

Введение в обучение с подкреплением

## Тема 3: Марковские процессы

Лектор: Кривошеин А.В.

## RL: общие понятия

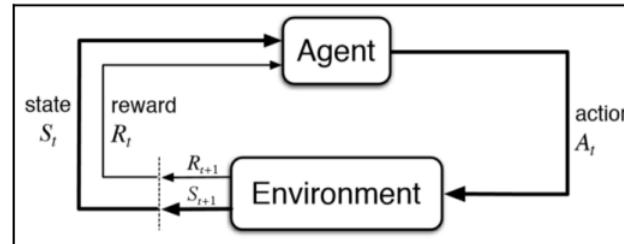
### Суть RL:

*научить обучаемого агента предпринимать "хорошие" последовательности действий в ходе взаимодействия со средой для эффективного достижения цели.*

**Цель агента:** поиск стратегии поведения, дающей максимальный доход.

### Схема взаимодействия:

обучаемый агент действует в окружающей его среде, среда реагирует на действия агента.



### Обозначения:

$\mathcal{S}$  — множество состояний,

$S_t$  — состояние в момент времени  $t$ ,

$\mathcal{A}$  — множество действий,

$A_t$  — действие в момент времени  $t$ ,

$\mathcal{R}$  — множество вознаграждений,

$R_t$  — вознаграждение в момент времени  $t$ .

Взаимодействие агента со средой — это **дискретный случайный процесс**:  $S_0, A_0, R_1, S_1, A_1, \dots$

Конкретная реализация этого процесса — это **траектория** взаимодействия агента со средой.

## Дискретный марковский процесс

Рассмотрим **частный случай** общей постановки задачи RL.

Агента нет. А среда меняет своё состояние в дискретные моменты времени.

Пусть  $\mathcal{S} = \{1, 2, \dots, N\}$  — конечное множество состояний среды,

а набор  $\{p_i\}_{i \in \mathcal{S}}$  задаёт вероятности появления того или иного состояния в начальный момент времени  $t = 0$ .

**Дискретный процесс смены состояний** — это последовательность случайных величин

$S_0, S_1, \dots, S_t, \dots$ , каждая из которых отвечает за состояние среды в момент времени  $t$ .

Этот процесс называют **марковским** (англ. Markov chain), если

$$\mathbb{P}(S_{t+1} = s_{t+1} | S_t = s_t, \dots, S_0 = s_0) = \mathbb{P}(S_{t+1} = s_{t+1} | S_t = s_t), \text{ где } s_0, \dots, s_{t+1} \in \mathcal{S}.$$

Иными словами, вероятность перехода среды в новое состояние  $s_{t+1}$

зависит только от текущего состояния  $s_t$

и не зависит от прошлых состояний среды.

Кроме того, будем рассматривать **стационарные** марковские процессы, то есть вероятность перехода среды из текущего состояния в новое состояние не меняется со временем.

Для полного описания **стационарной марковской среды** достаточность задать:

- матрицу вероятностей  $\{p_{i,j}\}_{i,j \in \mathcal{S}}$ , где  $p_{i,j}$  задаёт вероятность перехода среды из состояния  $i$  в состояние  $j$ ,

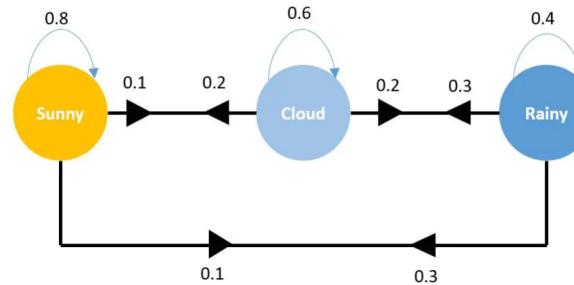
то есть  $p_{i,j} = \mathbb{P}(S_{t+1} = j | S_t = i)$ ,  $\sum_{j \in \mathcal{S}} p_{i,j} = 1$ , для каждого  $i \in \mathcal{S}$ ;

- вектор вероятностей  $\{p_i\}_{i \in \mathcal{S}}$ , где  $p_i$  задаёт вероятность нахождения среды в состоянии  $i$

в начальный момент времени  $t = 0$ .

## Дискретный марковский процесс: пример

Рассмотрим пример, моделирующий погоду на день в некотором городе. Выделим три состояния погоды: солнечно, облачно, идёт дождь. Переходы из одного состояния в другое моделируются с помощью следующей схемы.



