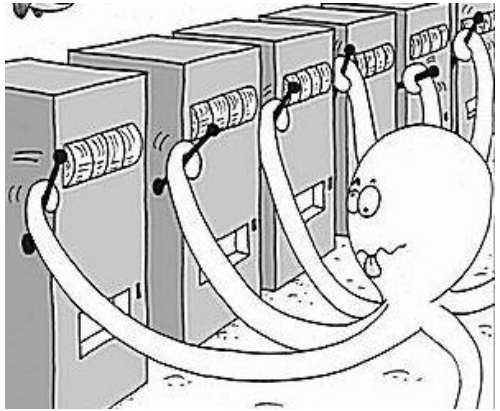


Multi-Armed Bandits

Chris Reger

Jon Courtney, Adam Richards, Frank Burkholder

Objectives

- Define Reinforcement Learning
- Explain the difference between *Exploration* and *Exploitation* and how they are related
- Explain what *Regret* is
- Discuss the following strategies:
  - Epsilon-Greedy
  - Softmax
  - UCB1
- Describe what a B*ayesian Bandit* is and how it relates to *online learning*

Review

- What is Supervised Learning
- What is Unsupervised Learning

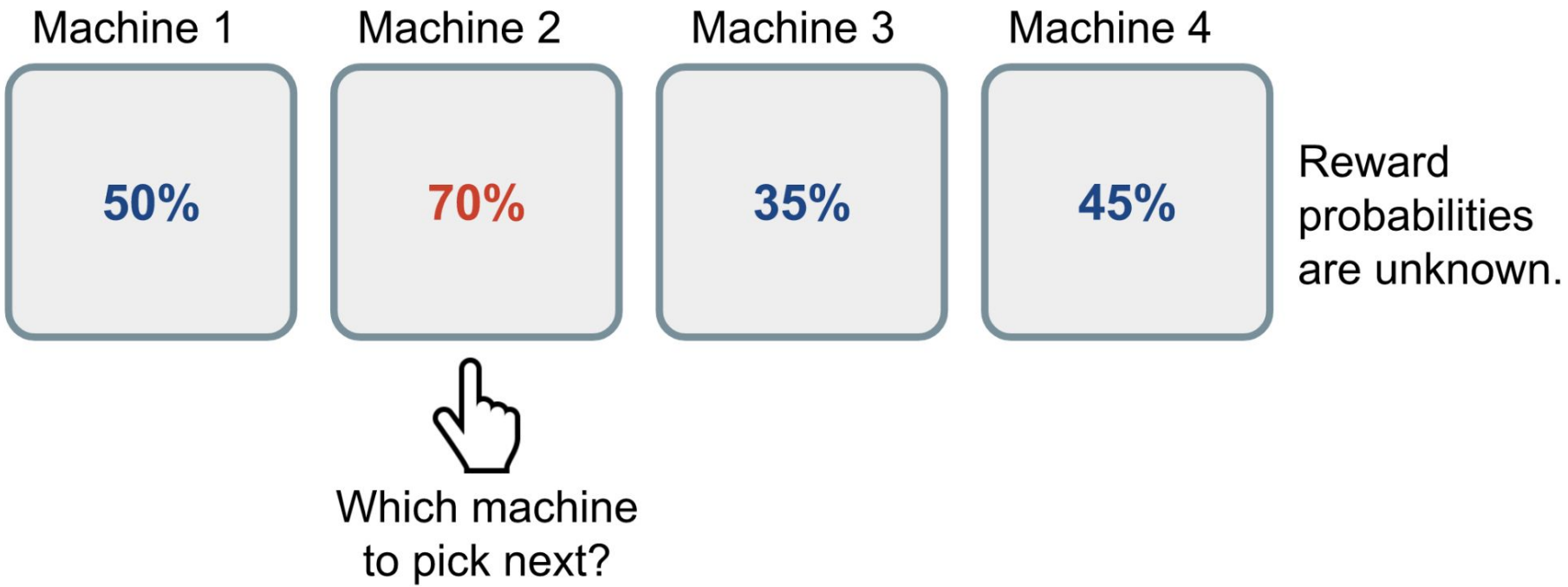
Reinforcement Learning

Definition (Wikipedia):

Reinforcement learning (RL) is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

The Bandits Problem

Suppose you are faced with N slot machines — so called “one-armed bandits”. Each “bandit” has an unknown probability of distributing a prize. (Assume for now the prizes are the same for each bandit; only the probabilities differ.) Some bandits are very generous, others not so much. Of course, you don't know what these probabilities are. By only choosing one bandit per round, your task is devise a strategy to maximize your winnings.



