The Bandit Report

Written by Brandon Rozek

Consider the following problem. You are at a casino, faced repeatedly with the choice of picking a slot machine to play on. You can't stop until you have played 1,000 times.

It's really a conundrum, as addictions are hard to break. Instead of trying to quit playing after each turn, another thought appears in your mind.

"How can I maximize the expected total reward that I'm going to get?"

This is the bandit problem.

Formulation of Ideas and Initial Estimation

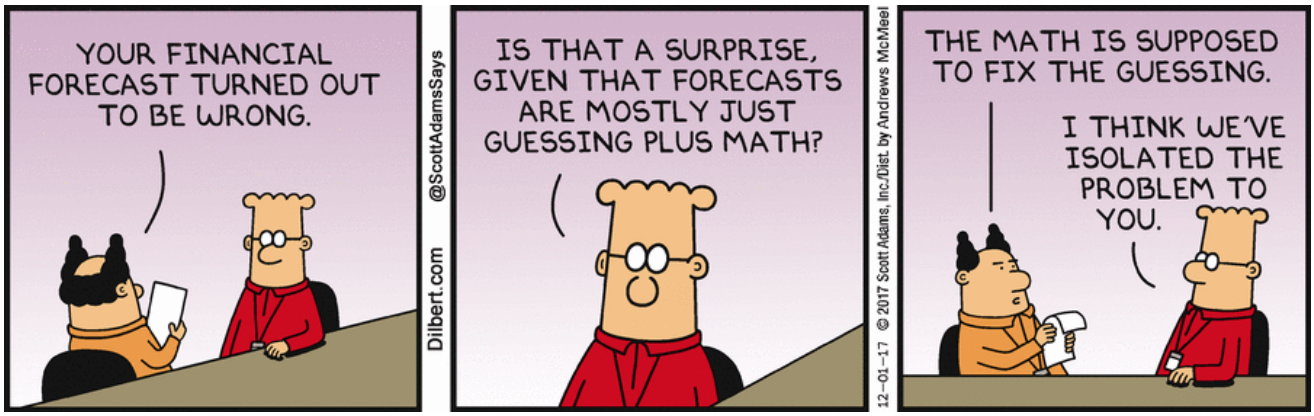
There are two key elements that will determine the success of your gambling trip:

- How good you are at estimating what a machine will give you.
- How you chose which machine to play on.

These two criteria are formally called *value estimation* and *action selection* respectively.

You walk up to the row of slot machines, otherwise called bandits. What's your *initial estimation* of the value of each of the machines? You've never played on any of them before and actually have no clue what to expect.

