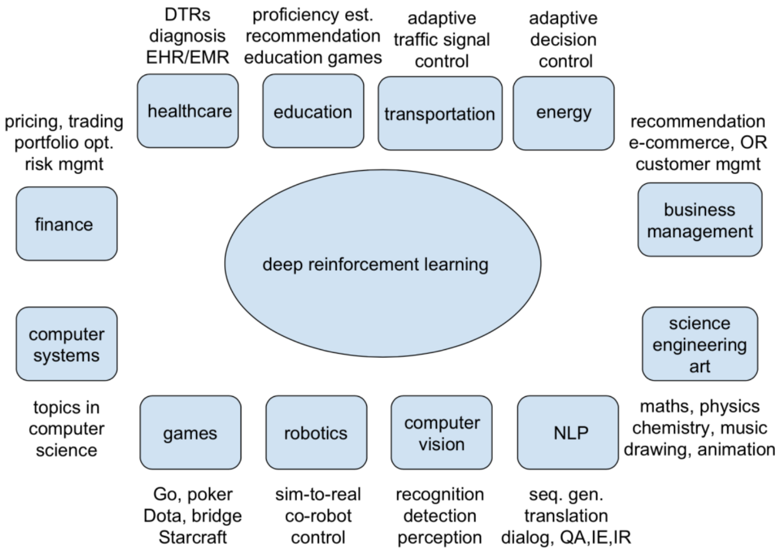
Глубокие нейронные сети

Обучение с подкреплением (Reinforcement learning)

Лекция 12

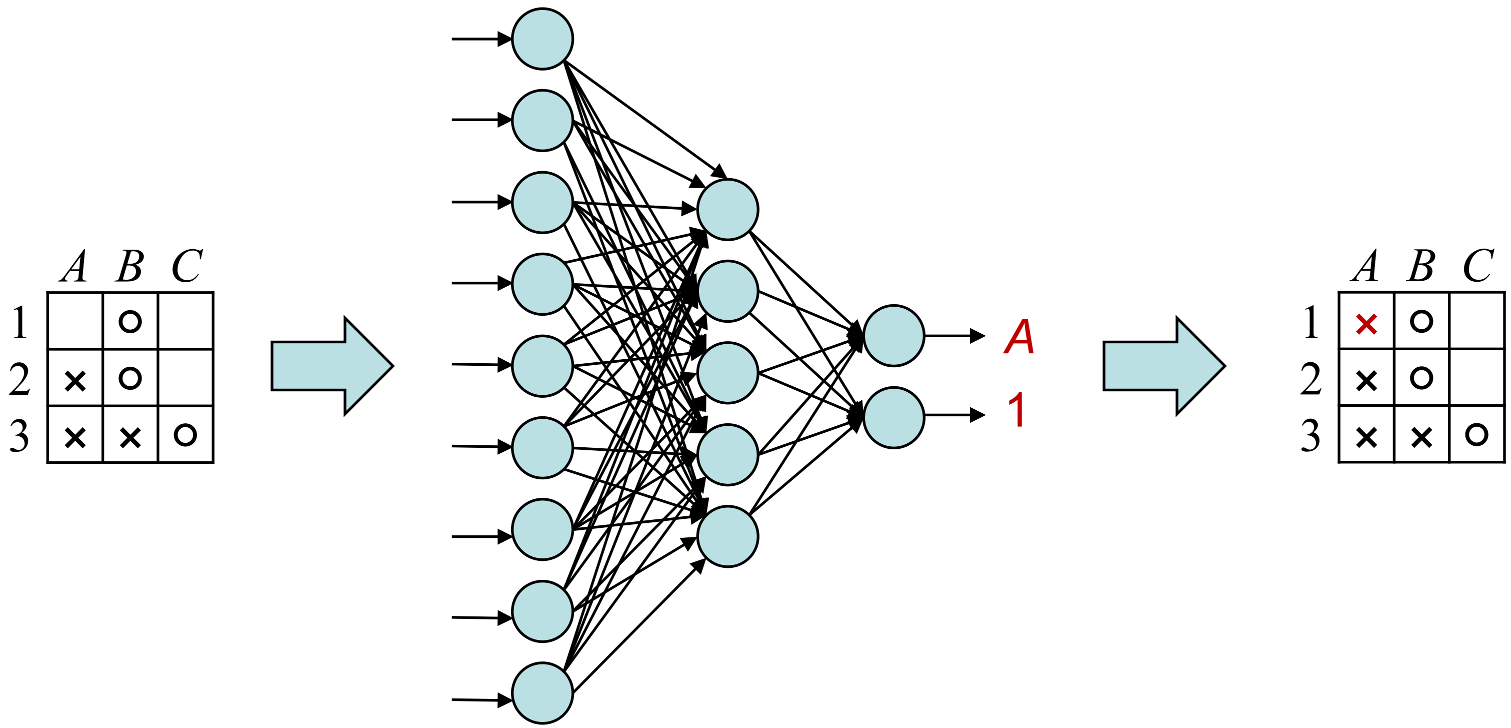
Где применяется обучение с подкреплением

- 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
- Здравоохранение
- …



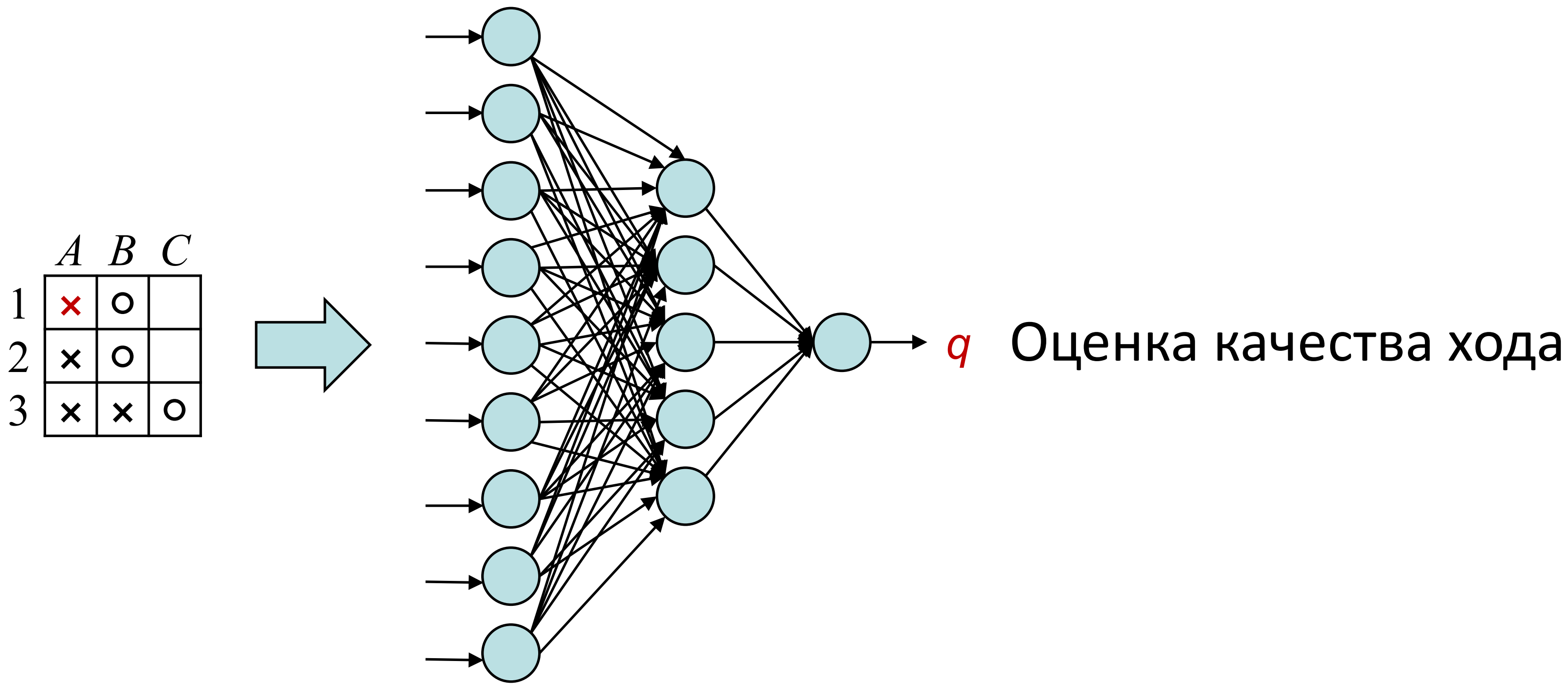
Пример с «крестиками-ноликами»

Интуитивный взгляд на глубокое обучение с подкреплением

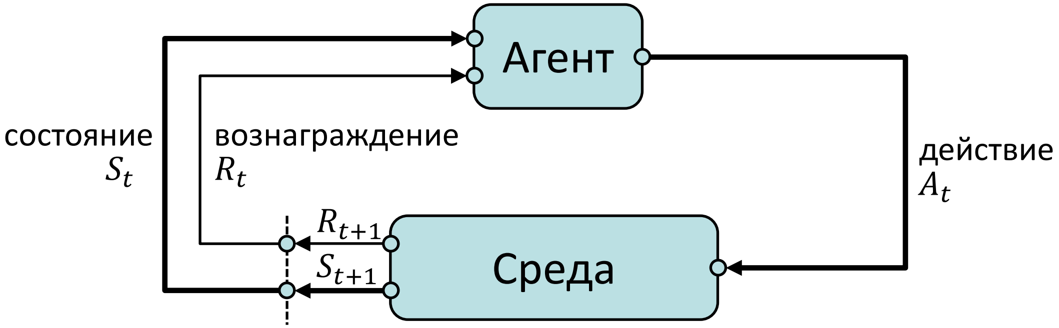


Пример с «крестиками-ноликами»

Реальный взгляд на глубокое обучение с подкреплением



Марковский процесс



© Соколинский Л.Б. Глубокие нейронные сети 16.12.2023

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.

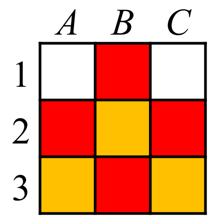
- А. А. Марков был сыном чиновника Андрея Григорьевича Маркова, служившего в Лесном департаменте в чине коллежского советника, а затем вышедшего в отставку и служившего в Санкт-Петербурге частным поверенным.
- Андрей Марков страдал туберкулёзом коленного сустава и до 10 лет ходил на костылях. После операции, 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.

