Time is Budget: A Heuristic for Reducing the Risk of Ruin in Multi-armed Gambler Bandits

Filipo Studzinski Perotto, Xavier Pucel, Jean-Loup Farges

Time is Budget: A Heuristic for Reducing the Risk of Ruin in Multi-armed Gambler Bandits

Filipo Studzinski Perotto(✉), Xavier Pucel, and Jean-Loup Farges

ONERA - The French AeroSpace Lab, DTIS - Information Processing and Systems Department, 2 Avenue Edouard Belin, 31055 Toulouse, France
filipo.perotto@onera.fr
https://www.onera.fr

Abstract. In this paper we consider *Multi-Armed Gambler Bandits* (MAGB), a stochastic random process in which an agent performs successive actions and either loses 1 unit from its budget after observing a *failure*, or earns 1 unit after a *success*. It constitutes a survival problem where the *risk of ruin* must be taken into account. The agent's initial *budget* evolves in time with the received rewards and must remain positive throughout the process. The contribution of this paper is the definition of an original heuristic which aims at improving the probability of survival in a MAGB by replacing the time by the budget as the factor that regulates exploration in UCB-like methods. The proposed strategy is then experimentally compared to standard algorithms presenting good results.

Keywords: Multi-armed bandits · Risk of ruin · Safety-critical systems

1 Introduction

Multi-Armed Bandits (MAB) constitute a classic framework to model online *sequential decision-making* while facing the *exploration-exploitation dilemma* [11,16]. A MAB is typically represented by an agent interacting with a random process. At each successive round t, the agent chooses an action A_t to perform among k possible actions and receives a corresponding reward R_t. The agent must estimate the reward functions associated to each action by sampling them. Rewards resulting from a same action are independent but identically distributed, and do not give any information about other actions. In that standard version, budget and risk are not taken into account. The objective is to maximise the expected future sum of rewards [1]. Different methods and guarantees have been proposed in the literature depending on the available information and on the assumptions on the reward distributions [3,8].

Survival Multi-Armed Bandits (SMAB) [13,14] and in particular *Multi-Armed Gambler Bandits* (MAGB) [12] are recent extensions of the standard

M. Bramer and F. Stahl (Eds.): SGAI-AI 2022, LNAI 13652, pp. 346–352, 2022.
https://doi.org/10.1007/978-3-031-21441-7_29

MAB problem in which the agent has a budget that must remain positive throughout the process, otherwise the agent is ruined. An initial budget $B_0 = b$ evolves with the received rewards while the agent is alive, so as $B_t = B_{t-1} + R_t$ if $B_{t-1} > 0$, otherwise $B_t = 0$, the agent is ruined and the process no longer evolves. In that scenario, the agent can either increase the probability of running the process indefinitely, becoming infinitely rich, or inversely, can increase the probability of ruin, until eventually running out of budget, which means that maximising the sum of rewards requires reducing the chances of being ruined.

This paper focuses on MAGB problems [12], where the rewards are limited to two values, $+1$ and -1, and the initial budget is a positive integer. When occasionally $B_t = 0$ is achieved for the first time, the agent is ruined. The rewards are drawn from underlying stationary Bernoulli distributions. Formally, $\{k \in \mathbb{N} \mid k \geq 2\}$ is the number of actions, $\{b \in \mathbb{N} \mid b > 0\}$ is the initial budget, $\{p_i \in \mathbb{R} \mid 0 \leq p_i \leq 1\}$ is the probability of *success* after executing action i, which returns reward $+1$, and $1 - p_i$ is the complementary probability of *failure*, with reward -1. It means that $X_t \sim Bern(p_i)$ and $R_t = 2X_t - 1$ for $A_t = i$. The expected mean reward of action i is $\mu_i = 2p_i - 1$.

There are few results concerning SMAB and MAGB into the literature [12–14], and the definition of an optimal algorithm is still an open problem. Related extensions like *Risk Averse* [4,7,15,18], and *Budgeted* MAB [2,6,20], even if sharing similar concerns, cannot be reduced to the survival setting [13,14].

2 Standard MAB Algorithms

Lets assume that the agent always plays each action once at the beginning of the process, so as $A_t = t$ for $1 \leq t \leq k$, in order to provide the decision algorithm with a first observation of them. `Empirical-Means` is a greedy algorithm which successively chooses the action with the best estimated mean reward, $A_{t+1} = \arg\max_i \frac{S_{i,t}}{N_{i,t}}$, where $N_{i,t}$ is the number of times the action i had been performed until round t, and $S_{i,t}$ is the sum of received rewards due to that action. That strategy is sub-optimal since no systematic exploration is performed, then it may not converge to the best action.

Exploration can be performed by introducing some non-determinism on the decision. ε`-Greedy` is a naive algorithm which chooses either the action with best estimated mean reward with probability ε, a hand-tuned parameter, or a random action otherwise. That strategy is sub-optimal since the exploration rate remains constant throughout the process [5,19].

