# Multi Agent Seminar : Assignment 2

Kenzo Clauw

Vrije Universiteit Brussel

kclauw@vub.ac.be.

## Introduction

We consider the 4-armed bandit problem, the optimal values are as following :

| Arm1 | Arm2 | Arm3 | Arm4 |
|---|---|---|---|
| 0.9 | 0.4 | 0.4 | 0.4 |

The parameters of each strategy are as following :

| | | |
|---|---|---|
| EpsilonGreedy | $\epsilon$ = 0.01 | |
| Softmax | $t$ = 0.1 | |
| Optimistic | Q = 100 | |
| UCB2 | $\alpha$ = 0.0001 | |
| EnGreedy | c = 0.2 | d = 0.3 |
| Thompson | $\alpha$ = 1 | $\beta$ = 1 |
| Efficient-UCBV | $p$ = 0.6 | $\psi = 156.25$ |
| EXP3 | $\alpha$ = 0.01 | |
| MOSS | $s$ = 1 | |

The results of the experiments are averaged over 20 runs.

The parameters are tuned, based on the aforementioned distribution.

### Run code

Run all the experiments as following :

Python ./bandits.py −−n 20

Parameter –n controls the amount of averages per experiment. The plots of each experiment are stored in the results folder.

### Extra strategies

- Evaluation and Analysis of the Performance of the EXP3 Algorithm in Stochastic EnvironmentsBurtini et al. (2015)
- Efficient-UCBV: An Almost Optimal Algorithm using Variance EstimatesMukherjee et al. (2017)
- Minimax policies for adversarial and stochastic banditsSeldin et al. (2013)Audibert and Bubeck (2009)

### Remarks

EXP3 is based on a Adversarial Bandit strategy, thus the regret is higher then other stochastic strategies. In the paper, they adapt EXP3 to the stochastic bandit case, but we ran out of time to test this.

MOSS is performing worse then expected.

Softmax is performing worse then Epsilon Greedy, this is because the first arm contains the highest probability.

