E1 245: Online Prediction & Learning	**Fall 2014**

Lecture 11 — September 9

Lecturer: Aditya Gopalan | *Scribe: Abhinav Das .N.V*

11.1 Recap

In the last lecture we analyzed the regret bound for Projected Online Gradient Descent for strongly convex losses and found that it is of O(log T). Also we studied the regret for Lipschitz loss functions with convex regularizer for the FTRL which turned out to be O($\sqrt{(\log\ N)*T}$) by suitably choosing the value of η. Towards the end we were on a new idea of Online Mirror Descent which is a different view point of FTRL. With linear cost functions ,ie. $f_t(x)$:=$\langle z_t, x\rangle$, we defined a link function $h : \mathbb{R}^d \to K$

$$h(\theta) = argmax_{w\in K}[\langle w, \theta\rangle - R(w)]$$

where R is the regularizer .
FTRL is same as
1)$\theta_1 = 0$
2)Predict: $w_t = h(\theta_t)$
3)Update: $\theta_{t+1} = \theta_t - z_t$

11.2 Geometric view of mirror descent

Let $R : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex function.
ie.$\forall x \neq y \in \mathbb{R}$ & $0 \leq \lambda \leq 1 : R(\lambda x + (1-\lambda)y) < \lambda R(x) + (1-\lambda)R(y)$
Even if the domain of the function R is not $\mathbb{R}^d$,but a convex set K, we can extend it to $\mathbb{R}^d$ by setting $R(x) = \infty\ \forall x \notin K$

11.2.1 Fenchel dual/Fenchel conjugate

It is defined for the function $R : \mathbb{R}^d \to \mathbb{R}$, as, $\forall\theta \in\mathbb{R}^d$:$R^*(\theta) = sup_{w\in\mathbb{R}^d}[\langle w, \theta\rangle - R(w)]$
INTUITION: We can represent a convex function f in two ways.
1)As the pairs $(x, f(x))$ which is the common representation.
2)As the pairs *(slope of the tangent,y intercept).* Fenchel conjugate is the function that relates between these two representations. Fig. 11.1 illustrates this idea.

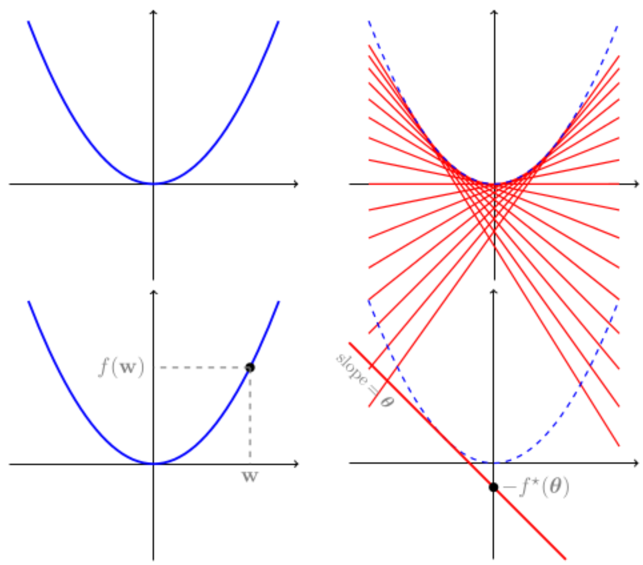

Fig. 11.1:Representation of a convex function in two different ways

11.2.2 Properties of the Fenchel dual

1)If R is convex and closed then $(R^*)^* = R$. In general $(R^*)^* \leq R$.

