Multi armed bandit

Tor Lattimor’s book and other miscellaneous sources

Presenter: Mathilda Nguyen

The name “Bandit”

“one-armed bandit” is an old name for a slot machine in a casino, because it has one arm and it steals your money.

Multi-arm bandit: imagine a casino with many (different) one-arm slot machines.

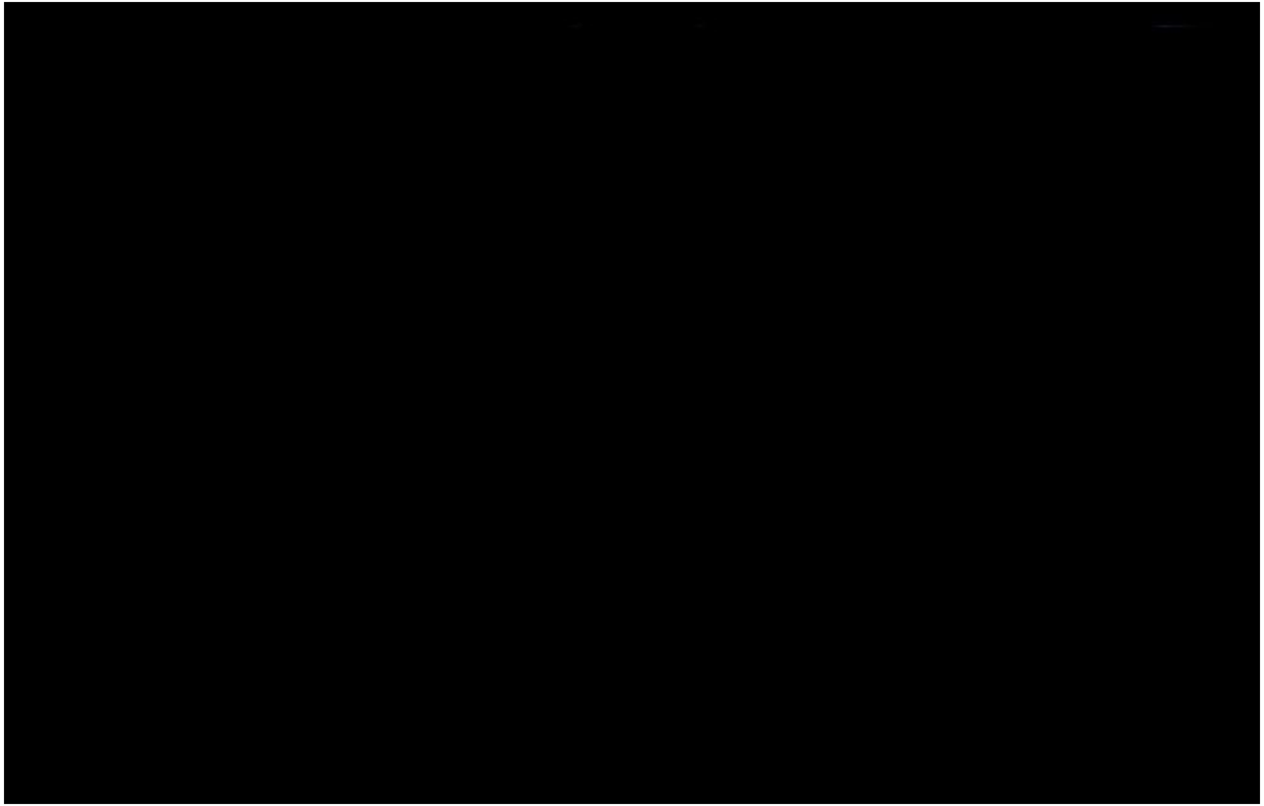

G+Deluxe
WIN IT AGAIN
G+Deluxe
$30.43
$324.10
VAMPIRE'S EMBRACE
WIN IT AGAIN
JACKPOT PARTY
$29.18
$350.16
NOW WITH WIN IT AGAIN
PLUS AN ALL NEW PARTY POOPER™ BONUS!
TOUCH HERE TO VIEW
GAME STATS
WMS
WMS
WMS

What is bandit?