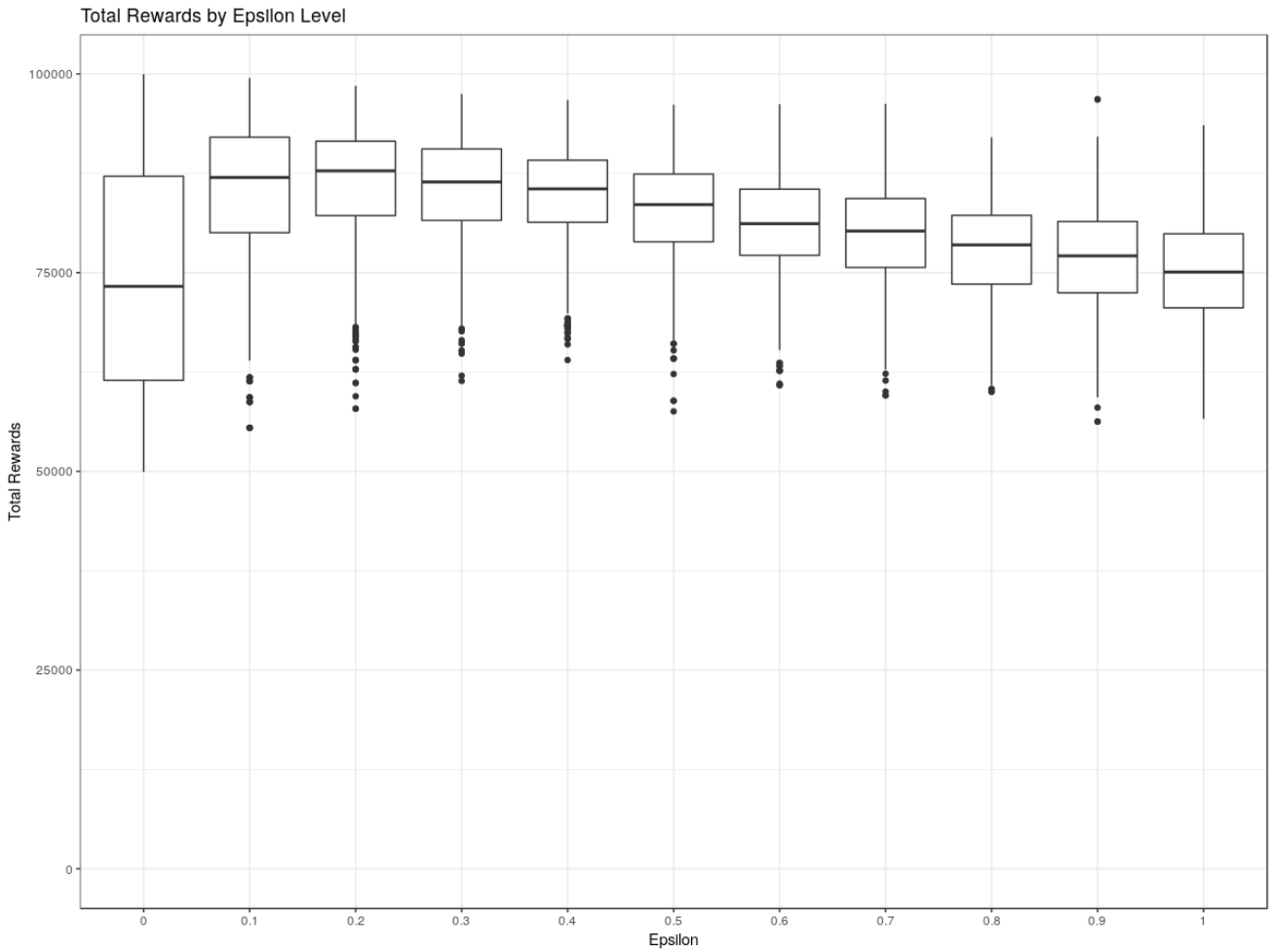
Now you start thinking about chickens and eggs. Man, you wish you had remembered to have eaten dinner before this. No worries, however. If your job has taught you anything: guessing reigns supreme.



Now that you have pretended to have a good understanding of what each bandit will give you, you need to act on your decision. Wait, you've spent all this time deliberating and no bandit was chosen yet!? Which one are you actually going to pick?

You feel the fabric of space-time morphing around you and multiple time lines of what you could do...

Choosing Greedily Sometimes...

