AI-3001: Advanced Topics in Machine Learning **Aug-Nov 2022**

Lecture 9: Follow The Leader and Follow The Regularized Leader

Lecturer: Ganesh Ghalme *Scribes: Ganesh Ghalme*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

9.1 Follow The Leader (FTL)

Follow the leader algorithm (also called as fictitious play in Game Theory) is simple; just take an action (i.e. pick a point) that minimizes the cumulative loss over loss functions revealed so far. In particular, we begin with an arbitrary point $x_1 \in \mathcal{K}$ and given previous loss functions pick a point

$$x_{t+1} = \arg\min_{x \in \mathcal{K}} \sum_{s=1}^{t} f_s(x) \tag{9.1}$$

Consider the following example where the FTL algorithm will fail to give sublinear regret. Let $\mathcal{K} = [-1, 1]$ and the loss function sequence is given as $f_1(x) = -x/2$ and $f_t(x) = (-1)^t x$ for $t \geq 2$. That is, the loss function is $-x$ for every odd time instance and x for every even time instance.

time	1	2	3	4	5	$\cdots$
x_t	x_1	1	-1	1	-1	$\cdots$
f_t	-x/2	x	-x	x	-x	$\cdots$
$f_t(x_t)$	-x_1/2	1	1	1	1	$\cdots$

Note that the regret of the FTL algorithm on the above instance is $O(T)$. The reason for this is the fact that the algorithm is forced to make *wild* changes in its decision points overfitting to the recent loss function. Next we will see that the regularization helps us in getting a sublinear regret. IN fact, we can get an optimal (in terms of its dependence on T) regret with regularization. Whats more, the regularization also allows us to construct appropriate functions catering to the geometry of the problem. Before we study Follow The Regularized Leader (FTRL) we will first prove an important lemma about FTL.

Lemma 9.1 (FTL-BTL lemma)**.** *Let $x_1, x_2, \cdots$ be the sequence of points chosen by FTL. Then, for any $u \in \mathcal{K}$ and for and stopping time T, we have*

$$\sum_{t=1}^{T} f_t(x_{t+1}) \leq \sum_{t=1}^{T} f_t(u) \tag{9.2}$$

Proof. We prove the lemma by induction. For the base case $t = 1$, the proof follows from the fact that $x_2 = \arg\min_{x \in \mathcal{K}} f_1(x)$. Now assume that it is true for some time τ. That is

$$\sum_{t=1}^{\tau-1} f_t(x_{t+1}) \leq \sum_{t=1}^{\tau-1} f_t(u) \quad \text{for all} \quad u \in \mathcal{K} \tag{9.3}$$

Adding $f_\tau(x_{\tau+1})$ on both sides we get

$$\sum_{t=1}^{\tau} f_t(x_{t+1}) \leq \sum_{t=1}^{\tau-1} f_t(u) + f_\tau(x_{\tau+1}) \quad \text{for all} \;\; u \in \mathcal{K}$$

The aboev equation holds for all values of u. Choosing $u = x_{\tau+1}$ we obtain

$$\sum_{t=1}^{\tau} f_t(x_{t+1}) \leq \sum_{t=1}^{\tau} f_t(x_{\tau+1}) \leq \sum_{t=1}^{\tau} f_t(x_{\tau+1}) \;\; \text{for all} \;\; u \in \mathcal{K}.$$

The second inequality above follows from the fact that $x_{\tau+1} = \arg\min_{x \in \mathcal{K}} \sum_{t=1}^{\tau} f_t(x)$. □

9.2 Follow the Regularized Leader

- The FTL is not stable as it tries to adjust its choices based on fluctuations in the loss functions
- FTRL introduces the regularization term that does not let the algorithms choices fluctuate wildly
- The regularizer considered is bounded, α-strongly convex
- We will consider R to be α-strongly convex functions (A slightly weaker condition also suffices but we will not study that in this course).
- update rule

$$x_{t+1} = \arg\min_{x \in \mathcal{K}} [\eta \sum_{s=1}^{t} f_s(x) + R(x)] \tag{9.4}$$

```
Algorithm 1: FTRL Algorithm
Input: convex set $\mathcal{K}$, function class $\mathcal{F}$, regularization function $R$
Initialize: $x_1 = \arg\min_{x \in \mathcal{K}} R(x)$ ;
for $t = 1, 2, \cdots$ do
    - Algorithm plays $x_t \in \mathcal{K}$ ;
    - Environment reveals $f_t \in \mathcal{F}$ ;
    - Algorithm incurs a loss $f_t(x_t) \in \mathbb{R}$;
    - Update
          $x_{t+1} := \arg\min_{x \in \mathcal{K}} [\eta \sum_{s=1}^{t} f_s(x) + R(x)]$      (9.5)
end
```

Lemma 9.2. *Let $x_1, x_2, \cdots$ be the sequence of points chosen by FTRL algorithm. Then for any $u \in \mathcal{K}$ and any $t \geq 1$ we have*

$$\sum_{s=1}^{t} f_s(x_{s+1}) - \sum_{s=1}^{t} f_s(u) \leq \frac{R(u) - R(x_1)}{\eta} \tag{9.6}$$

The proof of this lemma is similar to the lemma we proved for FTL and is left as an exercise.

