CSE599s, Spring 2014, Online Learning Lecture 10 - 05/01/2014

Contextual Bandits: From EXP3 to EXP4

Lecturer: Brendan McMahan *Scribe: Tianyi Zhou*

1 Motivation

We are going to talk about a generalization of EXP3 that arises in lots of real applications. For example, let's consider about advertising, in which the player (e.g., Google) needs to choose an Ad (i.e., an arm) from a huge pool of candidates according to a specific user's query.

- In EXP3' setting, the regret reflects the difference between the accumulated loss and the loss achieved by applying the best arm to all n rounds. The latter is a benchmark we expect to get. However, there obviously does not exists one unique best Ad for all queries from all users. Thus a small regret cannot indicates benefits from learning in this case. In addition, the huge number of Ads will lead to a large regret bound. We need a more smoothing regret?
- There is a great amount of useful contextual information in the query and user's profile, which is ignored in EXP3. How to take those contexts into account?

The basic idea to generalize EXP3 to more practical applications such as advertising is to run a copy of EXP3 independently on every query, just like what we did for "weekday" and "weekend" transactions in homework. However, how to capture the similarities between different queries such as "flower" and "buy flowers"?

We can define a mapping from contexts to arms for each expert. In particular, $e_\theta : (\text{query}, \theta) \to \Delta(A)$, where A is the set of arms (e.g., Ads), $\Delta(A)$ is the multinomial distribution over arms defined by a simplex, $\theta \in H$ corresponds to an expert. This leads to the setting of "contextual bandits". The following EXP4 is a contextual bandits method extended from EXP3.

2 Setting of EXP4

More specifically,

- A actions (i.e., arms) $\{1, 2, \ldots, A\}$, $\{e_1, e_2, \ldots, e_M\}$ is the set of M experts.
- At time t, for expert i, $e_{t,i} \in \Delta(A)$ is the associated probability over actions, $e_{t,i}(a) \in [0, 1]$ denotes the probability that expert i choose action a.
- Compare them: the regret bound for choosing the best action/arm such as in EXP3 is $\mathcal{O}(\sqrt{TA \log A})$; the regret bound for choosing the best expert is $\mathcal{O}(\sqrt{TM \log M})$, we will see this is a special case for EXP4; the regret bound for full feedback information is $\mathcal{O}(\sqrt{T \log M})$, this is another special case for EXP4.
- The general regret bound for EXP4 is $\mathcal{O}(\sqrt{TS \log M})$, where $S \leq \min\{A, M\}$. In advertising we always have $M \leq A$.

We consider the expected regret for EXP4:

$$\mathbb{E}(\text{Regret}(i^*)) = \mathbb{E}\left[\sum_{t=1}^{T} l_t(a_t) - \sum_{t=1}^{T}\sum_{a=1}^{A} e_{t,i^*}(a) l_t(a)\right]. \tag{1}$$

The first term is our expected loss by applying EXP4, while the second term is the expanded loss of the best expert i^*.

3 Algorithm of EXP4

The algorithm of EXP4 invokes exponential gradient (EG) as subroutine to update probability p_t over experts. Remark: $g_{t,i}$ is an unbiased estimate to the expected loss corresponding to expert i at time t, to

Algorithm 1 EXP4 for contextual bandits

Initialize $w_1 = (1, 1, \ldots, 1)$;
for $t = 1 \to T$ **do**
EG gives us probability over experts $p_t \in \Delta(M)$: $p_t = \frac{w_t}{\|w_t\|}$;
Compute probability q_t over actions by integrating out expert i: $\forall a, q_t(a) = \sum_{i=1}^{M} p_{t,i} e_{t,i}(a)$;
Draw action $a_t \sim q_t$, and incur loss $l_t(a_t)$;
Build the unbiased estimate for full feedback l_t:

$$\forall a, \hat{l}_t(a) = \begin{cases} l_t(a_t)/q_t(a_t), & \text{if } a = a_t; \\ 0, & \text{otherwise;} \end{cases}$$

Compute the expected loss g_t:

$$\forall i, g_{t,i} = \sum_{a=1}^{A} e_{t,i}(a) l_t(a) = \frac{e_{t,i}(a_t) l_t(a_t)}{q_t(a_t)};$$

EG update w_t from g_t: $\forall i, w_{t+1,i} = w_t \exp\left(-\eta \sum_{s=1}^{t} g_{t,i}\right)$;
end for

see this, we have

$$\mathbb{E}(g_{t,i}|p_t) = \sum_{a=1}^{A} q_t(a) \left(\frac{e_{t,i}(a) l_t(a)}{q_t(a)} \right) = \sum_{a=1}^{A} e_{t,i}(a) l_t(a). \tag{2}$$

The right hand side is the expected loss $l_t(a)$ at time t (by integrating out action a).