Applications

- **Internet display advertising**: What ad strategy will maximize sales? Naturally minimizing strategies that do not work (generalizes to A/B/C/D strategies).
- **Biology**: How do animals maximize fitness w.r.t energy?
- **Finance**: Which stock portfolio gives the highest return, under time-varying return profiles?
- **Psychology**: How does punishment and reward affect our behavior? How do humans learn?
- **Dating** : Play the field or settle down?

The Dilemma

- We don’t know the bandit with the best probability
- The task could be phrased as *Find the best bandit, and find it as quickly as possible*
- The task is complicated because of the stochastic nature of the bandits
- A poor bandit may give a good result, the best bandit might give a few bad results
- How good is good enough to be sure we’ve selected the ‘best’ bandit
- This is the *exploration vs exploitation dilemma*

Exploration

Trying out options in a search to determine the reward associated with a given bandit (acquiring more knowledge about our environment)

Discuss

What does exploration and exploitation look like in the following circumstances:

- Picking a restaurant
- Choosing an advertisement to run
- Choosing where to ski
- Choosing which slot machine to play

Comparison with A/B Testing

Consider the task of identifying the best of two websites based on click-thru rate (CTR):

- Decide your significance level and how many iterations to run, then...
- Start with pure exploration in which groups A and B are assigned equal number of users.
- Once you think you have determined the better site, switch to pure exploitation in which you stop the experiment and send all users to the better performer.

Issues with A/B

- Equal numbers of observations routed to both A and B for a preset amount of time/ observations
- Only after reaching stopping criteria do we evaluate and use better performer
- This will waste time/ money showing the less performant site
- Gives us no estimate of how likely it is that A is better or worse than B (just evidence at a given confidence level)

Multi Armed Bandit Benefits

- Shows a site that you think is best most of the time, based on what you know at the moment.
- As the experiment runs, we update the beliefs
- Run for however long until we are satisfied the experiment has determined the better site.
- Balances exploration and exploitation rather than doing only one or the other.

MODELING AND OPTIMIZATION OPPORTUNITY COSTS

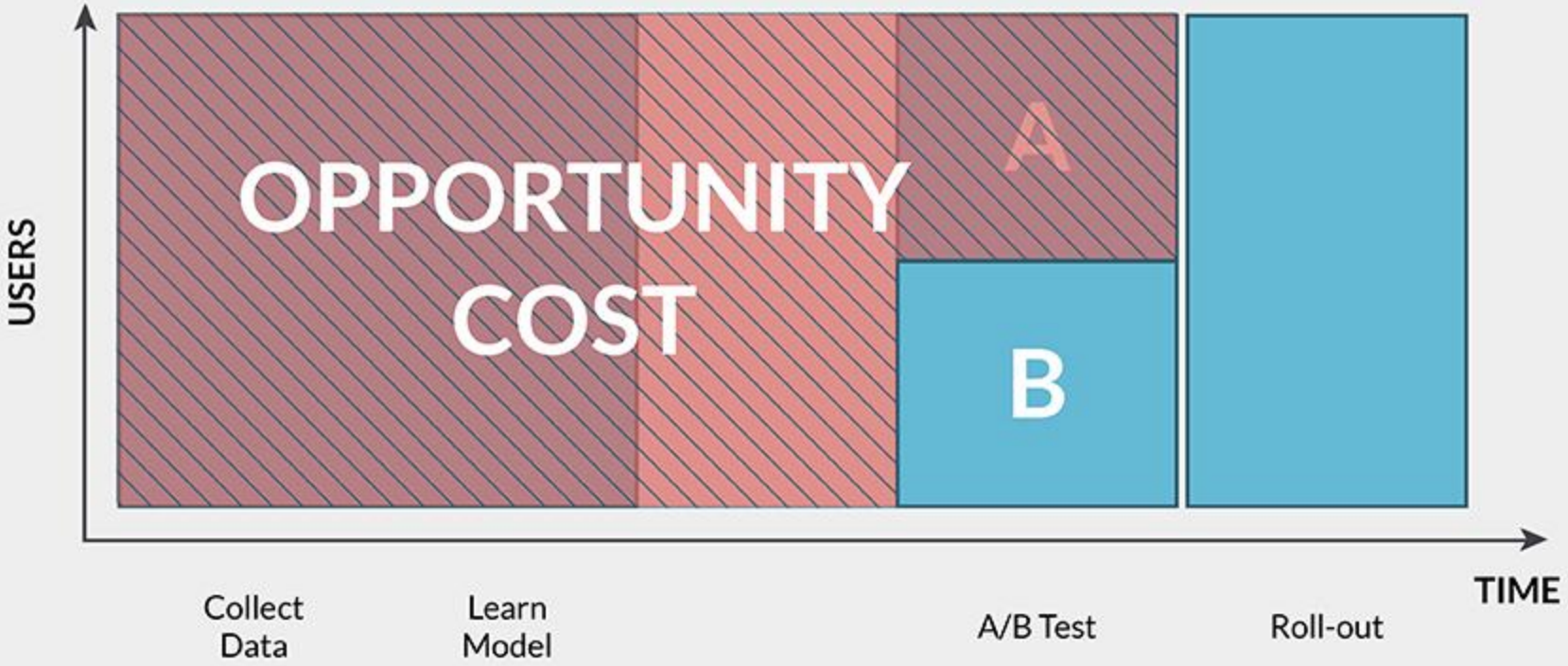
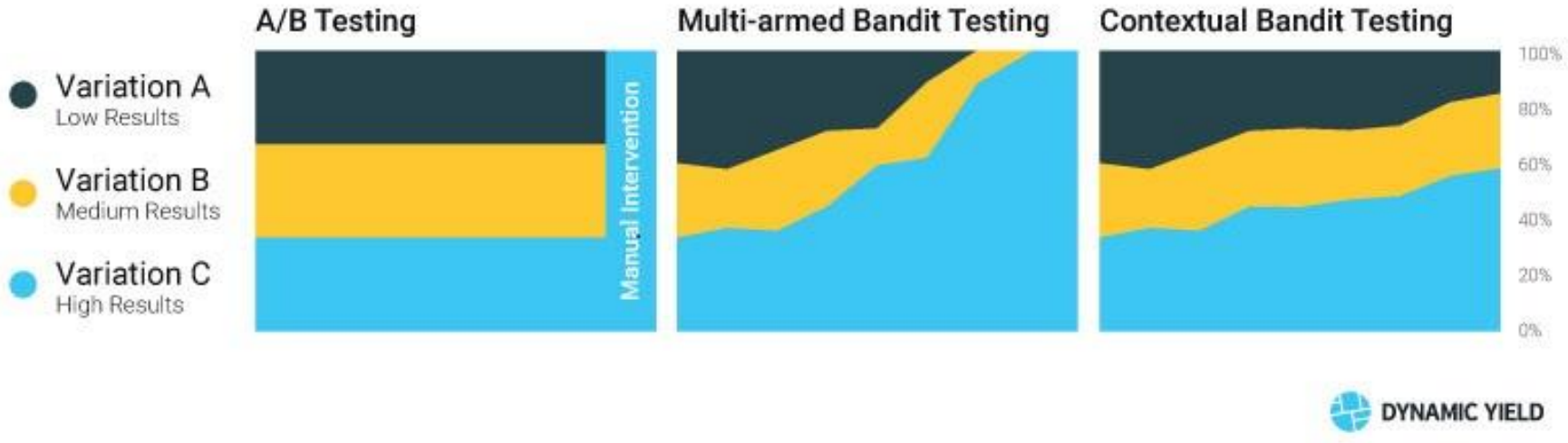


