Lecture #11: Convex Optimization, Gradient Descent

Instructor: Aditya Bhaskara *Scribe: Prasanth Yalamanchili*

CS 5966/6966: Theory of Machine Learning

February 15th, 2022

Abstract

This lecture covers the proof for vanilla analysis of gradient descent and projective gradient descent with basic inequalities and potential functions.

1 Gradient Descent Algorithm

Let us consider continuous differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ in domain. Gradient descent is an iterative algorithm. We start with a feasible point of w_0.Then, at each iteration, we take a step in the direction of the negative of the gradient at the current point from $t = (0, 1, 2, ..T - 1)$. The update of step is given by,

$$w_{(t+1)} = w_{(t)} - \eta \nabla f(w(t))$$

Where η is some fixed parameter determine learning rate .

If $w_{(t+1)}$ is outside the domain D .We take the closest point to that in our domain as the next point,which in general is the projection from that point onto the function.

There are few issues to address in this algorithm like how much to move in each step (aka learning rate). Which comes to:

- How should we set η ?
- Should η depend on t ?

Its natural to say that η should depend on t. which is what we wanted in practice.
Going further we will discuss much more about how to choose η.

2 Vanilla Analysis of Gradient Descent

Lipschitz property:- A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be L-Lipschitz if

$$|f(x) - f(y)| \leq L||x - y|| \quad \forall x, y.$$

Which is equivalent to saying, $||\nabla f(x)|| < L$
Intuitively, a Lipschitz function cannot change too fast.

The Vanilla Analysis is studying gradient descent algorithm under the two assumptions:

- f is L-Lipschitz, and domain $D = \mathbb{R}^d$
- The starting point w_0 is at a distance $\leq B$ from the optimum point w^*

Theorem:Consider running T steps of gradient descent with a fixed learning rate η, then we have

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

where w^* is the true minimizer of f i.e, $w^* = argmin_{x\in\mathbb{R}^d} f(x)$

If we set point $w_t^{'} = \frac{1}{T}(w_1 + w_2 + ... + w_t)$ and from definition of convex function we can write below inequality

$$f(w_t^{'}) \leq \frac{1}{T}(f(w_1) + f(w_2) + ... + f(w_t))$$

Now we can rewrite the above gradient descent equation as

$$f(w_t^{'}) - f(w^*) \leq \frac{\mathrm{B}^2}{2\eta T} + \frac{L^2\eta}{2}$$

If we take η to be very small like $\frac{\epsilon}{L^2}$ and T is very large. we get the value on the right side to be very small. which mean we are converging w_t to w^* (Point at which minimum value exists).

Which intuitively mean if we do small enough stepsize based on gradient value and enough number of steps we converge to minimum value.

3 Proof for Vanilla Analysis of Gradient Descent

According to the basic inequality of convexity, the function lies "above" the tangent plan at *any* point.

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle$$

As we move from any point on a convex function and move towards the minimizer, by inequality we can write the following:

$$f(w^*) - f(w_t) \geq \langle \nabla f(w_t), w^* - w_t \rangle \tag{2}$$

Which intuitively tells us keeping track of $||w^* - w_t||$ would be helpful to prove above theorem

So we define potential $\Phi_t = ||w^* - w_t||^2$
So our hope is that Φ_t reduces with time. we use this to analyse our algorithm.

Now we calculate $\Phi_t - \Phi_{t+1}$ using $w_{t+1} = w_t - \eta \nabla f(w(t))$

$$\Phi_t - \Phi_{t+1} = ||w^* - w_t||^2 - ||w^* - w_t + \eta \nabla f(w(t))||^2$$

$$\Phi_t - \Phi_{t+1} = -2\eta \langle\ w^* - w_t, \nabla f(w(t))\rangle - \eta^2 ||\nabla f(w(t))||^2$$

Using inequality-(2) for convex functions $f(w^*) - f(w_t) \geq \langle \nabla f(w_t), w^* - w_t \rangle$ we can write

$$-2\eta \langle\ w^* - w_t, \nabla f(w(t))\rangle - \eta^2 ||\nabla f(w(t))||^2 \geq -2\eta [f(w^*) - f(w_t))] - \eta^2 ||\nabla f(w(t))||^2$$

$$\Phi_t - \Phi_{t+1} \geq -2\eta [f(w^*) - f(w_t)] - \eta^2 ||\nabla f(w(t))||^2$$

If we reorganize above equation

$$2\eta [f(w_t) - f(w^*)] \leq \Phi_t - \Phi_{t+1} + \eta^2 ||\nabla f(w(t))||^2$$

$$f(w_t) - f(w^*) \leq \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \frac{\eta}{2} ||\nabla f(w(t))||^2$$

Since the function is L-Lipschitz $\forall$ x, $||\nabla f(x)|| \leq L$

$$f(w_t) - f(w^*) \leq \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \frac{\eta L^2}{2}$$

If we do summation of the above equation for all t values

$$\sum_{t=1}^{T} f(w_t) - f(w^*) \leq \frac{\Phi_0 - \Phi_T}{2\eta} + \frac{\eta T L^2}{2}$$

Since Φ_T is square which is always positive $\Phi_0 - \Phi_T \leq \Phi_0$ and $\Phi_0 \leq B^2$. So $\Phi_0 - \Phi_T \leq B^2$

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

Dividing by T on both sides,we proved the theorem

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

Now lets see what will be the best value of η for a given T . Since our right side bound need to be minimized. we need find η for which $\frac{\mathrm{B}^2}{2\eta T} + \frac{L^2\eta}{2}$ will be minimized.