4 Regret Bound of EXP4

Recall the expected regret bound defined in (1), we now study its upper bound.

By using the definition of p_t and g_t in Algorithm 1, the regret in (1) can be written as:

$$\mathbb{E}\left[\sum_{t=1}^{T} l_t(a_t) - \sum_{t=1}^{T} \sum_{a=1}^{A} e_{t,i^*}(a) l_t(a) \right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, g_t \rangle - g_{1:T,i^*} \right]. \tag{3}$$

From the regret bound analysis for EXP3, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, g_t \rangle - g_{1:T,i^*} \right] \leq \frac{1}{\eta} \log M + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{M} \mathbb{E}(p_{t,i} g_{t,i}^2). \tag{4}$$

In order to find the upper bound for $\sum_{i=1}^{M} \mathbb{E}(p_{t,i} g_{t,i}^2)$ in the last term, we firstly derive

$$\mathbb{E}(g_{t,i}^2|p_{t,i}) = \sum_{a=1}^{A} q_t(a) \left(\frac{e_{t,i}(a) l_t(a)}{q_t(a)} \right)^2 = G^2 \sum_{a=1}^{A} \frac{e_{t,i}^2(a)}{q_t(a)}. \tag{5}$$

Then the upper bound can be achieved by

$$\sum_{i=1}^{M} \mathbb{E}(p_{t,i} g_{t,i}^2) = \sum_{i=1}^{M} \mathbb{E}\left[p_{t,i} \mathbb{E}(g_{t,i}^2 | p_{t,i})\right] \leq G^2 \mathbb{E}\left[\sum_{a=1}^{A} \frac{\sum_{i=1}^{M} p_{t,i} e_{t,i}(a)}{q_t(a)} \cdot s_{t,a}\right] \leq G^2 \sum_{a=1}^{A} s_{t,a} \leq G^2 S, \tag{6}$$

where

$$s_{t,a} = \max_i e_{t,i}(a), S_t = \sum_{a=1}^{A} s_{t,a}, S = \max_t S_t. \tag{7}$$

The first inequality in (6) is the result of linearity of expectation operator, substituting (5) and $\forall i, e_{t,i}(a) \leq s_{t,a}$. The second inequality is due to the definition of $q_t(a)$ in Algorithm 1. The third inequality is due to (7).

Now applying (6) to the last term of (4) yields

$$\mathbb{E}(\text{Regret}(i^*)) = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, g_t \rangle - g_{1:T,i^*}\right] \leq \frac{1}{\eta} \log M + \frac{\eta}{2} G^2 T S. \tag{8}$$

Minimizing the right hand side w.r.t. learning rate η by setting the gradient to zero leads to

$$\text{Regret} \leq \mathcal{O}(\sqrt{TS \log M}). \tag{9}$$

There are three choices of the upper bound for S in the right hand side of (9). They results in three types of regret bound for EXP4.

- When all experts always agree with each other, $s_{a,t} = e_{t,i}(a)$ for arbitrary i, so $S_t = \sum_{a=1}^{A} e_{t,i}(a) = 1$ and thus $S = 1$. This results in regret bound of $\mathcal{O}(\sqrt{T \log M})$, which is identical to that in full feedback case.
- Since $s_{t,a} \leq 1$, we have $S_t \leq A$, so $S \leq A$. This results in regret bound of $\mathcal{O}(\sqrt{TA \log M})$. This is between the regret bounds for choosing the best action and choosing the best expert (recall Section 2).
- Since $s_{t,a} = \max_i e_{t,i}(a) \leq \sum_{i=1}^{M} e_{t,i}(a)$, we have $S_t \leq \sum_{a=1}^{A} s_{t,a} \leq M$, so $S \leq M$. This leads to regret bound of $\mathcal{O}(\sqrt{TM \log M})$, which is identical to the regret bound when choosing the best expert.

When $M \leq A$, EXP4's regret bound is always better than that of EXP3.