E1 245: Online Prediction & Learning	Fall 2014

Lecture 10 — Sept 04

Lecturer: Aditya Gopalan — *Scribe: Prakash Chandra*

10.1 Recap

1. (Projected) Online Gradient Descent at time t :

$$y_t := w_{t-1} - \eta \bigtriangledown f_{t-1}(w_{t-1})$$
$$w_t := \prod_k y_k$$

2. Theorem:

$$Regret_T(POGD) \leq DG\sqrt{T}$$

where, $D =$ diameter of convex space K and $G =$ bound on $||gradient||_2$

3. $\sigma-$ strongly convex function:

$$f(y) \geq f(x) + \langle \bigtriangledown f(x), y - x \rangle + \frac{\sigma}{2}||y-x||^2$$

10.2 OGD with strongly convex losses

10.2.1 Logarithmic regret with time-varying learning rate

Theorem 1: Let $\{f_t\}$ be $\sigma-$ strongly convex OGD with time-varying step size: $\eta_t = 1/\sigma_t$, gives:

$$Regret_T(POGD) \leq \frac{G^2}{2\sigma}(1+\log T) \tag{10.1}$$

Notes: 1. Extra curvature make regret $O(\sqrt{T})$ to $O(\log T)$.
2. Strong convexity + bounded gradients is weaker than exp-concave.

Proof: start with the same approach as for "$DG\sqrt{T}$".
let,

$$w^* = arg \min_{w \in K} \sum_{t=1}^{T} f_t(w)$$

$$f_t(w_t) - f_t(w^*) \leq \langle \overbrace{\bigtriangledown f_t(w_t)}^{g(t)}, w_t - w^* \rangle - \frac{\sigma}{2}||w_t - w^*||^2 \tag{10.2}$$

hence,

$$\langle g(t), w_t - w^* \rangle = \frac{1}{2\eta_t}[2\eta_t \langle (t), w_t - w^* \rangle$$

$$\Rightarrow \langle g(t), w_t - w^* \rangle \leq \frac{\eta_t}{2}||g(t)||^2 \frac{||w_t - w^*||^2 - ||w_{t+1} - w^*||^2}{2\eta_t} \tag{10.3}$$

where, $\eta_t = \dfrac{1}{\sigma t}$
use (10.2) together with (10.1) and sum over $t = 1, 2, ..., T$.

$$\sum_{t=1}^{T} [(w_t) - f_t(w^*)] \leq \frac{G^2}{2} \sum_{t=1}^{T} \eta_t + \frac{1}{2} \sum_{t=1}^{T} \left[\frac{\beta_t - \beta_{t+1}}{\eta_t} - \sigma \beta_t \right]$$

where, $\beta_t = ||w_t - w^*||^2$

$$= \frac{G^2}{2} \sum_{t=1}^{T} \left(\frac{1}{\sigma t} \right) + \frac{1}{2} \sum_{t=1}^{T} \left[\left(\frac{1}{eta_t} - \sigma \right) \beta_t - \frac{1}{\eta_t} \beta_{t+1} \right]$$

$$\leq \frac{G^2}{2\sigma}(\log(T) + 1) + \frac{1}{2} \underbrace{\left(\frac{1}{\eta_1} - \sigma \right)}_{=0} \beta_1 + \sum_{t=0}^{T} \left[\frac{1}{\eta_t} - \sigma - \frac{1}{\eta_{t-1}} \right] \beta_t$$

as $\dfrac{1}{\eta_t} - \sigma - \dfrac{1}{\eta_{t-1}} = \sigma t - \sigma - \sigma(t-1) = 0$

$$= \frac{G^2}{2\sigma}(1 + \log(T))$$

10.3 Impact of regularizer on FTRL performance

1. FTRL algorithm, regularizer $= ||.||^2 \Rightarrow$ Gradient descent.
2. OGD $N-$expert gives regret $= O\sqrt{NT}$.
3. "What's a good regularizer instead of" for my problem?

$$Regret_T^{FTRL}(u) \leq R(u) - R(w_1) + \sum_{t=1}^{T} [f_t(w_t) - f_t(w_{t+1})]$$

$$\therefore \textit{controlling}\ [f_t(w_t) - f_t(w_{t+1})]$$
$$\Rightarrow \textit{control Regret (general philosophy)}$$

Definition 2: [Lipschitz continuity]
$f :\to \mathbb{R}$ is Lipschitz continuous w.r.t a norm $||.||_\square$ if

$$|f(x) - f(y)| \leq L||x-y||_\square$$

where, $x, y \in K$

10.3.1 FTRL regret bound with Lipschitz losses + strongly convex regularizer

Theorem 3: [FTRL regret w/Lipschitz losses (w.r.t $||.||_\square$) + strongly-convex regularizer]

Suppose $f_1, f_2, ...$ is such that f_t is L_t-Lipschitz continuous (w.r.t $||.||_\square$). Let the regularizer R be σ-strongly convex w.r.t the same norm $||.||_\square$.