Bandit algorithm is a type of learning algorithm which the agent tried to balance out **exploration** (acquire new knowledge) and **exploitation** (optimized their decision based on knowledge) in order to **maximize rewards**.

Bandit provides a simple model of decision making under uncertainty.

What is bandit?



Real life example: exploration vs exploitation (and rewards)

Clinical trials: investigating effectiveness of different treatments while minimizing patient's loses

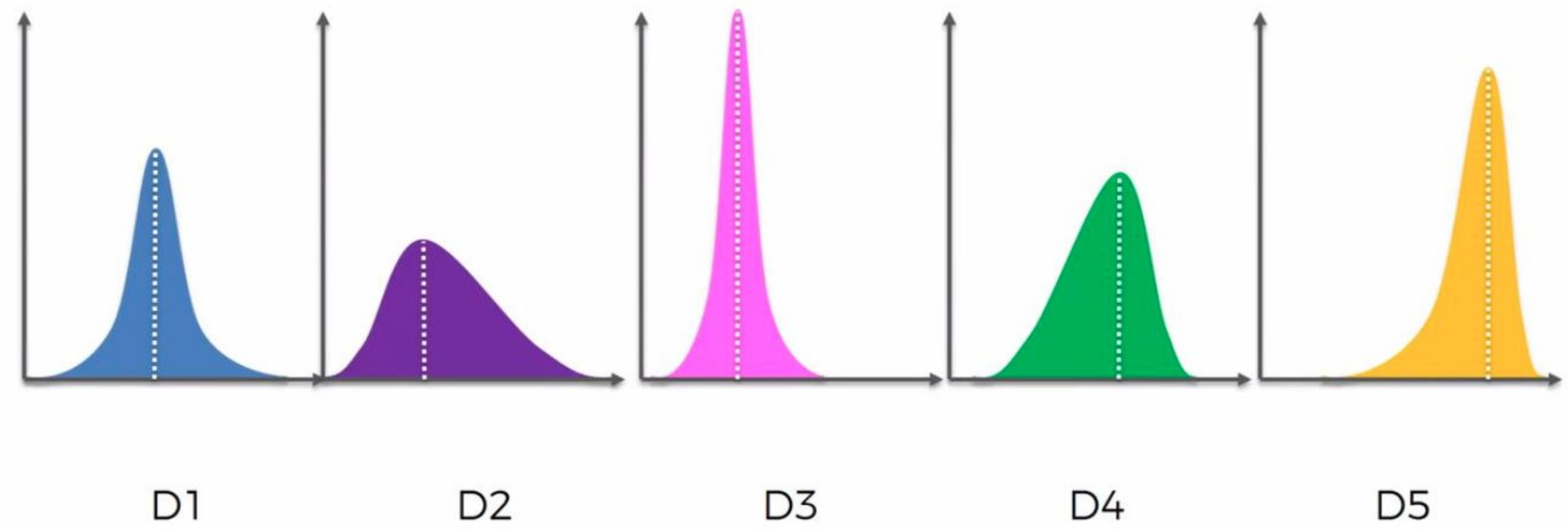
Ads placement/news recommendation: investigating effectiveness of different ads while maximize user clicking in the ads

Dynamic pricing: finding the best price for a product while maximizing buying potentials

Explore-then-commit (ETC)

Explores by playing each arm a fixed number of times and then exploits by committing to the arm that appeared best during exploration.

