Обучение с подкреплением

Особенности обучения с подкреплением

- Обучение происходит не на основе примеров «как надо делать», а путём самостоятельного поиска оптимального решения в процессе взаимодействия со средой
- Действия агента могут иметь отдалённые по времени последствия
- Баланс между использованием имеющихся знаний и получением новых
- Задача взаимодействия со средой решается в полном объёме и реальном времени

Основные элементы задачи обучения с подкреплением

- Агент
- Среда
- Стратегия
- Функция подкрепления
- Функция ценности
- Модель среды

N-рукий бандит

- Агент многократно выбирает из N действий.
- После каждого выбора подкрепление определяется из стационарного распределения, зависящего от выбора.
- Задача: максимизировать ожидаемое суммарное подкрепление за большое число (1000) попыток.

N-рукий бандит

- Ценность действия – ожидаемое (среднее) значение соответствующего распределения.

$$Q^*(a)$$

- Агент вычисляет оценки ценностей.

$$Q_t(a)$$

  - Выбор действия с максимальной оценкой – жадное действие
  - Выбор другого действия - исследование

Метод ценности действий

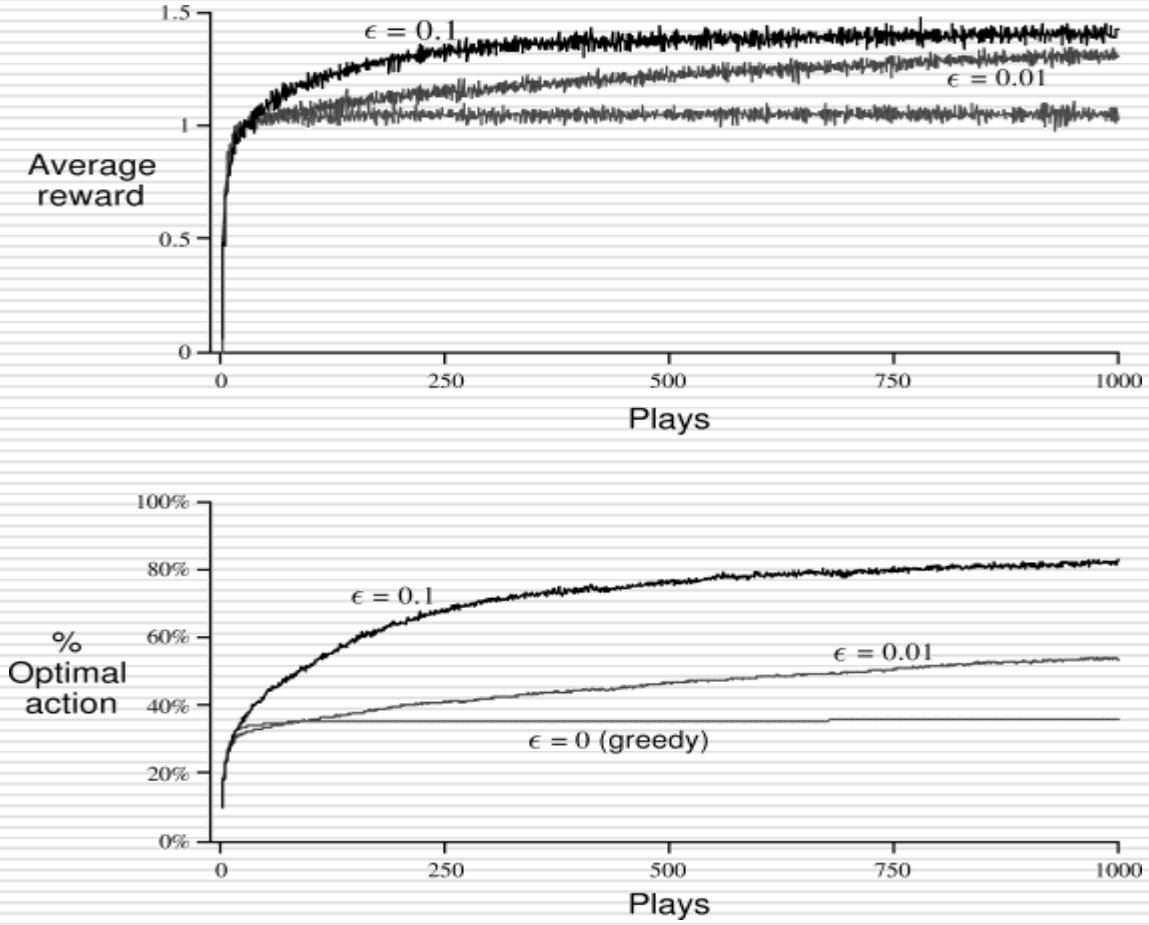
$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a}.$$

где k_a – число выбора действия a за первые t шагов, r_i – полученные в результате подкрепления.

