CSE599s, Spring 2014, Online Learning Lecture 99 - 05/08/2014

Adaptive Regret Bound

Lecturer: Brendan McMahan *Scribe: Tianqi Chen*

1 Recap

A normal regret bound for fixed learning rate is as follows

$$Regret(u) \le \frac{1}{2\eta}\|u\|^2 + \eta \sum_{t=1}^{T} \|g_t\|_2^2 \tag{1}$$

We should note that the regret depends on choice of η. If we do not know T in advance, this bound could be very bad. Even if we know T and set η properly, we need to wait until we get T to get the regret bound. Ideally, we want our bound to hold for any T, this is where we need to introduce adaptive update.

Our goal is prove a bound in the following style.

$$Regret \le B\sqrt{\sum_{t=1}^{T} \|g_t\|^2} << GB\sqrt{T} \tag{2}$$

We want to have a bound that does not need guess and double trick.

There are several class of related algorithms, that we will be discussed in a general framework OGD/Mirro Descent:

$$\hat{w}_{t+1} = \hat{w}_t - \eta_t g_t = \operatorname*{argmin}_w g_t w + \frac{1}{2\eta}\|w - w_t\|^2. \tag{3}$$

FTRL-Proximal

$$w_{t+1} = \operatorname*{argmin}_w f_{1:t}(w) + \sum_{s=1}^{t} \frac{\sigma_s}{2}\|w - w_s\|^2. \tag{4}$$

Dual-Averaging

$$w_{t+1} = \operatorname*{argmin}_w f_{1:t}(w) + \frac{\sigma_{1:t}}{2}\|w\|^2. \tag{5}$$

They are equivalent when we have no constraint. FTRL Proximal and dual averaging are equivalent when learning rate is constant.

2 General Framework for Adaptive Update

In this lecture, we will study the update rule in the following form:

$$w_{t+1} = \operatorname*{argmin}_w f_{1:t}(w) + r_{0:t}(w) = \operatorname*{argmin}_w h_{0:t}(w) \tag{6}$$

Note that we have $h_0(w) = r_0(w)$. Base on this update rule, we have a strong FTRL Lemma as follows

Lemma 1. *Strong FTRL Lemma*

$$Regret(u) \le r_{0:T}(u) + \sum_{t=1}^{T}[h_{0:t}(w_t) - h_{0:t}(w_{t+1}) - r_t(w_t)] \tag{7}$$

Proof. The bound can be proved by induction(see previous lecture note) □

Before we prove the main theorem, we will need the following lemma

Lemma 2. *Let*

$$w_1 = \operatorname*{argmin}_w \phi_1,$$

$$w_2 = \operatorname*{argmin}_w \phi_2 = \operatorname*{argmin}_w [\phi_1(w) + \psi(w)],$$

where ϕ_1 is 1 strongly convex function with respect to norm $\|.\|$, and $\psi(w)$ is convex function.

Let $b \in \partial\psi(w)$, then we will have

$$\phi_2(w_1) - \phi_2(w_2) \le \frac{1}{2}\|b\|_*,$$

$$\|w_1 - w_2\| \le \|b\|_*.$$

We can verify that for a special case where ϕ_1 is quadratic function, and ψ is linear: $\phi_1(w) = \frac{1}{2}\|w\|^2$, $\phi(w) = bw$, this bound is tight.

Theorem 3. *Assuming $r_t(w) \ge 0, r_t(w_t) = 0$, $h_{0:t}(w)$ is 1 strongly convex with respect to $\|.\|_t$. Then the regret of general framework can be bounded by*

$$Regret(u) \le r_{0:T}(u) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t,*)}. \tag{8}$$

Proof. For fixed round t, Let

$$\phi_1(w) = f_{1:t-1}(w) + r_{1:t-1}(w) + r_t(w) = h_{0:t-1}(w) + r_t(w_t)$$

Note that $w_t = \operatorname{argmin}_w r_t(w_t)$, we have $w_t = \operatorname{argmin}_w \phi_1(w)$. Let $\psi = f_t$, and $b = g_t, g_t \in \partial f_t(w)$. The following inequality holds follows because of Lemma **??**.

