IE613: Online Machine Learning **Jan-Apr 2016**

Lecture 17: Follow the Leader and Follow the Regularized Leader

Lecturer: M. K. Hanawal *Scribes: Setu Hitesh Dave*

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.*

17.1 Recap

Online Mirror Descent

- For each round t=1,2,3...T
- Player ω_t , Environment C_t
- Player update,$\nabla\Phi(\tilde{\omega}_{t+1}) \leftarrow \nabla\Phi(\omega_t) - \eta\nabla\Phi(\omega_t)$
- $\omega_{t+1} \leftarrow \underset{k}{Proj}(\tilde{\omega}_{t+1})$

17.2 Fenchel - Legendre Conjugate

- If $f : \tau \to \mathbb{R}$,where C is a convex set and f is convex function , $C \subseteq \mathbb{R}^d$
- F-L conjugate is, $f^*(y) = \sup\limits_{x \in C} \; (< u, x > -f(x))$
- **Note**: Supremum are not always defined.

$$f^*(\nabla f(x)) =< \nabla f(x), x > -f(x)$$

Young Inequality,

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

$$a, b \geq 0, \frac{1}{p} + \frac{1}{q} = 1; a, b \in \mathbb{R}$$

For p=q=2, it boils down to AM $\geq$ GM.

$$i.e \;\; ab \leq \frac{a^2}{2} + \frac{b^2}{2}$$

take,$a^2 = c; b^2 = d$

- Convex combination :$\lambda x + (1 - \lambda)y$, $(1 - \lambda)$ is conjugate of λ

17.3 Holder's Inequality

$$\sum_{i=1}^{\infty} x_i y_i \leq ||x||\,||y|| \tag{17.1}$$

- Cauchy schwarz inequality can be obtained from this holder's inequality 17.1 Through the 17.1, we are trying to emphasis the norm of dual space V^* by: $f^*(x) = \sup_{x \in V/\{0\}} \frac{f(x)}{||x||}$

17.4 Follow the leader

17.4.1 Regret for generalized cost function

- For round t, player $\omega_t = \arg\min_{\omega \in K} \sum_{s=1}^{t-1} C_s(\omega)$.
- Regret bound after revealing C_t,

$$Regret(u,T) = \sum_{t=1}^{T} C_t(\omega_t) - \sum_{t=1}^{T} C_t(u) \tag{17.2}$$

Lemma 17.1 *For any* $u \in K, \sum_{t=1}^{T} (C_t(\omega_t) - C_t(u)) \leq \sum_{t=1}^{T} (C_t(\omega_t) - C_t(\omega_{t+1}))$, $\omega_{t+1} = \arg\min_{\omega} \sum_{s=1}^{t} C_s(\omega)$

- After rearranging the inequality in Lemma 17.1 $\sum_{t=1}^{T} C_t(\omega_{t+1}) \leq \sum_{t=1}^{T} C_t(u)$
- For T=1, $C_1(\omega_2) \leq C_1(u) \forall u, \omega_2 = \arg\min_{\omega} C_1(\omega)$.
- After induction on T, $\sum_{t=1}^{T-1} C_t(\omega_{t+1}) \leq \sum_{t=1}^{T-1} C_t(u)$.
- For a particular $u = \omega_{T+1}$, $\sum_{t=1}^{T-1} C_t(\omega_{t+1}) \leq \sum_{t=1}^{T-1} C_t(\omega_{T+1})$.
- Add $C_T(\omega_{T+1})$ on both side ,

$$\sum_{t=1}^{T} C_t(\omega_{t+1}) \leq \sum_{t=1}^{T} C_t(\omega_{T+1})$$

$$\sum_{t=1}^{T} C_t(\omega_{t+1}) \leq \sum_{t=1}^{T} C_t(u) \forall u$$

- Hence, $\omega_{T+1} = \arg\min_{\omega} \sum_{t=1}^{T} C_t(\omega)$

17.4.2 Quadratic cost function

$$C_t(\omega) = \frac{1}{2(t-1)\omega}||\omega \; - \; z_t||^2 \; , \; z_t \in K, \; ||z_t|| \leq B \tag{17.3}$$

- By defination , $\omega_{T+1} = \arg\min_{\omega} \sum_{t=1}^{T} \frac{1}{1} \; ||\omega \; - \; z_t||^2$

