Введение в обучение с подкреплением

Тема 5: МППР и оптимальные стратегии

Лектор: Кривошеин А.В.

МППР: уравнения Беллмана

Пусть задана динамика переходов среды в рамках конечного МППР,

то есть задана функция $p(s', r \mid s, a)$.

Пусть агент действует по стратегии π, где $\pi(a \mid s)$ — вероятности выбора действия a в состоянии s.

Для оценки стратегии мы использовали функции ценности состояний и действий.

Эти функции удовлетворяют **уравнениям Беллмана**

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in S} \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, v_\pi(s')) = \mathbb{E}_\pi[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s],$$

$$q_\pi(s, a) = \sum_{r \in \mathcal{R}} \sum_{s' \in S} p(s', r \mid s, a) \left(r + \gamma \sum_{a' \in \mathcal{A}} \pi(a' \mid s')\, q_\pi(s', a') \right).$$

Связь между функциями $v_\pi(s)$ и $q_\pi(s, a)$:

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s)\, q_\pi(s, a).$$

Для различных стратегий получатся разные функции ценности.

Как среди множества стратегий выбрать лучшую?

МППР: улучшение стратегий

Будем говорить, что **одна стратегия лучше другой**, если ожидаемый доход при действии по стратегии π' выше при старте из любого состояния:

$\pi' \geq \pi$, если $v_{\pi'}(s) \geq v_\pi(s)$ для каждого $s \in \mathcal{S}$.

Далее, так как $v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s)\, q_\pi(s, a)$, то $\min_a q_\pi(s, a) \leq v_\pi(s) \leq \max_a q_\pi(s, a)$.

Пусть $a' = \arg\max_a q_\pi(s, a)$. Тогда $q_\pi(s, a') \geq v_\pi(s)$, то есть выбор действия a' в состоянии s (и следование стратегии π затем) приведёт большему ожидаемому доходу.

Но как связаны стратегия π и стратегия, при которой во всех встреченных состояниях s мы будем принимать действие a', а других состояниях действовать по стратегии π. Является ли одна лучше другой?

Теорема об улучшении стратегии. Пусть π, π' две стратегии, причём π' детерминирована и $q_\pi(s, \pi'(s)) \geq v_\pi(s)$ для всех $s \in \mathcal{S}$, $\gamma < 1$.

Тогда стратегия π' не хуже π. То есть $\pi' \geq \pi$ или $v_{\pi'}(s) \geq v_\pi(s)$.

На самом деле, можно установить чуть более общее утверждение.

Общая теорема об улучшении стратегии. Пусть π, π' две стратегии и $\sum_{a \in \mathcal{A}} \pi'(a \mid s)\, q_\pi(s, a) \geq v_\pi(s)$ для всех $s \in \mathcal{S}$, $\gamma < 1$.

Тогда стратегия π' не хуже π. То есть $\pi' \geq \pi$ или $v_{\pi'}(s) \geq v_\pi(s)$.

МППР: доказательство теоремы об улучшении стратегий

Доказательство. Рассмотрим цепочку неравенств

$$v_\pi(s) \le q_\pi(s, \pi'(s)) = \sum_{r\in\mathcal{R}}\sum_{s'\in S} p(s', r \mid s, \pi'(s))\,(r + \gamma\, v_\pi(s')) \le$$

$$\sum_{r\in\mathcal{R}}\sum_{s'\in S} p(s', r \mid s, \pi'(s))\,(r + \gamma\, q_\pi(s', \pi'(s'))) \le$$

$$\sum_{r\in\mathcal{R}}\sum_{s'\in S} p(s', r \mid s, \pi'(s))\left(r + \gamma \sum_{r\in\mathcal{R}}\sum_{s''\in S} p(s'', r \mid s', \pi'(s'))\,(r + \gamma\, v_\pi(s''))\right) \le \ldots,$$

где многоточие значит и дальше оценивать $v_\pi(s'') \le q_\pi(s'', \pi'(s''))$ и подставлять выражение для $q_\pi(s'', \pi'(s''))$. Получающиеся суммы не отличаются от итерации равенства

$$v_{\pi'}(s) = \sum_{s'\in S}\sum_{r\in\mathcal{R}} p(s', r \mid s, \pi'(s))\,(r + \gamma\, v_{\pi'}(s')) = \sum_{r\in\mathcal{R}}\sum_{s'\in S} p(s', r \mid s, \pi'(s))\left(r + \gamma \sum_{r\in\mathcal{R}}\sum_{s''\in S} p(s'', r \mid s', \pi'(s'))\,(r + \gamma\, v_{\pi'}(s''))\right)$$

кроме самого правого слагаемого:

в цепочке неравенств оно имеет вид $v_\pi\left(s^{(n)}\right)$, а в цепочке равенств оно имеет вид $v_{\pi'}\left(s^{(n)}\right)$,

где n число проделанных итераций. У слагаемых этого типа накапливается множитель γ^n.

