CS 596: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Timothy Yong

Lecture #21
December 1, 2016

1 Previous Lecture

In the last lecture we looked at convergence of gradient descent for functions with bounded gradiantes. A function f is G-bounded if $\forall x, \|\nabla f(x)\| \leq G$.

Theorem 1.1. *If f is convex and G-bounded, then after T steps of GD,*

$$\frac{1}{T}\sum_{i=1}^{T} f(x_i) - f(x^*) \leq \frac{RG}{NT}$$

where x^* is the optimal value for x, $R = \|x^*\|$, and $\eta = \frac{R}{G\sqrt{T}}$. One can get better rates for nicer functions, such as when f is β-Lipschitz. A function f is β-Lipschitz if

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

Theorem 1.2. *If f is convex and β-Lipschitz, then after T steps of GD and for $\eta = \frac{1}{\beta}$,*

$$f(x_t) - f(x^*) \leq \frac{\beta R^2}{2T}$$

β-Lipschitz functions have several advantages, such as a better rate of convergence, and the ability to use a fixed step size. This is due to the fact that as x approaches x^*, i.e., $\|x - x^*\| \leq \epsilon'$, we have that $\|\nabla f(x) - \nabla f(x^*)\| \leq \beta\epsilon$, and since $\nabla f(x^*) = 0$ we have that $\|\nabla f(x)\| \leq \beta\epsilon \leq 1$. Hence, the value of the gradients will make sure that we are taking smaller steps closer to the optimum.

2 Facts about β-Lipschitz functions

1. If f is β-Lipschitz, $f(y) \leq f(x) + \nabla f(x)(y - x) + \frac{\beta}{2}\|x - y\|^2$.
2. If f is β-Lipschitz and convex, $f(y) \geq f(x) + \nabla f(x)(y - x) + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2$

Proof. (of 1)

$$f(x) \geq f(y) + \nabla f(y)(x - y)$$

$$\Rightarrow f(y) - f(x) \leq \nabla f(y)(y - x)$$

$$\Rightarrow f(y) - f(x) - \nabla f(x)(y - x)$$

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

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

By Cauchy-Schwartz,

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

$$\leq \beta\|y - x\|^2$$

This can be further improved by applying Cauchy-Schwartz in a continuous fashion as follows

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

=

$$\int_0^1 \nabla f(x + t(y - x))(y - x)dt - \int_0^1 \nabla f(x)(y - x)dt$$

$$= \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))(y - x)dt$$

By Cauchy-Schwartz,

$$\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|\|y - x\|dt$$

$$\leq \int_0^1 \beta\|y - x\|^2 t dt = \frac{\beta}{2}\|y - x\|^2$$

□

Also notice that in the above proof with factor $\beta/2$ we never used the fact that f is convex.

Proof. (of 3)

We define $\phi(y) = f(y) - \nabla f(x)y$. Since f is a Lipschitz function and a linear combination of Lipschitz functions are always Lipschitz functions, then ϕ is also a Lipschitz function. Also notice that x is the minimizer of $\phi()$, i.e.

$$\forall y \neq x : f(y) - \nabla f(x)y \geq f(x) - \nabla f(x)x$$

$$\phi(y) \geq \phi(x)$$

Hence we have that

$$\phi(x) - \phi(y) \leq \phi(y - \frac{\nabla\phi(y)}{\beta}) - \phi(y)$$

$$\leq \nabla\phi(y)(\frac{-\nabla\phi(y)}{\beta}) + \frac{\beta}{2}\frac{\|\nabla\phi(y)\|^2}{\beta^2}$$

$$= \frac{-1}{2\beta}\|\nabla\phi(y)\|^2$$

$$= \frac{-1}{2\beta}\|\nabla f(y) - \nabla f(x)\|^2$$

Substituting $\phi(y)$ with $f(y) - \nabla f(x)y$:

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

□

Theorem 2.1. *If f is convex and β-Lipschitz, then for $\eta = \frac{1}{\beta}$, after T steps of GD, $f(x_t) - f(x) \leq \frac{R^2\beta}{2T}$. In other words, the higher the number of steps T, the closer $f(x_t)$ approximates $f(x)$.*

Proof. Step 1. Function values monotonically decrease

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

$$\leq f(x_t) + \nabla f(x_t)(\frac{-1}{\beta}\nabla f(x_t)) + \frac{\beta}{2}\frac{\|\nabla f(x_t)^2\|}{\beta^2}$$

$$\leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2 \tag{1}$$

Step 2. Using convexity, $y = x^*$, $x = x_t$

$$f(x^*) \geq f(x_t) + \nabla f(x_t)(x^* - x_t) \tag{2}$$

By rearranging (1)

$$\Rightarrow f(x_{t+1}) - f(x^*) \leq f(x_t) - f(x^*) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2$$

using (2), and by adding and subtracting $\frac{\beta}{2}\|x_t - x^*\|^2$,

$$\leq \nabla f(x_t)(x_t - x^*) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2 + \frac{\beta}{2}\|x_t - x^*\|^2 - \frac{\beta}{2}\|x_t - x^*\|^2$$

