Gambler Bandits and the Regret of Being Ruined

Filipo Studzinski Perotto
IRIT, University of Toulouse, France
filipo.perotto@irit.fr

Sattar Vakili
MediaTek Research, Cambridge, UK
sattar.vakili@mtkresearch.com

Pratik Gajane
DMIT, University of Leoben, Austria
pratik.gajane@unileoben.ac.at

Yaser Faghan
ISEG, University of Lisbon, Portugal
yaser.faghan@cemapre.pt

Mathieu Bourgais
LITIS, INSA of Rouen, France
mathieu.bourgais@insa-rouen.fr

ABSTRACT

In this paper we consider a particular class of problems called *multiarmed gambler bandits* (MAGB) which constitutes a modified version of the Bernoulli MAB problem where two new elements must be taken into account: the *budget* and the *risk of ruin*. The agent has an initial budget that evolves in time following the received rewards, which can be either +1 after a *success* or −1 after a *failure*. The problem can also be seen as a MAB version of the classic *gambler's ruin* game. The contribution of this paper is a preliminary analysis on the probability of being ruined given the current budget and observations, and the proposition of an alternative regret formulation, combining the classic regret notion with the expected loss due to the probability of being ruined. Finally, standard state-of-the-art methods are experimentally compared using the proposed metric.

MODIFIED PROBLEM

A *multiarmed gambler bandit* (MAGB) is a random process that exposes $k \in \mathbb{N}^+$ arms to an agent having an initial budget $b_0 \in \mathbb{N}^+$, which evolves in time with the received rewards:

$$B_h = b_0 + \sum_{t=1}^{h} R_t$$

Let $\mathcal{P} = \{p_1, \cdots, p_k\}$ be the set of parameters that regulate the underlying Bernoulli distributions from which the rewards $R_t \in \{+1, -1\}$ are drawn.

At each round $t \in \mathbb{N}^+$, the agent executes an action i, which either increases its budget B_t by 1 with stationary probability $p_i \in [0, 1]$, or decreases it by 1 with probability $1 - p_i$.

The game stops when $B_t = 0$ happens for the first time (the gambler is ruined), but it can be occasionally played forever if the initial conditions allow the budget to increase infinitely.

The probability of surviving, never being ruined, having a current budget B_t, and repeatedly pulling arm i, is:

$$\lim_{h\to\infty} \omega_{h,i} = \begin{cases} 1 - \left(\frac{1-p_i}{p_i}\right)^{B_t} & \text{if } p_i > 0.5\,, \\ 0 & \text{if } p_i \leq 0.5\,. \end{cases}$$

EXPERIMENTAL RESULTS

Setting:

- MAGB with $k = 10$ arms,
- half positive and half negative mean rewarded arms,
- p_i linearly distributed between 0.45 and 0.55,
- initial budget $b_0 = k = 10$,
- 2000 repetitions,
- time-horizon $h = 5000$.

Findings: `UCB1` presents a heavy regret due to its conservative behavior, which leads to intense exploration during the initial rounds, and often to ruin. The naive methods (`Empirical-Means`, `Empirical-Sum`, and ε-`Greedy`), which are classically sub-optimal, present better survival rates against the classically optimal algorithms (`Bayes-UCB`, `Thompson-Sampling`, and `KL-UCB`), which finally allows them to present better relative regret.

CONTRIBUTIONS

- The definition of a particular survival version of the *multi-armed bandits* problem called *multiarmed gambler bandits*;
- Experimental intuitions concerning the performance of standard methods in that context;
- The proposition of an alternative performance measure which integrates the *ruin cost* into the *regret*.

INTRODUCTION

The search for safety guarantees is receiving increased attention within the reinforcement learning community and in particular concerning multiarmed bandits.

Multiarmed bandits (MAB) constitute a framework to model online *sequential decision-making* while facing the *exploration-exploitation dilemma*.

A MAB is typically represented by an agent interacting with a discrete random process (or a "slot machine") by choosing, at each round t, some action $A_t = i$ to perform among k possible actions (or "arms"), then receiving a corresponding reward R_t, drawn from an unknown distribution.

The objective is to maximize the expected sum of rewards over a potentially infinite time-horizon.

PROPOSED METRIC

In contrast to the standard MAB, solving a MAGB involves a multi-objective optimization: in addition to minimizing the expected regret generated by the rounds when the best arm is not played (classic regret), the agent must also minimize the expected regret generated by the probability of being ruined.

