Lecture #: Topic

Instructor: Aditya Bhaskara *Scribe: Vinutha Raja*

CS 5966/6966: Theory of Machine Learning

February 22nd, 2022

Abstract

In this lecture we discuss about the Noisy Gradient Descent, Analysis of Gradient Descent for convex smooth functions and Non convex smooth functions.

1 Noisy Gradient Descent

Let's consider a setting where we cannot evaluate the gradient of the function at a given point. Here the gradient descent will be preformed using an oracle called Noisy gradient oracle. The oracle provides g(w) called noisy gradient, with the properties such that, the expected value of the g(w) is equal to the gradient:

$$E[g(w)] = \nabla f(w)$$

$$E[|g(w)|^2] \leq L^2$$

With the above properties of g(w), the convergence guarantees also holds good for noisy gradient.

Application: One of the applications of Noisy gradient descent is it is used in the field of federated learning, which is a machine learning technique that trains an algorithm across multiple decentralized servers holding local data samples, without exchanging them. Here the servers might not want to disclose the actual data but might provide a noisy gradient updates of the data.

2 Smoothness

The definition of smoothness relies on the notion of the gradient. Lets recall that the gradient of a differentiable function f : $R^D - > R$ at w, denoted $\nabla f(w)$, is the vector of partial derivatives of f.

Smoothness: A differentiable function f: $R^d - > R$ is M-smooth if its gradient is M-Lipschitz. i.e for all x, y we have,

$$||\nabla f(x) - \nabla f(y)|| \leq M||x - y||$$

this is equivalent to

$$||\nabla^2 f(x)||_2 <= M$$

In smooth functions the gradient doesn't suddenly shift. Another consequence of the smoothness is that the function can always be bounded by a parabola i.e. from the above definition of smoothness, if the function is M-smooth, we can write

$$f(y) \leq f(x) + < \nabla f(x), y - x > + M||y - x||^2 \quad - > Eq(1)$$

we know that, the update step of gradient descent is

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

using the above equation (1),

$$f(w_{t+1}) \leq f(w_t) + < \nabla f(w_t), w_t + 1 - w_t > + M||w_{t+1} - w_t||^2$$

by plugging

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

we get

$$f(w_{t+1}) \leq f(w_t) - \eta||\nabla f(w_t)||^2| + M\eta^2||\nabla f(w_t)||^2$$

Substituting

$$\eta = \frac{1}{2M}$$

$$f(w_{t+1}) \leq f(w_t) - \frac{1}{2M}||\nabla f(w_t)||^2 + \frac{1}{4M}||\nabla f(w_t)||^2$$

$$f(w_{t+1}) \leq f(w_t) - \frac{1}{4M}||\nabla f(w_t)||^2$$

$$f(w_{t+1}) \leq f(w_t) - \frac{\eta}{2}||\nabla f(w_t)||^2 \quad - > Equation(2)$$

So based on the Taylor Approximation, for smooth functions if η is small, the bound holds with some loss in the constant as it is $\eta/2$. The smoothness tells that the error in the Taylor approximation is small and which means that the drop we get in the function is very close to what we expect as the first order Taylor approximation.

If the function is M smooth, it holds in the neighborhood of radius about 1/2M.

After telescoping sum for all t in equation $f(w_{t+1}) \leq f(w_t) - \frac{1}{4M}||\nabla f(w_t)||^2$

$$\sum_t |\nabla f(w_t)|^2 \leq 4M(f(w_0) - f(w_{t+1}))$$

Since $f(w_{t+1}) \geq f(w*)$

$$\sum_t |\nabla f(w_t)|^2 \leq 4M(f(w_0) - f(w*))$$

3 Gradient Descent Analysis for convex smooth functions

Based on our earlier analysis, we know that:

$$\phi_t = ||w_t - w*||^2$$

$$\phi_t - \phi_{t+1} \geq 2\eta(f(w_t) - f(w*)) - \eta^2 ||\nabla f(w_t)||^2$$

From Eq (2)

$$||\nabla f(w_t)||^2 = 2/\eta f(w_t) - f(w_{t+1})$$

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

$$f(w_t) - f(w*) \leq \frac{\phi_t - \phi_t + 1}{2\eta} + (f(w_t) - f(w_{t+1})$$

By telescoping sum for all t, we get

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w*) \leq \frac{2M(\phi_0 - \phi_{t+1})}{T} + \frac{f(w_0) - f(w_t)}{T}$$

since $\phi_0 - \phi_{t+1} \leq \phi_0 \leq B^2$, where $B = |w_0 - w_*|$.

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w*) \leq \frac{2MB^2}{T} + \frac{f(w_0) - f(w_t)}{T}$$

If $f(w_0) - f(w_t) \leq C$, which means the starting function value and the minimum function value is C distance apart.

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

So here the error after T steps is proportional to $\frac{1}{T}$, where as in normal gradient descent, the error is error proportional to $\frac{1}{\sqrt{T}}$, so we converge at a faster rate than the normal gradient decent.

4 Non Convex Smooth function

If f is non convex but M-smooth, (smoothness is upper bounded by a quadratic), then analysis before (with $\eta = 1/2M$) implies that

$$||\nabla f(w_t)||^2 \leq 4M(f(w_t) - f(w_{t+1})$$

$$\frac{1}{T}\sum_{t=1}^{T} ||\nabla f(w_t)||^2 \leq \frac{4M(f(w_o) - f(w*))}{T}$$

The consequence of this is, there exists some t such that,

$$||\nabla f(w_t)||^2 \leq \frac{4MC}{T}$$

which means there exists say some point t where gradient converges to smaller value which is called as an approximate singular point.

So we can use this in the analysis of non convex functions, even if we perform gradient descent, we converge to a point where the gradient is close to zero. This can be either local maxima or local minima. But with majority of functions which we use always tends to local minima.