Let's first re-describe the bandit setting.

There are N arms, each with some unknown distribution of reward R_i. The goal is to find which arm has the maximum expected reward via trial and error. We assume that playing a trial with arm i yields a sample $r \sim R_i$ that is stationnary.

First we will consider the case where these distributions are bernoullis, $R_i \sim \text{Bern}(\mu_i)$, $R \in \{0, 1\}$.

The Bayesian approach to estimating the μ_is suggests that we use a prior distribution on the values of μ_i. A natural prior for $[0, 1]$ values is the Beta distribution.

The Beta distribution is defined over $[0, 1]$, Beta(α, β) has the following p.d.f.:

$$f(x, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

Note that the *natural prior* is the Beta(1,1) distribution since it is equivalent to the uniform distribution:

$$f(x, 1, 1) = \frac{1!}{0!0!} x^0(1 - x)^0 = 1$$

The mean and variance of Beta(α, β) are:

$$\mathbb{E} = \frac{\alpha}{\alpha + \beta}, \quad Var = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

As such, when α and β get high, the mass of this distribution is concentrated around its mean. In other words, increasing α and β increases the "confidence" that we have on the mean.

In our case, we are dealing with bernoulli arms. If sample from an arm S_i successes and F_i failures, the empirical estimate for μ_i is $\frac{S_i}{S_i+F_i}$, while the Bayesian posterior for the mean is $\mu_i \sim \text{Beta}(1 + S_i, 1 + F_i)$. In the limit of many samples, the two estimates will be the same, but the Thompson sampling algorithm allows better regret bounds.

The algorithm is as follows:

- sample $\mu_i \sim \text{Beta}_i$ for every arm
- play arm $\text{argmax}_i \mu_i$, observe reward $r_t \in \{0, 1\}$
- update Beta_i accordingly (add 1 to α for r_t a success, 1 to β for a failure)
- repeat to satisfaction

As the number of trials will grow, the confidence in the Beta estimate of each μ_i will be higher, and they will be closer to the true means of the distributions. In fact the expected regret can be proven to be logarithmic in the number of trials, $O(\ln T)$.

The following trick allows to achieve the same for rewards with support in $[0, 1]$.

The algorithm is almost exactly the same, except for an extra sampling step.

- sample $\mu_i \sim \text{Beta}_i$ for every arm
- play arm $\text{argmax}_i \mu_i$, observe reward $\hat{r}_t \in [0, 1]$
- **sample** $r_t \sim$ **Bern**$(\hat{r}_t)$
- update Beta_i accordingly (add 1 to α for r_t a success, 1 to β for a failure)
- repeat to satisfaction

This works because we only really care about finding the arm with the maximum *expected* reward, rather than truly estimating the underlying distribution. Note that the estimated mean μ_i of the unknown $[0, 1]$ distribution is equivalent simply because:

$$P(r_t = 1) = \int_0^1 \hat{r} f(\hat{r}) d\hat{r} = \mu_i$$

Additionally, because we reuse the same framework, it can almost trivially be proven that we achieve the same logarithmic regret.

It would be possible to update the Beta distribution as Beta$(1 + \alpha(t) + \hat{r}_t, 1 + \beta(t) + (1 - \hat{r}_t))$, but because $\hat{r}$ is not an integer, the analysis of the previous $\{0, 1\}$ case does not hold and it is not known how to calculate the regret bounds of such an estimate.