Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Обучение с подкреплением

Лекция #1

Введение. Табличные методы решения.

Санкт-Петербург, 2022

Введение

Общие понятия

Обучение с подкреплением – метод, который стремится отобразить ситуации на действия, чтобы максимизировать численный сигнал – вознаграждение. Для этого агенту необходимо понять, какие действия приносят максимальное вознаграждение, перебирая различные варианты. При этом действия могут влиять не только на непосредственное вознаграждение, но и на следующие ситуации. Эти две характеристики – поиск методом «поиск методом проб и ошибок» и отложенное вознаграждение – являются наиболее важными отличительными чертами обучения с подкреплением.

Обучающийся агент должен уметь в какой-то степени воспринимать состояние среды и предпринимать действия, изменяющие это состояние. Агента есть цель или несколько целей, связанных со стремлением изменять состояние окружающей среды. Любой метод, подходящий для решения таких задач, будет рассматриваться нами как метод обучения с подкреплением.

Обучение с подкреплением

Общие понятия

Отличие от обучения с учителем

Обучение с подкреплением отличается от *обучения с учителем*. В случае обучения с учителем имеется обучающий набор помеченных примеров, подготовленный квалифицированным внешним учителем. Каждый пример представляет собой описание ситуации и спецификацию – метку – правильного действия, которое система должна предпринять в этой ситуации.

Цель такого обучения – добиться, чтобы система смогла обобщить, свою реакцию на ситуации, которые не были предъявлены в обучающем наборе. Это важный вид обучения, но, взятый сам по себе, он не подходит для обучения с помощью взаимодействия. В интерактивных задачах часто практически невозможно получить примеры желаемого поведения, которые правильно представляли бы все ситуации, в которых агенту предстоит действовать. На неизведанной территории – там, где от обучения как раз и ожидают плодов, – агент должен уметь действовать, исходя из своего опыта.

Отличие от обучения без учителя

Обучение с подкреплением отличается и от *обучения без учителя, которое обычно имеет целью обнаружение структуры, скрытой в наборе* непомеченных данных. Может возникнуть соблазнительная мысль о том, что обучение с подкреплением – разновидность обучения без учителя, поскольку отсутствуют примеры правильного поведения. В действительности цель обучения с подкреплением – максимизировать вознаграждение, а не выявить скрытую структуру.

Выявление структуры в опыте агента может быть полезно, но само по себе не решает задачу максимизации вознаграждения, стоящую перед обучением с подкреплением.

Исследование и использование

Нахождение компромисса между исследованием и использованием, является одной из задач обучения с подкреплением. Чтобы получить большое вознаграждение, обучающийся с подкреплением агент должен предпочитать действия, которые были испробованы в прошлом и принесли вознаграждение. Но чтобы найти такие действия, он должен пробовать действия, которые раньше не выбирал. Агент должен *использовать уже приобретенный опыт, чтобы получить вознаграждение, но должен про*должать *исследования, чтобы выбирать более эффективные действия в будущем.*

Дилемма состоит в том, что одного лишь исследования или использования недостаточно для успешного решения задачи. Агент должен пробовать разные действия и неуклонно отдавать предпочтение тем, которые кажутся наилучшими. При этом каждое действие необходимо испробовать много раз, чтобы получить надежную оценку ожидаемого вознаграждения.

Исследование и использование. Продолжение

Важной чертой обучения с подкреплением – явное рассмотрение целостной проблемы целеустремленного агента, взаимодействующего с неопределенной окружающей средой. При этом планирование действий является очень важной частью данного подхода к обучению.

У всех обучающихся с подкреплением агентов имеются явные цели, выбирать действия, оказывающие влияние на среду. Агент должен действовать, невзирая на значительную неопределенность окружающей его среды. Поэтому если обучение с подкреплением включает планирование, то оно должно учитывать взаимное влияние планирования и выбора действий в реальном времени, а также ответить на вопрос о том, откуда поступают и как совершенствуются модели.

Элементы обучения с подкреплением. Стратегия

Стратегия определяет, как обучающийся агент поведет себе в данный момент времени. Она соответствует тому, что в психологии называется множеством правил, или ассоциаций, стимул–реакция.

Стратегия лежит в основе обучающегося с подкреплением агента, поскольку ее одной достаточно для определения поведения. В общем случае стратегии могут быть стохастическими, т. е. задавать вероятности каждого действия.

