CSE599s, Spring 2014, Online Learning Lecture 9 - 04/29/2014

EXP3 Regret Analysis

Lecturer: Ofer Dekel *Scribe: Travis Mandel*

1 Recap: Difference between Experts and Bandits

Recall that in past lectures we developed algorithms for the Expert problem, where at each round we had full informationabout how the other experts would have performed if we had selected them. One setting where this can occur in practice is the stock market. In fact the second programming assignment, which is currently being prepared, will involve running experts algorithms on stock market data that was very recently pulled from NASDAQ.

The leading algorithm in the experts setting is Exponentiated Gradient (EG), which in past lectures we showed has $O(\sqrt{T})$ regret.

However, we are now concerned with the bandit problem, in which we only get feedback from the expert (i.e. arm) we pulled. The question we are investigating in this unit is of the course is can we still do well given that some feedback has been removed?

In the last lecture we showed that we could use a "blocking" trick to segment time in such a way that separate samples were allocated to exploration and exploitation. We randomized when these phases occurred, which allowed us to rely on the assumption of an oblivious adversary to achieve a regret bound of $O(T^{2/3})$. However, this bound is pretty bad compared to $O(\sqrt{T})$ we have for EG in the experts problem.

In this lecture we want to recover $O(\sqrt{T})$ for the bandit case. The key idea is that we don't have to allocate separate samples to exploration and exploration, nor do we have to explicitly mix in some amount of uniform exploration. Instead it turns out that EG, given a small modification to work in the bandit setting, will have sufficient inherent exploration to give us the desired $O(\sqrt{T})$ regret bound.

2 The EXP3 Algorithm

EXP3 is a bandit algorithm developed around the year 2000. Before that, there had been quite a bit of work studying the stochastic multi-armed bandit problem, dating all the way back to Thompson in 1933. However it wasn't until the year 2000 that people developed good solutions for the **adversarial** multi-armed bandit problem.

EXP3 works by passing in a modified loss vector $\hat{\ell}_t$ to the experts algorithm EG, which is constructed as follows:

$$\hat{\ell}_t = \begin{pmatrix} \cdots \\ 0 \\ 0 \\ \frac{\ell_{t,I_t}}{p_{t,I_t}} \\ 0 \\ 0 \\ \cdots \end{pmatrix},$$

where I_t is the arm we pulled at time t. Note that $\hat{\ell}_t$ is an unbiased estimator of ℓ_t, in other words $\mathbb{E}\left[\hat{\ell}_t | p_t\right] = \ell_t$. This is true since:

$$\mathbb{E}\left[\hat{\ell}_{t,i} | p_t\right] = p_{t,i}\frac{\ell_{t,i}}{p_{t,i}} + (1 - p_{t,i})0 = \ell_{t,i}. \tag{1}$$

Beyond being an unbiased estimator, $\hat{\ell}_t$ is highly optimistic for arms that it has not pulled, hence it is likely to explore those arms, which intuitively adds a strong exploration component.

EXP3 can be precisely defined as follows for some learning rate η:

Algorithm 1 EXP3

$w_1 = (1, \ldots, 1)$
for $t = 1$ to T **do**
  Define $p_t = \frac{w_t}{\|w_t\|}$
  Draw $I_t \sim p_t$
  Observe $\ell_{t,I_t} \in [0, 1]$
  for $i = 1$ to d **do**
    if $i \neq I_t$ **then**
      $w_{t+1,i} = w_{t,i}$
    else
      $w_{t+1,i} = w_{t,i} e^{-\eta \frac{\ell_{t,i}}{p_{t,i}}}$
    end if
  end for
end for

Note that we could have equivalently written $w_{t+1,i} = w_{t,i} e^{-\eta \frac{\hat{\ell}_{t,i}}{p_{t,i}}}$ for all i, since $\hat{\ell}_{t,i} = 0$ if $i \neq I_t$.

3 Conditional Expectations and Measure Theory

Before proving the properties of EXP3, it might be useful to review conditional expectations. In other words, what does a quantity like $\mathbb{E}\left[\hat{\ell}_t | p_t\right]$ mean?

Assume Y and X are two real-valued random variables. Then $\mathbb{E}[Y]$ is just a scalar number.

How about $\mathbb{E}[Y|X = 3]$, assuming the event $X = 3$ has positive probability? It is also a scalar, as it is just the expectation of the distribution of Y when $X = 3$. Note that we can calculate that distribution as follows: $P(\cdot|X = 3) = \frac{P(\cdot)}{P(X=3)}$.

