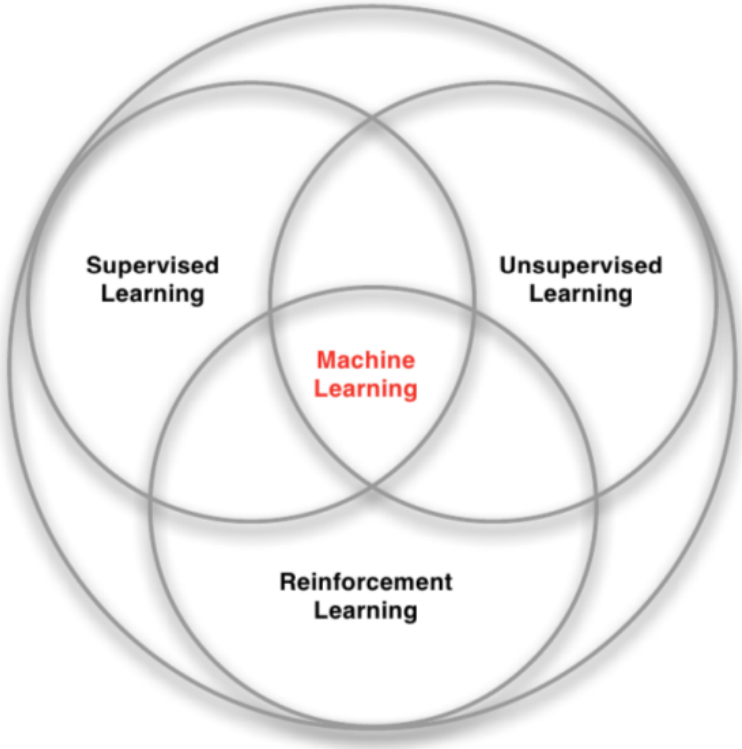
Reinforcement Learning (1)

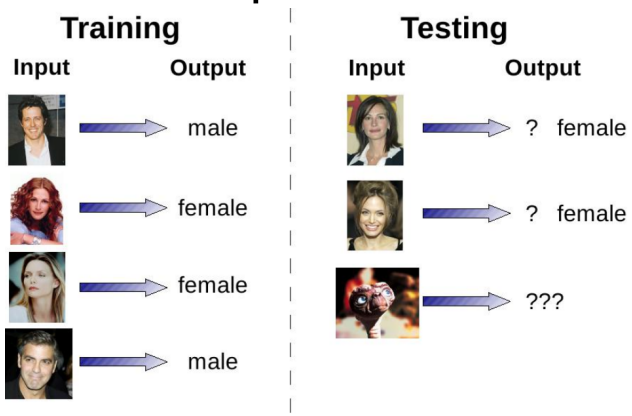
Maxime Berar

18 septembre 2025

Learnings . . .

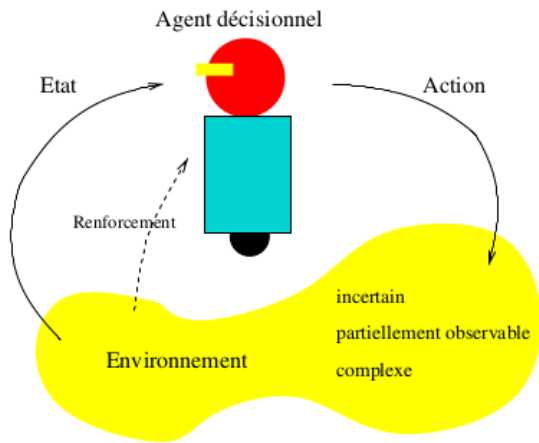


Supervised



Unsupervised

Reinforcement



A definition

from Richard S. Sutton and Andrew G. Barto
Reinforcement Learning : an Introduction
A Bradford Book,London, England
The MIT Press,Cambridge, Massachusetts

Reinforcement learning is learning what to do-how to map situations to actions-so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them.

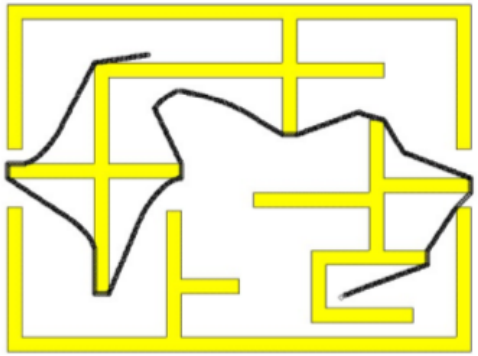
- RL : a problem, a class of solution methods that work well on the problem, and the field that studies this problems and its solution methods.
- distinction between problems and solution methods is very important

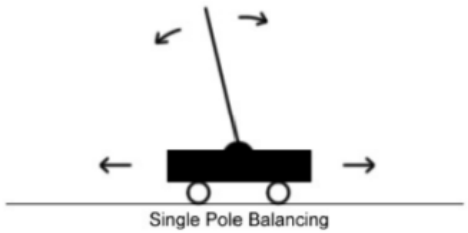
A broader definition

Modelization through dynamic system theories : the optimal control of incompletly-known Markov processes.

