3. ЛЕКЦИЯ. Классическая теория 1

Оценочные функции

Свойства траекторий

Как это всегда бывает, чем более общую задачу мы пытаемся решать, тем менее эффективный алгоритм мы можем придумать. В RL (обучение с подкреплением) мы сильно замахиваемся: хотим построить алгоритм, способный обучаться решению «произвольной» задачи, заданной средой с описанной функцией награды. Однако в формализме MDP (частично наблюдаемых Марковских процессов принятия решений) в постановке на самом деле внесли некоторые ограничения: марковость и стационарность. Эти предположения практически не ограничивают общность нашей задачи с точки зрения здравого смысла с одной стороны и при этом вносят в нашу задачу некоторую «структуру»; мы сможем придумать более эффективные алгоритмы решения за счёт эксплуатации этой структуры.

Что значит «структуру»? Представим, что мы решаем некоторую абстрактную задачу последовательного принятия решения, максимизируя некоторую кумулятивную награду. Вот мы находимся в некотором состоянии и должны выбрать некоторое действие. Интуитивно ясно, что на прошлое — ту награду, которую мы уже успели собрать — мы уже повлиять не можем, и нужно максимизировать награду в будущем. Более того, мы можем отбросить всю нашу предыдущую историю и задуматься лишь над тем, как максимизировать награду с учётом сложившейся ситуации — «текущего состояния».

Давайте сформулируем эту интуицию формальнее. Как и в обычных Марковских цепях, в средах благодаря марковости действует закон «независимости прошлого и будущего при известном настоящем». Формулируется он так:

Утверждение — Независимость прошлого и будущего при известном настоящем:

Пусть $\mathcal{T}_{:t} := \{s_0, a_0 \dots s_{t-1}, a_{t-1}\}$ — «прошлое», s_t — «настоящее», $\mathcal{T}_{t:} := \{a_t, s_{t+1}, a_{t+1\dots}\}$ — «будущее».

Тогда:

$$p(\mathcal{T}_{:t}, \mathcal{T}_{t:} \mid s_t) = p(\mathcal{T}_{:t} \mid s_t) p(\mathcal{T}_{t:} \mid s_t)$$

В силу марковости будущее зависит от настоящего и прошлого только через настоящее:

Для нас утверждение означает следующее: если мы сидим в момент времени t в состоянии s и хотим посчитать награду, которую получим в будущем (то есть величину, зависящую только от $\mathcal{T}_{t:}$), то нам совершенно не важна история

попадания в s. Это следует из свойства мат. ожиданий по независимым переменным:

$$\mathbb{E}_{\mathcal{T}|s_t=s} R(\mathcal{T}_{t:}) = \mathbb{E}_{\mathcal{T}_{:t}|s_t=s} \underbrace{\mathbb{E}_{\mathcal{T}_{t:}|s_t=s} R(\mathcal{T}_{t:})}_{\text{не зависит от } \mathcal{T}_{:t}} = \mathbb{E}_{\mathcal{T}_{t:}|s_t=s} R(\mathcal{T}_{t:}) \quad {}_{t=s} R(\mathcal{T}_{t:})$$

Определение: Для траектории $\mathcal{T}$ величина

$$R_t := R(\mathcal{T}_{t:}) = \sum_{\hat{t} \geq t} \gamma^{\hat{t}-t} \, \mathrm{r}_{\hat{t}}$$

называется reward-to-go (вознаграждение на ходу) с момента времени t.

Благодаря сделанному предположению, о стационарности (в том числе стационарности стратегии агента), получается, что будущее также не зависит от текущего момента времени t: всё определяется исключительно текущим состоянием. Иначе говоря, агенту неважно не только, как он добрался до текущего состояния и сколько награды встретил до настоящего момента, но и сколько шагов в траектории уже прошло. Формально это означает следующее: распределение будущих траекторий имеет в точности тот же вид, что и распределение всей траектории при условии заданного начала.

Утверждение: Будущее определено текущим состоянием:

$$p(\mathcal{T}_{t:} \,|s_t = s) \equiv p\big((\mathcal{T} \,|s_0 = s\big)$$

Утверждение: Для любого t и любой функции f от траекторий:

$$\mathbb{E}_{\mathcal{T}|s_0=s} f(T) \;=\; \mathbb{E}_{\mathcal{T}|s_t=s} f(\mathcal{T}_{t:})$$

V-функция

Итак, интуиция заключается в том, что, когда агент приходит в состояние s, прошлое не имеет значения, и оптимальный агент должен максимизировать в том числе и награду, которую он получит, стартуя из состояния s. Поэтому «обобщим» оптимизируемый функционал, варьируя стартовое состояние:

Определение: Для данного MDP V-функцией (value function – функции цены) или оценочной функцией состояний (state value function) для данной стратегии π называется величина

$$V^{\pi}(s) := \mathbb{E}_{\mathcal{T}\sim\pi|s_0=s} R(\mathcal{T}) \qquad (3.1)$$

По определению функция ценности состояния, или V-функция – это сколько набирает в среднем агент из состояния s. Причём в силу марковости и стационарности неважно, случился ли старт на нулевом шаге эпизода или на произвольном t-ом:

Утверждение: Для любого t верно

$$V^{\pi}(s) \;=\; \mathbb{E}_{\mathcal{T}\sim\pi|s_t=s} R_t$$

Утверждение: $V^{\pi}(s)$ ограничено

Утверждение: Для терминальных состояний $V^{\pi}(s) = 0$

Заметим, что любая политика π индуцирует V^{π}. То есть для данного MDP и данной стратегии π функция V^{π} однозначно задана своим определением; совсем другой вопрос, можем ли мы вычислить эту функцию.

Пример: Посчитаем V-функцию для MDP и стратегии π с рисунка, $\gamma = 0.8$. Её часто удобно считать «с конца», начиная с состояний, близких к терминальным, и замечая связи между значениями функции для разных состояний (рис. 3.1)

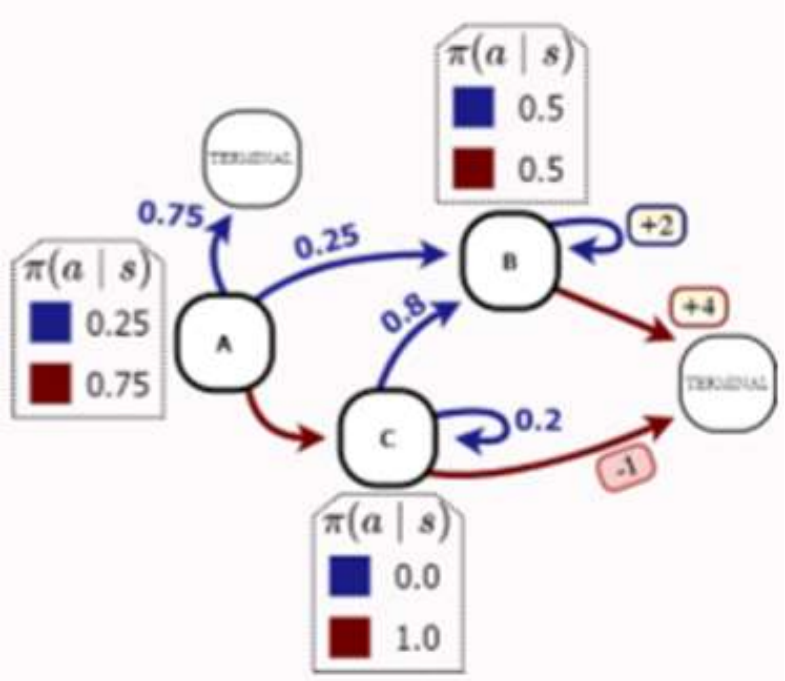


Рис.3.1

Начнём с состояния C: там агент всегда выбирает действие ■, получает -1, и эпизод заканчивается: $V^{\pi}(s = C) = -1$.

Для состояния B с вероятностью 0.5 агент выбирает действие ■ и получает $+4$. Иначе он получает $+2$ и возвращается снова в состояние B. Вся дальнейшая награда будет дисконтирована на $\gamma = 0.8$ и тоже равна $V^{\pi}(s = B)$ по определению. Итого:

$$V^{\pi}(s = B) = \underbrace{0.5 \cdot 4}_{\blacksquare} + \underbrace{0.5 \cdot (2 + \gamma V^{\pi}(s = B))}_{\blacksquare}$$

Решая это уравнение относительно $V^{\pi}(s = B)$, получаем ответ 5.

Для состояния A достаточно аналогично рассмотреть все дальнейшие события:

$$V^{\pi}(s = A) = \underbrace{0.25}_{\blacksquare} \cdot \left(\underbrace{0.75 \cdot 0}_{\text{terminal}} + \underbrace{0.25\gamma V^{\pi}(s = B)}_{B} \right) + \underbrace{0.75}_{\blacksquare} \underbrace{\gamma V^{\pi}(s = C)}_{C}$$

Подставляя значения, получаем ответ $V^{\pi}(s = A) = -0.35$.

