**STATS 710 – Seq. Dec. Making with mHealth Applications** Oct 18, 2016

# Lecture 12: EXP4 Algorithm

*Instructors: Susan Murphy and Ambuj Tewari* *Scribe: Huajie Qian*

## 1 Recap

Recall from Lecture 10 and 11 the following regret bounds for contextual bandits problem when competing against a class of deterministic policies $\Pi \subseteq \mathcal{X}^{\mathcal{A}}$, where $\mathcal{X}$ is the set of possible contexts and $\mathcal{A}$ is the set of actions. Let $K = |\mathcal{A}|$ be the total number of actions, and $T$ the number of rounds of play.

If the context set $\mathcal{X}$ is finite, i.e. $N = |\mathcal{X}| < \infty$, and $\Pi = \mathcal{X}^{\mathcal{A}}$, then one can achieve a regret bound $\sqrt{NKT \log K}$ by running EXP3 for each context $x \in \mathcal{X}$ separately (the $\log K$ factor can be removed by running a finer algorithm than EXP3). Rewrite the regret bound as $\sqrt{KT \log K^N}$, and note that $K^N = |\mathcal{X}^{\mathcal{A}}|$ is the number of policies that we are competing against. So it is tempting to guess that, which shall be proved below, when competing with an arbitrary class $\Pi \subset \mathcal{X}^{\mathcal{A}}$ there is some learning algorithm which gives a regret bound of $\sqrt{KT \log |\Pi|}$.

Otherwise, if $\mathcal{X}$ is a general set and the class of policies $\Pi$ has finite VC dimension VC-dim($\Pi$), then Epoch greedy algorithm (when $K = 2$) gives the regret bound $O((\text{VC-dim}(\Pi))^{1/3} T^{2/3})$.

## 2 Exponential-weight algorithm for Exploration and Exploitation using Expert advice (EXP4)

Now we revisit the non-stochastic bandits problem. We have shown that EXP3 gives the upper bound $\sqrt{2TK \log K}$ for expected regret against the best action, for any assignment of loss $l_{a,t} \in [0, 1]$. From another point of view, each action $a \in \mathcal{A}$ can be treated as an expert which always suggests taking action $a$ in each round, and the regret is defined against the best one of $K$ such experts. In this lecture, we move to a more general setting where there are an arbitrary and fixed number of experts (strategies, algorithms, policies) available who give advice on which action to choose. Here the way the experts generate their advice can be quite general. For example, it can be generated from MAB learning algorithms, policies in contextual bandits, and strategies such as sticking to a particular action in each round (the EXP3 case). The goal is to exploit their advice in such a way that the overall performance is close to that of the best expert. This is achieved by the EXP4 algorithm, which generalizes EXP3. The original paper on EXP4 is [ACBFS02].

### 2.1 Setup and the goal

Given a fixed but unknown set of loss $l_{a,t}, a \in \mathcal{A}$, the learner does the following

```
for t=1 to T do
    Receive "advice" from $N$ "experts": $\xi_t^1, \ldots, \xi_t^N$, $N$ probability measures over $\mathcal{A}$
    Combine $\xi_t^i$, $i = 1, \ldots, N$ into a single probability measure $p_t$ over $\mathcal{A}$
    Select an action $A_t \sim p_t$
    Suffer loss $l_{A_t,t}$
end for
```

The aim is to design a strategy of combining the advice so as to minimize the expected regret

$$\max_{k=1,\dots,N} \mathbb{E}\left[\sum_{t=1}^{T} l_{A_t,t} - Y_{k,T}\right] \tag{1}$$

against the (unknown) best expert, where $Y_{k,T} = \sum_{t=1}^{T} y_{k,t}$ with each $y_{k,t} = (\xi_t^k)^T l_t = \sum_{a\in\mathcal{A}} \xi_{a,t}^k l_{a,t}$ is the cumulative loss of the $k$-th expert. Notice that when $N = |\mathcal{A}|$ and $\xi_t^a$ is a point mass at $a \in \mathcal{A}$ for all $t$, then the regret (1) coincides with the one used for EXP3. As a side note, in contrast to Boosting which produces strong learners out of weak ones, the goal here is less ambitious, i.e., to do not much worse than the best choice.

