IE 598: Big Data Optimization **Fall 2016**

Lecture 8: Gradient Descent – September 15

Lecturer: Niao He *Scriber: Juho Kim*

Overview: In this lecture, we discussed the concept of the rate of convergence. We also started the discussion about smooth convex optimization. As the beginning of this, we studied the gradient descent method.

8.1 Recap

Recall that a typical form of an optimization problem is

$$\begin{aligned} \underset{x\in\mathcal{X}}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0,\ i = 1, \dots, m. \end{aligned}$$

A black-box oriented algorithm generates $x_1, x_2, \dots$ in a way that x_{t+1} depends on local information gathered along $x_1, x_2, \dots, x_t$.

We can define a measure of error $\mathcal{E}(x_t)$ evaluate the quality of solution obtained from any optimization algorithm. The error function should be nonnegative, and it goes to zero as $x_t \to x_*$. For example, we can use the following measures as our error function.

1. $\mathcal{E}(x_t) = \inf_{x_* \in \mathcal{X}_*} \|x_t - x_*\|$
2. $\mathcal{E}(x_t) = \max\{f(x_t) - f(x_*),\ [g_1(x_t)]_+,\ \dots,\ [g_m(x_t)]_+\}$

8.2 Convergence Rate

Consider that $\lim_{t\to\infty} \frac{\mathcal{E}(x_{t+1})}{\mathcal{E}(x_t)^p} \le q$.

1. Linear convergence

   If $p = 1$ and $q \in (0, 1)$, the convergence rate is linear. For example, $\mathcal{E}(x_t) = O(e^{-at})\ (a > 0)$ converges linearly.
   In this case, the number of iterations required to obtain a solution within error ϵ is of order $O(\log \frac{1}{\epsilon})$.

2. Sublinear convergence

   If $p = 1$ and $q = 1$, the convergence rate is sublinear, which is slower than linear convergence. For example, $\mathcal{E}(x_t) = \frac{1}{t^b}\ (b > 0)$ converges sublinearly.
   In this case, the number of iterations required to obtain a solution within error ϵ is of order $O(\frac{1}{\epsilon^{1/b}})$.

3. Superlinear convergence

   If $p = 1$ and $q = 0$, the convergence rate is superlinear, which is faster than any geometric decay. For instance, $\mathcal{E}(x_t) = O(e^{-at^2})\ (a > 0)$ converges superlinearly.

4. Convergence of order p

   If $p = 2$ and $q > 0$, the convergence rate is quadratic (e.g., Newton's method). If $p = 3$ and $q > 0$, the convergence rate is cubic. $\mathcal{E}(x_t) = O(e^{-at^p})\,(a > 0)$ is the example of this convergence rate.
   In this case, the number of iterations required to obtain a solution within error ϵ is of order $O(\log(\log \frac{1}{\epsilon}))$.

For comparison of the rate of convergence, it is convenient to use $log(\mathcal{E}(x_t))$ instead of $\mathcal{E}(x_t)$ with respect to the number of iterations.

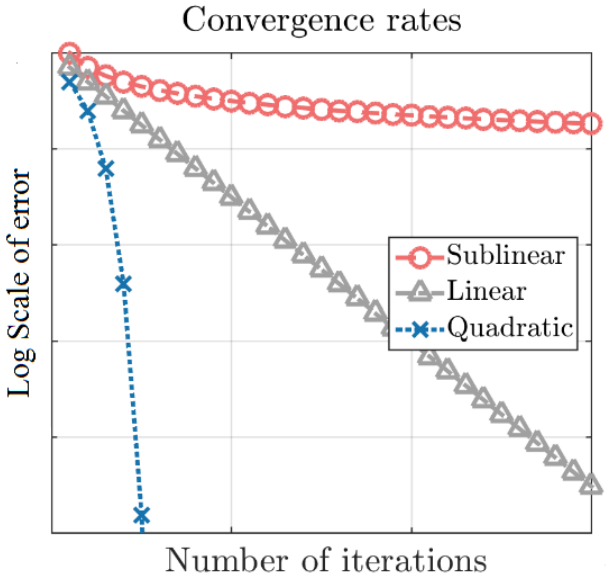

Figure 8.1: Comparison of the rate of convergence

The convergence rate of an optimization algorithm depends heavily on the structure of the problem (e.g. strong convexity, smoothness). We will first focus on smooth convex optimization.

8.3 Smooth Convex Optimization

Definition 8.1 ***(L-smooth)***
f is L-smooth (with a constant $L > 0$) on $\mathcal{X}$ if f is continuously differentiable and $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ for all $x, y \in \mathcal{X}$.

