CS 7545: Machine Learning Theory **Fall 2019**

Lecture 16: Adversarial + Stochastic Bandits

Lecturer: Jacob Abernethy *Scribes: Shunhao Oh, Nirbhay Modhe*

16.1 The EXP3 Algorithm and Proof

Recall that the EXP3 Algorithm is defined as follows:

Algorithm 1: EXP3 Algorithm [Auer, Cesa-Bianchi, Freund, and Schapire, 2003]

Fix some $\eta > 0$ Let $w_i^1 = 1$ for $i = 1, 2, \cdots, n$ **for** $t = 1, 2, \ldots T$ **do**
- Sets $p^t := w^t / \|w^t\|_1$;
- Sample $i_t \sim p^t$;
- Pays/observes loss $\ell_{i_t}^t \in [0, 1]$;
- Estimates $\widehat{\ell}^t := [0, \cdots, 0, \ell_{i_t}^t / p_{i_t}^t, 0, \cdots, 0]$; // $\widehat{\ell}^t$ is non-zero only on the i_tth entry
- Updates $w_i^{t+1} = w_i^t \exp\left(-\eta \widehat{\ell}_i^t\right)$ for all $i \in [n]$; // Note that $w_i^{t+1} = w_i^t$ for all $i \neq i_t$

end

Definition 16.1 (Regret) *We recall that **regret** in the adversarial setting is defined as follows:*

$$\text{Regret}_T := \sum_{t=1}^{T} \left(\ell_{i_t}^t - \ell_{i^*}^t\right)$$

Where $i^* \in [n]$ is the best arm in hindsight.

Claim 16.2 *EXP3 guarantees*

$$\mathbb{E}[\text{Regret}_T] \leq \frac{\log n}{\eta} + \frac{\eta}{2} nT$$

And with a properly tuned η (specifically, $\eta = \sqrt{\frac{2 \log n}{nT}}$)

$$\mathbb{E}[\text{Regret}_T] \leq \sqrt{2nT \log n}$$

Remark: compare this to the $O(\sqrt{T \log n})$ bound for the hedge setting. The extra $O(\sqrt{n})$ is the additional cost for not seeing the losses of the other arms.

Proof: As before, we define the potential function Φ_t as follows:

$$\Phi_t := \frac{1}{\eta} \log \left(\sum_{i=1}^{n} w_i^t \right)$$

We then proved in the previous lecture that:

$$\mathop{\mathbb{E}}_{i_t \sim p^t}[\Phi_{t+1} - \Phi_t \mid i_1, \cdots, i_{t-1}] \geq p^t \cdot \ell^t - \frac{\eta}{2} n$$

We can thus lower bound $\mathbb{E}[\Phi_{T+1} - \Phi_1]$ as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{i_1,\cdots,i_T}[\Phi_{T+1} - \Phi_1] &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{i_1,\cdots,i_T}[\Phi_{t+1} - \Phi_t] \\
&= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{i_1,\cdots,i_T}\left[\mathop{\mathbb{E}}_{i_1,\cdots,i_T}[\Phi_{t+1} - \Phi_t \mid i_1, \cdots, i_{t-1}]\right] && \text{by the tower rule} \\
&= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{i_1,\cdots,i_T}\left[\mathop{\mathbb{E}}_{i_t\sim p_t}[\Phi_{t+1} - \Phi_t \mid i_1, \cdots, i_{t-1}]\right] \\
&\geq \mathop{\mathbb{E}}_{i_1,\cdots,i_T}\left[\sum_{t=1}^{T}\left(p^t \cdot \ell^t - \frac{\eta}{2}n\right)\right]
\end{aligned}$$

