# Сведение к простым MDP

16 декабря 2018 г.

## 1 Ключевой вопрос

### 1.1 MDP

Рассмотрим MDP $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r)$, где:

- $\mathcal{S}$ — произвольное множество состояний
- $\mathcal{A}$ — **конечное** (и небольшое) множество действий
- $\mathcal{T}$ — transition probability, а то есть вероятностные распределения $p(s' \mid s, a)$, где $s, s' \in \mathcal{S}, a \in \mathcal{A}$, и которое нам неизвестно (из которого можем только сэмплировать в ходе интерактирования).
- $r : \mathcal{S} \to \mathbb{R}$ — функция награды, считаем, что детерминированная по состояниям.

### 1.2 Аппроксимация MDP

Попробуем *приблизить* этот MDP другим, более простым: $(\mathcal{S}_*, \mathcal{A}, \mathcal{T}_*, r_*)$, где:

- $\mathcal{S}_*$ — конечное множество из фиксированного числа элементов, ну там, 20.
- $\mathcal{A}$ — совпадает с $\mathcal{A}$ из предыдущего MDP
- $\mathcal{T}_*$ — распределения $p_*(s'_* \mid s_*, a)$, где $s_*, s'_* \in \mathcal{S}_*, a \in \mathcal{A}$. Для хранения этого распределения нам нужно хранить $|\mathcal{S}_*|^2|\mathcal{A}|$ чисел, что подъёмно при небольшом $|\mathcal{A}|$.
- $r_* : \mathcal{S}_* \to \mathbb{R}$ — аналогично детерминированная функция награды. Для её хранения требуется $|\mathcal{S}_*|$ чисел.

### 1.3 Эквивалентность двух MDP

Попробуем задать *эквивалентность* этих двух MDP. Назовём их эквивалентными, если существует такая функция $f \colon \mathcal{S} \to \mathcal{S}^*$, что:

- для любых $s \in \mathcal{S}, a \in \mathcal{A}, s'_* \in \mathcal{S}_*$ верно:

$$\sum_{s' \colon f(s')=s'_*} p(s' \mid s, a) = p(s'_* \mid f(s), a)$$

- для любых $s \in \mathcal{S}, a \in \mathcal{A}$ верно:

$$\sum_{s'} r(s')p(s' \mid s, a) = \sum_{s'_*} r_*(s'_*)p(s'_* \mid f(s), a)$$

### 1.4 Мягкая эквивалентность двух MDP

Нам в дальнейшем будет неудобно искать функцию $f$, выдающую дискретный выход вида «число от 1 до 20». Поэтому хочется, чтобы $f$ могло выдавать вероятность каждого из 20 состояний. Назовём MDP мягко эквивалентными, если существует такое вероятностное распределение $f(s_* \mid s), s \in \mathcal{S}, s_* \in \mathcal{S}^*$, что:

- для любых $s \in \mathcal{S}, a \in \mathcal{A}, s'_* \in \mathcal{S}_*$ верно:

$$\sum_{s'} f(s'_* \mid s')p(s' \mid s, a) = \sum_{s_*} p(s'_* \mid s_*, a)f(s_* \mid s)$$

- для любых $s \in \mathcal{S}, a \in \mathcal{A}$ верно:

$$\sum_{s'} r(s')p(s' \mid s, a) = \sum_{s'} \left[ r_*(s'_*) \sum_{s_*} p(s'_* \mid s_*, a)f(s_* \mid s) \right]$$

### 1.5 План

Будем обучать: $f_\theta(s)$ — нейросеть, которая по входу $s \in \mathcal{S}$ выдаёт вероятностное распределение на домене $\mathcal{S}_*$, табличку $|\mathcal{S}_*| \times |\mathcal{S}_*| \times |\mathcal{A}|$ чисел, моделирующую $\mathcal{T}_*$, а также функцию $r_*$.

Для этого нам бы хотелось ввести не эквивалентность, а условно метрику, которая была бы ноль при эквивалентности. В случае мягкой эквивалентности первое условие утверждает равенство двух распределений на домене $\mathcal{S}_*$, что наводит на мысль минимизировать дивергенцию между правой и левой частью.

