Lecture 9: Stochastic Gradient Descent

Instructor: Aditya Bhaskara *Scribe: Mohsen Abbasi*

CS 5966/6966: Theory of Machine Learning

February 8th, 2017

Abstract

In the first part of the lecture we will go over the convergence rate of Gradient Descent for strongly convex functions. Later on, we will cover another variant of GD called Stochastic Gradient Descent.

1 Introduction

In the last lecture we learned how to minimize convex functions. We showed that if f is a ρ-Lipschitz function, our starting point is at a distance $\leq B$ from the minimum and the learning rate is set to be $\eta = \frac{\epsilon}{\rho^2}$, after $T = \frac{\|x^{(0)} - x^*\|^2 \rho^2}{\epsilon^2}$ iterations:

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

in which $\epsilon = \frac{B^2}{2\eta T} + \frac{\eta}{2}\rho^2$.
Now the question is that can we get a better convergence rate? As it turns out, if the function is strongly convex, for an error of ϵ, we need only $log(\frac{1}{\epsilon})$ iterations.

1.1 definition. A function f is called γ-strongly convex if $\forall x, y$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\gamma}{2} \|y - x\|^2 \quad (1)$$

The definition of strongly convex functions is similar to that of smooth functions and just the direction of the inequality is reversed. It guarantees a curvature at every point i.e. the function can be lower bounded by a parabola. For example, a line is smooth and convex but not strongly convex. As shown in Figure **??**, The idea here is that a parabola gives a much better lower bound for the optimum point compared to a line.

A function f is called β-smooth if $\forall x, y$:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2$$

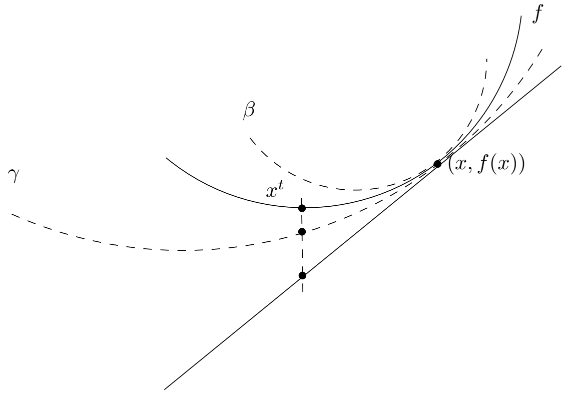

Figure 1: f The lower bound given by the lower parabola is closer to $(x^t, f(x^t))$ compared to the gradient.

1.2 claim. *If a function is strongly convex, for an error of ϵ, we need $log(\frac{1}{\epsilon})$ iterations.*

Proof. By definition we know:

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

To find out where the parabola is minimized, we should see what is the value of y for which the right hand side of the equation is minimized. This value would be:

$$y = x - \frac{1}{\gamma}\nabla f(x) \Rightarrow y - x = -\frac{1}{\gamma}\nabla f(x)$$

By plugging this value in the inequality we get:

$$(2) \quad f(x^*) \geq f(x) - \frac{1}{2\gamma}\|\nabla f(x)\|^2$$

Note that this means if the gradient is small, we're already close to the minimum and don't have to make a big progress.
From the last lecture we know that for β-smooth functions:

$$(3) \quad f(w^{(t+1)}) \leq f(w^{(t)}) - \frac{1}{2\beta}\|\nabla f(x^{(t)})\|^2$$

But now we know that this quantity is related to the distance from optimum. Also, knowing equation **??**:

$$\begin{aligned} f(w^{(t+1)}) - f(x^*) &\leq [f(w^{(t)}) - f(x^*)] - \frac{1}{2\beta}[2\gamma(f(w^{(t)}) - f(x^*))] \\ &= (1 - \frac{\gamma}{\beta})[f(w^{(t)}) - f(x^*))] \\ &\leq \cdots \leq (1 - \frac{\gamma}{\beta})^t[f(w^{(t)}) - f(x^*))] \end{aligned}$$

We know that $1 - \frac{\gamma}{\beta} \leq e^{-\frac{\gamma}{\beta}}$. Therefore, in order to have an error less than ϵ, it's enough to set $t = \frac{\beta}{\gamma}log(\frac{1}{\epsilon})$. □

2 Stochastic Gradient Descent

Learning is equivalent to empirical risk minimization which is finding a w for which sum of the losses over training examples is small. For a set of m training examples $x_1, \ldots, x_m$ with labels $y_1, \ldots, y_m$, minimizing this loss is defined as:

$$\arg\min_w f(w) = \arg\min_w \frac{1}{m} \sum_{i=1}^{m} l(w, x_i, y_i)$$

And the gradient of this function would be:

$$\nabla f(w) = \frac{1}{m} \sum_{i=1}^{m} \nabla l(w, x_i, y_i)$$

The idea behind SGD is that instead of going over all the examples in order to compute the gradient of f, a random sample is picked from the training set and the gradient of loss function is computed only at this point. If the training set is of size m, each iteration of Vanilla Gradient Descent takes $O(m)$ time while an iteration of SGD would take $O(1)$ time.

