Лекция 2

Обучение с подкреплением

Никита Юдин, iudin.ne@phystech.edu

Московский физико-технический институт
Физтех-школа прикладной математики и информатики

14 февраля 2024

ФПМИ

Марковский процесс принятия решений (MDP)

В обучении с подкреплением взаимодействия между агентом и окружающей средой часто описываются марковским процессом принятия решений (MDP). Различают:

- дисконтированный марковский процесс принятия решений (γ-*Discounted Markov Decision Process*, DMDP);
- MDP с усредненным вознаграждением (*infinite-horison Average reward Markov Decision Process*, AMDP);
- эпизодический марковский процесс принятия решений (*H-episodic Markov Decision Process*, HMDP);
- другие, в том числе и *частично наблюдаемые*.

Марковский процесс принятия решений (MDP)

Марковский процесс принятия решений представляет собой систему, которая со временем $(t = 0, 1, 2, \dots)$ претерпевает случайные изменения и обозначается кортежем $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$ со следующими объектами:

(i) $\mathcal{S}$ – пространство состояний, $S := |\mathcal{S}|$ – количество уникальных состояний.

(ii) $\mathcal{A}$ – пространство действий, $A := |\mathcal{A}|$ – количество уникальных действий.

(iii) $p\left(s, a; s'\right)$ – вероятность перехода из состояния $s \in \mathcal{S}$ в момент времени t с определенным действием $a \in \mathcal{A}$ в состояние $s' \in \mathcal{S}$ в момент $(t+1)$ (при этом $\sum\limits_{s' \in \mathcal{S}} p\left(s, a; s'\right) = 1$, $p\left(s, a; s'\right) \equiv \mathrm{P}\left(s' \middle| s, a\right)$). Функция вероятности $p(\cdot)$ вместе с функцией вероятности $\mathrm{P}(a|s)$ задают вероятности перехода для Марковского ядра, оно же ядро MDP.

Марковский процесс принятия решений (MDP)

(iv) Функция награды
$r_\xi(s,a) : \Omega \times \mathcal{S} \times \mathcal{A} \to [0,1]$, ($\mathbb{E}_\xi\left[r_\xi(s,a)\right] = r(s,a)$, где $\mathbb{E}[\cdot]$ – математическое ожидание). В зависимости от постановки задачи функция награды может зависеть от следующего за состоянием s состояния s':

$$r_\xi\left(s,a;s'\right) : \Omega \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1],\ \mathbb{E}_\xi\left[r_\xi\left(s,a;s'\right)\right] = r\left(s,a;s'\right).$$

Стоит отметить, что мы предполагаем в общем случае стохастическую природу функции награды в зависимости от случайной величины $\xi \in \Omega$. При детерминированном вычислении награды относительно фиксированных (s,a) или (s,a,s') мы просто опускаем обозначение ξ в $r_\xi(\cdot)$ и математическое ожидание по нему.

Марковский процесс принятия решений (MDP)

(iv) *(продолжение)* Здесь и далее используется предположение об ограниченности награды за каждое действие, поэтому без ограничений общности использованы приведённые выше определения функции $r(\cdot)$. В работе используются детерминированные относительно своих аргументов награды, если не оговорено иное.

(v) $\gamma \in (0, 1]$ – коэффициент дисконтирования для DMDP, для AMDP $\gamma = 1$, но просто положив $\gamma = 1$ из DMDP не сделать AMDP, понадобится ещё усреднение суммарной награды агента за взаимодействие с MDP по времени.

Нередко рассматривается более общая форма $M = (\mathcal{S}, \mathcal{A}, p, r, \mu_0, \gamma)$, в которой μ_0 – вероятностное распределение начального состояния $s_0 \sim \mu_0$, при явном отсутствии μ_0 происходит обуславливание всех вычислений на $s_0 \in \mathcal{S}$.

MDP. Принятие решений

- Здесь и далее приводятся результаты для дискретных $\mathcal{S}$ и $\mathcal{A}$ с конечными мощностями, однако они могут быть обобщены на непрерывный случай заменой суммы по переменной в области её непрерывности на соответствующий интеграл по области.
- Стратегией принятия решений или политику агента, принимающего решение в MDP, обозначим через символ π и присвоим ему отображения, задающие вероятностную меру на пространстве действий:

$$\pi(a|s) \equiv \mathrm{P}(a|s) \text{ в общем случае, } \hat{a} \sim \pi(\cdot|s) \text{ или } \pi(s) \sim \pi(\cdot|s);$$

$$\hat{a} := \pi(s) \textit{ в случае вырожденного распределения: } \mathrm{P}(\hat{a}|s) = 1.$$

MDP. Ядро

- Введённое распределение позволяет явно записать Марковское ядро перехода между состояниями $s \mapsto s'$, оно же ядро MDP, его также корректно называть Марковским ядром, обусловленным политикой π:

$$\mathrm{P}^{\pi}(s'|s) := \sum_{a \in \mathcal{A}} \pi(a|s)p(s,a;s').$$

- В процессах с конечным количеством состояний Марковское ядро можно задать с помощью матрицы:

$$P^{\pi} = \left(\sum_{a \in \mathcal{A}} \pi(a|s)p(s,a;s') \right)_{s \in \mathcal{S},\ s' \in \mathcal{S}}, \quad s \text{ – строка, } s' \text{ – столбец.}$$

MDP. Ядро

В процессе взаимодействия с марковским процессом стратегия π собирает траекторию $\tau_t := (s_0, a_0, r_0, ..., s_t, a_t, r_t)$. Её правдоподобие выражается следующим образом:

$$\mathrm{P}\left(\tau_{H-1} | \pi\right) = \mu_0(s_0) \prod_{t=0}^{H-2} \left(\pi(a_t | s_t) p(s_t, a_t; s_{t+1})\right) \pi(a_{H-1} | s_{H-1}).$$

DMDP. V-функция ценности

Для фиксированной политики и начального состояния $s_0 = s$ определяется V-функция значений (ценности) $V^\pi : \mathcal{S} \to \mathbb{R}$ как дисконтированная сумма будущих вознаграждений:

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a(s_t)) \middle| \pi,\ s_0 = s\right],$$

где s_t – состояние системы в момент времени t, $a(s_t)$ – выбор действия в соответствии с политикой $\pi(\cdot)$. Это есть средняя награда по политике π, если агент начинает действовать в момент времени t из состояния s. Иногда индекс политики опускают: $V(s) := V^\pi(s)$.

Замечание

В определении V-функции в левой части выражения отсутствует обозначение t в силу однородности MDP.

