Lecture 14

Instructor: Quanquan Gu *Date: Oct 12^{th}*

Last time we introduced a class of functions which has Lipschitz continuous gradient, and its property.

Lemma 1 *Let a function f has L-Lipschitz continuous gradient over* $\mathbf{dom} f$*, then for any* $\mathbf{x}$, $\mathbf{y} \in \mathbf{dom} f$*, we have*

$$|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|_2^2.$$

Proof: Let $g(t) \equiv f\big(\mathbf{y} + t(\mathbf{x} - \mathbf{y})\big)$. From calculus (the Fundamental Theorem of Calculus) we know that

$$\int_0^1 g'(t) dt = g(1) - g(0) = f(\mathbf{x}) - f(\mathbf{y}).$$

It then follows that

$$\begin{aligned}
&|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \\
&= \Big| \int_0^1 \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) dt - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \Big| \\
&= \Big| \int_0^1 \Big(\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}) \Big)^\top (\mathbf{x} - \mathbf{y}) \; dt \Big| \\
&\leq \Big| \int_0^1 \big\| \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}) \big\|_2 \cdot \big\| \mathbf{x} - \mathbf{y} \big\|_2 \; dt \Big|,
\end{aligned}$$

where the inequality follows from the Cauchy-Schwartz inequality. Since f has L-Lipschitz continuous gradient, we then have

$$\begin{aligned}
|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| &\leq \Big| L \|\mathbf{x} - \mathbf{y}\|_2^2 \int_0^1 t dt \Big| \\
&= \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|_2^2.
\end{aligned}$$

■

Now, we introduce a new class of functions, which is L-smooth. Although L-smooth is a weaker condition than L-Lipschitz continuous gradient, it can be implied by L-Lipschitz continuous gradient when a function is convex.

Definition 1 (L-smooth) *A function f is L-smooth if for any* $\mathbf{x}$, $\mathbf{y} \in \mathbf{dom} f$*, it holds that*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|_2^2.$$

Remark 1 *If f is twice differentiable, then there is an equivalent, and perhaps easier, definition of smoothness: f is smooth if there exists a constant $L > 0$ such that $\nabla^2 f(x) \preceq L\mathbf{I}$, where $\mathbf{I}$ is an identity matrix. In other words, the largest eigenvalue of the Hessian of f is uniformly upper bounded by L everywhere.*

Then, we turn to an important property of L-smooth functions.

Lemma 2 *If a function f is L-smooth and convex, then for any $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$, we have*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2.$$

Proof: Define a function $g_{\mathbf{y}}(\mathbf{x}) \equiv f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$. Since $f(\mathbf{x})$ is convex, $g_{\mathbf{y}}(\mathbf{x}) \geq 0$. In particular, $g_{\mathbf{y}}(\mathbf{y}) = 0$. Thus, we have

$$g_{\mathbf{y}}(\mathbf{y}) = \min_{\mathbf{x}} g_{\mathbf{y}}(\mathbf{x}) \text{ and } \nabla g_{\mathbf{y}}(\mathbf{y}) = 0.$$

From the optimality of $\mathbf{y}$, it then follows that

$$\begin{aligned} g_{\mathbf{y}}(\mathbf{y}) &\leq \min_{\eta} g_{\mathbf{y}}(\mathbf{x} - \eta \nabla g_{\mathbf{y}}(\mathbf{x})) \\ &= \min_{\eta} f\big(\mathbf{x} - \eta \nabla g_{\mathbf{y}}(\mathbf{x})\big) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top \big(\mathbf{x} - \eta \nabla g_{\mathbf{y}}(\mathbf{x}) - \mathbf{y}\big). \end{aligned} \tag{1}$$

By definition of L-smooth, we have

$$f\big(\mathbf{x} - \eta \nabla g_{\mathbf{y}}(\mathbf{x})\big) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \big(- \eta \nabla g_{\mathbf{y}}(\mathbf{x})\big) + \frac{L}{2} \|\eta \nabla g_{\mathbf{y}}(\mathbf{x})\|_2^2.$$

It then follows from (1) that

$$\begin{aligned} g_{\mathbf{y}}(\mathbf{y}) \leq \min_{\eta} & f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \big(- \eta \nabla g_{\mathbf{y}}(\mathbf{x})\big) + \frac{L}{2} \|\eta \nabla g_{\mathbf{y}}(\mathbf{x})\|_2^2 \\ & - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top \big(\mathbf{x} - \mathbf{y} - \eta \nabla g_{\mathbf{y}}(\mathbf{x})\big) \\ = \min_{\eta} & g_{\mathbf{y}}(\mathbf{x}) + \frac{L}{2} \|\eta \nabla g_{\mathbf{y}}(\mathbf{x})\|_2^2 - \eta \nabla g_{\mathbf{y}}(\mathbf{x})^\top \big(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\big) \\ = \min_{\eta} & g_{\mathbf{y}}(\mathbf{x}) + \frac{L}{2} \eta^2 \|\nabla g_{\mathbf{y}}(\mathbf{x})\|_2^2 - \eta \|\nabla g_{\mathbf{y}}(\mathbf{x})\|_2^2. \end{aligned}$$

It is easy to show that the minimum solution η to the above quadratic function minimization problem is $g_{\mathbf{y}}(\mathbf{x}) - \|\nabla g_{\mathbf{y}}(\mathbf{x})\|_2^2 / 2L$. Thus, from our definition of $g_{\mathbf{y}}(\mathbf{x})$, it immediately follows

$$0 \leq f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) - \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2.$$

■

Now, let's go back to the gradient descent algorithm. We will show that if a function is L-smooth and convex, then we have faster convergence rate when we apply the gradient descent algorithm to optimize the function.

