Structure and Sparsity of Stochastic Multi-Arm Bandits

Lilian Besson Lilian.Besson@CentraleSupelec.fr
CentraleSupélec (campus of Rennes), IETR, SCEE Team,
Avenue de la Boulaie – CS 47601, F-35576 Cesson-Sévigné, France

Emilie Kaufmann emilie.kaufmann@univ-lille1.fr
CNRS & Université de Lille, Inria SequeL team
UMR 9189 – CRIStAL, F-59000 Lille, France

Abstract

TODO

Keywords: Multi-Armed Bandits; Reinforcement Learning; Online Learning; TODO

Contents

1. Introduction

Multi-Armed Bandit (MAB) problems are well-studied sequential decision making problems in which an agent repeatedly chooses an action (the "arm" of a one-armed bandit) in order to maximize some total reward (Robbins, 1952; Lai and Robbins, 1985). Initial motivation for their study came from the modeling of clinical trials, as early as 1933 with the seminal work of Thompson (1933). In this example, arms correspond to different treatments with unknown, random effect. Since then, MAB models have been proved useful for many more applications, that range from cognitive radio (Jouini et al., 2009) to online content optimization (e.g., news article recommendation (Li et al., 2010), online advertising (Chapelle and Li, 2011), A/B

Testing (Kaufmann et al., 2014; Yang et al., 2017)), or portfolio optimization (Sani et al., 2012).

While the number of patients involved in a clinical study (and thus the number of treatments to select) is often decided in advance, in other contexts the total number of decisions to make (the horizon T) is unknown. It may correspond to the total number of visitors of a website optimizing its displays for a certain period of time, or to the number of attempted communications in a smart radio device. In such cases, it is thus crucial to devise anytime algorithms, that is algorithms that do no rely on the knowledge of this horizon T to sequentially select arms. A general way to turn any base algorithm into an anytime algorithm is the use of the so-called Doubling Trick, first proposed by Auer et al. (1995), that consists in repeatedly running the base algorithm with increasing horizons. Motivated by the frequent use of this technique and the absence of a generic study of its effect on the algorithm's efficiency, this paper investigates in details two families of doubling sequences (geometric and exponential), and shows that the former should be avoided for stochastic problems.

More formally, a MAB model is a set of K arms, each arm k being associated to a (unknown) reward stream $(Y_{k,t})_{t\in\mathbb{N}}$. Fix T a finite (possibly unknown) horizon. At each time step $t \in \{1,\dots,T\}$ an agent selects an arm $A(t) \in \{1,\dots,K\}$ and receives as a reward the current value of the associated reward stream, $r(t) := Y_{A(t),t}$. The agent's decision strategy (or bandit algorithm) $\mathcal{A}_T := (A(t), t \in \{1,\dots,T\})$ is such that $A(t)$ can only rely on the past observations $A(1), r(1), \dots, A(t-1), r(t-1)$, on external randomness and (possibly) on the knowledge of the horizon T. The objective of the agent is to find an algorithm $\mathcal{A}$ that maximizes the expected cumulated rewards, where the expectation is taken over the possible randomness used by the algorithm and the possible randomness in the generation of the rewards stream. In the oblivious case, in which the reward streams are independent of the algorithm's choice, this is equivalent to minimizing the regret, defined as

$$R_T(\mathcal{A}_T) := \max_{k\in\{1,\dots,K\}} \mathbb{E}\left[\sum_{t=1}^{T} \left(Y_{k,t} - Y_{A(t),t}\right)\right]. \tag{1}$$

This quantity, referred to as pseudo-regret in Bubeck et al. (2012), quantifies the difference between the expected cumulated reward of the best fixed action, and that of the strategy $\mathcal{A}_T$. For the general adversarial bandit problem (Auer et al., 2002b), in which the rewards streams are arbitrary (picked by an adversary), a worst-case lower bound has been given. It says that for every algorithm, there exists (stochastic) reward streams such that the regret is larger than $(1/20)\sqrt{KT}$ (Auer et al., 2002b). Besides, the EXP3 algorithm has been shown to have a regret of order $\sqrt{KT\log(K)}$.

Much smaller regret may be obtained in stochastic MAB models, in which the reward stream from each arm k is assumed to be i.i.d., from some (unknown) distribution ν_k, with mean μ_k. In that case, various algorithms have been proposed with problem-dependent regret upper bounds of the form $C(\boldsymbol{\nu})\log(T)$, where $C(\boldsymbol{\nu})$ is a constant that only depend on the arms distributions. Different assumptions on the

arms distributions lead to different problem-dependent constants. In particular, under some parametric assumptions (e.g., Gaussian distributions, exponential families), asymptotically optimal algorithms have been proposed and analyzed (e.g., kl-UCB (Cappé et al., 2013) or Thompson sampling (Agrawal and Goyal, 2012; Kaufmann et al., 2012)), for which the constant $C(\boldsymbol{\nu})$ obtained in the regret upper bound matches exactly that of the lower bound given by Lai and Robbins (1985). Under the non-parametric assumption that the ν_k are bounded in $[0,1]$, the regret of the UCB1 algorithm (Auer et al., 2002a) is of the above form with $C(\boldsymbol{\nu}) = 8 \times \sum_{k:\mu_k>\mu^*}(\mu^* - \mu_k)^{-1}$, where $\mu^* = \max_k \mu_k$ is the mean of the best arm. Like in this last example, all the available constants $C(\boldsymbol{\nu})$ become very large on "hard" instances, in which some arms are very close to the best arm. On such instances, $C(\boldsymbol{\nu})\log(T)$ may be much larger than the worst-case $(1/20)\sqrt{KT}$, and distribution-independent guarantees may actually be preferred.