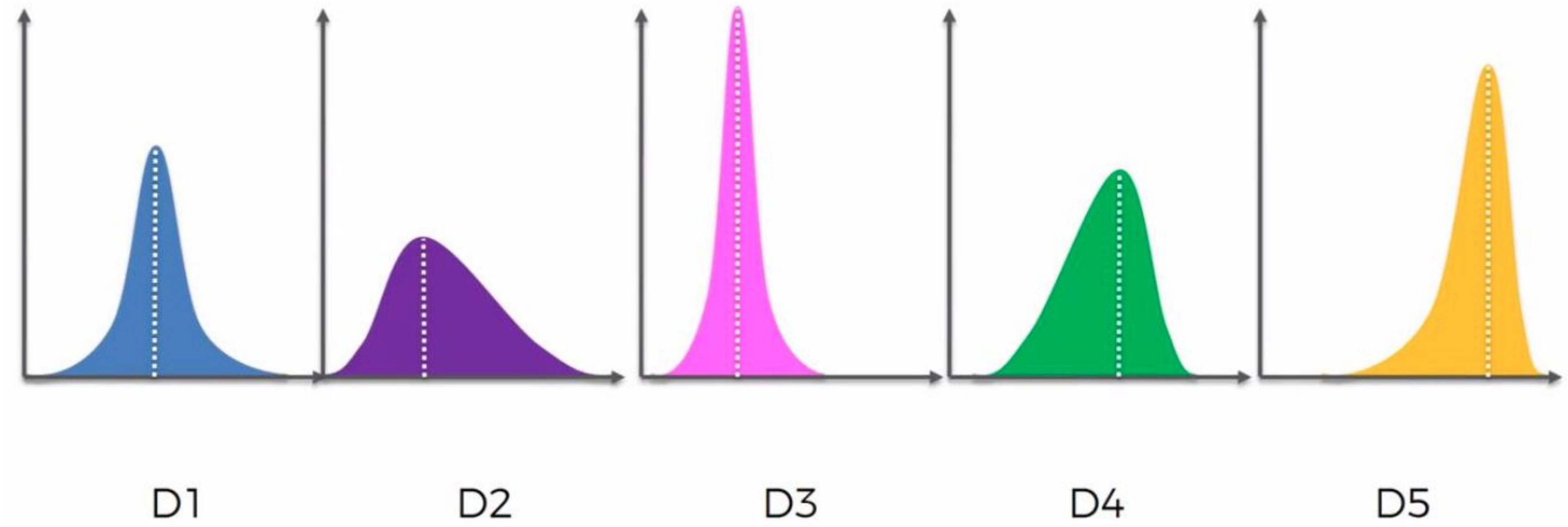
Characterised by the number of times it explores each arm



ε-Greedy Algorithm

Take the best action most of the time, random exploration occasionally.

"A randomised relative of ETC that in round t plays the empirically best arm with probability 1 − ε t and otherwise explores uniformly at random"



ε-Greedy Algorithm

Drawback:

- Choose bad arm if unlucky
- Exploration might not give the correct distribution

Better than ETC (?): no forced exploration (If we keep exploring for too long we are missing opportunities)

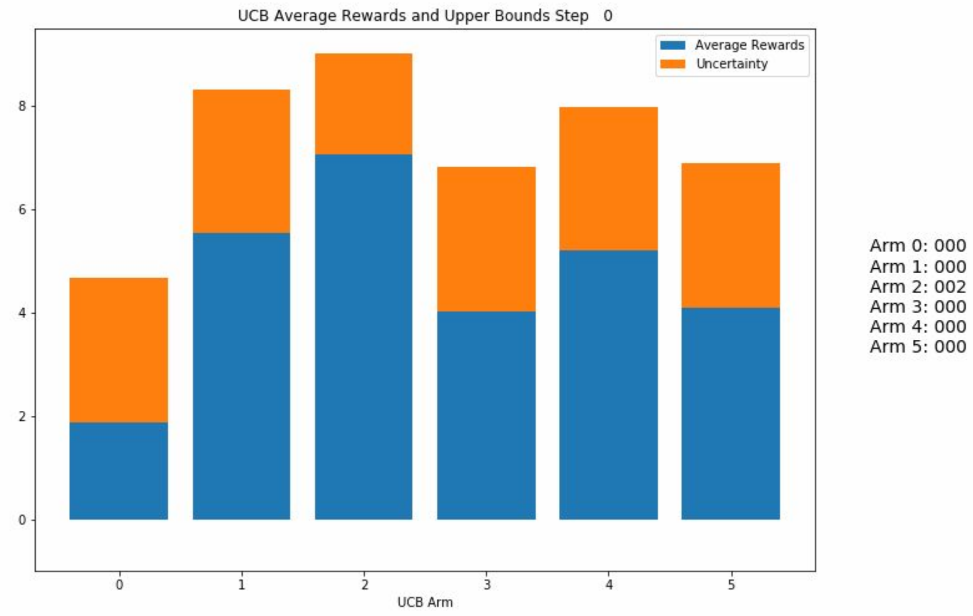
Introducing: Optimism

“The Upper Confidence Bound (UCB) algorithm is based on the principle of **optimism in the face of uncertainty**, which states that one should act as if the environment is as nice as plausibly possible.”