## 2.2 EXP4 algorithm

**Algorithm 1** EXP4 with $N$ experts

```
Tuning parameter: $\eta > 0$
Let the initial weights $q_{k,1} = \frac{1}{N}$ and initial cumulative loss $\tilde{Y}_{k,0} = 0$ for $1 \le k \le N$
for t=1 to T do
    Get $\xi_t^1, \dots, \xi_t^N$
    Let $p_t = \sum_{k=1}^{N} q_{k,t}\xi_t^k$
    Draw $A_t \sim p_t$
    Observe $l_{A_t,t}$
    Estimate the loss values $\tilde{l}_{a,t} = \frac{l_{a,t}}{p_{a,t}}\mathbb{1}(A_t = a)$ for $a \in \mathcal{A}$
    Estimate the loss each expert suffers $\tilde{y}_{k,t} = \sum_{a\in\mathcal{A}} \xi_{a,t}^k \tilde{l}_{a,t}$ for $1 \le k \le N$
    Update the cumulative estimated loss of each expert $\tilde{Y}_{k,t} = \tilde{Y}_{k,t-1} + \tilde{y}_{k,s}$ for $1 \le k \le N$
    Update $q_{k,t+1} = \exp(-\eta\tilde{Y}_{k,t}) / \sum_{k'=1}^{N} \exp(-\eta\tilde{Y}_{k',t})$
end for
```

Note that when $N = K$ and $\xi_t^a$ is a point mass at $a \in \mathcal{A}$ for all $t$, the algorithm essentially reduces to EXP3. We have the following theoretical guarantee for EXP4.

**Theorem 1.** *The expected regret* (1) *of EXP*4 *on any sequence of* $[0,1]$*-bounded loss is at most* $\sqrt{2TK\log N}$ *when* $\eta$ *is set to be* $\eta = \sqrt{\frac{2\log N}{KT}}$.

*Proof:* The proof resembles the one for EXP3. Let $q_t = (q_{1,t},\dots,q_{N,t})^T, \tilde{y}_t = (\tilde{y}_{1,t},\dots,\tilde{y}_{N,t})$. Let's first display some useful observations.

**1** $q_t^T\tilde{y}_t = l_{A_t,t}$. This can be shown by manipulating the sum

$$q_t^T\tilde{y}_t = \sum_{k=1}^{N} q_{k,t}\tilde{y}_{k,t} = \sum_{k=1}^{N} q_{k,t}\sum_{a\in\mathcal{A}} \xi_{a,t}^k\tilde{l}_{a,t} = \sum_{a\in\mathcal{A}} \tilde{l}_{a,t}\sum_{k=1}^{N} q_{k,t}\xi_{a,t}^k = \sum_{a\in\mathcal{A}} \tilde{l}_{a,t}p_{a,t} = l_{A_t,t}$$

**2** $\mathbb{E}\left[\tilde{l}_{a,t}|\mathcal{H}_t\right] = l_{a,t}$.

**3** $\mathbb{E}\left[\frac{1}{p_{A_t,t}}|\mathcal{H}_t\right] = \sum_{a\in\mathcal{A}} \frac{1}{p_{a,t}}p_{a,t} = K$.

Now we proceed to the main proof. Similar to the key inequality of the EXP3 proof from Lecture 9, the following counterpart for EXP4 can be derived using a parallel argument (think of experts as actions in EXP3)

$$q_t^T \tilde{y}_t \leq \frac{\eta}{2} q_t^T \tilde{y}_t^2 + \Phi_{t-1} - \Phi_t, \tag{2}$$

where the squaring operation is component-wise, the term $\frac{\eta}{2} q_t^T \tilde{y}_t^2$ comes from controlling the cumulant generating function, and the potential function $\Phi_t$ is defined as:

$$\Phi_t = \frac{1}{\eta} \log[\frac{1}{N} \sum_{k=1}^{N} \exp(-\eta \tilde{Y}_{k,t})].$$