DMDP. Q-функция ценности

Схожим образом задается Q-функция ценности $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a(s_t)) \middle| \pi,\; s_0 = s, a_0 = a\right].$$

То же, что V функция, только теперь из состояния $s_t = s$ обязательно сначала совершается действие $a_t = a$. Иногда индекс политики опускают: $Q(s, a) := Q^\pi(s, a)$.

Замечание

В определении V- и Q- функции в левой части выражения отсутствует обозначение t в силу однородности MDP.

DMDP. Дисконтированная награда

Дисконтированная кумулятивная награда за эпизод длины $H - t,\ t = \overline{0, H-1}$:

$$R_t^{H-1} := \sum_{j=t}^{H-1} \gamma^{j-t} r(s_j, a(s_j)) \text{ и } R_t := R_t^{\infty} := \sum_{j=t}^{\infty} \gamma^{j-t} r(s_j, a(s_j)).$$

При переходе к AMDP ($\gamma = 1$) наиболее естественным аналогом кумулятивной награды является среднее арифметическое наград по времени:

$$R_t^{H-1} := \frac{1}{H-t} \sum_{j=t}^{H-1} r(s_j, a(s_j)) \text{ и } R_t := R_t^{\infty} := \lim_{H \to \infty} \frac{1}{H-t} \sum_{j=t}^{H-1} r(s_j, a(s_j)).$$

DMDP. Дисконтированная награда

Мажоранта на $V^\pi(\cdot)$:

$$r(s,a) \in [0,1]: \quad 0 \le V^\pi(s) \le \frac{1}{(1-\gamma)} \quad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Q^π-функция обладает той же мажорантой, что и $V^\pi(\cdot)$:

$$r(s,a) \in [0,1]: \quad 0 \le Q^\pi(s,a) \le \frac{1}{(1-\gamma)} \quad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

DMDP. Уравнения Беллмана

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a(s_t)) \middle| \pi,\ s_0 = s\right] =$$

$$= \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} p(s, a; s') \left[r(s, a) + \gamma V^\pi(s')\right] = \mathbb{E}_\pi \left[Q^\pi(s, a)\right] =$$

$$= \mathbb{E}_\pi \left[r(s, a)\right] + \gamma \mathbb{E}_{p,\pi} \left[V^\pi(s')\right] = \mathbb{E}_\pi \left[r(s, a)\right] + \gamma \mathbb{E}_{p,\pi} \left[Q^\pi(s', a')\right];$$

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a(s_t)) \middle| \pi,\ s_0 = s, a_0 = a\right] =$$

$$= \sum_{s' \in \mathcal{S}} p(s, a; s') \left[r(s, a) + \gamma V^\pi(s')\right] =$$

$$= r(s, a) + \gamma \mathbb{E}_p \left[V^\pi(s')\right] = r(s, a) + \gamma \mathbb{E}_{p,\pi} \left[Q^\pi(s', a')\right],$$

$$a' \sim \pi(\cdot|s').$$

DMDP. Задача RL

Цель задачи обучения с подкреплением (Reinforcement Learning, RL) – поиск политики, позволяющей получить максимальное кумулятивное вознаграждение в долгосрочной перспективе. В большинстве практических случаев задача RL формулируется как задача оптимизации следующего формата:

$$\pi^* \in \operatorname*{Arg\,max}_{\pi \in \hat{\Pi}} \left\{ \mathbb{E}_{\mathrm{P}(\tau_{H-1}|\pi)} \left[R_0^{H-1} \right] = \mathbb{E}_{s \sim \mu_0} \left[V^\pi(s) \right] \right\};$$

$$\hat{\Pi} := \left\{ \pi \,\middle|\, \pi(a|s) \geq 0, \ \sum_{\hat{a} \in \mathcal{A}} \pi(\hat{a}|s) = 1, \ a \in \mathcal{A}, s \in \mathcal{S} \right\}.$$

DMDP. Задача RL

Следующее утверждение [1] (детали: глава 3, утв. 21 и предшествующие) задаёт подкласс оптимальных политик, в рамках которого достаточно производить поиск интересующей π.

Пусть Π – набор всех нестационарных и рандомизированных политик. $V^\pi(s)$, $Q^\pi(s,a)$ зажаты между 0 и $\frac{1}{1-\gamma}$, следовательно, существуют конечные

$$V^*(s) := \sup_{\pi \in \Pi} \left\{ V^\pi(s) \right\}, \qquad Q^*(s,a) := \sup_{\pi \in \Pi} \left\{ Q^\pi(s,a) \right\};$$

$\exists\ \pi$ – стационарная, детерминированная, такая, что $\forall s \in \mathcal{S},\ a \in \mathcal{A}:$

$$V^\pi(s) = V^*(s), \qquad Q^\pi(s,a) = Q^*(s,a),$$

а, значит, π – оптимальная политика.

DMDP. Задача RL

В данном утверждении мы можем легко заменить операцию sup на операцию max, как минимум, в случае наград с достижимыми верхними гранями:

$$V^*(s) = \max_{\pi \in \Pi} \left\{ V^\pi(s) \right\},$$

где V^* – оптимальная функция ценности. Введём обозначение класса всех отображений, описывающих детерминированные политики в данном процессе:

$$\mathbb{A} = \left\{ a(\cdot) \middle| \; a : \mathcal{S} \mapsto \mathcal{A} \right\}.$$

DMDP. Уравнения Беллмана

Если воспользоваться принципом динамического программирования, то удаётся вывести уравнение оптимальности Беллмана на V-функцию ценности:

$$
\begin{aligned}
V^*(s) &= \max_{a(\cdot)\in\mathbb{A}} \left\{ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a(s_t)) \right] \right\} = \\
&= \max_{a(\cdot)\in\mathbb{A}} \left\{ \mathbb{E}\left[r(s, a(s)) + \gamma \sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a(s_{t+1})) \right] \right\} = \\
&= \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}\left[V^*(s') \right] \right\} = \\
&= \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s,a;s') V^*(s') \right\}.
\end{aligned}
$$

DMDP. Уравнения Беллмана

Соответственно, мы можем провести аналогичные рассуждения для Q-функции ценности:

$$
\begin{aligned}
Q^*(s,a) &= \max_{\pi \in \Pi} \left\{ Q^\pi(s,a) \right\}; \\
Q^*(s,a) &= \mathbb{E}\left[r(s,a) + \gamma V^*(s') \middle| s_0 = s, a_0 = a \right] = \\
&= r(s,a) + \gamma \mathbb{E}\left[\max_{a' \in \mathcal{A}} \left\{ Q^*(s',a') \right\} \right] = \\
&= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') \max_{a' \in \mathcal{A}} \left\{ Q^*(s',a') \right\}.
\end{aligned}
$$

Критерий оптимальности относительно Q-функции ценности [1] (детали: глава 3.1.10.).