12. Обучение с подкреплением 59

- *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, …
- На каждом шаге *t* агент получает *состояние* (*state*) *среды* (ответный ход противника) $S_t \in \boldsymbol{\mathcal{S}}$, на основе которого выполняет *действие* (*action*) $A_t \in \boldsymbol{\mathcal{A}}$
- На следующем шаге $t+1$ агент получает *вознаграждение* $R_{t+1} \in \boldsymbol{\mathcal{R}} \subset \mathbb{R}$ и переходит в новое состояние $S_{t+1} \in \boldsymbol{\mathcal{S}}$

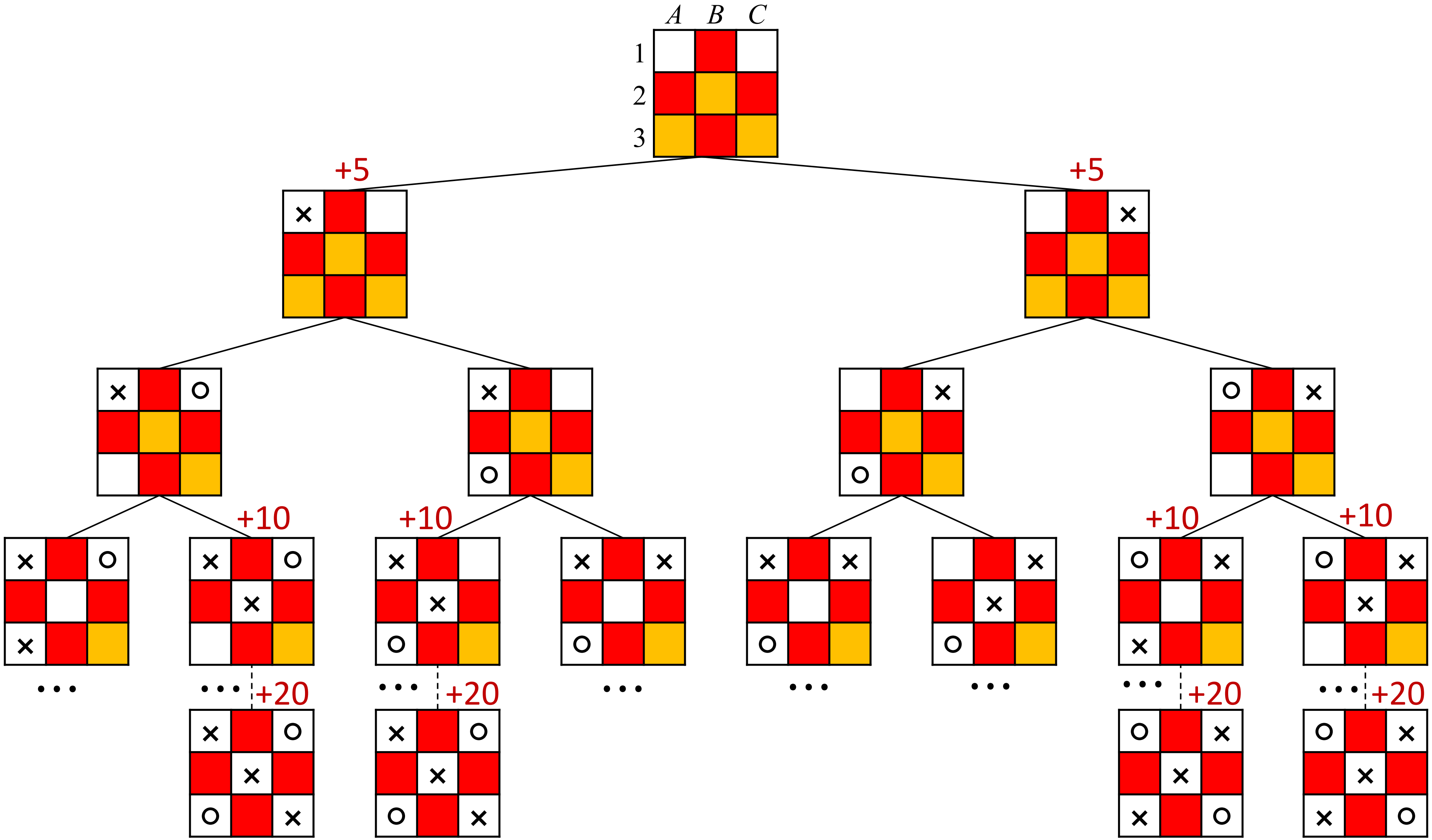
$$(S_0, A_0, 0) \rightarrow (S_1, A_1, R_1) \rightarrow (S_2, A_2, R_2) \rightarrow (S_3, A_3, R_3) \rightarrow \ldots$$

«Цветные крестики-нолики»

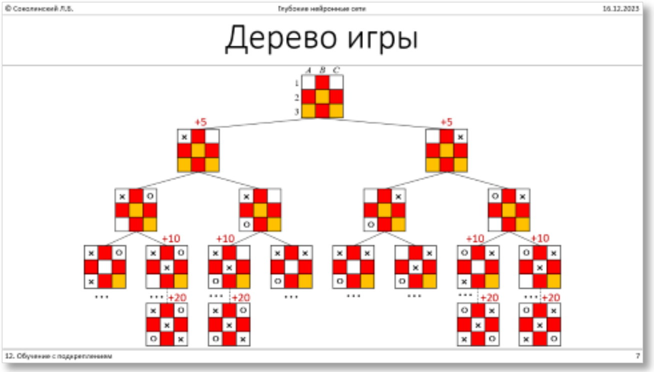


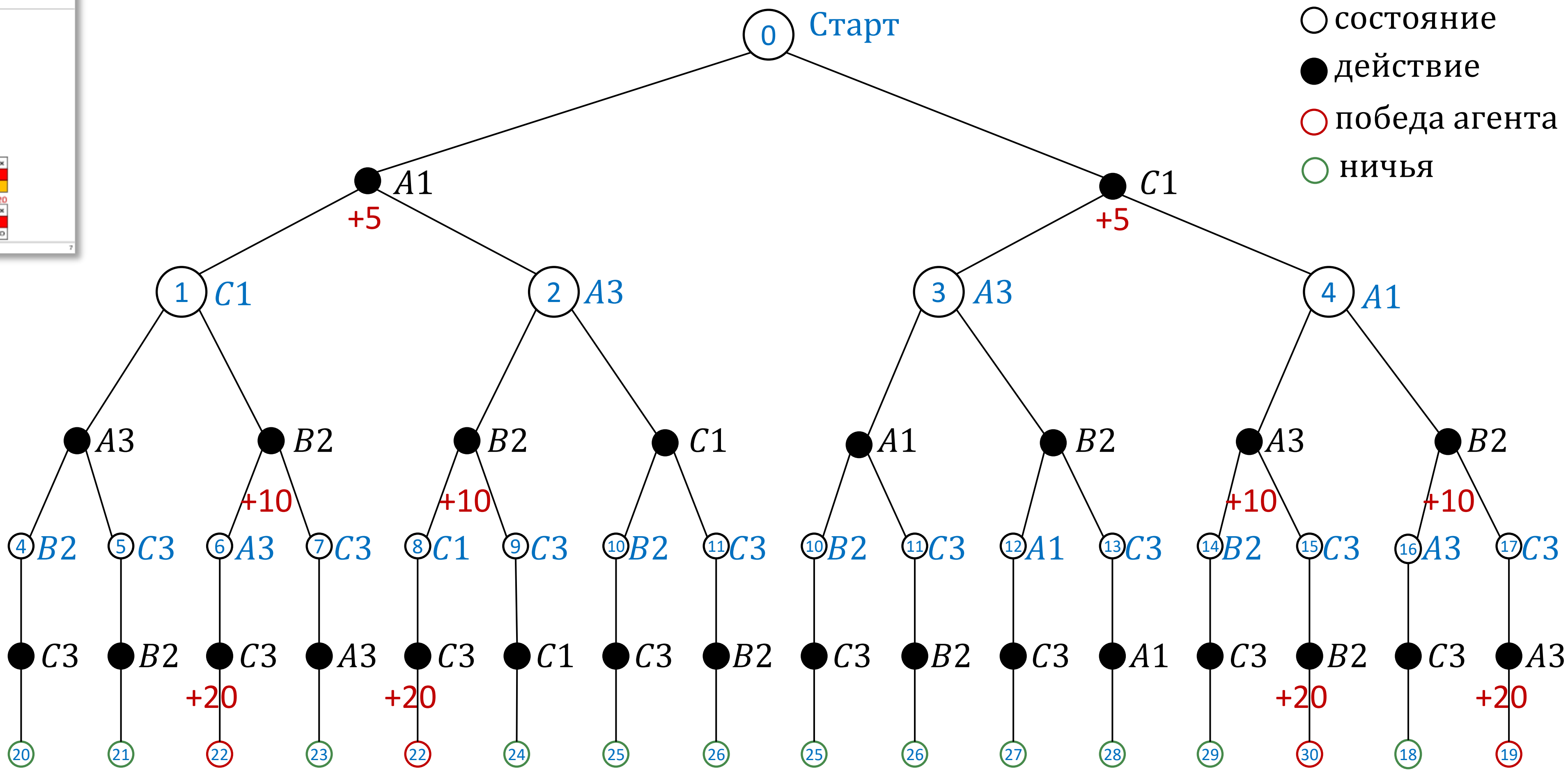
- Агент начинает игру и ставит «крестики»
- Противник (среда) ставит «нолики»
- Можно использовать только белые клетки
- Судья (среда) после каждого хода освобождает одну желтую клетку
- Побеждает тот, кто поставит три своих знака по диагонали, вертикали или горизонтали
- За правильные ходы начисляются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

