A comparison of multi-armed bandit algorithms

Farooq Ahmad, Federica Stolf, Gian Luca Vriz and Daniele Zago

Department of Statistical Sciences, University of Padova

Statistical Models 2021-2022

Università degli Studi di Padova

Abstract

The real world offers circumstances in which individuals must simultaneously explore their options or choices while also **maximizing some variables** such as their output, well-being or wealth.

In **economic activities**, this is translated into a profit perspective. Logged bandit dataset are useful in this regard, with several procedures available in the literature to deal with the revenue maximization problem.

Considering the **click-rate** of a large-scale fashion e-commerce platform, we will compare some common **algorithms**. This informal comparison may suggest the policies that yield the most profit.

Motivating application

In online tech companies, problems such as website advertising yield data that is **continuously collected over time**.

Suppose you are an advertiser and you can choose to show for each visitor one out of a collection of ads. If your goal is to maximize the **number of clicks** over time, switching to the more **successful option** in a short time can increase your revenue.

Question: which ad we should show to the user?

This choice can be seen as a problem of finding a balance between **exploration** of new options and **exploitation** of the experience already gained.

The data

Off-policy evaluation (OPE) aims to estimate the performance of hypothetical policies using data generated by a different policy.

We focus on a **logged bandit dataset** collected on a large-scale fashion e-commerce platform, ZOZOTOWN.

We consider the dataset obtained by **randomly selecting** which item (arm) to present to the user.

We also have information relative to the position in which the item was presented to the user: top, middle, or bottom of the screen.

Our goal is to evaluate the performance of an algorithm which is designed to maximize the **total reward** (click rate) when applied in place of the random policy.

Multi-Armed Bandits

Background and notation

The available **logged data** is the collection $D = (x_i, a_i, r_i)$ for $i = 1, \dots, T$.

- $x \in X$ is the contextual information that the user receives (e.g item position).
- a is the arm that is presented to the user, i.e. the fashion item.
- r is the reward, whether the presented fashion item results in a click.

Rewards and contexts are sampled from unknown distributions $p(r \mid x, a)$ and $p(x)$.

$\pi : X \to A$ is a **policy**, with $\pi(a \mid x)$ the probability of taking action a given context x.

A **bandit algorithm** determines the policy π (i.e. chooses the arms) in order to maximize the total reward $\mathbb{E}[\sum_{i=1}^{T} r_i]$.

Performance estimation

Performance is estimated by considering the **cumulative click rate** $\mathbb{E}[\sum_{i=1}^{T} r_i/T]$.

Using **simulated data**, computation is straightforward since we can simulate the choice by sampling from the reward distribution.

Using **real data**, the below algorithm has to be used to obtain a consistent estimate of the click rate Li et al. [2010].

Algorithm Replay Bandit.

$h_0 \leftarrow \varnothing$
$\widehat{r} \leftarrow 0$
$T \leftarrow 0$
for $t = 1, 2, 3, \dots$ **do**
  Get the t^{th} event $(x_t, a_t, r_{a,t})$
  if $\pi(h_{t-1}, x) = a$ **then**
    $h_t \leftarrow \text{CONCATENATE}\big(h_{t-1}, (x_t, a_t, r_{a,t})\big)$
    $\widehat{r} \leftarrow \widehat{r} + r_{a,t}$
    $T \leftarrow T + 1$
  else
    $h_t \leftarrow h_{t-1}$
  end if
end for
return $\widehat{r}/T$

Policies

We compare the following policies.

- **Random**: the best arm is selected uniformly at random.
- ϵ**–greedy**: the current best arm is selected with probability $1-\epsilon$ (exploitation), otherwise another arm is selected uniformly at random (exploration).
- **Softmax**: the arm is sampled at random with the probability that depends on the estimated click probability, using the softmax function.
- **Thompson-Sampling**: the softmax approach is generalized by considering a Bayesian prior distribution on the probability of click for each arm.
- **UCB** (Upper Confidence Bound): a confidence interval is assigned to the click probability of each arm, and the one with the largest upper limit is selected.
- **Exp3**: given a parameter $\gamma \in [0, 1]$, the algorithm spends a fraction $(1-\gamma)$ of the time performing a weighted exploration/exploitation based on the estimated actual reward.

Simulated data

We apply the bandit algorithms and simulate the arm rewards using the **marginal click probabilities** of the observed dataset.

