18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Haihao (Sean) Lu

Lecture 16
Nov. 2, 2015

Recall that in last lecture, we talked about prediction with expert advice. Remember that $l(e_j, z_t)$ means the loss of expert j at time t, where z_t is one adversary's move. In this lecture, for simplexity we replace the notation z_t and denote by z_t the loss associated to all experts at time t:

$$z_t = \begin{pmatrix} \ell(e_1, z_t) \\ \vdots \\ \ell(e_K, z_t) \end{pmatrix} ,$$

whereby for $p \in \Delta^K$, $p^\top z_t = \sum_{j=1}^K p_j \ell(e_j, z_t)$. This gives an alternative definition of $f_t(p)$ in last lecture. Actually it is easy to check $f_t(p) = p^\top z_t$, thus we can rewrite the theorem for exponential weighting(EW) strategy as

$$R_n \le \sum_{t=1}^n p_t^\top z_t - \min_{p \in \Delta^k} \sum_{t=1}^n p^\top z_t \le \sqrt{2n \log K},$$

where the first inequality is Jensen inequality:

$$\sum_{t=1}^n p_z^\top z_t \ge \sum_{t=1}^n \ell(p_z, z_t) \ .$$

We consider EW strategy for bounded convex losses. Without loss of generality, we assume $\ell(p, z) \in [0, 1]$, for all $(p, z) \in \Delta^K \times \mathcal{Z}$, thus in notation here, we expect $p_t \in \Delta^K$ and $z_t \in [0, 1]^K$. Indeed if $\ell(p, z) \in [m, M]$ then one can work with a rescaled loss $\bar{\ell}(a, z) = \frac{\ell(a,z)-m}{M-m}$. Note that now we have bounded gradient on p_t, since z_t is bounded.

2. FOLLOW THE PERTURBED LEADER (FPL)

In this section, we consider a different strategy, called Follow the Perturbed Leader.

At first, we introduce Follow the Leader strategy, and give an example to show that Follow the Leader can be hazardous sometimes. At time t, assume that choose

$$p_t = \operatorname*{argmin}_{p \in \Delta^K} \sum_{s=1}^{t-1} p^\top z_s.$$

Note that the function to be optimized is linear in p, whereby the optimal solution should be a vertex of the simplex. This method can be viewed as a greedy algorithm, however, it might not be a good strategy.

Consider the following example. Let $K = 2$, $z_1 = (0, \varepsilon)^\top$, $z_2 = (0, 1)^\top$, $z_3 = (1, 0)^\top$, $z_4 = (0, 1)^\top$ and so on (alternatively having $(0, 1)^\top$ and $(1, 0)^\top$ when $t \ge 2$), where ε is small enough. Then with Following the Leader Strategy, we have that p_1 is arbitrary and in the best case $p_1 = (1, 0)^\top$, and $p_2 = (1, 0)^\top$, $p_3 = (0, 1)^\top$, $p_4 = (1, 0)^\top$ and so on (alternatively having $(0, 1)^\top$ and $(1, 0)^\top$ when $t \ge 2$).

In the above example, we have

$$\sum_{t=1}^{n} p_t^\top z_t - \min_{p\in\Delta^k} \sum_{t=1}^{n} p^\top z_t \le n - 1 - \frac{n}{2} \le \frac{n}{2} - 1 \ ,$$

which gives raise to linear regret.

Now let's consider FPL. FPL regularizes FL by adding a small amount of noise, which can guarantee square root regret under oblivious adversary situation.

Algorithm 1 Follow the Perturbed Leader (FPL)

Input: Let ξ be a random variables uniformly drawn on $[0, \frac{1}{\eta}]^K$.
for $t = 1$ to n **do**

$$p_t = \operatorname*{argmin}_{p\in\Delta^K} \sum_{s=1}^{t-1} \big(p^\top z_s + \xi\big).$$

end for

We analyze this strategy in oblivious adversaries, which means the sequence z_t is chosen ahead of time, rather than adaptively given. The following theorem gives a bound for regret of FPL:

Theorem: FPL with $\eta = \frac{1}{\sqrt{kn}}$ yields expected regret:

$$\mathbb{E}_\xi[R_n] \le 2\sqrt{2nK} \ .$$

Before proving the theorem, we introduce the so-called Be-The-Leader Lemma at first.

