CS 294-128: Algorithms and Uncertainty	**Lecture 13**	**Date: Oct 12, 2016**
Instructor: Nikhil Bansal		**Scribe: Daniel Li**

1 Introduction

In this set of lecture notes we will go over the α-strongly convex proof, show a lower bound for gradient descent, examine regularization, and discuss a new group of algorithms to tackle the online convex optimization setting. Such algorithms include follow the leader (FTL), be the leader (BTL), and follow the regularized leader (FTRL).

2 Gradient Descent in α-strongly convex case

A function $f : X \to \mathbf{R}$ is α-strongly convex if:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

2.1 Proof

Theorem 1 *If we have a function f that is α-strongly convex and we set our learning rate for Gradient Descent to be $\eta_t = \frac{1}{\alpha t}$, then our regret is at most $O(\frac{G^2}{2\alpha}\ln T)$, where G is the bound on the gradient.*

Proof: Let $\mathbf{y}$ denote the fixed optimum solution in hindsight.

As f is α-strongly convex:

$$f_t(\mathbf{y}) \geq f_t(\mathbf{x})_t + \nabla f_t(\mathbf{x}_t)^T(\mathbf{y} - \mathbf{x}_t) + \frac{\alpha}{2}||\mathbf{y} - \mathbf{x}_t||_2^2$$

and hence

$$f_t(\mathbf{x}_t) - f_t(\mathbf{y}) \leq \nabla f_t(\mathbf{x}_t)^T(\mathbf{x}_t - \mathbf{y}) - \frac{\alpha}{2}||\mathbf{x}_t - \mathbf{y}||_2^2 \tag{1}$$

Consider the following potential function:

$$\Phi(t) = \frac{1}{2\eta_{t-1}}||\mathbf{x}_t - \mathbf{y}||_2^2$$

The change in potential is defined as the following:

$$\Phi(t+1) - \Phi(t) = \frac{1}{2\eta_t}||\mathbf{x}_{t+1} - \mathbf{y}||_2^2 - \frac{1}{2\eta_{t-1}}||\mathbf{x}_t - \mathbf{y}||_2^2$$

Recall from projecting onto a convex set K: $\mathbf{x}_{t+1} = \Pi_K(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t))$. We use the notation $\nabla_t := \nabla f_t(x_t)$ and plugging the previous in we get

$$
\begin{aligned}
\Phi(t+1)-\Phi(t) &\le \frac{1}{2\eta_t}||\mathbf{x}_t-\mathbf{y}-\eta_t\nabla_t||_2^2-\frac{1}{2\eta_t}||\mathbf{x}_t-\mathbf{y}||_2^2 \\
&\le \frac{1}{2\eta_t}(||\mathbf{x}_t-\mathbf{y}||_2^2+\eta_t^2\nabla_t^2-2\eta_t\nabla_t^T(\mathbf{x}_t-\mathbf{y}))-\frac{1}{2\eta_{t-1}}||\mathbf{x}_t-\mathbf{y}||_2^2 \\
&\le \frac{\alpha}{2}||\mathbf{x}_t-\mathbf{y}||_2^2+\frac{\eta_t}{2}\|\nabla_t\|^2-\nabla_t^T(\mathbf{x}_t-\mathbf{y})
\end{aligned} \tag{2}
$$

Note, the first line to the second line follows from expanding out the ℓ_2 norm. We also use $1/\eta_t = \alpha t$.