$$\begin{aligned} h_{0:t}(w_t) - h_{0:t}(w_{t+1}) &= \phi_1(w_t) + f_t(w_t) - \phi_1(w_{t+1}) - f_t(w_{t+1}) \\ &= \phi_2(w_t) - \phi_2(w_{t+1}) \le \frac{1}{2}\|g_t\|^2_{(t,*)}. \end{aligned} \tag{9}$$

Then the results follows by Lemma **??**. □

Now we need to make use Theorem **??** to analyze FTRL-Proximal algorithm. A first simple fact is that if r_t is σ_t strongly convex with respect to $\|.\|$, then $r_{0:t}$ is 1 strongly convex with respect to $\|u\|_t = \sqrt{\sigma_{1:t}}\|.\|$.

For FTRL-Proximal, we have

- $r_0(w) = I_W(w)$
- $r_t = \frac{\sigma_t}{2}\|w - w_t\|^2$, note $\eta_t = \frac{1}{\sigma_{1:t}}$
- $\|g\|_{t,*} = \frac{1}{\sqrt{\sigma_{1:t}}}\|g\|_2$

Applying Theorem **??**, we can get the following bound for adaptive learning rate.

$$Regret(u) \le \frac{(2B)^2}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \eta_t\|g_t\|^2 \tag{10}$$

We still need to decide how we can choose η_t, an important bound that we will use, is stated by following Lemma

Lemma 4. *For sequence $a_1, a_2, \cdots, a_n$, $a_i \geq 0$ the following inequality holds.*

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i} \tag{11}$$

Proof. Let $x_i = \sum_{j=1}^{i} a_j$, $x_0 = 0$, first note the integral equality

$$\int_0^{x_n} \frac{1}{\sqrt{z}} dz = 2\sqrt{x_n} - 2\sqrt{0} \tag{12}$$

This is because $2\partial_x \sqrt{x} = \frac{1}{\sqrt{x}}$. Then we can think how we can "compute" the integral in the left side numerically. We can first discretize the interval into small pieces of length $a_1, a_2, a_3 \cdots$, then take the right end of the function to approximate the function value in that interval. Note that the right end of function $\frac{1}{\sqrt{z}}$ is smaller than the functions in the interval, we can get a lower bound of integral:

$$\int_0^{x_n} \frac{1}{\sqrt{z}} dz = \sum_{i=0}^{n-1} \int_{x_{i+1}}^{x_i} \frac{1}{\sqrt{z}} dz \geq \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{\sqrt{x_{i+1}}} = \sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \tag{13}$$

□

As a special case (take $a_i = 1$), we have $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$. If we choose $\eta_t = \frac{\sqrt{2}B}{G\sqrt{t}}$, we have

$$Regret(u) \leq \frac{(2B)^2}{2\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \frac{\sqrt{2}B}{G\sqrt{t}} G^2 \leq 2\sqrt{2}GB\sqrt{T} \tag{14}$$

We can also let $a_i = \|g_t\|^2$, $\eta_t = \frac{\alpha}{\sqrt{\sum_{s=1}^{t} \|g_s\|^2}}$

$$\frac{1}{2} \sum_{t=1}^{T} \eta_t \|g_t\|^2 \leq \alpha \sqrt{\sum_{t=1}^{T} \|g_t\|^2} \tag{15}$$

The adaptive regret bound is given by

$$Regret(u) \leq \frac{(2B)^2}{2\alpha} \sqrt{\sum_{t=1}^{T} \|g_t\|^2} + \alpha \sqrt{\sum_{t=1}^{T} \|g_t\|^2} = \left(\frac{2B^2}{\alpha} + \alpha \right) \sqrt{\sum_{t=1}^{T} \|g_t\|^2} \tag{16}$$