ε-жадный алгоритм:

- с вероятностью 1-ε выбираем жадное действие $Q_t(a^*) = \max_a Q_t(a)$
- с вероятностью ε выбираем не жадное действие.

ε-жадный алгоритм



Мягкий максимум

Вероятность выбора действия пропорциональна его текущей оценке

$$\frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}},$$

Увеличивая температуру *t*, уменьшаем разницу между вероятностями выбора, уменьшая *t*, приближаем выбор к жадному.

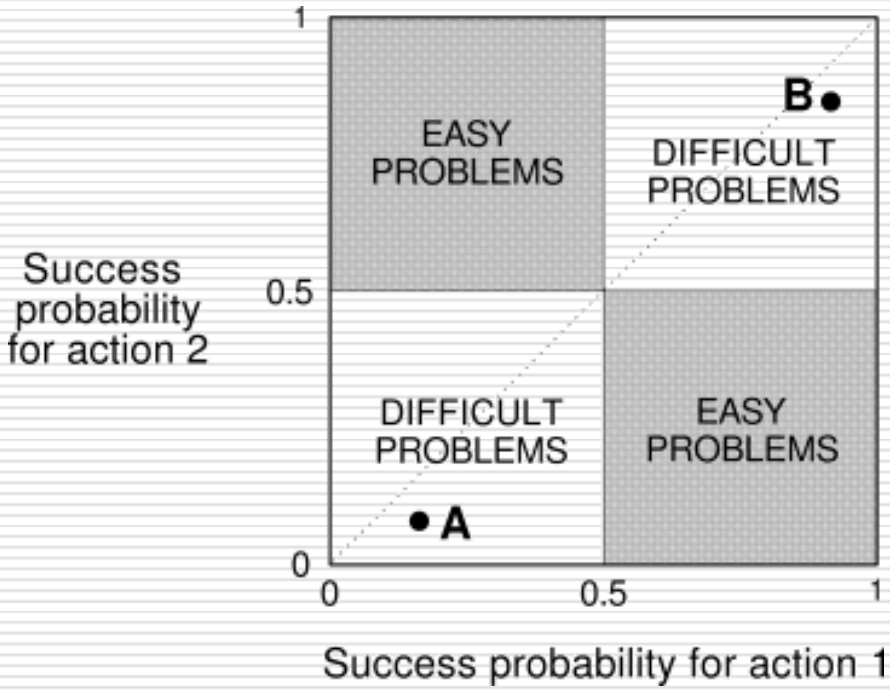
Двоичные бандиты

- Пусть есть два действия и два варианта подкрепления – удача и неудача.

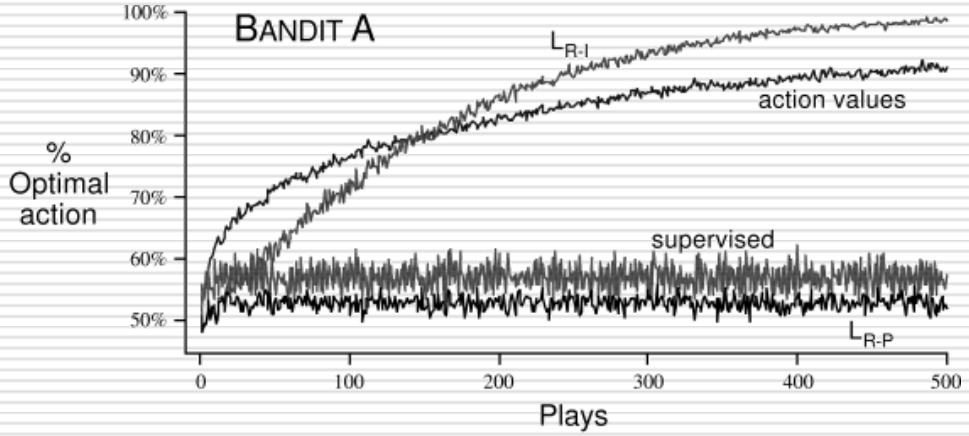
- Аналог обучения с учителем:
  - если результат – удача, значит правильным является выбранный вариант,
  - если неудача – правильно другое действие.
  - Выбираем действие, которое наиболее часто оказывается удачным.