Ok, then what about $\mathbb{E}[Y|X]$? Well, that is a deterministic function $f(x) = \mathbb{E}[Y|X = x]$ applied to a random variable X. Therefore this quantity is also a random variable. One way to think of it is a random variable which averages out Y's randomness but keeps X's.

Now, say we had some $Z = Y + X$ where X and Y are independent. By linearity of expectation, $\mathbb{E}[Z] = \mathbb{E}[Y] + \mathbb{E}[X]$

Similarly, $\mathbb{E}[Z|X] = \mathbb{E}[X|X] + \mathbb{E}[Y|X]$. $\mathbb{E}[X|X]$ simply returns X since $\mathbb{E}[X|X = x] = x$. Since Y and X are assumed independent, $\mathbb{E}[Y|X] = \mathbb{E}[Y]$. So we have:

$$\mathbb{E}[Z|X] = X + \mathbb{E}[Y].$$

Going back to our online learning setting, note that the random choice we made yesterday determined today's random distribution. So p_t is a random variable which depends on $I_1, \ldots, I_{t-1}$. And ℓ_t is a random variable which depends on $I_1, \ldots, I_t$

Without going too deep into measure theory, measure theorists would say that $\hat{\ell}_t$ is $\mathcal{F}_t$ measurable. Basically this means that in order to calculate $P(\hat{\ell}_t = \alpha)$ for some α, we need to expose all $I_1, \ldots, I_t$. We can't reason about $P(\hat{\ell}_t = \alpha)$ without knowing the behavior of some $I_j \in \{I_1, \ldots, I_t\}$.

4 Analysis of EXP3

First, we can look at the regret bound EG gives us if we plug $\hat{\ell}_t$ into EG. This will give us, for all $q \in \triangle_d$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_t(p_t - q)\right]. \tag{2}$$

EG plugs in the following regularizer to FTRL:

$$R = \frac{1}{\eta}(\sum_{i=1}^{d} q_i \log q_i + \log d),$$

which, from the results in past lectures, gives us the following upper bound on equation (2):

$$R(q) \leq \frac{1}{\eta}\log d + \eta \sum_{t=1}^{T}\left\|\hat{\ell}_t\right\|_2^2.$$

However, equation (2) is not the actual regret, because it uses $\hat{\ell}_t$ instead of ℓ_t. Therefore we need a lemma to relate this to the true regret.

Lemma 1. $\mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_t(p_t - q)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(p_t - q)\right]$

Proof.

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_t(p_t - q)\right]\\ =&\sum_{t=1}^{T}\mathbb{E}\left[\hat{\ell}_t(p_t - q)\right]\end{aligned}$$

Which is due to linearity of expectation. Now, the law of total expectation tell us that for all random variables Z, X $\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[Z|X]]$. So we can rewrite as:

$$=\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\hat{\ell}_t(p_t - q)|p_t\right]\right].$$

Now, observe that given a fixed p_t, the $p_t - q$ term is a constant. So therefore the random variable $p_t - q$ is conditionally independent of the random variable $\hat{\ell}_t$ given p_t. Therefore we can split apart the multiplication:

$$=\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\hat{\ell}_t|p_t\right]\mathbb{E}\left[(p_t - q)|p_t\right]\right]$$

We have already shown in equation (1) that $\mathbb{E}\left[\hat{\ell}_t|p_t\right] = \ell_t$. And then $\mathbb{E}[(p_t - q)|p_t] = \mathbb{E}[p_t|p_t] - q = p_t - q$. Therefore:

$$\begin{aligned}&=\sum_{t=1}^{T}\mathbb{E}\left[\ell_t(p_t - q)\right]\\ &=\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(p_t - q)\right]\end{aligned}$$

□

So now we have the following bound on the true regret:

$$R(q) \leq \frac{1}{\eta} \log d + \eta \sum_{t=1}^{T} \left\| \hat{\ell}_t \right\|_2^2.$$

However, in order to make this bound concrete we need to deal with the $\left\| \hat{\ell}_t \right\|_2^2$ term. For our EG analysis we could upper bound this by G since we knew $\|\ell_t\|_2^2 \leq G$. In this case we do in fact know $\|\ell_t\|_2^2 \leq 1$, however, this does not bound the 2 norm of $\hat{\ell}_t$. In fact, $\hat{\ell}_t$ can grow arbitrarily large if the selected arm I_t has low probability p_{t,I_t}.