Уравнения Беллмана

Если s_0 – стартовое состояние, то $V^{\pi}(s_0)$ по определению и есть функционал, который мы хотим оптимизировать. Формально, это единственная величина, которая нас действительно волнует, так как она нам явно задана в самой постановке задачи, но мы понимаем, что для максимизации $V^{\pi}(s_0)$ нам нужно промаксимизировать и $V^{\pi}(s)$. Другими словами, у нас в задаче есть подзадачи эквивалентной структуры: возможно, они, например, проще, и мы можем сначала их решить, а дальше как-то воспользоваться этими решениями для решения более сложной. Вот если граф MDP есть дерево, например, то очевидно, как считать V^{π}: посчитать значение в листьях (листья соответствуют терминальным состояниям — там ноль), затем в узлах перед листьями, ну и так далее индуктивно добраться до корня. Можно заметить, что в предыдущем примере на значения V-функции начали появляться рекурсивные соотношения. В этом и есть смысл введения понятия оценочных функций — «дополнительных переменных»: в том, что эти значения связаны между собой уравнениями Беллмана (Bellman equations).

Теорема – Уравнение Беллмана (Bellman expectation equation) для V^{π}:

$$V^{\pi}(s) = \mathbb{E}_a [r(s,a) + \gamma \mathbb{E}_{s'} s' V^{\pi}(s')]$$

Для формального доказательства раскладывается сумма по времени как первое слагаемое плюс сумма по времени и пользуемся утверждением о независимости V-функции от времени.

Пример: Выпишем уравнения Беллмана для MDP и стратегии π из предыдущего примера (рис.3.1). Число уравнений совпадает с числом состояний. Разберём подробно уравнение для состояния A:

$$V^{\pi}(A) = \underbrace{0.25}_{\blacksquare}(0 + \gamma 0.25 V^{\pi}(B)) + \underbrace{0.75}_{\blacksquare}(0 + \gamma V^{\pi}(C))$$

С вероятностью 0.25 будет выбрано действие ■, после чего случится дисконтирование на γ; с вероятностью 0.75 эпизод закончится и будет выдана нулевая награда, с вероятностью 0.25 агент перейдёт в состояние B. Второе слагаемое уравнения будет отвечать выбору действия ■; агент тогда перейдёт в состояние C и, начиная со следующего шага, получит в будущем $V^{\pi}(C)$. Аналогично расписываются два оставшихся уравнения.

$$V^{\pi}(A) = \frac{1}{16}\gamma V^{\pi}(B) + \frac{3}{4}\gamma V^{\pi}(C)$$
$$V^{\pi}(B) = 0.5\,(2 + \gamma V^{\pi}(B)) + 0.5 \cdot 4$$
$$V^{\pi}(C) = -1$$

Заметим, что мы получили систему из трёх линейных уравнений с тремя неизвестными.

Оптимальная стратегия

У нас есть конкретный функционал $J(\pi) = V^{\pi}(s_0)$, который мы хотим оптимизировать. Казалось бы, понятие оптимальной политики очевидно как вводить:

Определение: Политика π^* оптимальна, если $\forall\pi$: $V^{\pi^*}(s_0) \geq V^{\pi}(s_0)$.

Введём альтернативное определение:

Определение: Политика π^* оптимальна, если $\forall\pi, s$: $V^{\pi^*}(s) \geq V^{\pi}(s)$.

Теорема: Определения не эквивалентны.

Интуиция подсказывает, что различие между определениями проявляется только в состояниях, которые оптимальный агент будет избегать с вероятностью 1. Задавая оптимальность вторым определением, мы усложняем задачу, но упрощаем теоретический анализ: если бы мы оставили первое определение, у оптимальных политик могли бы быть разные V-функции; согласно второму определению, V-функция всех оптимальных политик совпадает.

Определение: Оптимальные стратегии будем обозначать π^*, а соответствующую им оптимальную V-функцию – V^*:

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

Пока нет никаких обоснований, что найдётся стратегия, которая максимизирует $V^{\pi}(s)$ сразу для всех состояний. Если в одних s макисмум достигается на одной стратегии, а в другом — на другой? Тогда оптимальных стратегий в сильном смысле вообще не существует, хотя формальная величина существует. Заметим лишь, что для ситуации, когда MDP — дерево, существование оптимальной стратегии в смысле второго определения можно опять показать «от листьев к корню».

Q-функция

V-функции нам не хватит. Если бы мы знали оптимальную value-функцию $V^*(s)$, мы не смогли бы восстановить хоть какую-то оптимальную политику из-за отсутствия в общем случае информации о динамике среды. Допустим, агент находится в некотором состоянии и знает его ценность $V^*(s)$, а также знает ценности всех других состояний; это не даёт понимания того, какие действия в

какие состояния приведут — мы никак не дифференцируем действия между собой. Поэтому увеличим количество переменных: введём схожее определение для ценности не состояний, но пар состояние-действие.

