STOR608 Final Report

Matthew Davison

Abstract: In Sprint 3 of STOR608 we investigated sequential decision making and decision making algorithms such as UCB and Thompson Sampling. In this report I focus on Thompson Sampling.

The Multi-Armed Bandit Problem

In a multi-armed bandit problem there is a set of K actions (or arms) and T rounds, where K and T are positive natural numbers. In each round $t \in \{1,\dots,T\}$ an action $a_t \in \{1,\dots,K\}$ is selected. Choosing action k in round t gives a stochastic reward $X_{k,t}$ [1].

We will assume that rewards are independent across actions, and that for action k,

$$X_{k,t} \sim \nu_k$$

are i.i.d. for $t \in \{1,\dots,T\}$, where ν is an unknown probability distribution. The aim is to identify a rule for selecting actions that maximises the expected cumulative reward over T rounds,

$$\max \sum_{t=1}^{T} \mathbb{E}(X_{a_t,t}).$$

At any particular round $t \in \{1,\dots,T\}$ we cannot look ahead at what reward we will get therefore choosing an action should depend on the history of actions we have chosen, and rewards that they gave us, in other rounds. This rule of choosing an action based on the history is called a policy [1].

Formally a policy, π, maps a history, $H_{t-1} = (a_1, X_{a_1,1}, \dots, a_1, X_{a_{t-1},t-1})$, to actions. That is,

$$\pi : \sigma(a_1, X_{a_1,1}, \dots, a_1, X_{a_{t-1},t-1}) \to \{1,\dots,K\}.$$

The optimal policy would be to play the arm with the largest expected value but since we don't know the distributions $\nu_1,\dots,\nu_K$ we cannot achieve this.

Regret

To measure the gap between the (unattainable) optimal policy and other policies we use a measure called regret.

Define $\mu_k = \mathbb{E}(X_{k,t})$ for $k \in \{1,\dots,K\}$ and $\mu^* = \max_{k\in\{1,\dots,K\}} \mu_k$. Then the regret of policy π in T is given by

$$Reg_\pi(T) = T\mu^* - \sum_{t-1}^{T} \mathbb{E}_\pi(\mu_{a_t}).$$

Minimising regret is equivalent to maximising reward [1]. Notice that we often cannot calculate regret in a real scenario. This is because again we don't know the distributions $\nu_1,\dots,\nu_K$. We can however use regret to measure how effective an algorithm is at finding a policy.

Thompson Sampling

The policy finding algorithm we will consider is called Thompson Sampling (also known as Local Thompson Sampling [3]). At each time $t \in \{1,\dots,T\}$ we need a mechanism that can, for each action $k \in \{1,\dots,K\}$, be used to sample from the posterior distribution $p_k(\mu_k|X_{k,1:t-1})$ of each action $k \in \{1,\dots,K\}$. We denote a random variable drawn from this distribution as $\tilde{\mu}_{k,t}$.

Algorithm 1: Thompson Sampling [4]

Input: Posterior distributions $\{p_k(\mu_k|X_{k,1:t-1}) : k \in \{1,\dots,K\}\}$
for *k=1* **to** K **do**
  Sample $\tilde{\mu}_{k,t} \sim p_k(\mu_k|X_{k,1:t-1})$
end
Sample a_t uniformly from $\text{argmax}_{k\in\{1,\dots,K\}}\tilde{\mu}_{k,t}$

Deriving Posterior for Thompson Sampling

Suppose we have a 2-armed Bernoulli bandit with mean parameters 0.5 and 0.55. That is to say that $\mathbb{P}(X_{2,t}=1)=0.55$, $\mathbb{P}(X_{2,t}=0)=0.45$ and $\mathbb{P}(X_{1,t}=1)=\mathbb{P}(X_{1,t}=0)=0.5$ for all $t \in \{1,\dots,T\}$. To create an appropriate posterior to sample from in the Thompson Sampling we require a likelihood function and a prior.

Since we know that $X_{k,t} \sim \text{Bernoulli}(\mu_k)$ where $k \in \{1,2\}$ then the likelihood of observing outcomes $X_{k,t}$ for t such that $a_t = k$ is given by

$$\prod_{t|a_t=k} \mu_k^{X_{k,t}}(1-\mu_k)^{1-X_{k,t}}.$$

We now need to select an appropriate prior. As the arms have a Bernoulli distribution then it is appropriate to select a $\text{Beta}(a_k, b_k)$ prior for arm k. This is because the Beta distribution is conjugate to the Binomial distribution [2] and the Bernoulli distribution is a special case of a Binomial distribution.

