CS861: Theoretical Foundations of Machine Learning Lecture 25 - 11/01/2024
University of Wisconsin–Madison, Fall 2024

Lecture 25: Follow the Perturbed Leader

Lecturer: Kirthevasan Kandasamy *Scribed by: Haoqun Cao, Sathya Kamesh Bhethanabhotla*

In this lecture we will continue where we left off in the previous lecture, by looking at the Experts problem but using negative entropy as the regularizer for the `FTRL` method.

Example 3: Experts Problem Revisited (`FTRL`)

This time around, we consider using the negative entropy as a regularizer,

$$\lambda(p) = -H(p) = \sum_{i=1}^{K} p(i)\log(p(i))$$

owing to the fact that a quadratic regularizer does not capture the geometry of the problem. Recall that $\lambda(p)$ is 1-strongly convex in $||\cdot||_1$. We can then bound our regret over T rounds in the following way,

$$\begin{aligned} R_T &\leq \frac{1}{\eta}\Big(\max_{\omega}\lambda(\omega) - \min_{\omega}\lambda(\omega)\Big) + \sum_{t=1}^{T}||g_t||_\star \\ &\leq \frac{1}{\eta}\Big(\underbrace{\max_{\omega} H(\omega)}_{\leq log(K)} - \min_{\omega} H(\omega)\Big) + \sum_{t=1}^{T}\underbrace{||\ell_t||_\infty}_{\leq 1} \\ &\leq \frac{\log(K)}{\eta} + \eta T \ \in \mathcal{O}(\sqrt{T\log(K)}) \quad \left(\text{for } \eta = \sqrt{\frac{\log(K)}{T}}\right) \end{aligned}$$

Given we have the bound, we can also derive the update rule for the problem,

$$p_t = \underset{p\,\in\Delta([K])}{\arg\min}\left(\sum_{s=1}^{t-1} l_s^\top p + \frac{1}{\eta}\sum_{i=1}^{K} p(i)\log(p(i))\right)$$

Since p is drawn from a simplex, we can frame this as the following optimization problem,

$$\underset{p}{\text{minimize}}\sum_{s=1}^{t-1} l_s^\top p + \frac{1}{\eta}\sum_{i=1}^{K} p(i)\log(p(i)) \ s.t.\ \mathbf{1}^\top p = 1, p \geq 0$$

We can now solve this by writing down its Lagrangian for the equality constraint and verify that the solution is non negative.

$$\mathcal{L} = \sum_{s=1}^{t-1} l_s^\top p + \frac{1}{\eta}\sum_{i=1}^{K} p(i)\log(p(i)) + \mu(p^\top\mathbf{1} - 1)$$

And we solve the Lagrangian by taking its derivative and setting it to 0, solving for $p(i)$,

$$\frac{\partial \mathcal{L}}{\partial p(i)} = \sum_{s=1}^{t-1} l_s(i) + \frac{1}{\eta}(1 + \log(p(i))) + \mu = 0$$
$$\Longrightarrow p_t(i) = e^{-\eta\mu} \exp(-\eta \sum_{s=1}^{t-1} l_s(i))$$
$$\text{(Here we know that } \sum_i p(i) = 1)$$
$$\Longrightarrow p_t(i) = \frac{\exp(-\eta \sum_{s=1}^{t-1} l_s(i))}{\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} l_s(j))}$$

Which is precisely the Hedge algorithm, an interesting result of using negative entropy as a regularizer for the experts problem.

1 Follow the Perturbed Leader

Lets now look at a different variation of follow the leader, ie Follow the perturbed leader `FTPL` that does not use a regularizer but creates a perturbed loss function which is then also used to optimize the weight vector at all time steps. Consider a time horizon T and a distribution D. We now want to sample a function $f_0 \sim D$. And when we compute the optimal weight ω at time step t, we choose,

$$\omega_t = \arg\min_{\omega \in \Omega} \sum_{s=0}^{t-1} f_s(\omega)$$

Remark Note here that this f_0 is the perturbation, and this function needs to be sampled in every round and included in the optimization problem associated with ω, and hence the index for the summation above begins with 0.

For the rest of this topic, we assume an oblivious adversary, for the sake of simplicity. The change involved with an adaptive adversary is that we need to sample $f_0 \sim D$ on every round.