**implement the epsilon-greedy, soft-max, and optimistic initialization action selection strategies**

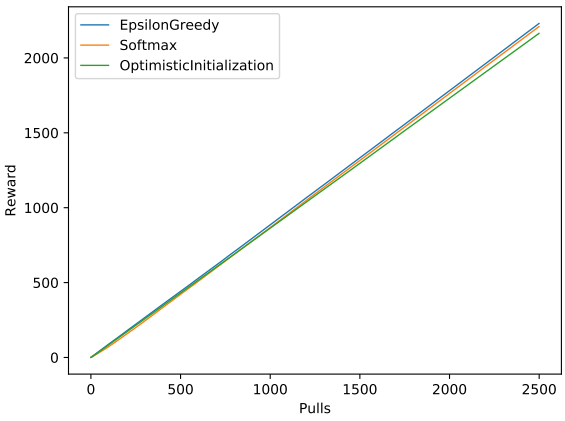


Figure 1: Average reward : distribution 1

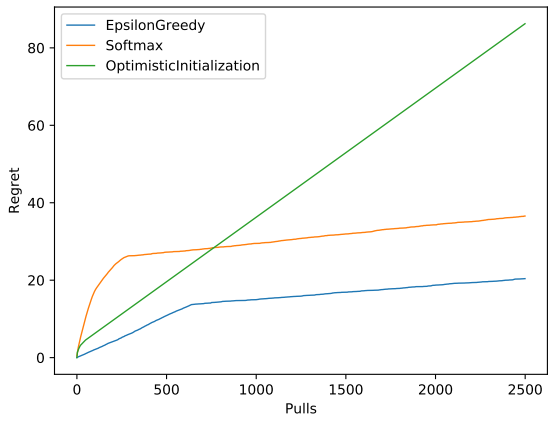


Figure 2: Average regret : distribution 1

**Implementation of UCB1, UCB2,n-greedy,Thompson Sampling, UCB Chernoff , MOSS, EXp3 and Efficient-UCBV**

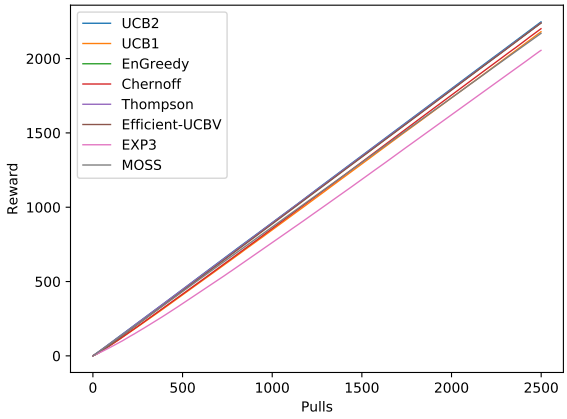


Figure 3: Average regret : distribution 1

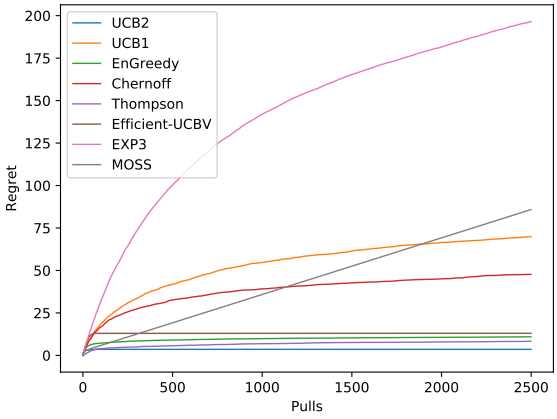


Figure 4: Average regret : distribution 1

## All experiments

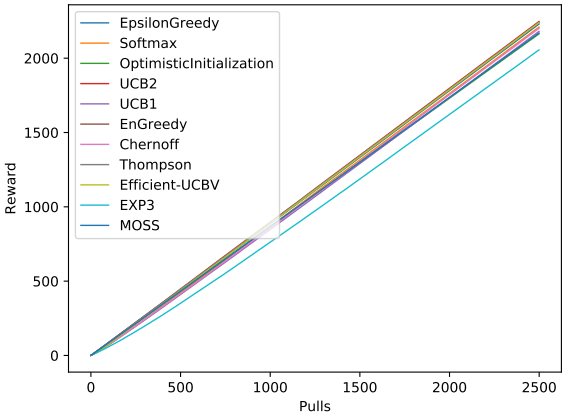


Figure 5: Average reward : distribution 1

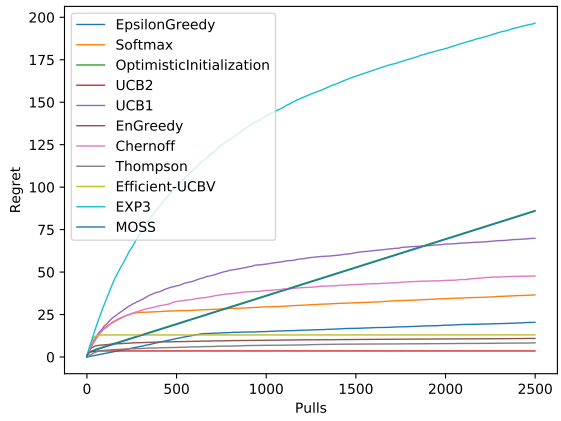