Определение: Для данного MDP Q-функцией (state-action value function, action quality function) для данной стратегии π называется

$$Q^{\pi}(s,a) := \mathbb{E}_{\mathcal{T}\sim\pi|s_0=s,a_0=a} \sum_{t\geq 0} \gamma^t r_t$$

Теорема – Связь оценочных функций: V-функции и Q-функции взаимозависимы, а именно:

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} V^{\pi}(s')$$
$$V^{\pi}(s) = \mathbb{E}_{a\sim\pi(a|s)} Q^{\pi}(s,a)$$

Итак, если V-функция — это сколько получит агент из некоторого состояния, то Q-функция — это сколько получит агент после выполнения данного действия из данного состояния. Как и V-функция, Q-функция не зависит от времени, ограничена по модулю при рассматриваемых требованиях к MDP, и, аналогично, для неё существует уравнение Беллмана:

Теорема – Уравнение Беллмана (Bellman expectation equation) для Q-функции:

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} \mathbb{E}_{a'} Q^{\pi}(s',a')$$

Пример: Q-функция получает на вход пару состояние-действие и ничего не говорится о том, что это действие должно быть как-то связано с оцениваемой стратегией π (рис. 3.2).

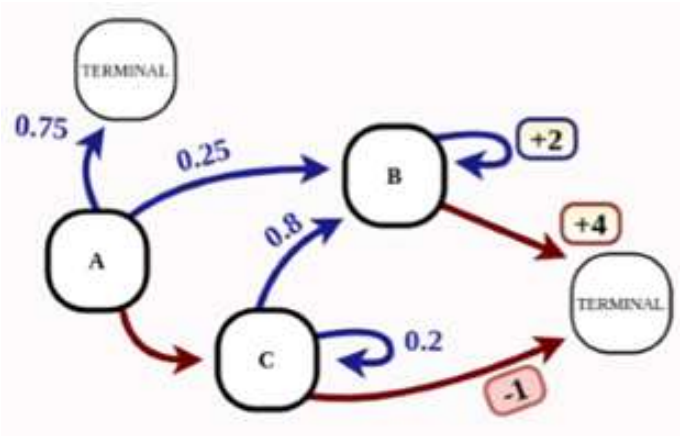


Рис.3.2

Давайте в MDP с рисунка рассмотрим стратегию π, которая всегда детерминировано выбирает действие ■. Мы тем не менее можем посчитать $Q^{\pi}(s,$ ■$)$ для любых состояний (например, для терминальных это значение формально равно нулю). Сделаем это при помощи QV уравнения:

$$Q^\pi(s=A, \blacksquare) = 0.25\gamma V^\pi(s=B)$$
$$Q^\pi(s=B, \blacksquare) = 2 + \gamma V^\pi(s=B)$$
$$Q^\pi(s=C, \blacksquare) = 0.8\gamma V^\pi(s=B) + 0.2\gamma V^\pi(s=C)$$

Внутри V^π сидит дальнейшее поведение при помощи стратегии π, то есть выбор исключительно действий ■: соответственно,

$V^\pi(s=B) = 4, V^\pi(s=C) = -1.$

Определение: Принцип оптимальности Беллмана: жадный выбор действия в предположении оптимальности дальнейшего поведения оптимален.

Догадку несложно доказать для случая, когда MDP является деревом: принятие решения в текущем состоянии s никак не связано с выбором действий в «поддеревьях». Если в поддереве, соответствующему одному действию, можно получить больше, чем в другом поддереве, то понятно, что выбирать нужно его. В общем случае, однако, нужно показать, что жадный выбор в s «позволит» в будущем набрать то $Q^*(s,a)$, которое мы выбрали — вдруг для того, чтобы получить в будущем $Q^*(s,a)$, нужно будет при попадании в то же состояние s выбирать действие как-то по-другому? Если бы это было так, было бы оптимально искать стратегию в классе нестационарных стратегий.

Тогда приходим к отказу от однородности, которая заключается в том, что будем искать максимум $\max_\pi V^\pi(s)$ не только среди стационарных, но и нестационарных стратегий. Мотивация в отказе от однородности заключается в том, что наше MDP теперь стало деревом: эквивалентно было бы сказать, что мы добавили в описание состояний время t. Теперь мы не оказываемся в одном состоянии несколько раз за эпизод; максимизация $Q_t^*(s,a)$ требует оптимальных выборов «в поддереве», то есть настройки π_{t+1}, π_{t+2} и так далее, а для $\pi_t(a|s)$ будет выгодно выбрать действие жадно.

