Indian Institute of Technology Kanpur
CS774 Optimization Techniques

Scribe: Abhishek Srivastava
Instructor: Purushottam Kar
Date: September 5, 2016

LECTURE
10

Strong Convexity and Strong Smoothness

1 Introduction

In the bounded gradient case, we derived that Gradient Descent converges with $\frac{1}{\sqrt{T}}$ rate for convex optimization problem. In order to improve the convergence rate, we are going to impose some more constrains on $f(x)$. we are going to introuduce the concepts of strongly covex and strongly smooth functions and then derive the convergence rate of Gradient Descent in either or both cases (Kakade et al., 2009).Below is gradient descent algorithm from last lecture, convergence for which is supposed to be analyzed.

Algorithm 1: Projected Gradient Descent

Input: Function f,Gradient Oracle $\nabla f(\mathbf{x})$ or Sub-gradient Oracle $g \in \partial_x f$, Projection oracle $\Pi_{\mathcal{C}}$, constraint set $\mathcal{C}$
Output: $\mathbf{x} \in \mathcal{C}$
1: $\mathbf{x}^0 \leftarrow \mathbf{0}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad \mathbf{z}^{t+1} \leftarrow \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t)$ [a] `//Gradient descent`
4: $\quad \mathbf{x}^{t+1} \leftarrow \Pi_c(\mathbf{z}^{t+1})$ `//Projection step`
5: **end for**
6: **return** $\hat{\mathbf{x}} = \begin{cases} \mathbf{x}^T \\ \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}^t \end{cases}$

[a] $\mathbf{z}^{t+1} \leftarrow \mathbf{x}^t - \eta_t g(\mathbf{x}^t)$ (Subgradient in the case gradient doesn't exist.)

2 Strong Convexity & Strong Smoothness

Definition 10.1. A function $f : \chi \to \mathbb{R}$ is said to be α-strongly convex (SC) if $\forall \mathbf{x}, \mathbf{y} \in \chi$, we have $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\alpha}{2} \|\mathbf{x} - \mathbf{y}\|^2$ provided it's a differntialbe function.

Remark 10.1. Strong convexity implies that the function values are supposed to lie above some quardratic curve wrt to the tangent. Some properties of convex functions related to strong convexity are following:

- $f(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|_2^2$ is 1-SC
- If f is α- SC then $c.f$ is $c.\alpha$ -strongly convex.

- If f is α- SC and g is β-SC, then $(f+g)$ is $(\alpha+\beta)$-SC
- Any Convex function is 0-SC

Definition 10.2. A differentiable function $f : \chi \to \mathbb{R}$ is said to be L-strongly smooth (SS) if $\forall \mathbf{x}, \mathbf{y} \in \chi$, we have $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2$

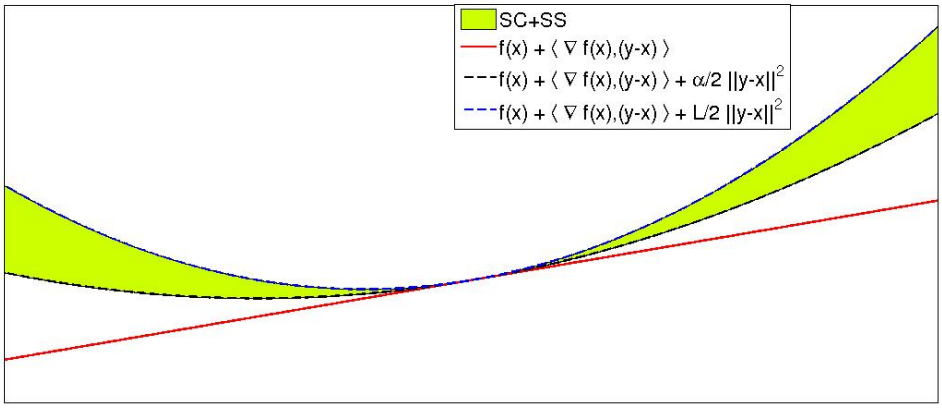

Figure 1: Functions lying above black dashed line are α-SC and functions lying between tangent and blue line are L-SS. Functions in shaded area are both SC & SS

Remark 10.2. Strong smoothness means that the functon values are bounded above its tangent by some quadratic curve.

Remark 10.3. If f is doubly differentiable, we can approximate the parameter of strong convexity α by minimizing the smallest eigenvalue of Hessian at differnt points in the domain of f. Similarly parameter of storng smoothness L for a convex function can be obtained my taking maximum of largest eigenvalues of Hessian at different points in χ.