Note that $\Phi_0 = 0$. Summing up (2) over $t = 1, \ldots, T$ and using observation 1 gives

$$\sum_{t=1}^{T} l_{A_t,t} \leq \frac{\eta}{2} \sum_{t=1}^{T} q_t^T \tilde{y}_t^2 + \Phi_0 - \Phi_T = \frac{\eta}{2} \sum_{t=1}^{T} q_t^T \tilde{y}_t^2 - \Phi_T. \tag{3}$$

Now it is clear that, in order to bound the cumulative loss of EXP4, we need to bound the negated potential

$$\begin{aligned} -\Phi_T &\leq -\frac{1}{\eta} \log[\frac{1}{N} \sum_{k=1}^{N} \exp(-\eta \tilde{Y}_{k,T})] \\ &\leq -\frac{1}{\eta} \log[\frac{1}{N} \exp(-\eta \tilde{Y}_{k,T})] \text{ for any } 1 \leq k \leq N \\ &\leq \frac{\log N}{\eta} + \tilde{Y}_{k,T} \text{ for any } 1 \leq k \leq N \end{aligned}$$

as well as

$$\begin{aligned} q_t^T \tilde{y}_t^2 &= \sum_{k=1}^{N} q_{k,t} \tilde{y}_{k,t}^2 \\ &= \sum_{k=1}^{N} q_{k,t} (\sum_{a \in \mathcal{A}} \xi_{a,t}^k \tilde{l}_{a,t})^2 && (4) \\ &\leq \sum_{k=1}^{N} q_{k,t} \sum_{a \in \mathcal{A}} \xi_{a,t}^k \tilde{l}_{a,t}^2 && (5) \\ &= \sum_{a \in \mathcal{A}} \sum_{k=1}^{N} q_{k,t} \xi_{a,t}^k \tilde{l}_{a,t}^2 \\ &= \sum_{a \in \mathcal{A}} \tilde{l}_{a,t}^2 \sum_{k=1}^{N} q_{k,t} \xi_{a,t}^k = \sum_{a \in \mathcal{A}} \tilde{l}_{a,t}^2 p_{a,t} = \frac{l_{A_t,t}^2}{p_{A_t,t}} \\ &\leq \frac{1}{p_{A_t,t}}, \end{aligned}$$

where in passing from (4) to (5) we used the fact that $(\mathbb{E}[Z])^2 \leq \mathbb{E}[Z^2]$ for the discrete random variable $Z$ that takes value $\tilde{l}_{a,t}$ with probability $\xi_{a,t}^k$ and in the last line the boundedness of loss is used. Use the above two bounds to further bound (3)

$$\sum_{t=1}^{T} l_{A_t,t} - \tilde{Y}_{k,T} \leq \frac{\eta}{2} \sum_{t=1}^{T} \frac{1}{p_{A_t,t}} + \frac{\log N}{\eta},$$

Note that, due to observation 2 and 3, it holds that $\mathbb{E}\left[\tilde{Y}_{k,T}\right] = \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \xi_{a,t}^{k} l_{a,t} = Y_{k,T}$ for all $1 \leq k \leq N$ and $\mathbb{E}\left[q_t^T \tilde{y}_t^2\right] \leq K$. So the expected regret satisfies

$$\mathbb{E}\left[\sum_{t=1}^{T} l_{A_t,t} - Y_{k,T}\right] = \mathbb{E}\left[\sum_{t=1}^{T} l_{A_t,t} - \tilde{Y}_{k,T}\right] \leq \frac{\eta T K}{2} + \frac{\log N}{\eta}, \text{ for all } 1 \leq k \leq N.$$

Therefore

$$\max_{k=1,\ldots,N} \mathbb{E}\left[\sum_{t=1}^{T} l_{A_t,t} - Y_{k,T}\right] \leq \frac{\eta T K}{2} + \frac{\log N}{\eta}.$$

Choosing $\eta = \sqrt{\frac{2 \log N}{KT}}$ gives the desired conclusion. □

## References

[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.