Proposition 8.2 *The followings are equivalent.*
(a) f is convex and L-smooth.
(b) $0 \leq f(y) - f(x) - \nabla f(x)^T(y - x) \leq \frac{L}{2}\|x - y\|_2^2$ for all $x, y \in \mathcal{X}$.
(c) $f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$ for all $x, y \in \mathcal{X}$.
(d) $\{\nabla f(x) - \nabla f(y)\}^T(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$ for all $x, y \in \mathcal{X}$.

Proof: (a) ⇒ (b) By the fundamental theorem of calculus,

$$\begin{aligned}
f(y) - f(x) - \nabla f(x)^T(y-x) &= \int_0^1 \nabla f(x + t(y-x))^T(y-x)\,dt - \nabla f(x)^T(y-x) \\
&= \int_0^1 [\nabla f(x + t(y-x)) - \nabla f(x)]^T(y-x)\,dt \\
&\le \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\|_2 \|y-x\|_2\,dt \quad \text{(by Cauchy-Schwarz inequality)} \\
&\le \int_0^1 L\|t(y-x)\|_2\|y-x\|_2\,dt \quad \text{(by L-smoothness of } f\text{)} \\
&= L\|(y-x)\|_2^2 \int_0^1 t\,dt \\
&= \frac{L}{2}\|(y-x)\|_2^2
\end{aligned}$$

(b) ⇒ (c) Let $z = y + \frac{1}{L}(\nabla f(x) - \nabla f(y))$

$$\begin{aligned}
f(y) - f(x) &= f(y) - f(z) + f(z) - f(x) \\
&\ge -\nabla f(y)^T(z-y) - \frac{L}{2}\|y-z\|_2^2 + \nabla f(x)^T(z-x) \\
&= \nabla f(x)^T(y-x) - \{\nabla f(x) - \nabla f(y)\}^T(y-z) - \frac{L}{2}\|y-z\|_2^2 \quad \text{(by plugging in z)} \\
&= \nabla f(x)^T(y-x) + \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2 \\
&= \nabla f(x)^T(y-x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2
\end{aligned}$$

(c) ⇒ (d) Suppose that

$$\begin{aligned}
f(y) &\ge f(x) + \nabla f(x)^T(y-x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\
f(x) &\ge f(y) + \nabla f(y)^T(x-y) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2
\end{aligned}$$

By summing up two inequalities, we can obtain

$$[\nabla f(x) - \nabla f(y)]^T(x-y) \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

(d) ⇒ (a) Suppose that $[\nabla f(x) - \nabla f(y)]^T(x-y) \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$.

By Cauchy-Schwarz inequality,

$$\|\nabla f(x) - \nabla f(y)\|_2\|x-y\|_2 \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

which implies f is L-smooth. The convexity of f is due to the following claim.
Therefore, (a), (b), (c), and (d) are equivalent. ■

Claim 8.3 *Suppose f is differentiable, then f is convex if and only if $[\nabla f(x) - \nabla f(y)]^T(x-y) \ge 0$.*

Proof:
($\Rightarrow$) Suppose f is convex. Then,

$$f(x) \geq f(y) + \nabla f(y)^T(x-y) \quad \textit{and} \quad f(y) \geq f(x) + \nabla f(x)^T(y-x)$$

which imply $[\nabla f(x) - \nabla f(y)]^T(x-y) \geq 0$.

($\Leftarrow$) Suppose $[\nabla f(x) - \nabla f(y)]^T(x-y) \geq 0$.

$$f(y) - f(x) - \nabla f(x)^T(y-x) = \int_0^1 \frac{1}{t}\{\nabla f(x+t(y-x)) - \nabla f(x)\}^T t(y-x)\, dt \geq 0$$

which implies f is convex. ■

Remark. In general, we call an operator $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ *monotonic* if $\langle F(x) - F(y),\, x-y\rangle \geq 0$. The above claim indicates that the gradient of a continuously differentiable convex function is an monotone operator.

8.4 Gradient Descent

Consider an unconstrained optimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)$$

where f is L-smooth and convex.

Gradient Descent (GD) Given a starting point $x_0 \in \operatorname{dom} f$, and step-size $\gamma > 0$, GD works as follows,

$$x_{t+1} = x_t - \gamma \nabla f(x_t), \quad t = 0, 1, 2, \ldots$$

until stopping criterion is satisfied.

