The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Convex Optimization — Fall 2019

Lecture 25

Aryan Mokhtari Monday, December 2, 2019.

Goal: In this lecture we talk about descent methods and in particular we focus on the convergence analysis of the gradient descent method.

1 Recap

1.1 Strong Convexity

Definition 1. *If there exists a constant $m > 0$ such that*

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

*for all $\mathbf{x}, \mathbf{y} \in S$, then the function f is m-*strongly convex *on S.*

Definition 2. *A twice differentiable function is m-strongly convex if*

$$\nabla^2 f(\mathbf{x}) \succeq m\mathbf{I}$$

Lemma 1. *Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ that is m-strongly convex. Then, the objective function sub-optimality $f(\mathbf{x}) - f^*$ is bounded above by*

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2$$

Lemma 2. *Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ that is m-strongly convex. Then, the optimality distance $\|\mathbf{x} - \mathbf{x}^*\|_2$ is bounded above by*

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \quad \leq \quad \frac{2}{m}\|\nabla f(\mathbf{x})\|_2 \tag{2}$$

where $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$ is the unique minimizer of f.

1.2 Smoothness (Lipschitz continuous gradients)

Definition 3. *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called M-smooth or has M-Lipschitz continuous gradients if for some $M > 0$ we have*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|.$$

Definition 4. *A twice differentiable function is M-smooth if*

$$-M\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I} \quad \Leftrightarrow \quad \|\nabla^2 f(\mathbf{x})\| \leq M$$

Lemma 3. *If f is M-smooth, then the following condition holds*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{M}{2}\|\mathbf{y} - \mathbf{x}\|^2 \tag{3}$$

Lemma 4. *If the function f is smooth, then*

$$f(\mathbf{x}) - f^* \geq \frac{1}{2M}\|\nabla f(\mathbf{x})\|_2^2$$

2 Gradient Descent Algorithm

2.1 Descent Methods

Definition 5. *A sequence* $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$ *descends if* $f(\mathbf{x}_1) > f(\mathbf{x}_2) > f(\mathbf{x}_3) \cdots$

Let us consider the following equation:

$$\mathbf{x}^+ = \mathbf{x} + \eta \Delta \mathbf{x} \tag{4}$$

In the above equation, $\mathbf{x}^+$ represents the new point while $\mathbf{x}$ denotes the point under consideration, η is the step size and $\Delta \mathbf{x}$ is the direction vector.
Intuitively, at each iterate, we would like to ensure that the next step taken by this algorithm results in a smaller function value at the next iterate. Thus, a descent method is the one in which $f(\mathbf{x}^+) < f(\mathbf{x})$ for every step.

2.2 Gradient descent

In the gradient descent algorithm, the descent direction is indicated by $\Delta \mathbf{x} = -\nabla f(\mathbf{x})$ while the step size is η. The motivation for using $\Delta \mathbf{x} = -\nabla f(\mathbf{x})$ lies in the fact that of all the directions available, $-\nabla f(\mathbf{x})$ represents the steepest direction of descent. The update equation for the gradient descent algorithm can be represented using the following equation :

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k) \tag{5}$$

In the above equation, the step size is assumed to be constant. In some case, the stepsize have to change over time to ensure convergence.

2.3 Advantages of gradient descent

1) Suitable for higher dimensions

   The gradient descent algorithm works for higher dimensions as well. The only problem may lie in computing the gradient in each dimension since the computations increase linearly with the number of dimensions.

2) Convergence

   The global convergence of the gradient descent algorithm is almost always guaranteed under standard assumptions and proper values of η.

3) Extensions

   The gradient descent technique has its analogue in stochastic optimization where stochastic gradient descent is used for minimizing the objective function that is written as a sum of differentiable functions.

2.4 Issues in Gradient descent

1) $\nabla f(\mathbf{x})$ does not exist for some x or is difficult to compute.

   The gradient of the function may not exist at certain points. In that case, we can use subgradient descent algorithms which will be covered in the next course. The gradient maybe difficult to compute as well for certain convex functions. Coordinate descent techniques may be used in that case.

