Gambler Bandits and the Regret of Being Ruined

Filipo Studzinski Perotto, Sattar Vakili, Pratik Gajane, Yaser Faghan, Mathieu Bourgais

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 *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.

KEYWORDS

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

ACM Reference Format:
Filipo Studzinski Perotto, Sattar Vakili, Pratik Gajane, Yaser Faghan, and Mathieu Bourgais. 2021. Gambler Bandits and the Regret of Being Ruined: Extended Abstract. In *Preprint - Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 4 pages.

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 *multiarmed 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 $B_h = b_0 + \sum_{t=1}^{h} R_t$. 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. 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:

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

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

$$\ell_{h,\pi} = \frac{\omega_{h,\pi}}{\omega_h^*} \cdot \underbrace{\sum_{i=1}^{k} \left[\frac{p^* - p_i}{p^*} \cdot \frac{\mathbb{E}[N_{i,h}]}{h} \right]}_{\text{normalized classic regret}} + \underbrace{\left(\frac{\omega_h^* - \omega_{h,\pi}}{\omega_h^*} \right)}_{\text{regret due to ruin}}, \quad (2)$$

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

Preprint - Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

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:

$$\hat{\ell}_{h,\pi} = 1 - B_{h,\pi}/B^*_h \,. \tag{3}$$

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 `ε-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 `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.

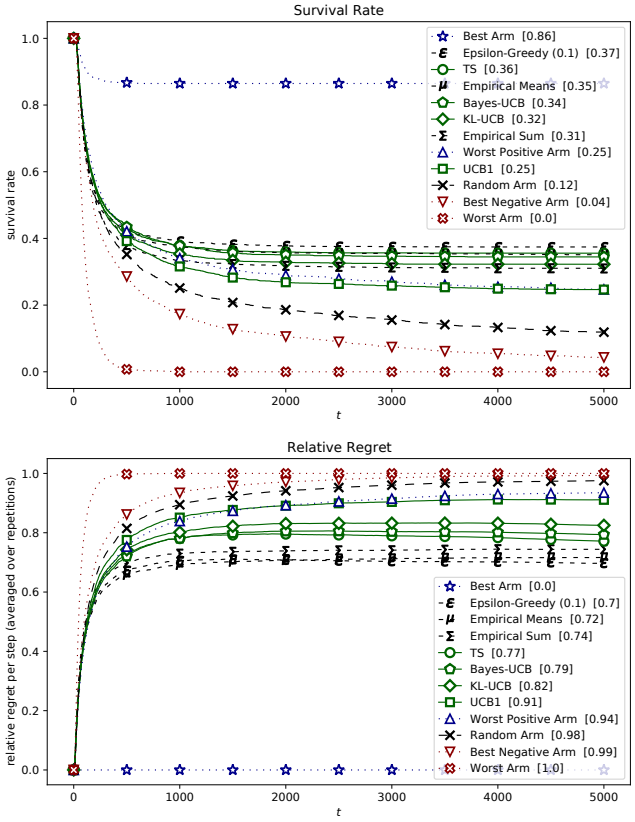

Figure 1: Survival rates and average empirical relative normalized regrets, $n = 2000$ episodes, time-horizon $h = 5000$.

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*.

REFERENCES

[1] Rajeev Agrawal. 1995. Sample Mean Based Index Policies with $O(\log n)$ Regret for the Multi-Armed Bandit Problem. *Advances in Applied Probability* 27, 4 (1995), 1054–1078.

[2] Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. 2016. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Proceedings of the 29th Annual Conference on Learning Theory, COLT 2016* (New York, USA, 2016, June 23-26), Vol. 49. PMLR, New York, USA, 4–18.

[3] Shipra Agrawal and Navin Goyal. 2017. Near-Optimal Regret Bounds for Thompson Sampling. *J. ACM* 64, 5 (2017), 24.

[4] Kook Jin Ahn and Sudipto Guha. 2009. Graph Sparsification in the Semi-streaming Model. In *Proceedings of the 36th International Colloquium on Automata, Languages and Programming, ICALP 2009, Part II* (Rhodes, Greece, July 5-12, 2009) *(Lecture Notes in Computer Science, Vol. 5556)*. Springer, Berlin, Heidelberg, 328–338.