Дерево игры

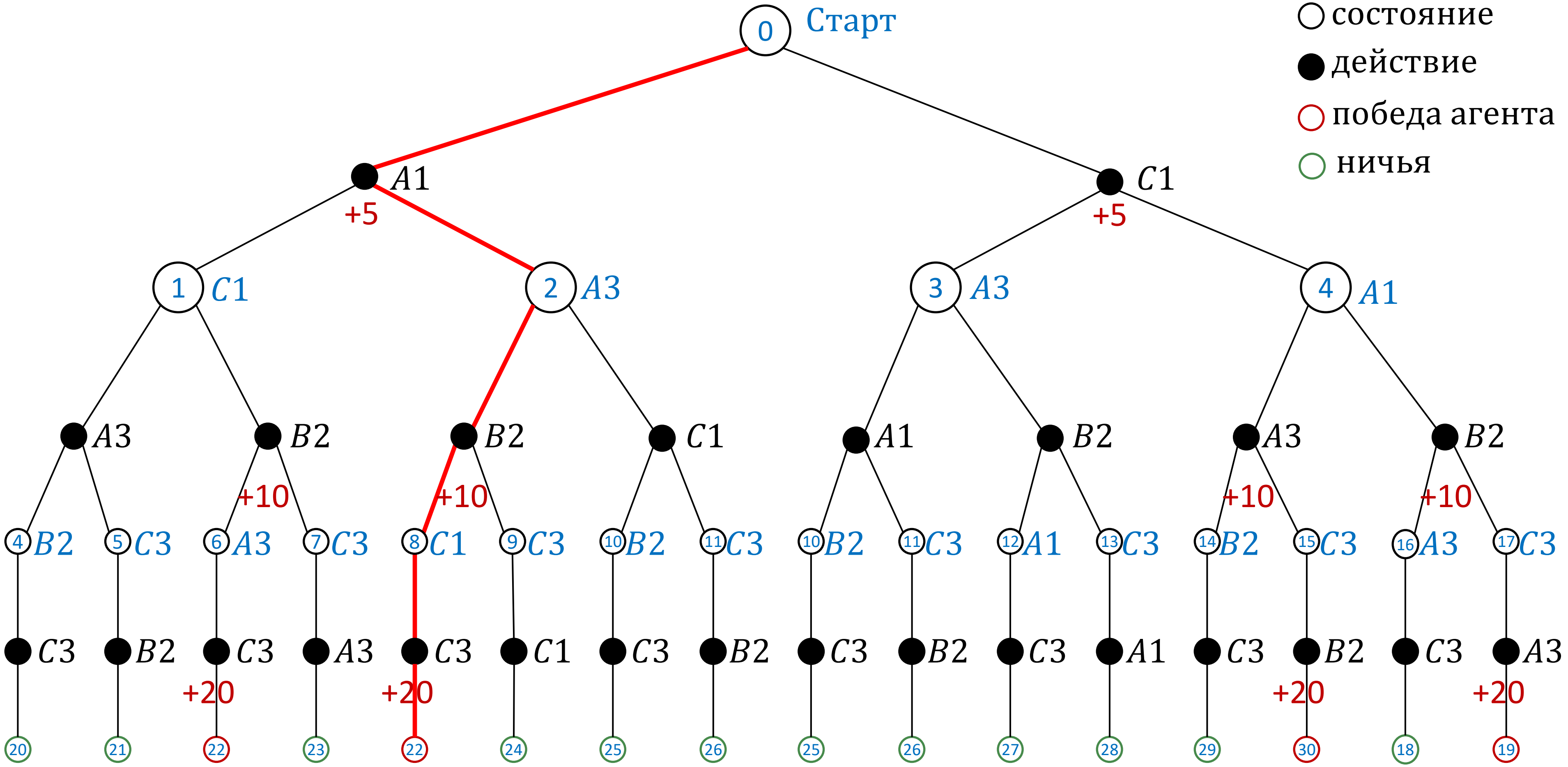


Дерево состояний





Пример марковского процесса



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

Марковское свойство

- $\Pr\{S_{t+1} = \acute{s}, R_{t+1} = \acute{r} | S_t = s, A_t = a\} \leftarrow$ вероятность перехода из состояния s в результате действия a в состояние $\acute{s}$ и получения вознаграждения $\acute{r}$
- Вероятность перехода в $\acute{s}$ и получения вознаграждение $\acute{r}$ на следующем шаге зависит только от предыдущего состояния s и действия a агента:
$$p(\acute{s}, \acute{r} | s, a) ::= \Pr\{S_{t+1} = \acute{s}, R_{t+1} = \acute{r} | S_t = s, A_t = a\}$$
- «Будущее» процесса не зависит от «прошлого» при известном «настоящем»

Распределение вероятностей

$p(\acute{s}, \acute{r}|s, a) ::= \Pr\{S_{t+1} = \acute{s}, R_{t+1} = \acute{r}|S_t = s, A_t = a\} \leftarrow$ вероятность перехода из состояния s в результате действия a в состояние $\acute{s}$ и получения вознаграждения $\acute{r}$

$$\forall s \in \boldsymbol{\mathcal{S}}, a \in \boldsymbol{\mathcal{A}}: \sum_{\acute{s}, \acute{r}} p(\acute{s}, \acute{r}|s, a) = 1$$

Эпизоды и доходы

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

Эпизод представляется в виде конечного марковского процесса:

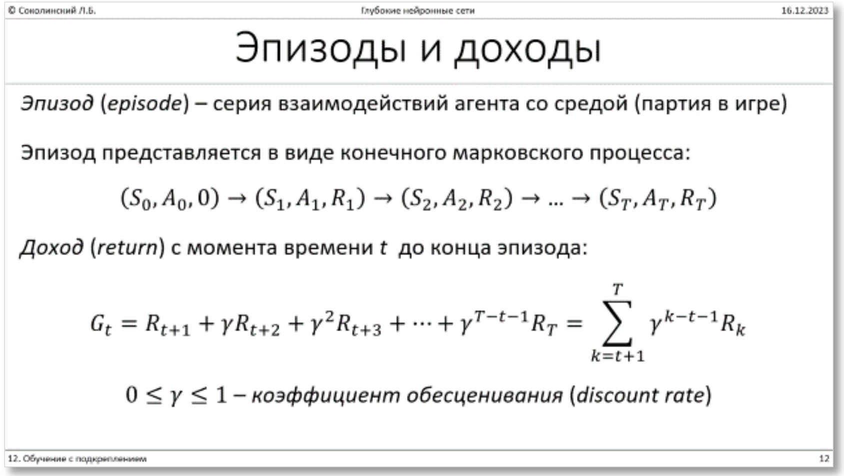
$$(S_0, A_0, 0) \to (S_1, A_1, R_1) \to (S_2, A_2, R_2) \to \dots \to (S_T, A_T, R_T)$$

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

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

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

Пример вычисления дохода



© Соколинский Л.Б. Глубокие нейронные сети 16.12.2023

Эпизоды и доходы

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

Эпизод представляется в виде конечного марковского процесса:

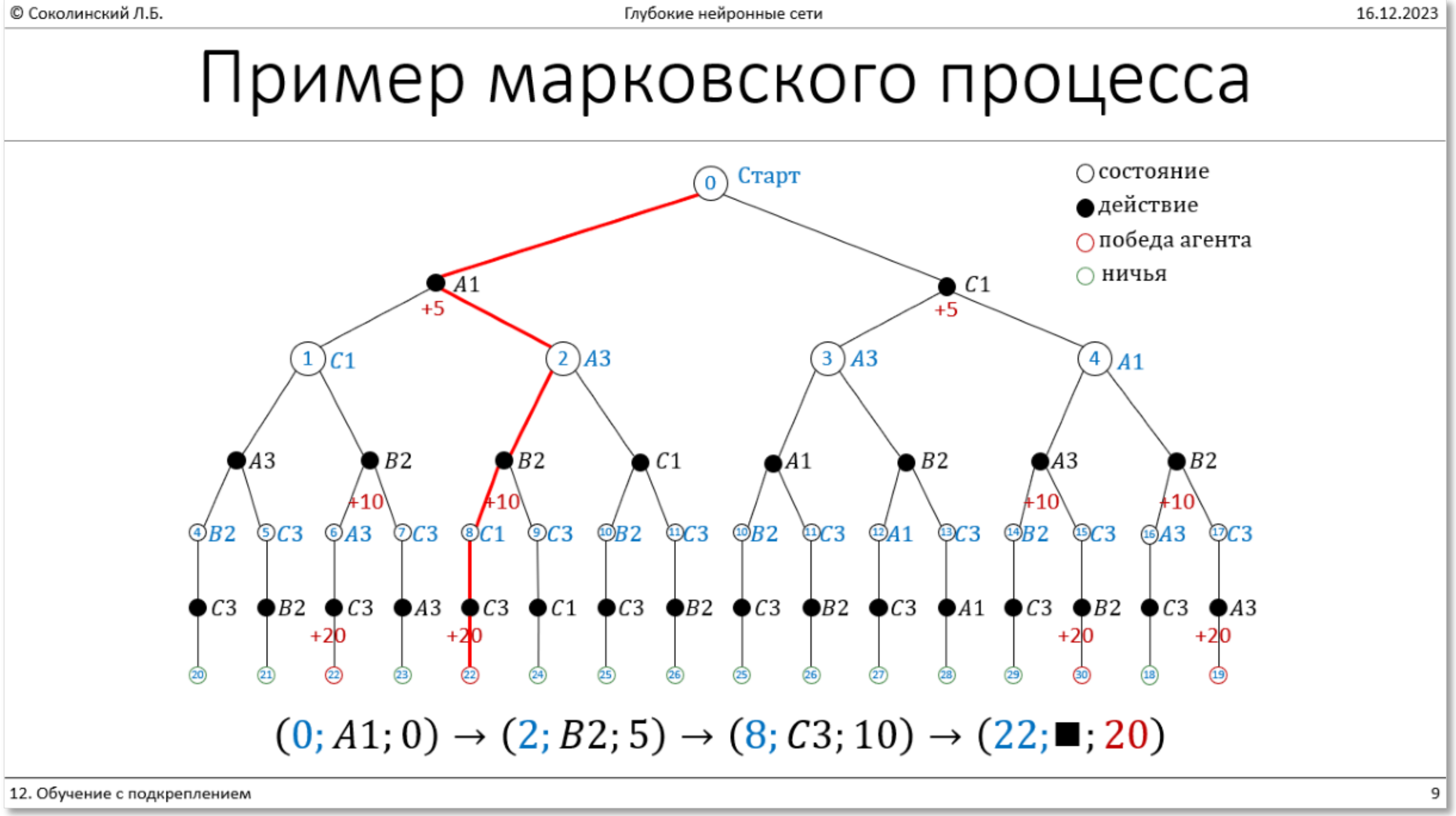
$$(S_0, A_0, 0) \to (S_1, A_1, R_1) \to (S_2, A_2, R_2) \to \dots \to (S_T, A_T, R_T)$$

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

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

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

12. Обучение с подкреплением 12



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

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

Свойство дохода

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$
$$= R_{t+1} + \gamma\left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots + \gamma^{T-(t+1)-2} R_T\right)$$
$$= R_{t+1} + \gamma G_{t+1}$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$

Доход G_t с момента времени t равен сумме вознаграждения в момент времени $t+1$ и обесцененного дохода с момента времени $t+1$

Стратегия (Policy)

$$\pi\colon \boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{S}} \to [0;1]$$

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

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

$\boldsymbol{\mathcal{A}}_s$ - 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	$\frac{1}{4}$

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

$$\pi(A1|s_0) = \pi(A3|s_0) = \pi(C1|s_0) = \pi(C3|s_0) = \tfrac{1}{4};$$
$$\pi(B1|s_0) = \pi(B2|s_0) = \pi(B3|s_0) = \pi(A2|s_0) =$$
$$= \pi(C2|s_0) = 0$$

s_1	A	B	C
1	×	0	0
2	0	○	0
3	0	0	1

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

$$\pi(C3|s_1) = 1;$$
$$\pi(A2|s_1) = \pi(A3|s_1) = \pi(B1|s_1) =$$
$$= \pi(B3|s_1) = \pi(C1|s_1) = \pi(C2|s_1) = 0$$

s_2	A	B	C
1	×	0	○
2	0	○	0
3	1	0	×

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

$$\pi(A3|s_2) = 1;$$
$$\pi(A2|s_2) = \pi(B1|s_2) =$$
$$= \pi(B3|s_2) = \pi(C2|s_2) = 0$$

s_3	A	B	C
1	×	0	○
2	○	○	0
3	×	1	×

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

$$\pi(B3|s_2) = 1;$$
$$\pi(B1|s_2) = \pi(C2|s_2) = 0$$

Функция ценности состояния (Value function)

Функция ценности $v_\pi(s)$ *состояния s* при стратегии π – это ожидаемый доход, когда агент начинает работу в состоянии *s* и в дальнейшем следует стратегии π :

$$v_\pi(s) = M_\pi[G_t | S_t = s] = M_\pi\left[\sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \middle| S_t = s\right]$$

M_π - математическое ожидание случайной величины при условии, что агент следует стратегии π

t – произвольный момент времени

В конечном состоянии функция ценности всегда полагается равной нулю

Задача 1. Вычислить $\boldsymbol{v_\pi(s_1)}$

- Дано:

