Upper Confidence Bounds Algorithm

Sungsoo Lim

December 7, 2020

1 (a)

The UCB hyperparameter α value was inverse varied from 10 to 0.001. The average regret for the training set and the bounds are shown:

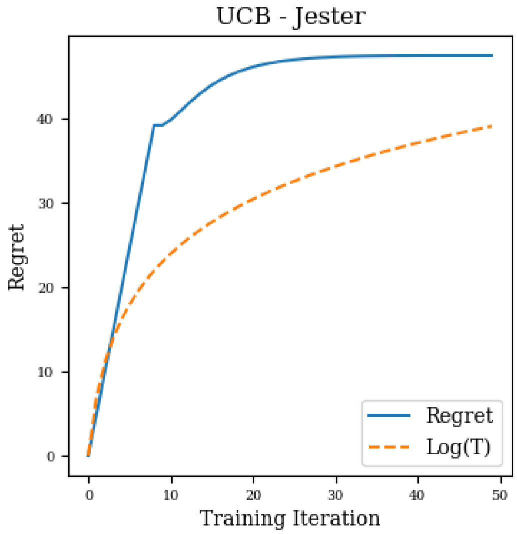


Figure 1: Averge regret for $N = 18000$ users.

The average rewards for the traininig set users are also shown:

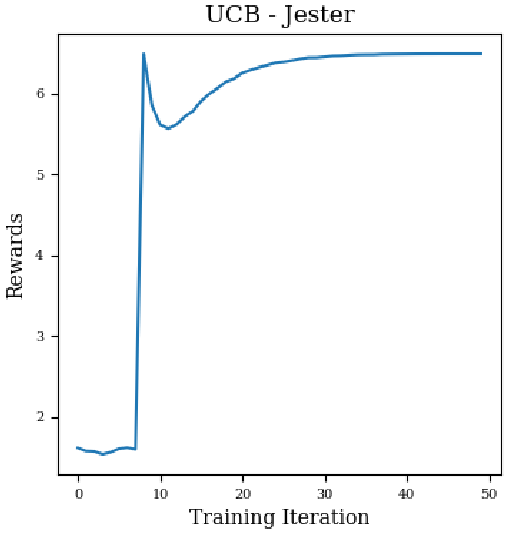
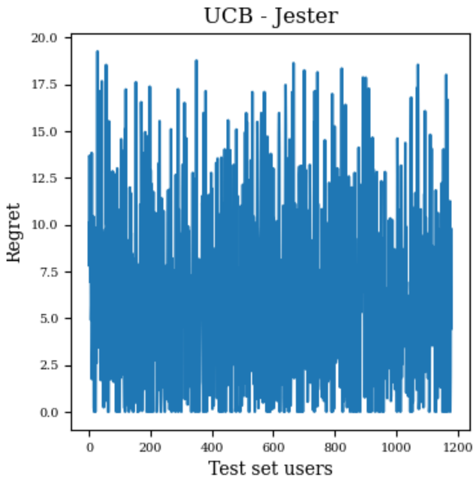


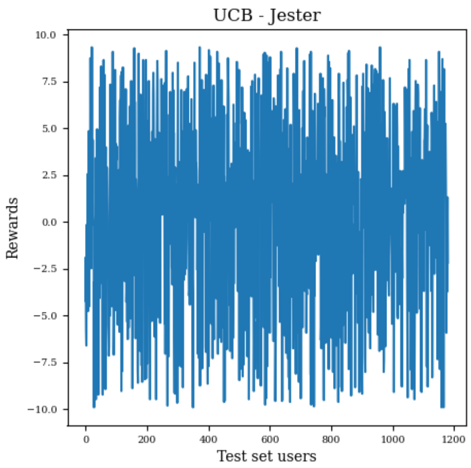
Figure 2: Averge rewards for $N = 18000$ users.

2 (b)

The trained A, b, and θ matrices were directly used to calculate the exploration-exploitation action choice for the remaining users. The average regret was 5.95, and the average reward was 0.44. The regret and rewards are shown for the test users:



(a) Regret for the test users



(b) Rewards for the test users