We then upper bound $\mathbb{E}[\Phi_{T+1} - \Phi_1]$ as follows:

$$\begin{aligned}
\Phi_{T+1} - \Phi_1 &= -\frac{1}{\eta}\log\left(\sum_{i=1}^{n} w_i^T\right) + \frac{1}{\eta}\log\left(\sum_{i=1}^{n} w_i^1\right) && \text{by definition of } \Phi_t \\
&\leq -\frac{1}{\eta}\log\left(w_{i^*}^T\right) + \frac{1}{\eta}\log\left(\sum_{i=1}^{n} w_i^1\right) && \text{as } \sum_{i=1}^{n} w_i^T \geq w_{i^*}^T \\
&= -\frac{1}{\eta}\log\left(\exp\left(-\eta\sum_{t=1}^{T}\widehat{\ell}_{i^*}^t\right)\right) + \frac{1}{\eta}\log(n) && \text{as } w_i^1 = 1 \text{ for all } i \in [n] \\
&= \sum_{t=1}^{T}\widehat{\ell}_{i^*}^t + \frac{1}{\eta}\log(n)
\end{aligned}$$

And thus, as for any i^*, $\widehat{\ell}_{i^*}$ is an unbiased estimator of ℓ_{i^*}, we have:

$$\mathop{\mathbb{E}}_{i_1,\cdots,i_T}[\Phi_{T+1} - \Phi_1] \leq \sum_{t=1}^{T}\ell_{i^*}^t + \frac{1}{\eta}\log(n)$$

Putting the upper and lower bounds together, we get:

$$\begin{aligned}
\mathop{\mathbb{E}}_{i_1,\cdots,i_T}\left[\sum_{t=1}^{T}\left(p^t \cdot \ell^t - \frac{\eta}{2}n\right)\right] &\leq \sum_{t=1}^{T}\ell_{i^*}^t + \frac{1}{\eta}\log(n) \\
\mathop{\mathbb{E}}_{i_1,\cdots,i_T}\left[\sum_{t=1}^{T} p^t \cdot \ell^t\right] - \sum_{t=1}^{T}\ell_{i^*}^t &\leq \frac{1}{\eta}\log(n) + \frac{\eta}{2}nT
\end{aligned}$$

We note that $p^t \cdot \ell^t = \mathbb{E}_{i_t\sim p^t}[\ell_{i^t}^t]$, so

$$\begin{aligned}
\mathop{\mathbb{E}}_{i_1,\cdots,i_T}\left[\sum_{t=1}^{T}\ell_{i_t}^t\right] - \sum_{t=1}^{T}\ell_{i^*}^t &\leq \frac{1}{\eta}\log(n) + \frac{\eta}{2}nT \\
\mathop{\mathbb{E}}_{i_1,\cdots,i_T}[\text{Regret}_T] &\leq \frac{1}{\eta}\log(n) + \frac{\eta}{2}nT
\end{aligned}$$

■

Remark: EXP3's regret bound of $O\left(\sqrt{Tn\log n}\right)$ is not the best possible bound. One way to obtain a better regret bound of $O\left(\sqrt{Tn}\right)$ is to use the Tsallis entropy with mirror descent. This meets the $\Omega\left(\sqrt{Tn}\right)$ regret lower bound for the problem (Adversarial Multi-Armed Bandits).

16.2 The Stochastic Bandit Setting

The stochastic bandit setting is the more commonly studied setting. In this setting, we assume stochasticity in the world. In other words, each arm $i \in \{1, 2, \cdots, n\}$ has a fixed distribution D_i, and the gain (opposite of "loss") for playing arm i on each time step is an independent sample from D_i. The sequence of gain vectors $X_1, X_2, \cdots X_T$ is thus a sequence of i.i.d. samples from the distribution $(D_1, \cdots D_n)$.

Algorithm 2: Stochastic Bandit Setting

Assume: Have n distributions $D_1, \ldots D_n$, $\mathop{\mathbb{E}}_{X \sim D_i}[X] = \mu_i$, $|\mu_i - \mu_j| \leq 1$
Assume: Distributions $D_1, \ldots D_n$ are sub-gaussian with variance proxy 1
for $t = 1, 2, \ldots T$ **do**
  Algorithm picks $i_t \in [n]$;
  Algorithm observes gain (opposite of "loss") $X^t_{i_t} \sim D_{i_t}$
end

Note: Algorithm makes deterministic choices when picking actions $i_t \in [n]$

Definition 16.3 (Regret) *Let $i^* := \arg\max_{i \in [n]}(\mu_i)$. The **regret** in the stochastic bandit setting is defined as:*

$$\text{Regret}_T := \sum_{t=1}^{T} \left(\mu_{i^*} - X^t_{i_t}\right) \tag{16.1}$$

