Preliminary Experiment

March 16, 2018

Simulation Details

$N = 200$ simulations were run.
The Bandit priors that were considered:

- Needle-in-haystack
    1. High (with three arms) - 2 arms with mean 0.50, 1 arm with mean 0.7 (+ 0.2)
    2. High (with twenty arms) - 19 arms with mean 0.50, 1 arm with mean 0.70 (+ 0.20)
    3. Medium-High (with twenty arms) - 18 arms with mean 0.50, 1 arm with 0.60, 1 arm 0.80

Algorithms considered:

1. ThompsonSampling with priors of $Beta(1, 1)$ for every arm.
2. DynamicGreedy with priors of $Beta(1, 1)$ for every arm
3. Bayesian Dynamic ϵ-greedy with priors of $Beta(1, 1)$ for every arm and $\epsilon = 0.05$
4. Non-Bayesian ϵ-greedy - the greedy decision was made based on empirical mean. When there were zero observations, assumed that the empirical mean was 0 (this seems questionable), $\epsilon = 0.05$
5. UCB1 (with constant 1)

Simulation Procedure

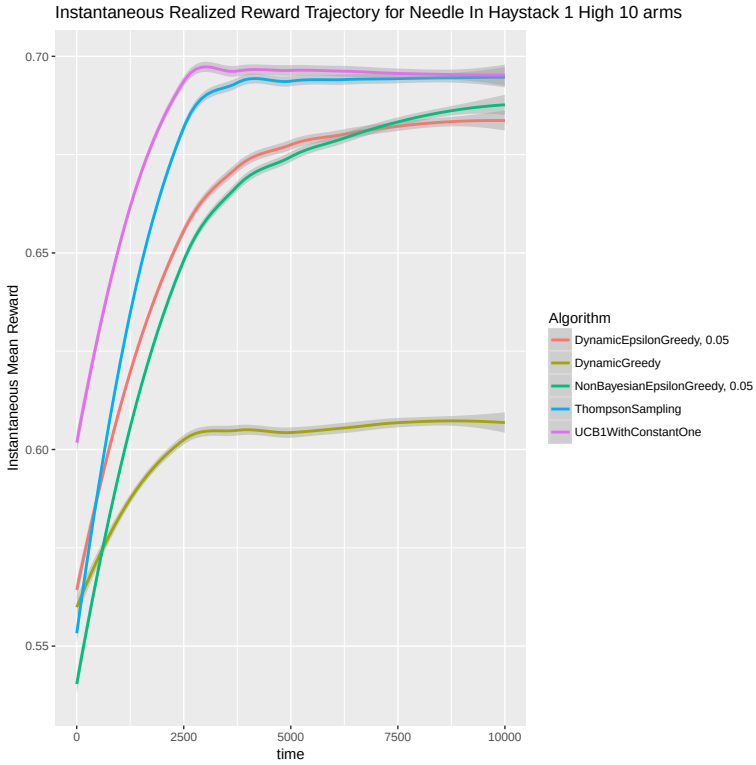
for Each prior p **do**
    for Each experiment i **do**
        Use p as the true distribution of the arms
        Generate realizations for each arm and round t and instantiate bandit instance
        for Each algorithm alg **do**
            Run simulation for T periods
        end for
    end for
end for