Матрица вероятностей перехода между состояниями имеет вид

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}.$$

## Марковский процесс вознаграждения

Добавим в рассмотренный выше процесс агента, который наблюдает переходы между состояниями среды и получает вознаграждения, не совершая при этом действий.

Этот процесс называют **марковским процессом вознаграждений** (МПВ, англ. Markov's reward process).

Для задания **дискретного** МПВ надо определить следующие понятия.

1. Конечное (или счётное) множество состояний среды  $S$ .
2. Модель среды или дискретный стационарный марковский процесс смены состояний  $S_0, S_1, S_2 \dots$

Для этого достаточно задать вероятности переходов между состояниями

$$p(s' | s) := \mathbb{P}(S_{t+1} = s' | S_t = s) \in [0, 1], \text{ где } s, s' \in S, \text{ причём } \sum_{s' \in S} p(s' | s) = 1.$$

3. Вознаграждение за пребывание в каждом состоянии.

В детерминированном случае достаточно задать функцию  $r: S \rightarrow \mathcal{R} \subset \mathbb{R}$ , определяющую значения  $r(s)$ .

В стохастическом случае надо задать набор распределений вероятностей  $p(r | s) := \mathbb{P}(R_{t+1} = r | S_t = s)$ .

При этом под  $r(s)$  удобно понимать математическое ожидание вознаграждения за пребывание в состоянии  $s$ :

$$r(s) = \mathbb{E}[R_{t+1} | S_t = s].$$

Под **конечным МПВ** будем понимать МПВ, где множества состояний  $S$  и вознаграждений  $\mathcal{R}$  конечны.

Далее, по умолчанию, будем рассматривать именно конечные МПВ.

## Марковский процесс вознаграждения

**Траекторией эпизодического** МПВ является реализация случайных величин

$$S_0, R_1, S_1, R_2, \dots R_T, S_T, \quad \text{где либо } S_T \in \mathcal{S}_+, \text{ либо } T = T_{\max}.$$

Здесь  $\mathcal{S}_+$  множество терминальных состояний, а  $T_{\max}$  — это максимальное число дискретных шагов.

**Траекторией бесконечного** МПВ является реализация последовательности случайных величин

$$S_0, R_1, S_1, R_2, \dots$$

В этой записи можно понимать и эпизодический МПВ с терминальными состояниями. Можно считать, что при попадании в терминальное состояние среда с вероятностью 1 переходит из этого состояния в него же, а агент получает нулевое вознаграждение.

**Доход** агента  $G_t$  — это взвешенная сумма всех полученных вознаграждений с момента времени  $t$  и до конца процесса или до бесконечности, то есть

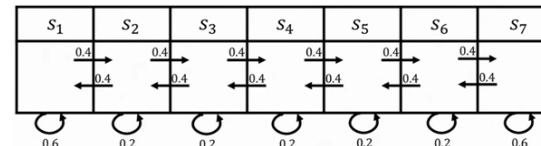
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}.$$

Здесь  $\gamma \in [0, 1]$  коэффициент обесценивания. Важно заметить, что в этом обозначении вознаграждения начинают накапливаться с  $R_{t+1}$ .

## МПВ: пример

Доход:  $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$

**Пример.** Рассмотрим пример с 7-ю состояниями, где вероятности смены состояний задаются схемой



Вознаграждения зададим следующим образом

$$r(s_1) = -1, \quad r(s_7) = 10, \quad r(s_i) = 0 \text{ при } i = 2, 3, 4, 5, 6.$$

Пусть  $\gamma = \frac{1}{2}$ . Для примера найдём доходы для нескольких 4-шаговых эпизодов с различными траекториями:

$$s_4, s_5, s_6, s_7: \quad G_0 = 0 + 0 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} + 10 \cdot \frac{1}{8} = 1.25,$$

$$s_4, s_5, s_4, s_3: \quad G_0 = 0 + 0 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} + 0 \cdot \frac{1}{8} = 0,$$

$$s_4, s_3, s_2, s_1: \quad G_0 = 0 + 0 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} - 1 \cdot \frac{1}{8} = -0.125.$$

## Вероятностное пространство и случайная величина

Напомним определение и некоторые свойства **математического ожидания случайных величин**.

