Open Problem: Risk of Ruin in Multiarmed Bandits

Filipo Studzinski Perotto, Mathieu Bourgais, Laurent Vercouter, Bruno Castro da Silva

Proceedings of Machine Learning Research vol 99:3194–3197, 2019 32nd Annual Conference on Learning Theory

Open Problem: Risk of Ruin in Multiarmed Bandits

Filipo Studzinski Perotto FILIPO.PEROTTO@LITISLAB.FR
Mathieu Bourgais MATHIEU.BOURGAIS@INSA-ROUEN.FR
Laurent Vercouter LAURENT.VERCOUTER@INSA-ROUEN.FR
Normandy University (UniRouen / UniHavre / INSA) - LITIS Lab, Rouen, France

Bruno Castro da Silva BSILVA@INF.UFRGS.BR
Federal University of Rio Grande do Sul (UFRGS) - Instituto de Informática, Porto Alegre, Brazil

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We formalize a particular class of problems called *survival multiarmed bandits* (S-MAB), which constitutes a modified version of *budgeted multiarmed bandits* (B-MAB) where a true *risk of ruin* must be considered, bringing it closer to *risk-averse multiarmed bandits* (RA-MAB). In a S-MAB, pulling an arm can result in both positive and negative rewards. The agent has an initial budget that evolves in time with the received rewards. The goal is finding a good *exploration-exploitation-safety* trade-off, maximizing rewards while minimizing the probability of getting ruined (i.e. hitting a negative budget). Such simple and until now neglected modification in the MAB statement changes the way to approach the problem, asking for adapted algorithms and specific analytical tools, and also make it more likely related to some important real-world applications. We are interested in the following open problems which stem from such new MAB definition: (a) how can the regret be meaningfully defined in formal terms for a S-MAB given its multiobjective optimization nature? (b) can a S-MAB be reduced to a RA-MAB or a B-MAB, transferring their theoretical guarantees? (c) what kind of method or strategy must an agent follow to optimally solve a S-MAB?

Keywords: Budgeted Multiarmed Bandits, Ruin Theory, Risk-Averse Decision Making

1. Introduction

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 process by choosing some action i to perform among k possible actions (or "arms"), then receiving a corresponding reward r. Because the complete information about the reward functions is not available, the agent must estimate their parameters by sampling them (i.e. by pulling the arms and observing the received rewards). Solving a MAB means maximizing the expected rewards to be received by the agent over time, finding the best trade-off between *exploration* (sampling arms) and *exploitation* (pulling the estimated best arm). Different methods and guarantees have been proposed in the literature depending on the available information and on the assumptions made about the reward functions (Bubeck et al., 2011). However, most of the existing bandit-related research ignores any notion of cost and the possible finitude of the resources for pulling arms.

In this paper we are interested in a class of problems that we call *survival multiarmed bandits* (S-MAB), which constitutes a version of MAB where a true *risk of ruin* must be considered. In a S-MAB, an initial budget b_0 evolves in time following the received rewards, so as $b_{t+1} = b_t + r_t$. The returned reward can just as well be positive as negative, and affects the budget, which corresponds

to the agent's "capital" or "fortune". Therefore, solving a S-MAB means estimating the best arm, maximizing the budget over time, but managing the costs of exploratory actions to avoid ruin. To do it, the agent must consider the uncertainty on the expected rewards, which can be a consequence of both high distribution variances and low confidence on the estimations.

If we see the budget as an *essential variable*, e.g. like an "energy level", a S-MAB becomes a survival problem, and can be associated to the cybernetic notion of *homeostasis*: the agent wants to learn the optimal action but remaining alive, i.e. preserving a positive budget during all its lifetime. The MAB problem defined as a S-MAB becomes more closely related to some general real-world problems: while trading-off between exploration and exploitation, trying to maximize their winnings, gamblers in a casino have some money and must avoid ruin, investors in the stock market have a bankroll and must avoid bankruptcy, organisms in the nature have essential variables, inherent to their structure, and must avoid death. The exploration-exploitation dilemma becomes more complex when such risk of ruin is taken into account.