Observation 1. *The FTRL algorithm satisfies the FTL-BTL lemma for the FTL algorithm on the sequence $f_0 = R/\eta, f_1, f_2, \cdots$.*

Proof. Rearrange the terms in the above lemma and observe that

$$\sum_{s=0}^{t} f_s(x_{s+1}) \leq \sum_{s=0}^{t} f_s(u) \tag{9.7}$$

□

9.2.1 Bregman Divergence and Dual norm

Definition 9.3. *Let R be a strictly convex and continuously differentiable function on closed and convex set $\mathcal{K}$. then the Bregman divergence is defined as the function*

$$B_R(x||y) := R(x) - R(y) - \nabla R(y)^T(x-y) \tag{9.8}$$

Examples:

- Let $R = \frac{1}{2}||x||^2$. Then, $B_R(x||y) = \frac{1}{2}||x-y||^2$. Proof is left as exercise.
- $R(x) = \sum_i x_i \log(x_i)$ where $x \in \Delta_n$. Then the Bregman divergence is the KL-divergence.

$$B_R(x||y) = \sum_i x_i \log\left(\frac{x_i}{y_i}\right) \tag{9.9}$$

Proof. Consider $n = 2$. Let $(x, 1-x)$ and $(y, 1-y)$ be distributions X and Y. We have

$$\begin{aligned} B_R(x||y) &= x\log(x) + (1-x)\log(1-x) - y\log(y) - (1-y)\log(1-y) - \begin{pmatrix} 1+\log(y) & 1+\log(1-y) \end{pmatrix} \begin{pmatrix} x-y \\ y-x \end{pmatrix} \\ &= x\log(x) + (1-x)\log(1-x) - y\log(y) - (1-y)\log(1-y) - (x-y)\log\left(\frac{y}{1-y}\right) \\ &= \log(1-x) + y\log(x) - (1-y)\log(1-y) - y\log(y) - y\log(1-x) \\ &= x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right) \end{aligned}$$

□

Properties of Bregman Divergence:

- It is strictly convex in the first argument
- Non-negative $B_R(x||y) \geq 0$. The proof follows from Taylors theorem.
- Asymmetric i.e. $B_R(x||y) \neq B_R(y||x)$. Example: KL-divergence.
- Non-convex in the second argument. Example $R(x) = -\log(x)$ and $B_R(x||y) = \log(y) - \log(x) - \frac{x-y}{y}$.
- $\frac{\partial B_R(x||y)}{\partial x} = \nabla R(x) - \nabla R(y)$ (proof left as exercise)
- Cosine inequality.

$$B_R(x||y) + B_R(y||z) = B_R(x||z) + \langle x-y, \nabla R(z) - \nabla R(y)\rangle \tag{9.10}$$

Proof.

$$\begin{aligned} LHS &= R(x) - R(y) - \nabla R(y)^T(x-y) + R(y) - R(z) - \nabla R(z)^T(y-z) \\ &= B_R(x||z) + \langle x-y, \nabla R(z) - \nabla R(y)\rangle \\ &= RHS \end{aligned}$$

□

9.2.2 Dual norm

Definition 9.4 (Norm). *Let $\mathcal{K} \subseteq \mathbb{R}^n$ be a vector space. We call a function $||.||\colon \mathcal{K} \to \mathbb{R}$ a norm if it satisfies the following properties (norm can also be defined on a subset of $\mathbb{R}^n$ that is a vector space)*

1. *(Non-negativity) $||x|| \geq 0$ for all $x \in \mathcal{K}$ and $||x|| = 0$ iff $x = 0$*
2. *(Scalar multiplication) $||\alpha x|| = |\alpha| \cdot ||x||$ for all $x \in \mathcal{K}$ and for all $\alpha \in \mathbb{R}$*
3. *(Trinagle Inequality) $||x+y|| \leq ||x|| + ||y||$ for all $x, y \in \mathcal{K}$.*

The dual norm (also called sup-norm) is defined as follows.