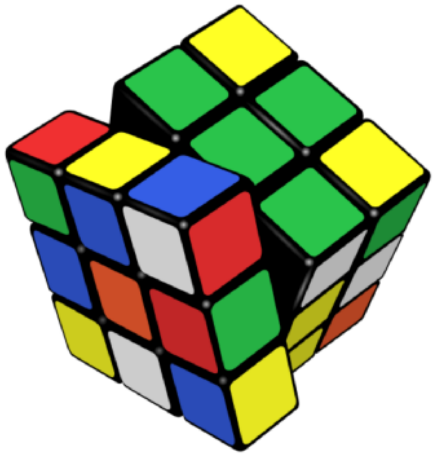
- The basic idea : to capture the most important aspects of the real problem facing a learning agent interacting over time with its environment to achieve a goal.
- Exploration/exploitation trade-off : to obtain a lot of reward, a RL agent must prefer actions that it has tried before and found to be effective in producing reward. But to discover such actions, it has to try actions that it has not selected before.
- a whole problem : goal-directed agent interacting with an uncertain environment.

Examples




Single Pole Balancing


BLACK JACK PAYS 3 TO 2
Dealer must draw to 16 and stand on all 17s
INSURANCE • PAYS • 2 TO 1




CLICK HERE!
Inflatable Fun For Everyone!
Bounce Magic
$77 - JUMP
explore&more
a children's museum
click
Pumpkinville
Fall Family Fun!
The Buffalo & Erie County Public Library presents
The Big Read and
Find the Falcon ... A Contest for All Ages!
Click here for more information
BIG READ

What have they in common ?

- interaction between an active decision-making *agent* and its *environment*. **Seeking a goal despite uncertainty**
- *uncertainty :* the effects of the action cannot be fully predicted, but experience can be used to improve the agent performance over time.

Elements of RL

- *policy* the agent's way of behaving at a given time
- *reward signal* single number given by the environment at a given time
- *value function* long-term desirability of given state
- *model* of the environment. Planning vs trial-error

Applications

- Helicopter
  `https://www.youtube.com/watch?v=3mvE_CeH9q0`
- Backgamon `https://en.wikipedia.org/wiki/TD-Gammon`,
- AlphaGo
- management of an action porfolio
- Humanoid robot walk
- Playing Atari games better than human (*Nature* paper)

MULTI-ARMED BANDITS

Multi-armed Bandits

Repeated choice of one action among k, after each choice you receive a numerical reward chosen from a stationary probability distribution that depends on the selected action.

- analogy to a slot machine, with k levers instead of 1
- each action selection is like a play of one of the levers
- the rewards are the payoffs for hitting the jackpot

A k-armed Bandit

Each action a has an expected or mean reward, called the value of the action

$$q_*(a) \doteq \mathbb{E}[R_t|A_t].$$

You do not known the action value with certainty, although you have estimates $Q_t(a)$, that you want as close as possible to $q_*(a)$.

- choosing the action with the highest estimated value, `greedy` action $\Rightarrow$ `exploitation`
- choosing another action $\Rightarrow$ `exploration`

It is not possible to both exploit and explore, algorithms must balance the choice between greedy and non-greedy action.

Stochastic bandits

- A stochastic bandit is a collection of distributions $\nu = (P_a : a \in \mathcal{A})$, where $\mathcal{A}$ is the set of available actions.
- learner and environment interact sequentially over T rounds.
- In each round $t \in \{1, \ldots, T\}$, the learner chooses an action $A_t \in \mathcal{A}$, which is fed to the environment. The environment then samples a reward $X_t \in \mathbb{R}$ from distribution P_{A_t} and reveals X_t to the learner.
- The interaction between the learner (or policy) and environment induces a probability measure on the sequence of outcomes $A_1, X_1, A_2, X_2, \ldots, A_T, X_T$.

Usually the horizon T is finite, but sometimes we allow the interaction to continue indefinitely ($T = \infty$).

Core Assumptions

(a) The conditional distribution of reward X_t given $A_1, X_1, \cdots, A_{t-1}, X_{t-1}, A_t$ is P_{A_t} , which captures the intuition that the environment samples X_t from P_{A_t} in round t.

(b) The conditional law of action A_t given $A_1, X_1, \ldots, A_{t-1}, X_{t-1}$ is $\pi_t(\cdot|A_1, X_1, \ldots, A_{t-1}, X_{t-1})$, where $\pi_1, \pi_2, \ldots$ is a sequence kernels that characterise the learner.

The most important element of this assumption is the intuitive fact that the learner cannot use the future observations in current decisions.

Learning Objective

The learner's goal is to maximise the total reward $S_T = \sum_{t=1}^{T} X_t$, which is a random quantity that depends on the actions of the learner and the rewards sampled by the environment. This is not an optimisation problem for three reasons :

1. What is the value of T for which we are maximising ? Occasionally prior knowledge of the horizon is reasonable, but very often the learner does not know ahead of time how many rounds are to be played.
2. The cumulative reward is a random quantity. Even if the reward distributions were known, then we require a measure of utility on distributions of S_T .
3. The learner does not know the distributions that govern the rewards for each arm.

The Regret

The regret of policy π on bandit instance ν is

$$R_T(\pi, \nu) = T\mu^*(\nu) - \mathbb{E}\left[\sum_{i=1}^{T} X_t\right]$$

where the expectation is taken with respect to the probability measure on outcomes induced by the interaction of π and ν.
Minimising the regret is equivalent to maximising the expectation of S_n, but the normalisation inherent in the definition of the regret is useful when stating results, which would otherwise need to be stated relative to the optimal action.
The regret is always non-negative, and for every bandit ν, there exists a policy π for which the regret vanishes.

Objectives

What can we hope for ?
A relatively weak objective is to find a policy π with sublinear regret on all $\nu \in \mathcal{E}$ (class of bandits).

$$\text{for all} \nu \in \mathcal{E}, \quad \lim_{T \to \infty} \frac{R_T(\pi, \nu)}{T} = 0$$

If the above holds, then at least the learner is choosing the optimal action almost all of the time as the horizon tends to infinity.