and using completing the square

$$= -(\frac{\beta}{2}\|x_t - x^*\|^2 + \frac{1}{2\beta}\|\nabla f(x_t)\|^2 - \nabla f(x_t)(x_t - x^*)) + \frac{\beta}{2}\|x_t - x^*\|^2$$

$$= -\frac{\beta}{2}(\|x_t - x^* - \frac{1}{\beta}\nabla f(x_t)\|^2) + \frac{\beta}{2}\|x_t - x^*\|^2$$

$$= -\frac{\beta}{2}\|x_{t+1} - x^*\|^2 + \frac{\beta}{2}\|x_t - x^*\|$$

This is now a telescoping sum, so we can rewrite this as

$$\frac{1}{T}\sum_{t=0}^{T} f(x_{t+1}) - f(x^*) \leq \frac{1}{T}\frac{\beta}{2}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$$

$$\leq \frac{\beta}{T^2}\|x_0 - x^*\|^2 \leq \frac{\beta R^2}{2T}$$

□

Theorem 2.2. *For any T, $\exists$ a β-Lipschitz convex function such that for any 1^{st} order method in T steps,*

$$f(x_T) - f(x^*) \geq \frac{3\beta R^2}{32T^2}$$

This theorem won't be proved in this set of notes. The tight bound is achieved for f be defined as

$$f(x_1, x_2, ..., x_{2T+1}) = \frac{\beta}{8}[x_1^2 + x_2^2 + ... + x_{2T+1}^2 - x_1x_2 - x_2x_3 + ... + (-x_ix_{i+1}) + ... + (-x_{2_T}x2_{T+1}) - x_1]$$

To see that f is convex, consider f in the following matrix form,

$$f(\vec{x}) = \vec{x}^T A\vec{x} - b^T\vec{x}$$

If the Hessian of f is positive semi-definite, then the function is convex. To compute the Hessian, we compute the derivative over x_i and x_j, where $n = 2T + 1$, and A is symmetric, so $A_{ij} = A_{ji}$. First we compute $\nabla_H \vec{x}^T A\vec{x}$

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_k}\vec{x}^T A\vec{x} = \frac{\partial}{\partial x_i}(\frac{\partial}{\partial x_k}\sum_{i=1,i\neq k}^{n} A_{ik}x_ix_k + \frac{\partial}{\partial x_k}\sum_{j=1,j\neq k}^{n} A_{kj}x_kx_j + \frac{\partial}{\partial x_k}A_{kk}x_k^2)$$

$$= \frac{\partial}{\partial x_i}(2\frac{\partial}{\partial x_k}\sum_{i=1,i\neq k}^{n} A_{ik}x_ix_k + \frac{\partial}{\partial x_k}A_{kk}x_k^2)$$

$$= \frac{\partial}{\partial x_i}(2\sum_{i=1,i\neq k}^{n} A_{ik}x_i + 2A_{kk}x_k) = 2A_{ik}$$

and then we can compute $\nabla_H b\vec{x}$

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_k}b\vec{x} = \frac{\partial}{\partial x_i}(\frac{\partial}{\partial x_k}\sum_{i=1}^{n} b_kx_k) = 0$$

And so the Hessian of f is $2A$, which is PSD and therefore convex. The gradient of f is

$$\nabla f(x) = 2AX - b$$

By Cauchy Schwartz,

$$\|2AX - 2AY\| \leq 2\|A\|\|X - Y\|$$

From this one can deduce that f is β-Lipschitz.

3 Momentum-based Methods

Gradient descent has a tendency to oscillate during the update of x, and so to reduce such oscillations, a momentum term may be introduced. The idea behind momentum is to "push" the gradient and attenuate for oscillations. Consider the gradient descent update with momentum

$$x_t = x_{t-1} - \eta \nabla f(x_t)$$

$$y_t = (1 - \gamma_t)x_t + m_t$$

where m_t is the momentum vector, and γ_t is the rate. There are several types of momentum based methods, such as conjugate gradient descent, heavy ball method, and accelerated gradient descent. The optimal one is Nesterov's accelerated gradient descent. The algorithm is as follows.

$$x_0 = 1$$

$$x_1 = x_0 - \frac{1}{\beta}\nabla f(x_0)$$

$$y_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$$

$$x_{t+1} = (1 - \gamma_t)y_{t+1} + \gamma_t y_t$$

Theorem 3.1. *For appropriate values of γ_t , accelerated gradient descent satisfies*

$$f(y_t) - f(x^*) \leq \frac{2\beta R^2}{t^2}, \eta = \frac{1}{\beta}$$

4 Additional Readings

- The book by Boyd and Vandenberghe. `http://stanford.edu/~boyd/cvxbook/`.
- Lecture notes by Sébastian Bubeck. `https://blogs.princeton.edu/imabandit/orf523-the-complexities-of-optimization/`.