Figure 6: Average regret : distribution 1

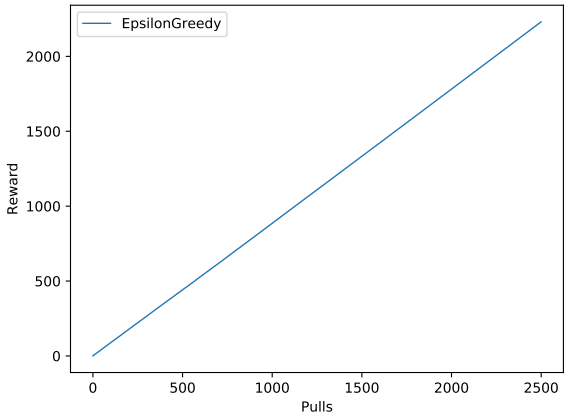


Figure 7: Average reward Epsilon Greedy: distribution 1

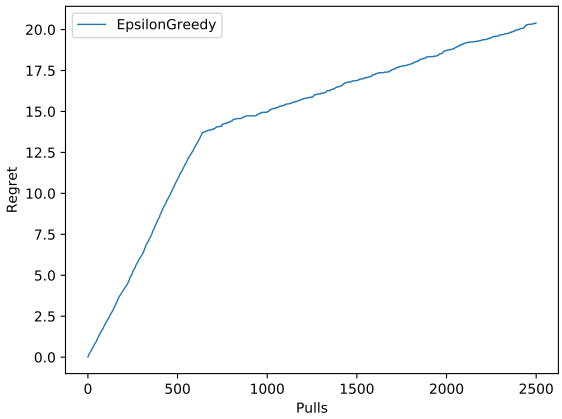


Figure 8: Average regret Epsilon Greedy: distribution 1

## Single Experiments

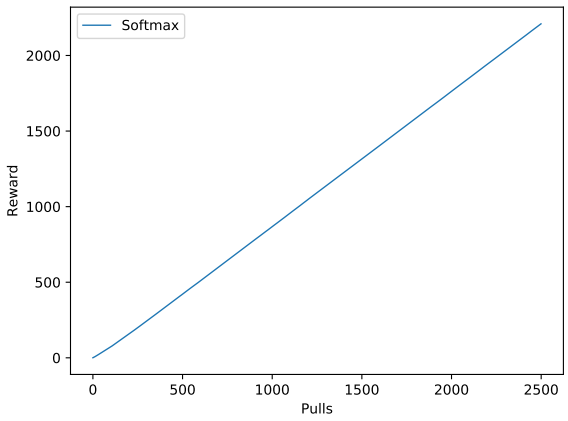


Figure 9: Average reward Softmax : distribution 1

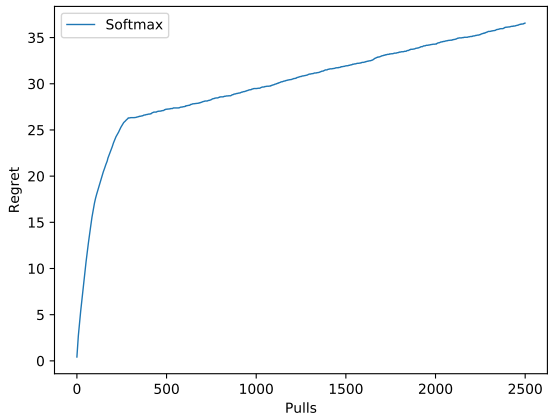


Figure 10: Average regret Softmax : distribution 1

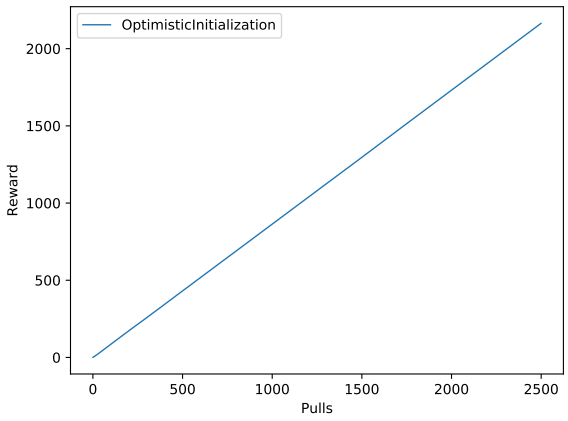


Figure 11: Average reward Optimistic Initialization : distribution 1

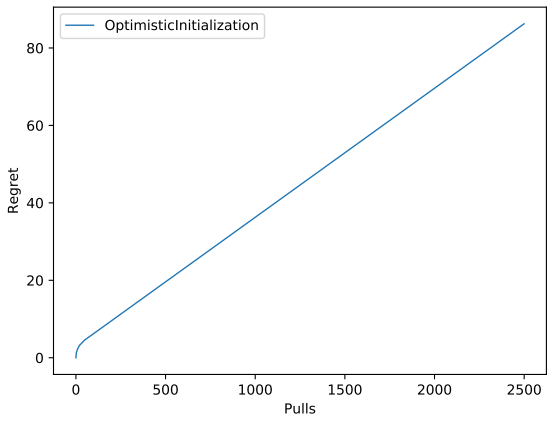


Figure 12: Average regret Optimistic Initialization : distribution 1