To analyze that, we define the notion of *expected normalized relative regret* $\ell \in [0, 1]$:

$$\ell_{h,\pi} = \frac{\omega_{h,\pi}}{\omega^*_h} \cdot \underbrace{\sum_{i=1}^{k} \left[\frac{p^* - p_i}{p^*} \cdot \frac{\mathbb{E}[N_{i,h}]}{h} \right]}_{\text{normalized classic regret}} + \underbrace{\left(\frac{\omega^*_h - \omega_{h,\pi}}{\omega^*_h} \right)}_{\text{regret due to ruin}},$$

where h is the considered (potentially infinite) time-horizon, p^* and p_i are, respectively, the underlying parameters of the optimal arm and of arm i, $\mathbb{E}[N_{i,h}]$ is the number of rounds arm i is expected to be pulled, and $\omega_{h,\pi}$ and ω^*_h are the probability of surviving, respectively, following a given strategy π, or always playing the best arm.

In finite-horizon experimental scenarios, after several independent repetitions, the expected normalized relative regret can be approximated empirically by averaging the normalized difference between the obtained final budget and the potentially best budget:

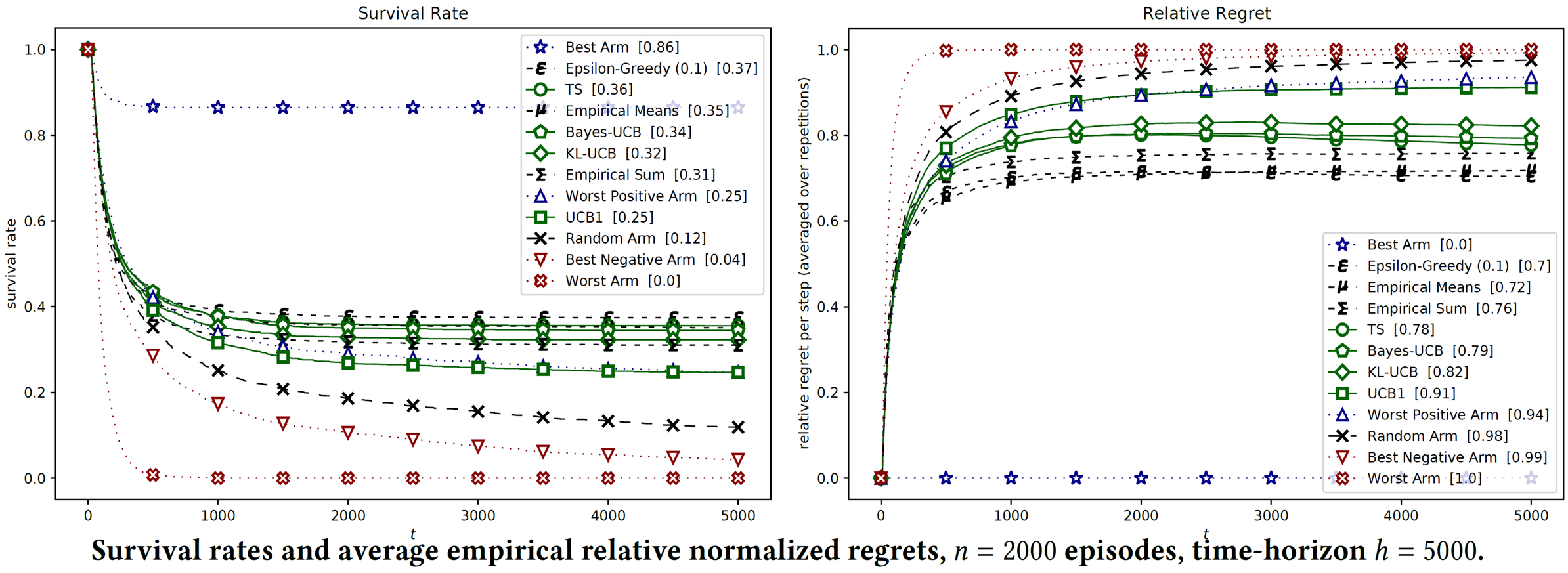
$$\hat{\ell}_{h,\pi} = 1 - B_{h,\pi} / B^*_h\,.$$

CONCLUSIONS

Taking the overall performance together, mixing the regret caused by sub-optimal choices (i.e. the regret in classic terms) and the regret caused by ruin, upsets the standard insights and strategies concerning MAB.

Intuitively, an algorithm for minimizing this alternative kind of regret must carefully coordinate the remaining budget with the confidence on the estimated distributions, seeking for minimizing the probability of ruin when the budget is relatively low, and gradually becoming classically optimal, as the budget increases.

Future works must include a more comprehensive set of experimental scenarios, a theoretical analysis about the regret bounds of the selected algorithms, and the extension of this survival setting to *Markovian Decision Processes*.



Survival rates and average empirical relative normalized regrets, $n = 2000$ episodes, time-horizon $h = 5000$.