Using standard results from Bayesian statistics [2] we have that

$$p_1(\mu_1|X_{1,1:t-1}) \sim \text{Beta}\left(a_1 + \sum_{i=1}^{t-1} X_{1,i}\mathbb{1}(a_i=1), b_1 + t - 1 - \sum_{i=1}^{t-1} X_{1,i}\mathbb{1}(a_i=1)\right),$$
$$p_2(\mu_2|X_{2,1:t-1}) \sim \text{Beta}\left(a_2 + \sum_{i=1}^{t-1} X_{2,i}\mathbb{1}(a_i=1), b_2 + t - 1 - \sum_{i=2}^{t-1} X_{2,i}\mathbb{1}(a_i=2)\right).$$

Prior Hyperparameters

In the above section we have selected priors that are Beta distributions. Since we have two arms in this example, we have four hyperparameters to specify. These should reflect our initial beliefs about the arms reward distribution.

	a_1	b_1	a_2	b_2
Case 1	30	30	36.6	30
Case 2	10.8	2.7	2.7	10.8
Case 3	1.5	1.5	1.8	1.5
Case 4	4.8	1.2	1.2	4.8
Case 5	1	1	1	1

Table 1: Table of the prior parameters for each case.

We will investigate the effect of the prior parameters on the performance of Thompson Sampling. Five cases are given in Table 1.

Case 1 represents a situation where we have assumed the correct mean for each arm and we are confident in this assumption (low variance). Case 2 represents a situation in which we have assumed the wrong mean for each arm and we are confident in this assumption.

Case 3 and Case 4 represent correct and wrong means respectively as in Case 1 and Case 2, except we are not confident in our assumption (higher variance).

Case 5 is a flat prior, representing the situation where we don't want to assume the mean of either arm. Case 1 to Case 4 are illustrated on Figure 5.

Thompson Sampling Experiment

In this section Thompson Sampling is used the priors given earlier and we measure the total regret. We run the algorithm for 1000 rounds and then 10000 rounds. Once the algorithm determines which arm gives the higher mean reward the total regret should not change.