Naive Method

Sample-average method

Use the current average $Q_t(a)$ of the value of the actions

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbf{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbf{1}_{A_i=a}}$$

As the denominator goes to infinity, $Q_t(a)$ converges to $q_*(a)$.

Proposition : `exploration` then `exploitation`

Uniform exploration

Explore-First with parameter N

1. Exploration phase: try each arm N times;
2. Select the arm $\hat{a}$ with the highest average reward (break ties arbitrarily);
3. Exploitation phase: play arm $\hat{a}$ in all remaining rounds.

Some elements ... for $K = 2$ arms

Let the average reward for each action a after exploration phase be denoted $\bar{\mu}(a)$. We want the average reward to be a good estimate of the true expected rewards, i.e. the following quantity should be small : $|\bar{\mu}(a) - \mu(a)|$. We bound it using the Hoeffding inequality in order to optain a *confidence interval* :

$$Pr\left[|\bar{\mu}(a) - \mu(a)| \leq \text{rad}\right] \geq 1 - 2/T^4, \text{where rad} := \sqrt{2\log(T)/N} \qquad (1)$$

We define the **clean event** to be the event that (1) holds for all arms simultaneously.

- one does not need to worry about probability in the rest of the analysis. Indeed, the probability has been taken care of by defining the clean event and observing that (1) holds therein.
- We do not need to worry about the bad event : essentially, because its probability is so tiny.

Some elements ... for $K = 2$ arms

Let the best arm be a^*, and suppose the algorithm chooses the other arm $a \neq a^*$. This must have been because its average reward was better than that of a^* : $\bar{\mu}(a) > \bar{\mu}(a^*)$. Since this is a clean event, we have :

$$\mu(a) + \text{rad} \geq \bar{\mu}(a) > \bar{\mu}(a^*) \geq \mu(a^*) - \text{rad}$$

$$\mu(a^*) - \mu(a) \leq 2\text{rad}$$

- each round in the exploitation phase contributes at most 2 rad to regret.
- each round in exploration trivially contributes at most 1.

$$R(T) \leq N + 2\text{rad} \cdot (T - 2N) < N + 2\text{rad} \cdot T$$

Some elements ... for $K = 2$ arms

$$R(T) \leq N + 2\text{rad} \cdot (T - 2N) < N + 2\text{rad} \cdot T$$

we can set N such that both summands are equal

$(\text{rad} = \sqrt{2\log(T)/N})$: For $N = \log(T)^{1/3} T^{2/3}$

$$R(T) \leq O(\log(T)^{1/3} T^{2/3})$$

It remains to analyze the "bad event". Since regret can be at most T (each round contributes at most 1), and the bad event happens with a very small probability, regret from this event can be neglected.

It is usually better to spread exploration more uniformly over time.

Explore-First

Explore-First with parameter N

1. Exploration phase: try each arm N times;
2. Select the arm $\hat{a}$ with the highest average reward (break ties arbitrarily);
3. Exploitation phase: play arm $\hat{a}$ in all remaining rounds.

Explore-first achieves regret $\mathbb{E}[R(T)] \leq T^{2/3} \times O(K \log T)^{1/3}$

The parameter N is fixed in advance : $N = (T/K)^{2/3} \cdot O(\log T)^{1/3}$

Spreading the exploration

`greedy` action

$$A_t \doteq \arg\max_a Q_t(a)$$

Near greedy action selection rule : $\varepsilon-$`greedy` selection

- ε : choose the action randomly with equal probability
- $1-\varepsilon$: choose the greedy action

$Q_t(a)$ converges to $q_*(a)$, as all actions will be evaluated as ε is constant