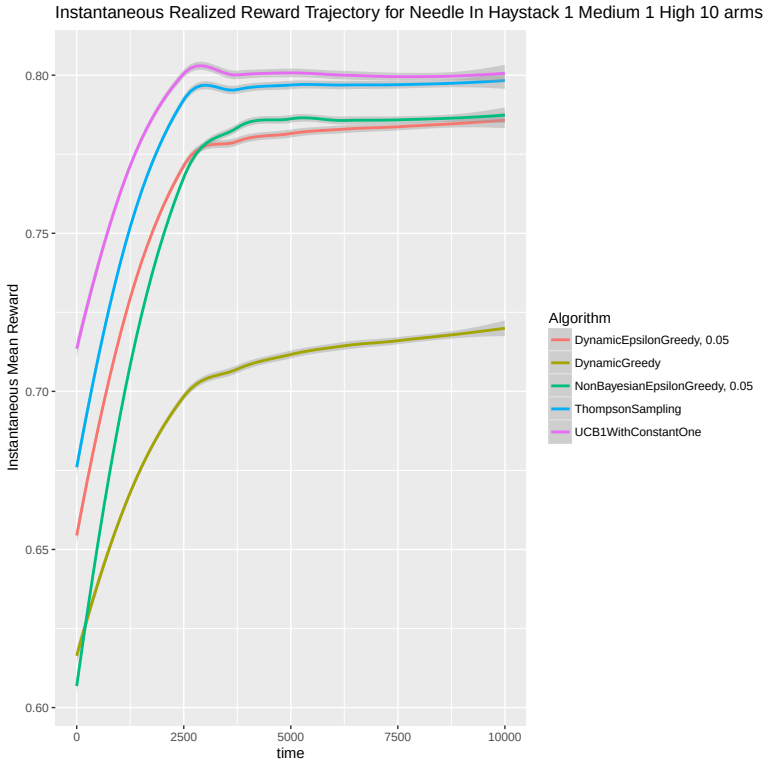
Results

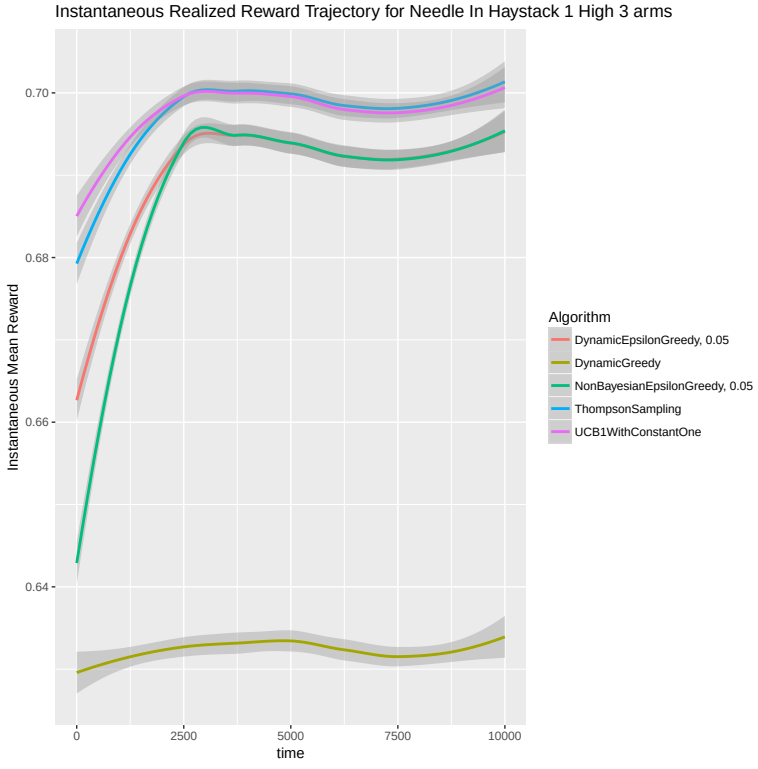
The confidence bands come from the standard error of the loess regression used to generate the curves. The actual datapoints, when plotted, are incredibly volatile but this volatility decreases as we increase N. Part of this comes from the fact that what is plotted is the actual realized reward and not the mean reward.

Instantaneous Realized Reward Plots

These plots contain the fitted lines for the instantaneous realized reward.

