Gambler Bandits and the Regret of Being Ruined
Filipo Studzinski Perotto, Sattar Vakili, Pratik Gajane, Yaser Faghan, Mathieu Bourgais

To cite this version:
Filipo Studzinski Perotto, Sattar Vakili, Pratik Gajane, Yaser Faghan, Mathieu Bourgais. Gambler Bandits and the Regret of Being Ruined. 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), International Foundation for Autonomous Agents and Multiagent Systems (IFAAAMAS), May 2021, London (fully virtual event), United Kingdom. hal-03120813

HAL Id: hal-03120813
https://hal.archives-ouvertes.fr/hal-03120813
Submitted on 17 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Gambler Bandits and the Regret of Being Ruined

Extended Abstract

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 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.

KEYWORDS

MAB; Ruin; Risk-Averse Decision Making; Safe RL

ACM Reference Format:

1 MAB AND MAGB

Multiarmed bandits (MAB) constitute a framework to model online sequential decision-making while facing the exploration-exploitation dilemma [37, 46, 48]. 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 $R_t$. Because the complete information about the reward functions is not available, the agent must estimate them by sampling (i.e. by pulling the arms and observing the received rewards). In the standard stochastic setting [7], the rewards originated from the same arm are independent but identically distributed, and observing an arm does not give any information about other arms. The objective is to maximize the expected sum of rewards over a potentially infinite time-horizon, finding a strategy that minimizes the expected regret (i.e. the cumulated difference between the rewards that could be obtained by always pulling the arm with highest mean, and the rewards the agent expects to receive following the given strategy). A good policy should guarantee sub-linear regret for any configuration of arms, i.e. the expected average regret per round must tend to zero asymptotically as time tends to infinity [1, 10, 28].

In this paper, we define a particular MAB variation called multi-armed gambler bandits (MAGB), which constitutes a subclass of survival MAB [44]. A MAGB can be formally defined as 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, so that $B_t = b_0 + \sum_{s=1}^{t} R_t$. Let $\mathcal{P} = \{ p_1, \ldots, p_k \}$ be the set of parameters that regulate the underlying Bernoulli distributions from which the rewards $R_t \in \{+1, -1\}$ are drawn. It means that, 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.

When taken separately, each arm within a MAGB can be seen as an instance of a gambler’s ruin game played against an infinitely rich adversary [20, 21, 29, 47]. For that reason, the probability of surviving, playing the game forever, and never being ruined, having a current budget $B_t$, and repeatedly pulling arm $i$, is:

$$\lim_{k \to \infty} \omega_{h,i} = \begin{cases} 1 - \left(\frac{1-p_i}{p_i}\right)^B_0 & \text{if } p_i > 0.5, \\ 0 & \text{if } p_i \leq 0.5. \end{cases}$$ (1)

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 $\ell \in [0, 1]$:

$$\ell_{h,\pi} = \frac{\omega_{h,\pi}}{\omega_h^*} \sum_{i=1}^{k} \left[ \frac{p_i^* - p_i}{p_i^*} \cdot \mathbb{E}[N_{i,h}] + \frac{\omega_h^* - \omega_h}{\omega_h^*} \right],$$ (2)

where $h$ is the considered (potentially infinite) time-horizon, $p_i^*$ 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 $\omega_h^*$ are the probability of surviving, respectively, following a given strategy $\pi$, 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:

\[ \hat{t}_{h,\pi} = 1 - B_{h,\pi}/B^*_h. \]

2 RELATED WORKS ON SAFE BANDITS

The search for safety guarantees is receiving increased attention within the reinforcement learning community [5, 9, 12, 14–16, 19, 25, 26, 42, 43, 54] and in particular concerning multiarmed bandits [23, 24]. In an alternative version of the problem called risk-averse MAB [13, 22, 40, 45, 51, 52, 58], the agent must take into account the expected variability on the expected rewards in order to identify (and avoid) unstable (then considered risky) actions, but without worrying about ruin, since the notion of budget is not considered. In this sense, the risk-reward trade-off can be tackled by using some risk-aware metric, such as the mean-variance, or the conditional value at risk. A MAGB cannot be reduced to the risk-averse setting due to the absence of notion of ruin, which leads to a simplified interpretation of safety as a synonym of reward constancy. In addition, in the Bernoulli case, both mean and variance are directly dependent on \( p \). In another variation of the problem called conservative bandits [24, 55], the agent knows, a priori, a default action with its underlying reward mean, and it is constrained to respect a threshold in the ongoing relative regret compared to that action.

In another modified version of the problem called budgeted MAB [2, 4, 8, 17, 18, 34, 35, 38, 49, 50, 56, 57], the player receives a reward but needs to pay a cost after pulling an arm, which is taken from a given initial budget. The process stops when the budget is over. In this setting, reward and cost are independent functions associated to each arm. The goal is to maximize cumulated rewards, constrained by a budget that limits the cumulated costs. The arm with best estimated reward-to-cost ratio should be preferred. Alternatively, the budget can be imposed only on a preliminary exploration phase [6, 11, 30, 39], and the question is how to spend the budget efficiently in order to identify the best arm. A MAGB cannot be reduced to any of those budgeted settings due to the explicit separation between rewards and costs, which does not exist in a MAGB.

3 FINDINGS AND CONCLUSIONS

In the experimental setting, a MAGB with \( k = 10 \) arms is instantiated, each one with a different parameter \( p_i \), linearly distributed between 0.45 and 0.55 (i.e. half positive and half negative mean rewarded arms). The initial budget is set to \( B_0 = k = 10 \), and the results are averaged after 2000 repetitions and over time-horizon \( h = 5000 \). The standard UCB1 method [7, 41] is compared with other state-of-the-art MAB algorithms, namely KL-UCB [27], Bayes-UCB [31], and Thompson-Sampling [3, 32, 33], which have proven to asymptotically achieve logarithmic regret for Bernoulli arms, matching the accepted theoretical lower bound [36], but also with classic naive sub-optimal methods, namely Empirical-Means and \( \epsilon \)-greedy [37, 48, 53], and with an original simple heuristic called Empirical-Sum, which chooses, at each round, the arm with highest observed sum of rewards. Finally, some fixed arm policies are included, namely Best-Arm (always pull the arm with highest mean), Worst-Arm (the arm with lowest mean), Worst-Positive-Arm (the arm with lowest \( p \) greater than 0.5), and Best-Negative-Arm (the arm with highest \( p \) lower than 0.5).

The methods are compared considering their survival rate, defined by the proportion of episodes that run without ruin until the predefined time-horizon, and considering their empirical normalized relative regret, given by Eq. (3), as shown in the Figure 1. 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 \( \epsilon \)-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.

![Figure 1: Survival rates and average empirical relative normalized regrets, \( n = 2000 \) episodes, time-horizon \( h = 5000 \).](image)

In conclusion, 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.