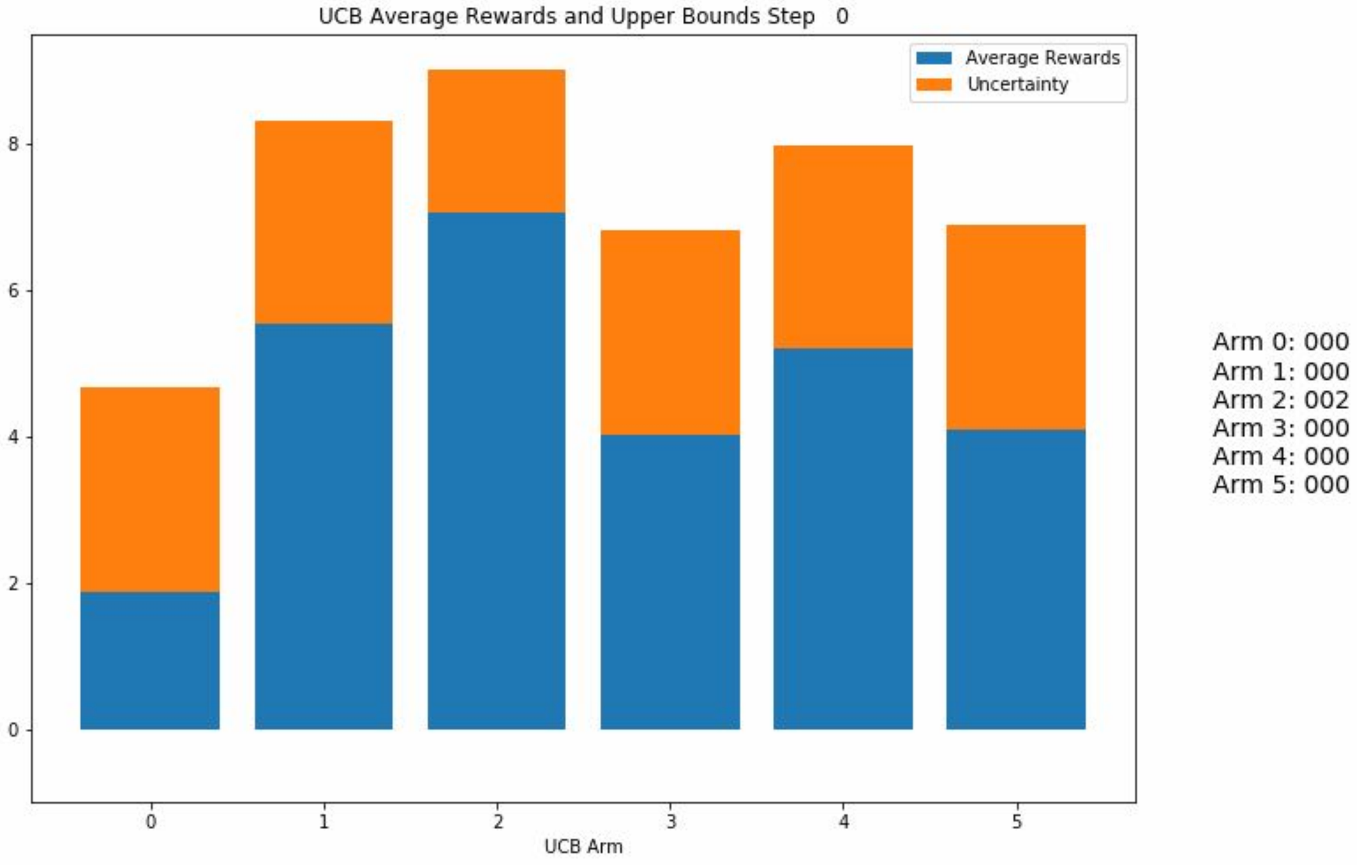
Assign to each arm a value, called the **upper confidence bound** that (with high probability) is an overestimate of the unknown mean.