The MOSS algorithm, proposed by Audibert and Bubeck (2009), is the first stochastic bandit algorithm to enjoy a problem-dependent logarithmic regret and to be optimal in a minimax sense, as its regret is proved to be upper bounded by $\sqrt{KT}$, for bandit models with rewards in $[0,1]$. However the corresponding constant $C(\boldsymbol{\nu})$ is proportional to $K/\Delta_{\min}$, where $\Delta_{\min} = \min_k(\mu^* - \mu_k)$ is the minimal gap, which worsen the constant of UCB1. Another drawback of MOSS is that it is not anytime. These two shortcoming have been overcame recently in two different works. On the one hand, the MOSS-anytime algorithm (Degenne and Perchet, 2016) is minimax optimal and anytime, but its problem-dependent regret does not improve that of MOSS. On the other hand, the kl-UCB^{++} algorithm (Ménard and Garivier, 2017) is simultaneously minimax optimal and asymptotically optimal (i.e., it has the best problem-dependent constant $C(\boldsymbol{\nu})$), but it is not anytime. A natural question is thus to know whether a Doubling Trick could overcome this limitation.

This question is the starting point of our comprehensive study of the Doubling Trick: can a single Doubling Trick be used to preserve both problem-dependent (logarithmic) regret and minimax (square-root) regret? We answer this question partially, by showing that two different types of Doubling Trick may actually be needed. In this paper, we investigate how algorithms enjoying regret guarantees of the generic form

$$\forall T \geq 1, \quad R_T(\mathcal{A}_T) \leq c\, T^{\gamma}(\log(T))^{\delta} + o(T^{\gamma}\left(\log(T)\right)^{\delta}) \tag{2}$$

may be turned into an anytime algorithm enjoying similar regret guarantees with an appropriate Doubling Trick. This does not come for free, and we exhibit a "price of Doubling Trick", that is a constant factor larger than 1, referred to as a constant manipulative overhead.

The rest of the paper is organized as follows. The Doubling Trick is formally defined in Section ??, along with a generic tool for its analysis. In Section ??, we present upper and lower bounds on the regret of algorithms to which a geometric Doubling Trick is applied. Section ?? investigates regret guarantees that can be obtained for

a "faster" exponential Doubling Trick. Experimental results are then reported in Section 5. Complementary elements of proofs are deferred to the appendix.

2. FIXME

TODO

3. FIXME

TODO

4. FIXME

TODO

5. Numerical Experiments

TODO

6. Conclusion

TODO

Acknowledgments

This work is supported by the French National Research Agency (ANR), under the project BADASS (grant coded: N ANR-16-CE40-0002), by the French Ministry of Higher Education and Research (MENESR) and ENS Paris-Saclay.

References

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the Multi-Armed Bandit problem. In Conference On Learning Theory. PMLR, 2012.

J-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In Conference on Learning Theory, pages 217–226. PMLR, 2009.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a Rigged Casino: The Adversarial Multi-Armed Bandit Problem. In Annual Symposium on Foundations of Computer Science, pages 322–331. IEEE, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multi-armed Bandit Problem. Machine Learning, 47(2):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The Nonstochastic Multiarmed Bandit Problem. SIAM journal on computing, 32(1):48–77, 2002b.

S. Bubeck, N. Cesa-Bianchi, et al. Regret Analysis of Stochastic and Non-Stochastic Multi-Armed Bandit Problems. Foundations and Trends® in Machine Learning, 5 (1), 2012.

O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. Annals of Statistics, 41(3):1516–1541, 2013.

O. Chapelle and L. Li. An Empirical Evaluation of Thompson Sampling. In Advances in Neural Information Processing Systems, pages 2249–2257. Curran Associates, Inc., 2011.

R. Degenne and V. Perchet. Anytime Optimal Algorithms In Stochastic Multi Armed Bandits. In International Conference on Machine Learning, pages 1587–1595, 2016.

W. Jouini, D. Ernst, C. Moy, and J. Palicot. Multi-Armed Bandit Based Policies for Cognitive Radio's Decision Making Issues. In International Conference Signals, Circuits and Systems. IEEE, 2009.

E. Kaufmann, N. Korda, and R. Munos. Thompson Sampling: an Asymptotically Optimal Finite-Time Analysis, pages 199–213. PMLR, 2012.

E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of A/B Testing. In Conference on Learning Theory, pages 461–481. PMLR, 2014.

T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. Advances in Applied Mathematics, 6(1):4–22, 1985.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In International Conference on World Wide Web, pages 661–670. ACM, 2010.

P. Ménard and A. Garivier. A Minimax and Asymptotically Optimal Algorithm for Stochastic Bandits. In Algorithmic Learning Theory, volume 76, pages 223–237. PMLR, 2017.

H. Robbins. Some Aspects of the Sequential Design of Experiments. Bulletin of the American Mathematical Society, 58(5):527–535, 1952.

A. Sani, A. Lazaric, and R. Munos. Risk-Aversion In Multi-Armed Bandits. In Advances in Neural Information Processing Systems, pages 3275–3283, 2012.

W. R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. Biometrika, 25, 1933.

F. Yang, A. Ramdas, K. Jamieson, and M. Wainwright. A framework for Multi-A(rmed)/B(andit) Testing with Online FDR Control. In Advances in Neural Information Processing Systems, pages 5957–5966. Curran Associates, Inc., 2017.

Note: the simulation code used for the experiments is using Python 3. It is open-sourced at https://GitHub.com/SMPyBandits/SMPyBandits and fully documented at
https://SMPyBandits.GitHub.io.

Appendix A. Omitted Proofs

We include here the proofs omitted in the main document.