Combining (1) and (2) we get:

$$
f_t(\mathbf{x}_t)-f_t(\mathbf{y})+\Phi(t+1)-\Phi(t) \le \frac{\eta_t}{2}||\nabla_t||_2^2
$$

Summing over time t,

$$
\sum_t^T(f_t(\mathbf{x}_t)-f_t(\mathbf{y})+\Phi(T+1)-\Phi(0)) \le \sum_{t=1}^T(\frac{\eta_t}{2}||\nabla_t||_2^2) \le \sum_t \frac{1}{(2\alpha t)}G^2 \le G^2\ln T/2\alpha
$$

Summing over t, the first term gives regret. Then we note that $\Phi(T+1)\ge 0$ and $\Phi(0)=0$. □

Theorem 2 *The bound $\Omega(DG\sqrt{T})$ is a tight bound for any algorithm for online convex optimization.*

Proof: We define the following:

1. $K :=$ hypercube where $\mathbf{x} = \{\pm 1\}^n$ vertices
2. $f_v(\mathbf{x}) = \mathbf{v}^T\mathbf{x}$ where $\mathbf{v} = [-1, 1, -1, 1...(\pm 1)^n]$. In other words, we have 2^n linear cost functions,one for each vertex in $\mathbf{v}$.
3. $D \le 2\sqrt{n}$ where D is the diameter. A sketch of this is as follows:

$$
D = \sup\{d(x,y) : x, y \in K\}
$$

$$
D \le \|\mathbf{x}-\mathbf{y}\|_2^2 \text{ where } \mathbf{x} = \{1\}^n, \mathbf{y} = \{-1\}^n
$$

$$
D \le \sqrt{\sum_{i=1}^n 2^2} = 2\sqrt{n}
$$

4. $G \le \sqrt{n}$ where G is the norm of the cost function gradients.

$$
G \le \sqrt{\sum_i^n(\pm 1)^2} = \sqrt{n}
$$

At each time step t, the adversary gives the function $f_t = f_{v_t}$ where v_t is picked uniformly at random. As the function is random at each step, and $E_v[f_v(x) = 0]$, the online algorithm has zero expected regret, no matter what it does.

We claim that the offline cost is less than $-cn\sqrt{T}$ for some constant c in expectation. This follows as if we consider the overall cost function $F = \sum_t f_t$, then in each coordinate i, it is just a sum of T random $\pm$ variables, there is constant probability that it is more than $c\sqrt{T}$ or less than $-c\sqrt{T}$ for some c. So the adversary can pick signs appropriately for each coordinate (and hence a vertex on the hypercube), so that the cost is at most $-c'n\sqrt{T}$, and hence the regret is $\Omega(DG\sqrt{T})$.

□

3 Follow the Leader, Be the Leader, and Regularization

3.1 Follow the Leader

Follow the Leader (FTL) is an algorithm that at each steps mimics the best offline solution. If the game were to end at time $t-1$, the offline would be at $argmin_x \sum_{s=1}^{t-1} f_s(x)$. So this is what online sets x_t to be.

3.1.1 Explanation

However, this procedure can be arbitrarily bad. Consider the following example. If we have a set $K : \{-1, .., 1\}$ and at time step $t = 1$, a function $f_1 \leftarrow (x/2)$, the logical choice would be to choose -1. However at time step $t = 2$, we have $f_2 \leftarrow -x$ it makes sense to choose 1. Now we have an online cost of at least T and offline cost of 0. This can be seen as "over optimizing". As a solution we introduce the concept of regularization were we add some $R(\mathbf{x})$ term to "regularize" and prevent too much changing to our function.

As a thought experiment, we pose the following algorithm called Be The Leader (BTL). It is a hypothetical algorithm assuming that the algorithm could see one time step in the future. But it has an interesting guarantee.

3.2 Be The Leader

As previously, let x_{t+1} be what FTL would play at time $t+1$.

$$x_{t+1} = \operatorname*{argmin}_{\mathbf{x} \in K} \sum_{i=1}^{t} f_i(x)$$

BTL plays x_{t+1} at time t.

Theorem 3

$$\sum_{t=1}^{T} f_t(x_{t+1}) \leq \sum_{t=1}^{T} f_t(\boldsymbol{u}) \quad \forall \boldsymbol{u} \in K$$

What this theorem is intuitively saying is that we have a lower bound on the cost of any fixed static optimum.

