ABSTRACT
In this paper we consider a particular class of problems called multi-armed 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 multi-armed gambler bandit (MAGB) is a random process that exposes k ∈ N* arms to an agent having an initial budget b0 ∈ N*, which evolves in time with the received rewards:

$$\tilde{b}_t = b_0 + \sum_{i=1}^{k} \tilde{r}_t$$

Let P = {p1, ..., pk} be the set of parameters that regulate the underlying Bernoulli distributions from which the rewards $\tilde{r}_t \in \{-1, 0, 1\}$ are drawn. At each round $t \in \mathbb{N}^*$, the agent executes an action $i$, which either increases its budget $\tilde{b}_t$ by 1 with stationary probability $p_i \in [0, 1]$, or decreases it by 1 with probability $1 - p_i$. The game stops when $\tilde{b}_t = 0$ 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 indefinitely.

The probability of surviving, never being ruined, having a current budget $\tilde{b}_t$, and repeatedly pulling arm $i$, is:

$$\lim_{h \to \infty} \phi_{hi} = \begin{cases} 1 - \frac{(1 - p_i)}{p_i} & \text{if } p_i > 0.5, \\ 0 & \text{if } p_i \leq 0.5. \end{cases}$$

EXPERIMENTAL RESULTS

- MAGB with k = 10 arms.
- Half positive and half mean reward arms.
- p_i.i.d. uniformly distributed between 0.45 and 0.55.
- Initial budget b0 = k = 10.
- 2000 simulations.
- Time horizon h = 5000.

Findings: UCBE presents a heavy regret due to its conservative behavior, which leads to intense exploration during the initial rounds, and often Collins. The naive methods (Empirical-Mean, Empirical-Sum and Random), which are classically sub-optimal, present lower survival rates against the classically optimal algorithms (Bayes–UCB, Thompson–Sampling, and KL–UCB), which finally allows them to present better regrets.

CONCLUSIONS
Taking the overall performance together, mixing the regret caused by sub-optimal choices (i.e., the regret in the classic bandit and the regret caused by ruin, upsetting the standard insights and strategies concerning MAGB).

Intuitively, an algorithm for minimizing this alternative kind of regret must carefully coordinate the remaining budget with the confidence level of 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.