Exploring the explore-exploit trade-off

Matthew Allcock - University of Sheffield

1 Plan

1.1 Details

This competition is open to Early Career Mathematicians (ECMs), with a prize of £250. We invite articles on any mathematical topic, including pure, applied, teaching, case studies, etc. between 1,000–2,000 words.

1.2 Possible titles

- Dora the exploit or explorer
- Dora the exploiter
- Dora the explorer or exploiter
- When should Dora explore?
- Exploring the explore-exploit trade-off
- To explore or to exploit

1.3 Structure

- Abstract
- Introduction: Examples of explore/exploit - Exploring possible partners, restaurants in a new city, career options to follow, reinforcement learning. We want a decision theory for when to explore and when to exploit

- modelling the trade-off - Multi-armed bandit problem.
- Lower bound - naive approach - see HAAISS notes. Expected value of information for next turn. Explain why this is a lower bound.
- Upper bound - Expected value of perfect information
- True expected value
- Markov chain monte carlo simulation - link to github repo?
- Conclusion

2 Abstract

3 Introduction

We are often faced with a decision between several options where we are uncertain as to how good each option is. We can *explore* the space of options to learn more about . However, exploring comes at a cost of time, money, energy, or all three. At some point, we will want to *exploit* the best option we know about, avoiding the costs of further exploration and reaping the rewards of what we believe to be the best option. Explore too long and you pay the opportunity cost of not exploiting the best known option. Exploit too early and you might be committing to a sub-optimal choice.

Let's say you move to a new city and want to find a good hair-dresser. One option is to try them all sequentially, then choose the one that gave you the best haircut to be your regular. This would work in a small town, where the option-space is small. You might only need to try three hair-dressers to exhaust all the options. You'd better hope that one of them did a good job. What if you move to a large city, where you have a hundred to choose from? It would take may years to explore them all. How should you choose when to stop exploring and to start getting your haircuts from the best known hair-dresser?

But it's not just a fun exercise in choosing who's going to make you look sharp. Understanding the explore-exploit trade-off is important for life biggest decisions. How long should you spend exploring different career options before going all-in on the one that think you can be most successful

in? How many people should you date before committing to a long-term partnership? How long should a mathematician explore the space of possible open research questions before tackling one? All of these examples involve a trade-off between exploring the space of options and exploiting the best known option.

Intuitively, we would expect that the optimum decision procedure is to explore the option-space until the expected value of exploring another option is less than the incurred cost. Let's put this trade-off on a mathematical footing.

4 Modelling the trade-off

Here's a game that models the explore-exploit trade-off. Each player draws a number from an unknown probability distribution, without seeing what other players draw. On every turn, each player can choose to either draw again or stop drawing altogether. The winner is the player with maximum score, which is calculated as the player's maximum draw minus a constant cost per number of draws taken.

(Discussion of why its a trade-off)

Here's the mathematical framework:

Let N be the number of players and X be a random variable over the real numbers with parameters unknown to each player. Player i's jth draw is given by x_{ij}, each of which is a realisation of X. Let c be the cost associated with drawing. Then we can define Player i's score after n draws as

$$s_{in} = m_{in} - nc, \quad \text{where} \quad m_{in} = \max_{1 \leq j \leq n} \{x_{ij}\} \tag{1}$$

Once all players have stopped drawing, the winner is Player i, whose final draw was draw number n_i and whose final score satisfies

$$s_{in_i} > s_{kn_k}, \quad \forall k \neq i. \tag{2}$$

The aim of this paper is to discuss optimal decision procedures - how to win the game!

This setup shares similar features to the multi-armed bandit problem in economics, the secretary problem in statistics, and optimal foraging in ecology, all of which has some interesting results and strategies (REFERENCES?).

5 When to draw: lower bound

Before the first draw, we know nothing about the underlying distribution. So we might have a uniform prior over the real numbers[1]. Then on each successive draw we can update our prior to a posterior distribution which tells us what we can expect the underlying distribution to be based on the draws taken. For the next draw, the previous posterior becomes the new prior. Each draw reduces our uncertainty about the underlying distribution from which we are drawing.

After the nth draw, to determine whether it is better to draw another or not we must determine the expected value of the score (s_n for brevity) after the next draw. To do this, we must determine a prior distribution from which we expect that we are drawing from. Notate by X'_n a random variable that follows this distribution, with probability density function f_n. Then we can expect the score after the $(n+1)$th draw to be

$$\mathbb{E}(\text{draw}) = \mathbb{E}(\max\{m_n, x'_{n+1}\}) - (n+1)c. \tag{3}$$

and if we choose not to draw then the score will be

$$\mathbb{E}(\text{no draw}) = m_n - nc. \tag{4}$$

But here we are only considering the value of the draw in increasing the score in the $(n+1)$th draw, as if we were definitely going to stop. We are missing the value of the new information about the distribution on all future draws.

6 When to draw: upper bound

7 Simulating the trade-off

8 Conclusion

[1]In reality, it would be more appropriate to have a weakly informative prior based on what distribution you would expect the game designer to select. For example, reasoning from our bias towards small numbers would tell us that a mean of 10 is more likely than a mean of 10^{100}. this would also avoid some of the hiccups that we might face when dealing with an improper prior.