Stochastic Contextual Linear Bandits

Problem Formulation

Consider a learning agent which, in each round $t \in \{1, 2, ..., T\}$, is faced with a decision set of arm vectors $D_t \subset R^d$ with $|D_t| = K$ for $K \geq 1$ from which the agent must choose an arm vector $x_t \in D_t$. Along with being associated with an action taken by the agent, the arm vector, also referred to as a feature vector, contains a typically time-varying "context", which can carry information about the action taken, or the surrounding environment in which the agent and its actions take place, or both. The agent then receives a reward

$$y_t = \langle x_t, \theta^* \rangle + \eta_t$$

where θ^* is some unknown but fixed parameter the agent would like to learn and η_t is random noise sampled i.i.d for every round with expected value zero. The agent's goal is to maximize its cumulative reward $\sum_{t=1}^{T} y_t$. To evaluate the performance of a learning algorithm, we can contrast the algorithm's performance with the performance of an optimal strategy that knows θ^*, and thus in every round is able to choose the optimal arm $x_t^* = \text{argmax}_{x \in D_t} \langle x, \theta^* \rangle$ that yields greatest expected reward, and receives optimal reward y_t^*. We can express this contrast in performance as the *regret* of the algorithm:

$$\hat{R}_T = \sum_{t=1}^{T} y_t^* - \sum_{t=1}^{T} y_t$$

Instead of the regret of an algorithm, it is often more useful to analyze the *pseudoregret* of an algorithm:

$$R_T = \left(\sum_{t=1}^{T} \langle x_t^*, \theta^* \rangle \right) - \left(\sum_{t=1}^{T} \langle x_t, \theta^* \rangle \right) = \sum_{t=1}^{T} \langle x_t^* - x_t, \theta^* \rangle$$

This is because the regret and pseudoregret have the same expected value, but the pseudoregret has lower variance and can be easier to work with since no random noise is involved. We can now rephrase the goal of the learning agent as minimizing its regret or pseudoregret.

References

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011, pp. 2312–2320.

[2] Peter Auer. "Using confidence bounds for exploitation-exploration trade-offs". In: *Journal of Machine Learning Research* 3 (2002), pp. 397–422.

[3] Abhimanyu Dubey and Alex Pentland. *Differentially-Private Federated Linear Bandits*. 2020. arXiv: `2010.11425 [cs.LG]`.

Nonstationary Bandits

Problem Formulation: Nonstochastic (Adversarial) Bandits

Consider a learning agent, which, in each round $t \in \{1, 2, ..., T\}$, is faced with a set of arms $[K] = \{1, 2, ..., K\}$ from which the agent chooses an arm $i(t) \in [K]$. Ahead of time, the environment, also referred to as an "oblivious" adversary[3], has already assigned a sequence of rewards $(\boldsymbol{r(1)}, ..., \boldsymbol{r(T)})$, where $\boldsymbol{r(t)}$ is a vector of predetermined rewards $(r_1(t), r_2(t), ..., r_K(t))$ for each arm at time t. One can also consider a situation where the sequence of rewards is not predetermined, but instead chosen in real time by a "reactive" adversary" [2] that has access to the agent's arm pick history. After choosing an arm, the agent receives a reward $r_{i(t)}(t)$ that exists as an entry in the vector $\boldsymbol{r(t)}$. The agent's goal is to maximize its cumulative reward

$\sum_{t=1}^{T} r_{i(t)}(t)$. To evaluate the agent's performance, we usually speak in terms of *regret* instead of cumulative reward. In the context of nonstochastic bandits, there are two main notions of regret we consider: *weak regret* and *worst-case regret*, both discussed in [1]. The weak regret of an agent is given by the difference between the cumulative reward of the single best arm in hindsight and the cumulative reward of the agent:

$$R_w = \max_{j \in [K]} \left(\sum_{t=1}^{T} r_j(t) \right) - \sum_{t=1}^{T} r_{i(t)}(t)$$

The worst-case regret of an agent is a very general notion of regret expressed as the difference between the cumulative reward of some policy that plays a given sequence of arms $(j_1, ..., j_T)$ and the cumulative reward of our agent:

$$R_W = \sum_{t=1}^{T} r_{j_t}(t) - \sum_{t=1}^{T} r_{i(t)}(t)$$

If the algorithm used by the agent is randomized, we can also discuss expected worst-case regret:

$$\mathbb{E}[R_W] = \sum_{t=1}^{T} r_{j_t}(t) - \mathbb{E}\left[\sum_{t=1}^{T} r_{i(t)}(t) \right]$$

We can now rephrase the agent's goal as minimizing any of the above notions of regret.

