

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)»

ноябрь 2012

Содержание

- 1 Многорукий бандит**
 - Простая постановка задачи
 - Жадные и полужадные стратегии
 - Адаптивные стратегии
- 2 Динамическое программирование**
 - Полная постановка задачи
 - Уравнения Беллмана
 - Динамическое программирование
- 3 Метод временных разностей**
 - Методы TD(0), SARSA, Q-обучения
 - Методы TD(λ), SARSA(λ), Q(λ)
 - Метод VDBE

Задача о многоруком бандите

A — множество возможных *действий*

$p_a(r)$ — неизвестное распределение *премии* $r \in \mathbb{R}$ за $\forall a \in A$

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

Игра агента со средой:

- 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 $r_t \sim p_{a_t}(r)$;
- 5: агент корректирует стратегию $\pi_{t+1}(a)$;

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

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

Жадная стратегия

Множество действий с максимальной текущей оценкой ценности:

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

Жадная стратегия — выбирать любое действие из A_t :

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

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

ε -жадная стратегия (компромисс «изучение—применение»):

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

Эвристика: параметр ε имеет смысл уменьшать со временем.

Стратегия softmax (распределение Гиббса)

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

$$\pi_{t+1}(a) = \frac{\exp(Q_t(a)/\tau)}{\sum_{b \in A} \exp(Q_t(b)/\tau)}$$

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 имеет смысл уменьшать со временем.

Какая из стратегий лучше?

- зависит от конкретной задачи,
- решается в эксперименте

Модельные эксперименты в обучении с подкреплением

«10-рукая испытательная среда»:

Генерируется 2000 задач, в каждой задаче

$$|A| = 10,$$

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

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

Строятся графики зависимости

— среднего вознаграждения (average reward),

— доли оптимальных действий (% optimal action),

от числа действий (сыгранных игр),

усреднённые по 2000 задачам.

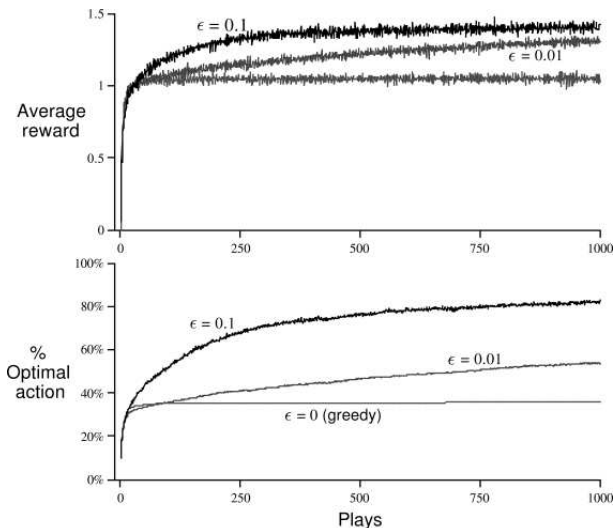
Richard Sutton, Andrew Barto. Reinforcement Learning: An Introduction. The MIT Press. 1998, 2004.

<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:

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

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

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

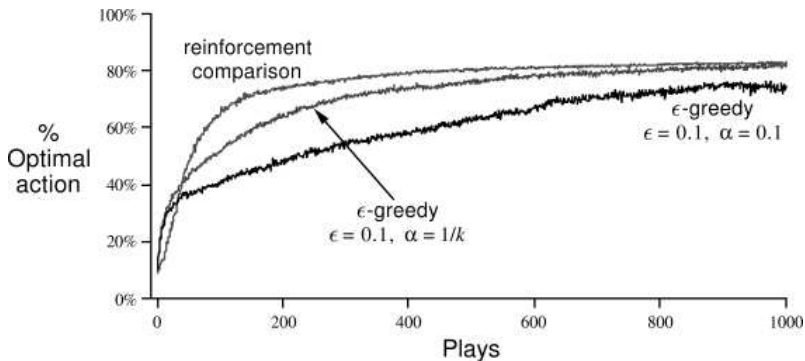
Начальное приближение r_0 : оптимистично завышенное стимулирует изучающие действия в начале

Экспериментальный факт:

метод сравнения с подкреплением обычно сходится быстрее.