Examples

- 2-armed bandit with $\varepsilon = 0.5$

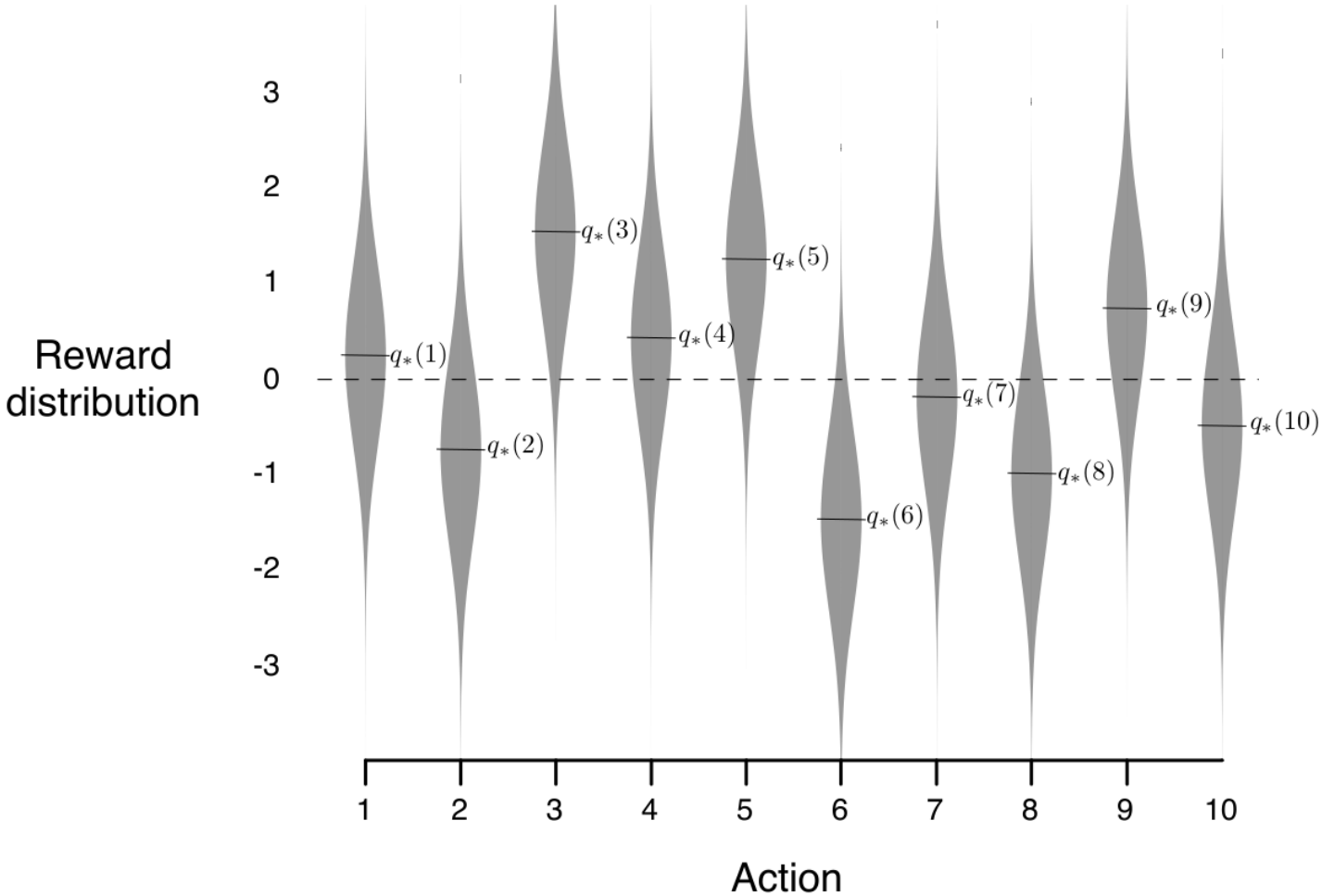
  The probability that the greedy action is selected
  $0.75 = 0.5 + 0.25$

- 4-armed bandit, at what time did the ε step occur ?
  $Q_1(a) = 0, \forall a$

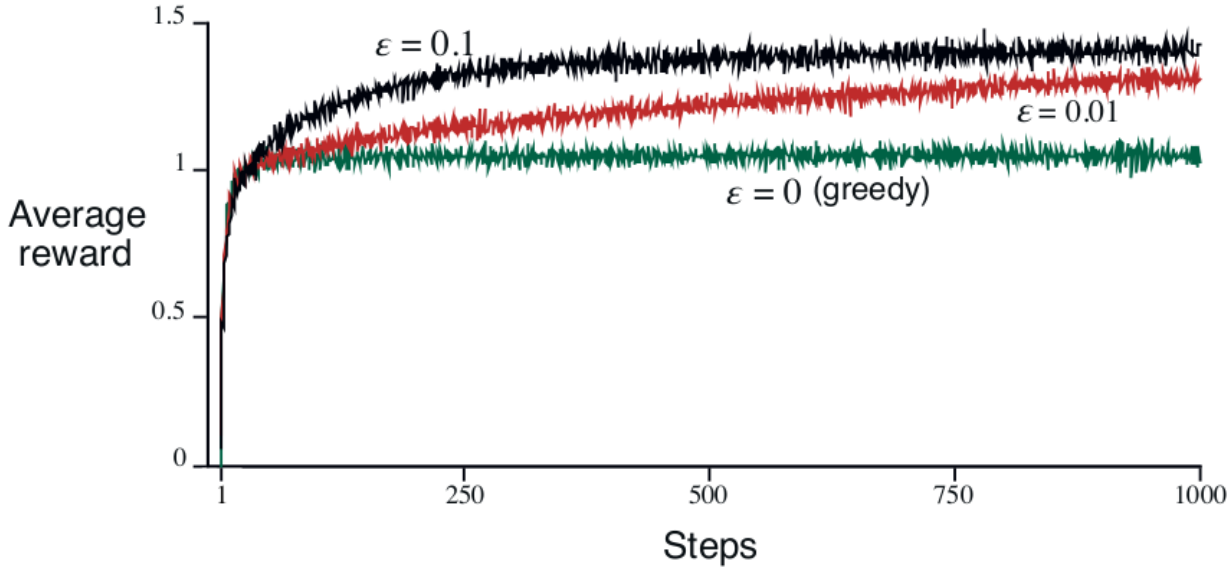
$A_1 = 1$	$R_1 = 1$	$Q_1(1) = 0$	$Q_1(2) = 0$	$Q_1(3) = 0$	$Q_1(4) = 0$
$A_2 = 2$	$R_2 = 1$	$Q_2(1) = 1$	$Q_2(2) = 0$	$Q_2(3) = 0$	$Q_2(4) = 0$
$A_3 = 2$	$R_3 = 2$	$Q_3(1) = 1$	$Q_3(2) = 1$	$Q_3(3) = 0$	$Q_3(4) = 0$
$A_4 = 2$	$R_4 = 2$	$Q_4(1) = 1$	$Q_4(2) = 1.5$	$Q_4(3) = 0$	$Q_4(4) = 0$
$A_5 = 3$	$R_5 = 0$	$Q_5(1) = 1$	$Q_5(2) = \frac{5}{3}$	$Q_5(3) = 0$	$Q_5(4) = 0$

