CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Maxwell Allman, Faidra Monachou

Lecture 15
November 12, 2018

1 Review and Overview

In the previous lecture, we introduced the following "Follow the Leader" algorithm:

"Follow The Leader" (FTL) algorithm
On each iteration $t = 1, \dots, T$, we select

$$w_t = \arg\min_{w \in \Omega} \sum_{i=1}^{t-1} f_i(w), \tag{1}$$

where $\sum_{i=1}^{t-1} f_i(w)$ is the sum of previous losses (up to the previous iteration $t-1$).

As we showed with an example for $N = 2$, FTL can perform very poorly, getting the worst possible regret. In today's lecture, our goal is to fix FTL.

2 "Be The Leader" Algorithm

To build some intuition about the next algorithm, we start with a "cheating" solution.

"Be The Leader" (BTL) algorithm
In the "Be The Leader" algorithm, we still find $w_t = \arg\min_{w\in\Omega} \sum_{i=1}^{t-1} f_i(w)$ for each time t but we now play w_{t+1} at iteration t.

The following Lemma shows that, if we were able to cheat and play w_{t+1} at time t, we would end up with zero regret.

Lemma 1 (BTL)**.** *For the regret of the BTL algorithm, it holds that*

$$\sum_{t=1}^{T} f_t(w_{t+1}) - \min_{w\in\Omega} \sum_{t=1}^{T} f_t(w) \leq 0. \tag{2}$$

Proof. First, note that $w_{T+1} = \arg\min \sum_{t=1}^{T} f_t(w)$. Hence, we can expand the sums in (2) and write them as follows

$$\sum_{t=1}^{T} f_t(w_{t+1}) - \min_{w\in\Omega} \sum_{t=1}^{T} f_t(w) = f_1(w_2) + \dots + \cancel{f_T(w_{T+1})} - (f_1(w_{T+1}) + \dots + \cancel{f_T(w_{T+1})}) =$$
$$= f_1(w_2) + \dots + f_{T-1}(w_T) - (f_1(w_{T+1}) + \dots + f_{T-1}(w_{T+1}))$$

By definition, we have that $w_{T+1} = \arg\min \sum_{t=1}^{T} f_t(w)$, thus we get the following inequality by replacing w_{T+1} by w_T:

$$\sum_{t=1}^{T} f_t(w_{t+1}) - \min_{w\in\Omega}\sum_{t=1}^{T} f_t(w) \le f_1(w_2) + \ldots + \cancel{f_{T-1}(w_T)} - (f_1(w_T) + \ldots + \cancel{f_{T-1}(w_T)})$$
$$\le f_1(w_2) + \ldots + f_{T-2}(w_{T-1}) - (f_1(w_T) + \ldots + f_{T-2}(w_T))$$

By recursively repeating the same argument, we eventually get that

$$\sum_{t=1}^{T} f_t(w_{t+1}) - \min_{w\in\Omega}\sum_{t=1}^{T} f_t(w) \le f_1(w_2) - f_1(w_3) \le 0.$$

□

By the BTL Lemma, we can write the regret as

$$R = \sum_{t=1}^{T} f_t(w_t) - \min_{w\in\Omega}\sum_{t=1}^{T} f_t(w) \le \sum_{t=1}^{T}(f_t(w_t) - f_t(w_{t+1})),$$

where each term $f_t(w_t) - f_t(w_{t+1})$ captures the stability of the algorithm. Thus, if we have stability, we can achieve better regret; for that reason, in the two expert problem that we saw in the previous lecture, we had larger regret.

3 "Follow The Regularized Leader" Algorithm

In this section, we introduce and study the properties of the "Follow the Regularized Leader" (FTRL) algorithm.

"Follow The Regularized Leader" (FTRL) algorithm
On each iteration $t = 1, \ldots, T$, we select

$$w_t = \arg\min_{w\in\Omega}\sum_{i=1}^{t-1} f_i(w) + \frac{1}{\eta}\phi(w), \tag{3}$$

where $\phi(\cdot)$ is the regularizer such that $\phi(w)$ is 1-strongly convex. We will define η later.

An important property that we will need in the analysis of the FTRL algorithm is that $\phi(w)$ is 1-strongly convex. Before we introduce the definition of a-strong convexity, note that convexity implies that

$$\forall x, y,\ \ f(x) - f(y) \ge \langle \nabla f(y), x - y\rangle.$$

We expand this property to define the notion of a-stong convexity.

Definition 2. *We say that the function $f : \Omega \to \mathbb{R}$ is α-strongly convex if*

$$\forall x, y,\ \ f(x) - f(y) \ge \langle \nabla f(y), x - y\rangle + \frac{\alpha}{2}\|x - y\|_2^2. \tag{4}$$

Remark. Using the second-order Taylor expansion, one can interpret this definition as follows:

$$f(x) - f(y) \simeq \langle \nabla f(y), x - y\rangle + \langle x - y, \nabla^2 f(y)(x - y)\rangle + \ldots$$

The following notation will be helpful for the application of the upcoming lemma. Let

$$F(w) \triangleq \sum_{i=1}^{t-1} f_i(w) + \frac{1}{\eta}\phi(w)$$

and

$$G(w) \triangleq \sum_{i=1}^{t} f_i(w) + \frac{1}{\eta}\phi(w) = F(w) + f_t(w).$$