Уравнения оптимальности Беллмана

Теорема – Связь оптимальных оценочных функций:

$$V^*(s) = \max_a Q^*(s,a)$$
$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} V^*(s')$$

Теперь V^* выражено через Q^* и наоборот. Значит, можно получить выражение для V^* через V^* и Q^* через Q^*:

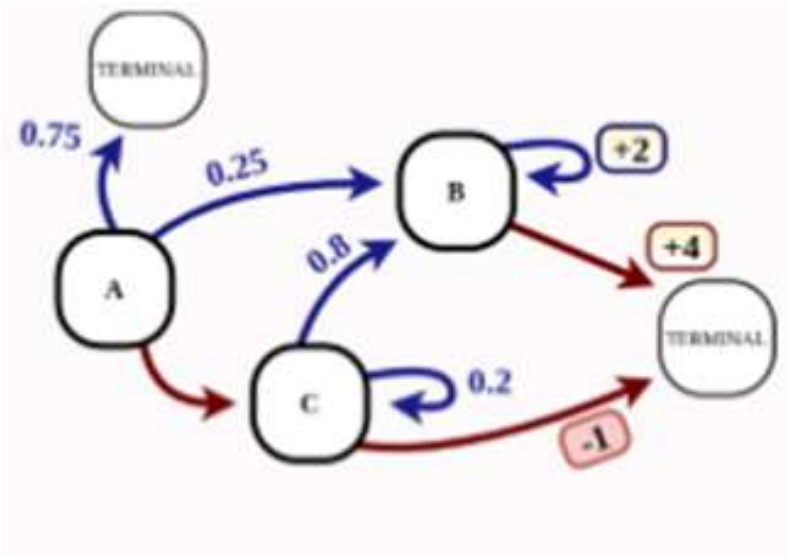
Теорема – Уравнения оптимальности Беллмана (Bellman optimality equation):

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'} \max_{a'} Q^*(s',a')$$

$$V^*(s) = \max_a [r(s,a) + \gamma \mathbb{E}_{s'} V^*(s')]$$

Уравнения оптимальности Беллмана крайне интуитивны. Для Q^*, например, можно рассудить так: что даст оптимальное поведение из состояния s после совершения действия a? Что с нами случится дальше: мы получим награду за этот выбор $r(s,a)$, на что уже повлиять не можем. Остальная награда будет дисконтирована. Затем среда переведёт нас в какое-то следующее состояние s' — нужно рассчитать мат. ожидание по функции переходов. После этого мы, пользуясь принципом Беллмана, просто выберем то действие, которое позволит в будущем набрать наибольшую награду, и тогда сможем получить $\max_{a'} Q^*(s', a')$.

Пример. Сформулируем для MDP с рисунка уравнения оптимальности Беллмана для V^*. Мы получим систему из трёх уравнений с трёмя неизвестными (рис. 3.3).



$$V^*(s=A) = \max\left(\underbrace{0.25\gamma V^*(s=B)}_{\blacksquare}, \underbrace{\gamma V^*(s=C)}_{\blacksquare}\right)$$

$$V^*(s=B) = \max\left(\underbrace{2+\gamma V^*(s=B)}_{\blacksquare}, \underbrace{4}_{\blacksquare}\right)$$

$$V^*(s=C) = \max\left(\underbrace{0.8\gamma V^*(s=B) + 0.2\gamma V^*(s=C)}_{\blacksquare}, \underbrace{-1}_{\blacksquare}\right)$$

Рис.3.3

Заметим, что в полученных уравнениях не присутствует мат.ожиданий по самим оптимальным стратегиям — предположение дальнейшей оптимальности поведения по сути «заменяет» их на взятие максимума по действиям. Более того,

оптимальные оценочные функции — единственные решения систем уравнений Беллмана. А значит, вместо поиска оптимальной стратегии можно искать оптимальные оценочные функции! Таким образом, мы свели задачу оптимизации нашего функционала к решению системы нелинейных уравнений особого вида. Беллман назвал данный подход «динамическое программирование» (dynamic programming).

Критерий оптимальности Беллмана

Теперь сформулируем критерий оптимальности стратегий в общей форме, описывающей вид всего множества оптимальных стратегий. Для доказательства нам понадобится факт: для данного MDP Q^* — единственная функция $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$, удовлетворяющая уравнениям оптимальности Беллмана.

Теорема – Критерий оптимальности Беллмана: π оптимальна тогда и только тогда, когда $\forall s, a: \pi(a|s) > 0$ верно:

$$a \in Arg \max_a Q^\pi(s, a)$$

Иначе говоря: теорема говорит, что оптимальны ровно те стратегии, которые пользуются принципом оптимальности Беллмана. Если в одном состоянии два действия позволят в будущем набрать максимальную награду, то между ними можно любым способом размазать вероятности выбора. Давайте при помощи этого критерия окончательно ответим на вопросы о том, существует ли оптимальная стратегия и сколько их вообще может быть.

Утверждение: Если $|\mathcal{A}| < +\infty$, всегда существует оптимальная стратегия.

Утверждение: Оптимальной стратегии может не существовать

Утверждение: Если существует хотя бы две различные оптимальные стратегии, то существует континуум (множество вещественных чисел) оптимальных стратегий.

Утверждение: Если существует хотя бы одна оптимальная стратегия, то существует детерминированная оптимальная стратегия

Пример. Найдём все оптимальные стратегии в MDP для $\gamma = 0.5$.