Observation 1 The gradient descent step can be regarded as minimizing a quadratic approximation of the objective function,

$$\begin{aligned} x_{t+1} &= \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \; \{f(x_t) + \nabla f(x_t)^T(x-x_t) + \frac{1}{2\gamma}\|x-x_t\|_2^2\} \\ &= \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \; \left\{\frac{1}{2\gamma}\|x-(x_t-\gamma\nabla f(x_t)\|_2^2\right\} \end{aligned}$$

When $\gamma \leq \frac{1}{L}$, the quadratic approximation is indeed an upper bound of the objective function.

Observation 2 The gradient descent step strictly reduces the objective function value at each iteration,

$$f(x_{t+1}) - f(x_t) \leq \nabla f(x_t)^T(-\gamma\nabla f(x_t)) + \frac{L}{2}\|-\gamma\nabla f(x_t)\|_2^2 = -\gamma(1-\frac{L}{2}\gamma)\|\nabla f(x_t)\|_2^2$$

It makes sense to seek $\gamma > 0$ that maximizes $\gamma(1-\frac{L}{2}\gamma)$ to ensure the most reduction. This takes place when $\gamma = \frac{1}{L}$ and moreover,

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L}\|\nabla f(x_t)\|_2^2.$$

Theorem 8.4 *Let f be convex and L-smooth, x_* be an optimal solution. With $\gamma = \frac{1}{L}$ for all $\gamma > 0$, the gradient descent satisfies*

$$f(x_t) - \min_{x \in \mathbb{R}^n} f(x) \leq \frac{2L\|x_0 - x_*\|_2^2}{t}$$

Proof: ([Nes03]) First of all, we have the following

(i) $f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L}\|\nabla f(x_t)\|_2^2$ from Observation 2.

(ii) $\|x_{t+1} - x_*\|_2 \leq \|x_t - x_*\|_2$
This is because

$$\begin{aligned}
\|x_{t+1} - x_*\|_2^2 &= \|x_t - \frac{1}{L}\nabla f(x_t) - x_*\|_2^2 \\
&= \|x_t - x_*\|_2^2 - \frac{2}{L}\nabla f(x_*)^T(x_t - x_*) + \frac{1}{L^2}\|\nabla f(x_t)\|_2^2 \\
&\leq \|x_t - x_*\|_2^2 - \frac{1}{L^2}\|\nabla f(x_t)\|_2^2 \quad \text{since } \nabla f(x_*)^T(x_t - x_*) \geq \frac{1}{L}\|\nabla f(x_t)\|_2^2 \\
&\leq \|x_t - x_*\|_2^2
\end{aligned}$$

(iii) $\|\nabla f(x_t)\|_2 \geq \frac{f(x_t) - f(x_*)}{\|x_t - x_*\|}$
since $f(x_t) - f(x_*) \leq \nabla f(x_t)^T(x_t - x_*) \leq \|\nabla f(x_t)\|_2\|x_t - x_*\|_2$.

By combining (i)-(iii), we arrive at

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L}\Big[\frac{f(x_t) - f(x_*)}{\|x_0 - x_*\|_2}\Big]^2.$$

Let $\epsilon_t = f(x_t) - f(x_*)$ and $\beta = \frac{1}{2L\|x_0 - x_*\|_2^2}$.

$$[f(x_{t+1}) - f(x_*)] - [f(x_t) - f(x_*)] = \epsilon_{t+1} - \epsilon_t \leq -\frac{1}{2L}\frac{\epsilon_t^2}{\|x_0 - x_*\|_2^2} = -\beta\epsilon_t^2$$

$$\frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t+1}} \leq -\beta\frac{\epsilon_t}{\epsilon_{t+1}} \leq -\beta$$

$$\begin{aligned}
&\Rightarrow \ \frac{1}{\epsilon_t} + \beta \leq \frac{1}{\epsilon_{t+1}} \\
&\Rightarrow \ \frac{1}{\epsilon_0} + \beta t \leq \frac{1}{\epsilon_t} \\
&\Rightarrow \ \beta t \leq \frac{1}{\epsilon_t} = \frac{1}{f(x_t) - f(x_*)}
\end{aligned}$$

which implies $f(x_t) - f(x_*) \leq \frac{2L\|x_0 - x_*\|_2^2}{t}$. ■

References

[Nes03] Y. Nesterov, "Introductory Lectures on Convex Optimization, A Basic Course," *Springer*, 2003.

[BV04] S. Boyd and L. Vandenberghe, "Convex Optimization," *Cambridge University Press*, 2004.