1000 Rounds

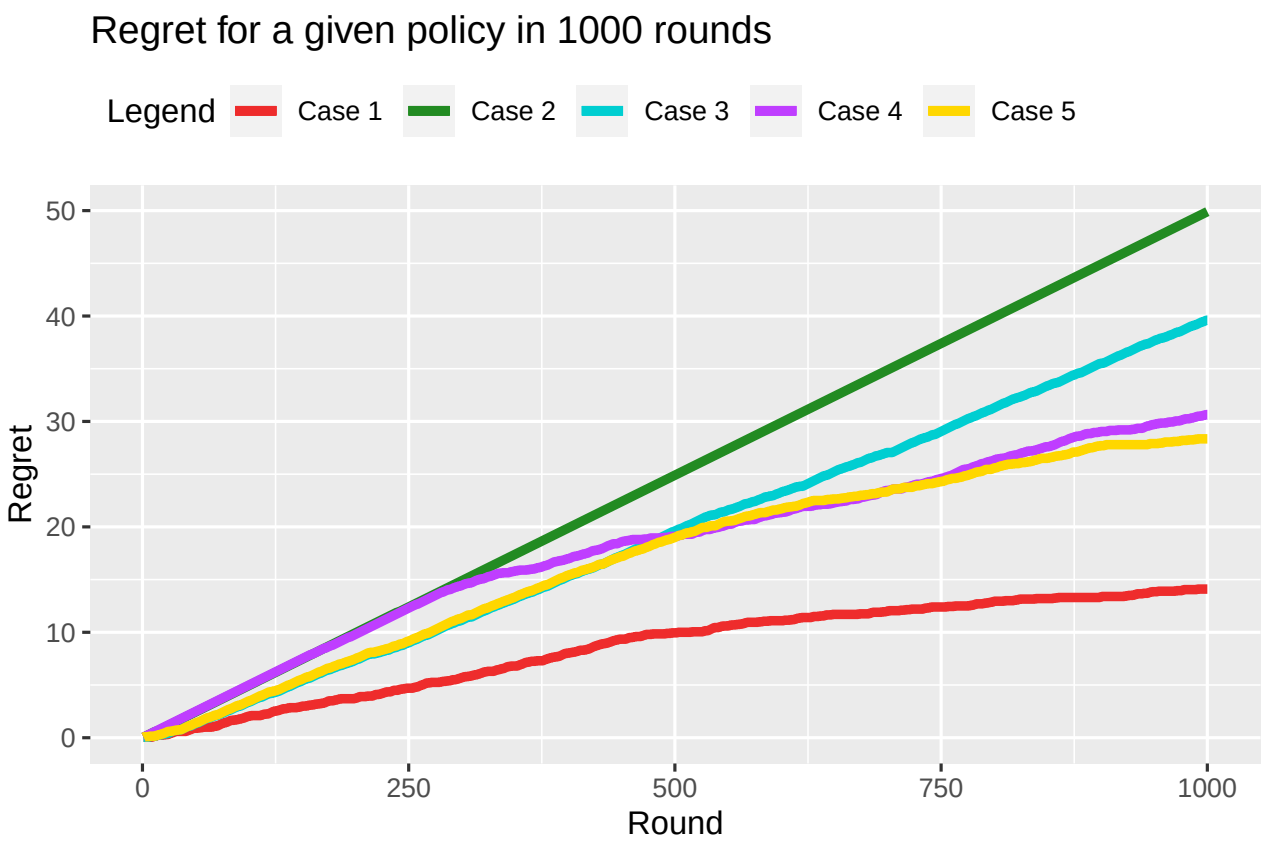

Figure 1: Regret of each case over 1000 rounds.

Figure 1 shows that Case 1 has the least regret and that Case 2 has the highest regret. Case 4 and 5 are nearly identical in regret and Case 3 is between Case 2 and 4. Notice that the high variance cases are close together in total regret, and the low variance cases are far apart. This suggests that the confidence in our assumptions effects our policy.

It is surprising that Case 4 has lower regret than Case 3, given that Case 3 has the correct mean rewards. After running the Thompson Sampling algorithm for Case 3 and Case 4 three hundred

times we can see from Figure 2 that on average Case 3 gives us a better policy. Therefore, what we assume about the mean reward for each arm does have an effect on total regret.

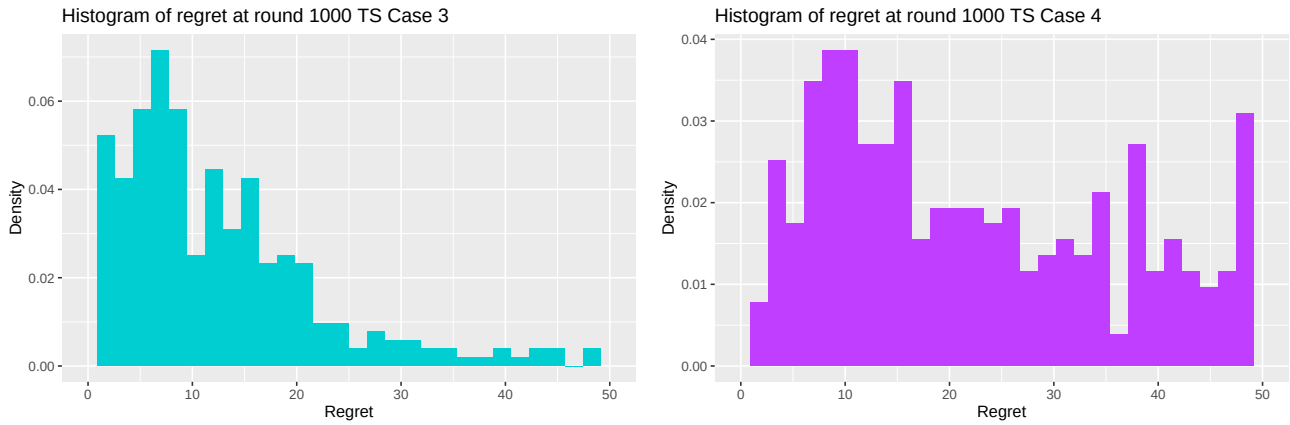

Figure 2: Histogram of total regret at round 1000 in 300 replications.