However, it turns out there is a special property of EG (and almost only EG) that allows us to still get a good regret bound. To prove it, we need another lemma.

Lemma 2. *For the EG algorithm,*

$$R(T) \leq \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} p_{ti} \ell_{ti}^2.$$

The intuition for why this is a good choice is that when we plug in $\hat{\ell}_{ti}$ the p_{ti} will cancel out the denominator in the $\hat{\ell}_{ti}$ term.

Proof. For some t, consider

$$\frac{1}{\eta} \log \sum_{i=1}^{d} p_{ti} e^{-\eta \ell_{ti}} \tag{3}$$

For a lower bound, we can plug EG's formula for p_{ti}. EG sets $p_t = \frac{w_t}{\|w_t\|_1}$, where $w_{ti} = \prod_{\tau=1}^{t} e^{-\eta \ell_{\tau,i}} = e^{-\eta \ell_{1:(t-1),i}}$ giving us a final result of

$$p_{ti} = \frac{e^{-\eta \ell_{1:(t-1),i}}}{\sum_{i=1}^{d} e^{-\eta \ell_{1:(t-1),i}}},$$

which gives us that equation (3) is equal to

$$\begin{aligned}
&\frac{1}{\eta} \log \sum_{i=1}^{d} \frac{e^{-\eta \ell_{1:(t-1),i}}}{\sum_{i=1}^{d} e^{-\eta \ell_{1:(t-1),i}}} e^{-\eta \ell_{ti}} \\
&= \frac{1}{\eta} \log \sum_{i=1}^{d} \frac{e^{-\eta \ell_{1:t,i}}}{\sum_{i=1}^{d} e^{-\eta \ell_{1:(t-1),i}}} \\
&= \frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:t,i}} - \frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:(t-1),i}}
\end{aligned}$$

This gives us (3) for one specific t, but if we want to sum (3) from 1 to T this becomes a telescoping sum. All terms will cancel but the terms for T and 1, giving us:

$$\sum_{t=1}^{T} \frac{1}{\eta} \log \sum_{i=1}^{d} p_{ti} e^{-\eta \ell_{ti}} = \frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:T,i}} - \frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:0,i}}.$$

$\ell_{1:0} = 0$, so $\sum_{i=1}^{d} e^{-\eta \ell_{1:0,i}} = d$, and we have:

$$= \frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:T,i}} - \frac{1}{\eta} \log d = \frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:T,i}} - \frac{\log d}{\eta} \tag{4}$$

Next we would like to upper bound equation (3). There are a couple of well-known and useful inequalities we will use to do so (they can be proved using the Taylor expansion):

$$\forall \alpha \geq 0, \log(\alpha) \leq \alpha - 1 \tag{5}$$

$$e^{-\alpha} \leq 1 - \alpha + \frac{\alpha^2}{2} \tag{6}$$

So since equation (3) is:

$$\frac{1}{\eta} \log \sum_{i=1}^{d} p_{ti} e^{-\eta \ell_{ti}},$$

we can use (5) to bound it by:

$$\leq \frac{1}{\eta} (\sum_{i=1}^{d} p_{ti} e^{-\eta \ell_{ti}} - 1).$$

Now we can apply (6) to bound it by:

$$\leq \frac{1}{\eta} (\sum_{i=1}^{d} p_{ti} (1 - \eta \ell_{ti} + \frac{\eta^2 \ell_{ti}^2}{2}) - 1)$$

$$= \frac{1}{\eta} \sum_{i=1}^{d} p_{ti} (\frac{\eta^2 \ell_{ti}^2 - \eta \ell_{ti}}{2})$$

$$= \frac{\eta}{2} \sum_{i=1}^{d} p_{ti} \ell_{ti}^2 - \sum_{i=1}^{d} p_{ti} \ell_{ti}$$

So, returning to equation (3) summed over timesteps:

$$\sum_{t=1}^{T} \frac{1}{\eta} \log \sum_{i=1}^{d} p_{ti} e^{-\eta \ell_{ti}} \leq \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} p_{ti} \ell_{ti}^2 - \sum_{t=1}^{T} \sum_{i=1}^{d} p_{ti} \ell_{ti}$$

And from equation (4):

$$\sum_{t=1}^{T} \frac{1}{\eta} \log \sum_{i=1}^{d} p_{ti} e^{-\eta \ell_{ti}} = \frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:T,i}} - \frac{\log d}{\eta}.$$