Элементы обучения с подкреплением. Сигнал вознаграждения

Сигнал вознаграждения определяет цель в задаче обучения с подкреплением. На каждом временном шаге среда посылает обучающемуся агенту одно число, называемое *вознаграждением.*

*Единственное стремление агента – максимизиро*вать полное вознаграждение, полученное в течение длительного времени работы.

Вознаграждения – прямые и определяющие характеристики проблемы, стоящей перед агентом.

Сигнал вознаграждения – главная причина изменения стратегии; если выбранное стратегией действие влечет низкое вознаграждение, то стратегию следует изменить, так чтобы в будущем в такой ситуации выбиралось другое действие.

Элементы обучения с подкреплением. Функция ценности

Если сигнал вознаграждения показывает, что хорошо прямо сейчас, то *функция ценности говорит, что хорошо в длительной перспективе.*

Вознаграждение определяет непосредственную внутренне присущую желательность состояний окружающей среды, а ценность – долговременную желательность состояний с учетом тех состояний, которые с большой вероятностью встретятся позже, и вознаграждений в этих состояниях.

Например, состояние может всегда приносить низкое немедленное вознаграждение, но при этом иметь высокую ценность, поскольку за ним регулярно следуют состояния, приносящие высокое вознаграждение.

Элементы обучения с подкреплением. Модель

Модель *имитирует поведение окружающей* среды. Модели используются для *планирования, под которым понимается любой способ выбора порядка действий* путем рассмотрения возможных будущих ситуаций, до того как они фактически произошли.

Методы решения задач обучения с подкреплением, в которых используются модели и планирование, называются *основанными на модели*.

Современное обучение с подкреплением охватывает весь спектр систем – от низкоуровневого обучения методом проб и ошибок до высокоуровневого обоснованного планирования.

Введение. Заключение.

Обучение с подкреплением – вычислительный подход к пониманию и автоматизации обучения и принятия решений, направляемых стремлением к достижению цели. Оно отличается от других вычислительных подходов упором на обучение агента в процессе прямого взаимодействия с окружающей средой, без посредничества учителя и без полной модели среды.

Табличные методы решения. Многорукие бандиты.

Задача о *k-руком бандите. Часть 1*

Рассмотрим следующую задачу обучения. Агент многократно стоит перед выбором из *k разных вариантов (действий). После каждого выбора получает чис*ленное вознаграждение, выбираемое из стационарного распределения вероятностей, которое зависит от выбранного действия. Цель – максимизировать ожидаемое полное вознаграждение за определённый период (k шагов).

Это исходная постановка задачи о *k-руком бандите, названной так по аналогии* с игровым автоматом, или «одноруким бандитом». Только в этом случае рычагов не один, а *k. Каждый выбор действия соответствует опусканию одного из рыча*гов игрового автомата, а вознаграждения – это выплаты за выпадение джекпота.

Задача о *k-руком бандите. Часть 2*

В задаче о k-руком бандите с каждым из k действий связано ожидаемое, вознаграждение при условии выбора этого действия; будем называть его ценностью действия. Обозначим A_t действие, выбранное на временном шаге t, а R_t – соответствующее ему вознаграждение. Тогда ценность произвольного действия a, обозначаемая q*(a), – это математическое ожидание вознаграждения при условии выбора a:

$q^*(a) = \mathbb{E}[R_t \mid A_t = a]$.

Если бы мы знали ценность каждого действия, то задача о k-руком бандите решалась бы тривиально. Предположим, что достоверно ценности действий неизвестны, но имеются оценки. Обозначим $Q_t(a)$ оценку ценности действия a на временном шаге t. Хотелось бы, чтобы $Q_t(a)$ было близко к q*(a).

Задача о *k-руком бандите. Часть 3*

Если оценки ценности действий запоминать, то на любом временном шаге имеется по крайней мере одно действие с максимальной оценкой. Назовем эти действия *жадными. Если же выбирается какое-то нежадное* действие, то выполняется *исследование*. Использование дает возможность максимизировать ожидаемое вознаграждение на одном шаге, а исследование может дать большее суммарное вознаграждение в длительной перспективе.

Если впереди много временных шагов, на которых можно выбирать действия, то может оказаться выгоднее исследовать нежадные действия и выяснить, какие из них лучше жадных. В краткосрочной перспективе вознаграждение во время исследования будет ниже, но в долгосрочной – выше, потому что, обнаружив лучшее действие, возможно использовать его много раз. Поскольку при выборе одного действия нельзя одновременно исследовать и использовать, получается «конфликт» между исследованием и использованием.