2. Survival MAB

Considering a stochastic approach, a S-MAB can be formally defined as $\mathcal{M} = \{I, F, b_0\}$, where $I = \{1, ..., k\}$ is the set of possible actions, $F = \{f_1, ..., f_k\}$ is the set of reward distribution functions, and $\{b_0 \in \mathbb{R} \mid b_0 \geq 0\}$ is the initial budget. The process evolves discretely over time. Each observation of a given arm is independent and identically distributed. Let $a_t = i$ be the action chosen (the arm pulled) and r_t be the immediate reward received at time t (drawn from f_i). A stationary reward distribution function $f_i \in F$ defines the probability of having reward $r_t \in \mathbb{R}$ after pulling arm i at time t, i.e. $\{r_t \sim f_i \mid a_t = i\}$. Received rewards impact the budget, increasing or decreasing it in the form $b_h = b_0 + \sum_{t=1}^{h} r_t$, where b_h is the budget in time h, and r_t is the reward received in time t.

In a S-MAB, the reward support $\mathcal{D} = [r_{min}, r_{max}]$ is given, where $\{r_{min} < 0 < r_{max}\}$. At least one arm presents a negative mean reward and at least one presents a positive mean, i.e. $\{\exists i, j \in I \mid \mu_i < 0 < \mu_j\}$. In such scenario, depending on the initial budget, on the arms' parameters, and on the chosen strategy, and supposing an infinite time-horizon ($h \to \infty$), the agent can either increase the probability of running the process indefinitely, becoming infinitely rich (i.e. $\tau \to \infty, b_\tau \to \infty$), or inversely, can increase the probability of ruin, until eventually getting broke (i.e. $\tau < \infty, b_\tau < 0$), with τ the time where the agent is ruined.

Classically, solving a MAB means finding a policy that can minimize the *reward regret* λ (i.e. the cumulated difference between the rewards expected by following the given strategy and the rewards that could be obtained by always pulling the arm with highest true mean reward, which is unknown to the agent) (Auer et al., 2002). Considering a given time-horizon h (which can be infinite), the *theoretical cumulative regret* is defined as $\lambda_h = \sum_{t=1}^{h} [\mu^* - \mu_{a_t}]$, where $\mu^* = \max_{i \in I}[\mu_i]$. However, in a S-MAB, the single reward regret minimization is not sufficient for evaluating the quality of an algorithm; strategies with very small regret can lead frequently to ruin. In fact, reducing the probability of ruin may imply an increasing regret, especially when considering a finite time-horizon. Thus, some other criterion must be also taken into account for optimizing survival. From a theoretical point of view, solving a S-MAB can be defined as the problem of finding an optimal sequence of actions $A_h^* = \{a_1, ..., a_h\}$ that minimize both the expected reward regret and the probability of being ruined, which can be formally stated as a multiobjective optimization: $A_h^* = \arg\min_{A_h}[\lambda_h, P_h^\dagger]$, where λ_h is the theoretical cumulative regret expected on the (poten-

tially infinite) time-horizon h and $P_h^\dagger$ is the probability of being ruined before the time-horizon h, while executing the sequence of actions A_h. Note that, such kind of optimization produces a Pareto-optimal frontier. This issue gives raise to our first question: *(a) how can the regret be meaningfully defined in formal terms for a S-MAB given its multiobjective optimization nature?*

3. Related Works: Budget and Risk

In the *budgeted multiarmed bandit* (B-MAB) setting, the player receives a reward but needs to pay a cost after pulling an arm. There are two independent functions associated to each action: *cost* and *reward*. The costs of pulling arms are taken from a given initial *budget* which limits the time-horizon of the process. No more arms can be pulled after running out of budget. When the budget is over, the game is over. The optimal arm can vary on time depending on the remaining budget. The agent needs to explore and estimate not only the reward function of an arm but also its cost function. Key results and algorithms developed for B-MAB are (Xia et al., 2017; Ding et al., 2013). In such references, the common strategy for evaluating arms considering these two different functions (rewards f_i and costs c_i) is based on a coefficient (called *density* or *reward-to-cost ratio*) which divides the expected reward $\hat{\mu}_{r_i}$ of arm i by its expected cost $\hat{\mu}_{c_i}$, i.e. $\hat{w}_{i,t} = \hat{\mu}_{r_i}/\hat{\mu}_{c_i}$.

In a similar setting called *bandits with knapsacks*, each arm is associated to d different costs that stochastically consume d different budgets (Badanidiyuru et al., 2018). Other variations of the budgeted problem considering specific scenarios can be found in (Zhou and Tomlin, 2018; Koren et al., 2017; Agrawal et al., 2016). Another version of the budgeted problem called "best arm identification" (Audibert et al., 2010) considers that pulling an arm has a cost but the budget is only imposed on a preliminary exploration phase. The subsequent exploitation phase is not associated with costs nor constrained by budgets, which constitutes a "pure exploration" problem.

