CSE599s, Spring 2014, Online Learning Lecture 8 - 04/24/2014

The Multi-armed Bandit Problem

Lecturer: Ofer Dekel *Scribe: Saghar Hosseini*

1 Recap: Follow the Regularized Leader with Entropic Regularizer in the Probability Simplex

Recall the problem of learning with expert advice. Let d be the number of experts. At each round t, the player can choose one expert I_t and observe the loss of all experts if they would have been chosen. This is a full feedback problem and the following algorithm was presented in the previous lecture to minimize the expected regret.

```
for $t = 1, 2, \ldots, T$ do
    $p_t = \arg\min_{p \in \mathbb{R}^d} \{p l_{1:t-1} + \frac{1}{\eta} \sum_{i=1}^{d} (p_i \log p_i + \log d) + I_{\Delta d}(p)\}$
    Draw $I_t \sim p_t$, and incur loss $l_{t,I_t}$
    Observe $l_t \in [0, d]^d$
end for
```

Moreover, the Exponentiated Gradient (EG) algorithm has been introduced to solve this problem:

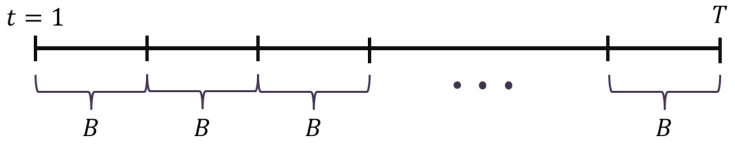
```
Initialize $w_1 = (1, 1, \ldots, 1)$
for $t = 1, 2, \ldots, T$ do
    Define $p_t = \frac{w_t}{||w_t||}$
    Draw $I_t \sim p_t$, and incur loss $l_{t,I_t}$
    Observe $l_t \in [0, d]^d$
    for $i = 1, 2, \ldots, d$ do
        Update $w_{t+1,i} = w_{t,i} e^{-\eta l_{t,i}} = e^{-\eta \sum_{s=1}^{t-1} l_{s,i}}$
    end for
end for
```

In some problems when the player choses one arm/experts he/she does not observe the whole loss vector $l_t \in [0, d]^d$. The player can only observe the loss associated with the expert that was chosen, i.e. l_{t,I_t}, and this is called a Bandits problem. In the next section, a method is presented to relate the "multi-arm bandits problem" to the "Experts" problem.

2 A (general) Reduction from "Bandits" to "Experts"

Blocking

Choose a block size B assuming that B divides T. Note that we are still assuming that the problem is in full feedback mode. In addition, let's assume we have a "Experts" algorithm called A with regret bound $c\sqrt{T}$



where c is a constant value.
We can present the blocks by an index set $\{1, 2, \dots, T/B\}$. Suppose player invokes algorithm A once per each block which means it chooses only one expert throughout of each block $b \in \{1, 2, \dots, T/B\}$. Therefore, when we run algorithm A, we choose an expert I_b from a probability vector p_b and play I_b throughout block b namely on $t = B(b-1)+1, \dots, Bb$. Note that $|\{B(b-1)+1, \dots, Bb\}| = B$.
Now, we choose τ_b uniformly at random from $\{B(b-1)+1, \dots, Bb\}$. Since algorithm A performs well on any arbitrarely sequence of $\{l_k\}$ and its regret is bounded by $c\sqrt{T}$, We know that the regret on $l_{\tau_1}, l_{\tau_2}, \dots, l_{\tau_{T/B}}$ is bounded as

$$\sum_{b=1}^{T/B} p_b l_{\tau_b} - \min_{p \in \Delta_d} \sum_{b=1}^{T/B} p_b l_{\tau_b} \leq c\sqrt{T/B}.$$

If we take the expectation of both side of the above inequality we have

$$\sum_{b=1}^{T/B} p_b \mathbb{E}[l_{\tau_b}] - \min_{p \in \Delta_d} \sum_{b=1}^{T/B} p_b \mathbb{E}[l_{\tau_b}] \leq c\sqrt{T/B}.$$

Moreover, we know that

$$\mathbb{E}[l_{I_{\tau_b}}] = \frac{1}{B} \sum_{t=B(b-1)+1}^{bB} l_t,$$

Thus, if we plug in $\mathbb{E}[l_{\tau_b}]$ in the regret bound we have

$$\frac{1}{B}(\sum_{b=1}^{T/B} \sum_{t=B(b-1)+1}^{Bb} p_b l_t - \min_{p \in \Delta_d} \sum_{t=1}^{T} p l_t) \leq c\sqrt{T/B}.$$

which implies

$$\frac{1}{B}(\sum_{t=1}^{T} \mathbb{E}[l_{t,I_t}] - \min_{p \in \Delta_d} \sum_{t=1}^{T} p l_t) \leq c\sqrt{T/B}.$$