$\pi(a_1|s_1) = 0.7$

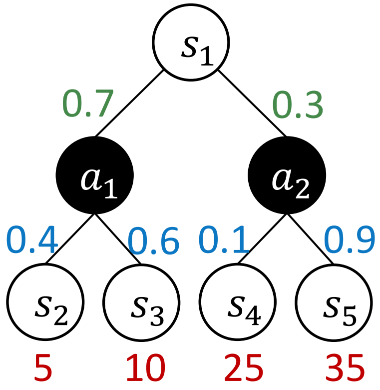
$\pi(a_2|s_1) = 0.3$

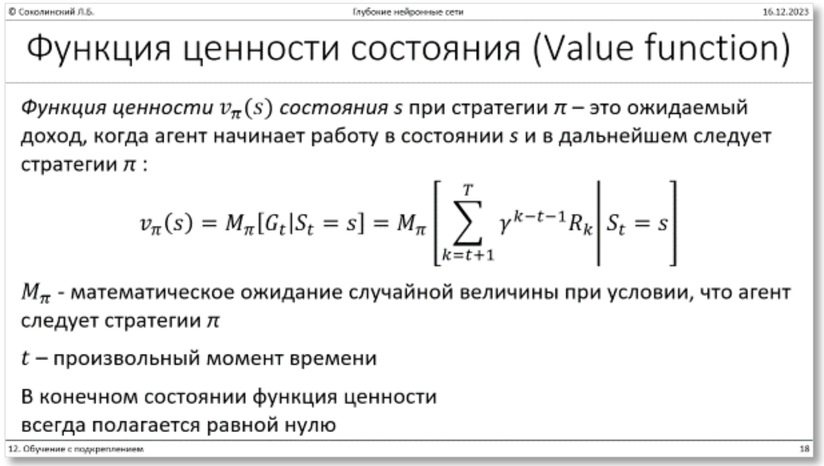
$p(s_2, 5|s_1, a_1) = 0.4$

$p(s_3, 10|s_1, a_1) = 0.6$

$p(s_4, 25|s_1, a_2) = 0.1$

$p(s_5, 35|s_1, a_2) = 0.9$




© Соколинский Л.Б. Глубокие нейронные сети 16.12.2023

Функция ценности состояния (Value function)

Функция ценности $v_\pi(s)$ состояния s при стратегии π – это ожидаемый доход, когда агент начинает работу в состоянии s и в дальнейшем следует стратегии π :

$$v_\pi(s) = M_\pi[G_t|S_t = s] = M_\pi\left[\sum_{k=t+1}^{T} \gamma^{k-t-1}R_k \middle| S_t = s\right]$$

M_π - математическое ожидание случайной величины при условии, что агент следует стратегии π

t – произвольный момент времени

В конечном состоянии функция ценности всегда полагается равной нулю

12. Обучение с подкреплением 18

- Решение:

$$v_\pi(s_1) = \pi(a_1|s_1)\big(5p(s_2, 5|s_1, a_1) + 10p(s_3, 10|s_1, a_1)\big) +$$
$$+ \pi(a_2|s_1)\big(25p(s_4, 25|s_1, a_2) + 35p(s_5, 35|s_1, a_2)\big) =$$
$$= 0.7 \cdot (0.4 \cdot 5 + 0.6 \cdot 10) + 0.3 \cdot (0.1 \cdot 25 + 0.9 \cdot 35) = \mathbf{15.8}$$

Функция качества действия (Q-функция)

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

$$q_\pi(s,a) = M_\pi[G_t | S_t = s, A_t = a] = M_\pi\left[\sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \middle| S_t = s, A_t = a\right]$$

M_π - математическое ожидание случайной величины при условии, что агент следует стратегии π

t – произвольный момент времени

Соотношение между качеством действия и ценностью состояния

$$q_\pi(s,a) = \sum_{\acute{s},\acute{r}} p(\acute{s},\acute{r}|s,a)\big(\acute{r} + \gamma v_\pi(\acute{s})\big)$$

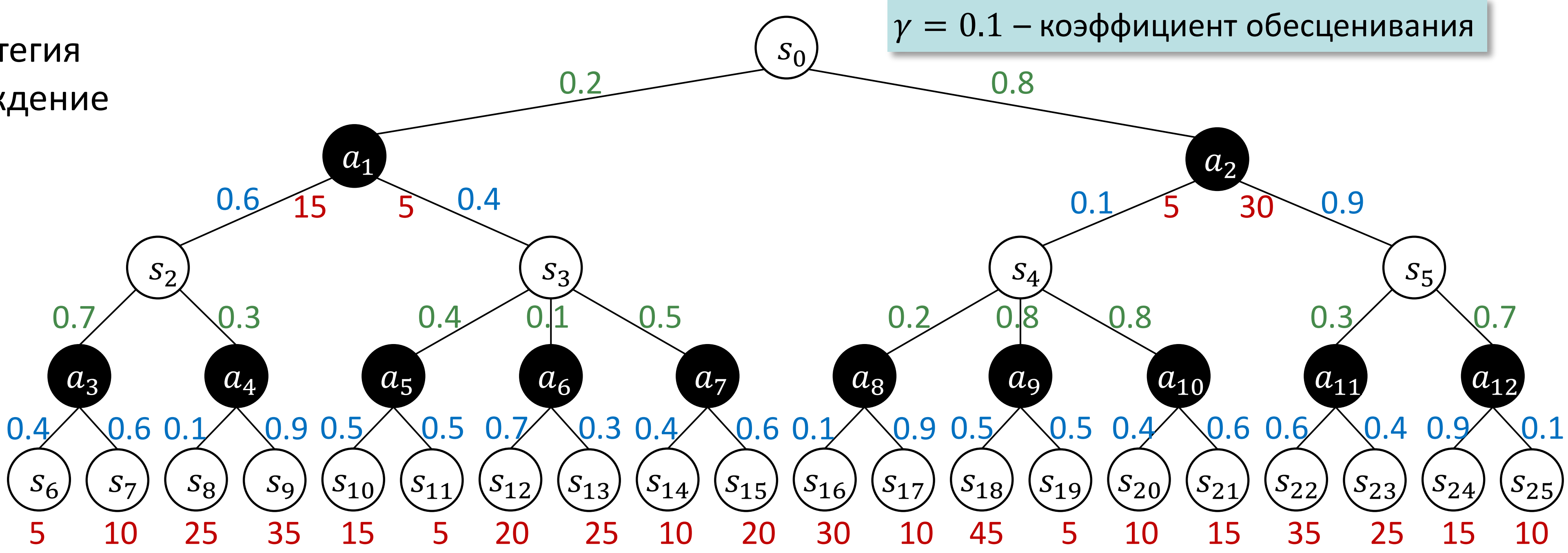
Вычисление качества действия

$p(\acute{s}, \acute{r}|s, a)$ - 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

$\gamma = 0.1$ – коэффициент обесценивания



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

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

$$q_\pi(s_0, a_1) = p(s_2, 15|s_0, a_1) \cdot \big(15 + \gamma v_\pi(s_2)\big) + p(s_3, 5|s_0, a_1) \cdot \big(5 + \gamma v_\pi(s_3)\big)$$

$$= 0.6 \cdot (15 + 0.1 \cdot 15.8) + 0.4 \cdot (5 + 0.1 \cdot 14.15) = 12.514$$

Соотношение между ценностью состояния и качеством действий

$$v_\pi(s) = \sum_a \pi(a|s) \cdot q_\pi(s, a)$$

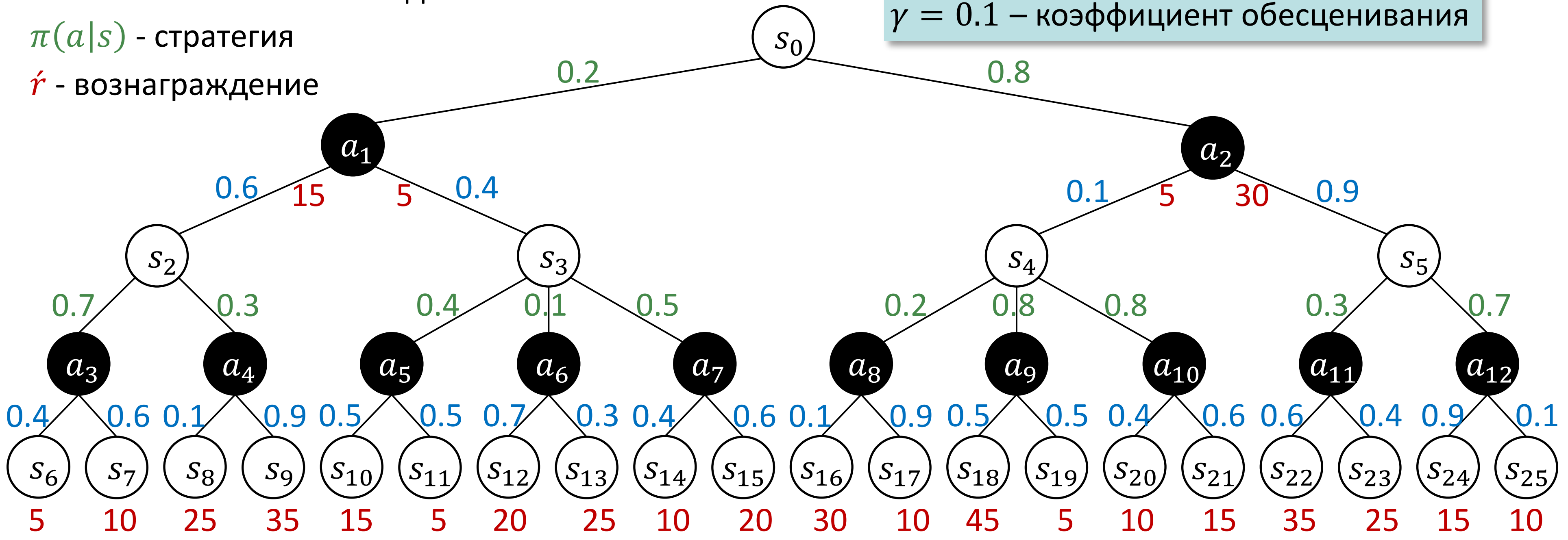
Соотношение между ценностью состояния s_0 и ценностями последующих состояний

$p(\acute{s},\acute{r}|s,a)$ - 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

$\gamma = 0.1$ – коэффициент обесценивания



$$v_\pi(s_0) =$$

$$= \pi(a_1|s_0)\cdot\Big(p(s_2,15|s_0,a_1)\cdot\big(15+\gamma v_\pi(s_2)\big)+p(s_3,5|s_0,a_1)\cdot\big(5+\gamma v_\pi(s_3)\big)\Big)$$

$$+\,\pi(a_2|s_0)\cdot\Big(p(s_4,5|s_0,a_2)\cdot\big(5+\gamma v_\pi(s_4)\big)+p(s_5,30|s_0,a_2)\cdot\big(30+\gamma v_\pi(s_5)\big)\Big)$$

Формула Беллмана для стратегии π

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{\acute{s},\acute{r}} p(\acute{s},\acute{r}|s,a)\left(\acute{r} + \gamma v_\pi(\acute{s})\right)$$

Отношение квазипорядка на множестве стратегий

$$\pi \succcurlyeq \acute{\pi} \Leftrightarrow \forall s \in \boldsymbol{\mathcal{S}}: v_{\pi}(s) \geq v_{\acute{\pi}}(s)$$

Функция ценности состояния (Value function)

Функция ценности $v_\pi(s)$ *состояния* s при стратегии π – это ожидаемый доход, когда агент начинает работу в состоянии s и в дальнейшем следует стратегии π :

$$v_\pi(s) = M_\pi[G_t|S_t = s] = M_\pi\left[\sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \middle| S_t = s\right]$$

M_π - математическое ожидание случайной величины при условии, что агент следует стратегии π

t – произвольный момент времени

В конечном состоянии функция ценности всегда полагается равной нулю

Отношение квазипорядка

- Рефлексивность

$$a \preccurlyeq a$$

- Транзитивность

$$a \preccurlyeq b \wedge b \preccurlyeq c \Rightarrow a \preccurlyeq c$$

Отношение (частичного) порядка

- Рефлексивность

$$a \preccurlyeq a$$

- Транзитивность

$$a \preccurlyeq b \wedge b \preccurlyeq c \Rightarrow a \preccurlyeq c$$

- Антисимметричность

$$a \preccurlyeq b \wedge b \preccurlyeq a \Rightarrow a = b$$

Оптимальная стратегия $\boldsymbol{\pi}_*$

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

Если дерево состояний конечно, то существует как минимум одна оптимальная стратегия π_*

