

Обучение с подкреплением (Reinforcement Learning)

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
<http://www.MachineLearning.ru/wiki>
«Машинное обучение (курс лекций, К.В.Воронцов)»

15 ноября 2016

1 Задача о многоруком бандите

- Простая постановка задачи
- Жадные и полужадные стратегии
- Адаптивные стратегии

2 Среда с контекстом

- Постановка задачи
- Линейная модель премий
- Оценивание модели по историческим данным

3 Среда с состояниями

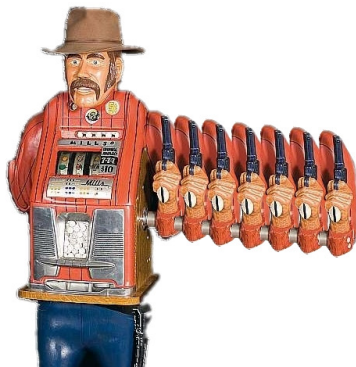
- Постановка задачи
- Ценность состояния и действия
- Методы временных разностей, SARSA, Q-обучения

Задача о многоруком бандите (multi-armed bandit)

Имеется множество допустимых *действий* (ручек, arm),
с различными распределениями размера *премии* (reward, payoff).

Как быстрее найти самое выгодное действие?

Какие возможны стратегии?



Задача о многоруком бандите

A — множество возможных *действий*

$p(r|a)$ — неизвестное распределение *премии* $r \in \mathbb{R}$ для $a \in A$

$\pi_t(a)$ — *стратегия* агента в момент t , распределение на A

Игра агента со средой (multi-armed bandit):

- 1: инициализация стратегии $\pi_1(a)$
- 2: **для всех** $t = 1, \dots, T, \dots$
- 3: агент выбирает действие $a_t \sim \pi_t(a)$;
- 4: среда генерирует премию $r_t \sim p(r|a_t)$;
- 5: агент корректирует стратегию $\pi_{t+1}(a)$;

$$Q_t(a) = \frac{\sum_{i=1}^t r_i [a_i = a]}{\sum_{i=1}^t [a_i = a]} \quad \text{— средняя премия в } t \text{ раундах}$$

$$Q^*(a) = \lim_{t \rightarrow \infty} Q_t(a) \rightarrow \max_{a \in A} \quad \text{— ценность действия } a$$

Примеры прикладных задач

- Рекомендация новостных статей пользователям
- Показ рекламы в Интернете
- Управление технологическими процессами
- Управление роботами
- Управление ценами и ассортиментом в сетях продаж
- Игра на бирже
- Маршрутизация в телекоммуникационных сетях
- Маршрутизация в беспроводных сенсорных сетях
- Логические игры (шашки, нарды, и т.д.)

Задача о многоруком бандите впервые рассмотрена в статье
H. Robbins. Some aspects of the sequential design of experiments.
Bulletin of the American Mathematics Society, 58:527–535, 1952.

Жадная стратегия

Множество действий с максимальной текущей оценкой ценности:

$$A_t = \text{Arg max}_{a \in A} Q_t(a)$$

Жадная стратегия — выбрать любое действие из A_t :

$$\pi_t(a) = \frac{1}{|A_t|} [a \in A_t]$$

Недостаток жадной стратегии — по некоторым действиям a можем так и не набрать статистику для оценки $Q_t(a)$.

Компромисс «изучение–применение» (exploration–exploitation)
 ϵ -жадная стратегия:

$$\pi_t(a) = \frac{1 - \epsilon}{|A_t|} [a \in A_t] + \frac{\epsilon}{|A|}$$

Эвристика: параметр ϵ уменьшать со временем.

Стратегия softmax (распределение Гиббса)

Мягкий вариант компромисса «изучение–применение»:
чем больше $Q_t(a)$, тем больше вероятность выбора a :

$$\pi_t(a) = \frac{\exp\left(\frac{1}{\tau} Q_t(a)\right)}{\sum_{b \in A} \exp\left(\frac{1}{\tau} Q_t(b)\right)}$$

где τ — параметр *температуры*,
при $\tau \rightarrow 0$ стратегия стремится к жадной,
при $\tau \rightarrow \infty$ — к равномерной, т.е. чисто исследовательской

Эвристика: параметр τ уменьшать со временем.