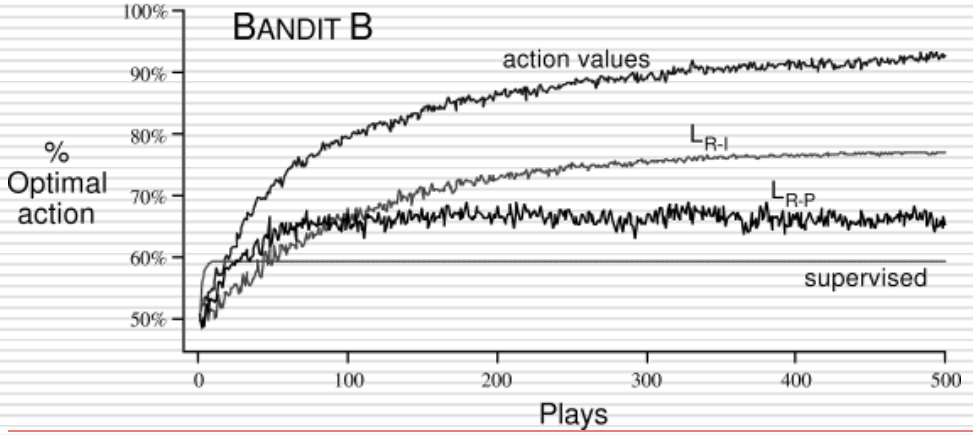
Стохастические двоичные бандиты

- Каждый бандит определяется вероятностями удачи для каждого из действий



Стохастические двоичные бандиты





- L_{R-P}:

Обновляем вероятность выбора «правильного» действия после каждой игры по закону

$$\pi_{t+1}(d_t) = \pi_t(d_t) + \alpha\Big[1 - \pi_t(d_t)\Big].$$

- L_{R-I}:

Обновляем вероятности выбора только после успешных игр

Оптимальное решение для двоичных бандитов

- Пусть в задаче n-рукого бандита для определения подкрепления используется распределение Бернулли с вероятностью p_i.
- Действие i назовём правильным, если $p_i = p + \epsilon$ и неправильным, если $p_i = p - \epsilon$ для $\epsilon \in (0,0.5)$ и $p \in [\epsilon, 1-\epsilon]$. Каждое действие может быть правильным с вероятностью α и неправильным с вероятностью $1-\alpha$.
- Следующий алгоритм гарантирует нахождение правильного действия с вероятностью $1-\delta$ за минимальное число шагов (K. Chandrasekaran, R. Karp, 2012):

1. $L_i = 1,\ i \in [1, \dots, n]$
2. *while* $L_i < \frac{(1-\alpha)(1-\delta)}{\alpha\delta}$
   a. *Выполнить действие* $i^* = arg \max_{i \in [1,\dots,n]} \{L_i\}$
   b. *Получить подкрепление* r_{i^*} *(0 или 1).*
   c. $L_{i^*} = L_{i^*} \left(\frac{p+\epsilon}{p-\epsilon}\right)^{r_{i^*}} \left(\frac{1-p-\epsilon}{1-p+\epsilon}\right)^{1-r_{i^*}}$.
3. *Выдать* $arg \max_{i \in [1,\dots,n]} \{L_i\}$.

Вычисление оценок

$$
\begin{aligned}
Q_{k+1} &= \frac{1}{k+1}\sum_{i=1}^{k+1} r_i \\
&= \frac{1}{k+1}\left(r_{k+1} + \sum_{i=1}^{n} r_i\right) \\
&= \frac{1}{k+1}\Big(r_{k+1} + kQ_k + Q_k - Q_k\Big) \\
&= \frac{1}{k+1}\Big(r_{k+1} + (k+1)Q_k - Q_k\Big) \\
&= Q_k + \frac{1}{k+1}\Big[r_{k+1} - Q_k\Big],
\end{aligned}
$$

Вычисление оценок

Новая оценка=Старая оценка +
Шаг * (Цель – Старая оценка).

(Цель – Старая оценка) – ошибка

У нас сейчас шаг = $1/k_a$.