Сравнение с подкреплением лучше ϵ -жадных стратегий

Эксперимент с 10-рукой испытательной средой:



Метод преследования (pursuit) жадной стратегии

Вместо собственно *жадной стратегии*

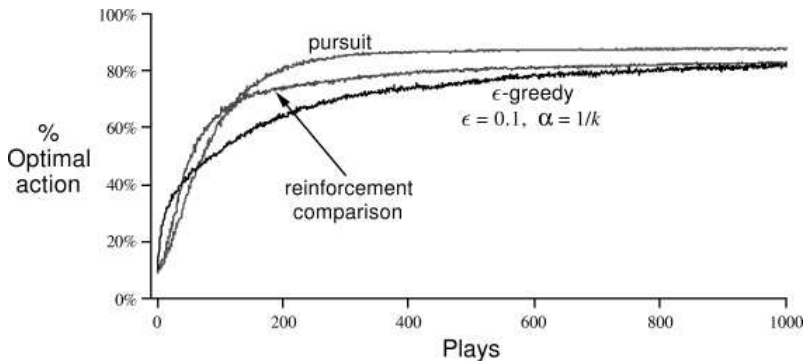
$$\pi_{t+1}(a) = \frac{[a \in A_t]}{|A_t|}$$

предлагается *преследование* (сглаживание) жадной стратегии:

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

Стратегия преследования ещё лучше

Эксперимент с 10-рукой испытательной средой:



Полная постановка задачи обучения с подкреплением

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

$$\begin{aligned} P(s_{t+1} = s', r_{t+1} = r \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, r_{t-1}, \dots, s_1, a_1) &= \\ = P(s_{t+1} = s', r_{t+1} = r \mid s_t, a_t) \end{aligned}$$

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

Выгода. Ценность состояния. Ценность действия

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

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

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

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

Уравнения Беллмана для V^π

Пусть имеется полная информация о среде (что не реально):

$\mathcal{P}_{ss'}^a = p(s_{t+1} = s' | s_t = s, a_t = a)$ — вероятности переходов $s \xrightarrow{a} s'$

$\mathcal{R}_{ss'}^a = E(r_{t+1} | s_t = s, a_t = a, s_{t+1} = s')$ — ожидаемые премии

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

$$(1) = \sum_a \sum_{s'} \mathcal{R}_{ss'}^a \mathcal{P}_{ss'}^a \pi(a|s)$$

$$(2) = \sum_a \sum_{s'} E_\pi(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s') \mathcal{P}_{ss'}^a \pi(a|s)$$

Отсюда уравнение Беллмана для ценности состояний V^π :

$$V^\pi(s) = \sum_a \pi(a|s) \sum_{s'} \mathcal{P}_{ss'}^a (\mathcal{R}_{ss'}^a + \gamma V^\pi(s'))$$

это система $|S|$ уравнений с $|S|$ неизвестными

Уравнения Беллмана для Q^π

Пусть имеется полная информация о среде (что не реально):

$\mathcal{P}_{ss'}^a = p(s_{t+1} = s' | s_t = s, a_t = a)$ — вероятности переходов $s \xrightarrow{a} s'$

$\mathcal{R}_{ss'}^a = E(r_{t+1} | s_t = s, a_t = a, s_{t+1} = s')$ — ожидаемые премии

$$Q^\pi(s, a) = E_\pi(r_{t+1} + \gamma \sum_k \gamma^k r_{t+k+2} | s_t = s, a_t = a) =$$

$$= \underbrace{E_\pi(r_{t+1} | s_t = s, a_t = a)}_{(1)} + \gamma \underbrace{E_\pi(\sum_k \gamma^k r_{t+k+2} | s_t = s, a_t = a)}_{(2)}$$

$$(1) = \sum_{s'} \mathcal{R}_{ss'}^a \mathcal{P}_{ss'}^a$$

$$(2) = \sum_{s'} E_\pi(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_{t+1} = s', a_t = a) \mathcal{P}_{ss'}^a.$$