2) Selecting the best step size

One of the most important parameters to control in gradient descent is the step size η. Very small values of η will cause our algorithm to converge very slowly and thus the computational time required is very large. On the other hand, larger values of η could cause our algorithm to overshoot the minima and may lead to oscillations around the optimum. Thus, there exists a tradeoff in choosing the best value of η between larger time required for computation and oscillations around the minima. This is illustrated in Figure 1 where for lower values of step size the function takes a long time to converge whereas for higher values the function oscillates around the optimum. As a result, it becomes very important to pick the best value of η which will deal with this trade-off. This will be mainly covered in the next lecture.

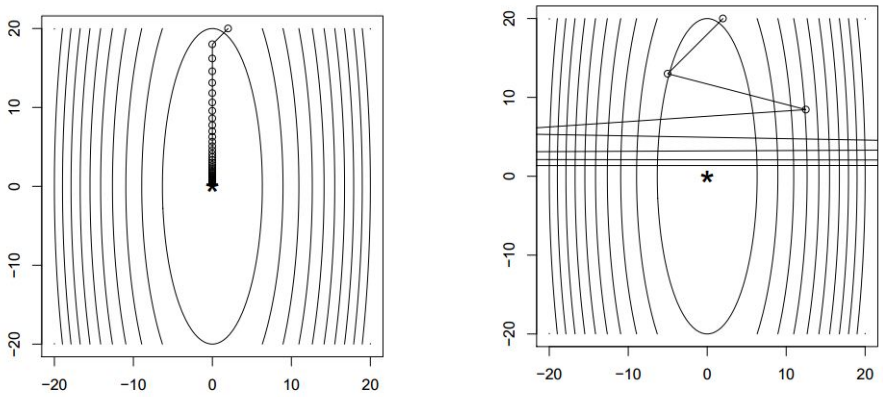

Figure 1: (left) a too small step size that leads to slow convergence; (right) a too large step size that leads to oscillations

3 Linear Convergence of GD

Definition 6. *A function $f(\mathbf{x})$ can be considered linearly convergent to f^* if:*

$$\lim_{k \to +\infty} \frac{f(\mathbf{x}_{k+1}) - f^*}{f(\mathbf{x}_k) - f^*} = c \, , \; c \in (0, 1) \tag{6}$$

This is helpful as it specifies a steady rate of convergence, rather than guaranteeing convergence but with no specification as to how long it might take.

Theorem 1. *For $f(\mathbf{x})$ so that $mI \preceq \nabla^2 f(\mathbf{x}) \preceq MI$, gradient descent with a step size of $\eta = \frac{1}{M}$ for k iterations will result in:*

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{m}{M}\right)^k (f(\mathbf{x}_0) - f^*) \tag{7}$$

Proof. The steps of the proof follows:

1. Since the function f is M-smooth we can write

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

from lemma 3.

2. Use this equation for a single step from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$, and recall $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)$:

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{M}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 && (8) \\ &\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (-\eta \nabla f(\mathbf{x}_k)) + \frac{M}{2} \| - \eta \nabla f(\mathbf{x}_k)\|^2 && (9) \\ &\leq f(\mathbf{x}_k) - \eta \|\nabla f(\mathbf{x}_k)\|^2 + \frac{M}{2} \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 && (10) \\ &\leq f(\mathbf{x}_k) - \eta \left(1 - \frac{\eta}{2} M\right) \|\nabla f(\mathbf{x}_k)\|^2 && (11) \end{aligned}$$

3. Setting step size $\eta = \frac{1}{M}$:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2M} \|\nabla f(\mathbf{x}_k)\|^2 \tag{12}$$

4. Since the function is m-strongly convex we have $f(\mathbf{x}^*) \geq f(\mathbf{x}) - \frac{1}{2m} \|\nabla f(\mathbf{x})\|^2$ from lemma 1, rearrange so:

$$- \|\nabla f(\mathbf{x})\|^2 \leq 2m(f(\mathbf{x}^*) - f(\mathbf{x})) \tag{13}$$