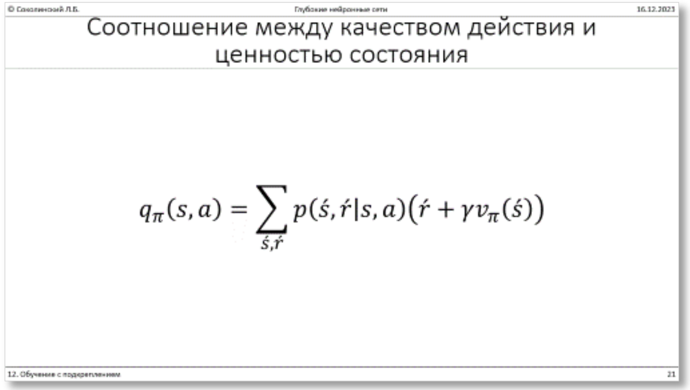
Оптимальные функция ценности состояния $\boldsymbol{v}_*(\boldsymbol{s})$ и функция качества действия $\boldsymbol{q}_*(\boldsymbol{s}, \boldsymbol{a})$

- Все оптимальные стратегии имеют одинаковую функцию ценности состояния, называемую *оптимальной*:

$$v_*(s) ::= \max_\pi v_\pi(s)$$

- У оптимальных стратегий также одна и та же *оптимальная* функция качества действий

$$q_*(s,a) = \sum_{\acute{s},\acute{r}} p(\acute{s},\acute{r}|s,a)\big(\acute{r} + \gamma v_*(\acute{s})\big)$$


© Соколинский Л.Б. Глубокие нейронные сети 16.12.2023

Соотношение между качеством действия и ценностью состояния

$$q_\pi(s,a) = \sum_{\acute{s},\acute{r}} p(\acute{s},\acute{r}|s,a)\big(\acute{r} + \gamma v_\pi(\acute{s})\big)$$

12. Обучение с подкреплением 21

Формула оптимальности Беллмана для ценности состояния

$$v_*(s) = \max_a \sum_{\acute{s},\acute{r}} p(\acute{s},\acute{r}|s,a)\big(\acute{r} + \gamma v_*(\acute{s})\big)$$

Задача 2. Вычислить $\boldsymbol{v_*(s_1)}$

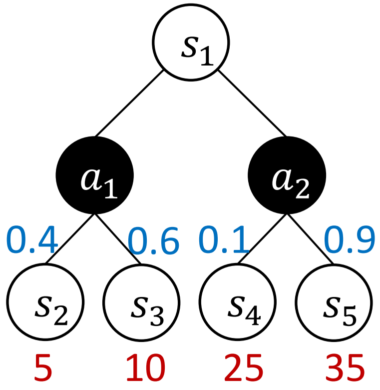
- Дано:

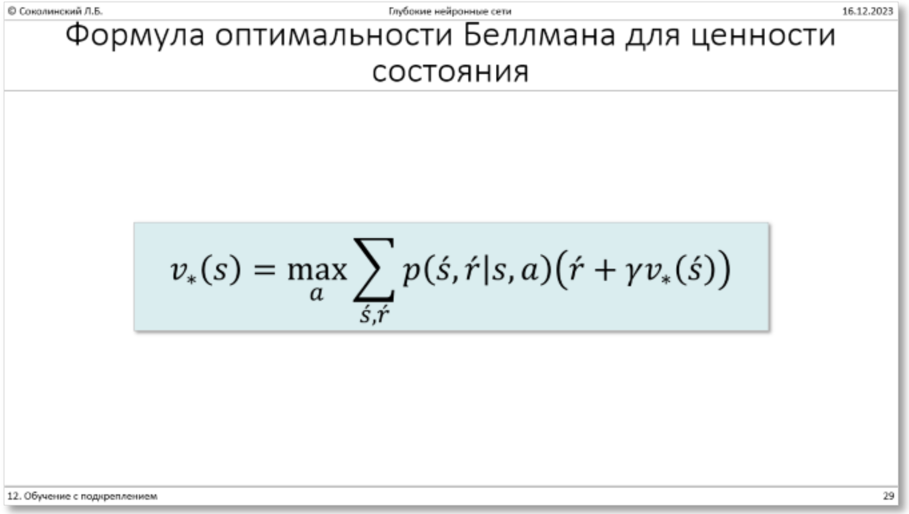
$$p(s_2, 5|s_1, a_1) = 0.4$$

$$p(s_3, 10|s_1, a_1) = 0.6$$

$$p(s_4, 25|s_1, a_2) = 0.1$$

$$p(s_5, 35|s_1, a_2) = 0.9$$




© Соколинский Л.Б. Глубокие нейронные сети 16.12.2023

Формула оптимальности Беллмана для ценности состояния

$$v_*(s) = \max_a \sum_{\acute{s},\acute{r}} p(\acute{s},\acute{r}|s,a)\left(\acute{r} + \gamma v_*(\acute{s})\right)$$

12. Обучение с подкреплением 29

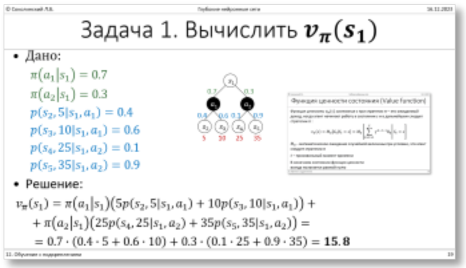
- Решение:

$$v_*(s_1) =$$

$$= \max\Big(\big(5p(s_2, 5|s_1, a_1) + 10p(s_3, 10|s_1, a_1)\big), \big(25p(s_4, 25|s_1, a_2)$$

$$+ 35p(s_5, 35|s_1, a_2)\big)\Big) = \max\big((0.4 \cdot 5 + 0.6 \cdot 10), (0.1 \cdot 25 + 0.9 \cdot 35)\big)$$

$$= \max(8, 34) = \mathbf{34}$$


Задача 1. Вычислить $\boldsymbol{v_\pi(s_1)}$

- Дано:

$\pi(a_1|s_1) = 0.7$

$\pi(a_2|s_1) = 0.3$

$p(s_2, 5|s_1, a_1) = 0.4$

$p(s_3, 10|s_1, a_1) = 0.6$

$p(s_4, 25|s_1, a_2) = 0.1$

$p(s_5, 35|s_1, a_2) = 0.9$

- Решение:

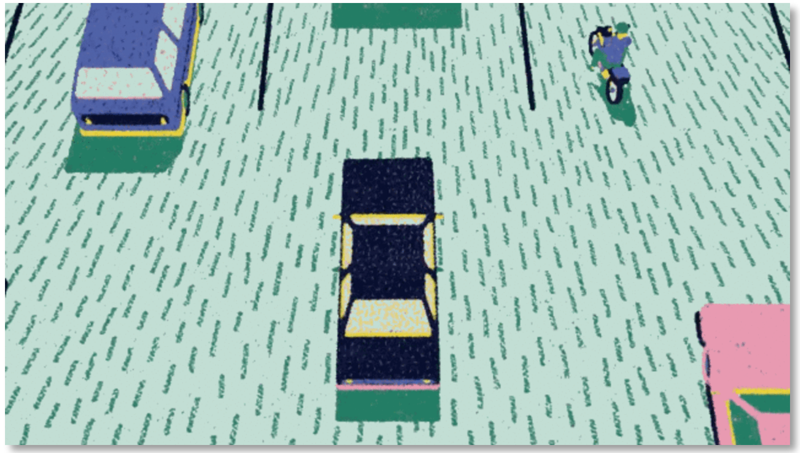
$v_\pi(s_1) = \pi(a_1|s_1)\big(5p(s_2, 5|s_1, a_1) + 10p(s_3, 10|s_1, a_1)\big) +$
$+ \pi(a_2|s_1)\big(25p(s_4, 25|s_1, a_2) + 35p(s_5, 35|s_1, a_2)\big) =$
$= 0.7 \cdot (0.4 \cdot 5 + 0.6 \cdot 10) + 0.3 \cdot (0.1 \cdot 25 + 0.9 \cdot 35) = \mathbf{15.8}$

Формула оптимальности Беллмана для качества действия

$$q_*(s, a) = \sum_{\acute{s}, \acute{r}} p(\acute{s}, \acute{r} | s, a) \left(\acute{r} + \gamma \max_{\acute{a}} q_*(\acute{s}, \acute{a}) \right)$$

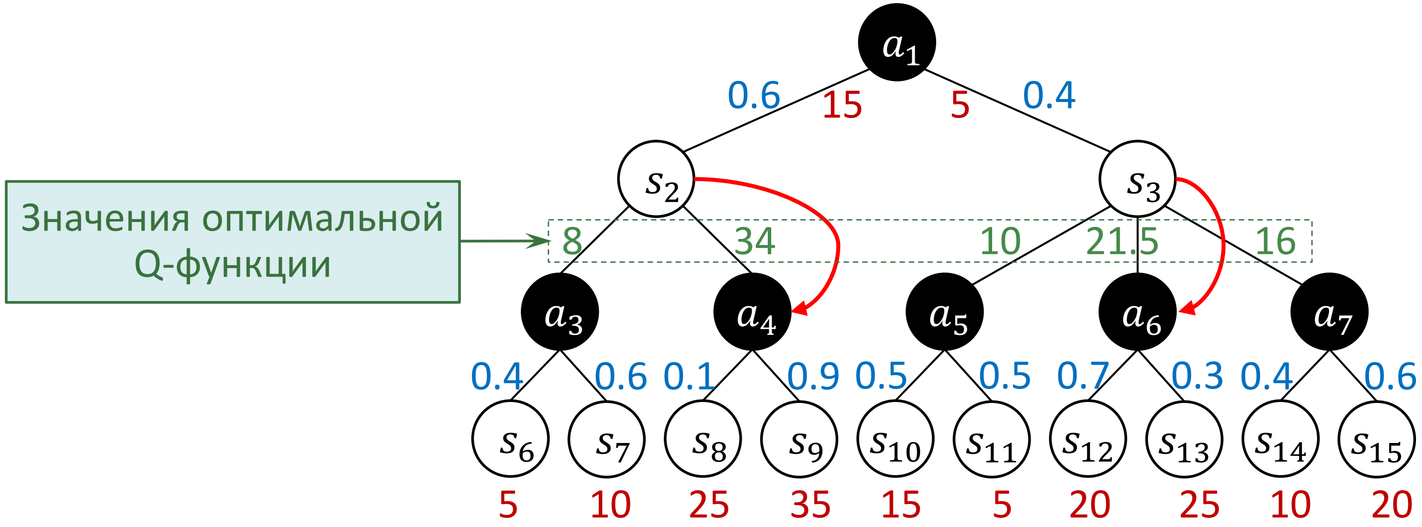
Модель среды

- Модель среды считается известной, если в дереве состояний известны вероятности всех переходов и все вознаграждения
- Если модель среды известна, то с помощью формулы оптимальности Беллмана для качества действия можно вычислить все значения оптимальной Q-функции
- Модель среды является основой для *обучения с подкреплением на основе модели* (Model-based reinforcement learning)