[5] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe Reinforcement Learning via Shielding. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI-18* (New Orleans, Louisiana, USA, February 2-7, 2018). AAAI Press, Palo Alto, California, USA, 2669–2678.

[6] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. 2010. Best Arm Identification in Multi-Armed Bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory, COLT 2010* (Haifa, Israel, June 27-29, 2010). Omnipress, 41–53.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47, 2-3 (2002), 235–256.

[8] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. 2018. Bandits with Knapsacks. *J. ACM* 65, 3 (2018), 13:1–13:55.

[9] F. Berkenkamp, M. Turchetta, A.P. Schoellig, and A. Krause. 2017. Safe Model-based Reinforcement Learning with Stability Guarantees. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Long Beach, California, USA, December 4-9, 2017) *(Advances in Neural Information Processing Systems, Vol. 30)*. Curran, Red Hook, NY, USA, 908–918.

[10] Sébastien Bubeck and Nicolò Cesa-Bianchi. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* 5, 1 (2012), 1–122.

[11] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. 2009. Pure Exploration in Multi-armed Bandits Problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory, ALT 2009* (Porto, Portugal, October 3-5, 2009) *(Lecture Notes in Artificial Intelligence, Vol. 5809)*. Springer, Berlin, Heidelberg, 23–37.

[12] S. Carpin, Y. Chow, and M. Pavone. 2016. Risk aversion in finite Markov Decision Processes using total cost criteria and average value at risk. In *Proceedings of ICRA*. IEEE, 335–342.

[13] Asaf Cassel, Shie Mannor, and Assaf Zeevi. 2018. A General Approach to Multi-Armed Bandits Under Risk Criteria. In *Proceedings of the 31st Annual Conference On Learning Theory, COLT 2018* (Stockholm, Sweden, July 6-9, 2018), Vol. 75. PMLR, Stockholm, Sweden, 1295–1306.

[14] R. Cheng, G. Orosz, R.M. Murray, and J.W. Burdick. 2019. End-to-End Safe Reinforcement Learning through Barrier Functions for Safety-Critical Continuous Control Tasks. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI-19* (Honolulu, Hawaii, USA, January 27 – February 1, 2019). AAAI Press, Palo Alto, California, USA, 3387–3395.

[15] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. 2018. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research* 18, 167 (2018), 1–51.

[16] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh. 2018. A Lyapunov-Based Approach to Safe Reinforcement Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NeurIPS'18* (Montréal, Canada) *(Advances in Neural Information Processing Systems, Vol. 31)*. Curran, Red Hook, NY, USA, 8103–8112.

[17] Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. 2014. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing, STOC'14* (New York, NY, USA, May 31 - June 03, 2014). ACM, 459–467.

[18] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. 2013. Multi-Armed Bandit with Budget Constraint and Variable Costs. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI-2013* (Bellevue, Washington, USA, July 14-18, 2013). AAAI Press, Palo Alto, California, USA, 232–238.

[19] Y. Efroni, S. Mannor, and M. Pirotta. 2020. Exploration-Exploitation in Constrained MDPs. *ArXiv* abs/2003.02189 (2020).

[20] Stewart N. Ethier. 2010. Gambler's Ruin. In *The Doctrine of Chances*. Springer, Berlin, Heidelberg, 241–274.

[21] W. Feller. 1966. *An Introduction to Probability Theory and Its Applications*. Wiley, New York, NY.

[22] Nicolas Galichet, Michèle Sebag, and Olivier Teytaud. 2013. Exploration vs Exploitation vs Safety: Risk-Aware Multi-Armed Bandits. In *Proceedings of the 5th Asian Conference on Machine Learning, ACML* (Canberra, Australia, November 13-15, 2013) *(Proceedings of Machine Learning Research, Vol. 29)*. PMLR, Canberra, Australia, 245–260.

[23] Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. 2020. Conservative Exploration in Reinforcement Learning. *CoRR* abs/2002.03218 (2020). arXiv:2002.03218

[24] Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. 2020. Improved Algorithms for Conservative Exploration in Bandits. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI-20*. AAAI Press, Palo Alto, California, USA, 3962–3969.

[25] Javier García and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16 (2015), 1437–1480.

[26] J. García and D. Shafie. 2020. Teaching a humanoid robot to walk faster through Safe Reinforcement Learning. *Engineering Applications of Artificial Intelligence* 88 (2020).