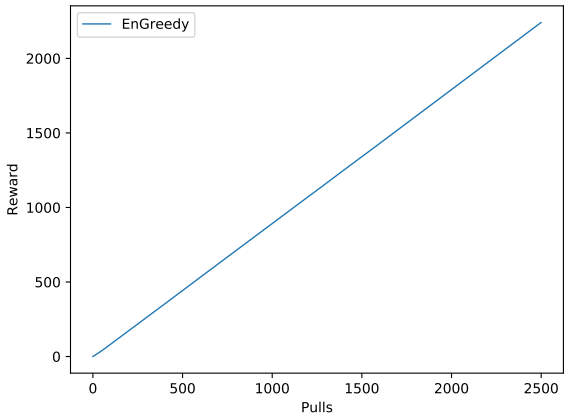


Figure 13: Average reward EnGreedy : distribution 1

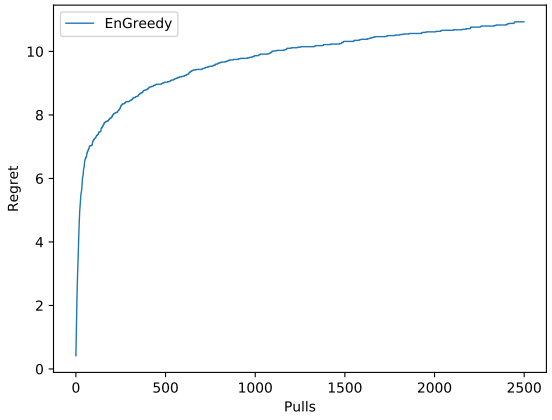


Figure 14: Average regret EnGreedy : distribution 1

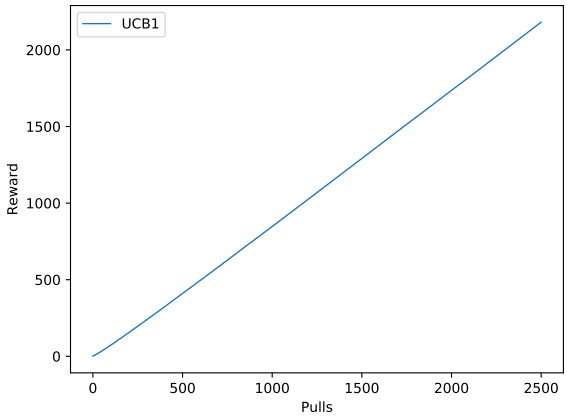


Figure 15: Average reward UCB1: distribution 1

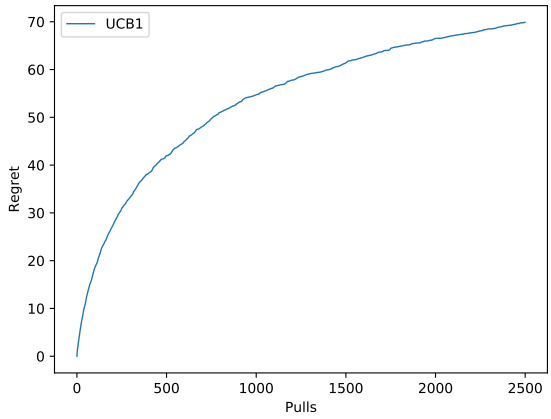


Figure 16: Average regret UCB1: distribution 1

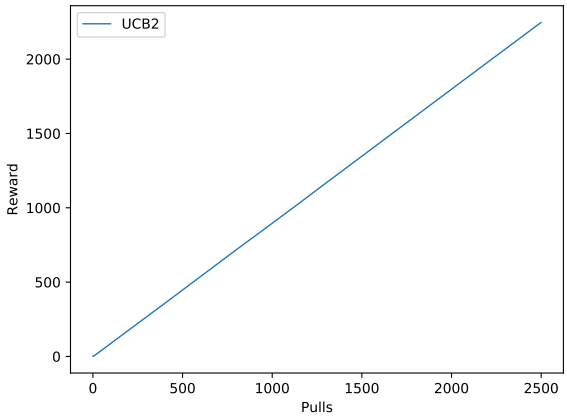


Figure 17: Average reward UCB2: distribution 1

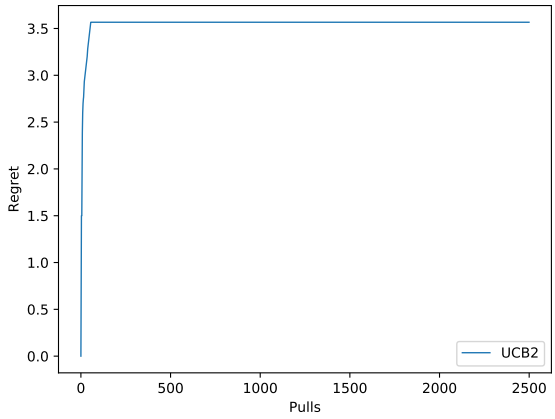


Figure 18: Average regret UCB2: distribution 1

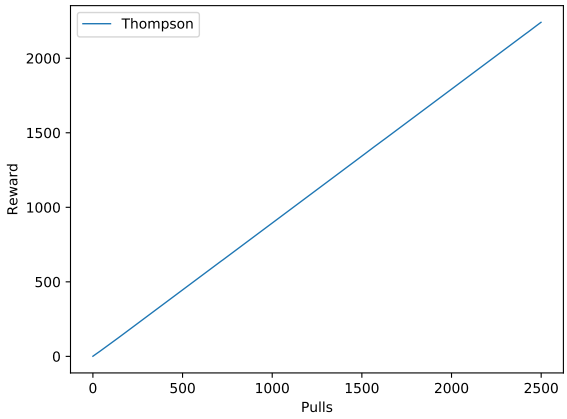


Figure 19: Average reward Thompson: distribution 1

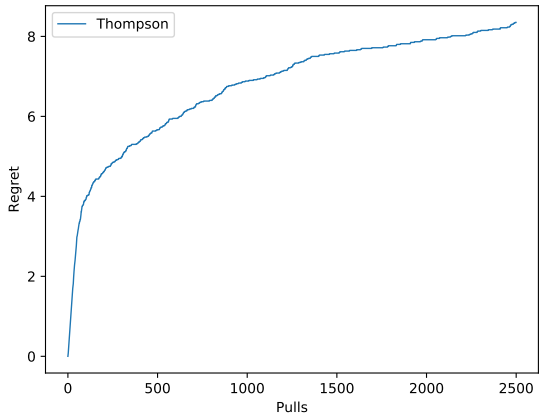