10000 Rounds

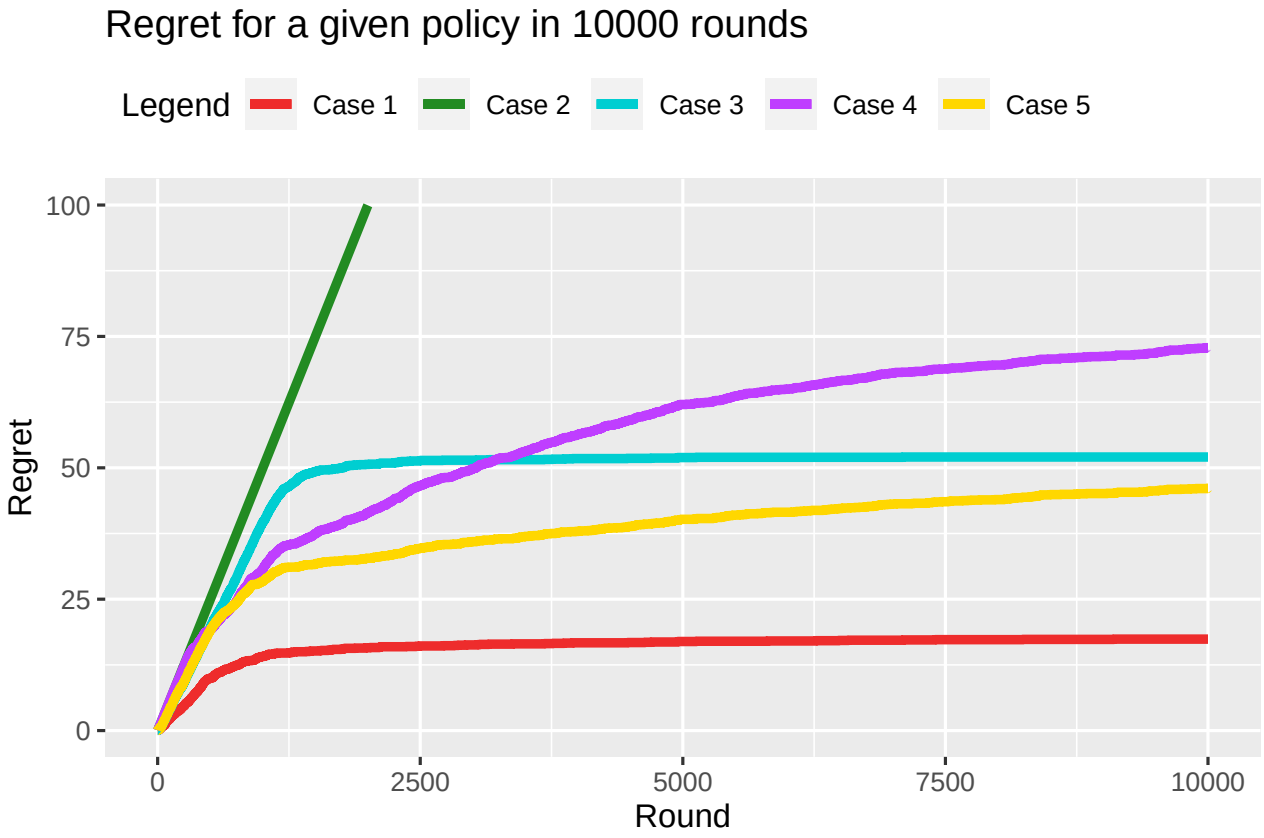

Figure 3: Regret of each case over 10000 rounds.

Figure 3 shows that once again Case 1 and Case 2 have the lowest and highest regret respectively. We notice that Case 2 only pulls Arm 1 for all 10000 rounds, therefore had to be cut from the graph so we can read the other results.

Case 1 and Case 3 have both plateaued, meaning they have both found an optimal policy (that is, only choose Arm 2) Case 4 and Case 5 are still increasing in regret after 10000 rounds. We can

see that Case 4 has accumulated more regret than Case 3 now that we have given the algorithm more time to run.

What is interesting is that the flat prior (Case 5) has performed better than Case 3, where we have assumed the correct arm means albeit with little confidence. Figure 4 shows that if we run the algorithm for 25000 rounds, Case 5 continues to accumulate more regret and eventually overtakes Case 3 at around 15000 rounds.