Зафиксируем $\varepsilon > 0$, тогда можно проделать столько итераций n, что $\gamma^n \max\limits_{s\in S} |v_\pi(s) - v_{\pi'}(s)| < \varepsilon$, а тогда

$$v_\pi(s) \le \sum_{r\in\mathcal{R}}\sum_{s'\in S} p(s', r \mid s, \pi'(s))\,(\ldots) = v_{\pi'}(s) + \left(\sum_{r\in\mathcal{R}}\sum_{s'\in S} p(s', r \mid s, \pi'(s))\,(\ldots) - v_{\pi'}(s)\right) \le v_{\pi'}(s) + \varepsilon.$$

и следовательно $v_\pi(s) \le v_{\pi'}(s)$. •

МППР: улучшение стратегий

В силу доказанной теоремы, чтобы по имеющейся стратегии **построить стратегию лучше,**
надо **новую стратегию определять жадным образом относительно** $q_\pi(s,\ a)$. А именно,

$\pi'(s) := \arg\max_a\ q_\pi(s, a)$. Эта стратегия будет не хуже, то есть $\pi \le \pi'$.

Таким образом, детерминированные стратегии всегда лучше стохастических.

Наилучшую стратегию следует искать среди детерминированных.

Для конечного МППР число детерминированных стратегий конечное число. Тогда в процессе монотонного улучшения стратегий возникнет момент, когда функции ценности по новой стратегии π' и старой стратегии π будут совпадать, то есть

$v_\pi = v_{\pi'}$ или $v_\pi(s) = \max_a q_\pi(s, a)$.

То есть ни в одном состоянии нет действия, которое ещё увеличит ожидаемый доход.
Полученное уравнение — это **уравнение оптимальности Беллмана:**

$$v_\pi(s) = \max_a q_\pi(s, a) = \max_a \sum_{r\in\mathcal{R}} \sum_{s'\in S} p(s', r \mid s, a)\,(r + \gamma\, v_\pi(s')) \text{ для всех } s \in \mathcal{S}.$$

МППР: оператор Беллмана

Введём оператор Беллмана, покажем, что он сжимающий на векторах $V = \{V(s)\}_{s\in\mathcal{S}}$.

$$B\,V(s) := \max_a \sum_{r\in\mathcal{R}}\sum_{s'\in\mathcal{S}} p(s', r \mid s, a)\,(r + \gamma\, V(s')).$$

Рассмотрим max-норму и провёдем оценку

$$\|B\,V - B\,W\| \le \max_{s\in\mathcal{S}} \left|\max_a \sum_{r\in\mathcal{R}}\sum_{s'\in\mathcal{S}} p(s', r \mid s, a)\,(r + \gamma\, V(s')) - \max_{a'} \sum_{r\in\mathcal{R}}\sum_{s'\in\mathcal{S}} p(s', r \mid s, a')\,(r + \gamma\, W(s'))\right| \le$$

$$\max_{s\in\mathcal{S}} \left|\max_a \left(\sum_{r\in\mathcal{R}}\sum_{s'\in\mathcal{S}} p(s', r \mid s, a)\,(r + \gamma\, V(s')) - \sum_{r\in\mathcal{R}}\sum_{s'\in\mathcal{S}} p(s', r \mid s, a')\,(r + \gamma\, W(s'))\right)\right| \le$$

$$\max_{s\in\mathcal{S}} \max_a \sum_{r\in\mathcal{R}}\sum_{s'\in\mathcal{S}} p(s', r \mid s, a)\,\gamma \max_{s'} |V(s') - W(s')| \le \gamma\,\|V - W\|.$$

Значит у**равнение оптимальности Беллмана** $B\,V = V$ **имеет единственное решение** при $\gamma < 1$.

В доказательстве используется неравенство : $\left|\max_x f(x) - \max_x g(x)\right| \le \max_x |f(x) - g(x)|$.

Оно верно по следующим причинам:

верно, что $f(x) \le |f(x) - g(x)| + g(x)$, а значит

$\max_x f(x) \le \max_x \left(|f(x) - g(x)| + g(x)\right) \le \max_x |f(x) - g(x)| + \max_x g(x)$.

Тем самым, $\max_x f(x) - \max_x g(x) \le \max_x |f(x) - g(x)|$.

Поменяв ролями f и g, получим

$\max_x g(x) - \max_x f(x) \le \max_x |f(x) - g(x)|$, что в итоге даёт треубемое неравенство.