Функция Q представляет собой оптимальную функцию ценности Q^*, если и только если она удовлетворяет уравнениям оптимальности Беллмана:

$$Q(s,a) = \mathbb{E}\left[r(s,a(s)) + \gamma \max_{a' \in \mathcal{A}} \left\{ Q(s',a') \right\} \middle| s_0 = s, a_0 = a \right] =$$
$$= \sum_{s' \in \mathcal{S}} p(s,a;s') \left[r(s,a) + \gamma \max_{a' \in \mathcal{A}} \left\{ Q(s',a') \right\} \right], \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}.$$

Кроме того, детерминированная политика, определенная как

$$\pi(s) \in \operatorname*{Arg\,max}_{a \in \mathcal{A}} \left\{ Q^*(s,a) \right\},$$

есть оптимальная политика.

DMDP. Уравнения оптимальности Беллмана

Таким образом, для оптимальной политики π^* выполнены следующие соотношения [2]:

1) $V^{\pi^*}(s) = \max\limits_{a \in \mathcal{A}} \left\{ Q^{\pi^*}(s,a) \right\}, \quad \forall s \in \mathcal{S};$
2) $Q^{\pi^*}(s,a) = r(s,a) + \gamma \sum\limits_{s' \in \mathcal{S}} p(s,a;s') V^{\pi^*}(s'), \quad \forall s \in \mathcal{S}, a \in \mathcal{A};$
3) $V^{\pi^*}(s) = \max\limits_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum\limits_{s' \in \mathcal{S}} p(s,a;s') V^{\pi^*}(s') \right\}, \quad \forall s \in \mathcal{S};$
4) $Q^{\pi^*}(s,a) = r(s,a) + \gamma \sum\limits_{s' \in \mathcal{S}} p(s,a;s') \max\limits_{a' \in \mathcal{A}} \left\{ Q^{\pi^*}(s',a') \right\}, \quad s \in \mathcal{S}, a \in \mathcal{A}.$

Соответствующая этим соотношениям детерминированная политика:

$$\pi^*(s) \in \operatorname*{Arg\,max}_{a \in \mathcal{A}} \left\{ Q^{\pi^*}(s,a) \right\} = \operatorname*{Arg\,max}_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') V^{\pi^*}(s') \right\}.$$

DMDP. Уравнения оптимальности Беллмана

Основное свойство оптимальных V/Q-функций

Если V^*, Q^* – оптимальные функции, то функции $V^\star(s) := V^*(s) + \alpha, Q^\star(s,a) := Q^*(s,a) + \alpha,\ \forall s \in \mathcal{S}, a \in \mathcal{A}, \alpha \in \mathbb{R}$ так же соответствуют оптимальной политике.

$$\pi^*(s) \in \operatorname*{Arg\,max}_{a \in \mathcal{A}} \{Q^*(s,a)\} = \operatorname*{Arg\,max}_{a \in \mathcal{A}} \{Q^*(s,a) + \alpha\} =$$

$$= \operatorname*{Arg\,max}_{a \in \mathcal{A}} \{Q^\star(s,a)\} = \operatorname*{Arg\,max}_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') V^*(s') \right\} =$$

$$= \operatorname*{Arg\,max}_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') \underbrace{\left(V^*(s') + \alpha\right)}_{=V^\star(s')} \right\},\ \forall s \in \mathcal{S}, \alpha \in \mathbb{R}$$

DMDP. Уравнения оптимальности Беллмана

Для уравнения Беллмана верна связь с константным сдвигом функции награды $\hat{r}(s,a) := r(s,a) + \alpha(1-\gamma)$, $\forall s \in \mathcal{S}, a \in \mathcal{A}, \alpha \in \mathbb{R}$:

$$
\begin{aligned}
Q^\star(s,a) = Q^*(s,a) + \alpha &= \left(r(s,a) + \alpha(1-\gamma)\right) + \\
&+ \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') \max_{a' \in \mathcal{A}} \left\{ Q^*(s',a') + \alpha \right\} = \\
&= \hat{r}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') \max_{a' \in \mathcal{A}} \left\{ Q^\star(s',a') \right\} = \\
&= \hat{r}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') V^\star(s');
\end{aligned}
$$

$$
\begin{aligned}
V^\star(s) &= \max_{a \in \mathcal{A}} \left\{ \left(r(s,a) + \alpha(1-\gamma)\right) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') \left(V^*(s') + \alpha \right) \right\} = \\
&= \max_{a \in \mathcal{A}} \left\{ \hat{r}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') V^\star(s') \right\} = \max_{a \in \mathcal{A}} \left\{ Q^\star(s,a) \right\}.
\end{aligned}
$$

DMDP. Уравнения оптимальности Беллмана

Аддитивное преобразование, не меняющее решение π^*

[a] $\hat{r}(s,a) := \beta\left(r(s,a) + f(s) - \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') f(s')\right)$, $\forall s \in \mathcal{S}, a \in \mathcal{A}, \beta > 0$, где $f : \mathcal{S} \mapsto \mathbb{R}$ – произвольное отображение.

[a] $\pi^*(s) = \arg\max_{a \in \mathcal{A}} \{Q^*(s,a)\} = \arg\max_{a \in \mathcal{A}} \{\beta\left(Q^*(s,a) + f(s)\right)\}, \ s \in \mathcal{S}.$

$$
\begin{aligned}
Q^\star(s,a) &:= \beta\left(Q^*(s,a) + f(s)\right) = \\
&= \beta\left(r(s,a) + f(s)\right) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') \max_{a' \in \mathcal{A}} \{\beta Q^*(s',a')\} = \\
&= \hat{r}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') \max_{a' \in \mathcal{A}} \{\beta\left(Q^*(s',a') + f(s')\right)\} = \\
&= \hat{r}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') \max_{a' \in \mathcal{A}} \{Q^\star(s',a')\}.
\end{aligned}
$$

DMDP. Уравнения оптимальности Беллмана

Для функции награды, зависящей от $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, аддитивное преобразование, инвариантное относительно оптимальной политики, выглядит проще: $\hat{r}(s, a, s') := \beta\left(r(s, a, s') + f(s) - \gamma f(s')\right)$.

