EECS598: Prediction and Learning: It's Only a Game Fall 2013

Lecture 17: FTRL and Applications of OCO

Prof. Jacob Abernethy *Scribe: Lianli Liu*

Announcements

- Sign up sheet for project discussion.
- HW2 presentation coming up.

11.1 Generic Bound of FTRL

11.1.1 Notation Switch

We will use $f_t(x)$ instead of $l_t(x)$ as the loss suffered in each round to avoid confusion with loss vectors. For example, in expert setting, we have $f_t(\underline{x}) = \underline{l}^t \cdot \underline{x}$; in portfolios,$f_t(\underline{x}) = -\log(\underline{b}^t \cdot \underline{x})$.

11.1.2 Analysis on Generic Bound of FTRL

In FTRL,

$$x_t = \arg\min_{x\in X} \sum_{s=1}^{t-1} f_s(x) + \frac{1}{\eta} R(x) \tag{11.1}$$

The generic bound is

$$\sum f_t(x_t) - \min_x \sum f_t(x) \le \frac{1}{\eta}(R(u) - R(x_1)) + \sum_{t=1}^{T} (f_t(x_t) - f_t(x_{t+1})) \tag{11.2}$$

The first term is a constant, to evaluate the generic bounds the second term needs to be studied.

By convexity,

$$f_t(x_t) - f_t(x_{t+1}) \le \nabla f_t(x_t)(x_t - x_{t+1}) \tag{*}$$

This is a variant of the standard definition of convexity,

$$f(x) - f(y) \ge \nabla f(y)(x - y) \tag{11.3}$$

We perform three different analysis on (*)

(a) By Cauchy-Schwartz inequality,

$$(*) \le \|\nabla f_t(x_t)\|_2 \|x_t - x_{t+1}\|_2 \tag{11.4}$$

When the regularized function is chosen as

$$R(x) = \frac{1}{2}\|x_1 - x\|_2^2 \tag{11.5}$$

we have

$$x_{t+1} \approx x_t - \eta \nabla f_t(x_t) \Rightarrow \|x_t - x_{t+1}\|_2 \le \|\eta \nabla f_t(x_t)\|_2 \tag{11.6}$$

thus

$$(*) \le \eta \|\nabla f_t(x_t)\|_2^2 \tag{11.7}$$

When $\|\nabla f_t(x_t)\|_2^2$ is bounded,

$$f_t(x_t) - f_t(x_{t+1}) = \mathcal{O}(\eta) \tag{11.8}$$

Therefore

$$\text{Regret} = \mathcal{O}(\frac{1}{\eta} + T\eta) \tag{11.9}$$

(b) By Hölder's inequality,

$$(*) \le \|\nabla f_t(x_t)\|_p \|x_{t+1} - x_t\|_q \tag{11.10}$$

where $\frac{1}{p} + \frac{1}{q} = 1$, $\|v\|_p$ is defined as $\|v\|_p = (\sum |v_i|^p)^{1/p}$

For expert setting, $R(x) = \sum_i x_i \log x_i$. Let $p = \infty, q = 1$

$$\|\nabla f_t(x_t)\|_\infty = \|\underline{l}^t\|_\infty = \mathcal{O}(1) \tag{11.11}$$

As $x_{t+1}^i = x_t^i \exp(-\eta l_t^i)$ (normalization term omitted for convenience)

$$\|x_t - x_{t+1}\|_1 \approx \sum_i x_t^i (1 - \exp(-\eta l_t^i)) \le \eta \sum_i x_t^i l_t^i \le \eta \tag{11.12}$$

The first inequality follows $1 - e^{-x} \le x$ and the second inequality holds as $x_t \in \triangle_n$ and $l_t \in [0,1]^n$

(c) In general, the most generic bound is given as

$$\text{Regret}_T \le \frac{1}{\eta}\Big((R(u) - R(x_1)) + \sum_{t=1}^{T} D_R(x_t, x_{t+1})\Big) \tag{11.13}$$

Bregman Divergence D_R in (11.13) stands for Bregman Divergence.

Definition Given any convex function f, the Bregman Divergence of f is defined as

$$D_f(x,y) = f(x) - f(y) - \nabla f(y)(x-y) \tag{11.14}$$

Bregman Divergence actually measures the "gap" in linear approximation, as illustrated in Fig.1

Property Bregman Divergence has the following properties

1. $D_f(x,y) \neq D_f(y,x)$.
   Equality is only true when f is quadratic, i.e.$D_f(x,y) = (x-y)^T \nabla^2 f(x-y)(x-y)$
2. $\forall x, y, D_f(x,y) \ge 0$, assuming f is convex
3. $D_f(x,y) = 0$, iff $x = y$. Holds when f is strictly convex
4. Quadratic approx of Bregman Divergence: if x is "close" to y,$D_f(x,y) \approx (x-y)^T \nabla^2 f(x-y)(x-y)$

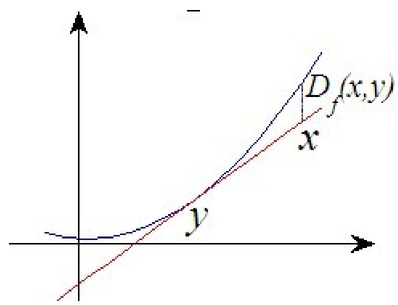

Figure 1: Bregman Divergence

11.2 Applications

11.2.1 Convex Optimization

In this application, we want to solve non-online online convex optimization problem, A.K.A Convex Optimization:

$$\min_x G(x), \text{ where } G \text{ is a convex function}$$

This problem can be reduced to OCO (Online Convex Opitimization),i.e. select a sequence of x_t's online.

