Allocating Divisible Resources on Arms with Unknown and Random Rewards (Extended Abstract)

Wenhao Li ZJULIWENHAO@GMAIL.COM
College of Business, Shanghai University of Finance and Economics Shanghai 200433, China

Ningyuan Chen NINGYUAN.CHEN@UTORONTO.CA
Department of Management, University of Toronto Mississauga, ON L5L 1C6, Canada
Rotman School of Management, University of Toronto, ON M5S 3E6, Canada

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider a decision maker allocating one unit of renewable and divisible resource in each period on a number of arms. The arms have unknown and random rewards whose means are proportional to the allocated resource and whose variances are proportional to an order b of the allocated resource. In particular, if the decision maker allocates resource $A_{ti} \in [0,1]$ to arm $i \in [K]$ in period t, then the reward Y_{ti} is $Y_{ti}(A_{ti}) = A_{ti}\mu_i + A_{ti}^b \xi_{ti}$, where μ_i is the unknown mean, the noise ξ_{ti} is independent and sub-Gaussian, and b reflects the signal-to-noise (SNR) ratio. When the order b ranges from 0 to 1, the framework smoothly bridges the standard stochastic multi-armed bandit problem and online learning with full feedback.

Developing theories upon the framework, this paper makes the following contributions to the literature. First, we develop two algorithms for the problem, inspired by the design principles of successive elimination and ϵ-greedy algorithms, for the gap-independent and gap-dependent regret, respectively. We show that the algorithms attain the optimal rate of gap-independent and gap-dependent regret for $b \in (0,1)$. (See the following table for the regret rates.) The regret leads to a number of interesting findings. (1) the regret displays completely different behavior for $b \le 1/2$ and $b > 1/2$ and thus phase transition at $b = 1/2$. (2) the gap-dependent regret is $O(\log T)$ for $b \le 1/2$ and *finite* for $b > 1/2$. For the gap-independent bound, a larger $b > 1/2$ reduces the regret in terms of the order of K but not T. (3) the regret smoothly bridges that of SMAB for small SNR ($0 \le b \le 1/2$) and that of online learning with full feedback for large SNR ($b = 1$). Second, in the theoretical analysis, we establish a novel concentration inequality that bounds a linear combination of sub-Gaussian random variables whose weights are fractional, adapted to the filtration, and monotonic. The concentration result has not been discovered in the literature and could be of independent interest. [1]

Keywords: renewable and divisible resource allocation, stochastic multi-armed bandit, gap-dependent (independent) regret

	Gap-independent	Gap-dependent
SMAB ($b = 0$) (Auer et al., 2002)	$O\left(\sqrt{TK}\right)$	$O\left(\log T \sum_i \Delta_i^{-1}\right)$
$b \in (0, 1/2]$ (this work)	$O\left(\sqrt{TK}\right)$	$O\left(\log T \sum_i \Delta_i^{-1}\right)$
$b \in (1/2, 1)$ (this work)	$O\left(\sqrt{T}K^{1-b}\right)$	$O(1)$
Full feedback ($b = 1$) (Degenne and Perchet, 2016)	$O\left(\sqrt{T \log K}\right)$	$O(1)$

1. Extended abstract. Full version appears as [arXiv:2306.16578, v1]

Acknowledgments

We thank Jinhui Han, Zhenghang Xu, Aiqi Zhang, Guan Wang and Jialin Li for helpful discussions.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1587–1595. PMLR, 2016.