Какая из стратегий лучше?

- зависит от конкретной задачи,
- решается в эксперименте

Метод UCB (upper confidence bound)

Выбор действия с максимальной верхней оценкой ценности:

$$A_t = \operatorname{Arg} \max_{a \in A} \left(Q_t(a) + \delta \sqrt{\frac{2 \ln t}{k_t(a)}} \right),$$

где $k_t(a) = \sum_{i=1}^t [a_i = a]$, δ — параметр expl/exp -компромисса.

Интерпретация:

чем меньше $k_t(a)$, тем менее исследована стратегия,
тем выше должна быть вероятность выбрать a ;

чем больше δ , тем стратегия более исследовательская.

Эвристика: параметр δ уменьшать со временем.

P. Auer, N. Cesa-Bianchi, P. Fischer. Finite-time analysis of the multiarmed bandit problem, Machine Learning, 2002.

Модельные эксперименты в обучении с подкреплением

«10-рукая испытательная среда»:

Генерируется 2000 задач, в каждой задаче

$$|A| = 10,$$

$$p(r|a) = \mathcal{N}(r; Q^*(a), 1),$$

$$Q^*(a) \sim \mathcal{N}(0, 1).$$

Строятся графики зависимости

— средней премии (average reward),

— доли оптимальных действий (% optimal action),

от числа шагов t , усреднённые по 2000 задачам.

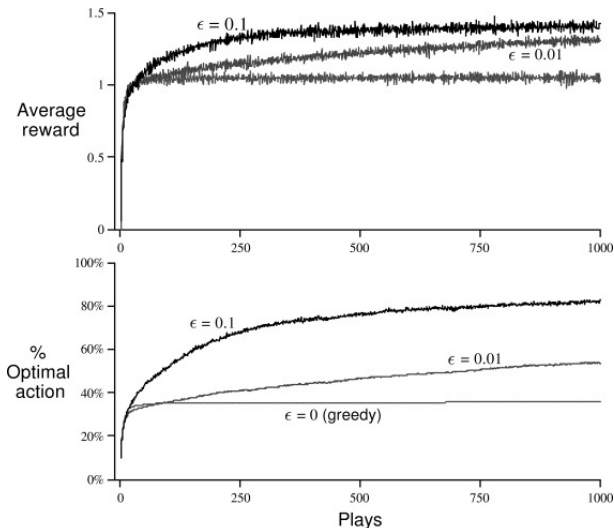
Richard Sutton, Andrew Barto. Reinforcement Learning: An Introduction.
The MIT Press. 1998, 2004.

<http://webdocs.cs.ualberta.ca/~sutton/book/ebook/the-book.html>

Русский перевод:

Р. Саттон, Э. Барто. Обучение с подкреплением. Изд-во «Бином». 2011.

Сравнение жадных и ϵ -жадных стратегий



Рекуррентная формула для эффективного вычисления средних

Общая формула вычисления Q_t для корректировки стратегии:

$$Q_{t+1}(a) = (1 - \alpha_t)Q_t(a) + \alpha_t r_{t+1} = Q_t(a) + \alpha_t (r_{t+1} - Q_t(a))$$

При $\alpha_t = \frac{1}{k_t(a)+1}$ это среднее арифметическое, $k_t(a) = \sum_{i=1}^t [a_i = a]$

При $\alpha_t = \text{const}$ это экспоненциальное скользящее среднее

Условие сходимости к среднему:

$$\sum_{t=1}^{\infty} \alpha_t = \infty, \quad \sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

Среднее арифметическое — для стационарных задач

Экспоненциальное скользящее среднее — для нестационарных
 (в этом случае сходимости нет, но она и не нужна)

Экспоненциальное скользящее среднее (напоминание)

Задача прогнозирования временного ряда y_0, \dots, y_t, \dots :