Then it follows that

$$w_t = \arg\min_{\omega \in \Omega} F(w)$$

and

$$w_{t+1} = \arg\min_{\omega \in \Omega} G(w).$$

Lemma 3. *Suppose F is α-strongly convex, f is convex, and let*

$$w = \arg\min_{z} F(z)$$

and

$$w' = \arg\min_{z} G(z).$$

Then,

$$0 \leq f(w) - f(w') \leq \frac{1}{\alpha}||\nabla f(w)||_2^2.$$

Proof. Since F is α-strongly convex,

$$F(w') - F(w) \geq \langle \nabla F(w), w' - w \rangle + \frac{\alpha}{2}||w - w'||_2^2.$$

By the optimality of w, it follows from convex analysis that $\langle \nabla F(w), w' - w \rangle \geq 0$, so

$$F(w') - F(w) \geq \frac{\alpha}{2}||w - w'||_2^2. \tag{5}$$

Similarly, we get

$$G(w') - G(w) \geq \frac{\alpha}{2}||w - w'||_2^2. \tag{6}$$

Combining (5) and (6) gives

$$f(w) - f(w') \geq \alpha||w - w'||_2^2 \geq 0 \tag{7}$$

and

$$\begin{aligned} f(w) - f(w') &\leq |\langle \nabla f(w), w - w' \rangle| \\ &\leq ||\nabla f(w)||_2 ||w - w'||_2 \\ &\leq ||\nabla f(w)||_2 \cdot \sqrt{\frac{1}{\alpha}(f(w) - f(w')} \end{aligned} \tag{8}$$

$$\implies f(w) - f(w') \leq \frac{1}{\alpha}||\nabla f(w)||_2^2$$

where the second inequality in (8) follows by Cauchy-Schwartz inequality. □

This lemma can be generalized to arbitrary norms, but we first need a definition.

Definition 4. *F is α-strongly convex w.r.t. norm $||\cdot||$ on Ω if $\forall x, y \in \Omega$,*

$$f(x) - f(y) \geq \langle \nabla f(x), x - y \rangle + \frac{\alpha}{2}||x - y||^2.$$

Lemma 5. *Suppose F is α-strongly convex, f is convex, and let*

$$w = \arg\min_z F(z)$$

and

$$w' = \arg\min_z G(z).$$

Then,

$$0 \leq f(w) - f(w') \leq \frac{1}{\alpha}||\nabla f(w)||_*^2,$$

where $||\cdot||_*$ is the dual norm of $||\cdot||$.

Proof. The proof follows analogously to the proof of lemma 3. □

We can now bound the regret of the FTRL algorithm.

Theorem 6 (Regret bound of FTRL)**.** *Suppose ϕ is 1-strongly convex w.r.t. $||\cdot||$. Then, the regret R of the FTRL algorithm* (1) *is bounded by*

$$R \leq \frac{D}{\eta} + \eta \sum_{t=1}^{T} ||\nabla f_t(w_t)||_*^2,$$

where $D = \max_{x\in\Omega} \phi(x) - \min_{x\in\Omega} \phi(x)$.

In addition, if $||\nabla f_t(w)||_* \leq G$ for all w and f_t, then taking

$$\eta = \sqrt{\frac{D}{TG^2}}$$

gives

$$R \leq O(G\sqrt{TD}).$$

Proof. Let

$$f_0(w) = \frac{\phi(w)}{\eta}, \quad w_t = \arg\min_{w\in\Omega} \sum_{i=0}^{t-1} f_i(w)$$

Then, using the BTL lemma,

$$\sum_{t=0}^{T} f_t(w_t) - \arg\min_{w\in\Omega} \sum_{t=0}^{T} f_i(w) \leq \sum_{t=0}^{T} (f_t(w_t) - f_t(w_{t+1}).$$

Thus, letting $w^* = \arg\min_{w\in\Omega} \sum_{t=1}^{T} f_t(w)$,

$$\begin{aligned}\sum_{t=0}^{T} f_t(w_t) - \arg\min_{w} \sum_{t=0}^{T} f_t(w) &\geq f_0(w_0) + \sum_{t=1}^{T} f_t(w_t) - \sum_{t=0}^{T} f_t(w^*) \\ &= \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*) + f_0(w_0) - f_0(w^*).\end{aligned}$$

so

$$\sum_{t=0}^{T}(f_t(w_t) - f_t(w_{t+1})) = f_0(w_0) - f_0(w_1) + \sum_{t=1}^{T}(f_t(w_t) - f_t(w_{t+1})).$$

By Lemma 5 with $F = \sum_{i=0}^{t-1} f_i$, $G = \sum_{i=0}^{t} f_i$, $f = f_t$ and $\alpha = \frac{1}{\eta}$, we get that

$$f_t(w_t) - f_t(w_{t+1}) \leq \eta \sum_{t=1}^{T} ||\nabla f_t(w_t)||_*^2.$$

Hence,

$$\begin{aligned} R &\leq f_0(w^*) - f_0(w_1) + \eta \sum_{t=0}^{T} ||\nabla f_t(w_t)||_*^2 \\ &\leq \frac{D}{\eta} + \eta \sum_{t=0}^{T} ||\nabla f_t(w_t)||_*^2. \end{aligned}$$

If $||\nabla f_t(w_t)||_*^2 \leq G$ then $R \leq \frac{D}{\eta} + \eta T G^2$. Setting $\eta = \sqrt{\frac{D}{TG^2}}$ gives $R \leq 2G\sqrt{TD}$. $\square$