Decisions, decisions

Redjon Cacani

The holiday

The concept is simple; you're on holiday in a secluded Albanian town and want to have the best experience possible. This best experience can be modelled on how much happiness you obtain throughout the trip on the decisions you make. Your choice of hotel, daily activity and even restaurant can change your final total output of happiness (FH). Focusing on the restaurant aspect, the town has 3 different restaurants: R_1 R_2 R_3. The decision of which to pick each day whilst on holiday is relatively blind with limited reliable information meaning you're unaware of:

- The possible average happiness gained from each restaurant
- The standard deviation of each (how bad or good their off days can get)

This form of decision making is characterised by the Multi Armed Bandit problem, where multiple options are presented to you, and you are blind to the expected reward and standard deviation of each; however, you can sample the choices multiple times before you run out of passes. In this problem, you will have to allocate a specific number of passes, or number of days in this case, to: exploring your options in attempt to find the best one overall, and, exploiting which option you believe is best to maximise your total output. The question is just how many passes should you dedicate to each process and how should you split them throughout your time frame?

Reverting to the restaurant analogy, we can model strategies for the specific scenario by first setting some parameters and key assumptions.

This holiday will be relatively lengthy, and you will spend 90 days on which you go to a restaurant once per day. Each visit is modelled as an independent draw from a fixed distribution.

Total Time Frame, TF = 90

The values of the average happiness you would gain, and standard deviation of each restaurant are hidden from you as you make your choices but are useful in modelling the outcomes of each strategy you may decide to take ,displaying how effective they will be. We assume the happiness gained from each restaurant will be measured on a positive scale and happiness gained (x) cannot be negative.

Where a = restaurant number

HR_a= average happiness gained, σ_a= standard deviation

HR_1= 10 σ_1= 4

HR_2= 7 σ_2= 1

$$HR_3 = 5\ \sigma_3 = 2$$

With the values for each option set, we can model the expected average maximum happiness output as:

$$MaxH \approx TF(HR_1) \approx 900$$

This displays the average expected output if you knew the data for each restaurant and chose the best option for the entire time frame. The actual maximum output will vary due to standard deviation of results; however, we will use MaxH.

To quantify how effective our strategy will be we can calculate how far our FH is from MaxH, modelling this as regret, ρ. The larger ρ, the worse the strategy.

$$\rho = 900 - FH$$

Now that our parameters and assumptions are established, we can look at the simplest ways to approach the problem.

Strategy A - Explore only

This strategy consists of randomly selecting a restaurant each day ignoring new information that would otherwise influence our choice. In expectation, each restaurant will be visited a third of the time, 30 days. This can be modelled as:

$$FH \approx 30HR_1 + 30HR_2 + 30HR_3$$

$$FH \approx 660$$

$$\rho \approx 240$$

This strategy results in an average approximate regret of 240, 26.7% of the maximum output, dependent on variation of happiness results and the allocation of visits to each restaurant.

Strategy B – Exploit only

This strategy consists of exploring each option once and then exploiting the best sample outcome for the remaining time. To understand its potential, we can consider an idealised scenario where variation is ignored and each restaurant always produces its average, meaning we correctly identify R_1 as best and exploit it for the remaining 88 days:

$$FH = 88HR_1 + HR_2 + HR_3$$

$$FH = 892$$

$$\rho = 8$$

Although displaying the regret being significantly lower at 8, 0.9% of the maximum output, our problem (and reality) does not have these properties and includes standard deviation.

When accounting for standard deviation the “best” restaurant in the first three sampling days can change depending on the probability of the restaurants falling behind and exceeding their

average outputs. More specifically the probability that the best restaurant (R_1) does not have the highest output in the first 3 days and therefore is not the one that is exploited.

Assuming that the restaurants follow a Gaussian distribution, we can model graphs of each restaurant's probability density $f(x)$ against their happiness output x as a probability density function (PDF curve) to visually understand our restaurant values:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-HR}{\sigma}\right)^2}$$

