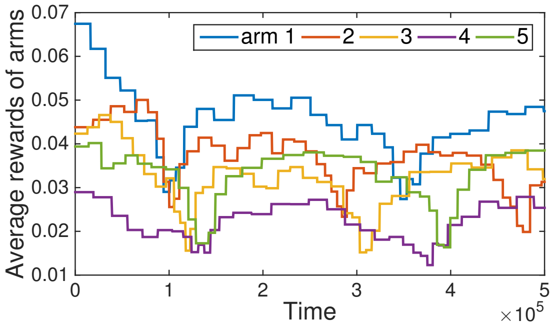
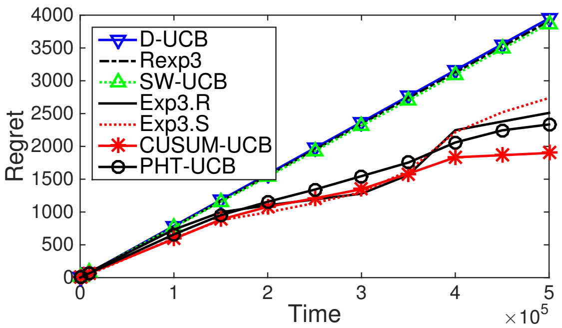


(a) Ground truth



(b) Regret

Figure 3: Rewards and regret over the Yahoo! dataset with $K = 5$

displayed on the Yahoo! Front Page (?). Given the arrival of a user, the goal is to select an article to present to the user, in order to maximize the expected click-through rate, where the reward is a binary value for user click. For the purpose of our experiment, we randomly select the set of 5 articles (i.e., $K = 5$) from a list of 100 permutations of possible articles which overlapped in time the most. To recover the ground truth of the expected click-through rates of the articles, we take the same approach as in **?** (?), where the click-through rates were estimated from the dataset by taking the mean of an article's click-through rate every 5000 time ticks (the length of a time tick is about one second), which is shown in Figure 3a.

The regret curves are shown in Figure 3b. We again fit the curves to the model $at^b + c$. The resulting exponents b of D-UCB, Rexp3, SW-UCB, Exp3.R, Exp3.S, CUSUM-UCB and PHT-UCB are 1, 1, 1, 0.81, 0.85, 0.69 and 0.79, respectively. The passively adaptive policies, D-UCB, SW-UCB and Rexp3, receive a linear regret for most of the time. CUSUM-UCB and PHT-UCB achieve much better performance and show sublinear regret, because of their active adaptation to changes. Another observation is that CUSUM-UCB outperforms PHT-UCB. The reason behind is that the Yahoo! dataset has more frequent breakpoints than the switching environment (i.e., high γ_T). Thus, the estimation $\hat{y}_k$ in PHT test may drift away before PHT detects the

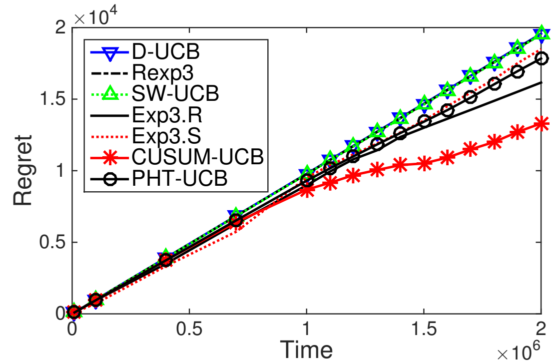


Figure 4: Regret over the Yahoo! dataset with $K = 100$

change, which in turn results in more detection misses and the higher regret.

Yahoo! Experiment 2 ($K = 100$). We repeat the above experiment with $K = 100$. The regret curves are shown in Figure 4. We again fit the curves to the model $at^b + c$. The resulting exponents b of D-UCB, Rexp3, SW-UCB, Exp3.R, Exp3.S, CUSUM-UCB and PHT-UCB are 1, 1, 1, 0.88, 0.9, 0.85 and 0.9, respectively. The passively adaptive policies, D-UCB, SW-UCB and Rexp3, receive a linear regret for most of the time. CUSUM-UCB and PHT-UCB show robust performance in this larger scale experiment.

7 Conclusion

We propose a change-detection based framework for multi-armed bandit problems in the non-stationary setting. We study a class of change-detection based policies, CD-UCB, and provide a general regret upper bound given the performance of change detection algorithms. We then develop CUSUM-UCB and PHT-UCB, that actively react to the environment by detecting breakpoints. We analytically show that the regret of CUSUM-UCB is $O(\sqrt{T\gamma_T \log \frac{T}{\gamma_T}})$, which is lower than the regret bound of existing policies for the non-stationary setting. To the best of our knowledge, this is the first regret bound for actively adaptive UCB policies. Finally, we demonstrate that CUSUM-UCB outperforms existing policies via extensive experiments over arbitrary Bernoulli rewards and the real world dataset of webpage click-through rates.

Acknowledgment

This work has been supported in part by grants from the Army Research Office W911NF-14-1-0368 W911NF-15-1-0277, and MURI W911NF-12-1-0385, DTRA grant HDTRA1-14-1-0058, and NSF grant CNS-1719371.

Ducimus velit quam esse adipisci quas voluptatibus, neque impedit nobis dolor dolorem in eius obcaecati distinctio, pariatur odio aperiam cumque qui asperiores libero nesciunt fugiat vero quam id?Voluptatum laudantium necessitatibus enim quos eaque nemo officiis sit dolorum aliquam, impedit accusantium corrupti id saepe vitae possimus reiciendis.Ducimus ad nostrum esse odio perferendis ratione iure rerum, dolore error quas atque ad neque debitis vitae iusto, at quos inventore.Eius corporis quaerat, obcaecati mollitia deleniti repellendus et odit voluptatibus