- простейшая регрессионная модель — константа $y_t = c$,
- наблюдения учитываются с весами, убывающими в прошлое,
- прогноз \hat{y}_{t+1} методом наименьших квадратов:

$$\sum_{i=0}^t w_{t-i} (y_i - c)^2 \rightarrow \min_c, \quad w_i = \beta^i, \quad \beta \in (0, 1)$$

Аналитическое решение — формула Надарая-Ватсона:

$$c \equiv \hat{y}_{t+1} = \frac{\sum_{i=0}^t \beta^i y_{t-i}}{\sum_{i=0}^t \beta^i}$$

Запишем аналогично \hat{y}_t , оценим $\sum_{i=0}^t \beta^i \approx \sum_{i=0}^{\infty} \beta^i = \frac{1}{1-\beta}$,

получим $\hat{y}_{t+1} = \hat{y}_t \beta + (1 - \beta) y_t$, заменим $\alpha = 1 - \beta$:

$$\hat{y}_{t+1} = (1 - \alpha) \hat{y}_t + \alpha y_t = \hat{y}_t + \alpha (y_t - \hat{y}_t)$$

Метод сравнения с подкреплением (reinforcement comparison)

Идея: использовать не сами значения премий, а их разности со средней (эталонной) премией:

$\bar{r}_{t+1} = \bar{r}_t + \alpha(r_t - \bar{r}_t)$ — средняя премия по всем действиям

$p_{t+1}(a_t) = p_t(a_t) + \beta(r_t - \bar{r}_t)$ — предпочтения действий

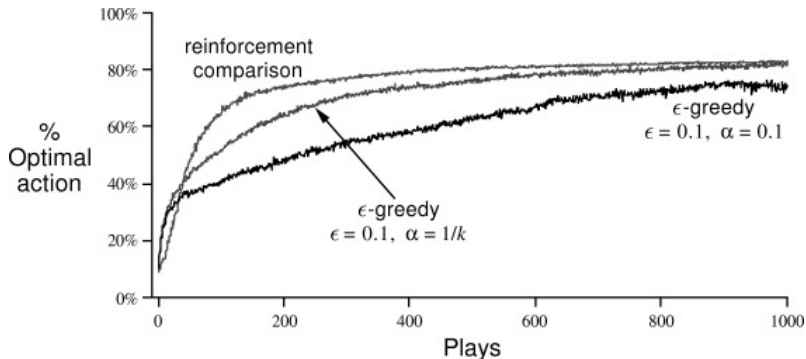
$\pi_{t+1}(a) = \frac{\exp(p_{t+1}(a))}{\sum_{b \in A} \exp(p_{t+1}(b))}$ — softmax-стратегия агента

Эвристика: оптимистично завышенное начальное \bar{r}_0 стимулирует изучающие действия в начале

Экспериментальный факт: сравнение с подкреплением сходится быстрее ε -жадных стратегий.

Сравнение с подкреплением лучше ϵ -жадных стратегий

Эксперимент с 10-рукой испытательной средой:



Richard Sutton, Andrew Barto. Reinforcement Learning: An Introduction. The MIT Press. 1998, 2004.

Р. Саттон, Э. Барто. Обучение с подкреплением. Изд-во «Бином». 2011.

Метод преследования (pursuit) жадной стратегии

Вместо собственно *жадной стратегии*

$$\pi_{t+1}(a) = \frac{[a \in A_t]}{|A_t|}$$

предлагается *преследование* (сглаживание) жадной стратегии:

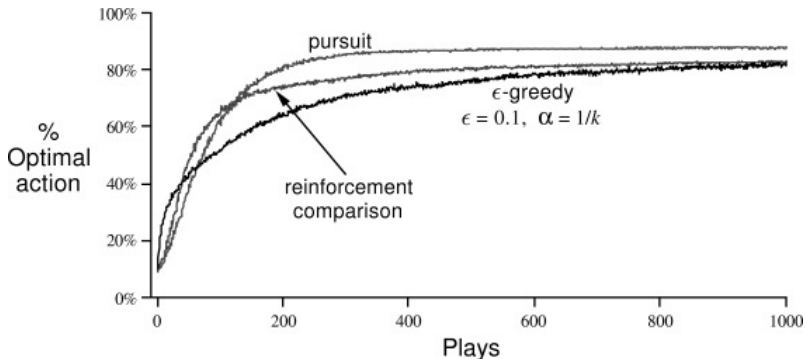
$$\pi_{t+1}(a) = \pi_t(a) + \beta \left(\frac{[a \in A_t]}{|A_t|} - \pi_t(a) \right)$$

Эвристика: начальное $\pi_0(a)$ можно взять равномерным.