Figure 20: Average regret Thompson: distribution 1

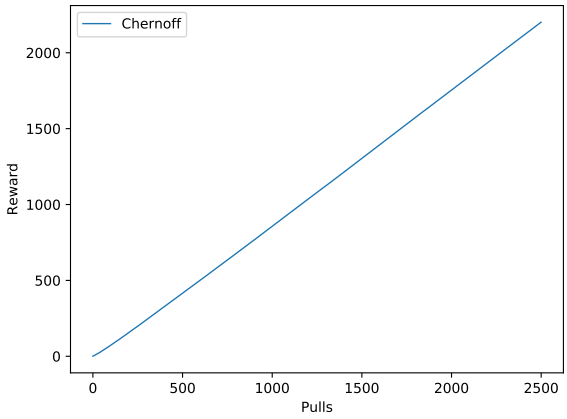


Figure 21: Average reward Chernoff: distribution 1

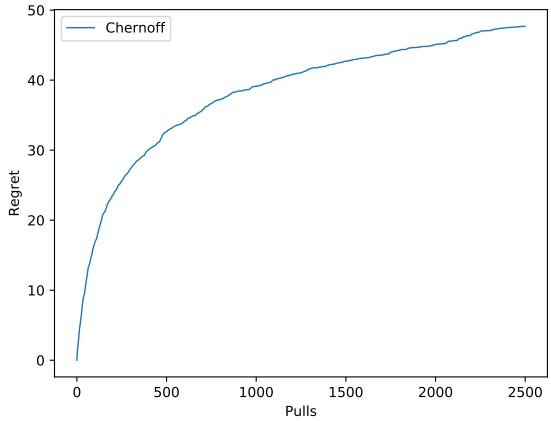


Figure 22: Average regret Chernoff: distribution 1

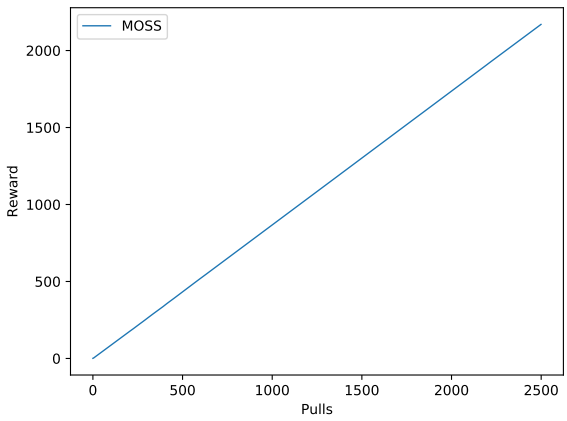


Figure 23: Average reward Moss: distribution 1

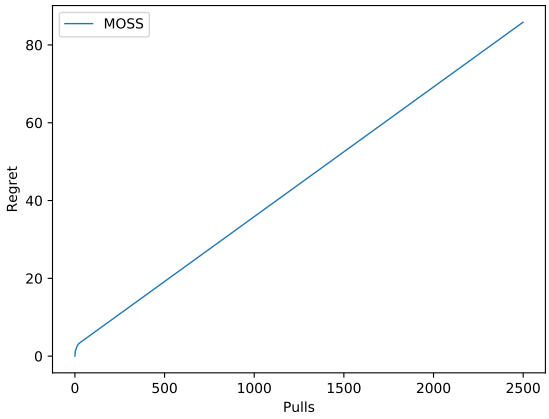


Figure 24: Average regret Moss: distribution 1

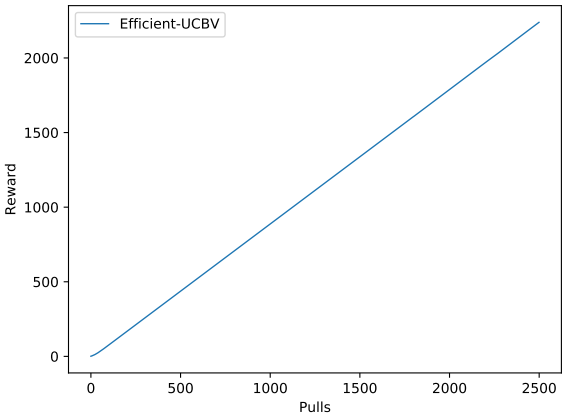


Figure 25: Average reward Efficient-UCBV : distribution 1

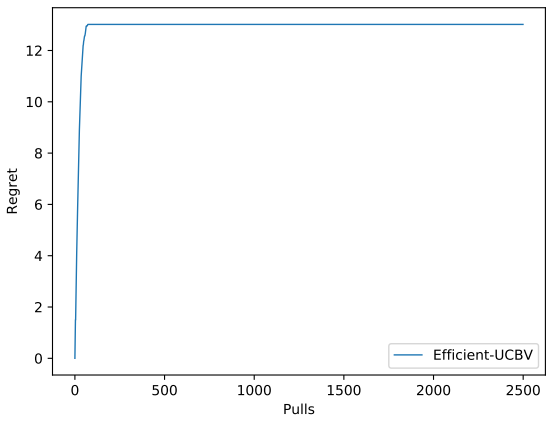


Figure 26: Average regret Efficient-UCBV : distribution 1

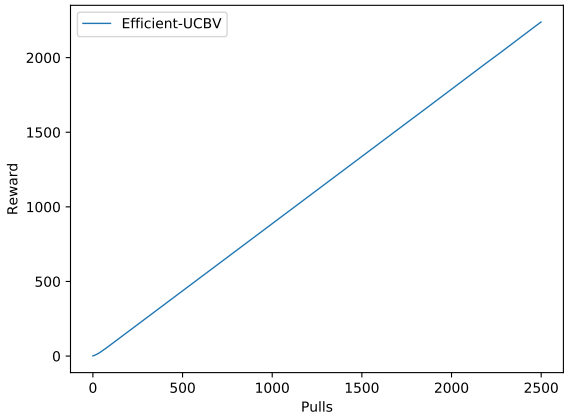