1.1 A comparison with `FTRL`

We can now compare `FTPL` with the method we previously studied. While `FTRL` is deterministic, in that the regularizer has no randomness associated with it, while `FTPL` has an inherent stochasticity in its approach due to the sampling of f_0 from D. In cases where the function f_0 is close to, or "looks like" all the f_t's, the optimization for `FTPL` can be simpler. However, the choice of D here, is not entirely straightforward, which makes the performance of these techniques depends on the specific use case being analysed. Let us look at an example of how `FTPL` changes the formulation of the optimization problem.

Example 1 (Online Linear Optimization in a Convex Polytope)**.** Let $f_t(\omega) = \ell_t^\top \omega$ be linear and $\Omega = \{\omega : A\omega \leq b\}$ where $A \in \mathbb{R}^{N \times d}, b \in \mathbb{R}^N$.
Say we run `FTRL` with a quadratic regularizer $\Lambda(\omega) = \frac{1}{2\eta}||\omega||_2^2$. Then we can see from an example in a previous lecture that we choose ω_t as,

$$\omega_t = \arg\min_{\omega \in \Omega} \left(\sum_{s=1}^{t-1} \omega^\top \ell_s + \frac{1}{2\eta}||\omega||_2^2 \right) = \arg\min_{\omega \in \Omega} ||\omega + \eta \sum_{s=1}^{t-1} \ell_t||_2$$

And this solution can be implemented with the following iterative scheme,

$$\begin{aligned} u_t &\longleftarrow u_{t-1} + \eta \ell_{t-1} \\ w_t &\longleftarrow \arg\min_{w \in \Omega} ||\omega - u_t||_2^2 \end{aligned}$$

This projection operation, while still a convex optimization problem, can be computationally expensive in a polytope. Let us now try `FTPL` here. We will sample $\ell_0 \sim D$ (For some appropriately chosen D). So for each round, our selection of ω_t can be written as,

$$\omega_t = \arg\min_{\omega \in \Omega} \left(\sum_{s=0}^{t-1} \ell_s^\top \omega \right)$$

And this can be implemented using a simple Linear program, which is computationally cheaper than a projection, given by,

$$\min_{\omega} \omega^\top \left[\sum_{s=0}^{t-1} \ell_s \right] \text{ s.t. } A\omega \le b$$

Remark Intuitively, FTPL can be viewed as "fooling" the adversary with the randomness in the loss ℓ_0

2 A preliminary bound for `FTPL`

In this section we will write an initial bound for the regret R_T when using `FTPL` and prove it.

Lemma 1. *let $f = (f_1, \ldots, f_T)$ be a sequence of losses. Then `FTPL` satisfies,*

$$\begin{aligned} R_T(\pi^{FTPL}, f) &\triangleq \mathbb{E}\left[\sum_{t=1}^{T} f_t(\omega_t) \right] - \min_{\omega \in \Omega} \sum_{t=1}^{T} f_t(\omega) \\ &\le \sum_{t=1}^{T} \mathbb{E}\Big[\underbrace{f_t(\omega_t) - f_t(\omega_{t+1})}_{\text{Fluctuations in the loss}} \Big] + \mathbb{E}\left[\max_{\omega \in \Omega} f_0(\omega) - \min_{\omega \in \Omega} f_0(\omega) \right] \end{aligned}$$

Remark *Here the expectation is w.r.t. $f_0 \sim D$. This lemma does not assume convexity of Ω, f_t, Λ. And note that we consider an oblivious adversary here.*

Proof: Recall the following bound for FTRL. For all $u \in \Omega$,

$$\sum_{t=1}^{T} f_t(\omega_t) - \sum_{t=1}^{T} f_t(u) \le \lambda(u) - \min_{\omega \in \Omega} \lambda(\omega) + \sum_{t=1}^{T} (f_t(\omega_t) - f_t(\omega_{t+1})).$$

For a given f_0, let us apply the above lemma with $\Lambda = f_0$ (setting the regularizer as our perturbation) and $u = \omega_\star = \arg\min_{\omega \in \Omega} \sum_{t=1}^{T} f_t(\omega)$, the best action in hindsight.

$$\sum_{t=1}^{T} f_t(\omega_t) - \sum_{t=1}^{T} f_t(\omega_\star) \le f_0(\omega_\star) - \min_{\omega \in \Omega} f_0(\omega) + \sum_{t=1}^{T} (f_t(\omega_t) - f_t(\omega_{t+1}))$$

The claim follows by noting that $f_0(\omega_\star) \le \max_{\omega \in \Omega} f_0(\omega)$ and taking an expectation on both sides.