Lemma: (Be-The-Leader)
For all loss function $\ell(p, z)$, let

$$p_t^* = \arg\min_{p\in\Delta^K} \sum_{s=1}^{t} \ell(p, z_s) \ ,$$

then we have

$$\sum_{t=1}^{n} \ell(p_t^*, z_t) \le \sum_{t=1}^{n} \ell(p_n^*, z_t)$$

Proof. The proof goes by induction on n. For $n = 1$, it is clearly true. From n to $n+1$, it follows from:

$$\begin{aligned}
\sum_{t=1}^{n+1} \ell(p_t^*, z_t) &= \sum_{i=1}^{n} \ell(p_t^*, z_t) + \ell(p_{n+1}^*, z_{n+1}) \\
&\leq \sum_{i=1}^{n} \ell(p_n^*, z_t) + \ell(p_{n+1}^*, z_{n+1}) \\
&\leq \sum_{i=1}^{n} \ell(p_{n+1}^*, z_t) + \ell(p_{n+1}^*, z_{n+1}) \ ,
\end{aligned}$$

where the first inequality uses induction and the second inequality follows from the definition of p_n^*. $\square$

Proof of Theorem. Define

$$q_t = \operatorname*{argmin}_{p \in \Delta^K} p^\top (\xi + \sum_{s=1}^{t} z_s) \ .$$

Using the Be-The-Leader Lemma with

$$\ell(p, z_t) = \begin{cases} p^T(\xi + z_1) & if\ t = 1 \\ p^T z_t & if\ t > 1 \ , \end{cases}$$

we have

$$q_1^\top \xi + \sum_{t=1}^{n} q_t^\top z_t \leq \min_{q \in \Delta^K} q^\top (\xi + \sum_{t=1}^{n} z_t) \ ,$$

whereby for any $q \in \Delta^K$,

$$\sum_{i=1}^{n} \left(q_t^\top z_t - q^\top z_t \right) \leq \left(q^\top - q_1^\top \right) \xi \leq \|q - q_1\|_1 \|\xi\|_\infty \leq \frac{2}{\eta} \ ,$$

where the second inequality uses Hölder's inequality and the third inequality is from the fact that q and q_1 are on the simplex and ξ is in the box.

Now let

$$q_t = \arg\min_{p \in \Delta^K} p^\top \left(\xi + z_t + \sum_{s=1}^{t} z_s \right)$$

and

$$p_t = \arg\min_{p \in \Delta^K} p^\top \left(\xi + 0 + \sum_{s=1}^{t} z_s \right) \ .$$

Therefore,

$$\begin{aligned}
\mathbb{E}[R_n] &\leq \sum_{i=1}^{n} p_t^\top z_t - \min_{p \in \Delta^k} \sum_{i=1}^{n} p^\top z_t \\
&\leq \sum_{i=1}^{n} \left(q_t^\top z_t - p^{*T} z_t \right) + \sum_{i=1}^{n} \mathbb{E}[(p_t - q_t)^\top z_t] \\
&\leq \frac{2}{\eta} + \sum_{i=1}^{n} \mathbb{E}[(p_t - q_t)^\top z_t] \ , \qquad (2.1)
\end{aligned}$$

where $p^* = \arg\min_{p \in \Delta^K} \sum_{t=1}^n p^\top z_t$.

Now let

$$h(\xi) = z_t^\top \left(\arg\min_{p \in \Delta^K} p^\top [\xi + \sum_{s=1}^{t-1} z_s] \right) \ ,$$

then we have a easy observation that

$$\mathbb{E}[z_t^\top (p_t - q_t)] = \mathbb{E}[h(\xi)] - \mathbb{E}[h(\xi + z_t)] \ .$$

Hence,

$$\begin{aligned}
\mathbb{E}[z_t^\top (p_t - q_t)] &= \eta^K \int_{\xi \in [0, \frac{1}{\eta}]^K} h(\xi) d\xi - \eta^K \int_{\xi \in z_t + [0, \frac{1}{\eta}]^K} h(\xi) d\xi \\
&\leq \eta^K \int_{\xi \in [0, \frac{1}{\eta}]^K \setminus \left\{ z_t + [0, \frac{1}{\eta}]^K \right\}} h(\xi) d\xi \\
&\leq \eta^K \int_{\xi \in [0, \frac{1}{\eta}]^K \setminus \left\{ z_t + [0, \frac{1}{\eta}]^K \right\}} 1 d\xi \\
&= \mathbb{P}\left(\exists i \in [K], \xi(i) \leq z_t(i) \right) \\
&\leq \sum_{i=1}^K \mathbb{P}\left(\mathbf{Unif}\left([0, \frac{1}{\eta}] \right) \leq z_t(i) \right) \\
&\leq \eta K z_t(i) \leq \eta K \ ,
\end{aligned} \tag{2.2}$$

where the first inequality is from the fact that $h(\xi) \geq 0$, the second inequality uses $h(\xi) \leq 1$, the second equation is just geometry and the last inequality is due to $z_t(i) \leq 1$.

Combining (2.1) and (2.2) together, we have

$$\mathbb{E}[R_n] \leq \frac{2}{\eta} + \eta K n \ .$$

In particular, with $\eta = \sqrt{\frac{2}{Kn}}$, we have

$$\mathbb{E}[R_n] \leq 2\sqrt{2Kn} \ ,$$

which completes the proof. $\square$