5. Replace $-\|\nabla f(\mathbf{x}_k)\|^2$ by its upper bound $2m(f(\mathbf{x}^*) - f(\mathbf{x}_k))$ to obtain

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \frac{1}{2M} (2m(f(\mathbf{x}^*) - f(\mathbf{x}_k)) && (14) \\ f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \frac{m}{M} f(\mathbf{x}_k) + \frac{m}{M} f(\mathbf{x}^*) && (15) \\ f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) &\leq \left(1 - \frac{m}{M}\right) (f(\mathbf{x}_k) - f(\mathbf{x}^*)) && (16) \end{aligned}$$

6. This is clearly linear convergence to $f(\mathbf{x}^*)$ with a rate of convergence of $c = 1 - \frac{m}{M}$. Were this iteration repeated k times, the difference from the minimum would be multiplied by c each time, resulting in Equation 7.

□

This result implies that after k iterations we have

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{m}{M}\right)^k (f(\mathbf{x}_0) - f^*)$$

Hence, to ensure that $f(\mathbf{x}_k) - f^* \leq \epsilon$ we need to ensure that

$$\left(1 - \frac{m}{M}\right)^k (f(\mathbf{x}_0) - f^*) \leq \epsilon$$

which happens after at most

$$\frac{\log(\frac{f(\mathbf{x}_0) - f^*}{\epsilon})}{\log \frac{1}{1 - \frac{m}{M}}}$$

iterations which can be upper bounded by

$$\frac{M}{m} \log \left(\frac{f(\mathbf{x}_0) - f^*}{\epsilon}\right)$$

since $(-\log(1 - x) > x)$ for $x \leq 1$.

3.1 Convergence of Gradient Descent

Although, the importance of step sizes in gradient descent method can be explained intuitively with simple figure examples (as shown in Figure 1), it is better to have a formal condition on a step size to guarantee the convergence. There exists a theorem which specifies the condition of a constant step size to have a converging behavior $\mathbf{x}_k \to \mathbf{x}^*$ regardless of the initial point. Before presenting the theorem, a useful condition of a function is required to be defined.
The following Theorem 2 provides a proper condition of step size for gradient descent when the step size is assumed to be fixed as a constant value η.

Theorem 2. *If f is M-Lipschitz and $\exists$ an optimum $\mathbf{x}^*$, i.e., $f^* = f(\mathbf{x}^*) = \min_{\mathbf{x}} f(\mathbf{x}) > -\infty$, then the gradient descent algorithm with fixed step size satisfying $\eta < \frac{2}{M}$ will converge to a stationary point (or $\mathbf{x}^*$ when f is convex) from any initial point.*

Proof. As we showed in the previous proof, if f is M-smooth then

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \eta \left(1 - \frac{\eta}{2} M\right) \|\nabla f(\mathbf{x}_k)\|^2$$

This leads to:

$$\|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{1}{\eta(1 - \frac{\eta M}{2})}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}))$$

Now, let's denote $\mathbf{x}_k$ a k^{th} value of $\mathbf{x}$ in gradient descent, where $\mathbf{x}_0$ is an initial value of $\mathbf{x}$. Then the above equation can be written as follows for for all $k \geq 0$.

$$\|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{1}{\eta(1 - \frac{\eta M}{2})}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}))$$

Summing these equations for $k = 0, \cdots, N-1$ gives:

$$\begin{aligned} \sum_{k=0}^{N-1} \|\nabla f(\mathbf{x}_k)\|^2 &\leq \frac{1}{\eta(1 - \frac{\eta}{2} M)}(f(\mathbf{x}_0) - f(\mathbf{x}^{(N)}) \\ &\leq \frac{1}{\eta(1 - \frac{\eta}{2} M)}(f(\mathbf{x}_0) - f^*) \end{aligned}$$

The last inequality comes from the fact that f^* is an optimal value. Since RHS of the inequality is a finite value, the summation of N sequences is bounded by finite number. This implies that $\lim_{k \to \infty} \nabla f(\mathbf{x}_k) = 0$. So, when f is a convex function, $\mathbf{x}_k \to \mathbf{x}^*$ as the gradient converges to 0. □