E1 245: Online Prediction & Learning **Fall 2015**

Lecture 17 — October 1

Lecturer: Aditya Gopalan *Scribe: Prakash Barman*

17.1 Recap:

Recall FTRL with linear loss function $\langle Z_t \rangle$ over $K \in \mathbb{R}^d \equiv \theta_1 = 0 \in \mathbb{R}^d \ \forall t = 1, 2, 3, ...$
Predict: $w_t = h(\theta_t)$
Update: $\theta_{t+1} = \theta_t - \nabla f_t(w_t)$; θ is updated in "Dual" space, actual decision w_t is *mirrored/linked* to K via h.

17.2 Fenchel dual

Definition 17.2.1. *Legendre function: Let $K \subseteq \mathbb{R}^d$ be a convex set. A function $R : K \to \mathbb{R}$ is said to be Legendre function if (i) R is strictly convex with continuous gradients over K and (ii) any sequence x approaching the boundary of K (δK) satisfies* $\lim_{x \to \delta K} \|\nabla R(x)\| = +\infty$

Define the "dual space" of K (with respect to the function R) to be $K^* = \{\nabla R(x) : x \in K\}$

Definition 17.2.2. *Fenchel dual: The Fenchel dual of R is $R^* : K^* \to \mathbb{R}$;*

$$R^*(\theta) = \sup_{x \in K} \{\langle x, \theta \rangle - R(x)\}$$

Properties of Fenchel Dual:

(1) If R is Legendre, $(R^*)^* \equiv R$
(2) $R(x) = \sup_{\theta \in K^*} \{\langle x, \theta \rangle - R^*(\theta)\}$
(3) Fenchel-Young inequality: $\forall \theta \in K^*, \forall x \in K : R^*(\theta) + R(x) \geq \langle x, \theta \rangle$, with equality if $x = \nabla R^*(\theta)$ or $\theta = \nabla R(x)$.

Note:

(*) $\forall \theta \in K^*, \nabla(\nabla R^*(\theta)) = \theta$
(*) $\forall x \in K, \nabla R^*(\nabla(R(x)) = x$ i.e.

$$\theta \in K^* \underset{\nabla R(x)}{\overset{\nabla R^*(\theta)}{\rightleftarrows}} x \in K$$

Examples:

Here are few examples of some convex functions and their Fenchel duals.
(a) Euclidean squared norm is its own dual. More generally, $\frac{1}{2}\|x\|_q^2 \leftrightarrow \frac{1}{2}\|x\|_p^2$; where $p,q \geq 1$ and $\frac{1}{p}+\frac{1}{q}=1$
(b) $\sum_{i=1}^{d} x_i \log x_i - \sum_{i=1}^{d} x_i \leftrightarrow \left(\sum_{i=1}^{d} e^{\theta_i}\right)$
(c) The Fenchel dual of negative entropy is $\sum_{i=1}^{d} x_i \log x_i \leftrightarrow \log\left(\sum_{i=1}^{d} e^{\theta_i}\right)$

17.3 Bregman Divergence

Definition 17.3.1. *Bregman Divergence: Let $R : K \to \mathbb{R}$ be a Legendre function. The Bregman Divergence corresponding to R, $\forall x, y \in K$ is defined as $D_R(x,y) = R(x) - R(y) - \langle \nabla R(y), (x-y)\rangle$*

Examples:

(a) $R(x) = \frac{1}{2}\|x\|_2^2 \Longrightarrow D_R(x,y) = \frac{1}{2}\|x-y\|_2^2$; In this case it's symmetric.
(b) $R(x) = \sum_i x_i(\log x_i - 1) \Longrightarrow D_R(x,y) = \sum_i x_i \log(\frac{x_i}{y_i}) - \sum_i (x_i - y_i)$; i.e. The unnormalized Kullback-Leibler divergence is the Bregman divergence induced by the unnormalized negative entropy.

Properties of Bregman Divergences:

(1) $D_R \geq 0$ (D_R is the difference between function value R at x and the linear approximation of R around point y evaluated at point x. Since R is convex, the difference is always non-negative.)
(2) For R and S convex and differentiable $\Longrightarrow D_{R+S}(x,y) = D_R(x,y) + D_S(x,y)$
(3) "3-point inequality" $\Longrightarrow \forall u,v,w$: $D_R(u,v) + D_R(v,w) = D_R(u,w) + \langle (u-v), \nabla R(w) - \nabla R(v)\rangle$
(4) [Projection] Bregman projection onto a convex set A exists and is unique. R being Legendre, $\forall w$, $w' = \operatorname{argmin}_{v \in A} D_R(v,w)$ is unique.
(5) "Generalized Pythagorean theorem" : Let, $w' = \operatorname{argmin}_{v \in A} D_R(v,w)$; then $\forall u \in A$, $D_R(u,w) \geq D_R(u,w') + D_R(w',w)$
(6) $\forall u,v$, R Legendre $\Longrightarrow D_R(u,v) = D_{R^*}(\nabla R(v), \nabla R(u))$; R^* is the Fenchel dual of R.
(7) $\nabla_x[D_R(x,y)] = \nabla R(x) - \nabla R(y)$
(8) If f is linear, then $D_f = 0 \Longrightarrow D_{R+f} = D_R$