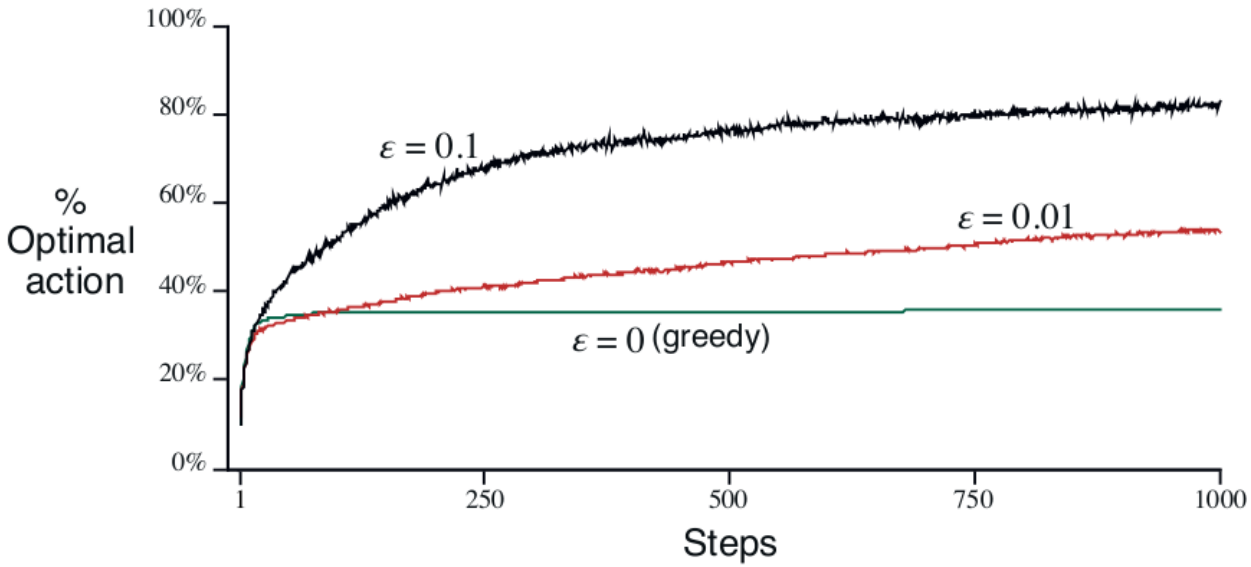
Example : 10-armed Testbed

- $q_*(a)$ selected according to a normal distribution $(0, 1)$
- R_t selected from a normal distribution $(q_*(A_t), 1)$



greedy or not ?





A simple bandit algorithm

incremental implementation, from $Q_t(a)$ to $Q_n(a)$:
$Q_n(a) = \frac{1}{N(a)} \sum_{i=1}^{N(a)} R_i$, $Q_{n+1} = Q_n + \frac{1}{n}[R_n - Q_n]$

Initialize, for $a = 1$ to k:
$Q(a) \leftarrow 0$
$N(a) \leftarrow 0$

Repeat forever:
$A \leftarrow \begin{cases} \arg\max_a Q(A) & \text{with probability } 1-\varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$
$R \leftarrow bandit(A)$
$N(A) \leftarrow N(A) + 1$
$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}[R - Q(A)]$

ϵ-greedy algorithm

Epsilon-greedy algorithm with exploration probabilities $\epsilon_t = t^{-1/3} \cdot (K \log t)^{1/3}$ achieves regret bound

$$\mathbb{E}[R(t)] \leq t^{2/3} \cdot O(K \log t)^{1/3} \text{ for each round } t$$

Hint : Fix round t and analyze $\mathbb{E}[\mu(a^*) - \mu(a_t)]$ for this round separately. Set up the "clean event" for rounds $1, \ldots, t$ treating t as the time horizon, but also include the number of exploration rounds up to time t.

Non-adaptive exploration

Explore-first and Epsilon-greedy do not adapt their exploration schedule to the history of the observed rewards.