Условия сходимости

- Чтобы исключить влияние начального значения

$$\sum_{k=1}^{\infty} \alpha_k(a) = \infty$$

$a_k = 1/k$ - выполняется
$a_k = \text{const}$ - выполняется

- Чтобы гарантировать сходимость

$$\sum_{k=1}^{\infty} \alpha_k^2(a) < \infty.$$

$a_k = 1/k$ - выполняется
$a_k = \text{const}$ - не выполняется

Нестационарные задачи

Пусть α-константа, $0<\alpha\leq 1$.

$$
\begin{aligned}
Q_k &= Q_{k-1} + \alpha\left[r_k - Q_{k-1}\right] \\
&= \alpha r_k + (1-\alpha)Q_{k-1} \\
&= \alpha r_k + (1-\alpha)\alpha r_{k-1} + (1-\alpha)^2 Q_{k-2} \\
&= \alpha r_k + (1-\alpha)\alpha r_{k-1} + (1-\alpha)^2 \alpha r_{k-2} + \\
&\quad \cdots + (1-\alpha)^{k-1}\alpha r_1 + (1-\alpha)^k Q_0 \\
&= (1-\alpha)^k Q_0 + \sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} r_i.
\end{aligned}
$$

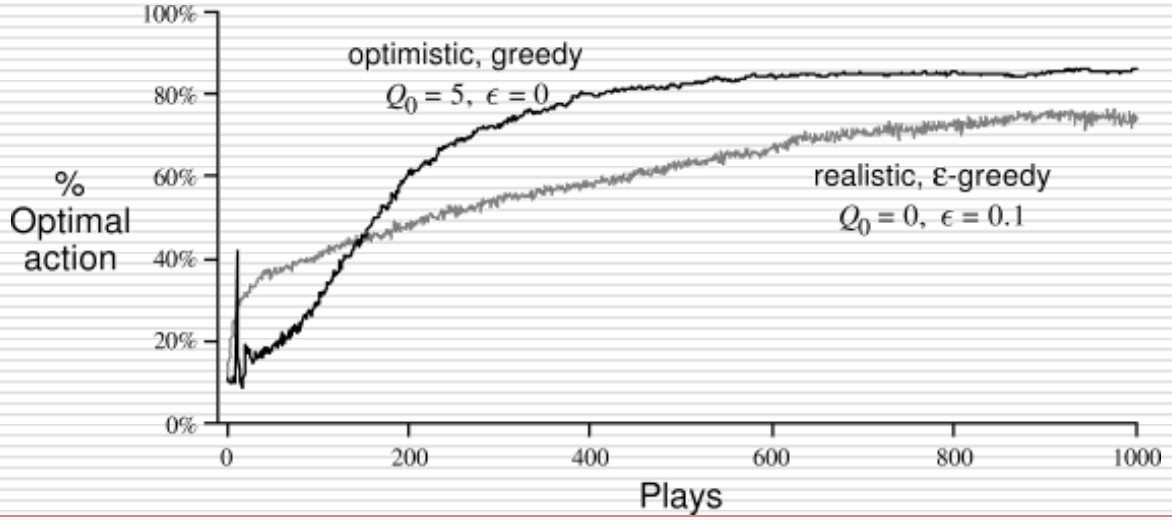
$(1-\alpha)^k + \sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} = 1$

Оптимистичные начальные значения

- Что будет, если мы сделаем начальные значения нереалистично большими?

Оптимистичные начальные значения

- Что будет, если мы сделаем начальные значения нереалистично большими?
  - Агент будет пробовать действия, и ухудшать их оценку. Следующим он выберет ещё не опробованное действие.
  - Мы получили способ поощрить исследования в начале обучения.



Сравнение подкреплений

Хорошее подкрепление - это то, которое лучше среднего.