Upper Confidence Bound Algorithm

In other word: favor *exploration* of actions with a strong potential to have a optimal value.

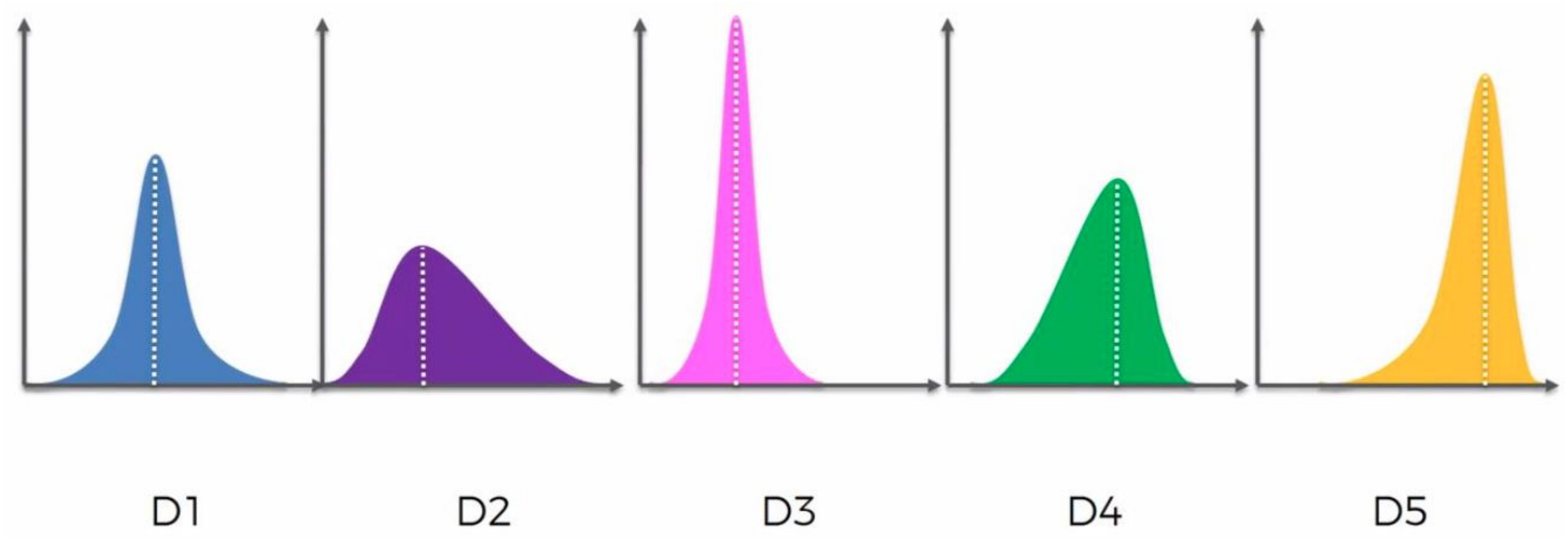


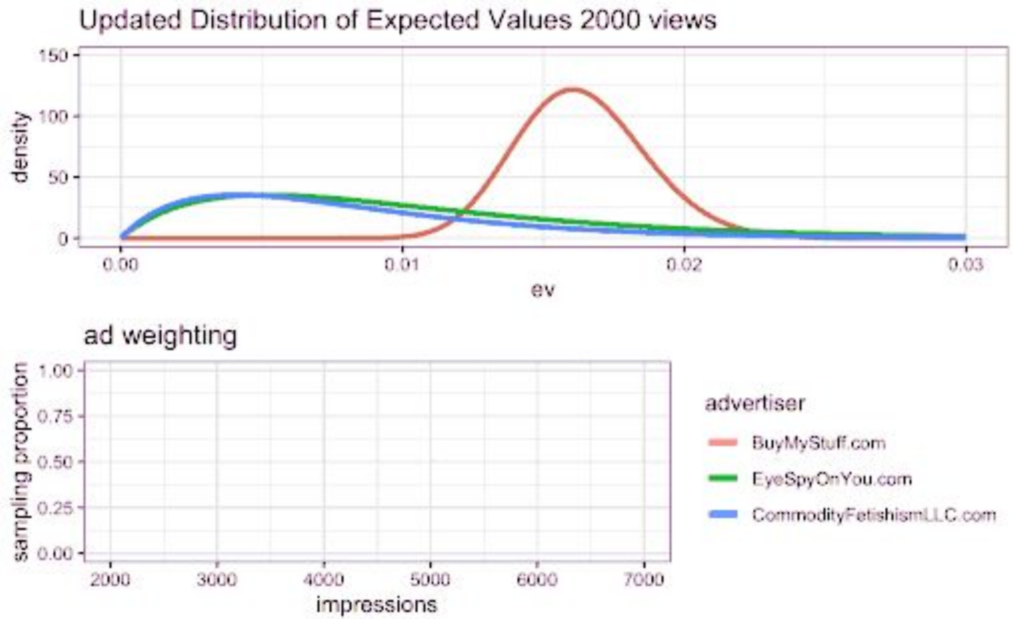
https://mwburke.github.io/images/ucb_race_gif.gif


UCB Average Rewards and Upper Bounds Step 0
Average Rewards
Uncertainty
0
2
4
6
8
0
1
2
3
4
5
UCB Arm
Arm 0: 000
Arm 1: 000
Arm 2: 002
Arm 3: 000
Arm 4: 000
Arm 5: 000