$$
\begin{aligned}
Q^{\star}(s, a) := \beta\left(Q^{*}(s, a) + f(s)\right) = \\
= \sum_{s' \in \mathcal{S}} p(s, a; s')\left(\beta\left(r(s, a, s') + f(s) - \gamma f(s')\right) + \right. \\
\left. + \gamma \max_{a' \in \mathcal{A}}\left\{\beta\left(Q^{*}(s', a') + f(s')\right)\right\}\right) = \\
= \sum_{s' \in \mathcal{S}} p(s, a; s')\left(\hat{r}(s, a, s') + \gamma \max_{a' \in \mathcal{A}}\left\{Q^{\star}(s', a')\right\}\right).
\end{aligned}
$$

Предположения

Будем считать, что о нашей среде нам известны следующие функции:

- $p(s'|s,a)$ — вероятность попасть в состояние s' из состояния s с помощью действия a
- $r(s,a)$ — награда за выполнения действия a в состоянии s

Замечание

Эти допущения существенно сужают круг задач, которые мы можем решить, однако алгоритмы, предложенные в этой лекции будут оптимальными в данной постановке.

Построение алгоритма обучения π

Сам процесс можно разделить на два этапа: оценка качества текущей политики и поиск следующего приближения оптимальной политики. *Дополнительно предположим дискретность и конечность пространств* $\mathcal{S}$ *и* $\mathcal{A}$. Начнём с оценивания — вычислим $V^\pi(s)$ и $Q^\pi(s,a)$:

$$V^\pi(s) = \underbrace{\underbrace{\mathbb{E}_{\pi(a|s)}\left[r(s,a)\right]}_{:=u(s)} + \gamma \mathbb{E}_{\pi(a|s)} \mathbb{E}_{p(s'|s,a)}\left[V^\pi(s')\right]}_{:=F(V^\pi(s))}, \quad \forall s \in S.$$

Представим выражение выше в виде матрично векторных операций. V^π — вектор ценностей состояний. P — матрица вероятностей: $P_{ss'} = \sum\limits_{a \in \mathcal{A}} \pi(a|s) p(s'|s,a) = p(s'|s)$. u – вектор средних наград за шаг.

Построение алгоритма обучения π

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a) V^\pi(s') =$$
$$= u(s) + \gamma \sum_{s'\in\mathcal{S}} V^\pi(s') p(s'|s) \Longrightarrow V^\pi = F(V^\pi) = u + \gamma P V^\pi.$$

Полученное отображение F является сжимающим, для произвольных векторов V и W:

$$\|F(V) - F(W)\|_\infty = \|u + \gamma PV - u - \gamma PW\|_\infty = \gamma \|P(V - W)\|_\infty \leq$$
$$\leq \gamma \|P\|_\infty \|V - W\|_\infty = \gamma \|V - W\|_\infty,$$

$$\|P\|_\infty = \max_{x\neq 0} \left\{ \frac{\|Px\|_\infty}{\|x\|_\infty} \right\} = \max_{\|x\|_\infty \leq 1} \max_{s\in\mathcal{S}} \left\{ \sum_{s'\in\mathcal{S}} p(s'|s) \underbrace{x_{s'}}_{=1} \right\} = 1.$$

Построение алгоритма обучения π

Получили алгортим Iterative Policy Evaluation для оценки политики π с заданной точностью $\varepsilon > 0$:

1) Инициализировать $V(s), \forall s \in \mathcal{S}$;
2) Повторять в цикле:

2.1) $\Delta := 0$

2.2) для всех $s \in \mathcal{S}$:

$$\delta := V(s);$$
$$V(s) := \mathbb{E}_{\pi(a|s)}\left[r(s,a)\right] + \gamma \mathbb{E}_{\pi(a|s)} \mathbb{E}_{p(s'|s,a)}\left[V(s')\right];$$
$$\Delta := \max\{\Delta, |\delta - V(s)|\}.$$

2.3) Если $\Delta \le \varepsilon$, то выход, иначе — переход на шаг 2.

Построение алгоритма обучения π

Теперь построим процедуру улучшения оценённой политики π:

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)}\left[V^{\pi}(s')\right].$$

Определение

Политика $\hat{\pi} \succeq \pi$ (монотонно лучше политики π), если $V^{\hat{\pi}}(s) \geq V^{\pi}(s), \quad \forall s \in \mathcal{S}$.

Таким образом, изменяя действие для одного состояния $\hat{s} \in \mathcal{S}$ мы производим улучшение π:

$$\hat{\pi}(a|\hat{s}) = \delta_{\left\{\arg\max\limits_{\hat{a}\in\mathcal{A}}\{Q^{\pi}(\hat{s},\hat{a})\}\right\}}(a),\ \hat{\pi}(\hat{s}) := \arg\max_{\hat{a}\in\mathcal{A}}\left\{Q^{\pi}(\hat{s},\hat{a})\right\};$$

$$\hat{\pi}(a|s) = \pi(a|s),\ \forall s \in \mathcal{S},\ s \neq \hat{s}.$$

То есть $Q^{\pi}(s,\hat{\pi}(s)) \geq V^{\pi}(s)$.

Построение алгоритма обучения π

Теорема об улучшении политики [1] (детали: глава 3.2.3, теорема 17)

Пусть π и π' – любая пара детерминированных политик, таких, что

$$\forall s \in \mathcal{S} \quad Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s). \tag{1}$$

Тогда π' должна быть не хуже, чем π, то есть ценность не хуже $\forall s \in \mathcal{S}$:

$$V^{\pi'}(s) \geq V^{\pi}(s). \tag{2}$$

Более того, если в каком-либо состоянии существует строгое (1), то и (2) должно быть строгим.

Построение алгоритма обучения π

Действительно,

$$V^{\pi}(s) \le Q^{\pi}(s, \underbrace{\hat{\pi}(s)}_{\text{детерминирован}}) = r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{p(s'|s,a)} \left[V^{\pi}(s') \right] \le$$

$$\le r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{p(s'|s,a)} \left[Q^{\pi}(s', \hat{\pi}(s')) \right] \le \ldots \le$$

$$\le r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{p(s'|s,\hat{\pi}(s))} \left[r(s', \hat{\pi}(s')) + \gamma^2 \ldots \right] = V^{\hat{\pi}}(s), \quad \forall s \in \mathcal{S}.$$

Шаг для обновления всей политики

$$\pi_{\text{new}}(s) := \arg\max_{a \in \mathcal{A}} \left\{ Q^{\pi_{\text{old}}}(s, a) \right\}, \quad \forall s \in \mathcal{S}.$$