- Предпочтения действий

$$\pi_t(a) = Pr\{a_t = a\} = \frac{e^{p_t(a)}}{\sum_{b=1}^{n} e^{p_t(b)}},$$

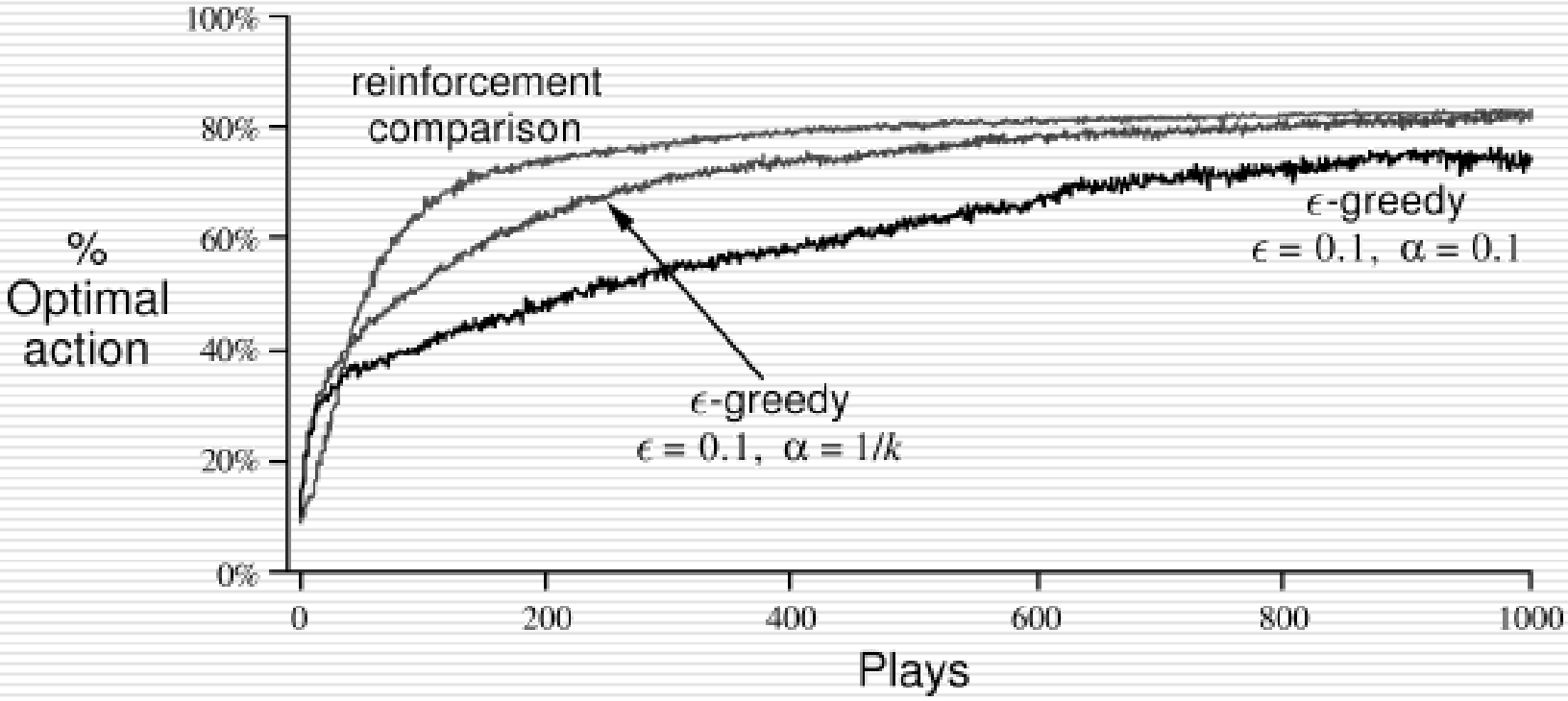
- Обновление предпочтений

$$p_{t+1}(a_t) = p_t(a_t) + \beta\left[r_t - \bar{r}_t\right],$$

- Опорное подкрепление

$$\bar{r}_{t+1} = \bar{r}_t + \alpha\left[r_t - \bar{r}_t\right],$$

Сравнение подкреплений



Методы преследования

Агент одновременно хранит и оценки ценностей действий, и предпочтения.

Предпочтения меняются, отслеживая изменения оценок:

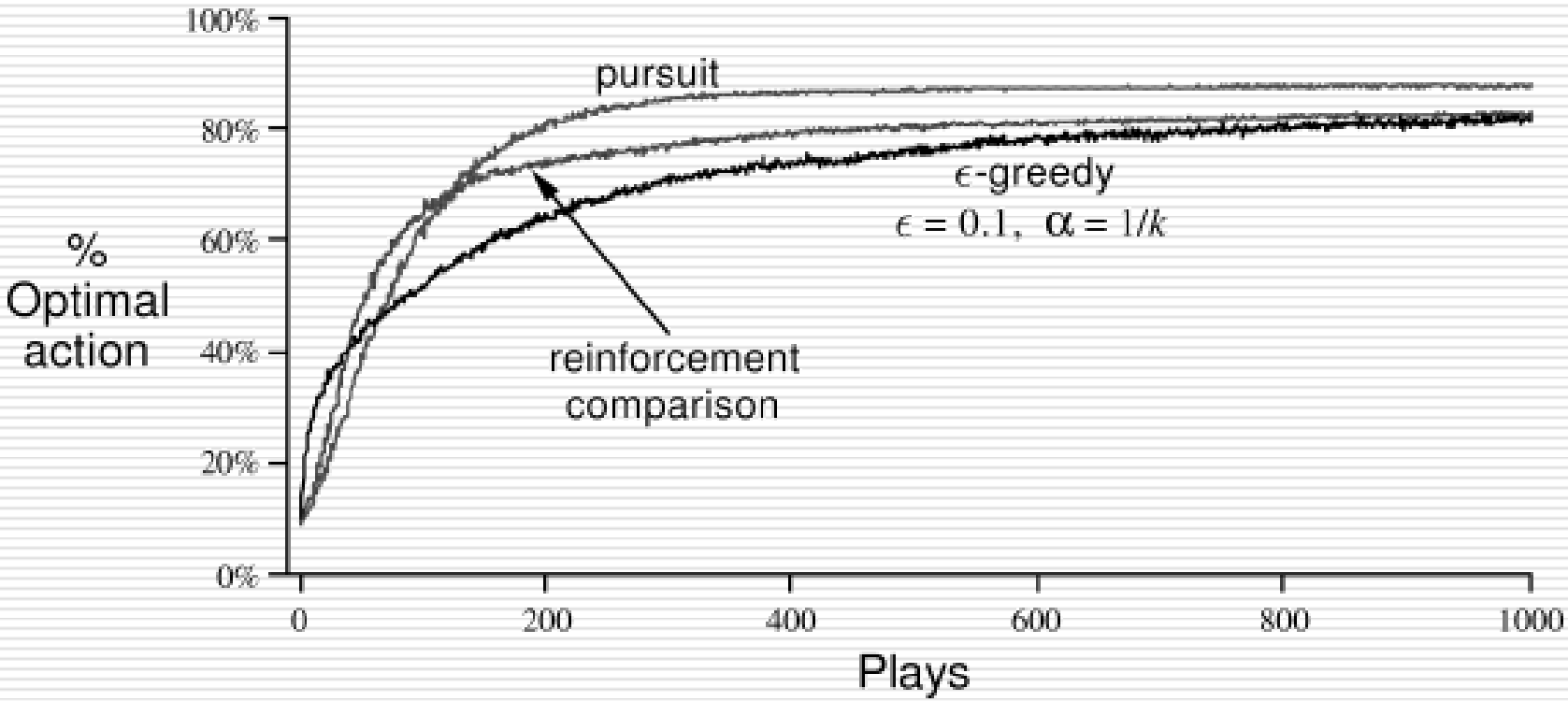
- для жадного действия $a^*_{t+1} = \arg\max_a Q_{t+1}(a)$

$$\pi_{t+1}(a^*_{t+1}) = \pi_t(a^*_{t+1}) + \beta\Big[1 - \pi_t(a^*_{t+1})\Big],$$

- для остальных действий

$$\pi_{t+1}(a) = \pi_t(a) + \beta\Big[0 - \pi_t(a)\Big], \qquad \text{for all } \; a \neq a^*_{t+1}.$$

Методы преследования



Ассоциативная задача

- В нашей задаче нет необходимости учитывать, что решения надо принимать в разных ситуациях.

- В общем случае, агент может попадать в разные условия и должен действовать в каждом из них оптимально (и, возможно, разным образом) – это называется ассоциативной задачей.

Ассоциативная задача

- Допустим, есть много задач N-рукого бандита, и агенту каждый раз предлагается одна случайная задача.
- Для агента такая ситуация выглядит как задача, в которой истинная ценность действий постоянно изменяется.
  - Если скорость изменения не велика, можно использовать методы, которые ориентированы на отслеживание изменений. Но в общем случае они будут не очень эффективны.
  - Если мы знаем, какую из задач мы сейчас имеем, то нам нужно хранить отдельные значения функции ценности для каждой задачи.
- Если действие агента влияет на вероятность выбора следующей задачи, то это полная задача обучения с подкреплением.

Выводы

- В условиях задачи обучения с подкреплением агент не знает, какое действие является правильным. Он пытается максимизировать сумму подкреплений, получаемых им от среды.
- Для того, чтобы определить «полезность» тех или иных действий, агент использует функции ценности.
- Агент старается вычислить оценку функции ценности таким образом, чтобы она приблизилась к истинной функции ценности.
- Большое значение имеет нахождение балланса между использованием имеющегося знания и исследования новых или плохо изученных ситуаций и действий.
  - ε-жадный алгоритм
  - Мягкий максимум
  - Метод сравнения подкреплений