Figure 1 shows the median click rate for each considered algorithm over 100 simulations, whereas Figure 2 shows the median click rate and pointwise 95% intervals for the best-performing four algorithms.

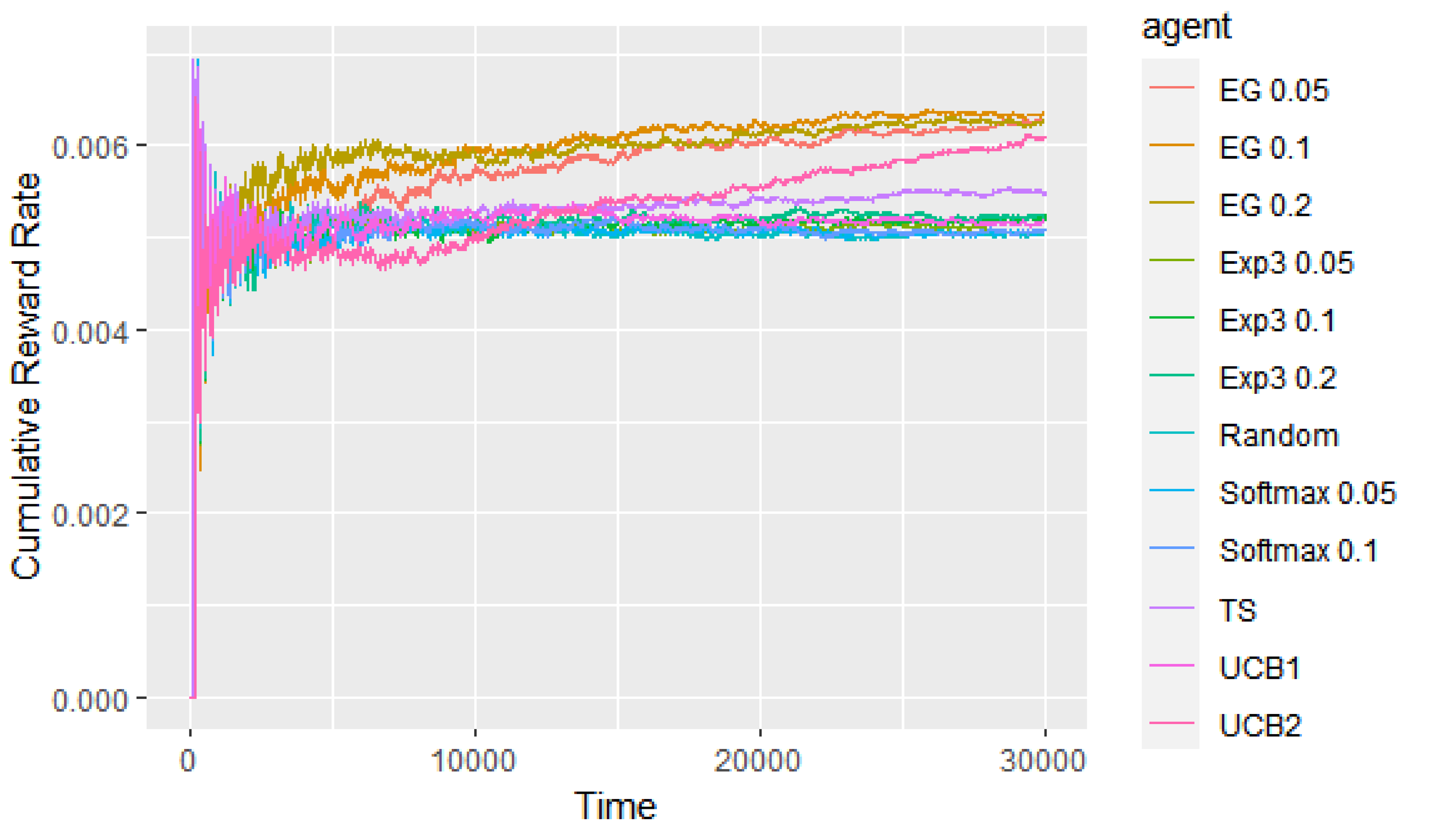

Figure 1: Median cumulative click rate for the bandit algorithms over 100 simulated runs.