Thompson sampling

Sampling from the posterior and playing the optimal action





Updated Distribution of Expected Values 2000 views
density
150
100
50
0
0.00
0.01
0.02
0.03
ev
ad weighting
sampling proportion
1.00
0.75
0.50
0.25
0.00
2000
3000
4000
5000
6000
7000
impressions
advertiser
BuyMyStuff.com
EyeSpyOnYou.com
CommodityFetishismLLC.com

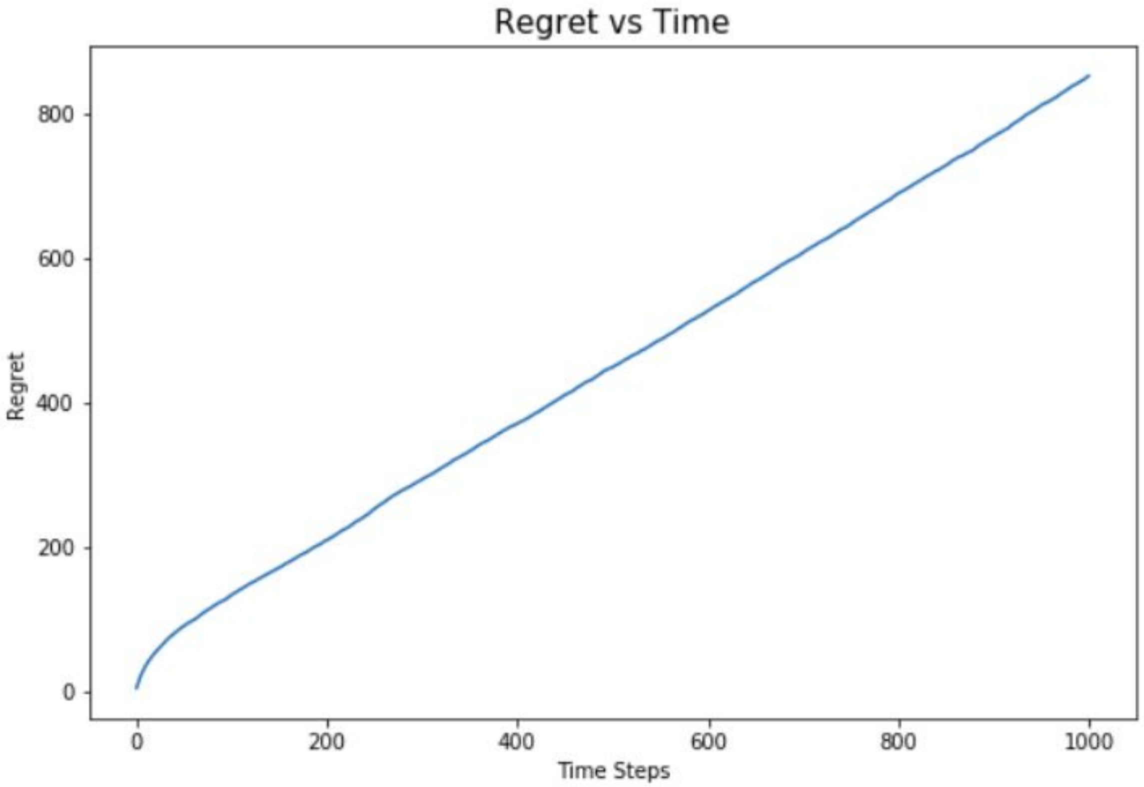
Measuring how good our algorithm is doing:

Regret (or *how better things could have been*) is the difference between the learner's action and the best action.

We aim to make the regret meaningful and small.

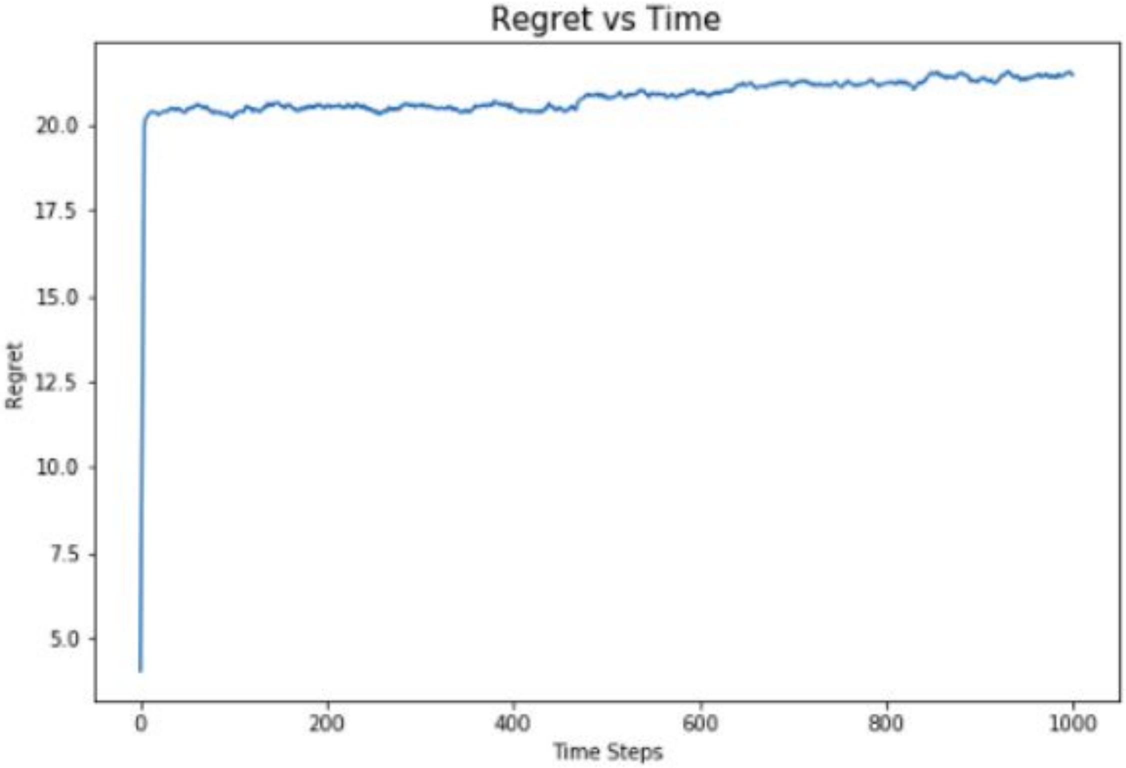
There exists a policy for which the regret vanishes (zero-regret strategies)