Theorem 1 *If a function f is L-smooth and convex, then the gradient descent algorithm with $0 \le \eta \le 1/L$ satisfies*

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \le \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2}{2\eta t}.$$

Proof: Let $\mathbf{x}^+ \equiv \mathbf{x}_{t+1}$ and $\mathbf{x} \equiv \mathbf{x}_t$. Since f is L-smooth, we have

$$f(\mathbf{x}^+) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{x}^+ - \mathbf{x}) + \frac{L}{2}\|\mathbf{x}^+ - \mathbf{x}\|_2^2.$$

Note that $\mathbf{x}^+ = \mathbf{x} - \eta\nabla f(\mathbf{x})$, it then follows that

$$\begin{aligned} f(\mathbf{x}^+) &\le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top\big(-\eta\nabla f(\mathbf{x})\big) + \frac{L}{2}\|-\eta\nabla f(\mathbf{x})\|_2^2 \\ &= f(\mathbf{x}) - \Big(1 - \frac{L\eta}{2}\Big)\eta\|\nabla f(\mathbf{x})\|_2^2 \\ &\le f(\mathbf{x}) - \frac{1}{2}\eta\|\nabla f(\mathbf{x})\|_2^2, \end{aligned} \tag{2}$$

where the last inequality follows from $0 \le \eta \le 1/L$. Note that (2) implies that the function value is monotonically decreasing.

In addition, since $f(\mathbf{x})$ is convex,

$$f(\mathbf{x}^*) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{x}^* - \mathbf{x}). \tag{3}$$

Combining (2) and (3), we get

$$f(\mathbf{x}^+) \le f(\mathbf{x}^*) - \nabla f(\mathbf{x})^\top(\mathbf{x}^* - \mathbf{x}) - \frac{1}{2}\eta\|\nabla f(\mathbf{x})\|_2^2.$$

Substituting $\nabla f(\mathbf{x})$ in the above inequality with $\frac{1}{\eta}(\mathbf{x} - \mathbf{x}^+)$, we have

$$\begin{aligned} f(\mathbf{x}^+) &\le f(\mathbf{x}^*) - \frac{1}{\eta}(\mathbf{x} - \mathbf{x}^+)^\top(\mathbf{x}^* - \mathbf{x}) - \frac{1}{2}\eta\Big\|\frac{1}{\eta}(\mathbf{x} - \mathbf{x}^+)\Big\|_2^2 \\ &= f(\mathbf{x}^*) - \frac{1}{\eta}(\mathbf{x} - \mathbf{x}^+)^\top(\mathbf{x}^* - \mathbf{x}) - \frac{1}{2\eta}\Big\|(\mathbf{x} - \mathbf{x}^+)\Big\|_2^2. \end{aligned}$$

Recall that $-2\mathbf{a}^\top\mathbf{b} = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - \|\mathbf{a} + \mathbf{b}\|_2^2$. Then, setting $\mathbf{a} = (\mathbf{x} - \mathbf{x}^+)$ and $\mathbf{b} = (\mathbf{x}^* - \mathbf{x})$ gives

$$\begin{aligned} f(\mathbf{x}^+) &\le f(\mathbf{x}^*) + \frac{1}{2\eta}\Big(\|\mathbf{x} - \mathbf{x}^+\|_2^2 + \|\mathbf{x}^* - \mathbf{x}\|_2^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_2^2\Big) - \frac{1}{2\eta}\Big\|(\mathbf{x} - \mathbf{x}^+)\Big\|_2^2 \\ &= f(\mathbf{x}^*) + \frac{1}{2\eta}\Big(\|\mathbf{x} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_2^2\Big). \end{aligned}$$

Note that this inequality holds for any positive integer t. Specifically, we have

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \le \frac{1}{2\eta}\Big(\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\Big),$$

$$\cdots$$

$$f(\mathbf{x}_2) - f(\mathbf{x}^*) \le \frac{1}{2\eta}\Big(\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_2 - \mathbf{x}^*\|_2^2\Big).$$

Adding these inequalities gives

$$\sum_{i=2}^{t+1} f(\mathbf{x}_i) - tf(\mathbf{x}^*) \leq \frac{1}{2\eta}\Big(\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\Big) \leq \frac{1}{2\eta}\Big(\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2\Big).$$

Recall from (2) we have proven that the function value is monotonically decreasing, so $tf(\mathbf{x}_{t+1}) \leq \sum_{i=2}^{t+1} f(\mathbf{x}_i)$. Therefore, the inequality above leads to

$$tf(\mathbf{x}_{t+1}) - tf(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2}{2\eta}$$

Dividing both sides of this inequality by t completes the proof. ■