Жадная оптимальная стратегия для известной оптимальной Q-функции

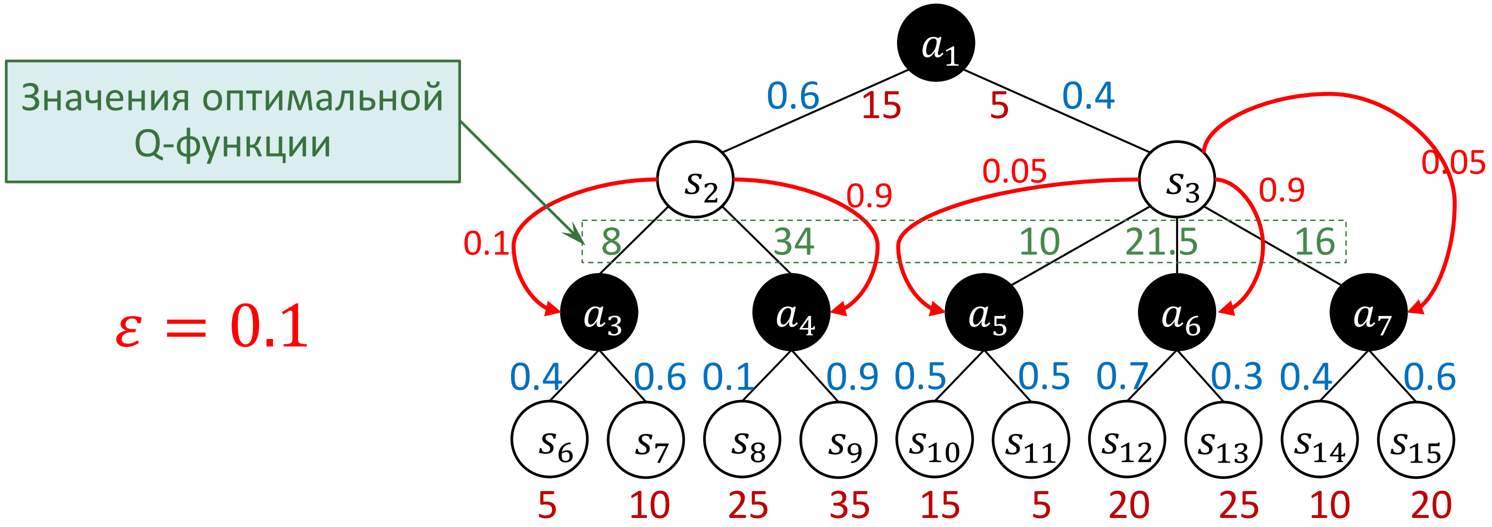
В состоянии s выбираем действие a, которое имеет максимальное качество



ε-жадная оптимальная стратегия для известной оптимальной Q-функции

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

$\varepsilon = 0.1$



Q-таблица

Действия

Состояния

	a_1	a_2	a_3
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

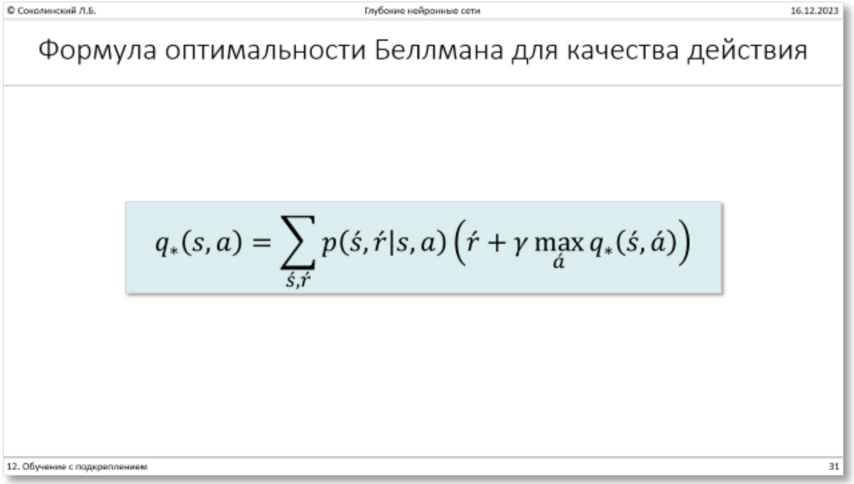
$$(S_1, A_1, R_1) \to (S_2, A_2, R_2) \to \dots \to (S_T, A_T, R_T)$$

обновляем соответствующие элементы Q-таблицы по формуле

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

η – скорость обучения

γ – коэффициент обесценивания



Алгоритм Q-обучения

Заполняем Q-таблицу нулями

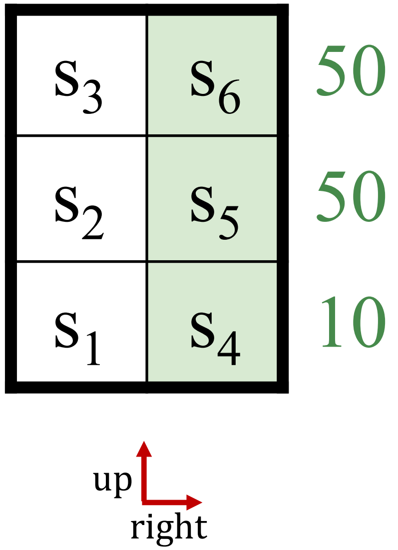
Если известна модель среды (*следующее состояние и вознаграждение однозначно определяются действием агента*), генерируем эпизоды, выбирая действия случайным образом и выполняем Q-обучение

Если модель среды неизвестна (*среда включает взаимодействие с контрагентом*), генерируем эпизоды с помощью ε-жадной[*] оптимальной стратегии и выполняем Q-обучение

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

Задача 3. Микрошашки (пример Q-обучения для известной модели среды)

- 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 $\eta = 0.1$, $\gamma = 1$)



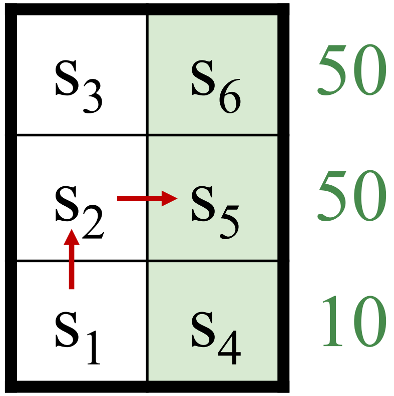
Эпизод 1

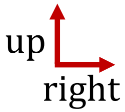
$$(s_1, \text{up}, 0) \to (s_2, \text{right}, 50) \to (s_5, \blacksquare, 0)$$

- 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 $= 0.1(-1)$	0	0
right	0	5 $= 0.1 \cdot 50$	0





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

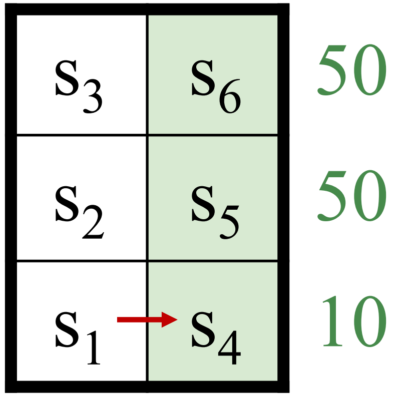
Эпизод 2

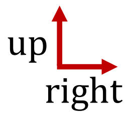
$$(s_1, \text{right}, 10) \rightarrow (s_4, \blacksquare, 0)$$

- 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)

	s1	s2	s3
up	-0.1	0	0
right	0	5	0

	s1	s2	s3
up	-0.1	0	0
right	1 $= 0.1 \cdot 10$	5	0





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

Эпизод 3

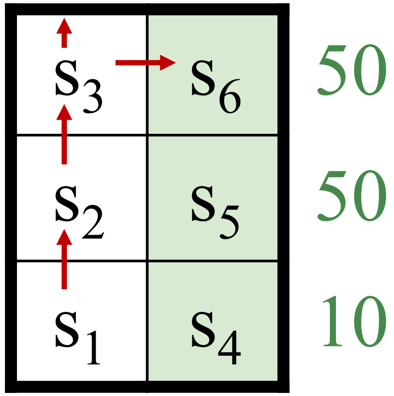
$$(s_1, \text{up}, 0) \rightarrow (s_2, \text{up}, -1) \rightarrow (s_3, \text{up}, -10) \rightarrow (s_3, \text{right}, 50) \rightarrow (s_6, \blacksquare, 0)$$

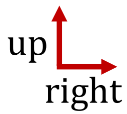
- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение 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 $= -0.1 + 0.1(-1 + 5 + 0.1)$	-0.1 $= 0.1(-1)$	-1 $= 0.1(-10)$
right	1	5	5 $= 0.1 \cdot 50$





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

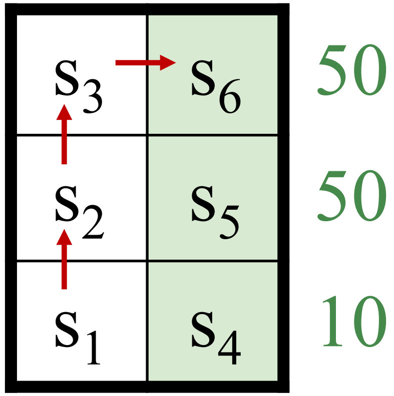
Эпизод 4

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

- 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 $= 0.31 + 0.1(-1 + 5 - 0.31)$	0.31 $= -0.1 + 0.1(-1 + 5 + 0.1)$	-1
right	1	5	9.5 $= 5 + 0.1(50 - 5)$



Порядок эпизодов важен!

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

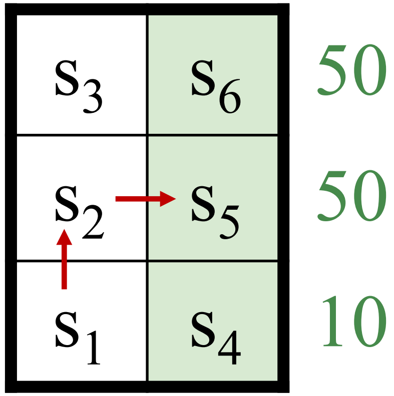
Эпизод 5

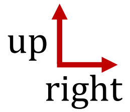
$$(s_1, \mathrm{up}, 0) \rightarrow (s_2, \mathrm{right}, 50) \rightarrow (s_5, \blacksquare, 0)$$

- 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 $= 0.679 + 0.1(-1 + 5 - 0.679)$	0.31	-1
right	1	9.5 $= 5 + 0.1(50 - 5)$	9.5





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

Эпизод 6

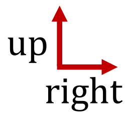
$$(s_1, \text{right}, 10) \rightarrow (s_4, \blacksquare, 0)$$

- 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 $= 1 + 0.1(10 - 1)$	9.5	9.5

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



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

Эпизод 7

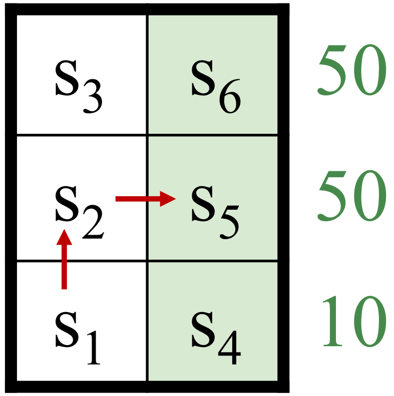
$$(s_1, \mathrm{up}, 0) \to (s_2, \mathrm{right}, 50) \to (s_5, \blacksquare, 0)$$