Выберем $\mathrm{KL}$-дивергенцию, но надо подумать, прямую или обратную.

$$\mathrm{KL}\left( \sum_{s'} f_\theta(s'_* \mid s')p(s' \mid s, a), \sum_{s_*} \mathcal{T}_*(s'_*, s_*, a)f_\theta(s_* \mid s) \right) \to \min_{\theta, \mathcal{T}_*}$$

Пользоваться нужно тем, что первая вероятность тут мат.ожидание — это важно, поскольку пользоваться мы сможем только сэмплами:

$$\mathrm{KL}\left( \mathbb{E}_{s' \sim p(s\ \mid s,a)} f_\theta(s'_* \mid s'), \sum_{s_*} \mathcal{T}_*(s'_*, s_*, a)f_\theta(s_* \mid s) \right) \to \min_{\theta, \mathcal{T}_*}$$

Раскроем определение:

$$\sum_{s'_*} \left( \mathbb{E}_{s' \sim p(s' \mid s,a)} f_\theta(s'_* \mid s') \log \frac{\mathbb{E}_{s' \sim p(s' \mid s,a)} f_\theta(s'_* \mid s')}{\sum_{s_*} \mathcal{T}_*(s'_*, s_*, a)f_\theta(s_* \mid s)} \right) \to \min_{\theta, \mathcal{T}_*}$$

Дело плохо: логарифм от мат.ожидания.

### 1.6 Попытки обойти. Вариант 1

У нас проблемы только с одним слагаемым, с энтропией:

$$\sum_{s'_*} \left( \mathbb{E}_{s' \sim p(s' \mid s,a)} f_\theta(s'_* \mid s') \log \mathbb{E}_{s' \sim p(s' \mid s,a)} f_\theta(s'_* \mid s') \right)$$

Вроде как (гугл) энтропию свёртки можно сверху оценить так:

$$\sum_{s'_*} \left(\mathbb{E}_{s'\sim p(s'|s,a)} f_\theta(s'_* \mid s') \log \mathbb{E}_{s'\sim p(s'|s,a)} f_\theta(s'_* \mid s')\right) \leq$$

$$\leq \sum_{s'_*} f_\theta(s'_* \mid s') \log f_\theta(s'_* \mid s') + \sum_{s'_*} p(s' \mid s, a) \log p(s' \mid s, a)$$

Второе от параметров не зависит, оставляем первое и минимизируем верхнюю оценку. Вроде чем разреженнее $f$, тем лучше оценка.

### 1.7 Попытки обойти. Вариант 2

Рассмотрим другую KL-дивергенцию.

$$\mathrm{KL}\left(\sum_{s_*} \mathcal{T}_*(s'_*, s_*, a) f_\theta(s_* \mid s), \sum_{s'} f_\theta(s'_* \mid s') p(s' \mid s, a)\right) \to \min_{\theta, \mathcal{T}_*}$$

Обозначим первое распределение за $q(s'_*)$ и распишем эту дивергенцию:

$$\sum_{s'_*} q(s'_*) \log q(s'_*) - \sum_{s'_*} \left(q(s'_*) \log \sum_{s'} f_\theta(s'_* \mid s') p(s' \mid s, a)\right) \to \min_{\theta, \mathcal{T}_*}$$

Оставляем первое слагаемое. Второе оценим сверху, заменив содержимое логарифма на оценку снизу.

$$\sum_{s'_*} q(s'_*) \log q(s'_*) - \sum_{s'_*} \left(q(s'_*) \log \prod_{s'} f_\theta(s'_* \mid s')^{p(s'|s,a)}\right) \to \min_{\theta, \mathcal{T}_*}$$

$$\sum_{s'_*} q(s'_*) \log q(s'_*) - \mathbb{E}_{s'\sim p(s'|s,a)} \sum_{s'_*} q(s'_*) \log f_\theta(s'_* \mid s') \to \min_{\theta, \mathcal{T}_*}$$

### 1.8 EMD

Воспользуемся тем, что домен - условно 20 наших никак не различаемых состояний, то есть кажется, что Earth Moving Distance тут имеет простой вид:

$$\mathrm{EMD}\left(\mathbb{E}_{s'\sim p(s\ |s,a)} f_\theta(s'_* \mid s'), \sum_{s_*} \mathcal{T}_*(s'_*, s_*, a) f_\theta(s_* \mid s)\right) \to \min_{\theta, \mathcal{T}_*}$$

$$\sum_{s'_*} \left|\mathbb{E}_{s'\sim p(s\ |s,a)} f_\theta(s'_* \mid s') - \sum_{s_*} \mathcal{T}_*(s'_*, s_*, a) f_\theta(s_* \mid s)\right| \to \min_{\theta, \mathcal{T}_*}$$

Но тут модуль мешается.