To simplify our notation later on, we assume without loss of generality that $i^* = 1$ (i.e. "first arm is best"). Let $\Delta_i = \mu_1 - \mu_i$ for $i = 2, 3, \ldots, n$. The expected regret of some algorithm choosing $i_1, i_2, \ldots i_T$ is as follows (note that N^t_i denotes "number of times i is chosen before time t"),

$$\mathbb{E}\left[\sum_{t=1}^{T} (\mu_1 - \mu_{i_t})\right] = \mathbb{E}\left[\sum_{i=2}^{n} N_i^{T+1} \Delta_i\right] \qquad \text{where } N^t_i := \sum_{s=1}^{t-1} \mathbb{1}_{[i_s = i]} \tag{16.2}$$

where $\mathbb{1}$ is the indicator function[1].

16.2.1 A Simple Algorithm for Stochastic Bandits

Algorithm 3: Simple Algorithm

Assume: $\Delta_* = \min_{i=2,\ldots,n} \Delta_i$ is known.
Let $K \leftarrow \left\lceil \frac{4 \log(nT)}{\Delta_*^2} \right\rceil$;
for $t = 1, 2, \ldots T$ **do**
  if $t \in [(i-1)K + 1, iK]$ *for some* $i \in [n]$ **then**
    `// explore`
    $i_t = i$;
  else
    `// exploit (for` $t > nK$`)`
    $i_t = \arg\max_i \hat{\mu}_i$; `// where` $\hat{\mu}_i := \frac{1}{K} \sum_{t=(i-1)K+1}^{iK} X^t_i$
  end
  Algorithm picks $i_t \in [n]$;
  Algorithm observes gain (opposite of "loss") $X^t_{i_t} \sim D_{i_t}$
end

---

[1] Indicator function is defined as: $\mathbb{1}_{[\text{statement}]} = \begin{cases} 1 & \text{if statement true;} \\ 0 & \text{if statement false} \end{cases}$

Claim 16.4 *Expected regret of simple algorithm [3] is:*

$$\mathbb{E}[\textit{Regret}_T(\textit{Simple Algorithm})] \leq \sum_{i=1}^{n} \frac{4\Delta_i \log(Tn)}{\Delta_*^2} + O(1) \tag{16.3}$$

16.2.2 Proof Sketch of Simple Algorithm

We give the main idea of the proof of Claim 16.4. The full proof will be given in the next lecture.

We consider 2 cases, which we will refer to as the **FOUND** and **NOT FOUND** events respectively.

- **Case 1 [FOUND]:** For $t > nK$, $i_t = 1$.
- **Case 2 [NOT FOUND]:** For $t > nK$, $i_t \neq 1$

In Case 1 [**FOUND**], N_i^{T+1} is at most K for all $i \neq 1$. In Case 2 [**NOT FOUND**], we can simply use a loose upper bound of T on the expected regret, using the assumption that $\Delta_i \leq 1$ for all $i \in [n]$. We can thus bound the regret as follows:

$$\begin{aligned}
\mathbb{E}[\text{Regret}_T] &= \mathbb{E}\left[\sum_{i=2}^{n} N_i^{T+1}\Delta_i\right] \\
&= \mathbb{E}\left[\mathbb{1}_{[\text{FOUND}]}\sum_{i=2}^{n} N_i^{T+1}\Delta_i + \mathbb{1}_{[\text{NOT FOUND}]}\sum_{i=2}^{n} N_i^{T+1}\Delta_i\right] \\
&\leq \mathbb{E}\left[\mathbb{1}_{[\text{FOUND}]}K\sum_{i=2}^{n}\Delta_i + \mathbb{1}_{[\text{NOT FOUND}]}T\right] \\
&\leq K\sum_{i=2}^{n}\Delta_i + \Pr[\text{NOT FOUND}]T
\end{aligned}$$

The remainder of the proof would be the use of Hoeffding's inequality to show that:

$$\Pr[\text{NOT FOUND}] \leq 1/T$$