Мы могли бы составить уравнения оптимальности Беллмана для Q^* и решать их, но сделаем чуть умнее и воспользуемся критерием оптимальности Беллмана. Например, в состоянии B (рис. 3.4) оптимально или выбирать какое-то одно из двух действий с вероятностью 1, или действия эквивалентны, и тогда оптимально любое поведение. Допустим, мы будем выбирать всегда синий, тогда мы получим $2/(1-\gamma) = 4$; если же будем выбирать красный, то получим +4. Значит, действия эквивалентны, оптимально любое поведение, и $V^*(s = B) = 4$.

Проведём аналогичное рассуждение для состояния C. Если оптимально

действие синее, то

$$Q^*(s = C, \blacksquare) = 0.2\gamma Q^*(s = C, \blacksquare) + 0.8\gamma V^*(s = B)$$

Решая это уравнение относительно неизвестного $Q^*(s = C, \blacksquare)$, получаем $\frac{16}{9} > Q^*(s = C, \blacksquare) = -1$.

Значит, в C оптимальная стратегия обязана выбирать $\blacksquare$, и $V^*(C) = \frac{16}{9}$.

Для состояния A достаточно сравнить

$$Q^*(s = A, \blacksquare) = 0.25\gamma V^*(s = B) = \frac{1}{4} \text{ и}$$

$$Q^*(s = A, \blacksquare) = \gamma V^*(s = C) = \frac{8}{9},$$

определив, что оптимальная стратегия должна выбирать красный.

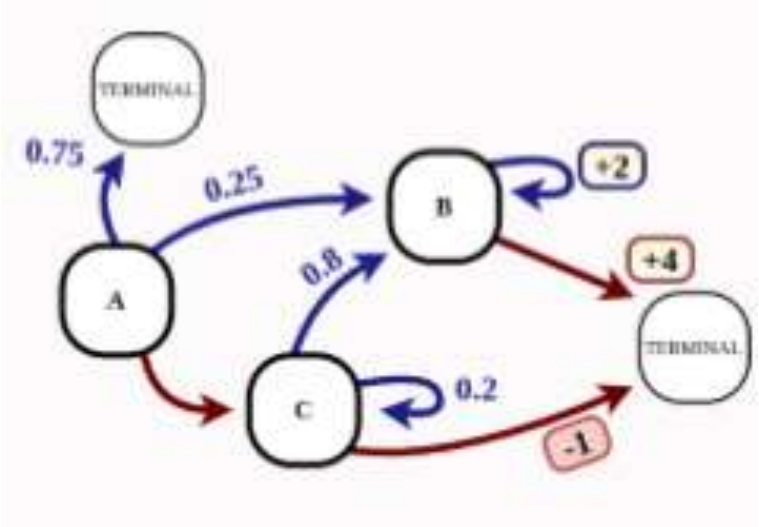


Рис.3.4

Улучшение политики

Advantage (выгода)-функция

Допустим, мы находились в некотором состоянии s, и засэмплировали $a \sim \pi(a|s)$ такое, что $Q^\pi(s,a) >> V^\pi(s)$. Что можно сказать о таком действии? Мы знаем, что вообще в среднем политика π набирает из данного состояния $V^\pi(s)$, но какой-то выбор действий даст в итоге награду больше $V^\pi(s)$, а какой-то меньше. Если $Q^\pi(s,a) > V^\pi(s)$, то после того, как мы выбрали действие a, «приняли решение», наша средняя будущая награда вдруг увеличилась. Эта проблема обучения с подкреплением, как credit assingment (выделение кредита), которая звучит примерно так: допустим, мы засэмплировали траекторию $s, a, s', a' \ldots$ до конца эпизода, и в конце в финальном состоянии через T шагов получили сигнал (награду) +1. Мы приняли T решений, но какое из всех этих

действий повлекло получение этого +1? «За что нас наградили?» Повлияло ли на получение +1 именно то действие a, которое мы засэмплировали в стартовом s? Вопрос нетривиальный, потому что в RL есть отложенный сигнал: возможно, именно действие a в состоянии s запустило какую-нибудь цепочку действий, которая дальше при любом выборе $a', a'', \cdots$ приводит к награде +1. Возможно, конечно, что первое действие и не имело никакого отношения к этой награде, и это поощрение именно за последний выбор. А ещё может быть такое, что имело место везение, и просто среда в какой-то момент перекинула нас в удачное состояние.

Но мы понимаем, что если какое-то действие началось получение награды через сто шагов, в промежуточных состояниях будет информация о том, сколько времени осталось до получения этой отложенной награды. Например, если мы выстрелили во вражеский инопланетный корабль, и через 100 шагов выстрел попадает во врага, давая агенту +1, мы будем видеть в состояниях расстояние от летящего выстрела до цели, и знать, что через такое-то время нас ждёт +1. Другими словами, вся необходимая информация лежит в идеальных оценочных функциях Q^π и V^π.

