Bandit on graphs : a bibliographic note

May 28, 2015

We review here a few models and algorithms presented in the literature that could be useful for tasks related to recommendation (or queries) in social networks.

Three kinds of bandit models on graphs appear in the literature:

- **bandits with side observations:** arms are nodes on a graph, after choosing an arm and receiving a reward, we also observe ("side-observation") the rewards of the neighbors.
- **spectral bandits:** arms are nodes on a graph, and the mean reward function is assumed to be smooth on the graph (e.g. does not change a lot from a node to its neighbor). No side observation.
- **communicating linear bandits:** ("gang of bandits") in this contextual setting each node (user in a social network) hosts a linear bandit algorithm, but the regression parameter of neighbors are assumed to be close, and the different algorithms can exchange information.

1 Bandit with side observation

n arms that are nodes on a graph $G = (V, E)$ $(V = \{1, \dots, n\})$. At time t, an agent :

- chooses a node A_t
- receives a reward $r_t = g_{A_t,t}$ where $g_{a,t}$ is the payoff of arm a at time t
- observes $(g_{a,t})$ for all a such that $(a, A_t) \in E$ (neighbors of the chosen node)

The observed extra rewards are not ”won”, since the regret is still

$$R_T = \max_a \mathbb{E}[\sum_{t=1}^{T} g_{a,t}] - \mathbb{E}[\sum_{t=1}^{T} g_{A_t,t}]$$

Algorithms have been proposed for this minimizing regret in the adversarial and stochastic setting. From far, they consist in adapting usual algorithms (EXP3, ϵ-greedy, UCB) by constraining exploration to well-chosen small set of arms that will allow us to observe all the nodes (e.g. take one arm per clique, or sample only a dominant set).

References

- Caron, Kveton, Lelarge, Bhagat, *Levaraging Side Observations in Stochastic Bandits*,
  UAI 2012 (stochastic)
- Buccapatnam, Eryilmaz, Shroff, *Stochastic Bandits with Side Observations on Networks*,
  SIGMETRICS 2014 (stochastic)
- Mannor, Shamir, *From Bandit to Expert : On the Value of Side-Observation*,
  NIPS 2011 (adversarial)
- Alon, Cesa-Bianchi, Gentile, Mannor, Mansour, Shamir *Nonstochastic Multi-Armed Bandits with Graph-Structured Feedback*, arxiv Preprint, 2014 (adversarial)

2 Spectral bandits

The setting is the same as before, expect that we do not receive side observation. Additionally, there is a stochastic assumption : $(g_{a,t})_{t\in\mathbb{N}}$ are i.i.d. and satisfy

$$\mathbb{E}[g_{a,t}] = f(a),$$

with f a function that is smooth on the graph $G = (V, E)$ formed by the nodes. From the theory of smooth graph function, it is known that f has a parametric form, $f = f_\alpha$ with $\alpha \in \mathbb{R}^n$ (n is the number of nodes) and

$$f_\alpha(a) = \langle X_{a,\cdot} | \alpha \rangle,$$

where $X_{a,\cdot}$ is the a-th row of the matrix $X \in \mathbb{R}^{n\times n}$ of eigenvectors of the Laplacian of G, sorted in increasing order of the associated eigenvalues.

The problem then becomes a linear bandit problems, with n arms and feature vectors that are of dimension n. Optimistic algorithms (Lin-UCB like) and Thompson Sampling are analyzed and the regret bound are shown to depend on some "effective dimension" rather than on n.

References

- Valko, Munos, Kveton, Kocak, *Spectral Bandits for Smooth Graph Functions*, ICML 2014
- Kocak, Valko, Munos, Agrawal, *Spectral Thompson Sampling*, NIPS 2014

3 Gang of bandits

This paper appears to be the best suited to recommendation in a social network, because it incorporates a contextual aspect (a vector may describe the items that we can recommend or the answers that we can give to a query), and uses also the graph structure on users to which a recommendation should be made. However the algorithm proposed seems not very scalable.

The model: a graph $G = (V, E)$ of n users, a regression parameter $u_i \in \mathbb{R}^d$ for each user. At time t,

- a user I_t arrives in the system
- a context set $C_t = (x_{1,t}, \cdots, x_{c_t,t})$ is build for this user (the descriptors of the items built for this specific user)
- we choose a context A_t to recommend to the user
- we observe a payoff

$$r_t = \langle u_{I_t} | x_{A_t,t} \rangle + \epsilon_t$$

The algorithm: consists in transforming the problem into a linear bandit problem with features and regression parameter in $\mathbb{R}^{d\times n}$. Moreover, the features are transformed using a specific matrix, so that information can be spread to all the graph.

Reference

- Cesa-Bianchi, Gentile, Zappella, *A gang of bandits*, NIPS 2013

4 Other papers

Jean-Michel pointed out to us this paper

- Fang, Tao, *Networked Bandits with Disjoint Linear Payoffs*, KDD 2014

The setup of this paper is similar to 1) but the side-observation becomes a ”side-reward”, that is when we draw an arm A_t, our reward is

$$\sum_{a:(a,A_t)\in E} g_{a,t}$$

(and the regret is modified similarly). It is thus a particular case of ”bandit with multiple play”, in which the set of arms that can be chosen are the neighborhoods of every node on a graph.

There is also a contextual aspect since it is assumed that $\mathbb{E}[g_{a,t}] = \langle u_a | x_{a,t} \rangle$, with $x_{a,t} \in \mathbb{R}^d$ a context associated to arm a at this moment and $u_a \in \mathbb{R}^d$ a regression parameter of arm a. The algorithm proposed is in fact a particular case of a known algorithm for multiple play (CUCB by Chen et al.) with the confidence intervals replaced by the one that are adapted for linear bandits.

Paul also pointed out this paper

- Li, Gentile, Karatzoglu, Zapella, *Online Context-Dependent Clustering in Recommendations based on Exploration-Exploitation Algorithms*, arxiv preprint 2015

that looks interesting as well for recommendation (and is also extensively based on optimistic algorithms for linear bandits...)