Open Problem: Finite-Time Instance Dependent Optimality for Stochastic Online Learning with Feedback Graphs

Teodor V. Marinov TVMARINOV@GOOGLE.COM
Google Research

Mehryar Mohri MOHRI@GOOGLE.COM
Courant Institute and Google Research

Julian Zimmert ZIMMERT@GOOGLE.COM
Google Research

Abstract

Both asymptotic and non-asymptotic instance-dependent regret bounds are known for the stochastic multi-armed bandit problem. Such regret bounds are known to be tight up to lower order terms in the setting of Gaussian rewards (Garivier et al., 2019). We revisit the related problem of stochastic online learning with feedback graphs, where asymptotically optimal instance dependent algorithms are known. Surprisingly, the notion of optimal finite-time regret is not a uniquely defined property in this context and in general, it is decoupled from the asymptotic rate. We pose two open problems. First we ask for a characterization of the finite-time instance-dependent optimal regret. Next, we ask for a characterization of the set of graphs for which the finite time regret is bounded by the asymptotically optimal rate, for reasonable values of the time horizon.

1. Introduction

The problem of stochastic online learning with feedback graphs is modeled as a sequential game. At every step of the game, the player selects an action/arm from a finite action set $[K]$ and then receives feedback. Each arm $i \in [K]$ admits a reward distribution with mean μ_i. The feedback consists of a random sample from the reward distributions of a subset of $[K]$. This feedback is specified by a graph $G = (V, E)$ with vertex set $V = [K]$, where an edge $(i, j) \in E$ indicates that the player observed the reward of arm j when selecting arm i. The goal of the player is to maximize their expected cumulative gain after T rounds of the game. The learning scenario above, introduced by (Mannor and Shamir, 2011), interpolates between the two classical settings of online learning with expert advice (Littlestone and Warmuth, 1994; Freund and Schapire, 1997), and the bandit setting (Lai et al., 1985; Auer et al., 2002a,b). The expert (full information) setting corresponds to an undirected complete feedback graph with self-loops, and the bandit setting corresponds to a feedback graph with edge set consisting only of self-loops.

The literature for both the full information and bandit settings is very rich. In the full information setting, it is possible to obtain *instance dependent* regret bounds, which only depend on the reward distribution of the arms (Mourtada and Gaïffas, 2019) and have no dependence on the time-horizon T. In the bandit setting it is possible to obtain similar regret bounds with only logarithmic dependence on T (Auer et al., 2002a). Further, tight asymptotic and non-asymptotic lower bounds are known for both settings (Mourtada and Gaïffas, 2019; Garivier et al., 2019) in the case when the reward distributions are Gaussian with constant variance. In both settings the algorithms also enjoy good finite time guarantees, matching the known lower bounds up to constant factors.

In the feedback graphs setting, algorithms which enjoy instance dependent bounds are also known Caron et al. (2012); Buccapatnam et al. (2014); Wu et al. (2015); Cohen et al. (2016); Buccapatnam et al. (2017); Lykouris et al. (2020); Li et al. (2020). Such bounds are more favorable than the optimal bounds in the bandit setting, and depend on the structure of the feedback graph. Furthermore, Wu et al. (2015); Li et al. (2020) propose algorithms which enjoy asymptotically optimal regret bounds. However, unlike in the bandit setting, the finite time component of these bounds might dominate the time-dependent component for exponentially many steps in K (Marinov et al., 2022). Surprisingly, it turns out that there exists feedback graphs on which any algorithm will fail to match the asymptotically optimal rate in finite time, for a large class of problem instances. Thus, we pose the question of characterizing the finite time instance-dependent rates for the stochastic online learning problem with feedback graphs.

2. Problem setup

In this note we are interested in fixed undirected feedback graphs, with self-loops at every arm. We assume that the distribution of the reward of each arm $i \in [K]$ is Gaussian with variance $\frac{1}{\sqrt{2}}$ and mean μ_i[1]. We assume that the means are in $[0, 1]$. For each arm i we define the gap $\Delta_i = \mu^* - \mu_i$, where $\mu^* = \max_{i\in[K]} \mu_i$ is the mean of the optimal round. We also denote the minimum gap by $\Delta_{\min}$. At each round $t \in [T]$ the player selects an arm i_t and observes a sample r_{t,i_t} drawn from the distribution of arm i_t. The player also observes samples $r_{t,j}$ for all neighbors $j \in N(i_t)$ of i_t, according to the feedback graph G. The objective of the player is to minimize their cumulative pseudo-regret[2]:

$$\mathsf{Reg}(T) = \mu^* T - \mathbb{E}\left[\sum_{t=1}^{T} r_{t,i_t}\right].$$

For simplicity, we assume that there exists a unique optimal arm with index i^*. Further, we assume the informed setting, in which the player has access to the full feedback graph at the start of the game.