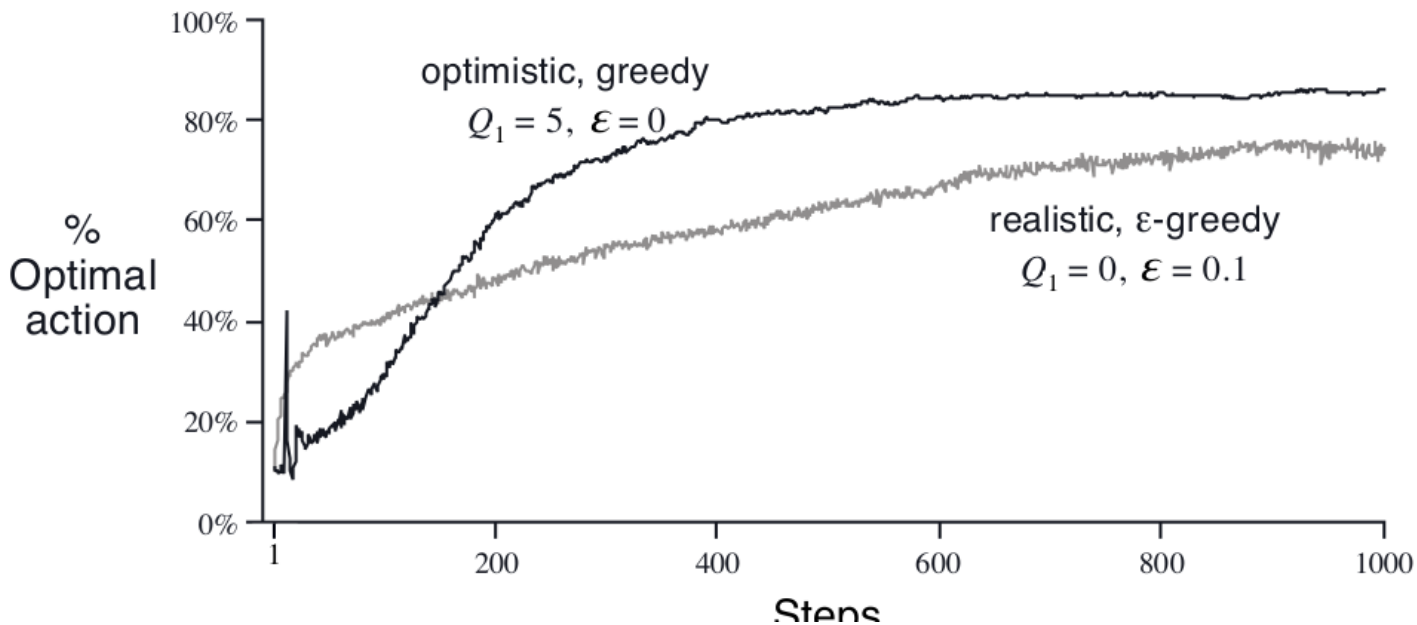
- non-adaptive exploration

A round t is an exploration round if the data (a_t, r_t) from this round is used by the algorithm in the future rounds
A deterministic algorithm satisfies non-adaptive exploration if the set of all exploration rounds and the choice of arms therein is fixed before round 1.
A randomized algorithm satisfies non-adaptive exploration if it does so for each realization of its random seed.

How to support exploration ?

Optimistic initial values

Choose $Q_1(a) \neq 0$, a simple way to encourage exploration (with $N_1(a) = 1$?)



Adaptive exploration

Two algorithms which achieve much better regret bounds. Both algorithms adapt exploration to the observations so that very under-performing arms are phased out sooner.
For each arm a at round t, we define upper and lower confidence bounds,

$$\mathrm{UCB}_t(a) = \bar{\mu}_t(a) + r_t(a),$$

$$\mathrm{LCB}_t(a) = \bar{\mu}_t(a) - r_t(a).$$

Clean event :

$$Pr\left[\mathcal{E}\right] \geq 1 - 2/T^2, \text{ where } \mathcal{E} := \{\forall a \forall t | \bar{\mu}_t(a) - \mu(a)| \leq r_t(a)\}$$

where $r_t(a) = 2\log(T)/n_t(a)$

Successive elimination algorithm

Let's come back to the case of $K = 2$ arms

`Alternate two arms until` $\mathrm{UCB}_t(a) < \mathrm{LCB}_t(a')$ `after some even round` t`;`

`Abandon arm` a`, and use arm` a' `forever since.`

Let t be the last round when we did not invoke the stopping rule, i.e., when the confidence intervals of the two arms still overlap

$$\Delta := |\mu(a) - \mu(a')| \leq 2(r_t(a) + r_t(a')).$$

Since the algorithm has been alternating the two arms before time t, we have $n_t(a) = t/2$

$$\Delta \leq 2(r_t(a) + r_t(a')) \leq 4\sqrt{2\log(T)/[t/2]} = O\left(\sqrt{\log(T)/t}\right)$$

then the total regret accumulated till round t is

$$R_t \leq \Delta \times t \leq O\left(\sqrt{t\log(T)}\right)$$

Since we've chosen the best arm from then on, we have $R(t) \leq O\left(\sqrt{t\log(T)}\right)$

Successive elimination algorithm

All arms are initially designated as **active**
loop *new phase*
  play each active arm once
  deactivate all arms a such that, letting t be the current round, $\mathrm{UCB}_t(a) < \mathrm{LCB}_t(a')$ for some other arm a' *deactivation rule*
end loop

$$\mathbb{E}[R(t)] = O\left(\sqrt{Kt \log T}\right) \text{ for all rounds } t \leq T.$$

Optimism under uncertainty (UCB1)

Try each arm once
for each round $t = 1, \ldots, T$ **do**
  pick arm some a which maximizes $\mathrm{UCB}_t(a)$.
end for
Clean event

$$\mu(a_t) + 2r_t(a_t) \geq \bar{\mu}(a_t) + r_t(a_t) = \mathrm{UCB}_t(a_t) \geq UCB_t(a^*) \geq \mu(a^*).$$

it follows that

$$\Delta(a_t) := \mu(a^*) - \mu(a_t) \leq 2r_t(a_t)$$

Regret bounds idem that successive elimination

Gradient Bandit Algorithm

numerical preference for each action $H_t(a)$ linked via soft-max distribution to the probabibility of choosing an action

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

Measure of performance : expectation of the value function over the policy

$$\mathbb{E}[R_t] = \sum_b \pi_t(b) q_*(b)$$

Problem : unknown q_*

Optimization of an "expectation" cost : *stochastic gradient ascent*

stochastic gradient ascent

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_b q_*(b) \frac{\partial \pi_t(b)}{\partial H_t(a)}$$

$$= \sum_b (q_*(b) - X_t) \frac{\partial \pi_t(b)}{\partial H_t(a)}$$

X_t any scalar independent from b, included here because $\sum_b \frac{\partial \pi_t(b)}{\partial H_t(a)} = 0$ (remember $\sum_b \pi_t(b) = 1$, the sum of the probability must remain 1).