3 Analysis of Strongly Convex function

In this section, we will analyze how Projected gradient descent performs when the function is strongly convex. In addition, we will continue to use the Bounded Gradient assumption.
We are denoting Potential function Φ_t and Auxillary function D_t as following

$$\Phi_t := f(\mathbf{x}^t) - f(\mathbf{x}^*)$$
$$D_t = \left\|\mathbf{x}^t - \mathbf{x}^*\right\|_2$$

Using the strong convexity assumption and incorporating convexity property, we get

$$f(\mathbf{x}^t) \leq f(\mathbf{x}^*) + \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^* \right\rangle \tag{1}$$

$$
\begin{aligned}
\Phi_t &\leq \left\langle \mathbf{g}^t, \mathbf{x}^t - \mathbf{x}^* \right\rangle - \frac{\alpha}{2} \left\| \mathbf{x}^t - \mathbf{x}^* \right\|^2 \\
&\leq \frac{D_t^2 - D_{t+1}^2}{2\eta_t} + \frac{\eta_2}{2} G^2 - \frac{\alpha}{2} D_t^2 \\
\sum_{t=1}^{T} \Phi_t &\leq D_0^2 \underbrace{\left(\frac{1}{2\eta_0} - \frac{\alpha}{2} \right)}_{\leq 0 \text{ if } \eta_0 = \frac{c}{\alpha}, c \geq 1} + \sum_{t=1}^{T} D_t^2 \underbrace{\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} - \frac{\alpha}{2} \right)}_{\leq 0 \text{ if } \eta_t = \frac{c}{\alpha t}, c \geq 1} + \frac{G^2}{2} \sum_{t=1}^{T} \eta_t
\end{aligned}
$$

If we set $\eta_t = \frac{1}{\alpha t}$ for $t \geq 1$ and $\eta_0 = \frac{1}{\alpha}$, The first and second term in RHS of above equation vanishes. Also, the third term can be approximated as $\frac{G^2}{2} \sum \eta_t = \frac{G^2}{2\alpha} \sum \frac{1}{t} \approx \frac{G^2}{2\alpha} \log T$. Dividing both sides by T, we finally get

$$
\frac{1}{T} \sum_{t=1}^{T} \Phi_t \leq \frac{G^2 \log T}{2\alpha T}
$$

Remark 10.4. Here we are getting $\frac{\log T}{T}$ rate fo convergence. To get rid of log-term, we can do the following

$$
\begin{aligned}
\sum_{t=1}^{T} t\Phi_t &\leq tD_0^2 \underbrace{\left(\frac{1}{2\eta_0} - \frac{\alpha}{2} \right)}_{=0 \text{ if } \eta_0 = \frac{1}{\alpha}} + \sum_{t=1}^{T} tD_t^2 \underbrace{\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} - \frac{\alpha}{2} \right)}_{=0 \text{ if } \eta_t = \frac{1}{\alpha t}} + \frac{G^2}{2} \sum_{t=1}^{T} t\eta_t \\
&= \frac{G^2}{2\alpha} T \\
\frac{2}{T(T+1)} \sum_{t=1}^{T} t\Phi_t &\leq \frac{G^2}{2\alpha(T+1)}
\end{aligned}
$$

Return $\hat{\mathbf{x}} = \frac{2}{T(T+1)} \sum_{t=1}^{T} t\mathbf{x}^t$ as the result. Here we are taking weighted average of potential funtion where we are giving more more weights to final values. This shows that Projected Gradient descent converges with $\frac{1}{T}$ rate.

4 Analysis of Strongly Smooth Function

In this section, we will analyze how the Projected Gradient Descent performs when the convex function is L-Strongly smooth. In this case we don't need to use the Bounded Gradient assumption. First we are going to prove that the value of Potential function decreases in every iteration.

4.1 Monotonicity

Let us use the condition that the function is L-strongly smooth.

$$\begin{aligned}
f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) &\leq \left\langle g^t, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
&\leq \frac{1}{\eta_t} \left\langle \mathbf{x}^t - \mathbf{z}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
&= -\frac{1}{\eta_t} \left\| \mathbf{x}^t - \mathbf{x}^{t+1} \right\|_2^2 + \frac{1}{\eta_t} \underbrace{\left\langle \mathbf{x}^{t+1} - \mathbf{z}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle}_{\leq 0 \text{ using projection property I}} + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) &\leq \left(\frac{L}{2} - \frac{1}{\eta_t} \right) \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) &\leq 0 \qquad \textbf{for} \quad \eta_t \leq \frac{2}{L}
\end{aligned}$$

Remark 10.5. If we choose $\eta_t \leq \frac{2}{L}$,we find that $f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)$. Hence can be sure that **the final answer is best answer**

4.2 Regret Analysis

$$\begin{aligned}
f(\mathbf{x}^t) &\leq f(\mathbf{x}^*) + \left\langle g^t, \mathbf{x}^t - \mathbf{x}^* \right\rangle \\
f(\mathbf{x}^{t+1}) &\leq f(\mathbf{x}^t) + \left\langle g^t, \mathbf{x}^{t+1} - \mathbf{x}^t) \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
&\leq f(\mathbf{x}^*) + \left\langle g^t, \mathbf{x}^t - \mathbf{x}^* \right\rangle + \left\langle g^t, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
\Phi^{t+1} &\leq \left\langle g^t, \mathbf{x}^t - \mathbf{x}^* \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
&\leq \frac{1}{\eta_t} \left\langle \mathbf{x}^t - \mathbf{z}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
&\leq \frac{1}{\eta_t} \underbrace{\left\langle \mathbf{x}^t - \mathbf{x}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle}_{A_t} + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
&\qquad\qquad D_t^2 = \left\| \mathbf{x}^t - \mathbf{x}^{t+1} \right\|_2^2 + D_{t+1}^2 + 2A_t \\
&\qquad\qquad A_t = \frac{D_t^2 - D_{t+1}^2 - \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2}{2} \\
\Phi^{t+1} &\leq \frac{1}{2\eta_t} \underbrace{(D_t^2 - D_{t+1}^2)}_{\text{Automatic Telescoping}} + \left(\frac{L}{2} - \frac{1}{2\eta_t} \right) \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2
\end{aligned}$$