Algorithms: Nonstochastic Bandits

When developing an algorithm for an adversarial bandit setting, it is important to notice that in order to achieve sublinear regret in terms of T, the learning algorithm must incorporate randomness into its choice of arms. This is because given a deterministic algorithm, it is possible for an adversary (reactive or oblivious), to choose a sequence of rewards that results in linear regret for the agent every time the algorithm is run [2]. On the other hand, blindly picking arms at random from a uniform distribution will also not yield good results. The answer given by [1] is to construct a probability distribution over all the arms that the algorithm samples from. The algorithm then uses its reward history for an arm to continually adjust, or "weight", the probability the arm is chosen according to how profitable the arm has been up to that point. The most famous algorithm to use this approach is the "Exp3" algorithm introduced by [1].

References

[1] Peter Auer et al. "The Non-Stochastic Multi-Armed Bandit Problem". In: *SIAM J. Comput.* 32 (Jan. 2002), pp. 48–77. DOI: 10.1137/S0097539701398375.

[2] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. DOI: 10.1017/9781108571401.

[3] Aleksandrs Slivkins. *Introduction to Multi-Armed Bandits*. 2019. arXiv: 1904.07272 [cs.LG].

Cooperative Nonstochastic Bandits

Problem Formulation

Consider N learning agents facing a nonstochastic multi-armed bandit problem that communicate over an undirected connected graph $G = (V, E)$ where $V = \{1, 2, ..., N\}$ is the set of vertices representing agents, and E is the set of edges representing communication links between agents. Let $\mathcal{N}(v)$ denote the neighborhood of an agent $v \in V$, including itself. More formally,

$$\mathcal{N}(v) = \{u \in V : (u, v) \in E\} \cup \{v\}$$

At each time step $t \in \{1, 2, ..., T\}$, the agents are faced with a common, fixed set of arms $[K] = \{1, 2, ..., K\}$, from which each agent must choose an arm and receive a loss. The arm chosen by agent $v \in V$ at time t is denoted by $I_t(v) \in [K]$, and the corresponding loss observed by the agent is denoted by $\ell_t(I_t(v))$. Because the agents face a nonstochastic bandit problem where losses can be thought of as being set ahead of time by an oblivious adversary, we can assume that for any serious algorithm there must be randomization in the agent's choice of arms [1, 4]. Specifically, we can assume that $I_t(v)$ is chosen from a distribution

$$\boldsymbol{p}_t^v = (p_t^v(1), p_t^v(2), ..., p_t^v(K))$$

where $p_t^v(k)$ is the probability of agent v choosing arm $k \in [K]$ at time t. How these probabilities are set and updated is up to the chosen algorithm, but usually some kind of exponential weighting scheme is used [1, 4]. At the end of a time step t, each agent v sends a message $m_t(v)$ to every agent in $\mathcal{N}(v)$ of the form

$$m_t(v) = (v, t, I_t(v), \ell_t(I_t(v)), \boldsymbol{p}_t^v)$$

and also receives messages from all its neighbors $m_t(v')$ for all $v' \in \mathcal{N}(v)$. The goal of the agents can be stated as minimizing their cumulative losses, but as usual it is more useful to think in terms of regret. Here we consider two different notions of regret. The first is the *average welfare regret* defined in [3] as

$$R_t^{coop} = \left(\frac{1}{N} \sum_{v \in V} \mathbb{E} \left[\sum_{t=1}^{T} \ell_t(I_t(v)) \right] - \min_{i \in [K]} \left(\sum_{t=1}^{T} \ell_t(i) \right) \right)$$

The second is the *individual regret* of an agent v, defined in [2] as

$$R_t(v) = \mathbb{E} \left[\sum_{t=1}^{T} \ell_t(I_t(v)) \right] - \min_{i \in [K]} \left(\sum_{t=1}^{T} \ell_t(i) \right)$$

Both notions of regret compare the performance of agents or agent to the performance of the single best arm in hindsight, and the expectation in both definitions is taken with respect to the randomization of the algorithm run by the agents. In [3] Cesa-Bianchi, et al. gave bounds for the average welfare regret in a setting with communication delays, but left bounds for the individual regret as an open problem. In [2] Baron and Mansour then provided bounds for the individual regret, but in a setting without delays identical to the one outlined here.

References

[1] Peter Auer et al. "The Non-Stochastic Multi-Armed Bandit Problem". In: *SIAM J. Comput.* 32 (Jan. 2002), pp. 48–77. DOI: 10.1137/S0097539701398375.

[2] Yogev Bar-On and Yishay Mansour. *Individual Regret in Cooperative Nonstochastic Multi-Armed Bandits*. 2019. arXiv: 1907.03346 [cs.LG].

[3] Nicolo' Cesa-Bianchi et al. *Delay and Cooperation in Nonstochastic Bandits*. 2016. arXiv: 1602.04741 [cs.LG].

[4] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. DOI: 10.1017/9781108571401.