2.1. Prior work

Caron et al. (2012) designed a UCB-type algorithm with guarantees which depend on the most favorable *clique covering* of the graph, that is guaranteed to be at least as favorable as the optimal bandit bound of order $\Theta(\log(T)\sum_{i\neq i^*}\frac{1}{\Delta_i})$. However, the bound depends on the ratio of the maximum and minimum mean reward gaps within each clique, which can be very large. Cohen et al. (2016) presented an action-elimination style algorithm (Even-Dar et al., 2006) and showed a regret bound of order $O(\log(T)\sum_{i\in\mathcal{I}\setminus\{i^*\}}\frac{1}{\Delta_i})$, where $\mathcal{I}$ is the least favorable independent set. Later, Lykouris et al. (2018) showed that the UCB-type algorithm from (Caron et al., 2012) also enjoys a similar guarantee up to a logarithmic factor in T. Marinov et al. (2022) show very simple instances in which these bounds are suboptimal. Buccapatnam et al. (2014) derive an asymptotic lower bound

[1] We make the Gaussian assumption, because it is easier to state instance-dependent optimality in terms of the reward gaps.

[2] For the rest of the note we refer to pseudo-regret simply as regret.

for the problem which is characterized by the value of the following LP

$$c^*(\Delta, G) := \min_{x \in \mathbb{R}_+^K} \langle x, \Delta \rangle \qquad s.t. \sum_{j \in N_i} x_j \geq \frac{1}{\Delta_i^2}, \ \forall i \in [K] \setminus I^*. \tag{LP1}$$

The lower bound states that the regret of any algorithm must satisfy $\lim_{T\to\infty} \frac{\mathsf{Reg}(T)}{\log(T)} \geq c^*(\Delta, G)$. Furthermore, Buccapatnam et al. (2014, 2017) design algorithms which solve a version of the above LP, and show a regret bound which depends on the domination number of G, as long as the fraction of minimum and maximum gaps is constant. There exist instances where the proposed algorithms are sub-optimal and would incur K-times more regret than an instance optimal algorithm. Wu et al. (2015); Li et al. (2020) also design algorithms based on LP1. These algorithms indeed enjoy asymptotically tight regret bounds, however, there are instances on which the finite time component dominates the bound for reasonable values of T. More precisely, the finite part of the regret bound is of order $\Omega(K/\Delta_{\min}^2)$, which dominates the time-dependent component of order $O(\log(T)/\Delta_{\min})$ for $T \leq O(\exp(K/\Delta_{\min}))$.

3. Instance-dependent finite-time bounds

We look for instance-dependent, finite-time bounds of the following form $\mathsf{Reg}(T) \leq f(T, \Delta, G)$, where f can be decomposed into a finite-time component and the asymptotically optimal rate $c^*(\Delta, G)\log(T)$, that is $f(T) = c^*(\Delta, G)\log(T) + d(\Delta, G)$ for some function d which only depends on the instance. We note that such a d exists both for the bandit setting, as bounds with $d \leq \sum_{i \neq i^*} \frac{1}{\Delta_i}$ are readily available (Lattimore and Szepesvári, 2020), and in the full information setting, where $c^* = 0$ and $d = \Theta\left(\frac{\log(K)}{\Delta_{\min}}\right)$ (Mourtada and Gaïffas, 2019).

For a family of graphs $\mathcal{G}$ and gap-vectors $\mathcal{D}$ we pose the following open problem:

Open problem: Characterize the functions $d \in \mathbb{R}^{\mathcal{D} \times \mathcal{G}}$ for which the following holds. For any graph $G \in \mathcal{G}$, gap vector $\Delta \in \mathcal{D}$, and algorithm $\mathcal{A}$, there exists a stochastic online learning problem instance with feedback graph G and rewards with means consistent with Δ so that the regret of $\mathcal{A}$ is lower bounded by

$$\frac{\mathsf{Reg}(T)}{c^*(\Delta, G)\log(T) + \frac{1}{\Delta_{\min}}} \geq d(G, \Delta) \wedge T^{\alpha}$$

for all $\alpha > 0, T \in \mathbb{N}$.

