Lecture 8: Analyzing Gradient Descent

Instructor: Aditya Bhaskara   *Scribe: Pruthuvi Maheshakya Wijewardena*

CS 5966/6966: Theory of Machine Learning

February 6th, 2017

Abstract

Assuming that objective function is ρ-Lipschitz, we show that the number of iterations GD algorithm requires to achieve error $\leq \epsilon$ is $O(1/\epsilon^2)$. We also explore obtain a better bound when the function is β-smooth. In this case, the number of iterations needed to achieve error $\leq \epsilon$ is $O(1/\epsilon)$.

1 Review and Introduction

In the last lecture, we started the discussion of optimization in learning. Further, we defined convex functions and sets and discussed properties of gradients of convex functions.

Recall that a function $f : \Re^d \Rightarrow \Re$ is defined to be ρ-Lipschitz if

$$\forall x, y : |f(x) - f(y)| \leq \rho \|x - y\|$$

This implies that the gradient of the function at any point is $\leq \rho$.
($\| \bigtriangledown f(x)| \ \leq \rho$).
A function $f : \Re^d \Rightarrow \Re$ is defined to be β-smooth if $\bigtriangledown f$ is β-Lipschitz.

$$\forall x, y : \| \bigtriangledown f(x) - \bigtriangledown f(y)\| \leq \beta \|x - y\|$$

We outlined the gradient descent algorithm in the last lecture. In this algorithm, in each iteration t, we do the following weight update to the vector:

$$x^{(t+1)} = x^{(t)} - \eta \bigtriangledown f(x^{(t)})$$

In this lecture we analyze the bounds of number of iterations this algorithm requires when the function f is ρ-Lipschitz and β-smooth.

2 Analyzing gradient descent algorithm

When analyzing the GD algorithm, we make the following assumptions:

1. Objective function f is ρ-Lipschitz
2. Value of the function at starting point is at a distance $\leq B$ from the optimal value of the function.

The result we want is, after a while we come to a point where $f(w^{(t)}) - f(w^*)$ is small where $f(w^{(t)})$ is the value at t^{th} iteration and $f(w^*)$ is the optimal value. In this analysis, we prove something stronger.

$$\frac{1}{T}\sum_{t=1}^{T}(f(w^{(t)})-f(w^*)) \leq \frac{B^2}{2\eta T}+\frac{\rho^2\eta}{2}$$

This is, the difference between average value of the function over T iterations and the optimal value is small (ϵ). We need to pick η small enough so that we take small steps that it is guaranteed to converge and large enough T.

In this analysis we argue about the following quantity.

$$\frac{1}{T}\sum_{t=1}^{T}(f(w^{(t)})-f(w^*))$$

By definition of a convex function we know that

$$f(w^*) \geq f(w^{(t)}) + < \bigtriangledown f(w^{(t)}), w^* - w^{(t)} >$$

$$\frac{1}{T}\sum_{t=1}^{T}(f(w^{(t)})-f(w^*)) \leq \frac{1}{T}\sum_{t=1}^{T} < \bigtriangledown f(w^{(t)}), w^{(t)} - w^* >$$

Consider the distances from iterations $t+1$ and t to w^*.

$$\|w^{(t+1)} - w^*\|^2 - \|w^{(t)} - w^*\|^2$$

$$= \|w^{(t)} - \eta \bigtriangledown f(w^{(t)}) - w^*\|^2 - \|w^{(t)} - w^*\|^2$$

$$= -2\eta < w^{(t)} - w^*, \bigtriangledown f(w^{(t)}) > +\eta^2 \| \bigtriangledown f(w^{(t)})\|^2$$

(by Lipschitz property)

$$\leq -2\eta < w^{(t)} - w^*, \bigtriangledown f(w^{(t)}) > +\eta^2\rho^2$$

Let $D_{t+1} = \|w^{(t+1)} - w^*\|^2$ and $D_t = \|w^{(t)} - w^*\|^2$. By substituting and rearranging the above expression, we get

$$< w^{(t)} - w^*, \bigtriangledown f(w^{(t)}) > \leq \frac{1}{2\eta}[D_t - D_{t+1}] + \frac{\eta\rho^2}{2}$$

Now we look at the inequality

$$\frac{1}{T}\sum_{t=1}^{T} < \bigtriangledown f(w^{(t)}), w^{(t)} - w^* > \leq \frac{1}{T}\sum_{t=1}^{T}\{\frac{1}{2\eta}[D_t - D_{t+1}] + \frac{\eta\rho^2}{2}\}$$