Так, если в некотором состоянии s засэмплировалось такое a, что $Q^\pi(s,a) = V^\pi(s)$, то мы можем заключить, что выбор действия на этом шаге не привёл ни к какой «неожиданной» награде. Если же $Q^\pi(s,a) >> V^\pi(s)$ – то мы приняли удачное решение, $Q^\pi(s,a) < V^\pi(s)$ – менее удачное, чем обычно. Если, например, $r(s,a) + V^\pi(s') > Q^\pi(s,a)$, то мы можем заключить, что имело место везение: среда засэмплировала такое s', что теперь мы получим больше награды, чем ожидали после выбора a в состоянии s. И так далее: мы сможем отследить, в какой конкретно момент случилось то событие (сэмплирование действия или ответ среды), за счёт которого получена награда.

Таким образом, идеальный «кредит» влияния действия a, выбранного в состоянии s, на будущую награду равен

$$Q^\pi(s,a) - V^\pi(s),$$

и именно эта величина на самом деле будет для нас ключевой. Поэтому из соображений удобства вводится ещё одно обозначение:

Определение: Для данного MDP Advantage-функцией политики π называется

$$A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$$

Отсюда выводятся ряд утверждений.

Утверждение: Для любой политики π и любого состояния s:

$$\mathbb{E}_{\pi(a|s)} A^\pi(s,a) = 0$$

Утверждение: Для любой политики π и любого состояния s:

$$\max_{a} A^{\pi}(s,a) \geq 0$$

Advantage (выгода) — это, если угодно, «центрированная» Q-функция. Если $A^{\pi}(s,a) \geq 0$ – действие a «лучше среднего» для нашей текущей политики в состоянии s, меньше нуля — хуже. И интуиция, что процесс обучения нужно строить на той простой идеи, что первые действия надо выбирать чаще, а вторые — реже, нас не обманывает.

Естественно, подвох в том, что на практике мы небудем знать точное значение оценочных функций, а значит, и истинное значение Advantage (выгоды). Решая вопрос оценки значения Advantage (выгоды) для данной пары s, a, мы фактически будем проводить credit assingment (выделение кредита) – это одна и та же задача.

Policy Improvement (Улучшение политики)

Определение: Будем говорить, что стратегия π_2 «не хуже» π_1 (запись: $\pi_2 \succcurlyeq \pi_1$), если $\forall s$:

$$V^{\pi_2}(s) \geq V^{\pi_1}(s),$$

и лучше (запись $\pi_2 \succ \pi_1$), если также найдётся s, для которого неравенство выполнено строго:

$$V^{\pi_2}(s) > V^{\pi_1}(s),$$

Мы ввели частичный порядок на множестве стратегий (понятно, что можно придумать две стратегии, которые будут «не сравнимы»: когда в одном состоянии одна будет набирать больше второй, в другом состоянии вторая будет набирать больше первой).

Зададимся следующим вопросом. Пусть для стратегии π_1 мы знаем оценочную функцию Q^{π_1}; тогда мы знаем и V^{π_1} из VQ уравнения и A^{π_1} по определению. Давайте попробуем построить $\pi_2 \succ \pi_1$. Для этого покажем более «классическим» способом, что стратегии π_2 достаточно лишь в среднем выбирать действия, дающие неотрицательный Advantage (выбор) стратегии π_1, чтобы быть не хуже.

Теорема – Policy Improvement: Пусть стратегии π_1 и π_2 таковы, что для всех состояний s выполняется:

$$\mathbb{E}_{\pi_2(a|s)} Q^{\pi_1}(s,a) \geq V^{\pi_1}(s),$$

или, в эквивалентной форме:

$$\mathbb{E}_{\pi_2(a|s)} A^{\pi_1}(s,a) \geq 0.$$

Тогда $\pi_2 \succcurlyeq \pi_1$; если хотя бы для одного s неравенство выполнено строго, то $\pi_2 \succ \pi_1$.

Что означает эта теорема? Знание оценочной функции позволяет улучшить стратегию. Улучшать стратегию можно прямо в отдельных состояниях, например, выбрав некоторое состояние s и сказав: неважно, как это повлияет на частоты посещения состояний, но будем конкретно в этом состоянии s выбирать действия так, что значение

$$\mathbb{E}_{\pi_2(a|s)} Q^{\pi_1}(s,a)$$

как можно больше. Тогда, если в s действие выбирается «новой» стратегией π_2, а в будущем агент будет вести себя не хуже, чем π_1, то и наберёт он в будущем не меньше $Q^{\pi_1}(s,a)$. Теорема показывает, что выражение $\mathbb{E}_{\pi_2(a|s)} Q^{\pi_1}(s,a)$ является нижней оценкой на награду, которую соберёт «новый» агент со стратегией π_2.