Claim 17.2

$$\omega_t = \frac{1}{t-1} \; \sum_{s=1}^{t-1} z_s \tag{17.4}$$

Proof:followed by the given equations

$$\sum_{t=1}^{T} \frac{1}{1} \; ||\omega \; - \; z_t||^2 = 0$$

$$\sum_{s=1}^{t}(\omega \; - \; z_s) = 0, \; (t-1)\omega = \sum_{s} z_s$$

$$\omega = \frac{1}{t-1}\sum_{s} z_s$$

Using induction on T,

$$\begin{aligned}
\Rightarrow \; \omega_{t+1} =& \frac{1}{t} \; \sum_{s=1}^{t} \; z_s \\
=& \frac{t-1}{t} \; \sum_{s=1}^{t} z_s \\
=& \frac{t-1}{t} \; \frac{1}{t-1} \; \sum_{s=1}^{t-1} z_s + \frac{1}{t} \; z_t \\
\omega_{t+1} =& \frac{t-1}{t}\omega_t + \frac{1}{t} z_t
\end{aligned}$$

Now for given cost function,

$$\begin{aligned}
\sum_{t=1}^{T}(C_t(\omega_t)) \; - \; (C_t(\omega_{t+1})) =& \frac{1}{2}\sum_{t=1}^{T} ||\omega_t - z_t||^2 \; - \; \frac{1}{2}\sum_{t=1}^{T} ||\omega_{t+1} - z_t||^2 \\
=& \frac{1}{2}\sum_{t=1}^{T} ||\omega_t - z_t||^2 \; - \; \frac{1}{2}\sum_{t=1}^{T} ||(\frac{t-1}{t})\omega_t - z_t + \frac{1}{t} z_t||^2 \\
=& \frac{1}{2}\sum_{t=1}^{T} ||\omega_t - z_t||^2 \; - \; \frac{1}{2}\sum_{t=1}^{T}(1-\frac{1}{t})^2||\omega_t - z_t||^2 - (1-\frac{1}{t})z_t \\
=& \frac{1}{2}\sum_{t=1}^{T}(1-(1-\frac{1}{t})^2)||\omega_t - z_t||^2
\end{aligned}$$

$$\sum_{t=1}^{T}(C_t(\omega_t)) \; - \; (C_t(\omega_{t+1})) = \frac{1}{2}\sum_{t=1}^{T}(1-(1-\frac{1}{t})^2)||\omega_t - z_t||^2 \tag{17.5}$$

17.4.3 Cauchy schwarz inequality

$$||\omega \; - \; z_t||^2 \; = \; ||\omega_t||^2 \; + \; ||z_t||^2 \; - \; 2 < \omega_t, z_t >$$

$$\leq ||\omega_t||^2 \; + \; ||z_t||^2 \; + \; 2||\omega_t|| \; ||z_t||$$

Using 17.5 , schwarz inequality and $||z_t|| \; \leq \; B$,

$$||\omega \; - \; z_t||^2 \; \leq \; \frac{1}{1}\sum_{t=1}^{T}(1-1-\frac{1}{t}+\frac{2}{t})\;(4B^2)$$

$$\leq \; 2B^2 \; \sum_{t=1}^{T}\frac{2}{t}$$

$$\leq \; 4B^2\;(1+\log(T))$$

Corollary 17.3 *Consider running FTL on an Online Quadratic Optimization problem with $S = R^d$ and let $B = \; max\, z_t$ Then, the regret of FTL with respect to all vectors $u \in R^d$ is at most $4B^2 \; (\log(T) + 1)$*

17.5 Follow the regularized leader(FTRL)

- Player at t round, $\omega_t = \arg\min_{\omega \in K} \; \sum_{s=1}^{T} C_s(\omega) \; + \; \Psi(\omega)$
- $\Psi(\omega)$ is regularized weighted scalar with η
- Environment chooses ,$C_t \; = < \omega_t, z_t >$
- Player chooses,$\Psi(\omega) = \frac{1}{2\eta} \; ||\omega||^2$.

