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

Lecture 9 — September 2

Lecturer: Aditya Gopalan — *Scribe: Rahul R*

9.1 Recap

In the last lecture, we introduced the framework of online convex optimization. A simple algorithm for the convex setting called Follow-The-Leader (FTL) was introduced and its general regret bound derived. It was illustrated that the performance of FTL is strongly dependent on the curvature of the loss functions. To stabilize the FTL algorithm, we introduced a new algorithm called Regularized FTL (FTRL). FTRL is shown to give different specialized algorithms and regret performance for different choices of the regularizer function. For example, the Euclidean regularizer results in Online Gradient Descent algorithm, whereas the Entropy regularizer function results in an EXP-WTS algorithm.

In this lecture, we will obtain a general regret bound for FTRL. We will then introduce the framework of constrained optimization and introduce the Projected OGD (POGD) algorithm. It will be shown that POGD gives suboptimal regret scaling when applied to the expert selection problem. We will then introduce the concept of strongly convex functions, which will be subsequently used to obtain better regret bounds.

9.2 Generic regret bound for FTRL

Theorem 9.1. *For a regularization function, $R : K \to R$, suppose FTRL predicts the sequence of vectors $w_1, w_2, w_3, \ldots$ such that $\forall t$, $w_t = \arg\min_{w \in K} \sum_{s=1}^{t-1} f_s(w) + R(w)$ then*

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

Proof: Running FTRL on $f_1, f_2, f_3, \ldots$ is equivalent to running FTL on $f^t = R, f_1, f_2, f_3, \ldots$, then by using the FTL regret lemma,

$$
\begin{aligned}
R_T^{FTL}(u) &\leq \sum_{t=0}^{T}[f_t(w_t) - f_t(w_{t+1})] \\
\therefore \sum_{t=0}^{T}[f_t(w_t) - f_t(u)] &\leq \sum_{t=0}^{T}[f_t(w_t) - f_t(w_{t+1})] \\
R(w_0) - R(u) + \sum_{t=1}^{T}[f_t(w_t) - f_t(u)] &\leq R(w_0) - R(w_1) + \sum_{t=1}^{T}[f_t(w_t) - f_t(w_{t+1})] \\
\implies \sum_{t=1}^{T}[f_t(w_t) - f_t(u)] &\leq R(u) - R(w_1) + \sum_{t=1}^{T}[f_t(w_t) - f_t(w_{t+1})]
\end{aligned}
$$

□

9.3 Regret bounds for unconstrained Online Gradient Descent (OGD)

Suppose we run FTRL on $K = R^d$, $R_\eta(w) = \frac{\|w\|_2^2}{2\eta}$ and linear loss function, $f_t(x) = \langle x, Z_t \rangle$. Note that, this particular choice of the regularization function would result in the Online Gradient Descent rule, $w_{t+1} = w_t - \eta Z_t = w_t - \eta \bigtriangledown f_t(w_t)$. We can now apply Thm(9.1) to obtain regret bounds for OGD.

Theorem 9.2.

$$
\forall u \in R^d,\, R_T^{OGD}(u) \leq \frac{\|u\|_2^2}{2\eta} + \eta \sum_{t=1}^{T} \|Z_t\|_2^2
$$

Proof: *Using FTRL lemma (9.1),*

$$
\begin{aligned}
R_T^{OGD}(u) &\leq R(u) - R(w_1) + \sum_{t=1}^{T}[f_t(w_t) - f_t(w_{t+1})] \\
&\leq \frac{\|u\|_2^2}{2\eta} + \sum_{t=1}^{T} < w_t - w_{t+1}, Z_t >, \quad [R(w_1) > 0] \\
&\leq \frac{\|u\|_2^2}{2\eta} + \eta \sum_{t=1}^{T} \|Z_t\|_2^2
\end{aligned}
$$

□

9.4 Projected Online Gradient Descent(POGD)

9.4.1 Constrained optimization

Suppose the convex decision space, $K \subset R^d$, then the POGD method projects the solution of OGD, y_t back into the decision space K to obtain w_t, here, η is the learning parameter. This ensures that the decision vector is in the constrained set K and also since w_t is closer to any member of K than y_t, it is also closer to the optimum decision w^*. Note that the loss function, $f_t : K \to R$ is assumed to be convex and hence $\forall t, \forall u, w_t \in K$; $f_t(w_t) - f_t(u) \leq \langle \bigtriangledown f_t(w_t), w_t - u \rangle$.

Projected Online Gradient Descent Algorithm:

$$\begin{aligned} y_t &:= w_{t-1} - \eta \bigtriangledown f_{t-1}(w_{t-1}) \\ w_t &:= \Pi_k y_t \\ &:= \arg\min_{w \in K} \|y_t - w\|_2 \end{aligned}$$

Theorem 9.3. *Regret of POGD*[4]

$$\begin{aligned} R_T^{POGD(\eta)} &:= \max_{u \in K} R_T^{POGD(\eta)}(u) \\ &\leq \frac{D^2}{2\eta} + \frac{\eta}{2} T G^2 \end{aligned}$$

where, $D := \max_{x,y \in K} \|x - y\|_2$, $G := \sup_{t \leq T, x \in K} \|\bigtriangledown f_t(x)\|$.