3 FTPL for experts problem

We will apply FTPL with $f_0(\cdot) = \ell_0^\top(\cdot)$, and where $\ell_0(a) \sim D(\eta)$ for all $a \in [K]$. On round t we will choose $p_t \in \operatorname{argmin}_{p\in\Delta([K])} \sum_{s=0}^{t-1} p^\top \ell_t$, which is equivalent to choosing $A_t \in \operatorname{argmin}_{a\in[K]} \sum_{s=0}^{t-1} \ell_t(a)$.

This gives rise to the following algorithm:

- Given: time horizon T, parameter η
- Sample $\ell_0(a) \sim D(\eta)$ for $a \in [K]$. $\quad \ell_0 \sim D$
- for $t = 1, \ldots, T$, $A_t \leftarrow \operatorname{argmin}_{a\in[K]} \sum_{s=0}^{t-1} \ell_s(\dot{a})$, $\quad \omega_t = \operatorname{argmin}_{\omega\in\Omega} \sum_{s=0}^{t-1} f_s(\omega)$

For all a, we will sample $\ell_0(a) = -Z(a)$ where $Z(a) \sim \text{Geom}(\eta)$.

3.1 Geometric distribution and related property

The Geom (η) distribution: the distribution of the number Z of Bern(η) coin flips to get the first 1 .

$$\text{pmf: for } k \in \{1,2,\ldots\} \quad p(k) = \mathbb{P}(Z=k) = (1-\eta)^{k-1}\eta.$$

Some useful properties: 1. $\mathbb{P}(Z \geq k+1 \mid Z \geq k) = \frac{\mathbb{P}(Z\geq k+1, Z\geq k)}{\mathbb{P}(Z\geq k)} = \frac{\mathbb{P}(Z\geq k+1)}{\mathbb{P}(Z\geq k)} = \frac{(1-\eta)^k}{(1-\eta)^{k-1}} = 1-\eta$. 2. Let $Z(a) \sim \text{Geom}(\eta)$ for $a \in [K]$. Then

$$\mathbb{E}\left[\|Z\|_\infty\right] = \mathbb{E}\left[\max_{a\in[K]} Z(a)\right] \leq 1 + \frac{H_K}{\eta}, \quad \text{where, } H_K = 1 + \frac{1}{2} + \cdots + \frac{1}{K}$$

3.2 FTPL for experts: negative geometric perturbation (cont'd)

Note that for the experts problem $\omega \in \Delta([K])$ is such that $\omega_t(a) = \mathbf{1}(A_t = i)$. Moreover,

$$\max_{\omega\in\Delta([K])} f_0(\omega) = \max_{\omega\in\Delta([K])} \ell_0^\top \omega = \max_{a\in[K]} \ell_0(a), \quad \min_{\omega\in\Delta([K])} f_0(\omega) = \min_{a\in[K]} \ell_0(a)$$

we have,

$$R_T\left(\pi^{\text{FTPL}}, f\right) \leq \sum_{t=1}^{T} \mathbb{E}\left[\ell_0\left(A_t\right) - \ell_0\left(A_{t+1}\right)\right] + \mathbb{E}\left[\max_{a\in[K]} \ell_0(a) - \min_{a\in[K]} \ell_0(a)\right]$$

Let us first bound the second term. Recall that $\ell_0(a) = -Z(a)$ where $Z(a) \sim \text{Geom}(\eta)$. As $Z(a) \geq 1$, we have

$$\max_a \ell_0(a) = \max_a -Z(a) \leq -1, \qquad -\min_a \ell_0(a) = \max_a Z(a) = \|Z\|_\infty$$

Therefore,

$$\mathbb{E}\left[\max_{a\in[K]} \ell_0(a) - \min_{a\in[K]} \ell_0(a)\right] \leq -1 + \mathbb{E}\left[\|Z\|_\infty\right] \leq -1 + 1 + \frac{H_K}{\eta} = \frac{H_K}{\eta}$$

Let us now bound the first term. Shortly, we will prove the following claim.

Claim 1. $\mathbb{P}\left(A_{t+1} = a \mid A_t = a\right) \geq 1 - \eta$ *for all* $a \in [K]$*, where* $\mathbb{P}$ *is with respect to* ℓ_0.