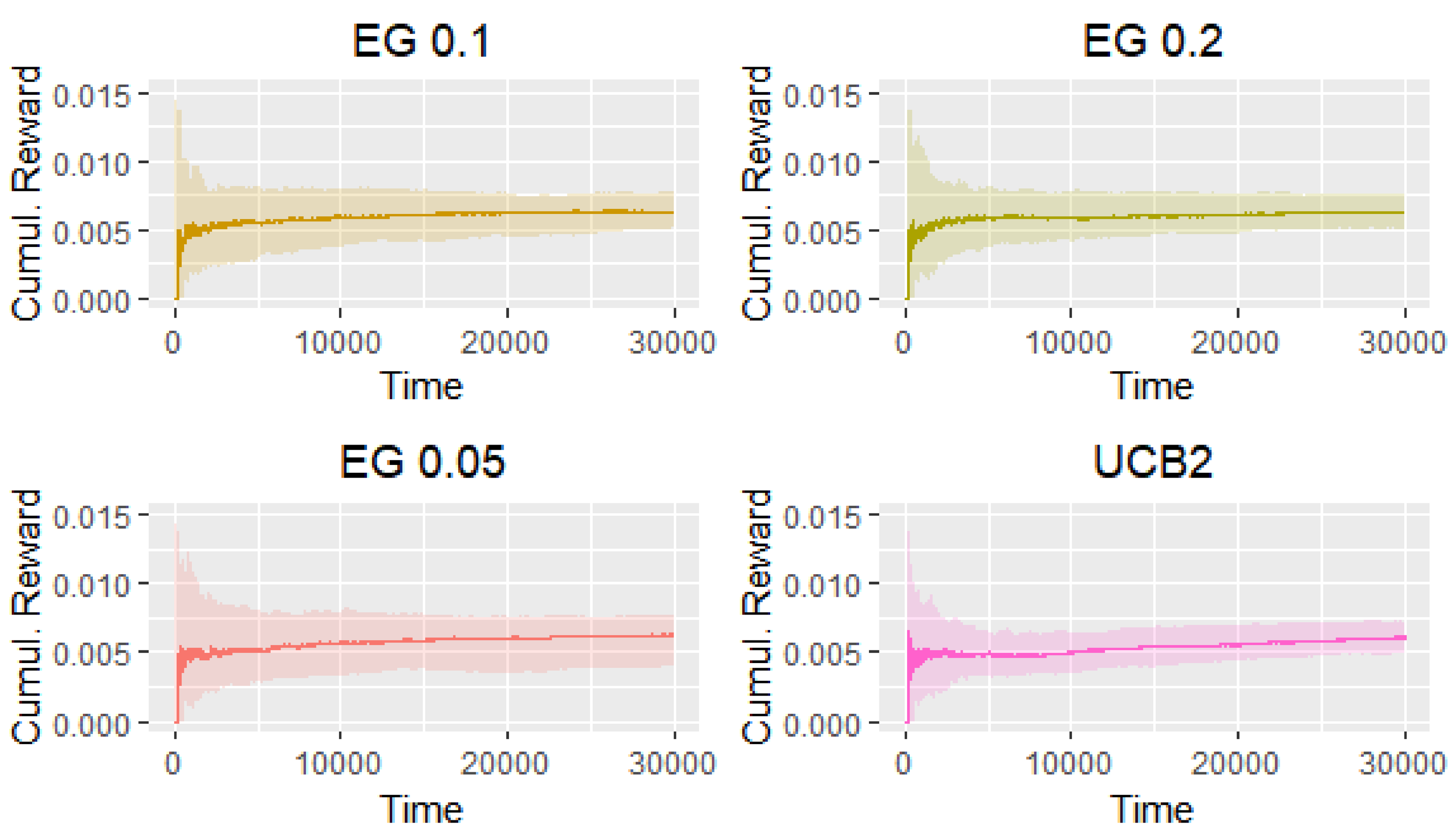

Figure 2: Median cumulative click rate and 95% pointwise interval for the best four bandit algorithms over 100 simulated runs.

Application to ZOZOTOWN Data

Analysis without covariates

We begin by considering multi-armed bandit policies **without covariates**.

For computational reasons (Replay Bandit algorithm) the temporal horizon is $T = 12000$.