Proof: *Let* $w^* = \arg\min_{w \in K} \sum_{t=1}^T f_t(w)$. *Then*

$$\begin{aligned} f_t(w_t) - f_t(w^*) &\leq \langle \bigtriangledown f_t(w_t), w_t - w^* \rangle \\ &= \frac{1}{2\eta} \langle 2\eta g_t, w_t - w^* \rangle \\ &= \frac{1}{2\eta} 2(w_t - y_{t+1})^T (w_t - w^*) \\ &= \frac{1}{2\eta} [\|w_t - y_{t+1}\|_2^2 + \|w_t - w^*\|_2^2 - \|w^* - y_{t+1}\|_2^2] \\ &= \frac{1}{2\eta} [\eta^2 \|g_t\|_2^2 + \|w_t - w^*\|_2^2 - \|w^* - y_{t+1}\|_2^2] \end{aligned}$$

Since, $w_{t+1} := \Pi_k y_{t+1}$, $\|y_{t+1} - w^*\| \geq \|w_{t+1} - w^*\|$,

$$
\begin{aligned}
f_t(w_t) - f_t(w^*) &\le \frac{1}{2\eta}\left[\eta^2\|g_t\|_2^2 + \|w_t - w^*\|_2^2 - \|w_{t+1} - w^*\|_2^2\right] \\
\therefore R_T^{POGD(\eta)} &= \sum_{t=1}^{T}[f_t(w_t) - f_t(w^*)] \\
&\le \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_2^2 + \frac{\|w_1 - w^*\|_2^2}{2\eta} \\
&\le \frac{\eta}{2}TG^2 + \frac{D^2}{2\eta}
\end{aligned}
$$

□

Note, setting $\eta = \frac{D}{G\sqrt{T}}$, gives, $R_T^{POGD} \le DG\sqrt{T}$.

Best expert problem

POGD can be applied to Best expert problem by setting $K = \triangle_N$ and convex loss function, $f_t(\pi) = \langle \pi, l_t \rangle$, where $l_t \in [0,1]^N$. Then, $D = \max_{x,y \in \triangle_N} \|x - y\|_2 = \sqrt{2}$ and $G = \max_{x \in \triangle_N, t \le T} \|\bigtriangledown f_t(x)\|_2 \le N$. This gives a regret bound of $R_T^{POGD} \le \sqrt{2NT}$. However, the EXP-WTS algorithm (obtained by choosing the entropy regularizer) is known to obtain a much better bound of $O(\sqrt{T\log(N)})$. Hence, using the euclidean regularizer, $\|.\|_2$ gives sub-optimal Regret.

9.5 Strongly Convex functions

A function is convex if it grows faster than a linear function everywhere. To be precise, for a convex function f, at any point w, the tangent at w does not exceed the functon, f at any point. A function f is strictly convex if, f is strictly above the tangent and the difference can be quantified as follows:

Definition (Strong Convexity)

Let K be a convex set. Then a function $f : K \to R$ is said to be strongly convex over K w.r.t a norm $\|.\|$ if, for any $w, v \in K$ and $\alpha \in [0,1]$

$$f(\alpha v + (1-\alpha)w) \le \alpha f(v) + (1-\alpha)f(w) - \frac{\sigma}{2}\alpha(1-\alpha)\|v - w\|^2$$

Equivalent definitions of strong convexity are,

$$\forall z \in \partial f(w),\;\; f(v) \ge f(w) + \langle z, v - w \rangle + \frac{\sigma}{2}\|v - w\|^2$$

If f is differentiable, then f is strongly convex iff,

$$\langle \bigtriangledown f(v) - \bigtriangledown f(w), v - w \rangle \geq \sigma \|v - w\|^2$$

Additionally, if f is twice differentiable then, a sufficient condition for strong convexity of f is

$$\forall w, x \in K,\ \langle \bigtriangledown^2 f(w)x, x \rangle \geq \sigma \|x\|^2$$

Example: (Euclidean Regularization) The function $R(w) = \frac{\|w\|_2^2}{2}$ is 1-strongly convex w.r.t to l_2 norm over R^d, since the Hessian of $R(w)$, $\bigtriangledown^2 R(w) = I$.

Example: (Entropy Regularization) The function $R(w) = \sum_{i=1}^d w(i) \log [w(i)]$ is 1-strongly convex w.r.t to l_1 norm over the probability simplex, since $\bigtriangledown^2 R(w) = diag\{\frac{1}{w(1)}, \frac{1}{w(2)}, \ldots, \frac{1}{w(d)}\}$ and

$$\begin{aligned}\langle \bigtriangledown^2 R(w)x, x \rangle &= \sum_{i=1}^d \frac{x(i)}{w(i)} \\ &= \frac{1}{\|w\|_1} (\sum_{i=1}^d w(i))(\sum_{i=1}^d \frac{x(i)}{w(i)})\end{aligned}$$

Then, by Cauchy-Schwartz inequality

$$\begin{aligned}\langle \bigtriangledown^2 R(w)x, x \rangle &\geq \frac{1}{\|w\|_1} (\sum_{i=1}^d \sqrt{w(i)} \frac{|x(i)|}{\sqrt{w(i)}})^2 \\ &= \frac{1}{\|w\|_1} \|x\|_1^2\end{aligned}$$

In particular for the probability simplex, $\|w\|_1^2 = 1$, and the result follows.

References:

[1] Shai Shalev-Shwartz, *Online Learning: Theory, Algorithms, and Applications*, Ph.D Thesis, Hebrew University, July 2007

[2] Sebastien Bubeck, *Introduction to Online Optimization (Chap 4)*, Princeton University, December 2011

[3] Shai Shalev-Shwartz, *Online Learning and Online Convex Optimization (Chap 2)*, 2011

[4] M. Zinkevich, *Online convex programming and generalized infinitesimal gradient ascent*, Twentieth International Conference on Machine Learning, 2003.