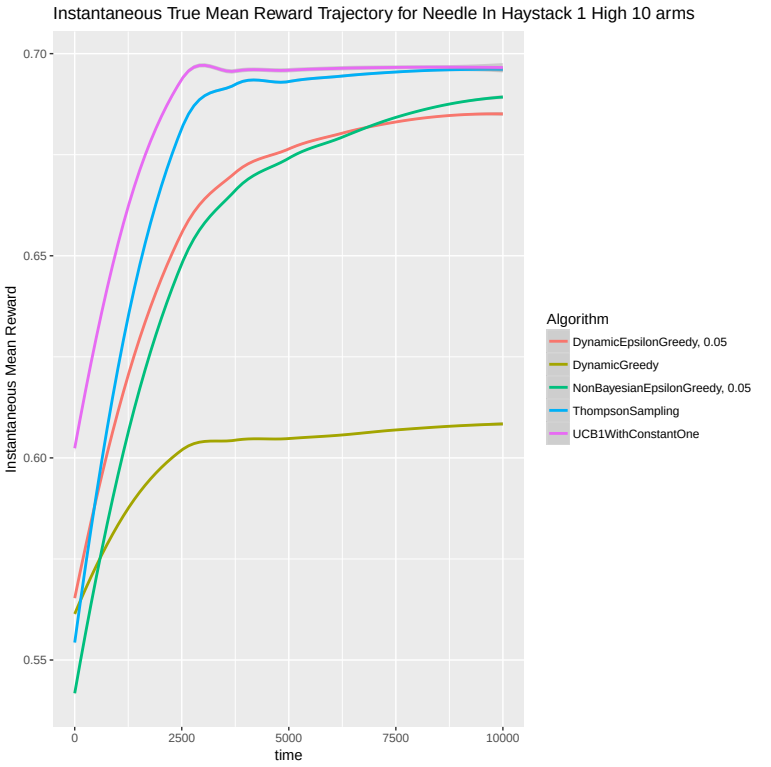


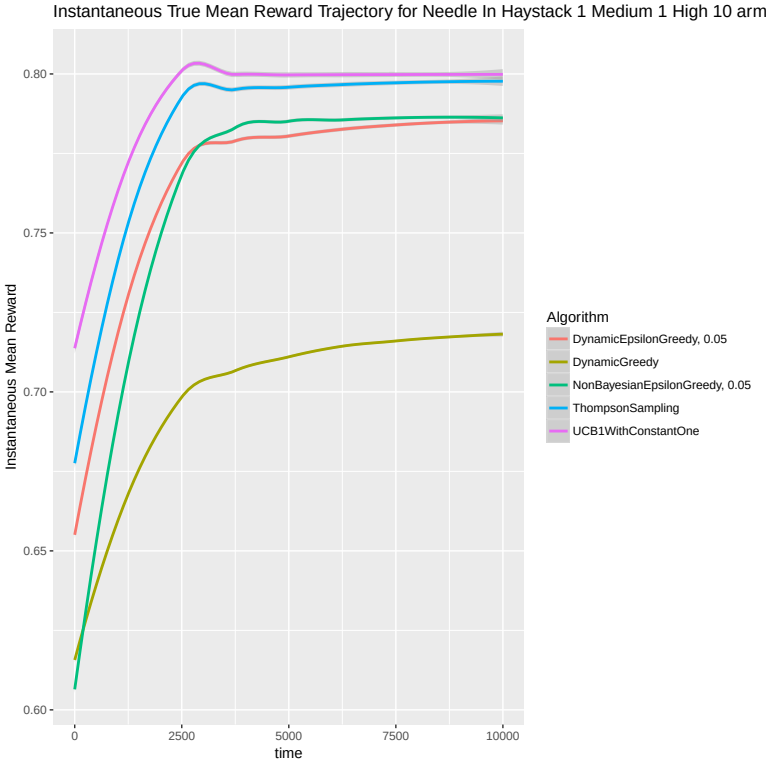

Instantaneous Realized Reward Trajectory for Needle In Haystack 1 Medium 1 High 10 arms
Instantaneous Mean Reward
0.80
0.75
0.70
0.65
0.60
0
2500
5000
7500
10000
time
Algorithm
DynamicEpsilonGreedy, 0.05
DynamicGreedy
NonBayesianEpsilonGreedy, 0.05
ThompsonSampling
UCB1WithConstantOne

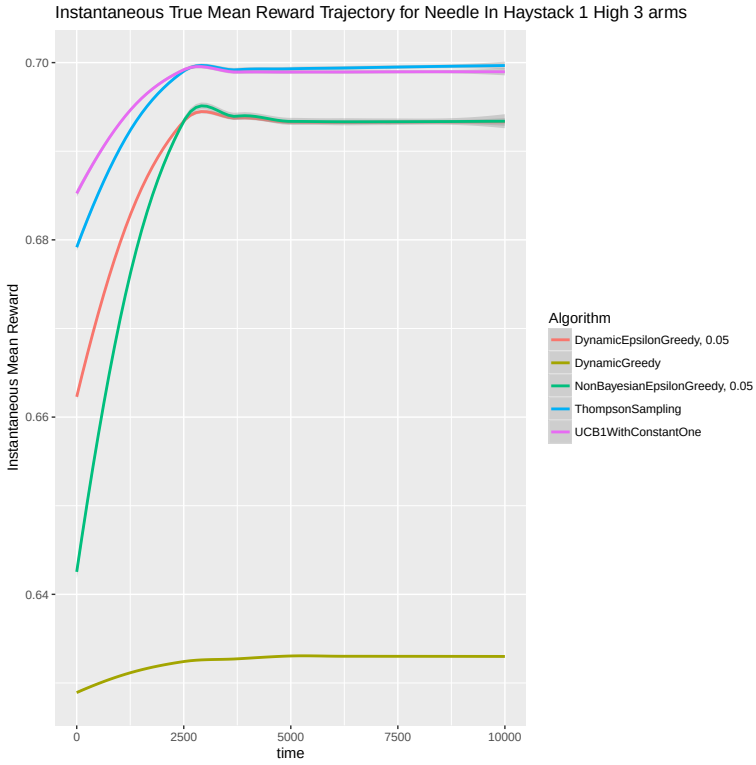

Instantaneous Realized Reward Trajectory for Needle In Haystack 1 High 3 arms
Instantaneous Mean Reward
0.70
0.68
0.66
0.64
0
2500
5000
7500
10000
time
Algorithm
DynamicEpsilonGreedy, 0.05
DynamicGreedy
NonBayesianEpsilonGreedy, 0.05
ThompsonSampling
UCB1WithConstantOne

Instantaneous True Mean Reward Plots

These plots contain the fitted lines for the TRUE mean reward for the selected arms.







Comments: It does appear that using 10 arms instead of 3 in this case does not appear too different. It appears that Thompson Sampling and UCB learn quickly (roughly seem to flatline at $t = 2000$) and DynamicGreedy slowly improves but does not seem to get to convergence even by $t = 10000$. However, the difference between the TS and DG trajectories is not vastly different between the 3 arm and 10 arm case. One difference to note is that it seems like one thing that does change between the two is that Dynamic ϵ-greedy seems to take longer to learn (doesn't even appear to converge by $t = 10000$) in the more complex cases but seems to converge as quickly as TS and UCB in the 3 arm case.