Методы ценности действий. Часть 1

Методы оценивания ценности действий и использования оценок при решении о выборе действия назовём *методами ценности действий. Ценность действия* – это среднее вознаграждение при условии выбора этого действия. Эту величину можно естественно оценить, усреднив фактически полученные вознаграждения:

$$Q_t(a) = (\sum R_i * \mathbb{1}(A_i = a))/ (\sum \mathbb{1} (A_i = a))$$

где $\mathbb{1}_{predicate}$ *обозначает случайную величину, равную 1, если предикат predicate ра*вен true, и 0 в противном случае. Если знаменатель равен 0, то в качестве *Qt(a)* принимается какое-нибудь значение по умолчанию, например 0. Когда знаменатель стремится к бесконечности, *Qt(a), по закону больших чисел, сходится к q*(a).* Назовем это методом *выборочного среднего для оценки ценности значений, по*скольку каждая оценка является средним по выборке релевантных действию вознаграждений.

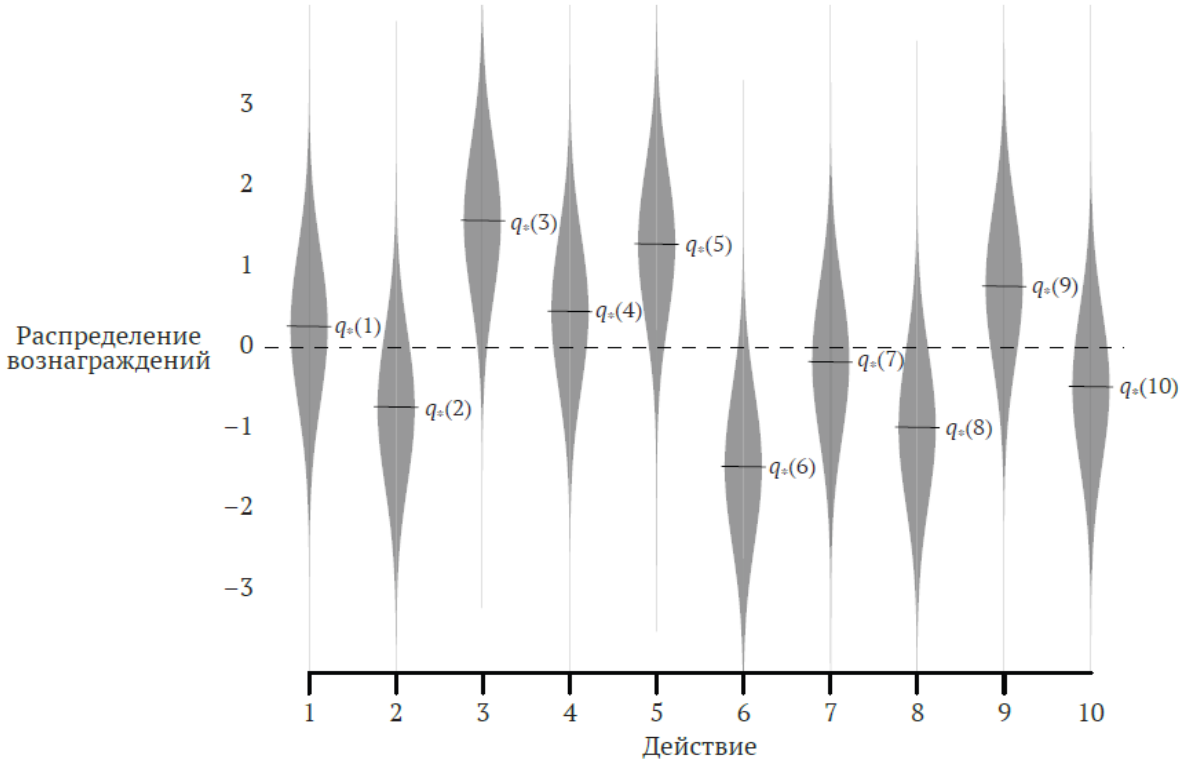
Методы ценности действий. Часть 2

Простейшее правило – выбирать действие с наибольшей оценкой ценности, т. е. одно из жадных действий в смысле предыдущего раздела. Если жадных действий несколько, то выбирается любое из них. Этот метод выбора жадного действия записывается как:
$A = argmax(Q_t(a))$,
где $argmax_a$ *обозначает действие, для которого следующее далее выражение при*нимает максимум.