Then, $\forall u \in K$

$$Regret_T^{FTRL}(u) \leq R(u) - \min_{v \in K} R(v) + \frac{TL^2}{\sigma} \tag{10.4}$$

Let's apply this(theorem 3) for expert advice problem W/N experts. i.e
$K = \triangle_N$, $f_t(\pi) := \langle \pi, Z_t = \textit{loss vector at time } t \in \mathbb{R}^N \textit{or} \in [0,1]^N \rangle$
use **entropic regularizer:**

$$R_\eta(w) := -\frac{1}{\eta} H(w) = \frac{1}{\eta} \sum_i w_i \log(w_i)$$

so

$$FTRL \Rightarrow Exp - weight$$

Let's first go through the following claim:
Claim 4: $R_1(w) = -H(w)$ is $\frac{1}{B}$-stronger convex over $K_B := w \in \mathbb{R}_+^N : ||w||_1 \leq B$ w.r.t $||.||_1$
Proof:
Lipschit continuity of $\{f_t\}$

$$f_t(\pi) - f_t(\Psi) = |\langle Z_t, \pi - \Psi \rangle|$$
$$\underbrace{\leq ||\pi - \Psi||_1 ||Z_t||_\infty}_{\textit{Holder's inequality}} \leq 1||\pi - \Psi||_1$$

$\Rightarrow \{f_t\}$ are 1-Lipschitz continuous w.r.t $||.||_1$
$\Rightarrow R_1$ is $\frac{1}{B}$-strongly convex.

Let's use this claim (claim 4) for entropic regularizer:
hence, $R_\eta = \frac{1}{\eta} R_1$
so R_η is $\frac{1}{\eta}$-strong convex over $\triangle_N$.
Apply theorem 3 with $L = 1$, $\sigma = \frac{1}{\eta}$ in above expert advice with N experts problem setup:

$$Regret_T^{Exp\text{-}wts(\eta)}(u) \le -\min_{\nu \in \triangle_N} R(\nu) + T\eta$$

$$= \max_{\nu \in \triangle_N} (\nu) + T\eta$$

$$\Rightarrow Regret_T^{Exp\text{-}wts(\eta)}(u) \le \frac{\log(N)}{\eta} + T\eta$$

note that if $\eta = \sqrt{\frac{\log(N)}{T}}$ then $Regret_T^{Exp\text{-}wts(\eta)}(u) \le \sqrt{T \log(N)}$.

-$O(\sqrt{T log(N)})$ is the "right scaling" for Exp-wts:
-BOTTOMLINE: choice of regularizer is important:
-Can/should depend on Lipschitz continuity of losses:
-Also depend on structure of K, i.e. how large (w.r.t$||.||_\square$)does K look.
e.g for best experts:

$$diam_{||.||}(K) \backsimeq diam_{||.||_1}(K) = O(1)$$

-But, cost functions are 1-Lip. w.r.t $||.||_1$, but $\sqrt{N}$-Lip. w.r.t $||.||_2$.

Proof of theorem 3:
$\forall t$,

$$\Phi_t(w) = \sum_{s=1}^{t-1} f_s(w) + R(w) \tag{10.5}$$

FTRL picks w_t such that:

$$w_t = arg \min_{w \in K} \Phi_t(w) \tag{10.6}$$

hence, R is σ-strongly convex w.r.t $||.||_\square$
$\Rightarrow \Phi_t$ is σ-strongly convex w.r.t $||.||_\square$
because adding σ-str-cvx + cvx is σ-str-cvx

Let's use the following lemma:

Lemma 5: if f is σ-str-convex and $x^* = \min_{x \in K}(x)$
then $f(x) - f(x^*) \geq \frac{\sigma}{2}||x - x^*||_{\Box}^2$

---

using lemma 5 and (10.6):

$$\Phi_t(w_{t+1}) - \Phi_t(w_t) \geq \frac{\sigma}{2}||w_{t+1} - w_t||_{\Box}^2 \tag{10.7}$$

$$\Phi_{t+1}(w_t) - \Phi_{t+1}(w_{t+1}) \geq \frac{\sigma}{2}||w_{t+1} - w_t||_{\Box}^2 \tag{10.8}$$

adding inequalities (10.7) and (10.8) and using (10.5):

$$f_t(w_t) - f_t(w_{t+1}) \geq \sigma||w_t - w_{t+1}||_{\Box}^2$$

$$\Rightarrow \sigma||w_t - w_{t+1}||_{\Box}^2 \leq \underbrace{f_t(w_t) - f_t(w_{t+1}) \leq L_t||w_t - w_{t+1}||_{\Box}}_{\textit{Lipschitz continuity}}$$

$$\Rightarrow ||w_t - w_{t+1}||_{\Box} \leq \frac{L_t}{\sigma} \tag{10.9}$$

and,

$$Regret_T^{FTRL}(u) \leq R(u) - R(w) + \sum_{t=1}^{T}[f_t(w_t) - f_t(w_{t+1})]$$

$$\leq R(u) - \min_{v \in K} R(v) + \sum_{t=1}^{T} L_t||w_t - w_{t+1}||_{\Box}$$

from (10.9)

$$\leq R(u) - \min_{v \in K} R(v) + \sum_{t=1}^{T} L_t \frac{L_t}{\sigma}$$

hence, $L^2 \geq \frac{1}{T}\sum_{t=1}^{T} L_t^2$

$$\Rightarrow Regret_T^{FTRL}(u) \leq R(u) - \min_{v \in K} R(v) + \frac{TL^2}{\sigma}$$