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_b \pi_t(b)(q_*(b) - X_t) \frac{\partial \pi_t(b)}{\partial H_t(a)} / \pi_t(b)$$

$$= \mathbb{E}\left[(q_*(b) - X_t) \frac{\partial \pi_t(b)}{\partial H_t(a)} / \pi_t(b)\right] = \mathbb{E}\left[(R_t - \bar{R}_t) \frac{\partial \pi_t(b)}{\partial H_t(a)} / \pi_t(b)\right]$$

as $\mathbb{E}[R_t | A_t] = q_*(A_t)$ and with $X_t = \bar{R}_t$ the baseline (empirical mean) reward

stochastic gradient ascent cont.

$$\frac{\partial \pi_t(b)}{\partial H_t(a)} = \pi_t(b)\left(\mathbf{1}_{a=b} - \pi_t(a)\right)$$

$$\begin{aligned}
\frac{\partial \pi_t(b)}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \frac{e^{H_t(b)}}{\sum_{c=1}^{k} e^{H_t(c)}} \\
&= \frac{\left(\sum_{c=1}^{k} e^{H_t(c)}\right) \frac{e^{H_t(b)}}{\partial H_t(a)} - e^{H_t(b)} \frac{\sum_{c=1}^{k} e^{H_t(c)}}{\partial H_t(a)}}{\left(\sum_{c=1}^{k} e^{H_t(c)}\right)^2} \\
&= \frac{\mathbf{1}_{a=b} e^{H_t(b)}}{\left(\sum_{c=1}^{k} e^{H_t(c)}\right)} - \frac{e^{H_t(b)} e^{H_t(a)}}{\left(\sum_{c=1}^{k} e^{H_t(c)}\right)^2} \\
&= \mathbf{1}_{a=b}\pi_t(b) - \pi_t(b)\pi_t(a) \\
&= \pi_t(b)\left(\mathbf{1}_{a=b} - \pi_t(a)\right)
\end{aligned}$$

stochastic gradient ascent

$$\frac{\partial \mathbb{E}\left[R_t\right]}{\partial H_t(a)} = \mathbb{E}\left[\left(R_t - \bar{R}_t\right)\left(\mathbf{1}_{a=b} - \pi_t(a)\right)\right]$$

Update rule

$$H_{t+1}(a) = H_t(a) + \alpha\left(R_t - \bar{R}_t\right)\left(\mathbf{1}_{a=b} - \pi_t(a)\right), \quad \forall a$$

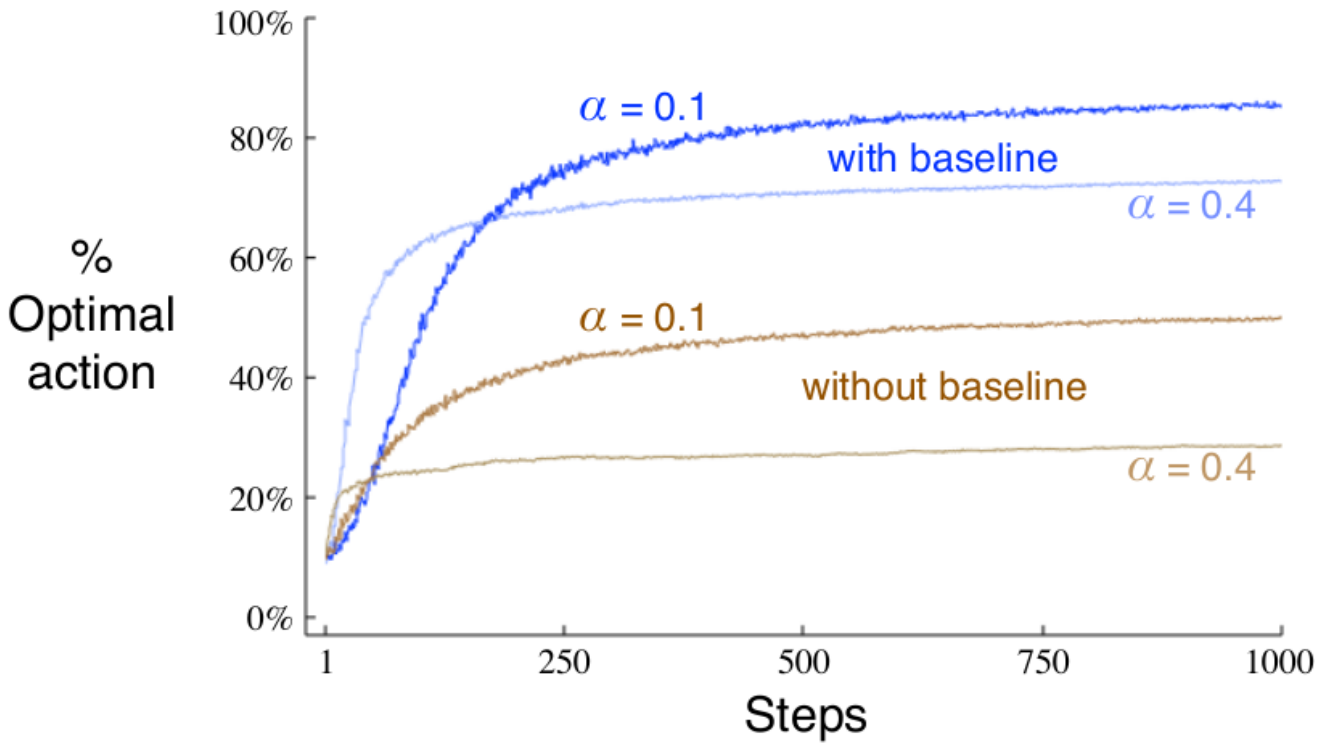
ie

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\left(R_t - \bar{R}_t\right)\left(1 - \pi_t(A_t)\right),$$

$$H_{t+1}(a) \doteq H_t(a) - \alpha\left(R_t - \bar{R}_t\right)\pi_t(a), \quad \forall a \neq A_t$$

Advantage : independent of the mean-value of the bandits distribution, due to the baseline term that will converge slowly to it. Can even be non-stationary with weighted averaging.