Image credit Netflix Tech Blog - **https://medium.com/netflix-techblog**

Variation Allocation in Different Test Methodologies Over Time



Formalization

- Model is given by a set of real distributions $B = R_1, \ldots, R_k$
- ...where each distribution is associated with a reward delivered by one of the K levers.
- We will let $\mu_1, \ldots, \mu_k$ be the mean values associated with these reward distributions.
- The gambler (i.e., the “agent”) plays one lever per round and observes the associated reward.
- The goal is to maximize the sum of the collective rewards, or (alternatively) to minimize the gambler/agent’s ***regret***.

Regret

- The *regret p* that an agent experiences after *T* rounds is the difference between the reward sum associated with an optimal strategy and the sum of collected rewards.
- *Regret* is simply a measure of how often you choose a suboptimal bandit. We can think of this as the cost function we are trying to minimize.

A Zero-Regret Strategy

- A zero-regret strategy is a strategy whose average regret per round ***p/T*** tends to zero when the number of rounds played tends toward infinity.
- A zero-regret strategy does not guarantee you will never choose a sub-optimal outcome, but rather guarantees that, over time, you will tend to choose the optimal outcome.

Bandit strategies

- Epsilon-greedy
- UCB1 (upper confidence bound)
- Softmax
- Bayesian bandit

and others.

“from one thing, know ten thousand things”
— Miyamoto Musashi, The Book of Five Rings: Miyamoto Musashi

Epsilon-Greedy Strategy

- Explore with some probability ε (often 10%).
- Exploit at all other times; i.e., choose the bandit with the best performance so far.
- After we choose a given bandit we update its performance based on the result.
- Exhibits linear regret for constant ε

UCB1: Upper Confidence Bound

For the UCB1 algorithm we will choose whichever bandit that has the largest value, where the value of bandit A is given as

$$\hat{\mu}_A + \sqrt{\frac{2logN}{n_A}}$$

- $\hat{\mu}_A$: the observed payout rate of bandit A
- n_A : The number of times bandit A has been played
- N : the total number of times any bandit has been played

Exhibits “optimism in the face of uncertainty” :
UCB1 gives weight to bandits that are relatively under-explored.