МППР: оптимальная стратегия

С одной стороны, начиная с каждой стратегии, можно начать процесс улучшения и добраться до такой стратегии π^*, у которой функция ценности v_{π^*} удовлетворяет уравнению оптимальности Беллмана.

С другой стороны, решение уравнения оптимальности Беллмана единственно. То есть

$v_{\pi^*}(s) \geq v_\pi(s)$ для всех состояний $s \in \mathcal{S}$ и стратегий π или $v_{\pi^*}(s) := \max_\pi v_\pi(s)$.

Введём обозначение : $v_*(s) := \max_\pi v_\pi(s)$.

Это и есть **наилучшая функция ценности** состояний из возможных.

Любую стратегию π^*, для которой $v_{\pi^*} = v_*$, называют **оптимальной**.

Аналогично, можно ввести **наилучшую функцию ценности действий**. Пусть π^* оптимальная стратегия, тогда

$$q_{\pi^*}(s, a) = \sum_{r \in \mathcal{R}} \sum_{s' \in S} p(s', r \mid s, a)\,(r + \gamma\, v_{\pi^*}(s')) \geq \sum_{r \in \mathcal{R}} \sum_{s' \in S} p(s', r \mid s, a)\,(r + \gamma\, v_\pi(s)) = q_\pi(s, a) \text{ или}$$

$q_{\pi^*}(s, a) \geq q_\pi(s, a)$ для всех $s \in \mathcal{S}$, $a \in \mathcal{A}$ и стратегий π.

Тогда наилучшая функция ценности действий : $q_*(s, a) := \max_\pi q_\pi(s, a) = q_{\pi^*}(s, a)$.

МППР: Итерация по стратегии

Соображения выше приводят к двум методам поиска оптимальной стратегии.

Способ 1. **Метод итерации по стратегиям.**

Общая идея алгоритма итерации по стратегии:

1. Инициализация начальной стратегии π_0 случайным образом, $i = 0$
2. Повторять, пока стратегия меняется:

   Оценить стратегию, то есть найти v_{π_i} по итеративному алгоритму

   Улучшить стратегию π_i, выбрав новую стратегию π_{i+1}

   $i = i + 1$

$$\pi_0 \xrightarrow{\text{вычислить}} v_{\pi_0} \xrightarrow{\text{улучшить}} \pi_1 \xrightarrow{\text{вычислить}} v_{\pi_1} \ldots \longrightarrow \pi_* \longrightarrow v_*.$$

Шаг улучшения заключается в следующем: найти $q_\pi(s,\ a)$ для всех s и a и определить

новую стратегию π' жадным образом относительно $q_\pi(s,\ a)$.

$$\pi'(s) := \arg\max_a\ q_\pi(s, a)\ = \arg\max_a \sum_{r \in \mathcal{R}} \sum_{s' \in S} p(s', r \mid s, a)\,(r\ +\ \gamma\, v_\pi(s')).$$

Эта стратегия будет не хуже, то есть $\pi \le \pi'$.

МППР: Итерация по стратегии, псевдокод

Шаг 1. Инициализировать:

число $\theta > 0$ (порог для критерия остановки итераций);

значения $V(s)$ для $s \in S$ произвольным образом;

значения $\pi(s) \in \mathcal{A}(s)$ произвольным образом для каждого s (начальная детерминированная стратегия).

Шаг 2. Оценка стратегии

Повторять:

$\Delta := 0$

Повторять для каждого $s \in S$:

$v := V(s)$

$$V(s) := \sum_{s' \in S} \sum_{r \in \mathcal{R}} p(s', r \mid s, \pi(s)) \, (r + \gamma V(s'))$$

$\Delta := \max\{\Delta, \mid v - V(s) \mid\}$

пока не окажется, что $\Delta < \theta$.

Шаг 3. Улучшение стратегии

IsOptimal := True

Повторять для каждого $s \in S$:

oldAction := $\pi(s)$

bestActions(s) := $\arg\max_a \sum_{r \in \mathcal{R}} \sum_{s' \in S} p(s', r \mid s, a) \, (r + \gamma V(s'))$ (формируем множество лучших действий)

Если oldAction $\notin$ bestActions(s) , то IsOptimal = False.

$\pi(s) = a$, где $a \in$ bestActions(s).

Если IsOptimal = True, то стоп, иначе перейти к Шагу 2.

Особенность псевдокода: при запуске Шага 2 для оценки стратегии, начальной оценкой является приближение функции ценности предыдущей стратегии (на практике это обычно увеличивает скорость сходимости).

МППР: Итерация по ценности

Способ 2. **Метод итерации по ценности**.