Proof: We do this through a proof by induction. Assume the following expression is true for $T-1$.

$$\sum_{t=1}^{T-1} f_t(\mathbf{x}_{t+1}) \leq \sum_{t=1}^{T-1} f_t(\mathbf{u}) \quad \forall u \in K$$

Now we set $\mathbf{u}$ to be $\mathbf{x}_{T+1}$. Now if add $f_T(\mathbf{x}_{T+1})$ to both sides we get the following:

$$\sum_{t=1}^{T-1} f_t(\mathbf{x}_{t+1}) + f_T(\mathbf{x}_{T+1}) \leq \sum_{t=1}^{T-1} f_t(\mathbf{x}_{T+1}) + f_T(\mathbf{x}_{T+1})$$

Now, the lhs is $\sum_{t=1}^{T} f_t(\mathbf{x}_{t+1})$. The rhs becomes $\sum_{t=1}^{T} f_t(\mathbf{x}_{T+1})$, but as x_{T+1} is the minimizer of $\sum_{t=1}^{T} f_t$, the rhs is at most $\sum_{t=1}^{T} f_t(u)$ for any $u \in K$. So,

$$\sum_{t=1}^{T} f_t(\mathbf{x}_{t+1}) \leq \sum_{t=1}^{T} f_t(\mathbf{u}) \quad \forall \mathbf{u} \in K$$

□

Now we state the significance of this theorem by providing a lower bound to our regret. As

$$\text{Regret } = \sum_t (f_t(\mathbf{x}_t) - \operatorname*{argmin}_{\mathbf{u}} \sum_t f_t(\mathbf{u}))$$

Because we have just showed:

$$\sum_{t=1}^{T} f_t(\mathbf{u}) \geq \sum_{t=1}^{T} f_t(\mathbf{x}_{t+1})$$

Which implies:

$$\text{Regret } \leq \sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1})$$

Now we move onto Follow The Regularized Leader (FTRL)

3.3 Follow the Regularized Leader

3.3.1 Introduction

The idea of adding a strongly convex regularizing term is to prevent excessive oscillating when we optimize.

3.3.2 Algorithm

We assume linear functions and adopt the convention $\nabla_i = \nabla f_i(\mathbf{x}_i)$. FTRL is defined as the following procedure:

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x} \in K} \ \eta(\nabla_1 + ... + \nabla_t)\mathbf{x} + R(\mathbf{x})$$

Theorem 4 *FTRL's regret is bounded by the following expression where* $\boldsymbol{y}$ *is the optimal solution in hindsight and* $\|\cdot\|_*$ *is the dual norm and* $R(x)$ *is* α*-strongly convex w.r.t. a norm* $\|\cdot\|$.

$$\textit{Regret} \leq \sum_t \frac{2\eta}{\alpha}\|\nabla_t\|_*^2 + \frac{R(\boldsymbol{y}) - R(\boldsymbol{x}_0)}{\eta}$$

Proof: Consider the following fake game as a thought experiment. At $t = 0$ we have the following function

$$g_0(x) = \frac{R(x)}{\eta} \quad \text{and} \quad g_t(\mathbf{x}) = \nabla_t^T\mathbf{x} \quad \forall t : t \geq 1$$

Thus by the previous discussion on FTL and BTL, we have the following bound on regret for FTL wrt costs g:

$$\text{Regret (FTL):} \leq \sum_{t=0}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1})$$

$$\sum_{t=0}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{u}) \leq \sum_{t=0}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1})$$

Now in the real game, FTRL does the same moves as above, and regret of FTRL that the RHS is bounded by the following

$$\text{Regret(FTRL)} \leq \sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(u) = \sum_{t=1}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{u})$$

Note that the summation is from $t = 1$, instead of $t = 0$ previously. But we have

$$\sum_{t=1}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{u}) \leq \sum_{t=0}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1}) + \frac{1}{\eta}(R(\mathbf{x}_0) - R(\mathbf{u}))$$