The standard approach for solving the exploration-exploitation dilemma is the *optimism in the face of uncertainty*. An intelligent exploration can be made by statistically controlling the confidence on the estimates. With similar mean reward, less explored actions should be preferred. `UCB1` [1] chooses, at each time t, the action that maximises the estimated mean plus the maximum estimation error given by a confidence bound that progressively increases over time, so as:

$$A_{t+1} = \arg\max_{1 \leq i \leq k} \left[\frac{S_{i,t}}{N_{i,t}} + \sqrt{\frac{\alpha \ln(t)}{N_{i,t}}} \right], \tag{1}$$

where α is the parameter regulating exploration. That strategy is asymptotically optimal if α is sufficiently high.

Estimating the parameters of a Bernoulli bandit corresponds to estimating the parameters of a binomial distribution. The binomial distribution represents the probability of a given number of successes on a sequence of Bernoulli trials when the parameter is known. The probability of having x successes in n trials, given p is $\mathbb{P}(x \mid p, n) = \text{Bin}(p, n) = \binom{n}{x} p^x (1-p)^{(n-x)}$. In a Bayesian approach, the beta distribution corresponds to the conjugate prior for the binomial distribution. Assuming a uniform prior, the posterior density function for p is given by $f(p \mid x, n) = \text{Beta}(x+1, n-x+1) = p^x(1-p)^{n-x}(n+1)\binom{n}{x}$. `Bayes-UCB` [9] is an improved UCB-like method designed for Bernoulli bandits that is also asymptotically optimal. It chooses the action that maximises the $1 - \frac{1}{t}$ quantile from the Beta posterior:

$$A_{t+1} = \arg\max_{1 \le i \le k} \left[\text{Q}_{1-1/t}\big(\text{Beta}(X_{i,t}+1, N_{i,t} - X_{i,t} + 1)\big) \right]. \tag{2}$$

Finally, `Thompson-Sampling` is another optimal Bayesian algorithm [10]. At each round, it draws a sample from the posterior of each action to decide which one to choose. This allows a non-optimal action to be sampled with a varying frequency, which dynamically balances exploration as the posterior becomes more precise:

$$A_{t+1} = \arg\max_{1 \le i \le k} \left[V_{i,t} \sim \text{Beta}(X_{i,t}+1, N_{i,t} - X_{i,t} + 1) \right]. \tag{3}$$

3 Our Contribution: The Gambler Methods

In a MAGB, the probability of being ruined by always performing action i is $\left(\frac{1-p_i}{p_i}\right)^b$ if $p_i > \frac{1}{2}$, and 1 if $p_i \le \frac{1}{2}$ [12]. The expected duration of the game is $\frac{b}{1-2p_i}$ if $p_i < \frac{1}{2}$, and ∞ if $p_i \ge \frac{1}{2}$. It means that, in a MAGB, the action with highest mean presents the best life expectancy and the best survival probability, independent of the current budget, and then, like in the classic MAB, the action with maximal mean reward is the optimal action, to which optimal methods must asymptotically converge. However, when exploring, the agent must consider the remaining budget and the estimated parameters of each action in order to compare the estimated ruin probabilities associated to them.

In this paper, we suggest a heuristic modification that can be applied to UCB-like methods, which consists in replacing t for B_t into the considered equations. The intuition is that, for maximising the chances of survival, the lowest is the budget, the more the agent must favour exploitation over exploration in order to increase its budget and avoid ruin. In contrast, the higher is the budget, the more the agent should prefer a classic optimal strategy.

In this way, the `Gambler-UCB` method modifies `UCB1` (Eq. (1)) by adding $\sqrt{\frac{\alpha \ln(B_t)}{N_{i,t}}}$ instead of $\sqrt{\frac{\alpha \ln(t)}{N_{i,t}}}$ to the estimated mean:

$$A_{t+1} = \arg\max_{1 \le i \le k} \left[\frac{S_{i,t}}{N_{i,t}} + \sqrt{\frac{\alpha \ln(B_t)}{N_{i,t}}} \right], \tag{4}$$

and the `Gambler-Bayes-UCB` method modifies `Bayes-UCB` (Eq. (2)) by taking the $1 - {}^1/_{B_t}$ quantile from the beta posterior, instead of the $1 - {}^1/_t$ quantile proposed on the original method:

$$A_{t+1} = \arg\max_{1 \le i \le k} \left[\mathrm{Q}_{1-{}^1/_{B_t}} \big(\mathrm{Beta}(X_{i,t} + 1, N_{i,t} - X_{i,t} + 1) \big) \right]. \qquad (5)$$