Algorithm 1 SGD algorithm

```
procedure SGD
    Initialize $w^{(0)}$ with any feasible $w$
    for $t = 1 \ldots T$ do
        Sample a random training example $i$ :
        Update $w^{(t+1)} = w^{(t)} - \eta \nabla l(w, x_i, y_i)$
    end for
    Output $w$
end procedure
```

In order to get an intuition for why SGD works, we should see what is the expected value of $w^{(t+1)}$.

$$E[w^{(t+1)}] = E[w^{(t)}] - \eta E[\nabla l(w, x_i, y_i)]]$$

The expected value of the loss for a random point is $\sum_{i=1}^{m} \frac{1}{m} \nabla l(w^{(t)}, x_i, y_i)$ which is $\nabla f(w^{(t)})$. Therefore, SGD is doing the same thing as GD meaning it goes to the right direction in expectation. The geometrical interpretation of this is shown in Figure **??**.

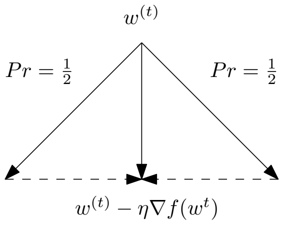

Figure 2: As long as we don't deviate too much, in expectation w is going to the right direction.

2.1 claim. *Under condition:*

$$\forall i, \|\nabla l(w^{(t)}, x_i, y_i)\| \leq \rho$$

If we set $\eta = \frac{\epsilon}{\rho^2}$, after running SGD for $T = \frac{B^2\rho^2}{\epsilon^2}$ iterations:

$$E[f(\bar{w}) - f(w^*)] \leq \epsilon$$

When $\epsilon = \frac{B^2}{2\eta T} + \frac{\eta}{2}\rho^2$.

Proof. We know from before:

$$f(w^{(t)}) - f(w^*) \leq \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$

We also bounded the sum:

$$\frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle \approx \|w^{(t+1)} - w^*\|^2 - \|w^{(t)} - w^*\|^2 + \dots$$

If at time t we make the step $w^{(t+1)} = w^{(t)} - \eta g(t)$ in which $g(t)$ is some chosen gradient then:

$$E[g(t)] = \nabla f(w^{(t)})$$

$$\begin{aligned} \|w^{(t+1)} - w^*\|^2 - \|w^{(t)} - w^*\|^2 &= \|w^{(t)} - \eta g(t) - w^*\|^2 - \|w^{(t)} - w^*\|^2 \\ &= -2\eta\langle w^{(t)} - w^*, g(t)\rangle + \eta^2\|g(t)\|^2 \end{aligned}$$

Setting $D_{t+1} = \|w^{(t+1)} - w^*\|^2$ and $D_t = \|w^{(t)} - w^*\|^2$:

$$\langle w^{(t)} - w^*, g(t)\rangle = \frac{1}{2\eta}[D_t - D_{t+1}] + \frac{\eta}{2}\|g(t)\|^2$$

$$\begin{aligned} \Rightarrow \sum_{t=1}^{T}\langle w^{(t)} - w^*, g(t)\rangle &= \frac{1}{2\eta}[D_t - D_{t+1}] + \frac{\eta}{2}\sum_{t=1}^{T}\|g(t)\|^2 \\ &\leq \frac{B^2}{2\eta} + \frac{\eta T}{2}\rho^2 \end{aligned}$$

It is left to show that

$$E[\frac{1}{T}\sum_{t=1}^{T}\langle w^{(t)} - w^*, \nabla f(w^{(t)})\rangle] \leq E[\frac{1}{T}\sum_{t=1}^{T}\langle w^{(t)} - w^*, g(t)\rangle] \tag{4}$$

But by the *law of total expectation* we see:

$$E[\langle w^{(t)} - w^*, g(t)\rangle | w^{(t)}] = \langle w^{(t)} - w^*, \nabla f(w^{(t)})\rangle$$

because looking at the expectation over $g(t)$, we can condition on all the values up to $t-1$ (think of them as being fixed) and take the average over the last one. □