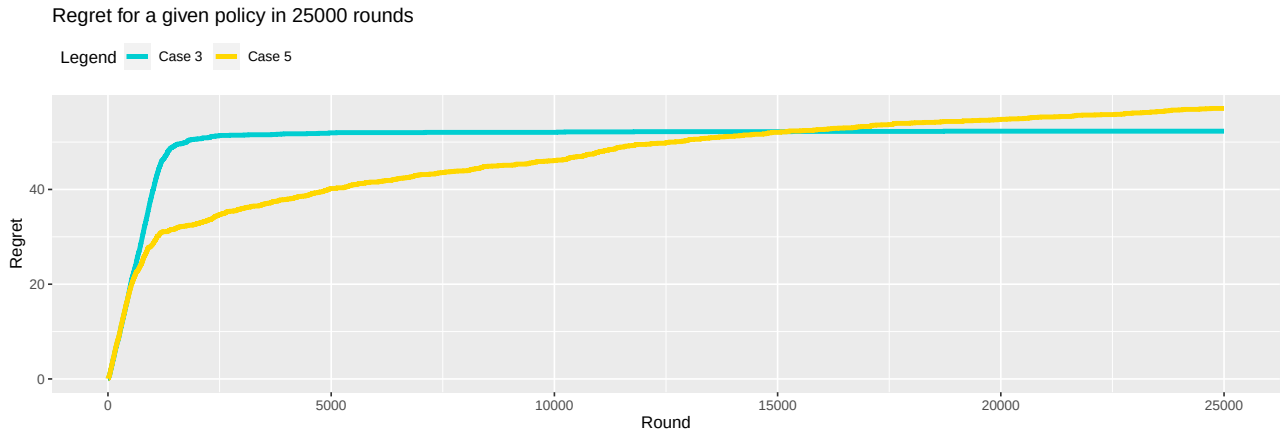

Figure 4: Running Case 3 and Case 5 for 25000 rounds.

Conclusion

When using the Thompson Sampling a prior needs to be chosen and the choice of prior is important. We found that the variance of the prior distributions played a large role in the decision making and that a low variance prior is only appropriate when you are confident that the mean of the prior is close to the mean of the arm.

To extend on that, we found that if you are not confident in your assumptions (right or wrong) then depending on how many rounds you are considering it may be better to choose a flat prior. This allows the algorithm to explore the arms and modify the posterior with the data.

The choice of prior only impacts earlier rounds of the algorithm as once you have a lot of data about the arms the algorithm uses this instead. Therefore, a good prior is useful in an instance where you have a limited number of rounds.

One way to avoid getting stuck on one arm as in Case 2 would be to implement Optimistic Bayesian Sampling (OBS), where the chance of playing an arm increases the more uncertain you are about the payoff of that arm [3].

References

[1] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[2] Peter M Lee. *Bayesian Statistics an Introduction*. John Wiley & Sons, 4th ed. edition, 2012.

[3] Benedict May, Nathan Korda, Anthony Lee, and David Leslie. Optimistic bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research*, 13:2069–2106, 2012.

