Confidence Intervals for Policy Evaluation in Adaptive Experiments Report

NAME: Rani Datta

1. **Implementing the Thompson Sampling Algorithm**

I implemented the Thompson Sampling Algorithm with the special case of having two arms with normal errors. Model 1 in the code is the environment. Model 2 is the actual posterior sampling algorithm. Model 2 runs for 1000 trials and picks the best arm based on the posterior distribution it calculates. The equations used are:

μ: sample mean, $\mu^{'}$: prior mean, $\mu^{'2}$: posterior mean
σ^2: sample variance, $\sigma^{'2}$: prior variance, $\sigma^{''2}$: posterior variance

1. Posterior Mean

$$\mu'' = \frac{\sigma^2\mu' + n\sigma^{'2}\mu}{n\sigma'^2 + \sigma^2}$$

2. Posterior Variance

$$\sigma^{''2} = \frac{\sigma^2\sigma^{'2}}{n\sigma'^2 + \sigma^2}$$

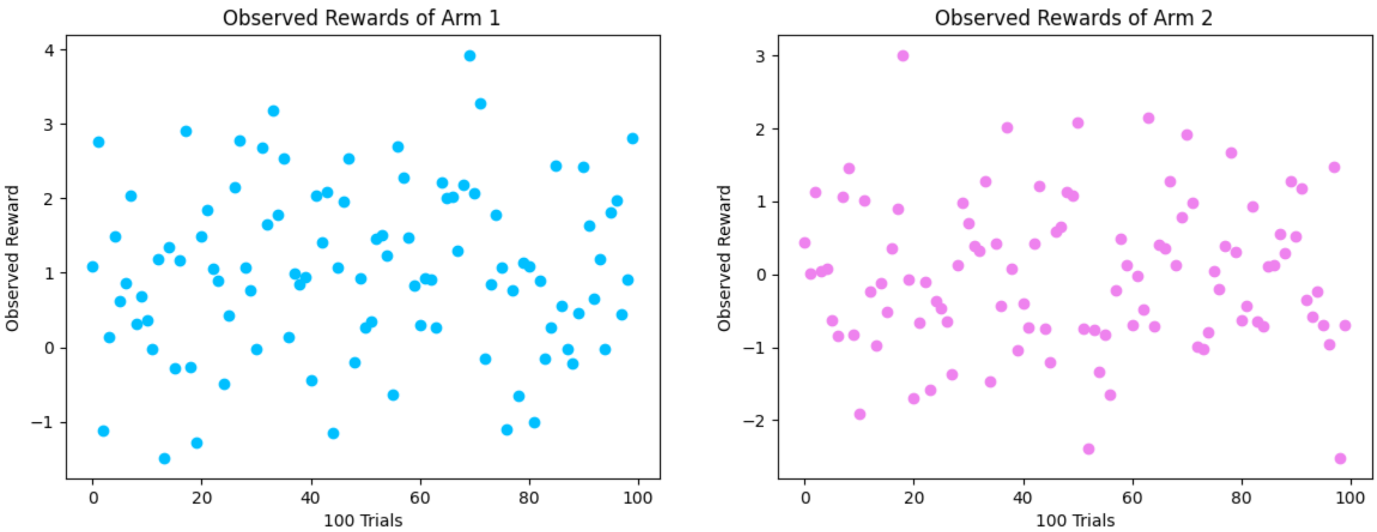
2. **Modified Thompson Sampling**

(1) **Update the posterior probability distribution to get $\hat{m}_t(w)$ and $\hat{\sigma}_t^2(w)$**

I ran the Thompson Sampling algorithm on two arms with normal distributions for 1000 trials. The posterior distribution had values $\hat{m}_t(w) = 0$ and $\hat{\sigma}_t^2(w) = 1$. After the trials were over, I found the best arm which is the arm with the highest average reward: Arm 2!

(2) **Draw $L = 100$ times for each arm from the posterior distribution**

I sampled 100 times from each of the arms' posterior probability distributions. Here are the data points for both of the arms



(3) **Compute "raw" Thompson Sampling probabilities**

This equation is for computing the "raw" Thompson Sampling probabilities using the data points observed in the previous step.

$$e_t(w) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{I}\{w = argmax\{y_1, y_2\}\}$$

Essentially, for each arm, find the amount of times it had the higher observed reward and divide it by L

(4) **Assign probability floor:** $x_t = 0.01$

This is for when e_t of an arm might be too small, so assign it to the floor and assign the other arm the complement

If $e_t(w_1) < x_t$, $e_t(w_1) = x_t$ and $e_t(w_2) = 1 - x_t$

(5) **Draw from Thompson Sampling probabilities with floor**

Now using the probabilities calculated for, we will draw from each of the arms with the probability they have. Since there are only two arms, the choosing is essentially a bernoilli flip trial.