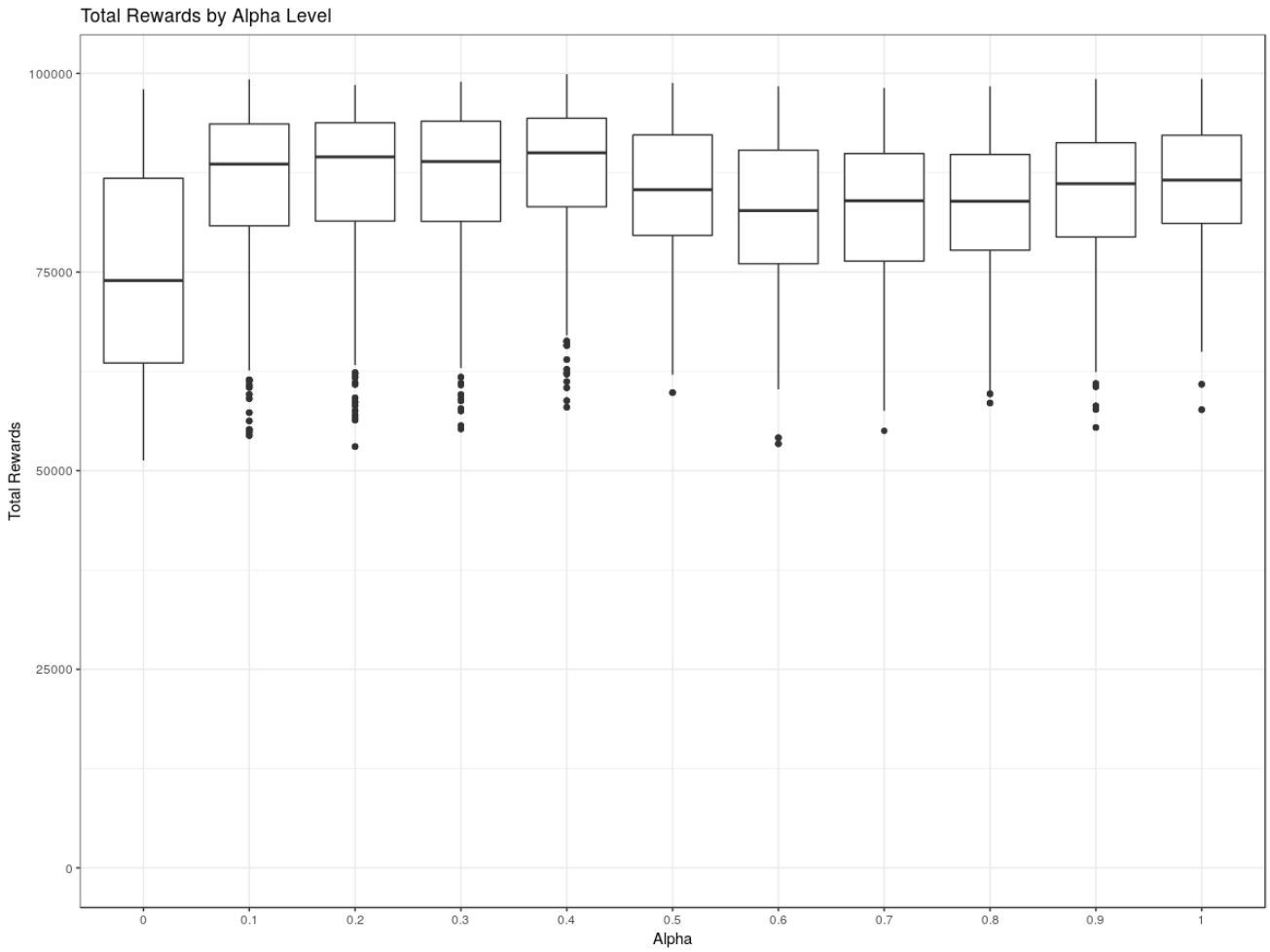


You have a guess at what each machine will provide to you. Why don't you just choose the best one each time? That is the world in which ϵ is zero otherwise known as *greedy action selection*. You watch multiple versions of yourself play with the bandits. Sometimes they hit it big, sometimes they don't. There appears to be a large variation in whether or not you succeed.

It's all about risk, and choosing the best each time is risky. The reason why is because you don't *actually* know what any of the machines' values are. You just have a guess. This leads to never exploring the different bandit options before you, which can be a good or bad thing. If you're exploring all the time $(\epsilon = 1)$, you never get to exploit the bandit with the highest expected reward.

The best case is exploring sometimes, but choosing greedily the rest of the time. This is formally known as *greedy-epsilon selection*. By watching your fellow selves play the slots, you realize the sweet spot for exploration is around 20% of the time. That gives you low variation, meaning you can expect to get around the average reward for that epsilon and, on average, you receive higher rewards from exploiting the bandits.

Reestimating Values



Okay, you have now played a bandit. Five dollars spit out of the machine. Congratulations on your win! Does that change your perspective of the bandit?