Построение алгоритма обучения π

Если после очередного шага обновления политики получилось, что $Q^\pi(s, \hat{\pi}(s)) = V^\pi(s),\ \forall s \in \mathcal{S}$, то это означает удовлетворение уравнению Беллмана:

$$\hat{\pi} = \pi;$$
$$Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{p(s'|s,a)} \left[V^\pi(s') \right], \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}.$$

Построение алгоритма обучения π

Мы получили алгоритм Policy Iteration:

1) Инициализируем $V(s)$ и $\pi(a|s)$ для всех $s \in \mathcal{S}$.
2) Оценить $V^\pi(s)$ для текущей π, используя Iterative Policy Evaluation.
3) Улучшаем политику:

3.1) Flag := True;

3.2) Для всех $s \in \mathcal{S}$:

$$
\begin{aligned}
& a = \pi(s); \\
& \pi(s) := \arg\max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \mathbb{E}_{p(s'|s,a)} \left[V^\pi(s') \right] \right\}; \\
& \text{если } a \neq \pi(s), \text{ то } \text{Flag} := \text{False}.
\end{aligned}
$$

4) Если Flag $=$ True, то выход, иначе — шаг 2.

Построение алгоритма обучения π

Описанную ранее процедуру поиска оптимальной политики можно представить как последовательность монотонно улучшающихся политик и функций ценности:

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{I} \ldots \xrightarrow{I} \pi^* \xrightarrow{E} V^*,$$

где $\xrightarrow{E}$ обозначает оценку политики, $\xrightarrow{I}$ – улучшение политики, то есть получен алгоритм итеративной оптимизации политики.

Построение алгоритма обучения π

Уравнение Беллмана относительно фиксированной политики по сути решается простым итеративным способом. Начальное приближение V_0 выбирается произвольно, а каждая последовательная итерация реализуется согласно уравнению Беллмана:

$$V_{t+1}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} p(s, a; s') \left(r(s, a) + \gamma V_t(s') \right),$$

где $V_t = V^\pi$ – фиксированная точка. Для получения каждого последующего приближения, V_{t+1} из V_t при итеративной оценке политики применяется та же операция к каждому состоянию s, и ее называют ожидаемым обновлением, а данный алгоритм – итерации функции ценности.

Построение алгоритма обучения π

Для решения уравнения оптимальности Беллмана можно проводить следующую процедуру:

$$V_{t+1}(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s, a; s') V_t(s') \right\}, \quad \forall s \in \mathcal{S};$$

$$\pi_{t+1}(s) \in \operatorname*{Arg\,max}_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s, a; s') V_{t+1}(s') \right\}, \quad \forall s \in \mathcal{S}.$$

Причём начальное значение V_0 может быть произвольным, а в качестве критерия останова может выступать $\| V_{t+1} - V_t \|_\infty \leq \varepsilon$.

Построение алгоритма обучения π

На предыдущем слайде предложен по сути частный случай алгоритма Policy Iteration — Value Iteration (вместо шагов 2 и 3 один шаг делаем):

1) Инициализируем $V(s)$ для всех $s \in \mathcal{S}$.
2) Повторять:
2.1) $\Delta := 0$;
2.2) Для всех $s \in \mathcal{S}$:

$$
\begin{aligned}
& v = V(s) \\
& V(s) := \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \left[V(s') \right] \right\} \\
& \Delta := \max\{\Delta, |v - V(s)|\}
\end{aligned}
$$

2.3) Если $\Delta \leq \varepsilon$, то выход, иначе — переход на шаг 2.

Мы по сути схлопнули Policy Iteration и Policy Improvement, не вычисляя π.

Построение алгоритма обучения π

В результате работы алгоритма Value Iteration оптимизированная политика вычисляется следующим образом:

$$\pi(s) = \arg\max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \mathbb{E}_{p(s'|s,a)} \left[V(s') \right] \right\}.$$

Сходимость алгоритма Value Iteration

Для оптимальной V-функции выполнено соотношение
$V^*(s) = V^{\pi^*}(s) = \max\limits_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \left[V^*(s') \right] \right\}, \ \forall s \in \mathcal{S}.$
Рассмотрим
$V^{\pi_{t+1}}(s), \ \pi_{t+1}(s) = \arg\max\limits_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \left[V^{\pi_{t+1}}(s') \right] \right\}.$
Построим оценку на невязку по V-функции,
$V^*(s) \geq V^{\pi_t}(s), \ t \in \mathbb{Z}_+, \ s \in \mathcal{S}$:

$$\max_{s\in\mathcal{S}} \left\{ \left| V^*(s) - V^{\pi_{t+1}}(s) \right| \right\} = \max_{s\in\mathcal{S}} \left\{ \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \left[V^*(s') \right] \right\} - \right.$$

$$\left. - \max_{a'\in\mathcal{A}} \left\{ r(s,a') + \gamma \mathbb{E}_{p(s''|s,a')} \left[V^{\pi_t}(s'') \right] \right\} \right\} \leq \gamma \max_{\substack{s \in \mathcal{S} \\ a \in \mathcal{A}}} \left\{ \mathbb{E}_{p(s'|s,a)} \left[V^*(s') \right] - \right.$$

$$\left. - \mathbb{E}_{p(s''|s,a)} \left[V^{\pi_t}(s'') \right] \right\} = \gamma \max_{s\in\mathcal{S}, a\in\mathcal{A}} \left\{ \mathbb{E}_{p(s'|s,a)} \left[V^*(s') - V^{\pi_t}(s') \right] \right\} \leq$$

Сходимость алгоритма Value Iteration

продолжим с последнего неравенства, раскрывая рекуррентную зависимость:

$$\leq \gamma \max_{s' \in \mathcal{S}} \left\{ V^*(s') - V^{\pi_t}(s') \right\} = \gamma \max_{s' \in \mathcal{S}} \left\{ \left| V^*(s') - V^{\pi_t}(s') \right| \right\} \Rightarrow$$

$$\Rightarrow \max_{s \in \mathcal{S}} \left\{ \left| V^{\pi_t}(s) - V^*(s) \right| \right\} \leq \gamma^t \max_{s' \in \mathcal{S}} \left\{ \left| V^{\pi_0}(s') - V^*(s') \right| \right\}.$$

В качестве $V^{\pi_0}(s),\ s \in \mathcal{S}$ можно взять произвольное начальное приближение, например, тождественную по всем состояниям константу. Имеем оценку относительно внешних итераций:

$$\max_{s \in \mathcal{S}} \left\{ \left| V^{\pi_t}(s) - V^*(s) \right| \right\} = \mathcal{O}\left(\gamma^t\right), \quad t \in \mathbb{Z}_+,\ \gamma \in (0, 1).$$

Value Iteration. Сложность решения

Введём следующий оператор $T : \mathbb{R}^S \mapsto \mathbb{R}^S$:
$T(V)_s := \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') V(s') \right\}$. Теперь обновление V-функции в Value Iteration выражается следующим образом: $V_{t+1} := T(V_t)$, а V-функция кодируется вещественным вектором.

