E1 245: Online Prediction & Learning	**Fall 2014**
Lecture 16 — September 1	
Lecturer: Aditya Gopalan	*Scribe: Shreyas S*

16.1 Another view of FTRL

This lecture we take a look at a difficult view of Follow-The-Regularized-Leader(FTRL) referred as "Dual Space" view or the "Mirror Decent" framework. Recall from the previous lecture that the FTRL prediction for $t+1$ round for linear loss functions over convex set $K \in \mathbb{R}^d$ and regularizer $R : \mathbb{R}^d \leftarrow \mathbb{R}$ which is a convex function is given by

$$\begin{aligned} w_{t+1} &= \underset{w \in K}{\operatorname{argmin}} \left(\sum_{s=1}^{t} \langle z_s, w \rangle + R(w) \right) \\ &= \underset{w \in K}{\operatorname{argmin}} \left(\langle z_{1,t}, w \rangle + R(w) \right) \text{ where } z_{1,t} = \sum_{s=1}^{t} z_s \\ &= \underset{w \in K}{\operatorname{argmax}} \left(\langle -z_{1,t}, w \rangle - R(w) \right) \end{aligned}$$

Define $h : \mathbb{R}^d \leftarrow K$

$$h(\theta) = \underset{w \in K}{\operatorname{argmax}} \left(\langle \theta, w \rangle - R(w) \right)$$

FTRL can be written as

1. $\theta_1 = 0 \in \mathbb{R}^d$.
2. For $t = 1, 2, 3, \cdots$
   (a) $w_t = h(\theta_t)$ [PREDICTION].
   (b) $\theta_{t+1} = \theta_t - z_t$ [UPDATE].

The space where θ updates and iterates occur is referred as Dual space and K is called as Prediction/Decision space. The above algorithm has the following interpretation while θ_t is being updated in Dual space and actual decision w_t is "MIRRORED/LINKED" to K via the link function $h(\theta_t)$ which maps θs from Dual space to K.

16.2 Why focus on Linear losses

Suppose we have and Online Convex Optimization algorithm that works with linear loss functions, then we can apply the same algorithm to work with general convex losses without incurring additional regret. Let the sequence of convex loss functions be $f_1, f_2, \cdots, f_T$ then for $t = 1, 2, \cdots, T$

1. Play $w_t \in K$.
2. Get to see f_t.
3. Feed $f_t' : \mathbb{R}^d \leftarrow \mathbb{R}$ to the O.C.O algorithm for Linear losses given by $f_t'(x) = \langle \nabla f_t(w_t), x \rangle$ and obtain the prediction w_{t+1} for next round.

The Regret with respect to $u \in \mathbb{R}^d$ is given by

$$\sum_{t=1}^{T} \{f_t(w_t) - f_t(u)\} \leq \sum_{t=1}^{T} \{\langle \nabla f_t(w_t), w_t \rangle - \langle \nabla f_t(w_t), u \rangle\}$$

The above equation is got by using the following property of convex functions $\{f_t(u) \geq f_t(w_t) + \langle \nabla f_t(w_t), u - w_t \rangle\}$. So in essence we concentrate only on linear losses because of

1. Linear losses are the "hardest" to play against i.e., they are same as playing against general convex function.
2. It gives clean regret bounds.

In general for any sequence of convex functions $f_1, f_2, \cdots, f_T$, MIRROR descent is given by

1. $\theta_1 = 0 \in \mathbb{R}^d$.
2. For $t = 1, 2, 3, \cdots$
   (a) $w_t = h(\theta_t)$ [PREDICTION].
   (b) $\theta_{t+1} = \theta_t - \nabla f_t(w_t)$ [UPDATE].

16.3 Dual View of FTRL

Let the regularizer $R : \mathbb{R}^d \leftarrow \mathbb{R}$ be such that R is strictly convex. The Fenchen Dual of R is $R^* : \mathbb{R}^d \leftarrow \mathbb{R}$ and $\forall \theta \in \mathbb{R}^d$ is given by

$$R^*(\theta) = \sup_{x \in \mathbb{R}^d} [\langle x, \theta \rangle - R(x)]$$

The above definition holds even if R is not convex. The R^* is always convex and if R is convex then $(R^*)^* = R$. So in general for any regularizer $(R^*)^* =$ " convex closure of R" (i.e., the tightest convex fit to R) or in other words

$$(R^*)^* = \sup \left[\text{convex functions } f : \ f(x) \leq R(x) \ , \forall x \in \mathbb{R}^d \right]$$

16.3.1 Properties of Fenchen Dual function

We now discuss some important properties of Frenchen Dual function.

1. **Fenchen-Young Dual inequality:**

$$\forall \theta \ , \forall x : R^*(\theta) + R(x) \geq \langle x, \theta \rangle$$

   The equality holds if $x = \nabla R^*(\theta)$ or if $\theta = \nabla R(x)$.

2. $$\nabla R^*(\theta) = \underset{x}{\operatorname{argmax}} \left[\langle x, \theta \rangle - R(x) \right].$$

3. $\theta = \nabla R(x^*)$ or $\theta = \nabla R(\nabla R^*(\theta))$ (got from above property); equivalently we say that inverse of ∇R is ∇R^* i.e., $(\nabla R)^{-1} = \nabla R^*$.

4. $$R(x) = \sup_{\theta} \left[\langle x, \theta \rangle - R^*(\theta) \right].$$

5. $$\nabla R(x) = \underset{w \in K}{\operatorname{argmax}} \left[\langle x, \theta \rangle - R^*(\theta) \right].$$

Recall Mirror descent over $K = R^d$ is given by

$$h(\theta) = \underset{w \in K}{\operatorname{argmax}} \left(\langle \theta, w \rangle - R(w) \right) = \nabla R^*(\theta).$$

Table 16.1: A table giving $R(w)$ and corresponding $R^*(\theta)$

$R(w)$	$R^*(\theta)$
$\frac{1}{2}\|\|w\|\|_2^2$	$\frac{1}{2}\|\|\theta\|\|_2^2$
$\frac{1}{2}\|\|w\|\|_p^2$	$\frac{1}{2}\|\|\theta\|\|_q^2$ where $\frac{1}{p}+\frac{1}{q}=1$ and $0 \leq p,q \leq 1$
$\sum_{i=1}^d w_i(\log w_i - 1)$	$\sum_{i=1}^d \exp\theta_i$
$\sum_{i=1}^d w_i \log w_i$ where $w \in \Delta_d$	$\log\left(\sum_{i=1}^d \exp\theta_i\right)$
$\frac{1}{\eta}R(x)$	$\frac{1}{\eta}R^*(\eta\theta)$

Bibliography

[1] Gabor Bartok, David Pal, Csaba Szepesvari, and Istvan Szita. Online learning - CMPUT 654 Course Notes. 2011.

[2] Shai Shalev-Shwartz. Online Learning and Online Convex Optimization. 2011.