Формально, сначала определяется **вероятностное пространство** ( $\Omega, \mathcal{A}, \mathbb{P}$ ):

$\Omega$  – множество **элементарных событий** (у нас это совокупность всевозможных траекторий конечного МПВ);

$\mathcal{A}$  – **сигма-алгебра** событий (в нашем случае  $\mathcal{A} = 2^\Omega$ );

$\mathbb{P}$  – **вероятность**, заданная на сигма-алгебре (то есть мера со свойством  $\mathbb{P}(\Omega) = 1$ ).

В случае конечного МПВ вероятность  $\mathbb{P}$  будет задана, если будет **задана вероятность появления каждой траектории**.

Для этого достаточно задать:

вероятности переходов между состояниями  $\{p(s' | s)\}_{s', s \in S}$ ,

вероятности  $\{p(s)\}_{s \in S}$  появления состояний в начальный момент времени  $t = 0$ ,

вероятности  $\{p(r | s)\}_{r \in R, s \in S}$ , вознаграждений за пребывание в каждом из состояний.

Тогда вероятность появления некоторой траектории  $\tau$  с конкретным набором состояний и вознаграждений вида:

$\tau = (s_0, r_1, s_1, \dots, r_T, s_T)$  будет равна  $\mathbb{P}(\tau) = p(s_0) p(r_1 | s_0) p(s_1 | s_0) p(r_2 | s_1) p(s_2 | s_1) \dots p(r_T | s_{T-1}) p(s_T | s_{T-1})$ .

**Случайная величина** – это (измеримое) отображение  $\xi: \Omega \rightarrow \mathbb{R}$ .

В случае конечного МПВ случайная величина задаётся на совокупности траекторий.

В частности, вознаграждение  $R_t$ , доход  $G_t$ , состояние  $S_t$  – это случайные величины (формально состояния, конечно, не обязаны быть числами, но все состояния можно проиндексировать и заменить состояния индексами).

Например, для траектории  $\tau$  вида  $\tau = (s_0, r_1, s_1, \dots, r_T, s_T)$  эти случайные величины принимают значения:

$$R_t(\tau) = r_t, \quad S_t(\tau) = s_t, \quad G_t(\tau) = r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{T-t-1} r_T.$$

## Математическое ожидание

**Математическое ожидание** — это интеграл от случайной величины по мере  $\mathbb{P}$ :

$$\mathbb{E}[\xi] = \int_{\Omega} \xi d\mathbb{P}.$$

Свойства математического ожидания — это свойства интеграла по мере (линейность, аддитивность и т.д.).

Пусть  $\xi$  — дискретная случайная величина, принимающая конечный набор значений  $x_1, \dots, x_N$  с вероятностями  $p_1, \dots, p_N$ , то есть  $\mathbb{P}(\xi = x_i) = p_i$ . Тогда математическое ожидание  $\xi$  имеет вид

$$\mathbb{E}[\xi] = \sum_{i=1}^N x_i p_i.$$

**Условное математическое ожидание** случайной величины  $\xi$  при условии события  $A \in \mathcal{A}$ :

введём обозначение для события  $B_i = \{\xi = x_i\} = \xi^{-1}(x_i)$ , тогда

$$\mathbb{E}[\xi|A] = \sum_{i=1}^N x_i \mathbb{P}(B_i|A) = \sum_{i=1}^N x_i \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \frac{1}{\mathbb{P}(A)} \mathbb{E}[\xi \chi_A].$$

## Математическое ожидание: примеры

Рассмотрим конечный МПВ. Найдём вероятность события, что начальное состояние равно  $s$ , то есть  $S_0 = s$ . Для этого надо просуммировать вероятности появления всех траекторий, у которых начальное состояние равно  $s$

$$\mathbb{P}(S_0 = s) = \sum_{\tau \in \Omega: S_0(\tau) = s} \mathbb{P}(\tau) = \sum_{r_1 \in \mathcal{R}} \sum_{s_1 \in \mathcal{S}} \dots \sum_{s_{T-1} \in \mathcal{S}} \sum_{r_T \in \mathcal{R}} \sum_{s_T \in \mathcal{S}} p(s) p(r_1 | s) p(s_1 | s_0) \dots p(r_T | s_{T-1}) p(s_T | s_{T-1}) = p(s).$$

Аналогичным образом можно показать, что, например,

$\mathbb{P}(S_{t+1} = s' | S_t = s)$  действительно равно  $p(s' | s)$ ,

а  $\mathbb{P}(R_{t+1} = r | S_t = s)$  действительно равно  $p(r | s)$ .