Softmax

- For the softmax algorithm we will choose the bandit randomly, in proportion to its estimated value relative to the other bandits

$$\frac{e^{\hat{\mu}_i/\tau}}{\sum_{j=1}^{k} e^{\hat{\mu}_j/\tau}}$$

- $\hat{\mu}_i$: the observed payout rate of bandit i
- τ : is a “temperature” parameter controlling the randomness of the choice (usually in range [0.01 - 0.1])
  - $\tau \to \infty$: Random exploration
  - $\tau \to 0$: Constant exploitation

Bayesian Bandits

The Bayesian bandit algorithm involves modeling each of our bandits with a *beta distribution* using the following shape parameters:

α : 1 + number of times bandit has won
β : 1 + number of times bandit has lost

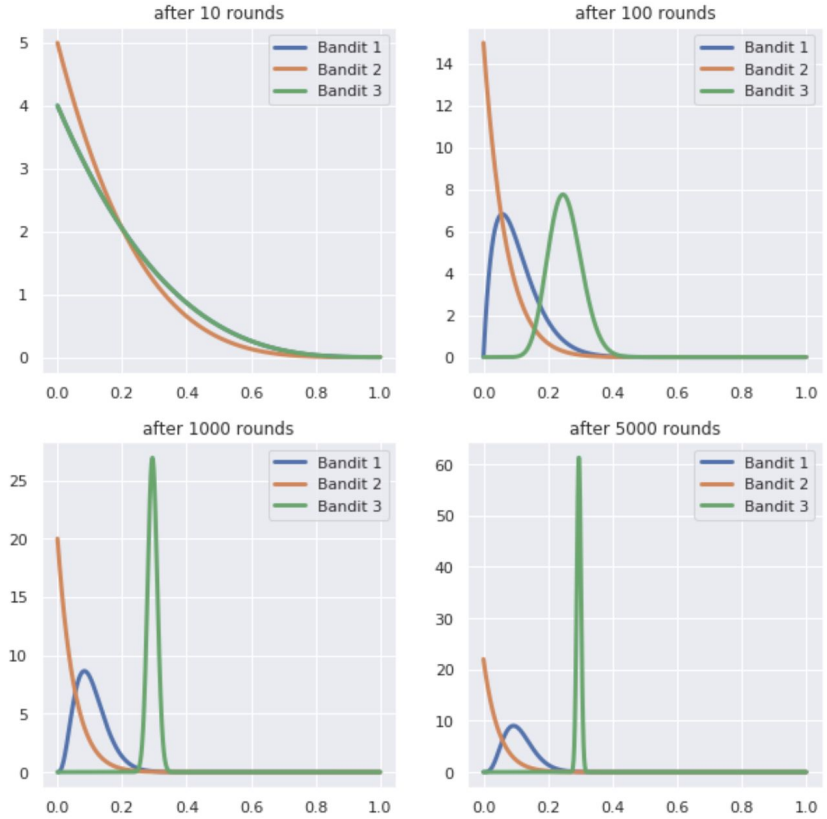
We then take a random sample from each bandit's distribution and choose the bandit with the highest value.

Bayesian bandits provide an approximately-optimal solution that scales and performs quite well.

Everyone loves a Bayesian bandit

Bayesian Bandit Performance

Sampling from 3 bandits with probability of winning [0.1, 0.2, 0.3].



Online Algorithms

These strategies are an example of *online / reinforcement learning algorithms* because they are continuously updated with new information. The algorithm starts in a state of ignorance, and begins to acquire data by testing the system. As it acquires data and results, it learns what the best and worst behaviors are. In this case, it learns which bandit is the best.

Contextual Bandits

- Like in a normal multi-arm problem, an agent must choose between arms during each iteration.
    - E.g., which online advertisement to serve
- In addition to bandit history, the agent also sees a *context vector* associated with the current iterations state.
    - E.g., the current user's profile information
- The agent uses the context vector and the history of past rewards to choose the arm to play in the current iteration.
- Over time, the aim is for the agent to learn how the context vectors relate to the associated rewards so as to pick the optimal arm.

Summary

- The problem is framed as a gambler/agent confronted with a row of slot machines (*aka one-armed bandits*)
- The agent has to decide which machines to play, how many times to play each machine, and it which order.
- Each bandit will provide a reward from an unknown distribution
- Objective is to maximize the sum of rewards earned through a series of lever pulls
- There are several classes of fairly optimal solutions

Review Questions

- What is *reinforcement learning*?
- What is the difference between *exploration* and *exploitation*, and how are they related?
- Explain (in plain English) what *regret* means in a bandits context?
- Explain the following strategies: *epsilon-greedy, softmax,* and *UCB1*.
- What is the *Bayesian Bandit* strategy and how is it an example of *online learning*?