By $A.M \geq G.M$ inequality.

$$\frac{\frac{\mathrm{B}^2}{2\eta T} + \frac{L^2\eta}{2}}{2} \geq \sqrt{\frac{\mathrm{B}^2}{2\eta T} . \frac{L^2\eta}{2}}$$

$$\frac{\mathrm{B}^2}{2\eta T} + \frac{L^2\eta}{2} \geq \frac{B.L}{\sqrt{T}}$$

Equality occur when both are equal $\frac{B^2}{2\eta T} = \frac{L^2\eta}{2}$

so $\eta = \frac{B}{L\sqrt{T}}$ is the best η we could take for a given T . So that

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^*) \leq \frac{B.L}{\sqrt{T}}$$

So the above equation imply. we start far from the minimum which mean B is large, In order to come close to minimum you need to have large T from the above equation.which means it takes more number of iterations.

Here in vanilla gradient descent if we do T iterations we are coming close to minimum by $\frac{1}{\sqrt{T}}$ times the original distance rather than $\frac{1}{T}$ times.In order to get $\frac{1}{T}$ times closer we need to have extra constrains on the function like smoothness, which we will be discussing on the further classes.

The same can be interpreted as if we want to get ϵ closer to minimum value we need to do $O(\frac{1}{\epsilon^2})$ iterations.

$$\frac{B.L}{\sqrt{T}} = \epsilon$$

$$T = \frac{B^2.L^2}{\epsilon^2}$$

4 Projected Gradient Descent

For a bounded domain D, Gradient descent definition is as follows.

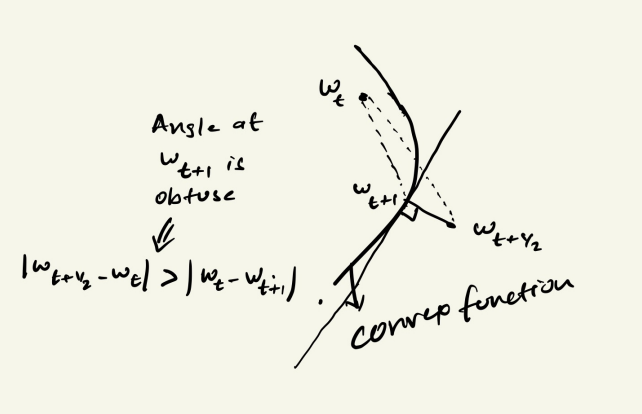
Suppose we have point w_t in Domain at t^{th} iteration and when we are going to the next iteration of the gradient descent algorithm and our new point was going outside of the domain. Then we take the projection of that point onto the domain and continue the process.

let $w_{t+\frac{1}{2}}$ be the point that went outside the domain.Now we need to take the projection of that point on to domain to find the w_{t+1}.

$$w_{t+\frac{1}{2}} = w_{(t)} - \eta\nabla f(w(t))$$

$$w_{t+1} = \Pi_D[w_{t+\frac{1}{2}}] = argmin_{x\in D}||w_{t+\frac{1}{2}} - x||$$

when we project the $w_{t+\frac{1}{2}}$ to w_{t+1} in the domain by taking gradient of the function at the projection w_{t+1} Since the function value is always above the gradient. By the property of the side opposite to the obtuse angle is the largest in the triangle we prove the below inequality.



$$||w^* - w_{t+1}||^2 \leq ||w^* - w_{t+\frac{1}{2}}||^2$$

$$\Phi_t - \Phi_{t+1} \leq \Phi_t - \Phi_{t+\frac{1}{2}}$$

Since we use $\Phi_t - \Phi_{t+\frac{1}{2}}$ with $w_{t+\frac{1}{2}} = w_t - \eta \nabla f(w(t))$ from gradient descent algorithm

$$\Phi_t - \Phi_{t+\frac{1}{2}} = ||w^* - w_t||^2 - ||w^* - w_t + \eta \nabla f(w(t))||^2$$

$$\Phi_t - \Phi_{t+\frac{1}{2}} = -2\eta \langle\, w^* - w_t, \nabla f(w(t))\rangle - \eta^2 ||\nabla f(w(t))||^2$$

$$\Phi_t - \Phi_{t+\frac{1}{2}} \geq -2\eta [f(w^*) - f(w_t)] - \eta^2 ||\nabla f(w(t))||^2$$

From the the equation $\Phi_t - \Phi_{t+1} \leq \Phi_t - \Phi_{t+\frac{1}{2}}$ we can say the below equation

$$\Phi_t - \Phi_{t+1} \geq \Phi_t - \Phi_{t+\frac{1}{2}} \geq -2\eta [f(w^*) - f(w_t)] - \eta^2 ||\nabla f(w(t))||^2$$

$$\Phi_t - \Phi_{t+1} \geq -2\eta [f(w^*) - f(w_t)] - \eta^2 ||\nabla f(w(t))||^2$$

Now we continue with the same proof as normal gradient descent algorithm which will prove the same analysis for projected gradient descent with the projection points gives the same bounds.

5 Extensions of Gradient Descent

What if functions at different steps are different in gradient descent ?

If all the functions are convex and L-Lipschitz and $f_t(x)$ is varied function which is taken at time-t. we can write $\forall w^* \in D$

$$f_t(w^*) - f_t(w_t) \geq \langle \nabla f_t(w_t), w^* - w_t \rangle$$

we can do similar analysis as gradient descent to this algorithm too and we get the below equation

$$\frac{1}{T} \sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq \frac{\mathrm{B}^2}{2\eta T} + \frac{L^2 \eta}{2}$$

This equation is exactly similar to online convex optimization which we will be discussing in the coming classes.
where we be will comparing the error from the best fixed strategy that can be used.

we also use the above algorithm in stochastic gradient descent, where in each step we take different examples and create a different loss function and try to reach minimum value by moving in opposite direction of gradient of that function at each step.

we will formally prove the above equation and do analysis on bounds in the next lecture.