I mean, it should, it gave you some money! The amount you got may have been higher or lower than you expected, but you received something nevertheless.

How do you resolve the initial estimate, the amount you just received, and any future amounts?

One option is to go with the arithmetic mean. Sum up all of the rewards you have received so far for each bandit and divide by the number of times that bandit has been used.

$$value_i = \frac{\sum_{j=1}^{n_i} reward_{j,i}}{n_i}$$

Where i is the bandit that you're currently considering and n_i is the number of times you played on said bandit.

It's headache-inducing to calculate this after each play, so you decide to derive a simpler formula

$$
\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}(R_n + \sum_{i=1}^{n-1} R_i) \\
&= \frac{1}{n}(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i) \\
&= \frac{1}{n}R_n + (n-1)Q_n \\
&= \frac{1}{n}(R_n + nQ_n - Q_n) \\
&= Q_n + \frac{1}{n}(R_n - Q_n) \qquad (2.3)
\end{aligned}
$$

That is so much better. Now you just need to remember the previous reward, how many times you played with the bandit, and what you just received from that bandit.

You decide to take a break. I mean after all, you just did math right? While eating some chicken and eggs, you immediately begin thinking about the problem at hand. How many hours has it been since you got here? Maybe you should contact some family or friends.

No, no, no, it's too early to stop. You must continue. Looking at the value estimation formula you derived, a sudden spark of brilliance emerges. This looks familiar...

$$NewEstimate = OldEstimate + StepSize(Reward - OldEstimate)$$

In the case of the mean, your $StepSize(\alpha)$ is $\frac{1}{n}$. Means have the nice property of converging to what we would consider the "true" value after experiencing a bandit many times. What if we tweak this, though?

To put an arbitrary bound on our problem, let's say we're considering step sizes between 0 and 1 inclusive. Start with a mental exercise. If our step size is zero, what would the new estimate look like?

$$NewEstimate = OldEstimate$$

It seems that in this way, your estimate never changes. Probably not the smartest way to play. (You confirm this by watching a version of you play many times with the bandits.) Yup, reminds me of the $\epsilon = 0$ decision previously.

Now let's consider if $\alpha = 1$.

$$
\begin{aligned}
NewEstimate &= OldEstimate + Reward - OldEstimate \\
&= Reward
\end{aligned}
$$

I like to consider this extreme short term memory. With this alpha level, you may as well forget all the rewards gotten previously from a bandit besides the last one. Probably a good strategy if you want to keep your happiness, but honestly, we both know you're in it to win it.