17.4 Equivalence of FTRL [constrained optimization ≡ unconstrained optimization + Bregman projection]

Lemma 17.1. *Let* $\phi_t(x) = \sum_{s=1}^{t-1} l_t(x) + R(x)$; *R is Legendre function. and Let,* $\prod_{\phi_t,K}(x) := \operatorname{argmin}_{v\in K} D_{\phi_t}(v,x)$ *be the Bregman projection of x onto K. Then* $\operatorname{argmin}_{w\in K} \phi_t(w) = \prod_{\phi_t,K}\left(\operatorname{argmin}_{w\in\mathbb{R}^d} \phi_t(w)\right)$

Proof: Let,

$$\begin{aligned}
\tilde{w}_t &:= \operatorname*{argmin}_{w\in\mathbb{R}^d} \phi_t(w) \\
w_t &:= \operatorname*{argmin}_{w\in K} \phi_t(w) \\
w'_t &:= \prod_{\phi_t,K}(\tilde{w}_t); w'_t \in K
\end{aligned}$$

So we require $w'_t = w_t$
$\because w_t$ is the minimizer of ϕ_t , So by definition, we have

$$\phi_t(w_t) \leq \phi_t(w'_t) \tag{17.1}$$

$\because \tilde{w}_t$ is the unconstrained minimizer of ϕ_t, we have $\nabla\phi_t(\tilde{w}_t) = 0$
$\therefore D_{\phi_t}(w,\tilde{w}_t) = \phi_t(w) - \phi_t(\tilde{w}_t); \forall w$
By definition of w'_t,

$$\begin{aligned}
&D_{\phi_t}(w'_t,\tilde{w}_t) \leq D_{\phi_t}(w_t,\tilde{w}_t) \\
&\phi_t(w'_t) - \phi_t(\tilde{w}_t) \leq \phi_t(w_t) - \phi_t(\tilde{w}_t) \\
&\phi_t(w'_t) \leq \phi_t(w_t)
\end{aligned} \tag{17.2}$$

By 17.1 and 17.2, we have $\phi_t(w'_t) = \phi_t(w_t) = \min_{w\in K} \phi_t(w)$

□

Mirror Descent form of FTRL:

Recall: Mirror Descent form of FTRL with linear losses $\langle Z_t\rangle$ over $K \subseteq \mathbb{R}^d \equiv \theta_1 = 0 \in \mathbb{R}^d$ $\forall t = 1,2,3,....$
Predict: $w_t = h(\theta_t) = \operatorname*{argmax}_{w\in K}[\underbrace{\langle w,\theta_t\rangle - R(w)}_{\phi_t(w)}]$
Update: $\theta_{t+1} = \theta_t - Z_t$;

$$h(\theta_t) = \prod_{R,K} \left(\underset{w \in \mathbb{R}^d}{\operatorname{argmax}} [\underbrace{\langle w, \theta_t \rangle - R(w)}_{\tilde{w}_t}] \right)$$

So $\forall t$, $\tilde{w}_t$ satisfies:

$$\begin{aligned}
&\sum_{s=1}^{t-1} Z_s + \nabla R(\tilde{w}_t) = 0 \\
&\sum_{s=1}^{t} Z_s + \nabla R(\tilde{w}_{t+1}) = 0 \\
&\nabla R(\tilde{w}_{t+1}) = \nabla R(\tilde{w}_t) - Z_t \\
&\tilde{w}_{t+1} = \nabla R^*[\nabla R(\tilde{w}_t) - Z_t] \\
&w_{t+1} = \prod_{R,K} [\nabla R^* \nabla R(\tilde{w}_t) - Z_t]
\end{aligned}$$

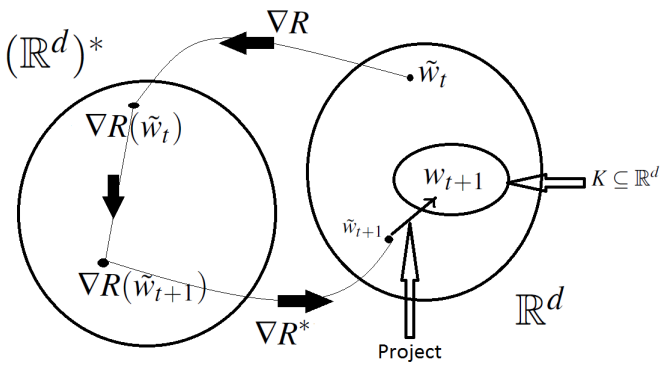

Figure 17.1. Mirror Descent form of FTRL

The figure above illustrates the online mirror descent form of FTRL. The iteration begins with the unconstrained (in $\mathbb{R}^d$) optimizer $\tilde{w}_t$ of the loss function. $\tilde{w}_t$ is mapped to the dual space via $\nabla R(\tilde{w}_t)$ and gradient step is taken in the dual space obtaining $\nabla R(\tilde{w}_{t+1})$. Then it's mapped back to $\tilde{w}_{t+1}$ using inverse mapping $\nabla R^*(\tilde{w}_{t+1})$. Finally the Bregman projection of $\tilde{w}_{t+1}$ onto K gives w_{t+1}.

References

[1] Nicolo Cesa-Bianchi and Gabor Lugosi, “Prediction, Learning and Games”, Cambridge University Press, 2006.

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