Отсюда уравнение Беллмана для ценности действий Q^π :

$$Q^\pi(s, a) = \sum_{s'} \mathcal{P}_{ss'}^a (\mathcal{R}_{ss'}^a + \gamma Q^\pi(s', a))$$

это также система $|S|$ уравнений с $|S|$ неизвестными

Уравнения оптимальности Беллмана

Вводится отношение частичного порядка на стратегиях:

$$\pi \leq \pi' \Leftrightarrow \forall s \in S \quad V^\pi(s) \leq V^{\pi'}(s)$$

$\forall \pi \quad \pi^* \not\leq \pi$ — оптимальная стратегия π^*

$V^*(s) = V^{\pi^*}(s)$ — оптимальная функция ценности состояния

$Q^*(s, a) = Q^{\pi^*}(s, a)$ — оптимальная функция ценности действия

Уравнения оптимальности Беллмана (две формы записи):

$$\left\{ \begin{array}{l} V^*(s) = \max_{a \in A} E(r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a) \\ Q^*(s, a) = E(r_{t+1} + \gamma \max_{a' \in A} Q^*(s_{t+1}, a') \mid s_t = s, a_t = a) \end{array} \right.$$

$$\left\{ \begin{array}{l} V^*(s) = \max_{a \in A} \sum_{s' \in S} \mathcal{P}_{ss'}^a (\mathcal{R}_{ss'}^a + \gamma V^*(s')) \\ Q^*(s, a) = \sum_{s' \in S} \mathcal{P}_{ss'}^a (\mathcal{R}_{ss'}^a + \gamma \max_{a' \in A} Q^*(s', a)) \end{array} \right.$$

Свойства оптимальных стратегий

Утв. 1. Для финитного МППР оптимальное решение $V^*(s)$, $Q^*(s, a)$ единственно.

Жадная по отношению к $V^*(s)$, $Q^*(s, a)$ стратегия π :
 выбирать то действие, на котором достигается максимум
 в уравнениях оптимальности Беллмана:

$$\pi(a|s) > 0 \Leftrightarrow V^*(s) = E(r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a)$$

$$\pi(a|s) > 0 \Leftrightarrow Q^*(s, a) = \max_{a'} Q^*(s, a')$$

Утв. 2. Жадная стратегия является оптимальной ($\pi = \pi^*$)

Проблема 1: как одновременно и решить систему, и определить оптимальную стратегию?

Проблема 2: как оценить правые части в условиях игры, когда $\mathcal{P}_{ss'}^a$, $\mathcal{R}_{ss'}^a$ не известны?

Динамическое программирование

Рассматриваем детерминированные стратегии $\pi: S \rightarrow A$

Теорема об улучшении стратегии.

Если π, π' таковы, что $\forall s \in S \quad Q^\pi(s, \pi'(s)) \geq V^\pi(s)$,
 то $\pi \leq \pi'$, т.е. $\forall s \in S \quad V^\pi(s) \leq V^{\pi'}(s)$;

Если в состоянии s выбрать a , затем следовать стратегии π , то

$$\begin{aligned} Q^\pi(s, a) &= E_\pi(r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s, a_t = a) \\ &= \sum_{s' \in S} \mathcal{P}_{ss'}^a (\mathcal{R}_{ss'}^a + \gamma V^\pi(s')) \end{aligned}$$

Жадное улучшение стратегии:

$$\pi(s) = \arg \max_a \sum_{s' \in S} \mathcal{P}_{ss'}^a (\mathcal{R}_{ss'}^a + \gamma V^\pi(s'))$$