10.3.2 Different View of FTRL: "Online Mirror Descent"

Recall FTRL applied to linear costs $\{f_t\}$.
$f_t(.) = \langle Z_t, . \rangle$ over $K \subseteq \mathbb{R}^d$, regularizer *Ri*
$= arg\min_{w \in K}[\langle w_t, Z_{1:t} \rangle + R(w)]$ where : $Z_{1:t} = \sum_{i=1}^{t} Z_i$
$= arg\max_{w \in K}[\langle w_t, -Z_{1:t} \rangle - R(w)]$

let $h : \mathbb{R}^d \to k$
$h(\theta) := arg\max_{w \in K}[\langle w, \theta \rangle - R(w)]$
Called "link function" or "prox function"

$\therefore$ FTRL can equivalently written as:
$\theta_1 = 0$ and $\forall t = 1, 2, 3....$

(1) $predict : w_t = h(\theta_t)$
(2) $update : \theta_{t+1} = \theta_t - Z_t$

- If dealing with general convex $\{f_t\}$
We can feed the gradients $\bigtriangledown f_t(w_t)$'
i.e. we use linear functions: $\tilde{f}_t \equiv \langle \bigtriangledown f_t(w_t), . \rangle$
to get regret:

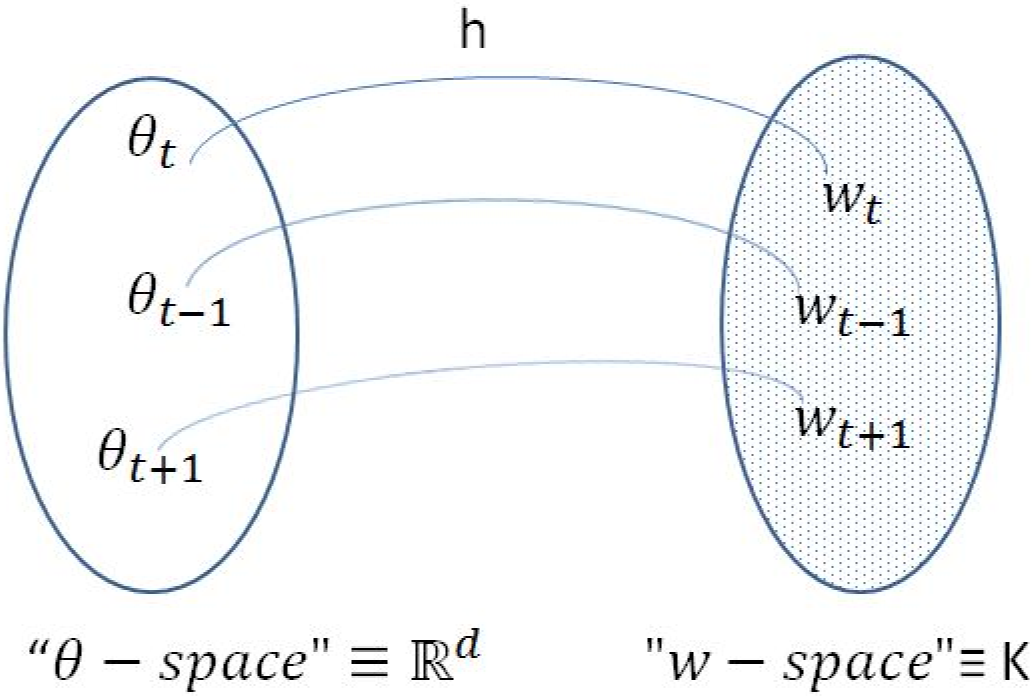
$$Regret_T(u) = \sum_{t=1}^{T}[f_t(x_t) - f_t(u)]$$

$$\leq \sum_{t=1}^{t}[\langle \bigtriangledown f_t(x_t), x_t \rangle - \langle \bigtriangledown f_t(x_t), u \rangle]$$

$$= \sum_{t=1}^{T}[\tilde{f}_t(x_t) - \tilde{f}_t(u)]$$

This generic reduction gives a general algorithm [mirror descent]:
Algorithm:
$\theta_1 = 0$
and $\forall t = 1, 2, 3....$
(1) $predict : w_t = h(\theta_t)$
(2) $update : \theta_{t+1} = \theta_t - \bigtriangledown f_t(w_t)$

Geometric Picture: ["Dual" spaces figure:]



Interpretation: θ is updated in dual space and prediction is linked/mirrored to the primal space (K).