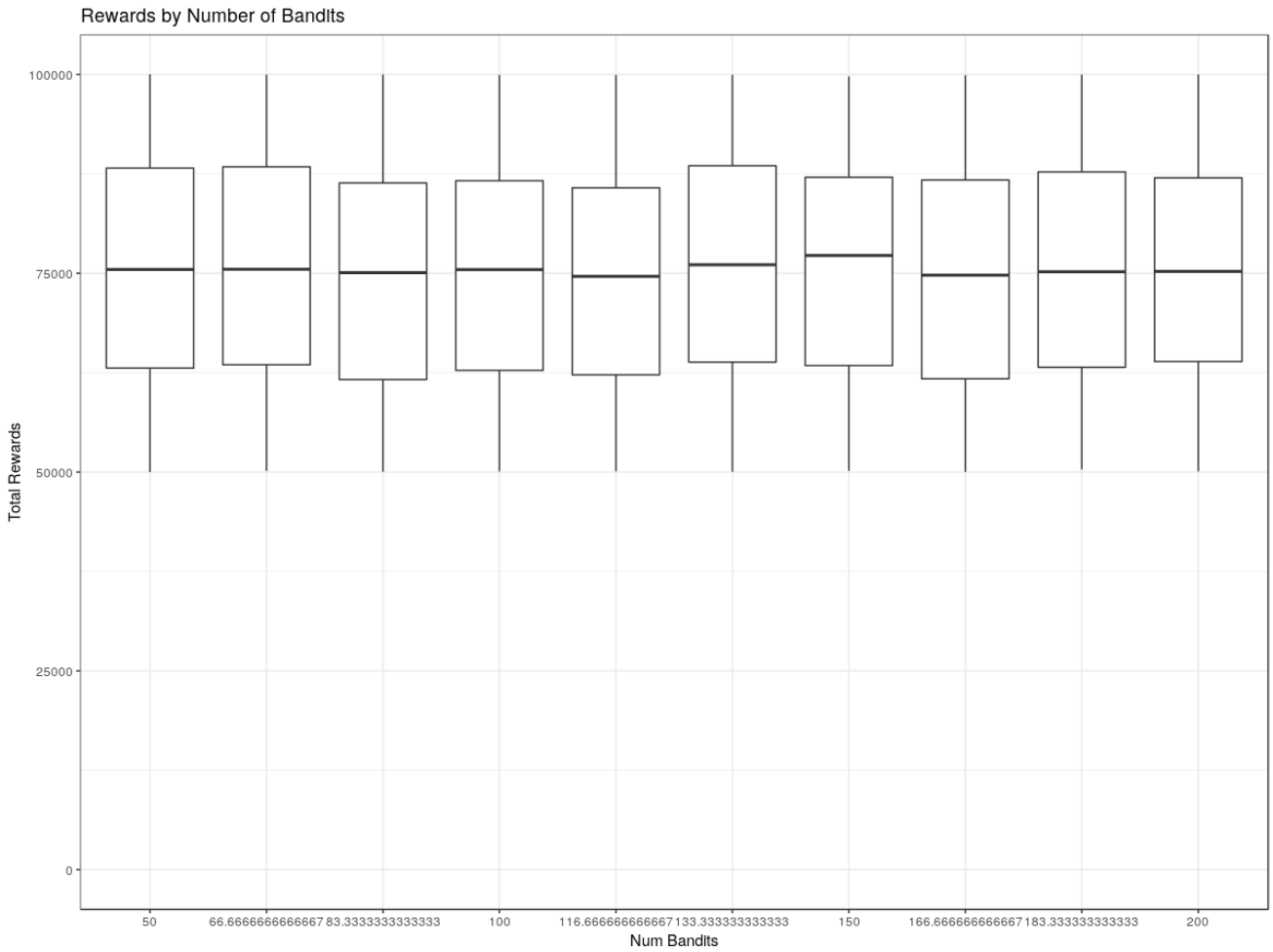
Versions upon versions of yourself play the slots with different alpha levels. Wait... This is confusing. Why do their potential earnings keep fluctuating?

You lost the guarantee of the law of large numbers. The law of large numbers states that if you calculate the mean of a large number of trials, the result of that should be close to the true value of the bandit.

The mean converges to a value while our arbitrary alpha does not.

You decide to stick with the mean for your value estimator.