Утв. Жадная стратегия сходится к оптимальной

Метод итераций по ценностям и стратегиям

Есть два вида шагов, которые можно по-разному сочетать:

1 Улучшение стратегии π

при фиксированной функции ценности V , в состоянии s :

$$\pi(s) := \arg \max_a \sum_{s' \in S} \mathcal{P}_{ss'}^a (\mathcal{R}_{ss'}^a + \gamma V(s'))$$

2 Один шаг метода простых итераций

для системы уравнений оптимальности Беллмана
 при фиксированной стратегии π , в состоянии s :

$$V(s) := \sum_{s' \in S} \mathcal{P}_{ss'}^{\pi(s)} (\mathcal{R}_{ss'}^{\pi(s)} + \gamma V(s'))$$

Эти шаги можно чередовать в произвольном порядке,
 (сходимость к оптимальной стратегии гарантируется)
 например, выполнять после каждого попадания в состояние s

Метод временных разностей TD(0)

Проблема 2: $\mathcal{P}_{ss'}^a$, $\mathcal{R}_{ss'}^a$ не известны, их надо оценивать из предшествующего опыта игры

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

Метод временных разностей TD (temporal difference)

После того, как выбрано a_t и стали известны r_{t+1} , s_{t+1} , оцениваем $V^\pi(s)$ 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 $V(s) \xrightarrow{\text{п.н.}} V^\pi(s)$, $t \rightarrow \infty$

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

Аналогично TD(0), но для функции ценности действий:

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

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 $a_t \sim \pi_t(a|s_t)$:
 $a_t = \arg \max_a Q(s_t, a)$ — жадная стратегия
 (но возможны и другие: ϵ -жадная, по Гиббсу, ...)
- 4: среда генерирует $r_{t+1} \sim p(r|a_t, s_t)$ и $s_{t+1} \sim p(s|a_t, s_t)$;
- 5: агент разыгрывает ещё один шаг: $a' \sim \pi_t(a|s_{t+1})$;
- 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

Аппроксимируем оптимальную функцию ценности действия:

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

Оценка $Q^*(s, a)$ 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 $Q \xrightarrow{\text{п.н.}} Q^*$, $t \rightarrow \infty$

В игре меняется только шаг b .

Многошаговое TD-прогнозирование

Хотелось бы иметь более надёжную оценку $V(s)$ или $Q(s, a)$ по большему числу шагов

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

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

...

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

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

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» $e(s)$:

будем корректировать $V(s)$ не только текущего s_t , но и недавно пройденных состояний, с коэффициентом затухания $\lambda \in [0, 1]$

Раньше делали так:

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

Теперь делаем так:

$$e(s_t) := e(s_t) + 1;$$

$$V(s) := V(s) + \alpha_t (r_{t+1} + \gamma V(s_{t+1}) - V(s)) e(s), \quad \forall s \in S: e(s) \neq 0;$$

$$e(s) := \gamma \lambda e(s), \quad \forall s \in S: e(s) \neq 0;$$

Возможны варианты обновления следов приемлемости:

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

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

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

Метод SARSA(λ)

Идея следов приемлемости легко переносится на метод SARSA:

Раньше делали так:

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

Теперь делаем так:

$$e(s_t, a_t) := e(s_t, a_t) + 1;$$

для всех $s \in S, a \in A: e(s, a) \neq 0$

$$Q(s, a) := Q(s, a) + \alpha_t (r_{t+1} + \gamma Q(s_{t+1}, a') - Q(s, a)) e(s, a);$$

$$e(s, a) := \gamma \lambda e(s, a);$$

Метод Q(λ)

Идея следов приемлемости легко переносится на Q-обучение:

Раньше делали так:

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha_t (r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t));$$