(6) **Store the vector of probabilities, the selected arm W_t and the observed reward Y_t**

I stored all of the observed data points in the form $(arm, reward)$. The tuple being recorded is the arm that was selected with a probability of $e_t(w_1)$

3. **Estimating the True Arm Values**

We will be estimating the true arm values using the following equations. These equations can be found on page 5.

3. IPW estimator

$$\hat{\Gamma}_t^{IPW}(w) := \frac{\mathbb{I}\{W_t = w\}}{e_t(w)} Y_t$$

4. Arm value estimator

$$\hat{Q}_T^{IPW}(w) := \frac{1}{T} \sum_{t=1}^{T} \hat{\Gamma}_t^{IPW}(w)$$

4. **Constructing Confidence Intervals for the True Arm Values**

We will now calculate the confidence intervals for the true arm values of each arm. Since we are constructing the 95% confidence interval, we will use a $z_{\alpha/2}$ value of 1.96. These equations can be found on page 12.

5. Variance Estimator

$$\hat{V}^{AVG}(w) := T_w^{-2} \sum_{t:W_t=w}^{T} (Y_T - \hat{Q}_T^{AVG}(w))^2$$

6. Confidence Interval

$$\hat{Q}_T \pm z_{\alpha/2} \hat{V}_T^{1/2}$$

5. Results

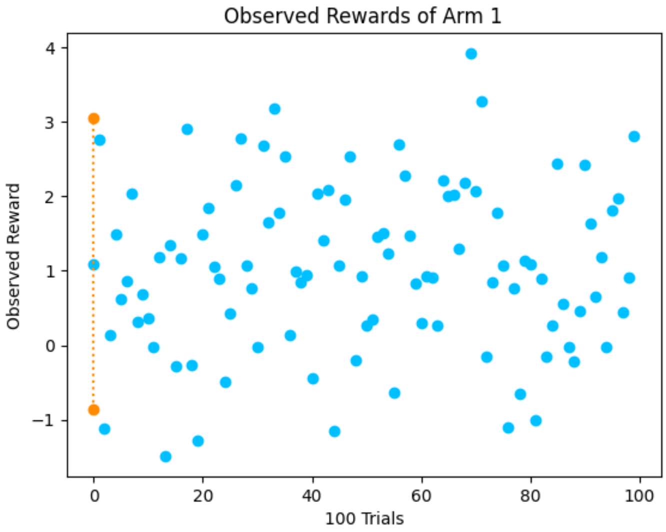
The true value of arm 1 is estimated to be 1.1 with the confidence interval of

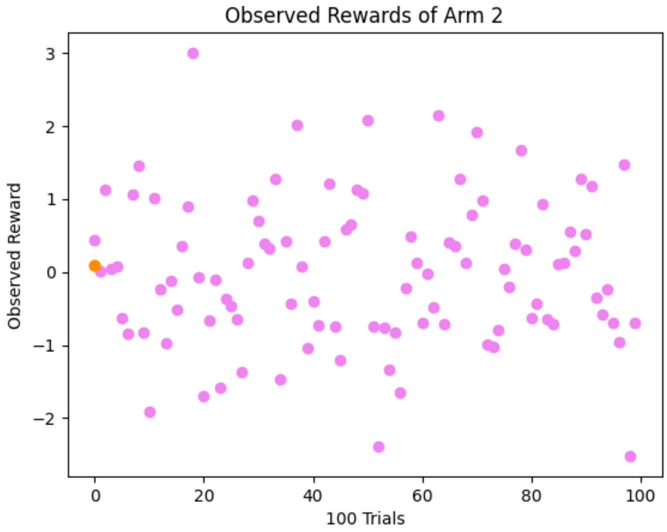
$$[-0.8556903715999997, 3.0556903716]$$

The true value of arm 2 is estimated to be 0.1 with the confidence interval of

$$[0.09999998040000001, 0.1000000196]$$

Here are the confidence intervals graphed.





It seems that arm 2 has MUCH less variance than arm 1, hence the lower arm value and the much smaller confidence interval. It seems that arm 1 has too much noise.

6. Conclusion

This is just one of the many ways of estimating means through adaptive reweighting. As a result, we can efficiently learn from our reinforcement learning algorithms and begin to apply them in real world settings! Hope you enjoyed <3