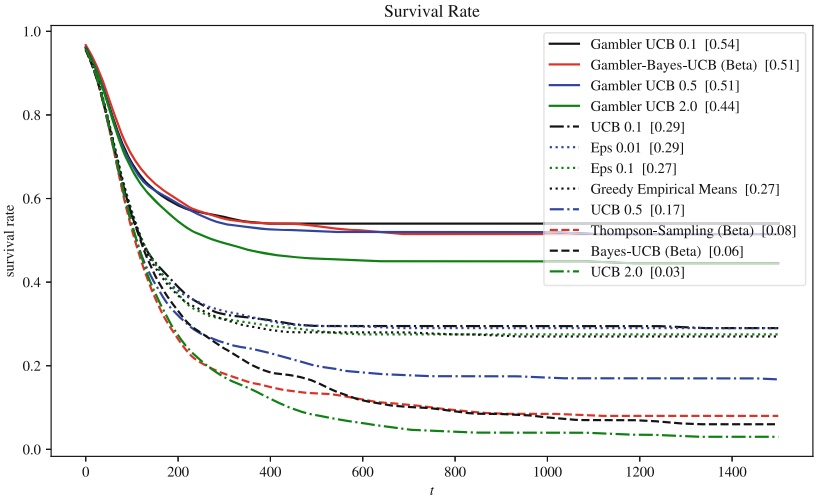

Fig. 1. Survival rate, i.e. the proportion of episodes in which the agent reaches the time-horizon $h = 1500$ without ruin in $n = 200$ episodes.

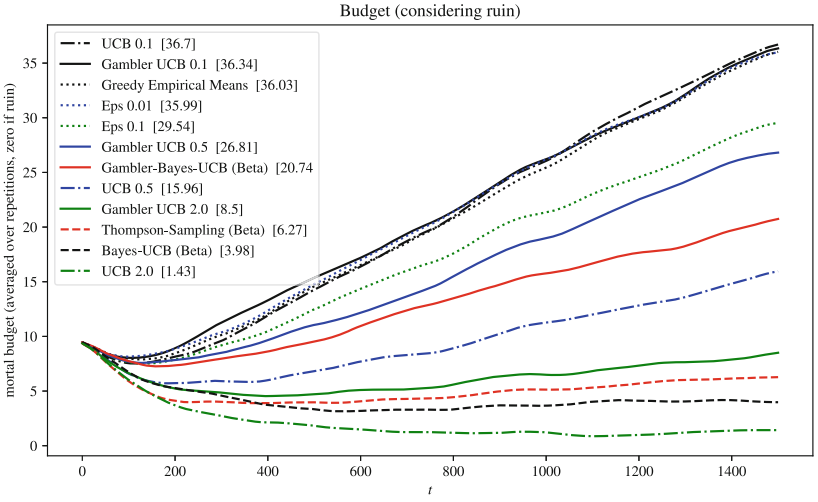

Fig. 2. Budget progression as a function of time, averaged over $n = 200$ episodes.

4 Experimental Results

The proposed algorithms, `Gambler-UCB` (Eq. (4)) and `Gambler-Bayes-UCB` (Eq. (5)) have been compared to classic and state-of-the-art MAB algorithms like `UCB1` [1], `Bayes-UCB` [9], and `Thompson-Sampling` [10], but also with naive methods like `Empirical-Means` and ε-`Greedy` [17]. The experimental scenario presents $k = 10$ actions, the first 8 of them parameterised by $p_1 = ... = p_8 = 0.45$, which corresponds to a slightly negative mean reward, leading the agent to ruin, and the two last ones defined as $p_9 = 0.525$ and $p_{10} = 0.55$, meaning that both are slightly positive, but only the last one is optimal. The initial budget was set to $b = k = 10$.

In the experiences, the survival rate, presented in the Figure 1, corresponds to the ratio between the number of episodes running until the defined time-horizon without ruin over the total number of episodes. The proposed methods performed significantly better than the other methods considering the survival rates, corroborating the intuition. `Gambler-UCB`, with parameter α varying between 0.1 and 2.0, as well as `Gambler-Bayes-UCB`, reached survival rates around 50%. The hypothetical oracle strategy (not shown in the graphic), which always plays the best action, ensures about 80% of survival. `UCB1` and ε-`Greedy` presented survival rates below 30%, even when the exploration parameters (α and ε, respectively) have been set to small values, losing theoretical guarantees of convergence. `Bayes-UCB` and `Thompson-Sampling`, both ensuring the best theoretical guarantees against the classic Bernoulli MAB problem, presented bad survival rates against the experienced MAGB, lower than 10%, due to intense exploration in the initial rounds. The greedy `Empirical-Means` method performs as well as the standard methods, reaching almost 30% of survival, corroborating the findings on [12].