Define $f_t(x) = \nabla G(x_t)(x - x_t) + G(x_t)$, we have the following observation

Observation 11.1. *By definition,* $f_t(x_t) = G(x_t)$

Observation 11.2. *By convexity,* $f_t(x) \leq G(x)$

Let ε_T be a bound on $\frac{\text{Regret}_T}{T}$, we would like to evaluate how "optimized" is $\frac{1}{T}\sum_{t=1}^{T} x_t =: \overline{x}_T$.

Denote x^* as the minimizer of G,by Jenson's Inequality

$$\begin{aligned} G(\overline{x}_T) &\leq \frac{1}{T}\sum G(x_t) \\ &= \frac{1}{T}\sum f_t(x_t) \\ &\leq \frac{1}{T}\sum f_t(x^*) + \varepsilon_T \\ &\leq \frac{1}{T}\sum G(x^*) + \varepsilon_T \\ &= G(x^*) + \varepsilon_T \end{aligned} \tag{11.15}$$

Notice:Maybe we do not want to apply FTRL to select x_t as it requires solving a minimization problem each round.

Instead,we may apply Online Gradient Descent(OGD) which requires $\mathcal{O}(dim)$ calculations each round, i.e.

$$x_{t+1} = x_t - \eta \nabla G(x_t) \tag{11.16}$$

11.2.2 Statistical Learning

Problem Statement

A canonical problem in statistical learning usually involves a Data Space Z, a Label Space Y, a Hypothesis Space H and a loss function $l : H \times Z \times Y \to \mathbb{R}$.

For $w \in H, (z, y) \in Z \times Y$,the loss of hypothesis w on(z, y) is denoted as $l(w, (z, y))$. In linear regression problem, where $w \in \mathbb{R}^n, z \in \mathbb{R}^n, y \in \mathbb{R}$, the loss function is the square error,i.e. $l(w, (z, y)) = (w \cdot z - y)^2$ Typically, H is assumed to be convex and $l()$ is convex in H.For simplicity, denote loss function as $l(w, z)$ where z contains both observation and label.

In statistical learning, the distribution D over $Z \times Y$ is unknown. We have access to i.i.d samples $z_1, \cdots z_n$ and the goal is to choose hypothesis w to minimize the risk of w, defined as

$$r(w) = \mathbb{E}_{z \sim D}[l(w, z)] \tag{11.17}$$

In general, we cannot compute $r(w)$ as the distribution D is unknown. A learning algorithm will solve the following optimization problem, known as Empirical Risk Minimization:

$$\hat{w}_n := \arg\min{}_{w \in H} \frac{1}{n} \sum_{t=1}^{n} l(w, z_t) \tag{11.18}$$

Define Bayes Risk as

$$\min{}_{w \in H} r(w) = r^* \tag{11.19}$$

Typical results in learning theory make the following statements

- $r(\hat{w}_n) \to r^*$, as $n \to \infty$
- $\hat{w}_n \to w^*$, as $n \to \infty$ (Consistency Statement)

Problem Solution: Online to Batch Conversion

Define $f_t(w) = l(w, z_t)$

Apply OCO to the sequence of samples, we will receive a sequence of w_t's.

Define $\bar{w}_n = \frac{1}{n} \sum_{t=1}^{n} w_t$. Let $\varepsilon_n = \frac{\text{Regret}_n}{n}$, we may analyze

$$\mathbb{E}_{d_{1 \cdots j}} r(\bar{w}_n) = \mathbb{E}_{d_{1 \cdots j}} \mathbb{E}_{z \sim D}[l(\bar{w}_n, z)]$$

where $d_{1\cdots j} = \{z_1, \cdots, z_j\}$. By Jenson's Inequality

$$
\begin{aligned}
\mathbb{E}_{d_{1\cdots j}} \mathbb{E}_{z\sim D}[l(\bar{w}_n, z)] &\leq \mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} l(w_t, z)] \\
&\leq \mathbb{E}_{d_{1\cdots n}\sim D}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{z\sim D}[l(w_t, z)|d_{1\cdots t-1}]] \\
&= \mathbb{E}_{d_{1\cdots n}\sim D}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{z_t} l(w_t, z_t)|d_{1\cdots t-1}] \\
&= \mathbb{E}_{d_{1\cdots n}\sim D}[\frac{1}{n}\sum_{t=1}^{n} l(w_t, z_t)] \\
&\leq \mathbb{E}_{d_{1\cdots n}\sim D}[\frac{1}{n}\sum_{t=1}^{n} l(u, z_t)] + \varepsilon_n \quad \text{(as } \varepsilon_n \text{ is the regret bound)} \\
&= \mathbb{E}_{z\sim D}[l(u, z)] + \varepsilon_n = r(u) + \varepsilon_n
\end{aligned}
\tag{11.20}
$$

The first equality in (11.20)holds by tower rule, since z and z_t have the same distribution on history $z_1 \cdots z_{t-1}$. The above inequality holds for any u, thus gives the bound on risk.