Then, we can write

$$\mathbb{E}\left[\ell_t\left(A_t\right) - \ell_t\left(A_{t+1}\right)\right] = \mathbb{E}[\underbrace{\ell_0\left(A_t\right) - \ell_0\left(A_{t+1}\right)}_{=0} \mid A_t = A_{t+1}]\mathbb{P}\left(A_t = A_{t+1}\right) +$$

$$\mathbb{E}[\underbrace{\ell_0\left(A_t\right) - \ell_0\left(A_{t+1}\right)}_{\leq 1 \text{ as } \ell_t\in[0,1]^K} \mid A_t \neq A_{t+1}]\mathbb{P}\left(A_t \neq A_{t+1}\right)$$

$$\leq \mathbb{P}\left(A_t \neq A_{t+1}\right) = \sum_{a=1}^{K} \underbrace{\mathbb{P}\left(A_{t+1} \neq a \mid A_t = a\right)}_{\leq \eta} \mathbb{P}\left(A_t = a\right) \leq \eta.$$

$$R_T\left(\pi^{\mathrm{FTPL}}, f\right) \leq \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\ell_0\left(A_t\right)-\ell_0\left(A_{t+1}\right)\right]}_{\leq \eta}+\underbrace{\mathbb{E}\left[\max_{a \in[K]} \ell_0(a)-\min_{a \in[K]} \ell_0(a)\right]}_{\leq H_K / \eta}$$

Therefore,

$$\begin{aligned}
R_T &\leq \eta T+\frac{H_K}{\eta} \\
&=2 \sqrt{T H_K} \quad \text{By choosing } \eta=\sqrt{H_K / T} \\
&\in \mathcal{O}(\sqrt{T \log (K)}).
\end{aligned}$$

We will now prove the Claim 1.

Proof Recall that $\ell_0(a)=-Z(a)$ where $Z(a) \sim \operatorname{Geom}(\eta)$. Let a be given. We will show that for every realization of $\{Z(j)\}_{j \neq a}$, we have

$$\mathbb{P}\left(A_{t+1}=a \mid A_t=a,\{Z(j)\}_{j \neq a}\right) \geq 1-\eta \Longrightarrow \mathbb{P}\left(A_{t+1}=a \mid A_t=a\right) \geq 1-\eta$$

Fix the values of $\{Z(j)\}_{j \neq a}$. First observe,

$$A_t=a \Longleftrightarrow \sum_{s=0}^{t-1} \ell_s(a)=\sum_{s=1}^{t-1} \ell_s(a)-Z(a) \leq \sum_{s=1}^{t-1} \ell_s(j)-Z(j) \quad \forall j \neq a$$

Now define $L_{t-1}(j) \triangleq \sum_{s=1}^{t-1} \ell_s(j)$ and $J_{t-1} \triangleq \min_{j \neq a}\left(L_{t-1}(j)-Z(j)\right)$. Therefore,

$$A_t=a \Longleftrightarrow Z(a) \geq L_{t-1}(a)-J_{t-1}$$

Now let us consider $A_{t+1}=j$. We can write,

$$\begin{aligned}
A_{t+1}=a \Longleftrightarrow & \sum_{s=0}^{t} \ell_s(a)=\sum_{s=1}^{t} \ell_s(a)-Z(a) \leq \sum_{s=1}^{t} \ell_s(j)-Z(j) \quad \forall j \neq a \\
& \Longleftrightarrow Z(a) \geq L_{t-1}(a)+\ell_t(a)-\left(\sum_{s=1}^{t-1} \ell_s(j)+\ell_t(j)-Z(j)\right) \quad \forall j \neq a \\
& \Longleftrightarrow Z(a) \geq L_{t-1}(a)-\underbrace{\left(\sum_{s=1}^{t-1} \ell_s(j)-Z(j)\right)}_{\geq J_{t-1}}+\underbrace{\ell_t(a)-\ell_t(j)}_{\leq 1 \text{ as } \ell_t \in[0,1]} \forall j \neq a \\
& \Longleftrightarrow Z(a) \geq L_{t-1}(a)-J_{t-1}+1
\end{aligned}$$

We therefore have,

$$\begin{aligned}
& \mathbb{P}\left(A_{t+1}=a \mid A_t=a,\{Z(j)\}_{j \neq a}\right) \\
& \quad \geq \mathbb{P}\left(Z(a) \geq L_{t-1}(a)-J_{t-1}+1 \mid Z(a) \geq L_{t-1}(a)-J_{t-1},\{Z(j)\}_{j \neq a}\right)=1-\eta
\end{aligned}$$

Hence, $\mathbb{P}\left(A_{t+1}=a \mid A_t=a\right) \geq 1-\eta$. □