2)Fenchel-Young inequality

$\forall \theta, x \in \mathbb{R}^d : R^*(\theta) \geq [\langle x, \theta \rangle - R(x)]$
It is obvious from the definition of $R^*(\theta)$.
In particular if R and $R^*$are differentiable,the equality will be achieved when $x = \nabla R^*(\theta)$ or when $\theta = \nabla R(x)$
3)Duality
$argmax_{x \in \mathbb{R}^d}(\langle x, \theta \rangle - R(x)) = \nabla R^*(\theta)$
$argmax_{\theta \in \mathbb{R}^d}(\langle x, \theta \rangle - R^*(\theta)) = \nabla R(x)$
Let $x^* = argmax_{x \in \mathbb{R}^d}(\langle x, \theta \rangle - R(x))$

$\Rightarrow \theta = \nabla R(x^*)$=$\nabla R(\nabla R^*(\theta))$

$\Rightarrow (\nabla R)^{-1} = \nabla R^*$ or equivalently $(\nabla R^*)^{-1} = \nabla R$

$R(x)$	$R^*(\theta)$
$\frac{1}{2}\|\|x\|\|_2^2$	$\frac{1}{2}\|\|\theta\|\|_2^2$
$\frac{1}{2}\|\|x\|\|_p^2$	$\frac{1}{2}\|\|\theta\|\|_q^2$where $\frac{1}{p} + \frac{1}{q} = 1$
$\sum_{i=1}^{d} x(i)(\log(x(i)) - 1)$	$\sum_{i=1}^{d} e^{\theta(i)}$
$\sum_{i=1}^{d} x(i)(\log(x(i)) + I_{\Delta d}(x)$	$\log(\sum_{i=1}^{d} e^{\theta(i)})$
$\frac{1}{\eta} R(x)$	$\frac{1}{\eta} R^*(\eta\theta)$

Table 11.1:Some Fenchel dual pairs

11.2.3 Bregman Divergence

Let $R : \mathbb{R}^d \to \mathbb{R}$ be convex function.
The Bregman Divergence of R is defined as
$D_R(x,y) = R(x) - R(y) - \langle \nabla R(y), x-y \rangle \ \forall x,y \in \mathbb{R}^d$
We can see that it is the difference between the function value at x and first order Taylor series approximation of $R(x)$ around y.

11.2.4 Properties of Bregman Divergence

1. For a convex function R, $D_R(x,y) \geq 0$
2. $D_{R+S}(x,y) = D_R(x,y) + D_s(x,y)$
3. $D_R(u,v) + +D_R(v,w) = D_R(u,w) + \langle u-v, \nabla R(w) - \nabla R(v) \rangle$
4. Bregman projection to a convex set K

   $\forall w \in \mathbb{R}^d, \exists a$ unique $w^{'}$such that $w^{'} = argmin_{v \in K}(D_R(v,w))$

   We represent this $w^{'}$as $\prod_{R,K}(w)$.
5. Generalised Pythagorean theorem

   $\forall w \in \mathbb{R}^d$ and $w^{'} = \prod_{R,K}(w)$, $D_R(u,w) \geq D_R(u,w^{'}) + D_R(w^{'},w)$
6. Bregman Divergence through the dual space

   $D_R(u,w) = D_{R^*}(\nabla R(u), \nabla R(v))$
7. Gradient of the Bregman Divergence

   $\nabla_x D_R(x,y) = \nabla R(x) - \nabla R(y)$
8. Bregman Divergence for a line

   $D_{linear}(x,y) = 0$ (This is obvious since first order approximation on a line will not make any errors.)

11.3 Theorem

Theorem 11.1. *For linear cost functions FTRL is equivalent to performing the unconstrained minimization over entire $\mathbb{R}^d$ and then taking the Bregman projection to the convex decision space. Formally, let R be a strictly convex function which is the FTRL regularizer.*

$\Phi_t(x) = \sum_{s=1}^{t-1} \langle z_s, x \rangle + R(x)$,*then*

$argmin_{w \in K} \Phi_t(w) = \prod_{\Phi_t,K}(argmin_{w \in \mathbb{R}^d}(\Phi_t(w)))$

Proof: The first term in the $\Phi_t(x)$ is linear.
$\Rightarrow D_{\Phi_t} = D_R$
Let
$w_t^* := argmin_{w \in \mathbb{R}^d}(\Phi_t(w))$ ie.$w_t^*$is the universal minimizer.
$w_t := argmin_{w \in K}(\Phi_t(w))$ ie.w_t is the minimizer in the set K.
$w_t^{'} := \prod_{\Phi_t, K}(w_t^*)$ ie.$w_t^{'}$is the Bregman projection of the universal minimizer to the set K.