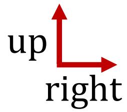
- Двигать шашку можно на одну клетку вверх (up) или вправо (right)
- При попадании в зеленую клетку начисляется указанное справа от нее вознаграждение и игра заканчивается
- При попытке выхода за границы доски, шашка остается в той же клетке и начисляется вознаграждение (-10)
- За каждое перемещение 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 $\approx 1.0111 + 0.1(-1 + 9.5 - 1.0111)$	0.31	-1
right	1.9	13.55 $= 9.5 + 0.1(50 - 9.5)$	9.5





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

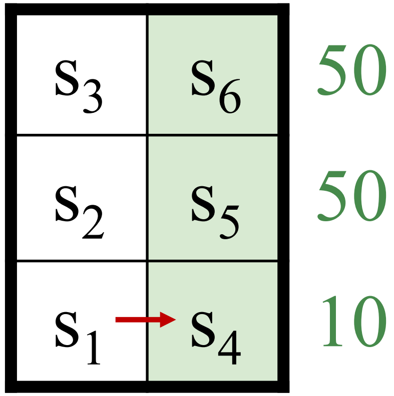
Эпизод 8

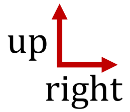
$$(s_1, \text{right}, 10) \to (s_4, \blacksquare, 0)$$

- 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 $= 1.9 + 0.1(10 - 1.9)$	13.55	9.5





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

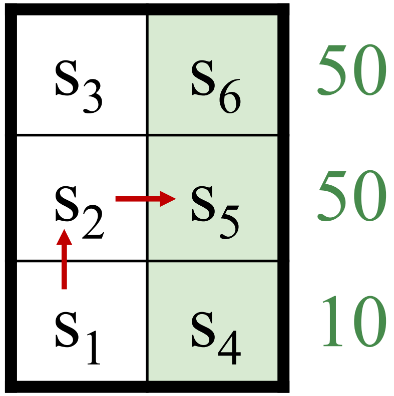
Эпизод 9

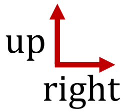
$$(s_1, \mathrm{up}, 0) \rightarrow (s_2, \mathrm{right}, 50) \rightarrow (s_5, \blacksquare, 0)$$

- 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 $\approx 1.76 + 0.1(-1 + 13.55 - 1.76)$	0.31	-1
right	2.71	17.2 $\approx 13.55 + 0.1(50 - 13.55)$	9.5



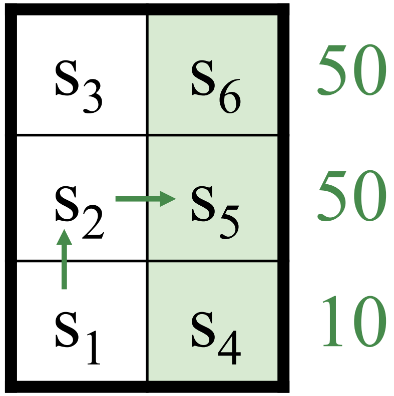


$$Q_{S_t,A_t} \coloneqq Q_{S_t,A_t} + 0.1\left(R_{t+1} + \max_a Q_{S_{t+1},a} - Q_{S_t,A_t}\right)$$

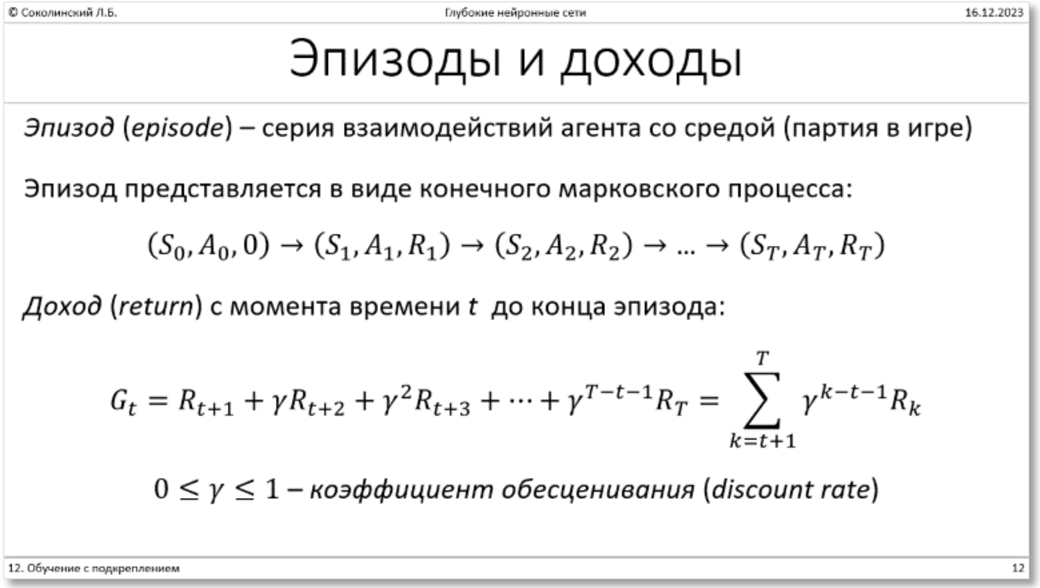
После Q-обучения жадная оптимальная стратегия обеспечивает максимальный доход

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

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



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


© Соколинский Л.Б. Глубокие нейронные сети 16.12.2023

Эпизоды и доходы

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

Эпизод представляется в виде конечного марковского процесса:

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

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

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

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

12. Обучение с подкреплением 12

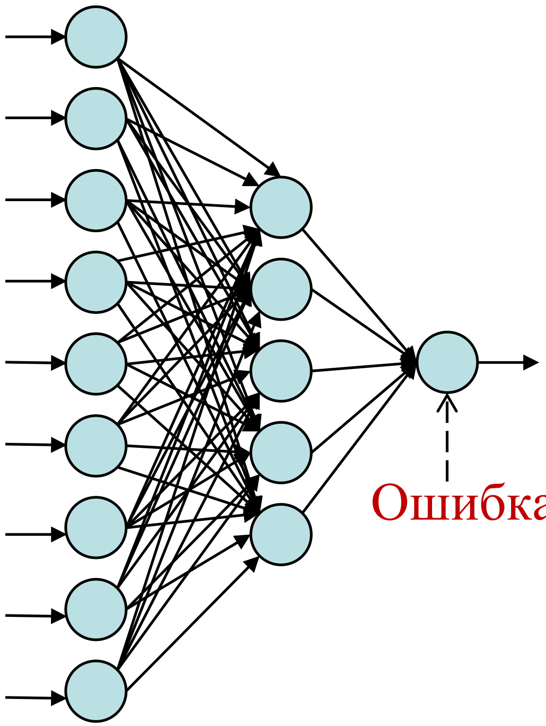
Проблема комбинаторного взрыва

- 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

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

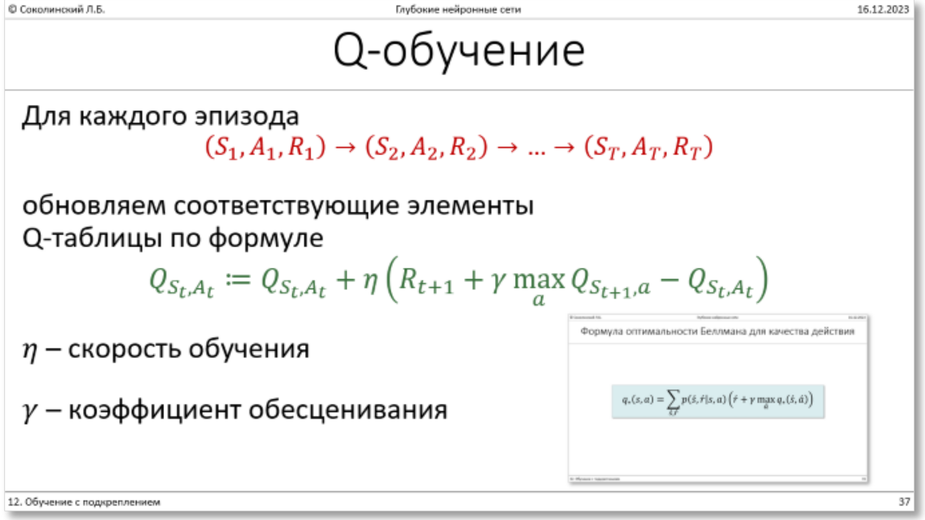
Состояние и действие: (S_t, A_t)



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

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

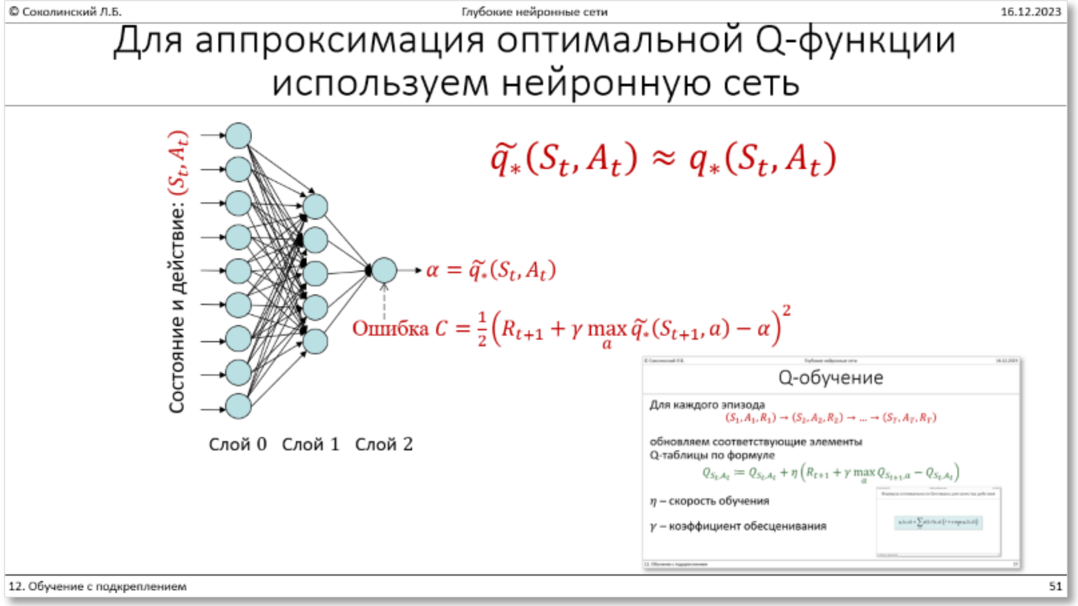
Слой 0 Слой 1 Слой 2



Для построения эпизода используем ε-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



Глубокое обучение с подкреплением

1. Инициализируем веса $\boldsymbol{w}$ каким-либо образом

2. С помощью $\boldsymbol{\varepsilon}$-жадной оптимальной стратегии генерируем эпизод, вычисляя качество возможных действий с помощью нейронной сети:

$$(S_0, A_0, 0) \to (S_1, A_1, R_1) \to \cdots \to (S_T, \blacksquare, R_T)$$

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

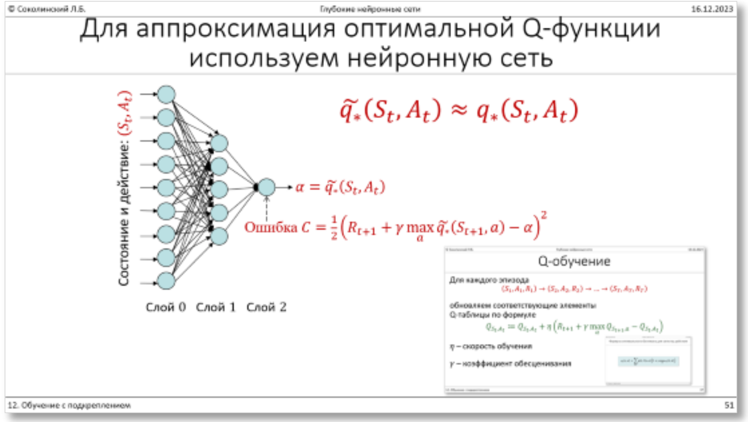
$$\boldsymbol{w} := \boldsymbol{w} - \eta \nabla_{\boldsymbol{w}} C$$