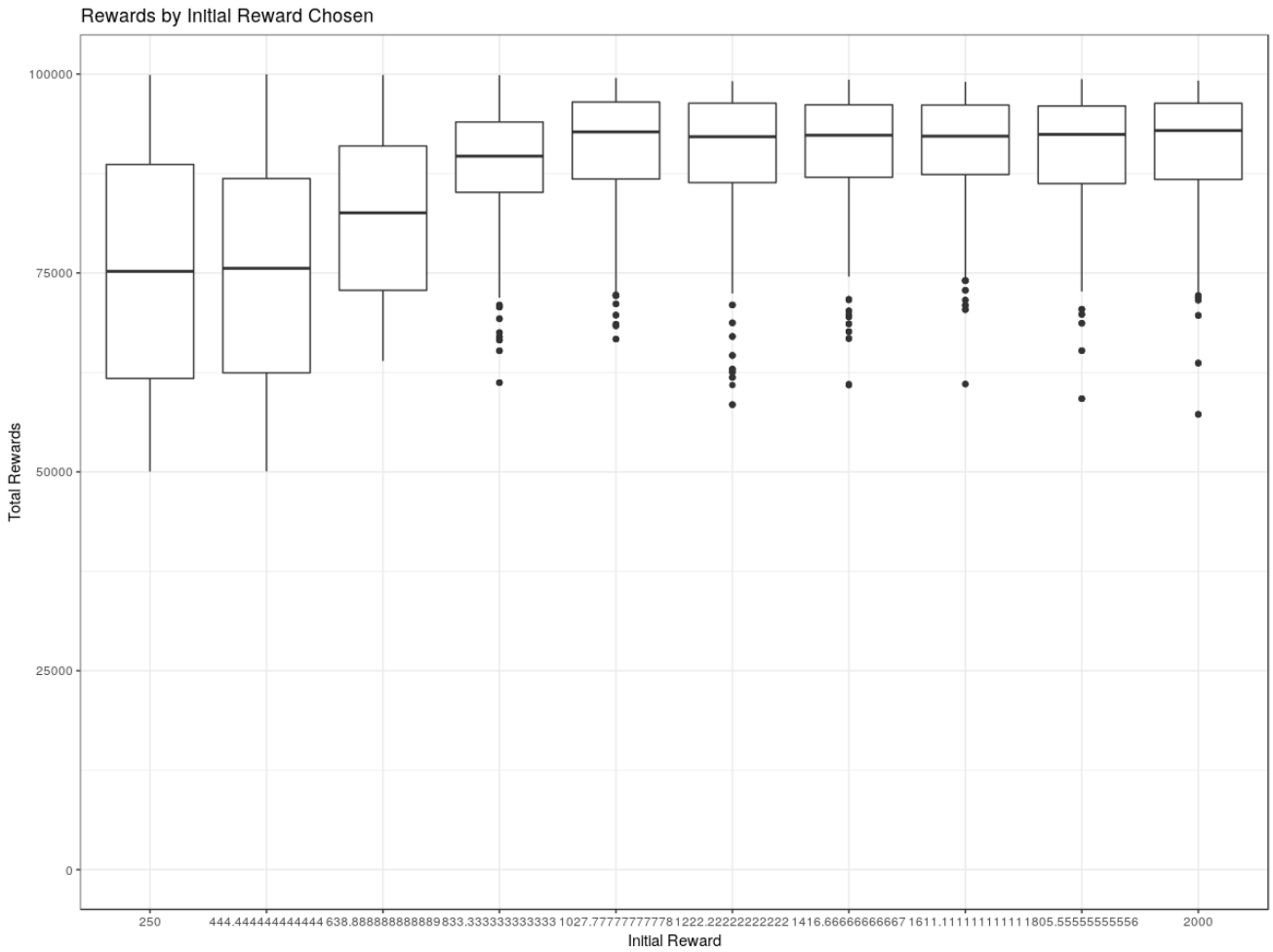
Changing the number of bandits



There are many slot machines at the casino. Rather than consider all of them, why don't you limit the number you choose at a time and see how your performance goes?

Conclusion: The number of bandits does not have a strong affect on the reward.

Changing Initial Estimates



Remember that first guess you made at the beginning of this story? Let's change that to be a bit less of a guess. After running through the machines with a fresh, clean mind, you notice that the more you initially overestimate, the better your performance tends to be. Formally, this is called an *optimistic initial value*.

Why does this work? Perhaps your thoughts are something like this...

"I can't wait to try this machine, it's going to give me sooo much money!"

"Drats, that was horrible! Lemme try this machine, I bet this one will give me lots of money."

"Ah! Not again. Now this other one..."

After trying out every machine:

"Okay. Now that I have chosen the best machine out of all of these sad, sad ones, I can start mining the money!"

Optimistic initial value is a great incentive for exploration in the beginning stages of your gambling session, while preserving a high rate of exploitation afterwards.

Conclusion

After a week has passed, you have finally determined a winning strategy. Compiling your notes, you come up with the following tips:

- Choose randomly 20% of the time.

- Be optimistic.
- Stick with the good ol' mean.

With these tips in mind, you walk up to the first bandit machine, ready to start making the big bucks.