  Theorem 17.4 *Consider running FoReL on a sequence of linear functions, $f_t(\omega) =< \omega, z_t >$ for all t with $S = R^d$, and with the regularizer $r(\omega) = \frac{1}{2\eta}||\omega||^2$, which yields the predictions given in 17.6 Then, for all u we have*

$$R_{FTRL}(u,T) \; \leq \; \frac{1}{2\eta}\;||u||^2 \; + \; \frac{\eta}{2}\sum_{t=1}^{T}||z_t||^2, \; \eta = \frac{u}{\sqrt[2]{T\;B}} \tag{17.6}$$

- When $\eta = \frac{u}{\sqrt[2]{T\;B}}$, Regret as per 17.6 will be bounded by $\sqrt[2]{T}\;B\;u$
- Trying to convert FTRL to FTL, for $t \geq 1$ and round t $= 0$,$\omega_0 = \arg\min_u \; \Psi(u)$

- For $t \geq 1$, environement reveals C_t, $\omega_t = \arg\min_{\omega} \ \Psi(\omega) \ + \ \sum_{s=1}^{t-1} C_s(\omega)$

- For FTL,$\sum_{t=1}^{T} (C_t(\omega_t) \ - \ C_t(u)) \ \leq \ \sum_{t=1}^{T} (C_t(\omega_t) \ - \ C_t(\omega_{t+1}))$

  For FTRL,

$$\begin{aligned}
\Psi(\omega_0) \ - \ \Psi(u) \ + \ \sum_{t=1}^{T} (C_t(\omega_t) \ - \ C_t(u)) \ &\leq \ \sum_{t=1}^{T} (C_t(\omega_t) \ - \ C_t(\omega_{t+1})) \ + \ \Psi(\omega_0) \ - \ \Psi(\omega_1) \\
\sum_{t=1}^{T} (C_t(\omega_t) \ - \ C_t(u)) \ &\leq \ \sum_{t=1}^{T} (C_t(\omega_t) \ - \ C_t(\omega_{t+1})) \ + \ \Psi(u) \ - \ \Psi(\omega_1)
\end{aligned}$$

- Now ,$\Psi(\omega) \ = \ \frac{1}{2\eta} \, ||\omega_1||^2 \ \geq \ 0$

$$\sum_{t=1}^{T} (C_t(\omega_t) \ - \ C_t(\omega_{t+1})) \ + \ \Psi(u) \ - \ \Psi(\omega_1) \ \leq \ \frac{1}{2\eta} \, ||u||^2 \ + \ \sum_{t=1}^{T} (C_t(\omega_t) \ - \ C_t(\omega_{t+1})) \tag{17.7}$$

$$\begin{aligned}
C_t(\omega_t) \ - \ C_t(\omega_{t+1}) \ &= \ <\omega_t, z_t> - <\omega_{t+1}, z_t> \\
&= \ <\omega_t \ - \ \omega_{t+1}, z_t> \\
&\leq \ ||\omega_t \ - \ \omega_{t+1}|| \ ||z_t|| \\
&\leq \ \eta \, ||z_t|| \ ||z_t||, \ \ \eta \ ||z_t|| = ||\omega_t \ - \ \omega_{t+1}||
\end{aligned}$$

- In FTL algorithm ,$C_t(\omega) \ = <\omega_t, z_t>$

$$\begin{aligned}
\omega_t &= \arg\min_{\omega} \sum_{s=1}^{t-1} C_s(\omega) \\
&= \arg\min_{\omega} \sum_{s=1}^{t-1} <\omega_t, z_t>
\end{aligned}$$

- $\omega_t \ = \ \underset{k}{Proj} \ [-\,\eta \sum_{s=1}^{t-1} z_s] \ \ ,\omega_{t+1} \ = \ \underset{k}{Proj} \ [-\,\eta \sum_{s=1}^{t} z_s]$

- Using inequality , $||\underset{k}{Proj}(u) - \underset{k}{Proj}(v)|| \ \ \leq \ \ ||u \ - \ v||$

$$\begin{aligned}
||\omega_t - \omega_{t+1}|| \ &= ||\underset{k}{Proj}[-\,\eta \textstyle\sum_{s=1}^{t-1} z_s] \ \ + \ \ \eta \textstyle\sum_{s=1}^{t-1} z_s|| \\
&\leq ||-\eta \sum_{s=1}^{t-1} z_s \ \ + \ \ \eta \sum_{s=1}^{t} z_s|| \\
&= \eta \ \ ||z_t||
\end{aligned}$$

Differentiate $\omega_t = \arg\min_{\omega \in K} \sum_{s=1}^{T} C_s(\omega) + \Psi(\omega)$ w.r.t ω ,

$$\frac{1}{\eta}\omega + \sum_{s=1}^{t-1} z_s = 0$$

$$\omega = -\eta \sum_{s=1}^{t-1} z_s$$

$$\omega = \underset{k}{Proj}[-\eta \ \textstyle\sum_{s=1}^{t-1} z_s]$$

- This is called a lazy projection since we dont project the iterates t but only project when we need to make a prediction. This is also called Nesterovs Dual Averaging algorithm.