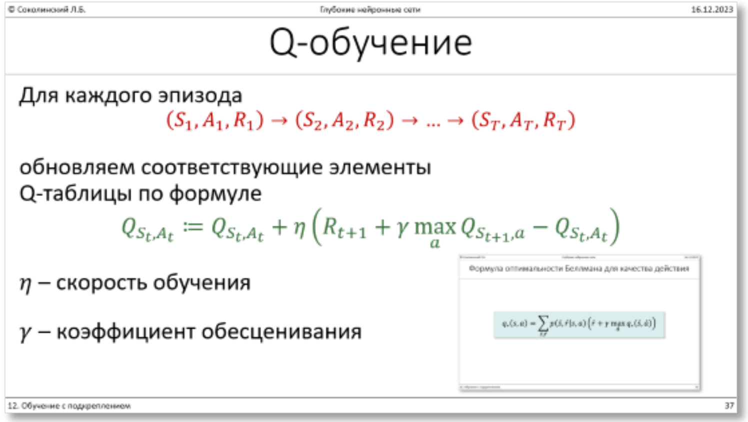
- Ошибка вычисляется по формуле

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

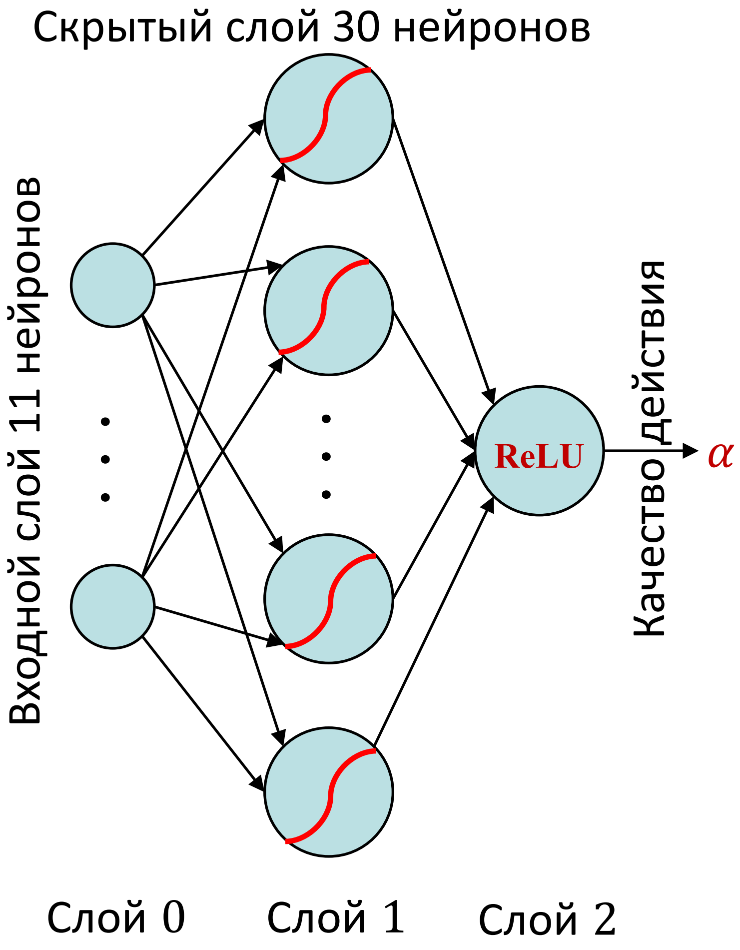
- Для вычисления градиентов используем метод обратного распространения ошибки

4. Повторяем шаги 2-3 много раз





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 $\gamma = 1, \eta = 0.1$

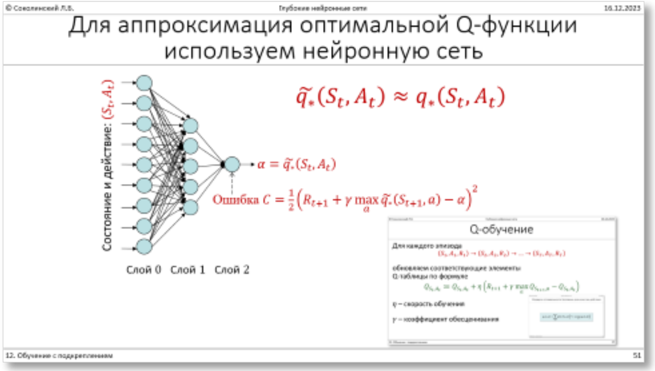
Обратное распространение ошибки

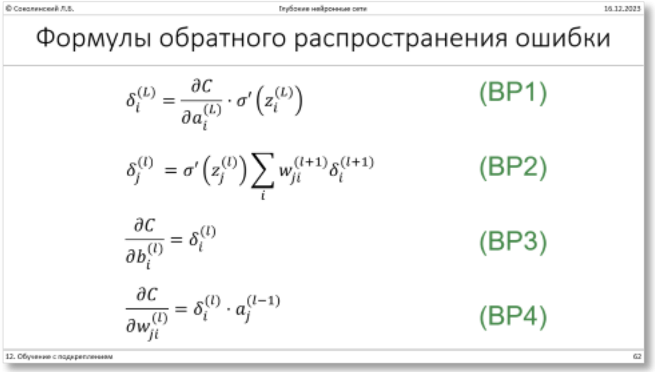
$$\delta_1^{(2)} = \frac{\partial C}{\partial \alpha} = \frac{\partial\left(\frac{1}{2}\left(R_{t+1} + \max_a \tilde{q}_*(S_{t+1}, a) - \alpha\right)^2\right)}{\partial \alpha} = \alpha - R_{t+1} - \max_a \tilde{q}_*(S_{t+1}, a)$$

$$\delta_j^{(1)} = \acute{\sigma}\left(z_j^{(1)}\right) w_{j1}^{(2)} \delta^{(2)}$$

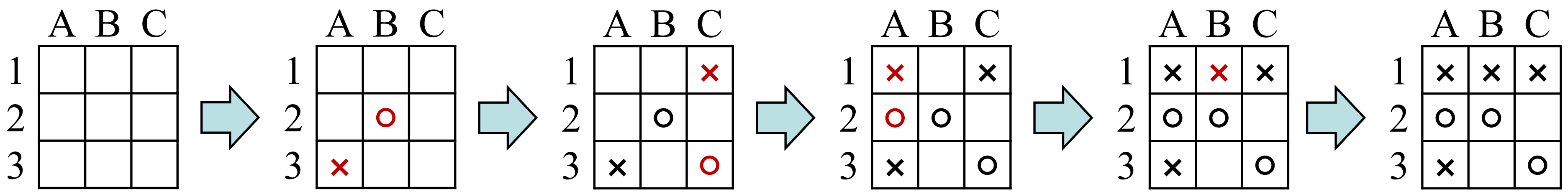
$$\frac{\partial C}{\partial w_{j1}^{(2)}} = \delta_1^{(2)} \cdot \sigma\left(z_j^{(1)}\right)$$

$$\frac{\partial C}{\partial w_{ji}^{(1)}} = \delta_i^{(1)} \cdot x_j$$





Генерация эпизодов и обучение



$(S_0, A3,0) \to (S_1, C1,0) \to (S_2, C1,0) \to (S_3, A1,0) \to (S_4, B1,0) \to (S_4, \blacksquare, 10)$

- 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: $\acute{\boldsymbol{w}} \coloneqq \boldsymbol{w}$
    - На протяжении одной игры ходы делаются на основе $\boldsymbol{w}$, а корректируется $\acute{\boldsymbol{w}}$
    - По завершению игры выполняется присваивание: $\boldsymbol{w} \coloneqq \acute{\boldsymbol{w}}$
  - Рекомендуется постепенно уменьшать скорость обучения η
- Обучение с использованием модели среды

Конец лекции 12

Вспомогательные слайды

Андрей Андреевич Марков

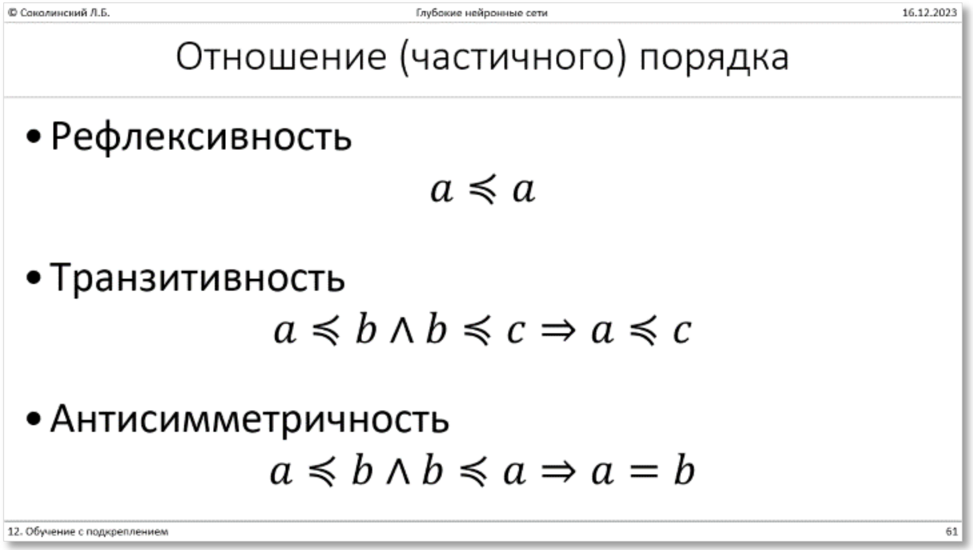
Андрей Андреевич Марков
(2 июня 1856 — 20 июля 1922) русский математик, академик, внесший большой вклад в теорию вероятностей, математический анализ и теорию чисел.

- 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

$$a \preccurlyeq a$$



- Транзитивность

$$a \preccurlyeq b \wedge b \preccurlyeq c \Rightarrow a \preccurlyeq c$$

Отношение (частичного) порядка

- Рефлексивность

$$a \preccurlyeq a$$

- Транзитивность

$$a \preccurlyeq b \wedge b \preccurlyeq c \Rightarrow a \preccurlyeq c$$

- Антисимметричность

$$a \preccurlyeq b \wedge b \preccurlyeq a \Rightarrow a = b$$

Формулы обратного распространения ошибки

$$\delta_i^{(L)} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \sigma'\left(z_i^{(L)}\right) \qquad \text{(BP1)}$$

$$\delta_j^{(l)} = \sigma'\left(z_j^{(l)}\right) \sum_i w_{ji}^{(l+1)} \delta_i^{(l+1)} \qquad \text{(BP2)}$$

$$\frac{\partial C}{\partial b_i^{(l)}} = \delta_i^{(l)} \qquad \text{(BP3)}$$

$$\frac{\partial C}{\partial w_{ji}^{(l)}} = \delta_i^{(l)} \cdot a_j^{(l-1)} \qquad \text{(BP4)}$$