Далее, можно также вычислять математические ожидания от вознаграждений при условии, что среда находится в некотором заданном состоянии:

$$r(s) = \mathbb{E}[R_{t+1} | S_t = s] = \sum_{r \in \mathcal{R}} r \mathbb{P}(R_{t+1} = r | S_t = s) = \sum_{r \in \mathcal{R}} r p(r | s).$$

## Математическое ожидание: формула полной вероятности

Рассмотрим разбиение множества  $\Omega$  на конечное число частей  $B_1, \dots, B_n$ , где

$$\Omega = \bigcup_{i=1}^n B_i, \quad B_i \cap B_j = \emptyset, \quad B_i \in \mathbb{A}.$$

Формула полной вероятности для условного математического ожидания  $\xi$  при условии события  $A$  имеет вид

$$\mathbb{E}[\xi|A] = \sum_{i=1}^n P(B_i|A) \mathbb{E}[\xi|A \cap B_i].$$

В силу определения условного математического ожидания и формулы полной вероятности верны равенства

$$\begin{aligned} \mathbb{E}[\xi|A] &= \frac{1}{\mathbb{P}(A)} \mathbb{E}[\xi \chi_A] = \frac{1}{\mathbb{P}(A)} \mathbb{E}\left[\sum_{i=1}^n \xi \chi_A \chi_{B_i}\right] = \\ &= \sum_{i=1}^n \frac{\mathbb{P}(A \cap B_i)}{\mathbb{P}(A)} \frac{1}{\mathbb{P}(A \cap B_i)} \mathbb{E}[\xi \chi_{A \cap B_i}] = \sum_{i=1}^n \mathbb{P}(B_i|A) \mathbb{E}[\xi|A \cap B_i]. \end{aligned}$$

## МПВ: ценности состояний

**Функция ценности состояния**  $v(s)$  определяется как ожидаемый доход по всем возможным траекториям при старте из состояния  $s$ :

$$v(s) := \mathbb{E}[G_t | S_t = s] = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right] = \mathbb{E}[R_{t+1} + \gamma G_{t+1} | S_t = s] = r(s) + \gamma \mathbb{E}[G_{t+1} | S_t = s].$$

Целый ряд методов RL использует функции ценности для нахождения оптимальных стратегий. Поэтому важно уметь вычислять их или оценивать.

Для вычисления функции ценности состояний в рамках МПВ можно предложить два метода.

**Первый метод** основан на том, чтобы сформировать много траекторий и усреднить доходы, полученные из каждого состояния. Для реализации этого метода модель среды (то есть знание вероятностей перехода между состояниями) не требуется. Более того, не требуется даже, чтобы среда была марковской.

**Второй метод** позволяет получить точное решение, но для этого среда должна быть марковской и модель среды должны быть точно известна. Этот метод основан на том, чтобы сформировать уравнения, в которых участвуют искомые величины  $v(s)$ .

## МПВ: вычисление ценности состояний

Перепишем определение функции ценности состояний в виде уравнения относительно значений  $v(s)$ :

$$\begin{aligned} v(s) &= \mathbb{E}[G_t | S_t = s] = r(s) + \gamma \mathbb{E}[G_{t+1} | S_t = s] \\ &= r(s) + \gamma \sum_{s' \in S} p(s' | s) v(s'), \quad s \in S. \end{aligned}$$

Равенство

$$\mathbb{E}[G_{t+1} | S_t = s] = \sum_{s' \in S} p(s' | s) v(s') = \sum_{s' \in S} \mathbb{P}(S_{t+1} = s' | S_t = s) \mathbb{E}[G_{t+1} | S_t = s, S_{t+1} = s']$$

верно в силу формулы полной вероятности для условного математического ожидания, где

$$\xi = G_{t+1}, \quad A = \{\tau: S_t = s\}, \quad B_{s'} = \{\tau: S_{t+1} = s'\}.$$

Осталось заметить, что в силу марковости среды (будущее зависит лишь от текущего состояния), верно

$$\mathbb{E}[G_{t+1} | S_{t+1} = s', S_t = s] = \mathbb{E}[G_{t+1} | S_{t+1} = s'] = v(s').$$