Поскольку решение уравнения оптимальности Беллмана — это нахождение неподвижной точки сжимающего оператора, то решение, то есть оптимальную функцию ценности состояний v_*, можно искать итеративным методом.

$$V_{k+1} = B\,V_k \quad \text{или} \quad V_{k+1}(s) = \max_a \sum_{s' \in S} \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, V_k(s')).$$

Как восстановить оптимальную стратегию π^*?

Для этого можно вычислить оптимальную функцию ценности действий:

$$q_*(s, a) = \sum_{r \in \mathcal{R}} \sum_{s' \in S} p(s', r \mid s, a)\,(r + \gamma\, v_*(s'))$$

и **выбирать действия жадно относительно** $q_*(s, a)$:

$$\pi^*(s) := \arg\max_a q_*(s, a) = \operatorname{argmax}_a \sum_{s' \in S} \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, v_*(s')).$$

Общий алгоритм **итерации по ценности:**

1. Инициализировать случайным образом функцию ценности для всех состояний $V_0(s)$.
2. Повторять, пока не получится хорошее приближение:

   $V_{k+1} = B\,V_k$.
3. Выделить оптимальную стратегию.

Если изобразить схему алгоритма, аналогичную итерации по стратегии, то она имеет вид

$$V_0 \longrightarrow V_1 \longrightarrow V_2 \longrightarrow \ldots \longrightarrow v_*.$$

МППР: Итерация по ценности, псевдокод

Шаг 1. Инициализировать:

число $\theta > 0$ (порог для критерия остановки итераций);

значения $V(s)$ для $s \in S$ произвольным образом;

Шаг 2. Решение уравнения оптимальности Беллмана:

Повторять:

$\Delta := 0$

Повторять для каждого $s \in S$:

$v := V(s)$

$$V(s) := \max_a \sum_{s' \in S} \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, V(s'))$$

$\Delta := \max\{\Delta, \mid v - V(s) \mid\}$

пока не окажется, что $\Delta < \theta$.

Шаг 3. Выделение оптимальной стратегии:

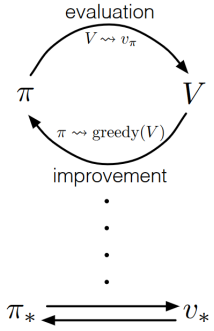
$$\pi(s) = \operatorname*{argmax}_a \sum_{s' \in S} \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, V(s'))$$

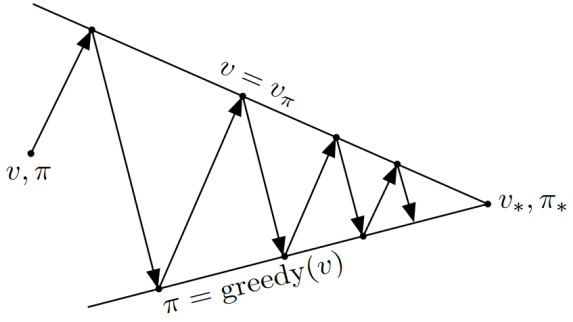
МППР: диаграммы

Итерация по стратегиям означает чередование двух шагов:

1. Оценка: найти истинную функцию ценности v_π для текущей стратегии;
2. Улучшение: выбрать стратегию жадно, относительно текущей функции ценности v_π.

Out[ ]=





Итерация по ценности в некотором смысле также чередует эти два шага: по сути вычисляется одна итерация оценки Q-функции и при необходимости можно получить улучшенную стратегию жадным выбором действия по этой оценке.

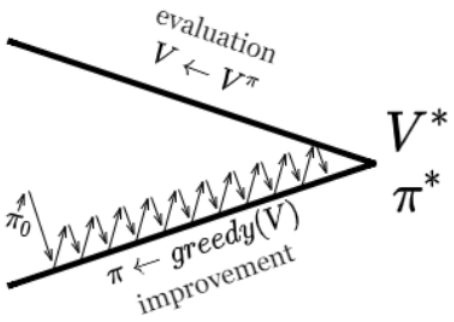
При вычисления in-place чередование шагов происходит ещё чаще: обновляется оценка ценности для одного состояния и сразу улучшается стратегия.

Обобщённая итерация по стратегиям (ОИС) — это чередование шагов оценки и улучшения стратегии произвольным образом.

Шаг улучшения стратегии жадно выбирает стратегию по текущей функции ценности и делает эту функцию некорректной для новой стратегии.

Шаг обновления оценки для текущей функции ценности приводит к тому, что новая стратегия перестаёт быть жадной.

Однако, совместно эти два процесса находят единственную оптимальную функцию ценности.


evaluation
$V \leftarrow V^{\pi}$
V^*
π^*
π_0
$\pi \leftarrow greedy(V)$
improvement