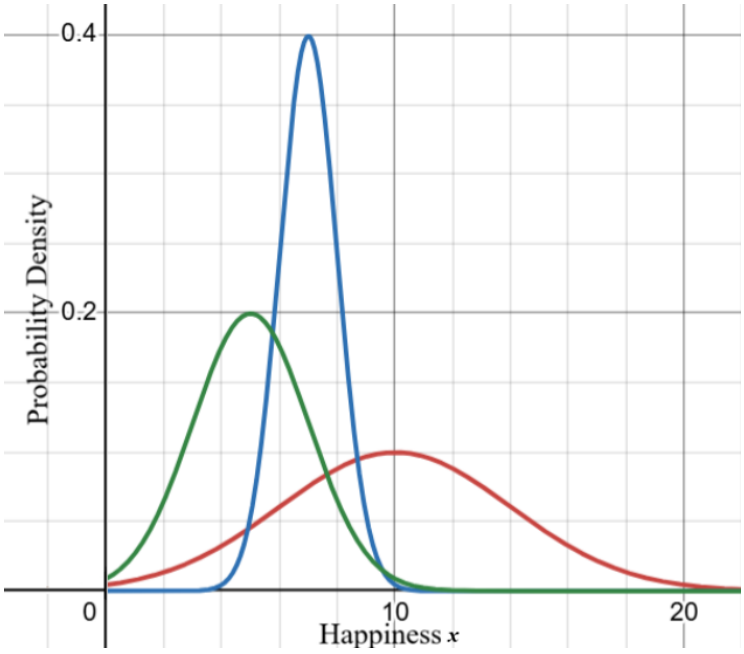
Where ***HR*** is the average happiness output and $\boldsymbol{\sigma}$ is the standard deviation, dependent on restaurant.

This however does not account for our previous assumption that the happiness gained (x) is positive. This means we have to truncate the curve at $x > 0$ and usually renormalise by dividing by $\boldsymbol{P}(\boldsymbol{X} > 0)$ to preserve the total probability, removing all probability mass below 0 and then rescaling the remaining probabilities giving:

$$f_T(x) = \frac{f(x)}{P(X > 0)}, \quad x > 0$$

Where $\boldsymbol{P}(\boldsymbol{X} > 0)$ is the probability that the x output is larger than 0

However, in our scenario the probability of R_1 ,R_2 ,R_3 having negative outputs are $\sim 0.006, \approx 0, \sim 0.006$. This means that while we can renormalise the curve, the values of probability below 0 negligible and we ignore negative outputs plotting $f(x)$ above 0:

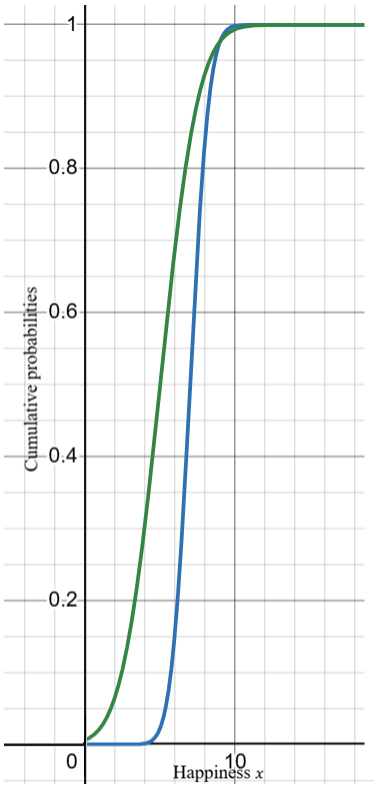


R_1 ,R_2 ,R_3

By interpreting the curves, we can see that due to restaurant 1 having a higher standard deviation compared to the other options, it has a flatter tail on the left and right of the mean. Visually this shows that there are instances in which R_1 does not have the highest output when sampling each restaurant.

This graph was sketched using Desmos and the function: normaldist(HR_a,σ_a)

To identify the probability of these instances occurring, and thus the exploit strategy being increasingly suboptimal, we must first calculate the probability that R_1 has the highest output in the first three passes and subtract that from 1. To do this we will find the probability that R_1 lands on the value x and R_2 and R_3 land on values lower than x ,in which we must introduce the cumulative distribution functions (CDFs) of R_2 and R_3 ,being F_2 and F_3. The CDF curve shows the probability of a variable having a value between 0 and X_a in this case. F_2 and F_3 are sketched below with the following equations:



$$R_2 \text{ value} = X_2$$
$$R_3 \text{ value} = X_3$$

$$F_2(x) = P(X_2 \leq x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-7}{\sqrt{2}}\right)\right]$$

$$F_3(x) = P(X_3 \leq x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-5}{2\sqrt{2}}\right)\right]$$

Where $\boldsymbol{P(X_a \leq x)}$ is the probability the R_a value is $\leq R_1$ value x.
The Gaussian curve's integral cannot be solved using basic algebra, so we introduce the error function (erf). This function is defined as:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

This is simply a standard function used to compute cumulative probabilities of normal distributions.