By definition, we know that

$$\Phi_t(w_t) \leq \Phi_t(w_t^{'}) \tag{11.1}$$

Also by definition we have

$$\nabla\Phi_t(w_t^*) = 0$$

$$\Rightarrow D_{\Phi_t}(w, w_t^*) = \Phi_t(w) - \Phi_t(w_t^*)$$

$$D(w_t^{'}, w_t^*) \leq D_{\Phi_t}(w_t, w_t^*)$$

$$\Phi_t(w_t^{'}) - \Phi_t(w_t^*) \leq \Phi_t(w_t) - \Phi_t(w_t^*) \tag{11.2}$$

(11.1) and (11.2) $\Rightarrow \Phi_t(w_t) = \Phi(w_t^{'})$ ie.Bregman projection and the minimizer in K are equal.
Also by strict convexity of Φ_t this minimizer must be unique.

□

11.3.1 FTRL in dual space

Unconstrained FTRL with linear loss function is according to

$w_{t+1}^* := argmin_{w \in \mathbb{R}^d}[\sum_{s=1}^{t} \langle z_s, w \rangle + R(w)]$
$\Rightarrow \sum_{s=1}^{t} z_s + \nabla R(w_{t+1}^*) = 0$
and $\sum_{s=1}^{t-1} z_s + \nabla R(w_t^*) = 0$
$\nabla R(w_{t+1}^*) = \nabla R(w_t^*) - z_t$
$w_{t+1}^* = \nabla R^*(\nabla R(w_t^*) - z_t)$ (Taking the inverse)
From this, constrained FTRL can be seen as
$w_{t+1}^* = \prod_{R,K}(\nabla R^*(\nabla R(w_t^*) - z_t))$
We can summarize the mirror descent update as
$\forall t = 1, 2, 3 \cdots$
1)$\nabla R(w_t^*) = \nabla R(w_{t-1}^*) - z_{t-1}$ (The reference pont is w_{t-1}^*)
2)$w_t = \prod_{R,K}(w_t^*)$

This is also called "Lazy version of OMD". Here our updation is done in the w_t^* space and it is projected back to K.

Another type of updation,called as "Active version of OMD", is also there. But it is not equal to FTRL.

$\forall t = 1,2,3\cdots$
1)$\nabla R(y_t) = \nabla R(w_{t-1}) - z_{t-1}$ (The reference pont is w_{t-1})
2)$w_t = \prod_{R,K}(y_t)$

Note:

1)When $R(w) = \frac{1}{2\eta}||w||_2^2$
a)$K = \mathbb{R}^d \Rightarrow$ Lazy OMD=Active OMD=OGD(η)
b)$K \subsetneq \mathbb{R}^d \Rightarrow$ Active OMD=Projected OGD
2)When $R(w) = \frac{1}{\eta}\sum_{i=1}^{d} w(i)(\log(w(i))$
Lazy OMD=Active OMD=EXP-WTS(η)

Theorem 11.2. *[Regret bound for Active version of OMD]*
$\forall u \in K$ $Regret_T(u) \le D_R(u,w_1) - D_R(u,w_{T+1}) + \sum_{t=1}^{T} D_R(w_t, y_{t+1})$

11.4 References

1)Online Learning and Online Convex Optimization By Shai Shalev-Shwartz:Chapter 2-sections-2.3,2.4,2.6,2.7

`http://www.cs.huji.ac.il/~shais/papers/OLsurvey.pdf`

2)Introduction to Online Optimization by Śebastien Bubeck:Chapter 5-sections-5.1,5.2

`http://www.princeton.edu/~sbubeck/BubeckLectureNotes.pdf`