So:

$$\frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:T,i}} - \frac{\log d}{\eta} \leq \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} p_{ti} \ell_{ti}^2 - \sum_{t=1}^{T} \sum_{i=1}^{d} p_{ti} \ell_{ti} \tag{7}$$

$$\sum_{t=1}^{T} \sum_{i=1}^{d} p_{ti} \ell_{ti} + \frac{1}{\eta} \log \sum_{i=1}^{d} e^{-\eta \ell_{1:T,i}} \leq \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} p_{ti} \ell_{ti}^2 \tag{8}$$

Now, observe:

$$\frac{1}{\eta}\log\sum_{i=1}^{d} e^{-\eta\ell_{1:T,i}} \geq \frac{1}{\eta}\log e^{-\eta\min_i \ell_{1:T,i}},$$

since the sum must be greater than its minimum element. We can cancel the log to get:

$$\begin{aligned}
&= \frac{1}{\eta} - \eta\min_i \ell_{1:T,i} \\
&\geq -\eta\min_i \ell_{1:T,i}
\end{aligned}$$

Now, plugging this smaller quantity back in to equation (8) we have:

$$\sum_{t=1}^{T}\sum_{i=1}^{d} p_{ti}\ell_{ti} - \eta\min_i \ell_{1:T,i} \leq \frac{\log d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} p_{ti}\ell_{ti}^2.$$

The equation on the left is just the formula for regret in the experts setting, so we are done. □

Theorem 3. *EXP3 has regret bound* $\sqrt{2Td\log d}$.

Proof. By Lemma 2 EG has a regret bound of

$$\frac{\log d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} p_{ti}\ell_{ti}^2.$$

Now if we use EXP3, we have simply changed ℓ_{ti} to $\hat{\ell}_{ti}$. Lemma 1 tells us that

$$\begin{aligned}
\mathbb{E}\left[R(T)\right] =& \mathbb{E}\left[\frac{\log d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} p_{ti}\hat{\ell}_{ti}^2\right] \\
=& \frac{\log d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} \mathbb{E}\left[p_{ti}\hat{\ell}_{ti}^2\right] \\
=& \frac{\log d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} \mathbb{E}\left[\mathbb{E}\left[p_{ti}\hat{\ell}_{ti}^2|p_t\right]\right]
\end{aligned}$$

Similar to the proof of Lemma 1, observe that $\mathbb{E}\left[p_{ti}|p_t\right]$ is just a constant, so it is conditionally independent of $\hat{\ell}_{ti}^2$ given p_t.

$$= \frac{\log d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d} \mathbb{E}\left[p_{ti}\,\mathbb{E}\left[\hat{\ell}_{ti}^2|p_t\right]\right] \tag{9}$$

Now, observe that, since p_{ti} is the probability of filling in element i of the vector $\hat{\ell}_{ti}$:

$$\mathbb{E}\left[\hat{\ell}_{ti}^2|p_t\right] = p_{ti}\frac{\ell_{ti}^2}{p_{ti}^2} + (1-p_{ti})0 = \frac{\ell_{ti}^2}{p_{ti}}$$

. Plugging into (9) we have:

$$= \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} \mathbb{E}\left[p_{ti} \frac{\ell_{ti}^2}{p_{ti}} \right] \tag{10}$$

$$= \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} \mathbb{E}\left[\ell_{ti}^2 \right] \tag{11}$$

Next, ℓ_{ti} is chosen by the adversary beforehand and so does not involve any randomness once the game starts. So the expectation is just the fixed value.

$$= \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{d} \ell_{ti}^2 \tag{12}$$

We know $\ell_{ti} \in [0, 1]$, so $\ell_{ti}^2 \in [0, 1]$. Therefore $T \sum_{i=1}^{d} \ell_{ti}^2 \leq Td$, so:

$$R(T) \leq \frac{\log d}{\eta} + \frac{\eta}{2} Td.$$

All that remains is to choose η as a function of T to get the best bound possible. One good choice is

$$\eta = \sqrt{\frac{2 \log d}{Td}}.$$

Plugging this in, we have:

$$R(T) \leq \log d \frac{\sqrt{Td}}{\sqrt{2 \log d}} + \frac{1}{2} \sqrt{\frac{2 \log d}{Td}} Td$$

$$= \frac{1}{\sqrt{2}} \sqrt{Td} \sqrt{\log d} + \frac{1}{2} \sqrt{2} \sqrt{\log d} \sqrt{Td}$$

$$R(T) \leq \sqrt{2Td \log d}$$

□