Экспериментальный факт: метод преследования, сравнение с подкреплением и ϵ -жадные стратегии имеют каждый свою область применения.

Стратегия преследования ещё лучше

Эксперимент с 10-рукой испытательной средой:



Richard Sutton, Andrew Barto. Reinforcement Learning: An Introduction. The MIT Press. 1998, 2004.

Р. Саттон, Э. Барто. Обучение с подкреплением. Изд-во «Бином». 2011.

Постановка задачи в случае, когда имеется информация о среде

A — множество возможных *действий*

X — пространство контекстов, описаний состояния среды

$x_{ta} \in X$ — состояние среды в раунде t в случае выбора $a \in A$

$p(r | a, x)$ — неизвестное распределение премии $r \in \mathbb{R}$ для $a \in A$

$\pi_t(a | x)$ — стратегия агента в момент t , распределение на A

Игра агента со средой (contextual bandit):

- 1: инициализация стратегии $\pi_1(a)$
- 2: для всех $t = 1, \dots, T, \dots$
- 3: агенту сообщается контекст x_{ta} для всех $a \in A$;
- 4: агент выбирает действие $a_t \sim \pi_t(a | x_{ta})$;
- 5: среда генерирует премию $r_t \sim p(r | a_t, x_{ta})$;
- 6: агент корректирует стратегию $\pi_{t+1}(a | x)$;

Context-free bandit — когда $\pi_t(a | x) = \pi_t(a)$, т.е. не зависит от x

Регрессия с инкрементным обучением и доверительной оценкой

$r(a, x)$ — функция премии за действие a в контексте x ,
 $\hat{r}(a, x)$ — регрессионная оценка этой функции,
 $UCB(a, x)$ — верхняя оценка отклонения $\hat{r} - r$,
 δ — параметр (чем больше, тем больше exploration).

Игра агента со средой (contextual bandit):

- 1: инициализация стратегии $\pi_1(a)$
- 2: **для всех** $t = 1, \dots, T, \dots$
- 3: агенту сообщается контекст x_{ta} для всех $a \in A$;
- 4: агент выбирает действие
$$a_t = \arg \max_{a \in A} \left(\hat{r}(a, x_{ta}) + \delta UCB(a, x_{ta}) \right);$$
- 5: среда генерирует премию $r_t = r(a_t, x_{ta_t})$;
- 6: **регрессия $\hat{r}(a, x)$ дообучается на точке $(a_t, x_{ta_t}; r_t)$;**

Пример. Рекомендация новостных статей пользователям



Агент — рекомендательная система для персонализации показов новостных статей (Yahoo! Today).

F1..F4 — позиции для показа заголовков новостей.

A — новостные статьи, действия системы;

$x_{ta} \in X$ — признаковое описание пары (u_t, a) ;

u_t — пользователь, которому агент даёт рекомендацию;

$r_t \in \{0, 1\}$ — пользователь u_t кликнул на предложенную статью;

$Q_t(a)$ — средняя премия, CTR (click-through rate) статьи.

Цель — повышение среднего CTR и «счастья пользователя».

Lihong Li, Wei Chu, John Langford, Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. WWW-2010.

Линейная модель премий и гребневая регрессия

Пусть $x_{ta} \in X = \mathbb{R}^n$, $w \in \mathbb{R}^n$.