[27] Aurélien Garivier and Olivier Cappé. 2011. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Proceedings of the 24th Annual Conference on Learning Theory, COLT 2011* (Budapest, Hungary, June 9-11, 2011), Vol. 19. PMLR, Budapest, Hungary, 359–376.

[28] Aurélien Garivier, Hédi Hadiji, Pierre Ménard, and Gilles Stoltz. 2018. KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *CoRR* abs/1805.05071 (2018). arXiv:1805.05071

[29] Prakash Gorroochurn. 2012. Huygens and the Gambler's Ruin (1657). In *Classic Problems of Probability*. Wiley, Chapter 5, 39–48.

[30] Sudipto Guha and Kamesh Munagala. 2007. Approximation algorithms for budgeted learning problems. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, STOC'07* (San Diego, California, USA, June 11-13, 2007). ACM, 104–113.

[31] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2012. On Bayesian Upper Confidence Bounds for Bandit Problems. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics, AISTATS 2012* (La Palma, Canary Islands, Spain, April 21-23, 2012) *(Proceedings of Machine Learning Research, Vol. 22)*. PMLR, La Palma, Canary Islands, Spain, 592–600.

[32] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. 2012. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory, ALT 2012* (Lyon, France, October 29-31, 2012) *(Lecture Notes in Artificial Intelligence, Vol. 7568)*. Springer, Berlin, Heidelberg, 199–213. https://doi.org/10.1007/978-3-642-34106-9_18

[33] Nathaniel Korda, Emilie Kaufmann, and Remi Munos. 2013. Thompson Sampling for 1-Dimensional Exponential Family Bandits. In *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13* (Lake Tahoe, Nevada) *(Advances in Neural Information Processing Systems, Vol. 25)*. Curran, Red Hook, NY, USA, 1448–1456.

[34] Tomer Koren, Roi Livni, and Yishay Mansour. 2017. Bandits with Movement Costs and Adaptive Pricing. In *Proceedings of the 30th Annual Conference on Learning Theory, COLT 2017* (Amsterdam, The Netherlands, July 7-10, 2017), Vol. 65. PMLR, Amsterdam, The Netherlands, 1242–1268.

[35] Tomer Koren, Roi Livni, and Yishay Mansour. 2017. Multi-Armed Bandits with Metric Movement Costs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Long Beach, California, USA, December 4-9, 2017) *(Advances in Neural Information Processing Systems, Vol. 30)*. Curran, Red Hook, NY, USA, 4122–4131.

[36] T.L Lai and Herbert Robbins. 1985. Asymptotically Efficient Adaptive Allocation Rules. *Adv. Appl. Math.* 6, 1 (1985), 4–22.

[37] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

[38] A. Luedtke, E. Kaufmann, and A. Chambaz. 2019. Asymptotically optimal algorithms for budgeted multiple play bandits. *Machine Learning* 108 (2019), 1919–1949. https://doi.org/10.1007/s10994-019-05799-x

[39] Omid Madani, Daniel J. Lizotte, and Russell Greiner. 2004. The Budgeted Multi-armed Bandit Problem. In *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004* (Banff, Canada, July 1-4, 2004) *(Lecture Notes in Computer Science, Vol. 3120)*. Springer, Berlin, Heidelberg, 643–645.

[40] Odalric-Ambrym Maillard. 2013. Robust Risk-Averse Stochastic Multi-armed Bandits. In *Proceedings of the 24th International Conference on Algorithmic Learning Theory, ALT 2013* (Singapore, October 6-9, October 6-9, 2013) *(Lecture Notes in Artificial Intelligence, Vol. 8139)*. Springer, Berlin, Heidelberg, 218–233.

[41] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. 2011. A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In *Proceedings of the 24th Annual Conference on Learning Theory, COLT 2011* (Budapest, Hungary, June 9-11, 2011), Vol. 19. PMLR, Budapest, Hungary, 497–514.

[42] A. Majumdar and M. Pavone. 2020. How Should a Robot Assess Risk? Towards an Axiomatic Theory of Risk in Robotics. *Robotics Research* 10 (2020), 75–84.

[43] Filipo Studzinski Perotto. 2015. Looking for the Right Time to Shift Strategy in the Exploration-exploitation Dilemma. *Schedae Informaticae* 24 (2015), 73–82.