Definition 9.5. *Given a norm $||.||_*$ we call $||x||^* := \sup_{y \in \mathcal{K}, ||y|| \leq 1} \langle x, y\rangle = \sup_{y \in \mathcal{K}} \langle x, \frac{y}{||y||}\rangle$ as the dual norm.*

Claim 9.6. *Dual norm is a norm.*

The formal proof is left as an exercise. Few hints: 1. to check the non-negativity observe that sup should be at-least as much as when we put $y = x$ in the inner product, 2. to check scalar multiplication property, replace $y = -x$ (we can do this since $\mathcal{K}$ is a vector space) and 3. to check triangle inequality you need to prove that

$$\sup_{z \in \mathcal{K}, ||z|| \leq 1} \langle x+y, z\rangle = \sup_{z \in \mathcal{K}, ||z|| \leq 1} [\langle x, z\rangle + \langle y, z\rangle] \leq \sup_{z \in \mathcal{K}, ||z|| \leq 1} \langle x, z\rangle + \sup_{z \in \mathcal{K}, ||z|| \leq 1} \langle x, z\rangle$$

Next we introduce the norms introduced by positive definite matrix A over $\mathbb{R}^n$.

Definition 9.7. *Let A be a positive definite matrix. Define $||x||_A = \sqrt{x^T A x}$.*

Use the property of positive definite matrices that $A = B^T B$ where B is positive definite matrix (Cholesky factorization).

Lemma 9.8. *Let $||x||_A$ be the norm induced by a symmetric positive definite matrix A then we have $||x||_A^* = \sqrt{x^T A^{-1} x}$.*

Proof. Write the problem as the following optimization problem.

$$\max_{x \in \mathbb{R}^n} \langle x, y\rangle$$
$$\text{Subject to, } \; x^T A x = 1$$

Write the lagrangian function

$$L(x, \lambda) = \Big[\langle x, y\rangle - \lambda(x^T A x - 1)\Big]$$

differentiating the lagrangian (wrt x) and equating to 0, we get

$$\begin{aligned} & y = 2\lambda A x^* \\ \implies & \langle x^*, y\rangle = 2\lambda x^{*T} A x^* = 2\lambda \qquad \text{(since } x^{*T}Ax^* = 1) \end{aligned}$$

Here x^* is the optimal value of x. We have that the optimal objective value is 2λ. Next, we have

$$\begin{aligned} x^{*T}Ax^* &= \frac{1}{4\lambda^2} y^T A^{-1T} A A^{-1} y = 1 \\ \implies \lambda &= 1/2\sqrt{y^T A^{-1T} y} \end{aligned}$$

Hence we have that the objective value is $2\lambda = \sqrt{y^T A^{-1T} y} = \sqrt{y^T A^{-1} y}$, as stated. As inverse of a symmetric positive definite matrix is also symmetric. □

9.2.3 Convex Optimization

Lemma 9.9. *Let f be a convex function on a closed convex set $\mathcal{K}$ and let x^* be the minima then for all $x \in \mathcal{K}$ we have*

$$\langle \nabla f(x^*), y - x^*\rangle \geq 0 \tag{9.11}$$

Lemma 9.10. *Let f be a convex function on a closed convex set $\mathcal{K}$ then for all $x, y \in \mathcal{K}$ we have*

$$\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq 0 \tag{9.12}$$

9.2.4 Regret Guarantee of FTRL

Theorem 9.11. *For any $u \in \mathcal{K}$, the FTRL algorithm satisfies the following*

$$\mathcal{R}_T(FTRL) \leq \eta \sum_{t=1}^{T} ||\nabla_t||_t^{*2} + \frac{R(u) - R(x_1)}{\eta} \tag{9.13}$$

Proof. We will begin by defining the notation used in the proof. First note that from Taylors approximaiton theorem we have

$$R(x) = R(y) + \langle \nabla R(y), x - y\rangle + 1/2(x-y)^T \nabla^2 R(z)(x-y) \tag{9.14}$$

Here $z = \beta x + (1-\beta)y$ for some $\beta \in [0, 1]$.

Consider two consecutive points chosen by FTRL i.e. x_t and x_{t+1} and some point z_t between them so that the the above equation (Eq 9.14) holds. Denote by $||x||_t = x^T \nabla^2 R(z_t) x$ to be the norm induced by the Hessian at z_t and $||x||_t^*$ be the dual norm. We have that $||x||_t^* = x^T \nabla^{-2} R(z_t) x$ (see preliminaries). The regret of FTRL is given as

$$\begin{aligned} \mathcal{R}_T(FTRL) &= \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \\ &\leq \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x_{t+1}) + \frac{R(u) - R(x_1)}{\eta} && \text{(follows from FTL-BTL lemma)} \\ &= \sum_{t=1}^{T} \langle \nabla_t, x_t - x_{t+1}\rangle + \frac{R(u) - R(x_1)}{\eta} && \text{(Since losses are assumed to be linear)} \end{aligned}$$

We complete the proof by showing that the inner product in the first summation is less than the $2\eta||\nabla_t||_t^{*2}$ for every t. Towards this we prove the below lemma.