## МПВ: вычисление ценности состояний

Итак, функция ценности состояний  $v(s)$  удовлетворяет уравнению:

$$v(s) = r(s) + \gamma \sum_{s' \in S} p(s' | s) v(s'), \quad s \in S.$$

Это уравнение для ценности состояний  $v(s)$  можно записать в матричном виде, положив

$$V = (v(s_1), \dots, v(s_N))^T,$$

$$R = (r(s_1), \dots, r(s_N))^T,$$

$$P = \{p(s' | s)\}_{s, s' \in S}.$$

Тогда

$$V = R + \gamma P V \text{ или } (I - \gamma P) V = R. \quad \textbf{Точное решение: } V = (I - \gamma P)^{-1} R.$$

Для больших размерностей обращение матрицы может быть вычислительно не устойчивым и не эффективным.

**Итеративный способ.** Пусть  $V_0(s) = 0$  для всех  $s$ . Для  $k \in \mathbb{N}$  полагаем,

$$V_{k+1} := R + \gamma P V_k.$$

Решение считается полученным, если для некоторого  $k \in \mathbb{N}$  выполняется оценка  $\|V_{k-1} - V_k\| < \varepsilon$  при заданном  $\varepsilon > 0$ .

## МПВ: вычисление ценности состояний

Гарантия сходимости итеративного метода следует из **принципа сжимающих отображений**. Пусть оператор  $T$  действует на векторах со значениями ценностей состояний  $V = (\nu(s_1), \dots, \nu(s_N))$  по правилу

$$T(V) := R + \gamma P V. \quad \text{Итерации имеют вид: } V_{k+1} = T(V_k).$$

Если  $\gamma < 1$ , то **оператор  $T$  является сжимающим** в max-норме  $\|V\| = \max_s |\nu(s)|$ .

Действительно, рассмотрим два вектора

$V = (\nu(s_1), \dots, \nu(s_N))$  и  $W = (w(s_1), \dots, w(s_N))$  и проведём оценку

$$\|T(V) - T(W)\| \leq \|\gamma P(V - W)\| \leq$$

$$\gamma \max_s \sum_{s' \in S} p(s' | s) |\nu(s') - w(s')| \leq$$

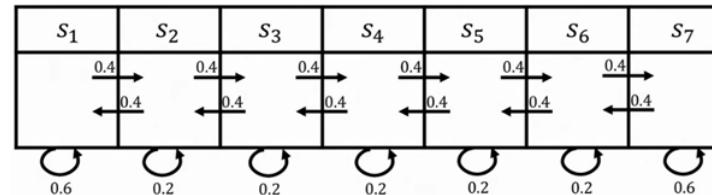
$$\gamma \|V - W\|, \text{ так как } \sum_{s' \in S} p(s' | s) = 1 \text{ для любого } s.$$

Тогда теорема о сжимающем отображении гарантирует сходимость итеративной последовательности приближений  $V_k$  к неподвижной точке оператора  $T$ , то есть к вектору  $V$  удовлетворяющему равенству  $T(V) = V$ , а это и есть искомое решение.

$$T(V) = R + \gamma P V = V$$

## МПВ: пример

**Пример.** Рассмотрим пример с 7-ю состояниями, где вероятности смены состояний задаются схемой:



Вознаграждения зададим следующим образом

$$r(s_1) = -1, \quad r(s_7) = 10, \quad r(s_i) = 0 \text{ при } i = 2, 3, 4, 5, 6.$$

Пусть  $\gamma = \frac{1}{2}$ . Найдём функцию ценности состояний по точному методу:

```

In[=] R = {-1, 0, 0, 0, 0, 0, 10};
P = {{0.6, 0.4, 0, 0, 0, 0, 0}, {0.4, 0.2, 0.4, 0, 0, 0, 0}, {0, 0.4, 0.2, 0.4, 0, 0, 0}, {0, 0, 0.4, 0.2, 0.4, 0, 0}, {0, 0, 0, 0.4, 0.2, 0.4, 0}, {0, 0, 0, 0, 0.4, 0.2, 0.4}, {0, 0, 0, 0, 0, 0.4, 0.6}};
gamma = 0.5;
Inverse[IdentityMatrix[7] - gamma P].R

```

Out[=]= { -1.52799, -0.347969, -0.037869, 0.177559, 0.836883, 3.58841, 15.311 }