Теперь делаем так:

$$e(s_t, a_t) := e(s_t, a_t) + 1;$$

для всех $s \in S$, $a \in A$: $e(s, a) \neq 0$

$$Q(s, a) := Q(s, a) + \alpha_t (r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s, a)) e(s, a);$$

$$e(s, a) := \gamma \lambda e(s, a);$$

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

Адаптивный ε -жадный метод временных разностей

Идея: чем сильнее колебания (дисперсия) $Q_t(s, a)$, тем больше должна быть вероятность ε_t исследовательских действий.

$$f(s, a) = \left| \frac{\exp(Q_t(s, a)/\sigma) - \exp(Q_{t+1}(s, a)/\sigma)}{\exp(Q_t(s, a)/\sigma) + \exp(Q_{t+1}(s, a)/\sigma)} \right|$$

$$\varepsilon_{t+1}(s) = \varepsilon_t(s) + \delta(f(s_t, a_t) - \varepsilon_t(s))$$

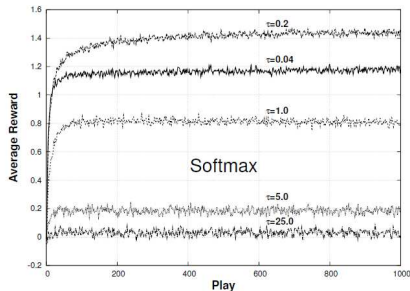
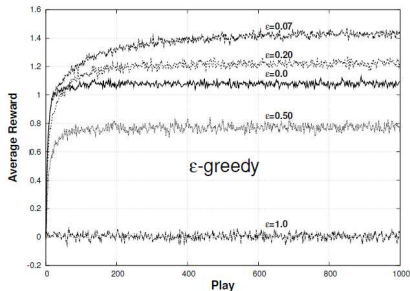
Рекомендации:

$\delta = 1/|A(s)|$, $A(s)$ — число возможных действий в состоянии s
 σ — обратная чувствительность (inverse sensitivity),
 при $\sigma \rightarrow 0$ — чисто исследовательская стратегия

Инициализация: $\varepsilon_1(t) \equiv 1$ — чисто исследовательская стратегия

Michel Tokic. Adaptive ε -greedy exploration in reinforcement learning based on value differences // 33rd German conference on Advances in artificial intelligence. 2010. Pp.203–210.

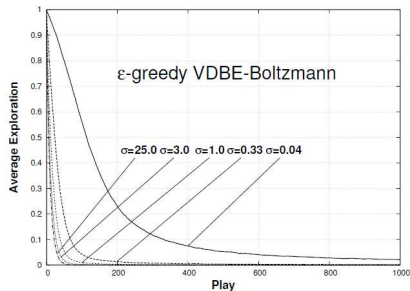
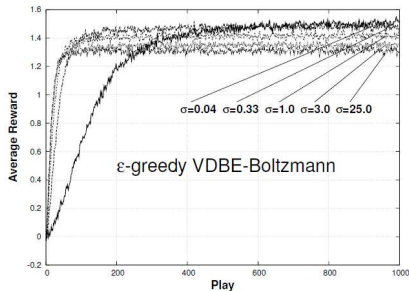
ϵ -жадные стратегии и softmax



ϵ -жадные стратегии и softmax

чувствительны к выбору параметров ϵ и τ соответственно

Адаптивный ϵ -жадный метод VDBE



Метод VDBE (value differences based exploration)

- обгоняет ϵ -жадные стратегии и softmax;
- постепенно уменьшает долю исследований ϵ ;
- может легко сочетаться с другими методами

Резюме в конце лекции

- В обучении с подкреплением нет ответов учителя, есть только ответная реакция среды
- Задача о многоруком бандите — это простой случай среды с одним состоянием
- Методы динамического программирования хорошо обоснованы, но требуют полного знания среды $\mathcal{P}_{ss'}^a, \mathcal{R}_{ss'}^a$
- Методы временных разностей лишь косвенно опираются на уравнения Беллмана (Q-метод) и основаны на оценках экспоненциальных скользящих средних