Простая альтернатива – вести себя жадно большую часть времени, но иногда, скажем с малой вероятностью *ε, случайным* образом выбирать какое-то из прочих действий с одинаковой вероятностью, не зависящей от оценок ценности действий. Будем называть методы, в которых применяется такое почти жадное правило выбора действия, *ε-жадными. Их преим*ущество состоит в том, что в пределе, когда число шагов стремится к бесконечности, каждое действие будет случайно выбрано бесконечное число раз, поэтому все $Q_t(a)$ *сходятся к* $q^*(a)$. *Отсюда, конечно, следует, что вероятность выбора оп*тимального действия сходится к числу, большему 1 – *ε, т. е. оптимальный выбор* почти гарантирован.

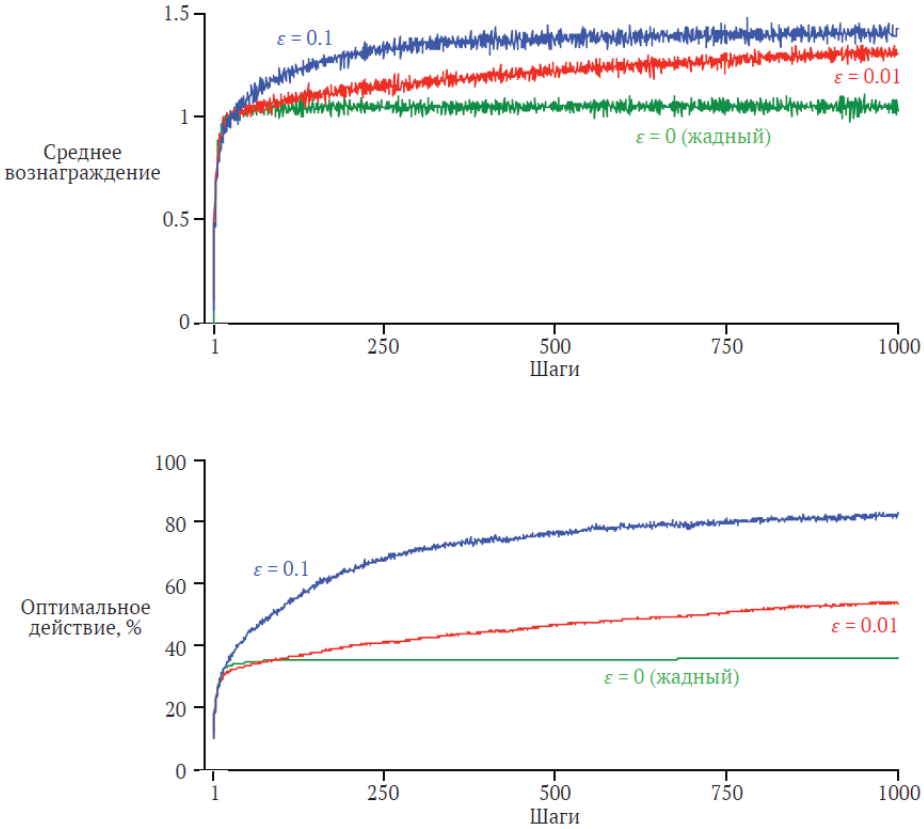
10-рукий испытательный стенд. Часть 1

Чтобы грубо оценить относительную эффективность жадных и *ε-жадных методов* ценности действий, сравним их численно на наборе тестовых задач. В набор вошли 2000 случайно сгенерированных задач о *k-руких бандитах при k = 10.* Такой набор тестовых задач мы будем называть *10-руким испытательным стендом.*



10-рукий испытательный стенд. Часть 2

Результаты сравнения жадного метода с двумя описанными выше *ε-жадными (при ε = 0.01 и ε = 0.1) на 10-руком испытательном стен*де. Во всех методах оценки ценности значений были получены усреднением по выборке.



Инкрементная реализация

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}.$$

$$
\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)Q_n\right) \\
&= \frac{1}{n}(R_n + nQ_n - Q_n) \\
&= Q_n + \frac{1}{n}[R_n - Q_n],
\end{aligned}
$$

НоваяОценка ← СтараяОценка + РазмерШага[Цель − СтараяОценка]

Простой алгоритм бандита

Инициализировать для a от 1 до k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Повторять бесконечно:

$$A \leftarrow \begin{cases} \operatorname{argmax}_a Q(a) & \text{с вероятностью } 1-\varepsilon \\ \text{случайное действие} & \text{с вероятностью } \varepsilon \end{cases} \quad \begin{array}{r}(\text{неоднозначность} \\ \text{разрешается} \\ \text{случайным образом})\end{array}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + 1/N(A)\,[R - Q(A)]$