[44] Filipo Studzinski Perotto, Mathieu Bourgais, Bruno Castro da Silva, and Laurent Vercouter. 2019. Open Problem: Risk of Ruin in Multiarmed Bandits. In *Proceedings*

of the 32nd Annual Conference on Learning Theory, COLT 2019 (Phoenix, AZ, USA, June 25-28, 2019) *(Proceedings of Machine Learning Research, Vol. 99)*. PMLR, Phoenix, AZ, USA, 3194–3197.

[45] Amir Sani, Alessandro Lazaric, and Rémi Munos. 2012. Risk-Aversion in Multi-armed Bandits. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems, NIPS'12* (Lake Tahoe, Nevada, USA, December 3-6, 2012) *(Advances in Neural Information Processing Systems, Vol. 25)*. Curran, Red Hook, NY, USA, 3284–3292.

[46] Aleksandrs Slivkins. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends in Machine Learning* 12, 1-2 (2019), 1–286.

[47] Seongjoo Song and Jongwoo Song. 2013. A note on the history of the gambler's ruin problem. *Communications for Statistical Applications and Methods* 20, 2 (2013), 157–168.

[48] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (2 ed.). MIT Press.

[49] Long Tran-Thanh, Archie C. Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas R. Jennings. 2010. Epsilon-First Policies for Budget-Limited Multi-Armed Bandits. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence, AAAI-2010* (Atlanta, Georgia, USA, July 11-15, 2010). AAAI Press, Palo Alto, California, USA, 1211–1216.

[50] Long Tran-Thanh, Archie C. Chapman, Alex Rogers, and Nicholas R. Jennings. 2012. Knapsack Based Optimal Policies for Budget-Limited Multi-Armed Bandits. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence, AAAI-2012* (Toronto, Ontario, Canada, July 22-26, 2012). AAAI Press, Palo Alto, California, USA, 1134–1140.

[51] Sattar Vakili, A. Boukouvalas, and Q. Zhao. 2019. Decision Variance in Risk-Averse Online Learning. In *Proceedings of the 58th Conference on Decision and Control, CDC*. IEEE, 2738–2744. https://doi.org/10.1109/CDC40024.2019.9029461

[52] Sattar Vakili and Qing Zhao. 2016. Risk-Averse Multi-Armed Bandit Problems Under Mean-Variance Measure. *IEEE Journal of Selected Topics in Signal Processing* 10, 6 (2016), 1093–1111.

[53] Joannès Vermorel and Mehryar Mohri. 2005. Multi-armed Bandit Algorithms and Empirical Evaluation. In *Proceedings of the 16th European Conference on Machine Learning, ECML 2005* (Porto, Portugal, October 3-7, 2005) *(Lecture Notes in Computer Science, Vol. 3720)*. Springer, Berlin, Heidelberg, 437–448.

[54] D. Wu, X. Chen, X. Yang, H. Wang, Q. Tan, X. Zhang, J. Xu, and K. Gai. 2018. Budget Constrained Bidding by Model-Free Reinforcement Learning in Display Advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM'18* (Turin, Italy, October 22-26, 2018). ACM, New York, USA, 1443–1451.

[55] Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvari. 2016. Conservative Bandits. In *Proceedings of the 33rd International Conference on Machine Learning, ICML'16* (New York, USA, June 20-22, 2016) *(Proceedings of Machine Learning Research, Vol. 48)*. PMLR, New York, USA, 1254–1262.

[56] Yingce Xia, Tao Qin, Wenkui Ding, Haifang Li, Xu-Dong Zhang, Nenghai Yu, and Tie-Yan Liu. 2017. Finite budget analysis of multi-armed bandit problems. *Neurocomputing* 258 (2017), 13–29.

[57] Datong P. Zhou and Claire J. Tomlin. 2018. Budget-Constrained Multi-Armed Bandits With Multiple Plays. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI-18* (New Orleans, Louisiana, USA, February 2-7, 2018). AAAI Press, Palo Alto, California, USA, 4572–4579.

[58] Qiuyu Zhu and Vincent Tan. 2020. Thompson Sampling Algorithms for Mean-Variance Bandits. In *Proceedings of the 37th International Conference on Machine Learning, ICML* (virtual, July 13–18, 2020) *(Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 11599–11608.