The family of graphs, $\mathcal{G}$, we are interested in is all undirected graphs with self-loops and the family of gap-vectors is $\mathcal{D} = [0,1]^{K-1}$. We are also interested in algorithms which admit regret bounds of the order $\mathsf{Reg}(T) = O(c^* \log(T) + d(\Delta, G))$. The term T^{ϵ} is needed in the statement above due to the following. Because $d(G, \Delta)$ is independent of T, and $\mathsf{Reg}(T)$ might have polynomial behavior in T for small T (e.g. Theorem 3 (Garivier et al., 2019)), we would like to take the smaller of the polynomial behavior of $\mathsf{Reg}(T)$ and $d(G, \Delta)$ as the lower bound. For example, in the full information setting, where $c^* = 0$, the characterization of d becomes $d \leq O(\log(K))$, as for any $d > \omega(\log(K))$ we can find a T and α such that the above inequality is not satisfied.

3.1. Preliminary progress

Marinov et al. (2022) show the following result.

Theorem 1 *For any $\Delta_{\min}$ and K, there exists a graph G such that for any algorithm, there exists an instance with a unique optimal arm, such that the regret of the algorithm satisfies*

$$\frac{\mathsf{Reg}(T)}{c^* \log(T) + \frac{1}{\Delta_{\min}}} = \tilde{\Omega}\left(K^{\frac{1}{8}}\right),$$

for any T s.t. $O\left(\exp(K^{1/8})\right) \geq T \geq \Omega\left(\frac{K^{3/4}}{\Delta_{\min}}\right)$.

The above theorem shows that there exists at least one family of problem instances in which $c^*(\Delta, G)$ is dominated by $d(\Delta, G)$ for $T = O(\exp(K^{1/8}))$. The construction in the theorem is general in the sense that the smallest gap $\Delta_{\min}$ is independent of the graph G, however, G is fixed and has very specific properties. Marinov et al. (2022) also give the following characterization of d. For a fixed problem instance let μ denote the vector of rewards with i-th coordinate μ_i.

Lemma 2 *Fix any instance with rewards vector $\mu \in [0, 1/2]^K$ and arbitrary feedback graph G. Let $\Lambda_s(\mu)$ be the set of problem instances with means $\mu' \in \mu + [0, 1/2^{s-1}]^K$. Then for any algorithm and all $s \in \mathbb{R}_+$, there exists an instance in $\Lambda_s(\mu)$ such that the regret of the algorithm is bounded by the following LP*

$$\min_{x \in \mathbb{R}^{[K]}} \langle \Delta^s, x \rangle \qquad s.t. \sum_{j \in N_i} x_j \geq \frac{1}{2^{2s}},\ \forall i \in \Gamma_s ., \tag{LP2}$$

where $\Gamma_s = \{i \in [K] : \Delta_i \leq 1/2^{s-1}\}$.

The result in the above lemma can now be used by taking a maximum over all $s \in [0, \log_2(1/\epsilon)]$ to characterize finite time rates. However, this result has a min-max flavor in the sense that all possible instances which are close to μ are considered and thus does not really answer the more complicated question of a purely instance dependent quantity.

Marinov et al. (2022) also show a family of graphs, different from the bandit graph, for which $d(\Delta, G) \leq c^*(\Delta, G)$ for all gap vectors Δ. This motivates our second open problem.

Open problem: What is a necessary and sufficient condition on G for which we can guarantee $d = O(c^*)$?

References

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B Shroff. Stochastic bandits with side observations on networks. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 289–300, 2014.

Swapna Buccapatnam, Fang Liu, Atilla Eryilmaz, and Ness B Shroff. Reward maximization under uncertainty: Leveraging side-observations on networks. *arXiv preprint arXiv:1704.07943*, 2017.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. *arXiv preprint arXiv:1210.4839*, 2012.

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. In *International Conference on Machine Learning*, pages 811–819. PMLR, 2016.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, 7:1079–1105, 2006.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Shuai Li, Wei Chen, Zheng Wen, and Kwong-Sak Leung. Stochastic online learning with probabilistic graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4675–4682, 2020.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Small-loss bounds for online learning with partial information. In *Conference on Learning Theory*, pages 979–986. PMLR, 2018.

Thodoris Lykouris, Éva Tardos, and Drishti Wali. Feedback graph regret bounds for Thompson sampling and UCB. In *Algorithmic Learning Theory*, pages 592–614. PMLR, 2020.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24:684–692, 2011.

Teodor V. Marinov, Mehryar Mohri, and Julian Zimmert. Stochastic online learning with feedback graphs: Finite-time and asymptotic optimality. *arXiv preprint*, 2022.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20:1–28, 2019.

Yifan Wu, András György, and Csaba Szepesvári. Online learning with gaussian payoffs and side observations. *arXiv preprint arXiv:1510.08108*, 2015.