Теорема [3]

Существует DMDP, для которого последовательность оценок V-функции удовлетворяет:
$V_0 = 0_S, V_{n+1} \in \operatorname{span}\{V_0, \ldots, V_n, T(V_0), \ldots, T(V_n)\}$, $n \in \mathbb{Z}_+$, со следующим свойством $\forall n = \overline{0, N-1}$:

$$\|V_n - V^*\|_\infty \geq \frac{\gamma^n}{1+\gamma}, \; V^* = T(V^*).$$

Value Iteration. Сложность решения

Доказательство. Для произвольного $\gamma \in (0,1)$ предложим DMDP с N состояниями и с одним действием. Награда для первого состояния $r_1 := 1$, для остальных состояний $r_i := 0, i = \overline{2, N}$. Действие из первого состояния оставляет в нём же, действие из $(i+1)$-го состояния переводит в i-ое. Оптимальное значение V-функции следующее: $V^*(i) = \frac{\gamma^{i-1}}{1-\gamma}$. Рассмотрим последовательность векторов $(V_n)_{n \geq 0}, V_0 = 0_S$ и

$$V_{n+1} \in \operatorname{span}\left\{V_0, \ldots, V_n, T(V_0), \ldots, T(V_n)\right\}, n \geq 0.$$

Докажем через раскрытие рекурсии, что $\forall n \geq 0, i \in \mathcal{S}$ имеем $V_n(i) = 0$, если $i \geq n+1$. Это верно для $n = 0$, так как $V_0 = 0_S$. Предположим, что верно и для $V_0, \ldots V_{n-1}$. По определению оператора T и в силу того, что $r_i = 0, i \geq 2$, имеем $T(V_t)_i = 0$, если $i \geq t+2, \forall t \leq n-1$.

Value Iteration. Сложность решения

Следовательно, в силу $V_{n+1} \in \text{span}\{V_0, \ldots, V_n, T(V_0), \ldots, T(V_n)\}$ заметим, что $V_n(i) = 0$, если $i \geq n+1$, и мы доказали нашу рекурсию. $r_1 > 0$ – единственное, по сути мы доказали, что для любого метода первого порядка требуется $n-1$ шаг для распространения награды первого состояния до состояния $1 \leq n \leq N$.

Value Iteration. Сложность решения

Теперь мы имеем для $1 \leq n \leq N-1$:

$$
\begin{aligned}
\|\mathrm{V}_n - T(\mathrm{V}_n)\|_\infty &= \|\mathrm{V}_n - T(\mathrm{V}_n) - (\mathrm{V}^* - T(\mathrm{V}^*))\|_\infty && (3)\\
&\geq (1-\gamma)\|\mathrm{V}_n - \mathrm{V}^*\|_\infty && (4)\\
&= (1-\gamma) \max_{1 \leq i \leq N} \{|V_n(i) - V^*(i)|\} \\
&\geq (1-\gamma) \max_{n+1 \leq i \leq N} \{|V_n(i) - V^*(i)|\} \\
&\geq (1-\gamma) \max_{n+1 \leq i \leq N} \{|V^*(i)|\} && (5)\\
&\geq (1-\gamma) \max_{n+1 \leq i \leq N} \left\{\frac{\gamma^{i-1}}{1-\gamma}\right\} \\
&\geq \gamma^n,
\end{aligned}
$$

где (3) следует из $\mathrm{V}^* = T(\mathrm{V}^*)$,

Value Iteration. Сложность решения

(4) следует из (6), и (5) следует из $V_n(i) = 0$ для $i \geq n+1$. Можем заключить в силу (6):

$$\|\mathrm{V}_n - \mathrm{V}^*\|_\infty \geq \frac{1}{1+\gamma} \cdot \|\mathrm{V}_n - T(\mathrm{V}_n) - (\mathrm{V}^* - T(\mathrm{V}^*))\|_\infty = \\ = \frac{1}{1+\gamma} \cdot \|\mathrm{V}_n - T(\mathrm{V}_n)\|_\infty \geq \frac{\gamma^n}{1+\gamma}.$$

$$\forall V, W \in \mathbb{R}^S \text{ (проверяется непосредственно)} : \\ (1-\gamma) \cdot \|\mathrm{V} - \mathrm{W}\|_\infty \leq \|(\mathrm{I} - T)(\mathrm{V}) - (\mathrm{I} - T)(\mathrm{W})\|_\infty; \\ \|(\mathrm{I} - T)(\mathrm{V}) - (\mathrm{I} - T)(\mathrm{W})\|_\infty \leq (1+\gamma) \cdot \|\mathrm{V} - \mathrm{W}\|_\infty. \tag{6}$$

Утверждение доказано.

Value Iteration. Сложность решения

Из доказанного утверждения вместе с утверждением о сходимости Value Iteration следует оптимальность алгоритма Value Iteration:

$$\frac{\gamma^t}{1-\gamma} = \gamma^t \|V_0 - V^*\|_\infty \geq \|V_t - V^*\|_\infty \geq \frac{\gamma^t}{1+\gamma}, \; t \in \mathbb{Z}_+, \; \gamma \in (0,1);$$

$$\|V_t - V^*\|_\infty = \Theta\left(\gamma^t\right), \quad V_0 = 0_N.$$

LP-релаксация MDP. AMDP [4]

Для AMDP LP-задача вводится следующим образом:

$$V^*(s) = \max_{a(\cdot)\in\mathbb{A}} \left\{ \lim_{H\to\infty} \frac{1}{H} \mathbb{E} \left[\sum_{t=0}^{H-1} r(s_t, a_t(s_t)) \middle| s_0 = s \right] \right\},$$

где H – эпизодическое ограничение, то есть максимальная длина эпизода. В случае эпизодов конечной длины предел опускается и используется максимальное значение H. Для политики $\pi(a|s)$ можно определить стационарное распределение:

$$\nu_\pi(s') = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} p(s,a;s')\pi(a|s)\nu_\pi(s), \quad s' \in \mathcal{S},$$

которое соответствует своему вектору из вероятностей $\nu_\pi = (\nu_\pi(s))_{s\in\mathcal{S}}$.

