STATS 710 – Seq. Dec. Making with mHealth Applications Oct 06, 2016

Lecture 9: Proof of the Sublinear Regret of EXP3

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Julie Deeke*

1 Sublinear Regret of EXP3

1.1 Setup

Theorem 1. *EXP3 run with* $\eta = \sqrt{\frac{2\log K}{TK}}$ *enjoys* $\max_b \mathbb{E}\left[\sum_{t=1}^T l_{A_t,t} - \sum_{t=1}^T l_{b,t}\right] \leq \sqrt{2TK\log K}$.

The algorithm for EXP3 and the beginning of this proof can be found in sections 1.3 and 1.4 of Lecture 8. We will use the four useful facts and the key fact that was shown in section 1.4 in the proof below.

1.2 Continuation of Proof

From last lecture, the key idea was to write:

$$p_t^T \tilde{l}_t = \underbrace{\frac{1}{\eta}[\log p_t^T \exp(-\eta\tilde{l}_t) + \eta p_t^T \tilde{l}_t]}_{:=A} - \underbrace{\frac{1}{\eta}\log p_t^T \exp(-\eta\tilde{l}_t)}_{:=B}$$

where every function is calculated component-wise.
The first part of the expression (A) can be bounded by:

$$\begin{aligned}
\frac{1}{\eta}[\log p_t^T \exp(-\eta\tilde{l}_t) + \eta p_t^T \tilde{l}_t] &\leq \frac{1}{\eta}[p_t^T(\exp(-\eta\tilde{l}_t) - 1 + \eta p_t^T \tilde{l}_t)] \quad (\because \log(x) \leq x-1) \\
&= \frac{p_t^T}{\eta}[\exp(-\eta\tilde{l}_t) - \overrightarrow{1} + \eta\tilde{l}_t] \quad (\because p_t^T\overrightarrow{1} = 1) \\
&\leq \frac{p_t^T}{\eta}\frac{\eta^2\tilde{l}_t^2}{2} \quad (\because e^{-x} - 1 + x \leq \frac{x^2}{2},\ \forall x \geq 0) \\
&= \frac{\eta p_t^T \tilde{l}_t^2}{2} \\
&= \frac{\eta}{2}\frac{l_{A_t,t}^2}{p_{A_t,t}} \quad (\because \text{observation 3, } p_t^T\tilde{l}_t^2 = l_{A_t,t}^2/p_{A_t,t}) \\
&\leq \frac{\eta}{2p_{A_t,t}} \quad (\because l_{A_t,t} \leq 1)
\end{aligned}$$

The second part of this expression (B) is equivalent to:

$$
\begin{aligned}
-\frac{1}{\eta}[\log p_t^T \exp(-\eta \tilde{l}_t)] &= -\frac{1}{\eta}[\log \sum_{a\in\mathcal{A}} p_{a,t} \exp(-\eta \tilde{l}_{a,t})] \\
&= -\frac{1}{\eta}[\log \sum_{a\in\mathcal{A}} \frac{\exp(-\eta \tilde{L}_{a,t-1})}{\sum_{a^*\in\mathcal{A}} \exp(-\eta \tilde{L}_{a^*,t-1})} \exp(-\eta \tilde{l}_{a,t})] \quad (\because \text{definition of } p_{a,t+1}) \\
&= -\frac{1}{\eta}[\log \frac{\sum_{a\in\mathcal{A}} \exp(-\eta \tilde{L}_{a,t})}{\sum_{a^*\in\mathcal{A}} \exp(-\eta \tilde{L}_{a^*,t-1})}] \\
&= -\frac{1}{\eta}[\log \frac{\frac{1}{K}\sum_{a\in\mathcal{A}} \exp(-\eta \tilde{L}_{a,t})}{\frac{1}{K}\sum_{a\in\mathcal{A}} \exp(-\eta \tilde{L}_{a,t-1})}] \\
&= \Phi_{t-1} - \Phi_t
\end{aligned}
$$

where $\Phi_t := \frac{1}{\eta}\log[\frac{1}{K}\sum_{a\in\mathcal{A}} \exp(-\eta \tilde{L}_{a,t})], \tilde{L}_{a,0} := 0, \Phi_0 := 0$. This will result in a telescoping sum. Φ_t is called a potential function.