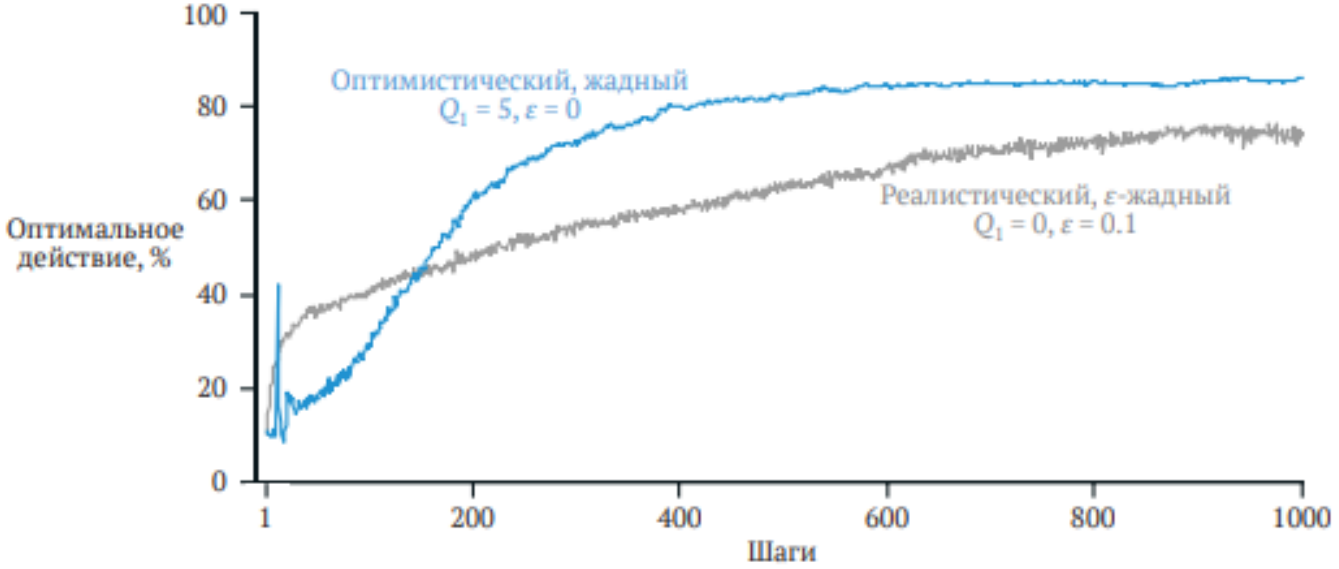
Нестационарная задача

$$Q_{n+1} \doteq Q_n + \alpha[R_n - Q_n],$$

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \quad \text{и} \quad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty.$$

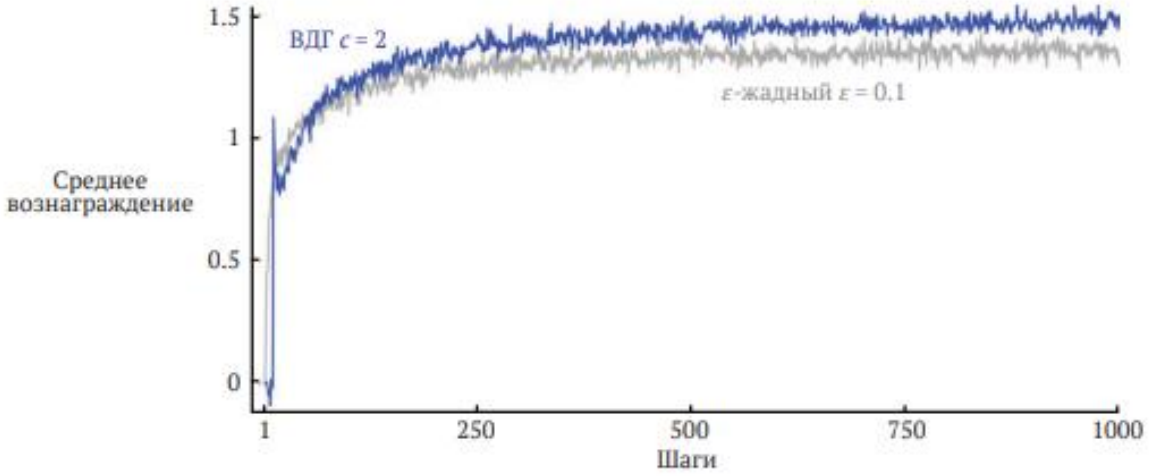
$$\begin{aligned}
Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\
&= \alpha R_n + (1-\alpha)Q_n \\
&= \alpha R_n + (1-\alpha)[\alpha R_{n-1} + (1-\alpha)Q_{n-1}] \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \cdots + (1-\alpha)^{n-1}\alpha R_1 + (1-\alpha)^n Q_1 \\
&= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i.
\end{aligned}$$