This graph was sketched using Desmos and the function: normaldist(HR_a,σ_a).cdf(x)

Now that we have equations for the cumulative distribution functions of R_2 and R_3 , we can calculate the probability of R_1 having the highest output in the first 3 days at a single moment by multiplying the probability that R_1 produced the specific happiness output x (The PDF curve of R_1) by the probability that R_2 and R_3 produced values lower than x (the CDF curves of each). This shows how concentrated the chances of R_1 being the highest are at a specific happiness output e.g. 12. This is displayed by the function:

$$P(R_1 \text{ is highest output at } x) = f_1(x) \cdot F_2(x) \cdot F_3(x)$$

To calculate the probability of R_1 having the highest outcome in the first 3 days as a whole we must integrate the function above from 0 to ∞:

$$P(R_1 \text{ is highest output overall}) = \int_0^\infty f_1(x) \cdot F_2(x) \cdot F_3(x)\,dx$$

This integral has no closed-form solution, so we evaluate it numerically using a method called quadrature. This gives $\boldsymbol{P}(\boldsymbol{R}_1$ is highest output overall) ≈ 0.7492, approximately 74.9%. Meaning the probability that we choose the wrong restaurant to exploit after our first 3 exploration days is 25.1%. This shows the greedy exploit strategy is sub-optimal 25.1% of the time.

To find the realistic range of values for regret (ρ), we first find the probabilities of R_2 and R_3 having the highest output in the first three days. To do this we use the same logic as before, integrating the function made with the PDF curve of the restaurant which would be the highest and the CDF curves of the other two, all multiplied together. This gives 0.1953 for R_2 and 0.0555 for R_3 (19.5% and 5.6% approximately).

Now we can model the regret for each scenario. Assuming that since each restaurant is sampled once in the first three days, the expected total happiness for those days is the sum of their means, which is 22. Although outcomes vary, we expect these deviations to balance out, so we approximate the total as 22 regardless of which restaurant performs best. In the case R_1 is exploited the average expected regret ρ will be 8, as established previously. In the scenario R_2 is exploited, the remaining 87 days will be spent there at an average output of 7 giving the final happiness and regret as:

$$FH \approx 87HR_2 + 22 \approx 631$$

$$\rho \approx 269$$

In the scenario R_3 is exploited:

$$FH \approx 87HR_3 + 22 \approx 457$$

$$\rho \approx 443$$

Now we take an average expected regret for the strategy as a whole by weighting them by their probabilities:

$$\text{Average expected } \rho \approx 0.749\times8 + 0.195\times269 + 0.056\times443 \approx 83.255$$

*An important nuance is that this is a weighted average prediction, not what regret you will obtain from the strategy.

With our results finally obtained we can observe that on average the exploit strategy is a much better choice than explore only, for this problem specifically, however, does exhibit high variance, with very low or very high regret possible. Inferring this, we can deduce a strategy with more exploitation and less exploration is better for less variation and vice versa. An optimal strategy for all problems, due to variance of options being hidden, would be one that balances exploration and exploitation, adapting as more information is gained and confidence is built.

Upper confidence bound 1

The upper confidence bound 1 (UCB) strategy is widely used within MAB mathematics as a strategy that balances explore/exploit processes automatically and accounts for uncertainty. Similarly to the exploit only strategy, the UCB method begins by sampling each option once, but unlike it, continues to explore based on uncertainty rather than committing immediately. It does this by favouring restaurants that performed well historically, whilst still giving a chance to restaurants with uncertainty. Mathematically the equation is shown as:

$$\mathrm{UCB}_a(t) = \bar{x}_a(t) + \sqrt{\frac{2\ln t}{n_a(t)}},$$

Where t is the number of days passed, $\bar{x}_a(t)$ is the average reward from restaurant a up to day t, $n_a(t)$ is the number of times restaurant a has been chosen up to t.

We select the restaurant that gives the highest $\mathrm{UCB}_a(t)$ value for our next choice.

The first term represents exploitation; it prioritises options with higher observed rewards. The second term (formally known as Hoeffding bound) acts as an uncertainty balance, which is larger for restaurants that have been chosen fewer times, encouraging exploration of lesser-known choices. As n_a increases, this bonus shrinks, meaning the strategy gradually becomes more confident in its estimates and focuses on the best-performing options.

Intuitively, UCB chooses the option with the highest optimistic estimate of its true value. Rather than treating unknown options as risky, it temporarily treats them as potentially better than they appear, ensuring they are explored sufficiently before being ruled out, while still accounting for observed rewards. Unlike explore only, it avoids wasting time on poor options, and unlike exploit only, it avoids premature commitment.

Conclusion

Ultimately, the problem is not about choosing the option with the highest observed reward but is centred around managing uncertainty. The optimal strategy does not treat exploration of options as a waste of resources but as an investment to measure uncertainty and build confidence overtime, exploiting the true best option.