Линейная модель премий для действия $a \in A$:

$$Q^*(a) = E[r_t | x_{ta}] = \langle x_{ta}, w_a \rangle.$$

Гребневая регрессия: обучение w_a для действия a в момент t :

$$\sum_{i=1}^t [a_i = a] (\langle x_{ia}, w \rangle - r_i)^2 + \frac{\tau}{2} \|w\|^2 \rightarrow \min_w.$$

$w_a = (F_a^T F_a + \tau I_n)^{-1} F_a^T y_a$ — решение задачи МНК, где

$F_a = (x_{ia})_{i=1: a_i=a}^t$ — $\ell \times n$ -матрица объекты–признаки,

$y_a = (r_i)_{i=1: a_i=a}^t$ — $\ell \times 1$ -вектор ответов,

$\ell = k_t(a) = \sum_{i=1}^t [a_i = a]$ — объём обучающей выборки.

LinUCB: линейная модель с верхней доверительной оценкой

Доверительный интервал с коэффициентом доверия $1 - \alpha$ для линейной модели регрессии:

$$y = \langle x, w \rangle \pm \hat{\sigma} t_{\ell-n, 1-\frac{\alpha}{2}} \sqrt{x^\top (F^\top F)^{-1} x},$$

$t_{\ell-n, 1-\frac{\alpha}{2}}$ — квантиль распределения Стьюдента,
 $\hat{\sigma}^2 = \frac{1}{\ell-n} RSS$ — оценка дисперсии отклика y .

Стратегия выбора действия с максимальной верхней оценкой ценности UCB (upper confidence bound):

$$A_t = \text{Arg max}_{a \in A} \left(\langle x_{ta}, w_a \rangle + \delta \hat{\sigma} t_{\ell-n, 1-\frac{\alpha}{2}} \sqrt{x_{ta}^\top (F_a^\top F_a + \tau I_n)^{-1} x_{ta}} \right).$$

Чем больше параметр δ , тем больше исследования.

LinUCB: особенности реализации и обобщения

- *Инкрементный алгоритм* пересчёта w_a и матрицы $(F_a^T F_a + \tau I_n)^{-1}$ при добавлении каждой строки в F_a .
- Гибридная линейная модель $Q^*(a) = \langle z_{ta}, v \rangle + \langle x_{ta}, w_a \rangle$, где v — часть контекста, не зависящая от действия a .
- «Сырые признаки»:
пользователи: 12 соцдем, 200 география, ~ 1000 категорий,
статьи: ~ 100 категорий.
- Используется кластеризация и *понижение размерности*:
 $\dim w_a = 6$, $\dim v = 36$.
- Можно было бы использовать любую другую модель с инкрементным обучением и доверительными оценками.

Lihong Li, Wei Chu, John Langford, Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. WWW-2010.

Оценивание модели по историческим данным

Проблема off-line оценивания стратегии π :

исторические данные накоплены при использовании другой стратегии (logging policy) $\pi_0(a)$, отличной от π

Идея:

для оценивания $Q_t(a)$ отбираются только те события (x_{ta}, a, r_t) , для которых стратегии π и π_0 выбирали одинаковое действие:

$$a = \arg \max_a \pi(a, x_{ta}) = \arg \max_a \pi_0(a)$$

(для этого нужны очень большие данные)

Утв. Если $\pi_0(a)$ — равномерное распределение, то оценка $Q_t(a)$ по отобранной выборке является несмещённой.

Lihong Li, Wei Chu, John Langford, Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. WWW-2010.

Постановка задачи в случае, когда агент влияет на среду

A — множество возможных *действий*

S — множество состояний среды

Игра агента со средой:

- 1: инициализация стратегии $\pi_1(a | s)$ и состояния среды s_1
- 2: для всех $t = 1, \dots, T, \dots$
- 3: агент выбирает действие $a_t \sim \pi_t(a | s_t)$;
- 4: среда генерирует премию $r_{t+1} \sim p(r | a_t, s_t)$
и новое состояние $s_{t+1} \sim p(s | a_t, s_t)$;
- 5: агент корректирует стратегию $\pi_{t+1}(a | s)$;

Это *марковский процесс принятия решений* (МППР), если

$$\begin{aligned} P(s_{t+1}, r_{t+1} | s_t, a_t, r_t, s_{t-1}, a_{t-1}, r_{t-1}, \dots, s_1, a_1) = \\ = P(s_{t+1}, r_{t+1} | s_t, a_t) \end{aligned}$$

МППР называется *финитным*, если $|A| < \infty$, $|S| < \infty$.