LP-релаксация MDP. AMDP [4]

И если MDP равномерно эргодично, то:

$$V(\pi) := \lim_{H\to\infty} \frac{1}{H}\mathbb{E}\left[\sum_{t=0}^{H-1} r(s_t, a_t(s_t))\right] = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} r(s,a)\pi(a|s)\nu_\pi(s).$$

Напоминаем, что в данном случае равномерная эргодичность соответствует:

$$\max_{i=\overline{1,S}} \left\{\left\|\left(P^{\pi}\right)^n e_i - \nu_\pi\right\|_\infty\right\} \to 0,\ n\to\infty, \quad e_i^\top = (0,\ldots,0,\underbrace{1}_{i},0,\ldots,0).$$

LP-релаксация MDP. AMDP [4]

Вводится распределение действий по состояниям – $\mu(s,a) = \nu_\pi(s)\pi(a|s)$, следовательно, можно переписать задачу поиска оптимальной политики в AMDP как задачу LP со смыслом оценки ценности политики по распределению μ:

$$\max_{\mu \in \Delta^{\mathcal{S}\times\mathcal{A}}} \left[V(\mu) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} r(s,a)\mu(s,a) = \langle r, \mu \rangle : \sum_{b\in\mathcal{A}} \mu(s',b) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} p(s,a;s')\mu(s,a),\ s' \in \mathcal{S} \right];$$

$$\Delta^{\mathcal{S}\times\mathcal{A}} = \left\{ \mu : \mu(s,a) \geq 0, \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu(s,a) = 1 \right\},\ \pi_\mu(a|s) = \frac{\mu(s,a)}{\sum\limits_{b\in\mathcal{A}} \mu(s,b)}.$$

LP-релаксация MDP. AMDP [4]

Данную задачу можно напрямую переписать в матричной форме:

$$\max_{\mu \in \Delta^{\mathcal{S}\times\mathcal{A}}} \langle r, \mu \rangle ;$$
$$s.t.\ (\widehat{I} - P)\mu = 0.$$

Единичная матрица $\widehat{I}$ имеет нестандартный формат: это прямоугольная матрица размера $S \times (SA)$, на каждой строке $s \in \mathcal{S}$ только элементы, соответствующие паре (s, a), $a \in \mathcal{A}$, равняются единице, остальные элементы данной строки равняются нулю, то есть на каждой строке $\widehat{I}$ ровно A единиц. У матрицы P размера $S \times (SA)$ в каждом столбце $(s, a) \in \mathcal{S} \times \mathcal{A}$ записано распределение $\mathrm{P}(\cdot|s, a)$.

LP-релаксация MDP. AMDP [4]

Для этой задачи LP напрямую строится двойственная задача, с условием, что $\mu \geq 0$, которая имеет смысл оценки ценности оптимальной политики через V-функцию:

$$\min_{\overline{V} \in \mathbb{R}, V \in \mathbb{R}^{|\mathcal{S}|}} \overline{V};$$
$$s.t.\ r - \overline{V} \cdot 1_{SA} - (\widehat{I} - P)^\top V \leq 0.$$

Таким образом, имеет место уравнение оптимальности Беллмана со средним вознаграждением:

$$V(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) - V^* + \sum_{s' \in \mathcal{S}} p(s,a;s') V(s') \right\}, \quad V^* = \langle r, \mu^* \rangle,$$

полученное из ограничений вида неравенства:

$$\widehat{I}^\top V \geq r - \overline{V} \cdot 1_{SA} + P^\top V.$$

LP-релаксация MDP. DMDP [4]

Для DMDP задача LP записывается в следующем виде (q – распределение начального состояния μ_0 в виде вектора):

$$\min_{V \in \mathbb{R}^{|\mathcal{S}|}} \langle q, V \rangle ;$$
$$s.t.\ r - (\widehat{I} - \gamma P)^\top V \leq 0.$$

И ей соответствует такая двойственная задача:

$$\max_{\mu \in \Delta^{\mathcal{S} \times \mathcal{A}}} \langle r, \mu \rangle ;$$
$$s.t.\ (\widehat{I} - \gamma P)\mu = q.$$

LP-релаксация MDP. DMDP [5]

Существует также постановка задачи LP для ограниченного DMDP – Constrained Markov Decision Process, CMDP:

$$\max_{\mu \in \Delta^{\mathcal{S} \times \mathcal{A}}} \langle r, \mu \rangle \,;$$
$$s.t.\ (\widehat{I} - \gamma P)\mu = q,\ D\mu \geq c.$$

По сравнению с предыдущими задачами линейного программирования вводится дополнительно аффинное ограничение вида неравенства: $D\mu \geq c$.

LP-релаксация MDP. DMDP [5]

Заметим, что вместо обозначенного ранее распределения $\mu(s, a) = \nu_\pi(s)\pi(a|s)$ может быть полезно рассмотреть:

$$\mu(s, a) := \mu^\pi(s, a) = \mathbb{E}_{s_0 \sim \mu_0}\left[\sum_{t=0}^{\infty} \gamma^t \, \mathrm{P}(s_t = s, a_t = a | s_0)\right].$$

Или даже масштабированную сумму сверху в виде корректно определённой вероятностной меры:

$$\mu(s, a) := \tilde{\mu}^\pi(s, a) = (1 - \gamma)\mu^\pi(s, a) = \\ = \mathbb{E}_{s_0 \sim \mu_0}\left[(1 - \gamma)\sum_{t=0}^{\infty} \gamma^t \, \mathrm{P}(s_t = s, a_t = a | s_0)\right].$$

LP-релаксация MDP. DMDP [5]

В обоих случаях получается одна и та же политика:

$$\pi(a|s) = \frac{\mu^\pi(s,a)}{\sum\limits_{b \in \mathcal{A}} \mu^\pi(s,b)} = \frac{\tilde{\mu}^\pi(s,a)}{\sum\limits_{b \in \mathcal{A}} \tilde{\mu}^\pi(s,b)}.$$

DMDP. Пример

Задача о разборчивой невесте

В некотором царстве, в некотором государстве пришло время принцессе выбирать себе жениха. В назначенный день явились 1000 (достаточно большое количество) царевичей (расстановки претендентов равновероятны). Их построили в очередь в случайном порядке и стали по одному приглашать к принцессе. Про любых двух претендентов принцесса, познакомившись с ними, может сказать, какой из них лучше. Познакомившись с претендентом, принцесса может либо принять предложение (и тогда выбор сделан навсегда), либо отвергнуть его (и тогда претендент потерян: царевичи гордые и не возвращаются). Какой стратегии должна придерживаться принцесса, чтобы с наибольшей вероятностью выбрать лучшего?