Если эта нижняя оценка поднята выше $V^{\pi_1}(s)$, то стратегию удалось улучшить: и тогда какой бы ни была π_1, мы точно имеем гарантии $\pi_2 \geqslant \pi_1$. Важно отметить, что такой policy improvement (улучшение политики) работает всегда: и для «плохих» стратегий, близких к случайному поведению, и для уже умеющих что-то делать разумное.

В частности, мы можем попробовать нижнюю оценку максимально поднять, то есть провести жадный (greedy) policy improvement. Для этого мы формально решаем такую задачу оптимизации:

$$\mathbb{E}_{\pi_2(a|s)} Q^{\pi_1}(s,a) \to \max_{\pi_2},$$

и понятно, что решение находится в детерминированной π_2:

$$\pi_2(s) = \arg\max_a Q^{\pi_1}(s,a) = \arg\max_a A^{\pi_1}(s,a)$$

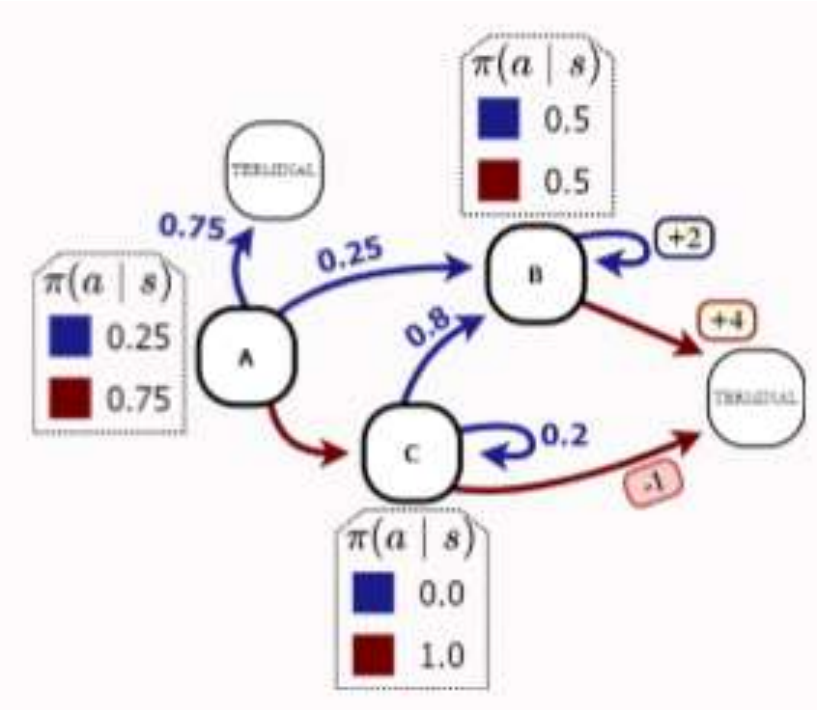
Конечно, мы так не получим «за один ход» сразу оптимальную стратегию, поскольку выбор $\pi_2(a|s)$ сколь угодно хитро может изменить распределение траекторий, но тем не менее.

Пример: Попробуем улучшить стратегию π из рассмотренного примера, γ = 0.8.

Например, в состоянии C она выбирает ■ с вероятностью 1, и получает -1; попробуем посчитать $Q^\pi(s = C, ■)$:

$$Q^\pi(s = C, ■) = 0.2\gamma V^\pi(C) + 0.8\gamma V^\pi(B)$$

Подставляя ранее подсчитанные $V^\pi(C) = -1, V^\pi(B) = 5$, видим, что действие ■ принесло бы нашей стратегии π куда больше -1, а именно $Q^\pi(s = C, ■) = 3.04$. Давайте построим π_2, скопировав π в A и B, а в C будем с вероятностью 1 выбирать ■.



Что говорит нам теория? Важно, что она не даёт нам значение $V^{\pi_2}(C)$; в частности, нельзя утверждать, что $Q^{\pi_2}(s = C, \blacksquare) = \mathbf{3.04}$, и повторение вычислений подтвердит, что это не так. Однако у нас есть гарантии, что, во-первых, $Q^{\pi_2}(s = C, \blacksquare) \geq \mathbf{3.04}$, и, что важнее, из состояния C мы начали набирать больше награды: $V^{\pi_2}(C) > V^{\pi_1}(C)$ строго. Во-вторых, есть гарантии, что мы не «сломали» стратегию в других состояниях: во всех остальных состояниях гарантированно $V^{\pi_2}(s) \geq V^{\pi_1}(s)$. Для Q-функции, как можно показать, выполняются аналогичные неравенства.