Note that $I_t = I_b$ when b is the block that contains t. Therefore, the expected regret is bounded as

$$\text{Regret} = \sum_{t=1}^{T} \mathbb{E}[l_{t,I_t}] - \min_{p \in \Delta_d} \sum_{t=1}^{T} p l_t \leq c\sqrt{BT}.$$

Suppose we chose distinct $\tau_{b,1}, \tau_{b,2}, \dots, \tau_{b,d}$, all in $\{B(b-1)+1, \dots, Bb\}$ uniformly, where $B \geq d$. In other words, at the end of each block b, algorithm A observes the loss $[l_{\tau_{b,1}}, l_{\tau_{b,2}}, \dots, l_{\tau_{b,d}}]^T$. Thus, we have

$$\mathbb{E}[l_{\tau_{b,i}}] = \frac{1}{B} \sum_{t=B(b-1)+1}^{bB} l_{t,i} \quad \text{for all } i \in \{1, \dots, d\}.$$

Bandit Feedback

In this section we assume the problem is in "Bandits" mode. In order to observe $l_{t,i}$, we must choose expert i on round t. Therefore, we explore on d rounds in block b and we exploits on $B-d$ of the rounds, i.e., we play I_b suggested by algorithm A for $B-d$ rounds. Since $l_{t,i} \in [0,1]^d$ the cost of exploration on d rounds is upper bounded as

$$\text{Experts loss on exploration} \leq \frac{T}{B} d,$$

and subsequently we have

$$\text{Experts loss} \leq \mathbb{E}[l_{t,I_b}] + \frac{T}{B} d,$$

which implies

$$\text{Regret} =\leq c\sqrt{BT} + \frac{T}{B}d.$$

Now, we need to find B such that it minimizes $c\sqrt{BT} + \frac{T}{B}d$. By setting the gradient of $c\sqrt{BT} + \frac{T}{B}d$ with respect to B to zero, we have

$$\frac{1}{2}c\sqrt{T}B^{-1/2} - TdB^{-2} = 0$$

$$\Rightarrow B = (\frac{2d}{c})^{2/3}T^{1/3}.$$

Since Regret $\leq O(T^{2/3})$, the algorithm has to run for $T^{2/3}$ rounds to achieve the desired accuracy. Note that this approach only works in oblivious adversary setting. An example for this type of problem is showing one news story at each day on yahoo page.

An algorithm to solve the multi-arm bandits problem is called EXP3 algorithm which is similar to EG algorithm. In EXP3 algorithm the player chooses I_t and observes l_{t,I_t}. Then, he/she constructs an unbiased estimate of l_t named $\hat{l}_t$:

$$\hat{l}_t = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ l_{t,I_t}/p_{t,I_t} \\ \vdots \\ 0 \end{pmatrix},$$

where the non-zero element of $\hat{l}_t$ is associated with expert I_t. Note that we are very optimistic regarding the experts that are not playing. Moreover, we have

$$\mathbb{E}[\hat{l}_{t,i}|p_t] = l_{t,i} \quad \text{for all } i \in \{1, \dots, d\}.$$

Therefore the EXP3 algorithm can be presented as

```
Initialize $w_1 = (1, 1, \dots, 1)$
for $t = 1, 2, \dots, T$ do
  Define $p_t = \frac{w_t}{||w_t||}$
  Draw $I_t \sim p_t$
  Observe $l_{t,I_t} \in [0, d]^d$
  Construct $\hat{l}_t = [0, 0, \dots, l_{t,I_t}/p_{t,I_t}, \dots, 0]$
  for $i = 1, 2, \dots, d$ do
    Update $w_{t+1,i} = w_{t,i}e^{-\eta\hat{l}_{t,i}}$
  end for
end for
```

Analysis

Let's pretend that the adversary chose $\hat{l}_1, \dots, \hat{l}_T$. Therefore, for any constant vector $q \in \Delta_d$ we have

$$\mathbb{E}[\sum_{t=1}^{T} \hat{l}_t(p_t - q)] \leq R(q) + \eta \sum_{t=1}^{T} ||\hat{l}_t||_\infty^2,$$

where R is a strongly convex regularization function and η is the learning rate. Since $||\hat{l}_t||_\infty \leq G$, we have

$$\mathbb{E}[\sum_{t=1}^{T} \hat{l}_t(p_t - q)] \leq R(q) + \eta T G^2.$$

Note that random vector $p_t \in \Delta_d$ is $\mathcal{F}_t - 1$-measurable, where $\mathcal{F}_k$ is a σ-field and

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \mathcal{F}_T$$

In addition, random vector $\hat{l}_t \in \mathbb{R}^d$ is $\mathcal{F}_t$-measurable and subsequently

$$\mathbb{E}[\hat{l}_t p_t] \neq \mathbb{E}[\hat{l}_t]\mathbb{E}[p_t]$$

To be continued ...