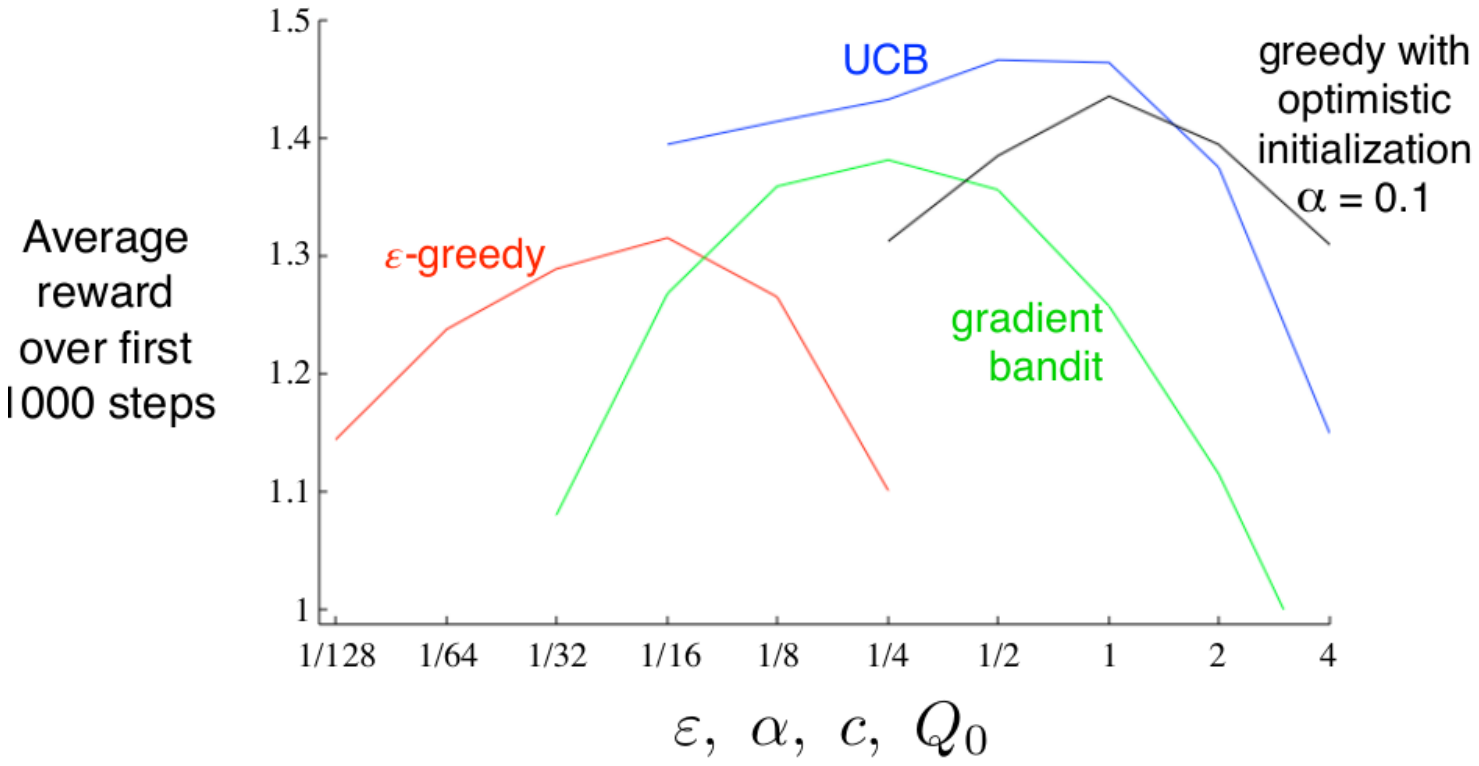
Effect of the baseline



baseline is the empirical mean.

$$H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t)\left(\mathbf{1}_{a=b} - \pi_t(a)\right), \quad \forall a$$

Comparison


Average reward over first 1000 steps
1.5
1.4
1.3
1.2
1.1
1
UCB
greedy with optimistic initialization
α = 0.1
ε-greedy
gradient bandit
1/128
1/64
1/32
1/16
1/8
1/4
1/2
1
2
4
$\varepsilon,\ \alpha,\ c,\ Q_0$

Tracking a non-stationary problem

$Q_{n+1} = Q_n + \alpha[R_n - Q_n]$, with $\alpha \in (0, 1]$ constant

$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$$

Conditions to assure convergence

$$\sum_{i=1}^{\infty} \alpha_n(a) = \infty \quad \text{and} \quad \sum_{i=1}^{\infty} \alpha_n^2(a) < \infty$$

- $\alpha_n(a) = \frac{1}{n}$ ok
- $\alpha_n(a) = \alpha$ not ok

Lower Bounds

What bandit algorithms cannot do

- fundamental results which imply that the regret rates seen previously are essentially the best possible.
- prove that any algorithm suffers regret $\Omega(KT)$ on some problem instance.
- this lower bound is "worst-case",

Fix time horizon T and the number of arms K. For any bandit algorithm, there exists a problem instance such that

$$\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT}).$$