Оптимистические начальные значения



Выбор действия, дающего верхнюю доверительную границу

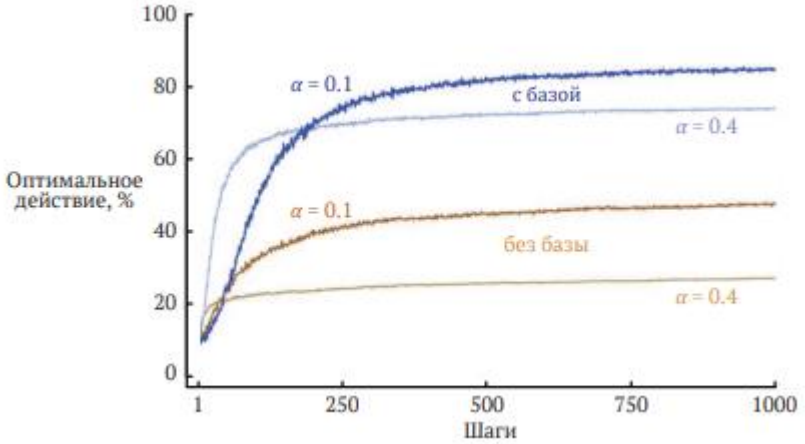
$$A_t \doteq \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right],$$



Градиентные алгоритмы бандита

$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a),$$

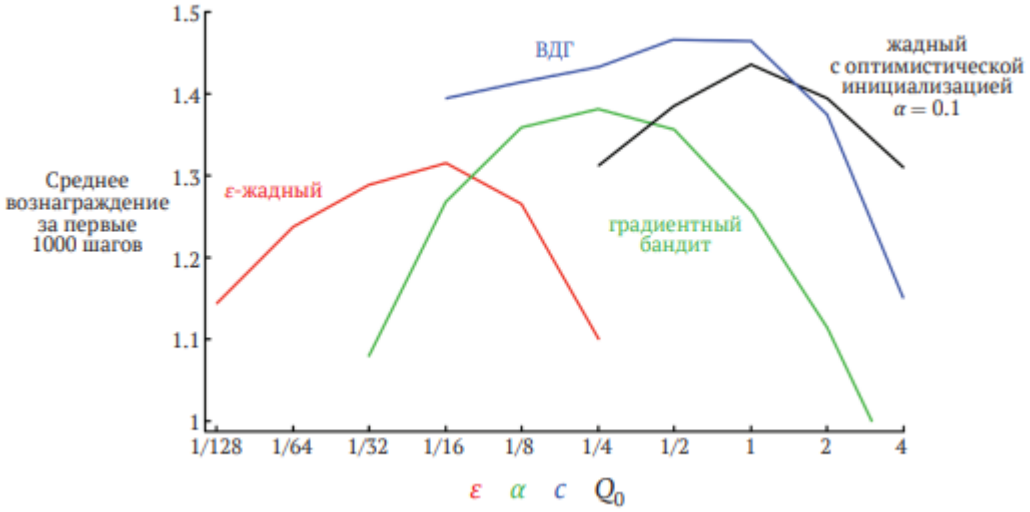
$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)) \quad \text{и}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a) \quad \text{для всех } a \neq A_t,$$



Ассоциативный поиск

Назван так, потому что сочетает как обучение методом проб и ошибок, направленное на поиск наилучших действий, так и ассоциирование этих действий с ситуациями, в которых они являются наилучшими. В литературе задачи ассоциативного поиска часто называются контекстуальными бандитами

Заключение


1.5
1.4
1.3
1.2
1.1
1
Среднее вознаграждение за первые 1000 шагов
ВДГ
жадный с оптимистической инициализацией $\alpha = 0.1$
ε-жадный
градиентный бандит
1/128
1/64
1/32
1/16
1/8
1/4
1/2
1
2
4
ε α c Q_0

Спасибо за внимание!

Вопросы?