From series sum of $(D_t - D_{t+1})$

$$\frac{1}{T}\sum_{t=1}^{T}\{\frac{1}{2\eta}[D_t - D_{t+1}] + \frac{\eta\rho^2}{2}\} = \frac{1}{2\eta T}(D_1 - D_{T+1}) + \frac{\eta\rho^2}{2}$$

$D_1 - D_{t+1} \leq D_1$ and we know that $D_1 = B^2$. Therefore we can write

$$\frac{1}{T}\sum_{t=1}^{T}(f(w^{(t)})-f(w^*)) \leq \frac{B^2}{2\eta T}+\frac{\eta\rho^2}{2}$$

By setting the terms $\frac{B^2}{2\eta T} = \epsilon/2$ and $\frac{\eta\rho^2}{2} = \epsilon/2$, we obtain $\eta = \frac{\epsilon}{\rho^2}$ and $T =$

$\frac{B^2\rho^2}{\epsilon^2}$. From this we can conclude that to have ϵ- accuracy at the end, we need $O(1/\epsilon^2)$ iterations when the function is ρ-Lipschitz.

We know that once the weight is updated, $x^{(t+1)}$ may not be in the valid set. In that case we project the vector back into the set.

$$x^{(t+\frac{1}{2})} = x^{(t)} - \eta \bigtriangledown f(x^t)$$

$$x^{(t+1)} = \prod x^{(t+\frac{1}{2})}$$

We know that $\|x^{(t+\frac{1}{2})} - x^*\| \geq \|x^{(t+1)} - x^*\|$. It can be proved that with this projection, a similar result on the number of iterations can be obtained (full proof can be found in the book).

Next we explore the number of iterations required to have ϵ- accuracy when the function is β-smooth as well. We do not prove this, but the result can be obtained by setting $\eta = 1/\beta$ in the weight update step of GD algorithm.

$$x^{(t+1)} = x^t - \frac{1}{\beta} \bigtriangledown f(x^t)$$

We know that when a function is β-smooth it has the following property.

$$f(x^{(t+1)}) \leq f(x^{(t)}) + < \bigtriangledown f(x^{(t)}), x^{(t+1)} - x^{(t)} > + \frac{\beta}{2}\|x^{(t+1)} - x^{(t)}\|^2$$

By substituting $x^{(t+1)}$ from the weight update rule, we get

$$f(x^{(t+1)}) \leq f(x^{(t)}) - \frac{1}{2\beta}\| \bigtriangledown f(x^{(t)})\|^2$$

Using this property, we can show that after T iterations, difference between function value and the optimal is small.

$$f(w^{t)}) - f(x^*) \leq \frac{2\beta B^2}{T}$$

By setting $\frac{2\beta B^2}{T} = \epsilon$, we get $T = \frac{2\beta B^2}{\epsilon}$. Therefore, we can conclude that to have ϵ-accuracy at the end, we need $O(1/\epsilon)$ iterations when the function is β-smooth.

3 Sub-Gradients

Sometimes we may be interested in functions that are convex, but not differentiable at every point. e.g: absolute value function at 0.

We define sub-gradient of a function at point $(x, f(x))$ as any direction v, that satisfies $\forall y$

$$f(y) \geq f(x) + < y - x, v >$$

For example in $|x|$, any of $v1$, $v2$, $v3$ can be considered as a sub-gradient of $|x|$ at point 0 (figure 1).

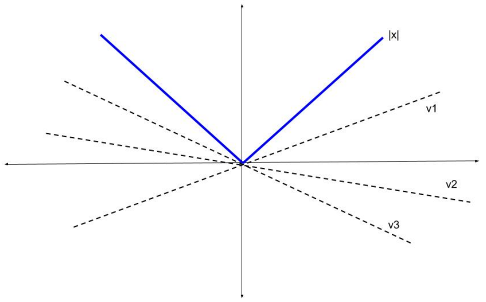

Figure 1: Sub-gradients of absolute value function at 0

4 Nesterovs accelerated Gradient Descent

This algorithm gives the optimal convergence for smooth functions.

$$f(w^{(t)}) - f(w^*) \leq O(\frac{\beta B^2}{T^2})$$

According to this, if we want ϵ-accuracy, we need $O(1/\sqrt{\epsilon})$ iterations.
The weight update rule is a slight variant of GD in this algorithm.

$$x^{(t+1)} = x^t - c_1 \bigtriangledown f(x^t) - c_2 \bigtriangledown f(x^{t-1})$$

$(t+1)^{th}$ term depends on both $(t)^{th}$ and $(t-1)^{th}$ terms. The constants c_1 and c_2 needs to be carefully chosen. This is the best possible guarantee for a GD algorithm.