Выгода. Ценность состояния. Ценность действия

$R_t = r_{t+1} + r_{t+2} + \dots + r_{t+k} + \dots$ — суммарная выгода

Обобщение — дисконтированная выгода:

$$R_t = r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{k-1} r_{t+k} + \dots$$

$\gamma \in [0, 1]$ — коэффициент дисконтирования:

чем выше γ , тем более агент дальновидный

Функция ценности состояния s при стратегии π :

$$V^\pi(s) = E_\pi(R_t | s_t = s) = E_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right)$$

Функция ценности действия a в состоянии s при стратегии π :

$$Q^\pi(s, a) = E_\pi(R_t | s_t = s, a_t = a) = E_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right)$$

E_π — мат.ожидание при условии, что агент следует стратегии π

Рекуррентные формулы для функций ценности

Рекуррентная формула для ценности состояния $V^\pi(s)$:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi \left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right) = \\ &= \mathbb{E}_\pi \left(r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right) = \\ &= \mathbb{E}_\pi \left(r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s \right) \end{aligned}$$

Рекуррентная формула для ценности действия $Q^\pi(s, a)$:

$$\begin{aligned} Q^\pi(s, a) &= \mathbb{E}_\pi \left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a \right) = \\ &= \mathbb{E}_\pi \left(r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s, a_t = a \right) = \\ &= \mathbb{E}_\pi \left(r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1}) \mid s_t = s, a_t = a \right) \end{aligned}$$

Жадные стратегии максимизации ценности

$V^*(s)$, $Q^*(s, a)$ — оптимальные функции ценности.

Уравнения оптимальности Беллмана:

$$V^*(s) = \max_{a \in A} E_{\pi}(r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a)$$

$$Q^*(s, a) = E_{\pi}(r_{t+1} + \gamma \max_{a' \in A} Q^*(s_{t+1}, a') \mid s_t = s, a_t = a)$$

Жадные стратегии π относительно $V^*(s)$ или $Q^*(s, a)$:
выбирать то действие, на котором достигается максимум
в уравнениях оптимальности Беллмана:

$$A_t = \text{Arg max}_{a \in A} E_{\pi}(r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t, a)$$

$$A_t = \text{Arg max}_{a \in A} Q^*(s_t, a)$$

Утв. Эти стратегии являются оптимальными.

Метод временных разностей TD(0)

Рекуррентная формула для ценности состояния $V^\pi(s)$:

$$V^\pi(s) = E_\pi(r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s)$$

Нужна эмпирическая оценка математического ожидания E_π .

Метод временных разностей TD (temporal difference)

После того, как выбрано a_t и стали известны r_{t+1} , s_{t+1} , оцениваем $V^\pi(s_t)$ экспоненциальным скользящим средним:

$$V(s_t) := V(s_t) + \alpha_t(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

Утв. Если α_t уменьшается ($\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$), и все s посещаются бесконечное число раз, то $V(s) \xrightarrow{\text{п.п.}} V^\pi(s)$, $t \rightarrow \infty$

Метод SARSA (state–action–reward–state–action)

Рекуррентная формула для ценности действия $Q^\pi(s, a)$:

$$Q^\pi(s, a) = E_\pi(r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1}) \mid s_t = s, a_t = a)$$

Игра агента со средой:

- 1: инициализация стратегии $\pi_1(a \mid s)$ и состояния среды s_1
- 2: **для всех** $t = 1, \dots, T, \dots$
- 3: агент выбирает действие $a_t \sim \pi_t(a \mid s_t)$:
 $a_t = \arg \max_a Q(s_t, a)$ — жадная стратегия
 (но возможны и другие: ϵ -жадная, по Гиббсу, ...)
- 4: среда генерирует $r_{t+1} \sim p(r \mid a_t, s_t)$ и $s_{t+1} \sim p(s \mid a_t, s_t)$;
- 5: агент разыгрывает ещё один шаг: $a' \sim \pi_t(a \mid s_{t+1})$;
- 6: $Q(s_t, a_t) := Q(s_t, a_t) + \alpha_t (r_{t+1} + \gamma Q(s_{t+1}, a') - Q(s_t, a_t))$;