Figure 27: Average reward Efficient-UCBV : distribution 1

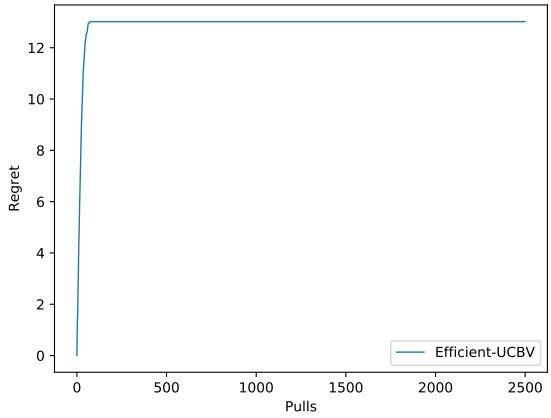


Figure 28: Average regret Efficient-UCBV : distribution 1

## References


Audibert, J. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits.

Burtini, G., Loeppky, J., and Lawrence, R. (2015). A survey of online experiment design with the stochastic multi-armed bandit. *CoRR*, abs/1510.00757.

Mukherjee, S., Naveen, K. P., Sudarsanam, N., and Ravindran, B. (2017). Efficient-ucbv: An almost optimal algorithm using variance estimates. *CoRR*, abs/1711.03591.

Seldin, Y., Szepesvri, C., Auer, P., and Abbasi-Yadkori, Y. (2013). Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments. 24:103–116.