The Figure 2 presents the average budget progression, which is affected by the survival rate. If the agent is ruined during an episode, its budget remains $B_t = 0$ until the simulation reaches the predefined time-horizon $h = 1500$. In terms of budget, the performance of the proposed methods is disappointing. `UCB1` and ε-`Greedy` with low exploration, as well as `Empirical-Means`, presented the best performance on the proposed setting. Even if one instance of `Gambler-UCB` reaches similar performance, the fact of having superior survival rates indicates that it is making sub-optimal choices too often. It means that the proposed heuristics are not converging to the optimal action, or are converging too slowly.

5 Conclusion and Perspectives

This paper approaches a Multi-Armed Bandit setting called MAGB, a specific survival MAB problem, in which, in addition to solve the classical exploration-exploitation dilemma, the agent must find a good trade-off between safety and risk to avoid ruin, still trying to maximise the sum of rewards. Two algorithms have been proposed, `Gambler-UCB` and `Gambler-Bayes-UCB`, modifying respectively `UCB1` and `Bayes-UCB` by replacing the time by the budget as the parameter

that regulates exploration. The new methods presented good results in experimental simulations considering the survival rate, but are apparently sub-optimal in terms of convergence to the best action. Both methods are the result of a simple and intuitive heuristic, that seems to be efficient for preserving the agent alive during the initial rounds of the process, when it is more vulnerable to ruin, but the modified equations do not ensure gradative convergence to the best action. The results are nevertheless very promising, and the proposed heuristics should be the subject of theoretical analyses in future works, in order to find the necessary adjustments for ensuring optimality.

References

1. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Mach. Learn. **47**(2–3), 235–256 (2002)
2. Badanidiyuru, A., Kleinberg, R., Slivkins, A.: Bandits with knapsacks. J. ACM **65**(3), 13:1-13:55 (2018)
3. Bubeck, S., Munos, R., Stoltz, G., Szepesvári, C.: X-armed bandits. JMLR **12**, 1655–1695 (2011)
4. Cassel, A., Mannor, S., Zeevi, A.: A general approach to multi-armed bandits under risk criteria. In: Proceedings of the COLT 2018, pp. 1295–1306 (2018)
5. Cesa-Bianchi, N., Fischer, P.: Finite-time regret bounds for the multiarmed Bandit problem. In: Proceedings of the 5th ICML, pp. 100–108. Morgan Kaufmann (1998)
6. Ding, W., Qin, T., Zhang, X., Liu, T.: Multi-armed bandit with budget constraint and variable costs. In: Proceedings of the 27th AAAI (2013)
7. Galichet, N., Sebag, M., Teytaud, O.: Exploration vs exploitation vs safety: risk-aware multi-armed bandits. In: Proceedings of the 5th ACML, vol. 29, pp. 245–260. PMLR (2013)
8. Garivier, A., Hadiji, H., Ménard, P., Stoltz, G.: KL-UCB-Switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. CoRR abs/1805.05071 (2018)
9. Kaufmann, E., Cappé, O., Garivier, A.: On Bayesian upper confidence bounds for bandit problems. In: Proceedings of the 15th AISTATS, pp. 592–600 (2012)
10. Kaufmann, E., Korda, N., Munos, R.: Thompson sampling: an asymptotically optimal finite-time analysis. In: Proceedings of the 23rd ALT, pp. 199–213 (2012)
11. Lattimore, T., Szepesvári, C.: Bandit Algorithms. Cambridge University Press, Cambridge (2020)
12. Perotto, F.S.: Gambler bandits and the regret of being ruined. In: Proceedings of the 20th AAMAS, pp. 1664–1667 (2021)
13. Perotto, F.S., Bourgais, M., Silva, B.C., Vercouter, L.: Open problem: risk of ruin in multiarmed Bandits. In: Proceedings of the COLT 2019, pp. 3194–3197 (2019)
14. Riou, C., Honda, J., Sugiyama, M.: The survival bandit problem (2022). https://doi.org/10.48550/ARXIV.2206.03019
15. Sani, A., Lazaric, A., Munos, R.: Risk-aversion in multi-armed bandits. In: Proceedings of the 26th NIPS, pp. 3284–3292 (2012)
16. Slivkins, A.: Introduction to multi-armed Bandits. Found. Trends Mach. Learn. **12**(1–2), 1–286 (2019)
17. Sutton, R., Barto, A.: Introduction to Reinforcement Learning. MIT Press, Cambridge (1998)

18. Vakili, S., Zhao, Q.: Risk-averse multi-armed bandit problems under mean-variance measure. J. Sel. Top. Signal Process. **10**(6), 1093–1111 (2016)
19. Vermorel, J., Mohri, M.: Multi-armed Bandit Algorithms and Empirical Evaluation. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 437–448. Springer, Heidelberg (2005). https://doi.org/10.1007/11564096_42
20. Xia, Y., et al.: Finite budget analysis of multi-armed bandit problems. Neurocomputing **258**, 13–29 (2017)