So, we focus on bounding $\sum_{t=1}^{T} g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1})$. By definition of g_t we have for $t \geq 1$,

$$\begin{aligned} g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1}) &= \nabla_t^T f_t(\mathbf{x}_t)^T(\mathbf{x}_t - \mathbf{x}_{t+1}) \\ &= \frac{(\nabla_t^T(\mathbf{x}_t - \mathbf{x}_{x+1}))^2}{\nabla_t^T(\mathbf{x}_t - \mathbf{x}_{x+1})} \end{aligned} \tag{3}$$

Let Φ_t denote the function that FTRL is minimizing (the symbol Φ should not be confused with any potential function here).

$$\Phi_t(x) := \eta(\nabla_1 + ... + \nabla_t)\mathbf{x} + R(\mathbf{x})$$

As $\Phi_t = \Phi_{t-1} + \eta\nabla_t$,

$$\Phi_t(\mathbf{x}_t) - \Phi_t(\mathbf{x}_{t+1}) = \Phi_{t-1}(\mathbf{x}_t) - \eta\nabla_t^T\mathbf{x}_t - \Phi_{t-1}(\mathbf{x}_{t+1}) - \eta\nabla_t^T\mathbf{x}_{t+1}.$$

As x_t is the minimizer of Φ_{t-1}, we have $\Phi_{t-1}(x_t) \leq \Phi_{t-1}(x_{t+1})$ and thus,

$$\Phi_t(\mathbf{x}_t) - \Phi_t(\mathbf{x}_{t+1}) \leq \eta\nabla_t^T(\mathbf{x}_t - \mathbf{x}_{t+1})$$

By strong convexity we also have the following:

$$\Phi_t(\mathbf{x}_t) - \Phi_t(\mathbf{x}_{t+1}) \geq \nabla_t^T \Phi_t(\mathbf{x}_{t+1})^T(\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{\alpha}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1})\|_2^2$$

Now, as x_{t+1} is the minimizer of Φ_t, because standard optimality conditions, $\nabla_t^T \Phi_t(\mathbf{x}_{t+1})^T(y - \mathbf{x}_{t+1}) \geq 0$ for any $y \in K$ otherwise, one could decrease $\Phi_t(x_{t+1})$ by moving slightly in the direction of $y - x_{t+1}$.

So, $\nabla_t^T \Phi_t(\mathbf{x}_{t+1})^T(\mathbf{x}_t - \mathbf{x}_{t+1}) \geq 0$ and putting everything together gives,

$$\eta\nabla_t^T(\mathbf{x}_t - \mathbf{x}_{t+1}) \geq \Phi_t(\mathbf{x}_t) - \Phi_t(\mathbf{x}_{t+1}) \geq \frac{\alpha}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2 \tag{4}$$

Returning back to (3)

$$g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1}) = \frac{(\nabla_t^T(\mathbf{x}_t - \mathbf{x}_{x+1}))^2}{\nabla_t^T(\mathbf{x}_t - \mathbf{x}_{x+1})} \leq \frac{(\|\nabla_t\|_*^T|\mathbf{x}_t - \mathbf{x}_{t+1}|)^2}{\frac{\alpha}{2\eta}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2}$$

Here we are upper bounding the numerator using the definition of dual norms,

$$\nabla_t^T(\mathbf{x}_t - \mathbf{x}_{x+1}) \leq \|\nabla_t\|_*|x_t - x_{t+1}|$$

and lower bounding the denominator using (4). This gives,

$$g_t(\mathbf{x}_t) - g_t(\mathbf{x}_{t+1}) \leq \frac{2\eta}{\alpha}^T \|\nabla_t\|_*^2$$

which finishes the proof. □

References

[1] Elad Hazan. Introduction to Online Convex Optimization, Foundations and Trends in Optimization, 2015.