[4] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Appendix

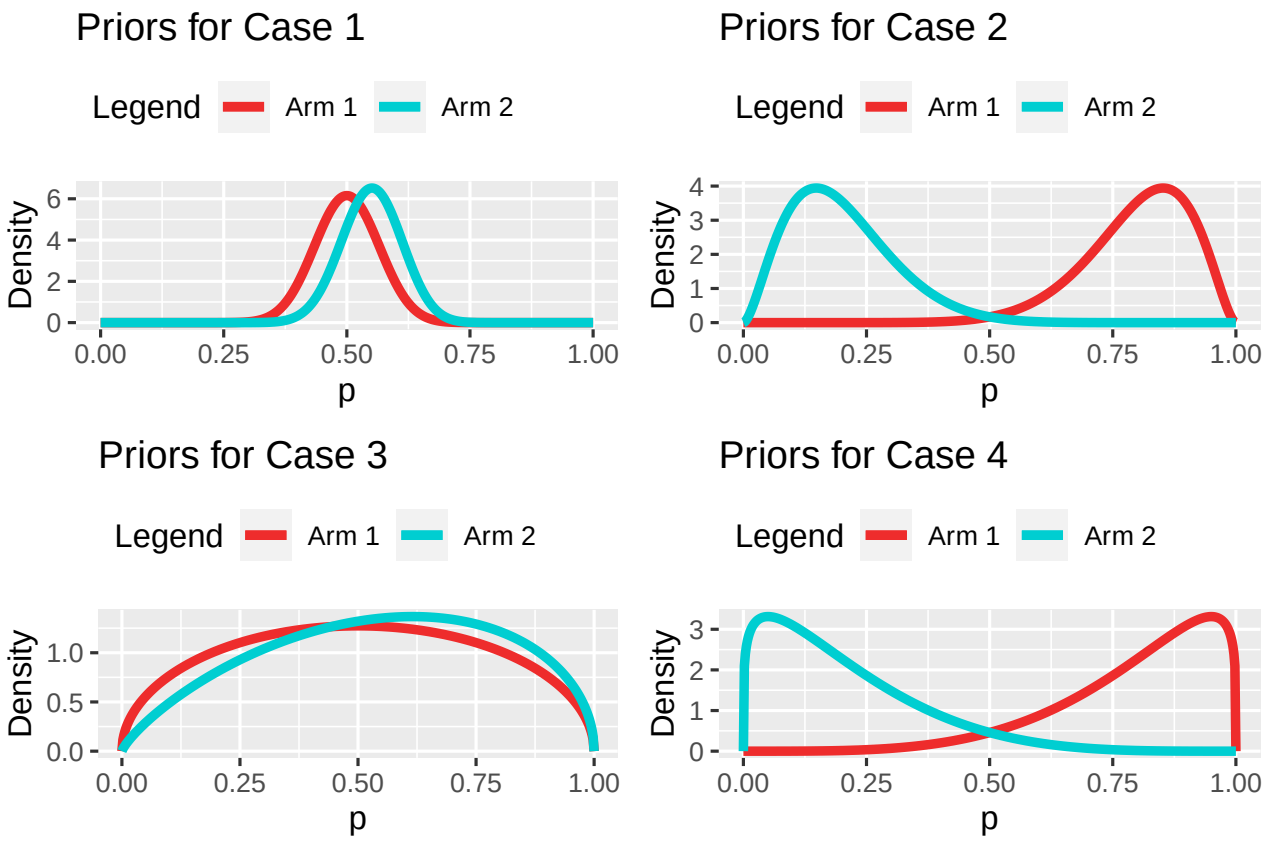

Figure 5: Priors for the different cases given in Table 1.

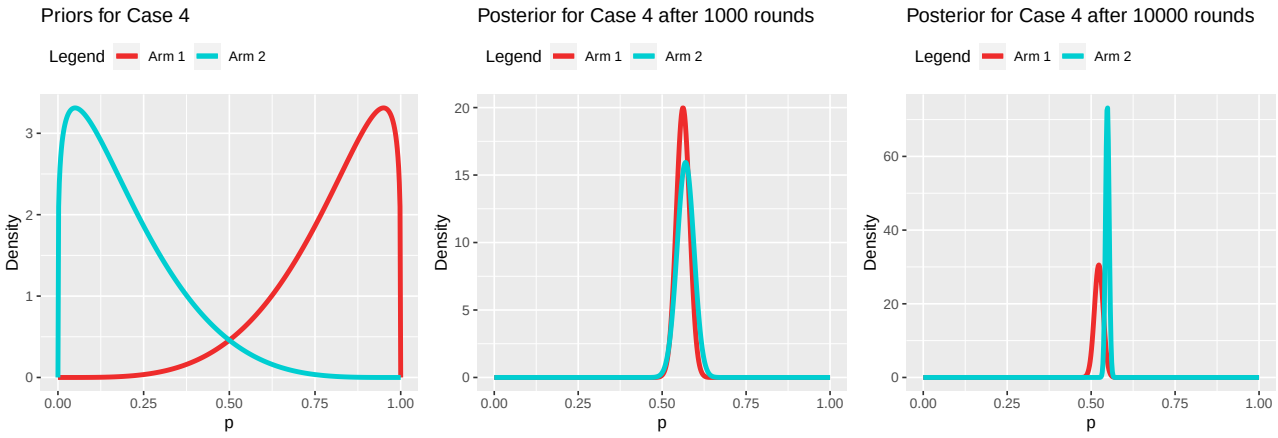

Figure 6: How the data changes the shape of the prior of Case 4 after 1000 and 10000 rounds.