Epsilon-greedy - the regret



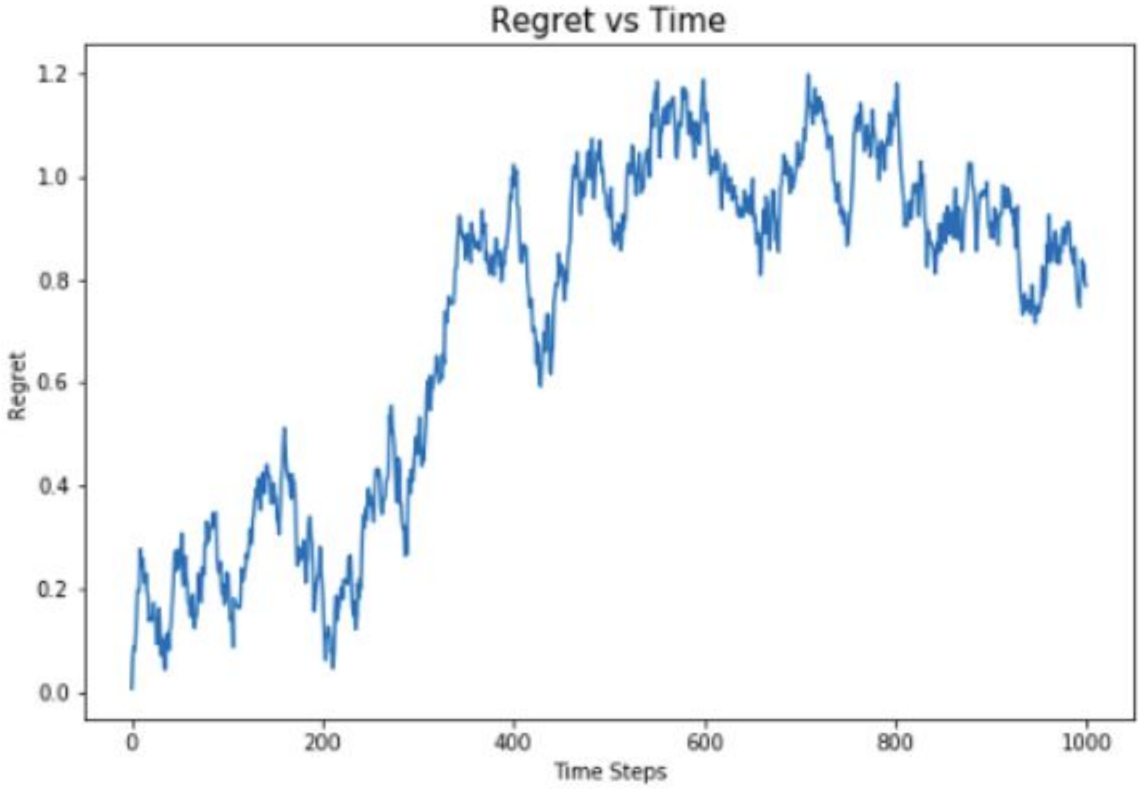
https://towardsdatascience.com/bandit-algorithms-34fd7890cb18

UCB - regret



https://towardsdatascience.com/the-upper-confidence-bound-ucb-bandit-algorithm-c05c2bf4c13f

Thompson sampling - regret



https://towardsdatascience.com/thompson-sampling-fc28817eacb8

Some formalization

Generally, a Bandit Algorithm is one with: limited number of round, agent has to receive some sort of feedback (rewards), and the agent can not peak into the future.

Stochastic bandit: each action corresponds to an IID reward (aka has an underlying distr that it samples from when an action is selected). Mean rewards do not shift (significantly) over time → *simply need to explore the arm until it can properly get the distr*

More:

- Non-stationary: relaxation of the stochastic setting (with a cost)

- Adversarial: we are in the dark. Rewards are worst-case results to throw off a learner. Strategy: randomization is key.

- Contextual: the learner has access to additional information that may help predict the quality of the actions

- Linear: reward is an unknown linear function (stochastic bandit problem can be seen as a special case of the linear bandit problem)

- Gaussian, Rotting, Restless, Dueling, Firing Bandit