From the formulation for regret with respect to some fixed action $b \in \mathcal{A}$ found in the last lecture and the work above:

$$
\begin{aligned}
\sum_{t=1}^{T} l_{A_t,t} - \sum_{t=1}^{T} l_{b,t} &= \sum_{t=1}^{T} p_t^T \tilde{l}_t - \sum_{t=1}^{T} \mathbb{E}\left[\tilde{l}_{b,t} \mid \mathcal{H}_t\right] \\
&\leq \sum_{t=1}^{T} \frac{\eta}{2p_{A_t,t}} + \sum_{t=1}^{T} (\Phi_{t-1} - \Phi_t) - \sum_{t=1}^{T} \mathbb{E}\left[\tilde{l}_{b,t} \mid \mathcal{H}_t\right] \\
&= \sum_{t=1}^{T} \frac{\eta}{2p_{A_t,t}} + \Phi_0 - \Phi_T - \sum_{t=1}^{T} \mathbb{E}\left[\tilde{l}_{b,t} \mid \mathcal{H}_T\right]
\end{aligned}
$$

From above, $\Phi_0 := 0$ and

$$
\begin{aligned}
-\Phi_T &= \frac{\log K}{\eta} - \frac{1}{\eta}\log[\sum_{a\in\mathcal{A}} \exp(-\eta \tilde{L}_{a,T})] \\
&\leq \frac{\log K}{\eta} - \frac{1}{\eta}\log[\exp(-\eta \tilde{L}_{b,T})] \quad (\because \text{sum is greater than the single term involving } b) \\
&= \frac{\log K}{\eta} + \tilde{L}_{b,T}
\end{aligned}
$$

Thus, combining the above two formulas, the expected regret becomes:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} l_{A_t,t} - \sum_{t=1}^{T} l_{b,t}\right] &\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{\eta}{2p_{A_t,t}}\right] + \frac{\log K}{\eta} + \mathbb{E}\left[\tilde{L}_{b,T}\right] - \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\tilde{l}_{b,t} \mid \mathcal{H}_t\right]\right] \\
&= \sum_{t=1}^{T} \frac{\eta}{2} K + \frac{\log K}{\eta} \quad (\because \text{the last two terms cancel and fact 4}) \\
&= \frac{\eta}{2} KT + \frac{\log K}{\eta}
\end{aligned}
$$

When we tune $\eta := \sqrt{\frac{2 \log K}{TK}}$, we get the theorem statement and the regret bound.

The new advance of Auer et al. over the corresponding full information problem (aka "experts problem" where the entire vector l_t is revealed at tne end of each round) was the algorithm presented and the regret bound [ACBFS02]. The algorithm does require that we pay an additional price (of the order$\sqrt{K}$), which comes from Fact 4. In the full information case, the regret bound is $O(\sqrt{T \log K})$. In addition, this proof is brittle to departures of unbiasedness in the estimated losses.

The lower bound for the expected regret is $c\sqrt{TK}$ where c is some constant. The lower bound was shown to be tight in 2009 via the INF algorithm [AB09].

2 Introduction to Contextual Bandits

A brief introduction to contextual bandits will be found here. Multi-armed bandits do not use background information, but there is usually additional information in the real world. In the context of health care, if we were to be taking actions with regards to a patient's health, we would normally have other information about the patient and would want to take into account that information when deciding what actions to perform. Health care data in particular lends itself to a contextual bandit, with examples of context including number of steps, activity level, and number of activities on the calendar.

The general flow of information for a multi-armed bandit is:

1: **for** $t = 1$ to T **do**
2: Learner selects action $A_t \in \mathcal{A}$
3: Learner receives reward $R_t = R_t^{A_t}$
4: **end for**

In the case of contextual bandits, in addition to the stream of rewards for a particular action, there is a stream of contexts. These $X_1, X_2, \ldots$ are contexts that can be viewed as iid and stochastic. The action selected will be influenced by the context. If the actions were to impact the context, then we would be moving into the realm of reinforcement learning.

The contextual bandit flow of information is modified from the multi-armed bandit to:

1: **for** $t = 1$ to T **do**
2: Learner sees $X_t \in \mathcal{X}$
3: Learner selects action $A_t \in \mathcal{A}$ with consideration to X_t
4: Learner receives reward $R_t = R_t^{A_t}$
5: **end for**

References

[AB09] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 217–226, 2009.

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The non-stochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.