Lemma 9.12. $\langle \nabla_t, x_t - x_{t+1}\rangle \leq 2\eta||\nabla_t||_t^{*2}$

Proof of the lemma. Let $g_t(x) = \eta \sum_{s=1}^{t} \nabla_s^T x + R(x)$ with $g_0(x) = 0$ for all $x \in \mathcal{K}$. Throughput this proof we will consider that the functions $\{f_t\}$ are linear. We begin with the following observation for linear loss functions.

Observation 2. $B_{g_t}(x_t||x_{t+1}) = B_R(x_t||x_{t+1})$

Proof.

$$\begin{aligned} LHS &= \eta \sum_{s=1}^{t} \langle \nabla_s, x_t\rangle + R(x_t) - \eta \sum_{s=1}^{t} \langle \nabla_s, x_{t+1}\rangle - R(x_{t+1}) - \langle \eta \sum_{s=1}^{t} \nabla_s + \nabla R(x_t), x_t - x_{t+1}\rangle \\ &= R(x_t) - R(x_{t+1}) - \langle \nabla R(x_t), x_t - x_{t+1}\rangle = RHS \end{aligned}$$

□

Write,

$$\begin{aligned} g_t(x_t) &= g_t(x_{t+1}) + (x_t - x_{t+1})\nabla g_t(x_{t+1}) + B_{g_t}(x_t||x_{t+1}) \\ &\geq g_t(x_{t+1}) + B_{g_t}(x_t||x_{t+1}) \qquad \text{(follows from optimality of } x_{t+1} \text{ and Eq. 9.11)} \\ &= g_t(x_{t+1}) + B_R(x_t||x_{t+1}) \\ &\qquad \text{(follows from linearity of loss functions; see observation above)} \\ \implies B_R(x_t||x_{t+1}) &\leq g_t(x_t) - g_t(x_{t+1}) \end{aligned}$$

We have

$$\begin{aligned} B_R(x_t||x_{t+1}) &\leq g_t(x_t) - g_t(x_{t+1}) = g_{t-1}(x_t) - g_{t-1}(x_{t+1}) + \eta\langle \nabla_t, x_t - x_{t+1}\rangle \\ &\leq \eta\langle \nabla_t, x_t - x_{t+1}\rangle \end{aligned} \tag{9.15}$$

The last inequality Follows from that fact that x_t is a minima of $g_{t-1}(.)$.

$$\tag{9.16}$$

Now we upper bound the RHS,

$$\begin{aligned} \langle \nabla_t, x_t - x_{t+1}\rangle &\leq ||\nabla_t||^*_{\nabla^2 R(z_t)} ||x_t - x_{t+1}||_{\nabla^2 R(z_t)} \\ &\qquad \text{(for } z_t \in \mathcal{K} \text{ s.t. } z_t = \alpha x_t + (1-\alpha)x_{t+1} \text{ for some } \alpha \in [0,1]) \\ &= ||\nabla_t||_t^* ||x_t - x_{t+1}||_t \qquad \text{(define } ||.||_t := ||.||_{\nabla^2 R(z_t)} \text{ and } ||.||_t^* = ||.||^*_{\nabla^{-2} R(z_t)}) \\ &= ||\nabla_t||_t^* \sqrt{2B_R(x_t||x_{t+1})} \end{aligned} \tag{9.17}$$

The last equality follows from the finite Taylor series expansion of twice differentiable function. i.e.

$$R(x) = R(y) + \langle \nabla R(y), x - y\rangle + \frac{1}{2}(x-y)^T \nabla^2 R(z)(x-y) \text{ where } z = \alpha x + (1-\alpha)y \text{ for some } \alpha \in [0,1]$$

We complete the proof using equation 9.15 and 9.17

$$\begin{aligned} \langle \nabla_t, x_t - x_{t+1} \rangle &\leq ||\nabla_t||_t^* \sqrt{2B_R(x_t||x_{t+1})} && \text{(from 9.17)} \\ &\leq ||\nabla_t||_t^* \sqrt{2\eta\langle \nabla_t, x_t - x_{t+1} \rangle} \\ \implies \langle \nabla_t, x_t - x_{t+1} \rangle &\leq 2\eta||\nabla_t||_t^{*2} \end{aligned}$$

□

This completes the proof of the theorem. □