DMDP. Пример

Оптимальная стратегия невесты: Пропустить первых $1/e$ ($e \simeq 2.718$) претендентов и затем выбрать первого наилучшего (среди пропущенных мог быть самый лучший – в таком случае, никого лучше невеста уже не встретит). Такая стратегия позволяет невесте выбрать наилучшего жениха с вероятностью $1/e$.
Введем управляемую марковскую систему с $\gamma = 1$, $\mathcal{S} = \{1, 2, ..., N, End\}$, $N = 1000$. Состоянию s соответствует s-й претендент, оказавшийся наилучшим на данный момент.
Определим множество стратегий. Возможны всего два действия $\mathcal{A} =$ (не выбрать, выбрать). «Фиктивное» состояние *End* наступает на следующем шаге, после того, как невеста пропустила лучшего жениха, или когда невеста сделала свой выбор. Исходя из этого, можно посчитать соответствующие функции вознаграждения и вероятности переходов (P – вероятность):

DMDP. Пример

$$r(s = End,\ a_1 = \text{выбрать}) = 0,\ \ r(s,\ a_2 = \text{не выбрать}) = 0;$$

$$r(s,\ a_1) = s/N,\ \ s = 1, \ldots, N,$$

поскольку

$$r_\xi(s, a_1) = \begin{cases} 1, & \text{с вероятностью } s/N, \\ 0, & \text{с вероятностью } 1 - s/N. \end{cases}$$

DMDP. Пример

С переходными вероятностями немного сложнее:

$$p(s,\, a_1;\, s') = 0,\ s' \neq End, \qquad p(s,\, a_1;\, s' = End) = 1;$$

$$p(s = End,\, a;\, s' = End) = 1;$$

$$p(s,\, a_2;\, s' = End) = \mathrm{P}\left(\begin{array}{c} s\text{-й лучший, если известно,} \\ \text{что он лучше предыдущих} \end{array}\right) = \frac{s}{N};$$

$$p(s,\, a_2;\, s') = \mathrm{P}\left(\begin{array}{c} s'\text{-й – первый кто, лучше} \\ s\text{-го, если } s\text{-й был лучшим} \end{array}\right) =$$

$$= \frac{\mathrm{P}\left(\begin{array}{c} s\text{-й лучше предыдущих и} \\ s'\text{-й – первый кто, лучше } s\text{-го} \end{array}\right)}{\mathrm{P}(s\text{-й лучше предыдущих})} = \frac{s}{s'(s'-1)},$$

DMDP. Пример

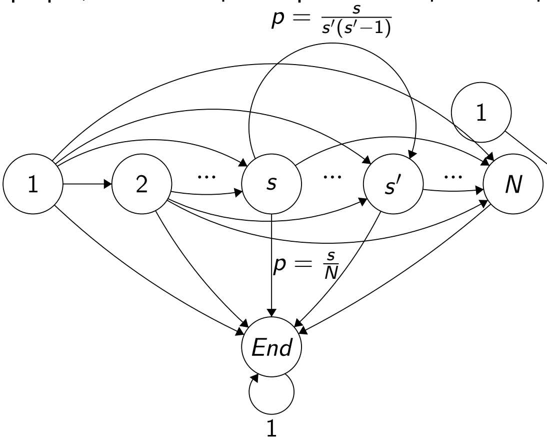
поскольку

$$\mathrm{P}(s\text{-й лучше предыдущих}) = \frac{(s-1)!}{s!} = \frac{1}{s},$$

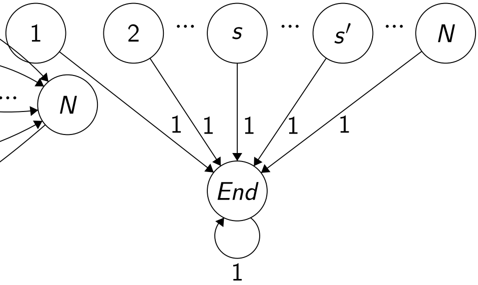
$$\mathrm{P}\left(\begin{array}{c} s\text{-й лучше предыдущих и} \\ s'\text{-й – первый кто, лучше } s\text{-го} \end{array}\right) = \frac{(s'-2)!}{s'!} = \frac{1}{s'(s'-1)}.$$

DMDP. Пример

Графы, отвечающие марковской цепи в задаче о разборчивой невесте:



a = не выбрать



a = выбрать

DMDP. Пример

Функция $V^*(s)$ удовлетворяет уравнению Беллмана, в данном случае получившееся выражение называют ещё уравнением Вальда—Беллмана, а оптимальная стратегия может быть найдена из условия:

$$a(s) = \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s,a;s') V^*(s') \right\} =>$$

$$V^*(s) = \max \left\{ \frac{s}{N}; \sum_{s'=s+1}^{N} \frac{s}{s'(s'-1)} V^*(s') \right\}, \; s = 1, \ldots, N-1; \;\; V^*(N) = 1.$$

DMDP. Пример

Если максимум в $V^*(s)$ достигается на первом аргументе, то $a(s) =$ выбрать, если на втором, то $a(s) =$ не выбрать. $s^*(N) \simeq \left[\frac{N}{e}\right]$:

$$\frac{1}{s^*} + \frac{1}{s^*+1} + \dots + \frac{1}{N-1} \leq 1 \leq \frac{1}{s^*-1} + \frac{1}{s^*} + \dots + \frac{1}{N-1}.$$

Подставим $s^*(N)$ в уравнение на функцию значений:

$$V^* = \frac{s^*(N)-1}{N}\left(\frac{1}{s^*(N)-1} + \frac{1}{s^*(N)} + \dots + \frac{1}{N-1}\right) \simeq \frac{1}{e},$$

$$=> V^*(s) = \begin{cases} V^*, 1 \leq s \leq s^*(N), \\ s/N, s \geq s^*(N). \end{cases}$$

Источники I

[1] S. Ivanov, "Reinforcement learning textbook," arXiv preprint arXiv:2201.09746, настольная книга по данному курсу, 2022.

[2] D. Bertsekas, Reinforcement learning and optimal control. Athena Scientific, 2019.

[3] V. Goyal and J. Grand-Clement, "A first-order approach to accelerated value iteration," Operations Research, vol. 71, no. 2, pp. 517–535, 2023.

[4] M. L. Puterman, Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, 2014.

Источники II

[5] T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian, "Policy optimization for constrained mdps with provable fast global convergence," arXiv preprint arXiv:2111.00552, 2021.