If we set $\eta_t \leq \frac{1}{L}$, The second term will in RHS of above equation will become negative. We are setting $\eta_t = \frac{1}{L}$ which is also less than $\frac{2}{L}$. After taking the sum over $1, ..., T$, we get

$$\frac{1}{T} \sum_{t=1}^{T} \Phi_t \quad \leq \quad \frac{D_0^2 L}{2T}$$

It can be noted that we don't require bounded gradient assumption as strong-smoothness already guarantees that. Also, we have proved that the function value is always decreasing. Hence we get guarantee of $\frac{1}{T}$ rate of convergence with the last value as our optimum.

5 Analysis of SC & SS

Since we are employing both restrictions of Strongly Convex and Strongly Smooth, we can continue to use the results derived in previous section which states that **final answer is best answer** if $\eta_t \leq \frac{2}{L}$. we only need to do little modification in previous proof to obtain the convergence bound.

5.1 Regreat Analysis

$$
\begin{aligned}
f(\mathbf{x}^t) &\leq f(\mathbf{x}^*) + \left\langle g^t, \mathbf{x}^t - \mathbf{x}^* \right\rangle \\
f(\mathbf{x}^{t+1}) &\leq f(\mathbf{x}^t) + \left\langle g^t, \mathbf{x}^{t+1} - \mathbf{x}^t) \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 - \frac{\alpha}{2} D_t^2 \\
&\leq f(\mathbf{x}^*) + \left\langle g^t, \mathbf{x}^t - \mathbf{x}^* \right\rangle + \left\langle g^t, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 - \frac{\alpha}{2} D_t^2 \\
\Phi^{t+1} &\leq \left\langle g^t, \mathbf{x}^t - \mathbf{x}^* \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 - \frac{\alpha}{2} D_t^2 \\
&\leq \frac{1}{\eta_t} \left\langle \mathbf{x}^t - \mathbf{z}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 - \frac{\alpha}{2} D_t^2 \\
&\leq \frac{1}{\eta_t} \left\langle \mathbf{x}^t - \mathbf{x}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{L}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 - \frac{\alpha}{2} D_t^2 \\
\Phi^{t+1} &\leq \frac{1}{2\eta_t} (D_t^2 - D_{t+1}^2) + \left(\frac{L}{2} - \frac{1}{2\eta_t} \right) \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
0 \leq \Phi^{t+1} &\leq \frac{L}{2} (D_t^2 - D_{t+1}^2) - \frac{\alpha}{2} D_t^2 \qquad [\text{By taking } \eta_t = \frac{1}{L}]
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \quad D_{t+1}^2 \cdot \frac{L}{2} &\leq \frac{L}{2} D_t^2 - \frac{\alpha}{2} d_t^2 \\
D_{t+1}^2 &\leq \frac{2}{L} \frac{(L-\alpha)}{2} D_t^2 \quad = \left(1 - \frac{\alpha}{L}\right) D_t^2 \\
D_T^2 &\leq \left(1 - \frac{\alpha}{L}\right)^T D_0^2 \\
&\leq e^{-T\alpha/2} D_0^2
\end{aligned}
$$

We find that there is an exponential decrease in distance to the optimum. This is called linear rate of convergence because in binary expansion, the error term will loose one significant digit in constant number of iterations. In other words, the number of zeros after decimal in error will increase at a linear rate.

6 Summary

following table concludes the convergence bounds for Projected gradient descent in different scenarios. (Pattanaik and Kar, 2015)

Function type	Convergence	Selector	Example
Cvx + Bounded Gradient	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$	Average	Perceptron
SC + Bounded Gradient	$\mathcal{O}\left(\frac{\log T}{T}\right)$ or $\mathcal{O}\left(\frac{1}{T}\right)$	Avg. /weighted avg.	SVM
Cvx + SS	$\mathcal{O}\left(\frac{1}{T}\right)$	Final Value	Least square
SC+SS	$\mathcal{O}\left(e^{-T}\right)$	Final Value	Ridge regression

Table 1: Comparing the convergence Rates

References

Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2009.

Anay Pattanaik and Purushottam Kar. Introduction to convex optimization ii, lecture notes. *http://web.cse.iitk.ac.in/users/purushot/courses/olo/2015-16-w/material/scribes/lec9.pdf*, 2015.