Метод Q-обучения

Оптимальную функцию ценности действия

$$Q^*(s, a) = E_{\pi}(r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') \mid s_t = s, a_t = a)$$

аппроксимируем экспоненциальным скользящим средним:

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha_t (r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t))$$

Утв. Если α_t уменьшается ($\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$), и все s посещаются бесконечное число раз, то $Q \xrightarrow{\text{п.н.}} Q^*$, $t \rightarrow \infty$

Отличия от SARSA: выбрасывается шаг 5 и меняется шаг 6.

Многошаговое TD-прогнозирование

Хотелось бы иметь более надёжную оценку $V(s)$ или $Q(s, a)$, приближающуюся к дисконтированной выгоде R_t

$$R_t^{(1)} = r_{t+1} + \gamma V(s_{t+1})$$

$$R_t^{(2)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 V(s_{t+2})$$

...

$$R_t^{(n)} = r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{n-1} r_{t+n} + \gamma^n V(s_{t+n})$$

$$R_t = r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{n-1} r_{t+n} + \dots$$

Премии r_{t+2}, r_{t+3}, \dots в момент t неизвестны, но, оказывается, можно усреднять прошлые, а не будущие наблюдения, и асимптотически это приводит к тому же результату!

Метод временных разностей TD(λ)

Идея «следов приемлемости» $e(s)$ (eligibility traces)
 будем корректировать $V(s)$ не только текущего s_t , но и недавно пройденных состояний, с коэффициентом затухания $\lambda \in [0, 1]$

Обновление $V(s)$ теперь не только для $s = s_t$:

- 1: $e(s_t) := e(s_t) + 1$;
- 2: для всех $s \in S$, $e(s) \neq 0$
- 3: $V(s) := V(s) + e(s) \cdot \alpha_t (r_{t+1} + \gamma V(s_{t+1}) - V(s))$;
- 4: $e(s) := \gamma \lambda e(s)$;

Возможны варианты обновления следов приемлемости:

$e(s) := [s = s_t]$ — получаем метод TD(0)

$e(s) := \min\{\gamma \lambda e(s), 1\}$ — «заметаящий след»

$e(s) := (e(s) < \varepsilon) ? 0 : e(s)$ — обнуление слишком старых следов

При $\lambda = 0$ имеем TD(0), при $\lambda = 1$ приближаемся к оценке R_t

Методы SARSA(λ) и Q(λ)

Идея следов приемлемости легко переносится на метод SARSA:

Обновление $Q(s, a)$ теперь не только для $s = s_t$:

- 1: $e(s_t, a_t) := e(s_t, a_t) + 1$;
- 2: **для всех** $s \in S, a \in A$: $e(s, a) \neq 0$
- 3: $Q(s, a) := Q(s, a) + e(s, a) \cdot \alpha_t (r_{t+1} + \gamma Q(s_{t+1}, a') - Q(s, a))$;
- 4: $e(s, a) := \gamma \lambda e(s, a)$;

... и на Q-обучение, если положить
 $a' := \arg \max_a Q(s_{t+1}, a)$;

Важная деталь: исследовательские действия должны прерывать следы приемлемости, иначе будут строиться неверные оценки оптимальной стратегии.

- В обучении с подкреплением нет ответов учителя, есть только ответная реакция среды
- В контекстных бандитах используются модели машинного обучения, удовлетворяющие двум требованиям:
 - существует эффективный инкрементный метод обучения
 - существуют доверительные оценки средней премии $Q^t(a)$
- В марковских процессах принятия решений накапливается информация о ценности отдельных состояний и действий
- Компромисс «изучение–применение» при любом обучении с подкреплением подбирается экспериментальным путём
- Объём исследовательских действий приходится уменьшать в случае конечного горизонта игры