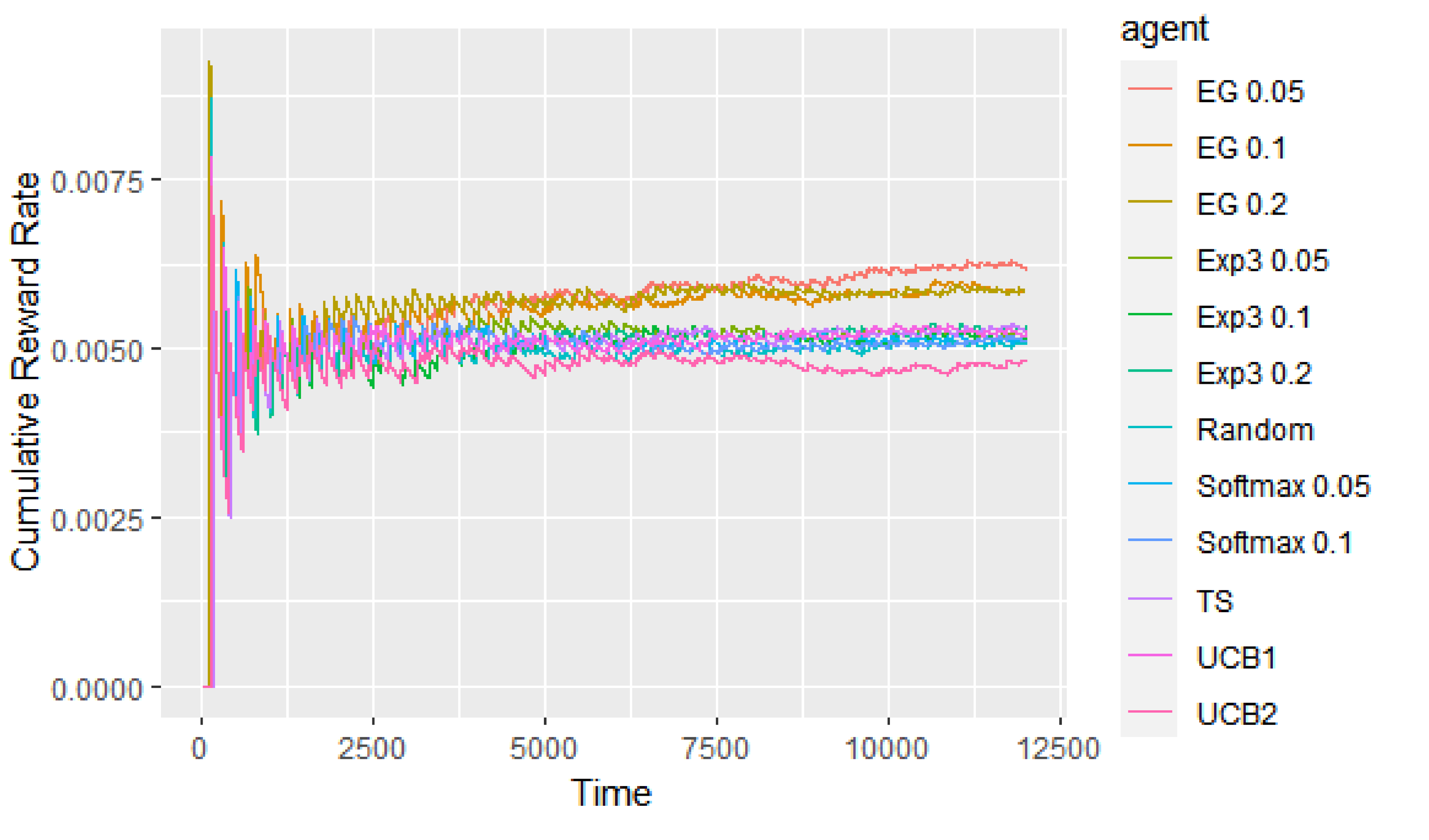

Figure 3: Median cumulative click rate for the bandit algorithms over 100 replications of the Replay Bandit on random subsets of the data.

Among the best performing algorithms on the ZOZOTOWN dataset, **EG 0.05** and **EG 0.1** are also present in the best performing algorithms on the simulated data.

Although simple, the ε-greedy algorithms seem to perform very well on arms with **low reward probability**.

Analysis with covariates

Performing multi-armed bandit policies **with position as covariate** does not add any significant improvement.

Scan the QR code for the Git repository!

References

Burtini, G., Loeppky, J. and Lawrence, R. (2015). A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit *arXiv:1510.00757.*

Saito, Y., Shunsuke, A., Megumi, M. and Yusuke, N.(2020). Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. *arXiv:2008.07146.*

Li, Z., Zhang, J. and Wang, Z. (2010). Self-Starting Control Chart for Simultaneously Monitoring Process Mean and Variance. *International Journal of Production Research*, **48**, 4537–4553.