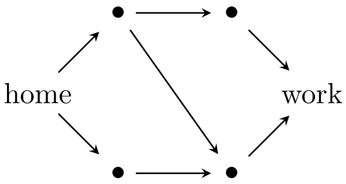
STATS 710 – Seq. Dec. Making with mHealth Applications Nov 29, 2016

Lecture 21: Partial feedback and combinatorial action space

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Yutong Wang*

1 Motivation

Suppose that the learner commutes from home to work along routes shown by the graph below.



Each arrow represents a subroute and has some amount of congestion. The learner's goal is to pick a path from home to work that has low congestion. Let $\mathcal{A}$ denote all path from home to work. We can think of this as a bandit problem where the arms are the paths $\mathcal{A}$. In this example, the rewards could be highly correlated between the arms. Thus, we may get much better regret bounds if we consider this correlation structure. Furthermore, it may be the case that the learner gets only partial feedback, e.g. only the congestion on a subroute of the path is revealed. Finally, combinatorial action spaces such as the above tend to be large, which means applying bandit algorithms out of the box could be impractical. For all the preceding reasons, the partial feedback and combinatorial action space setting deserves some special attention.

2 Online ranking with top-1 feedback

We consider the problem of online ranking of K items by relevance where we only receive feedback for the top ranked item. For a real world scenario, suppose that we own a restaurant that serves K dishes and we would like the menu to display the most relevant dishes on top. The menu is digital and can rearrange itself in real time. In order to not annoy the customers, we only collect feedback for the top menu item. The goal to learn a ranking that ranks the relevant dishes as high as possible.

Below, let K be a positive integer. Following group theory notations, let $\mathcal{A} = S_K$ be the set of permutations on $\{1, \ldots, K\}$. In other words, an element σ of S_K is a bijection from $\{1, \ldots, K\}$ to itself. For each $i = 1, \ldots, K$, we think of $\sigma(i)$ as the ranking of the i-th item (lower is better). Let T denote the time horizon. The protocol is as follows.

For $t = 1, \ldots, T$:

1. Learner takes action $\sigma_t \in S_K$
2. Nature selects $\rho_t \in \{0,1\}^K$, where $\rho_t(i) = 1$ if and only if the i-th item is relevant.
3. Learner observes $\rho_t(\sigma_t^{-1}(1))$, i.e. only the relevance of the top ranked item is revealed.
4. Learner suffers loss $\rho_t^T \sigma_t := \sum_{i=1}^K \rho_t(i)\sigma_t(i) = \sum_{i:\rho_t(i)=1} \sigma_t(i)$

The expected regret is defined as

$$\mathcal{R}(\mathcal{L}, T) \equiv \mathbb{E}\left[\sum_{t=1}^{T} \rho_t^T \sigma_t\right] - \min_{\sigma \in S_K} \sum_{t=1}^{T} \rho_t^T \sigma \tag{1}$$

Using algorithm 1 below as the learning algorithm $\mathcal{L}$, we can achieve

$$\mathcal{R}(\mathcal{L}, T) \sim O(\text{poly}(K) T^{2/3}) \tag{2}$$

which is in fact optimal and strictly harder than a bandit problem. Note that in the bandit feedback setting, i.e. we have knowledge of $\rho_t^T \sigma_T$ in step 3, then $\mathcal{R}(\text{EXP}(3), T) \sim O(\exp(K)\sqrt{T})$.

2.1 A learning algorithm

We split the T rounds into L blocks, with L to be determined later. Write $B_i = \left((i-1)\frac{T}{L} + 1, \ldots, i\frac{T}{L}\right)$ to denote the i-th block. For each i, sample K indices without replacement from B_i and denote these indices by $I_1^{(i)}, \ldots, I_K^{(i)}$. During the $I_j^{(i)}$-th round, we rank the j-th item on top and rank the rest arbitrarily, i.e. choose $\rho_{I_j^{(i)}} \in S_K$ such that $\rho_{I_j^{(i)}}(j) = 1$. Let $\hat{\rho}_i = \left(\rho_{I_1^{(i)}}(1), \ldots, \rho_{I_K^{(i)}}(K)\right)$.

Lemma 1. *$\hat{\rho}_i$ is an unbiased estimator for $\sum_{t \in B_i} \rho_t / |B_i|$.*

Here is the overall learning algorithm from [CT15]

Algorithm 1

```
1: Given parameters $T, L, \epsilon, B_i$ as above
2: Initialize $\hat{s}_0 = \vec{0}$ as the zero vector
3: for $i = 1, \ldots, L$ do
4:     $\hat{s}_{i-1} = \hat{s}_{i-2} + \hat{\rho}_{i-1}$
5:     for $t \in B_i$ do
6:         Sample $I_1^{(i)}, \ldots, I_K^{(i)}$ from $B_i$ without replacement.
7:         if $t = I_j^{(i)}$ for some $j = 1, \ldots, K$ then
8:             Output any $\sigma_t$ such that $\sigma_t(j) = 1$
9:             Receive $\rho_{I_j^{(i)}}$
10:        else
11:            Draw $Z_t \in [0, 1/\epsilon]^K$ so that the coordinates are independently and uniformly sampled from $[0, 1/\epsilon]$
12:            Output $\sigma_t = \text{argmin}_{\sigma \in S_K} \sigma^T (\hat{s}_{i-1} + Z_t)$
```

Line 12 is known as "follow the perturbed leader" (FTPL) [KV05] where Z_t is the perturbation. In the stochastic setting, it is okay to leave out Z_t, which is known as simply "follow the leader" (FTL). However, in the adversarial setting, the adversary may exploit the determinism in FTL and result expected regret linear in T.

2.2 Full information version of the ranking problem

From the perspective of the outer for loop of algorithm 1, the protocol is as follows. For $t = 1, \ldots, L$:

1. Learner takes action $\sigma_t \in S_K$
2. Nature reveals the entire vector $\rho_t \in \{0,1\}^K$
3. Learner suffers loss $\rho_t^T \sigma_t := \sum_{i=1}^K \rho_t(i)\sigma_t(i) = \sum_{i:\rho_t(i)=1} \sigma_t(i)$

The regret is still defined as in (1).

Theorem 2 ([KV05]). *Consider FTPL* $\sigma_\tau = \operatorname{argmin}_{\sigma \in S_K} \sigma^T(\sum_{s=1}^{\tau-1} \rho_s + Z_t)$ *as a learning algorithm for the protocol above. If* $\|\sigma_\tau\|_1 \le B_\sigma$*,* $|\sigma_\tau^T \rho_\tau| \le B_{loss}$ *and* $\|\rho_\tau\|_1 \le B_\rho$ *then for* $\epsilon = \sqrt{\frac{B_\sigma}{B_{loss} B_\rho L}}$*, the expected regret of FTPL is* $2\sqrt{B_\sigma B_{loss} B_\rho L}$*.*

Note that to use the theorem for the outer loop in algorithm 1, we can let $B_\sigma = B_{\text{loss}} = K^2$ and $B_\rho = K$ in the theorem. Next time, we will apply the theorem to our setting to prove (2).

References

[CT15] Sougata Chaudhuri and Ambuj Tewari. Online ranking with top-1 feedback. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *JMLR Workshop and Conference Proceedings*, pages 129–137, 2015.

[KV05] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.