Alternatively, in the *risk-averse multiarmed bandit* (RA-MAB) setting, the agent must take into account the expected variability on the expected rewards in order to identify (and avoid) less predictable (then considered risky) actions, but without worrying about ruin since no notion of budget is considered. The notion of "safety" is then understood as "stability" or "predictability" of the rewards. In this sense, the risk-reward trade-off can be tackled by using some mean-variance metric like $\hat{w}_{i,t} = \hat{\mu}_i - \varrho\hat{\sigma}_i^2$, where $\{\varrho \in \mathbb{R} \mid 0 < \varrho < \infty\}$ represents the risk aversion, $\hat{\mu}_i$ is the estimated mean reward of arm i, and $\hat{\sigma}_i^2$ is its estimated variance (Sani et al., 2012; Vakili and Zhao, 2016). Another criterion for evaluating the riskiness of an arm is its *conditional value at risk*, based on the quantiles (or lower-tails) of the reward distribution. In such proposition, the utility of an arm i is not based on its estimated mean reward, but on $\Phi_i(\alpha)$, its estimated *cumulative distribution function*, where $\{\alpha \in \mathbb{R} \mid 0 \leq \alpha \leq 1\}$ represents the risk tolerance, and indicates the reference quantile (Galichet et al., 2013; Maillard, 2013; Cassel et al., 2018). In another variation of the problem called *conservative bandits* (Wu et al., 2016), the agent also have a budget determined by the obtained rewards, but the constraint is also related to the regret. The agent must minimize regret in the long-term, but such regret must not exceed some relative amount. The idea is to limit wild exploration by defining the notion of immediate acceptable loss.

What makes S-MAB different of B-MAB is that the same single reward function impacts the budget and defines the objective (cumulated reward maximization). There is no separated cost function, and for such reason, there is no cost-reward ratio. Inversely, RA-MAB does not consider any notion of budget. The risk is understood as unpredictability, whereas in S-MAB the risk is the ruin. Such differences motivate the last two questions: *(b) can a S-MAB be reduced to a RA-MAB or*

a B-MAB, transferring their theoretical guarantees? and finally *(c) what kind of method or strategy must an agent follow to optimally solve a S-MAB?*

References

S. Agrawal, N.R. Devanur, and L. Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In 29^{th} *Conf. on Learning Theory (COLT), June 23–26, New York*, pages 4–18, 2016.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In 23^{rd} *Conf. on Learning Theory (COLT), June 27–29, Haifa*, pages 41–53, 2010.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. *J. ACM*, 65(3):13:1–13:55, 2018.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. *X*-armed bandits. *JMLR*, 12:1655–1695, 2011.

A. Cassel, S. Mannor, and A. Zeevi. A general approach to multi-armed bandits under risk criteria. In 31^{st} *Conf. on Learning Theory (COLT), July 6–9, Stockholm*, pages 1295–1306, 2018.

W. Ding, T. Qin, X.-D. Zhang, and T.-Y. Liu. Multi-armed bandit with budget constraint and variable costs. In 27^{th} *AAAI Conf. on Artificial Intelligence, July 14-18, Bellevue, WA, USA*, 2013.

N. Galichet, M. Sebag, and O. Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In 5^{th} *Asian Conf. on Mach. Learn. (ACML), Nov. 13–15, Canberra*, pages 245–260, 2013.

T. Koren, R. Livni, and Y. Mansour. Multi-armed bandits with metric movement costs. In 31^{st} *Conf. on Neural Information Processing Systems (NIPS), Dec. 4–9, Long Beach*, pages 4122–4131, 2017.

O.-A. Maillard. Robust risk-averse stochastic multi-armed bandits. In 24^{th} *Int. Conf. on Algorithmic Learning Theory (ALT), Oct. 6–9, Singapore*, pages 218–233, 2013.

A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In 26^{th} *Conf. on Neural Information Processing Systems (NIPS), Dec. 3–6, Lake Tahoe*, pages 3284–3292, 2012.

S. Vakili and Q. Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *Signal Processing*, 10(6):1093–1111, 2016.

Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvari. Conservative bandits. In 33^{rd} *Int. Conf. on Mach. Learn. (ICML), June 20–22, New York*, pages 1254–1262, 2016.

Y. Xia, T. Qin, W. Ding, H. Li, X.-D. Zhang, N. Yu, and T.-Y. Liu. Finite budget analysis of multi-armed bandit problems. *Neurocomputing*, 258:13–29, 2017.

D.P. Zhou and C.J. Tomlin. Budget-constrained multi-armed bandits with multiple plays. In 32^{nd} *AAAI Conf. on Artificial Intelligence, Feb. 2–7, New Orleans*, 2018.