Optimization

Fourth Class

2020 - 2021

By

Dr. Jawad Mahmoud Jassim

Dept. of Math.

Education College for Pure Sciences

University of Basrah

Iraq

Chapter Two

Line Search

Lecture 2

3: Convergence Theory for Exact Line Search

The general form of an unconstrained optimization algorithm is as follows.

Algorithm (2): (General Form of Unconstrained Optimization Algorithm)

First: Initial step

Given $X_0 \epsilon R^n$, $0 \le \varepsilon \le 1$.

Second: $k^{th} - Step$

1: $Compute\ the\ descent\ direction\ d_k$.

2: Compute the step size $\alpha_k\ such\ that$

$$f(X_k + \alpha_k d_k) = \min_{\alpha \ge 0} f(X_k + \alpha d_k) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

3: Set $X_{k+1} = X_k + \alpha_k d_k \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$

4: If $\|g(X_{k+1})\| \le \varepsilon$ **stop, where** $g(X_{k+1}\ \ is\ the\ gradient\ vector\ at\ X_{k+1}$. **Otherwise repeated the above steps.**

Note (8):

We denote $\Phi(\alpha) = f(X_k + \alpha d_k) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (8)$

Obviously, we have from Algorithm (2) that

$\Phi(0) = f(X_k), \Phi(\alpha) \leq \Phi(0).$

Note (9):

The equation (6) in Algorithm (2) is to find the global minimizer of $\Phi(\alpha)$ which is rather difficult. Instead, we take α_k such that

$$\alpha_k = min\left\{\alpha \geq 0 :\ g_k^T d_k = 0\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (9)$$

$where\ g\ is\ the\ gradient\ vector\ of\ \Phi(\alpha)\ which\ is\ given\ in\ (8).$

Since by (6) and (9), we find the exact minimizer of $\Phi(\alpha)$ respectively. We say that (6) and (9) are exact line searches.

Note (10):

$Let\ \theta_k\ be\ the\ angle\ between\ d_k\ and - g_k\ , then$

$$\cos\theta_k = -\frac{d_k^T g_k}{\|d_k\|\ \|g_k\|} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ (10)$$

Theorem (2):

Let $\alpha_k > 0$ be the solution of the equation

$$f(X_k + \alpha_k d_k) = \min_{\alpha \geq 0} f(X_k + \alpha d_k), \text{ and } \|G(X_k + \alpha d_k)\| \leq M, where\ M\ is$$

$some\ positive\ number\ and\ G\ is\ the\ Hessian\ matrix$. Then

$$f(X_k) - f(X_k + \alpha_k d_k) \geq \frac{\left(d_k^T g(X_k)\right)^2}{2M\|d_k\|^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. \quad (11)$$

Note (11):

Theorem (2) means that $f(X_k + \alpha_k d_k) < f(X_k)$.

Definition (7): (Neighborhood)

Given a point $X \epsilon R^n$ and a $\delta > 0$. $The\ \delta - neighborhood$ of X is defined as $N_\delta(X) = \{ Y \epsilon R^n\colon\ \|Y - X\| < \delta\}$.

Definition (8): (Accumulation Point)

The point $X \epsilon D \subset R^n$ is said to be an accumulation point if for each $\delta > 0$, $D \cap N_\delta(X) \neq \emptyset, where\ \emptyset\ is\ an\ empty\ set$.

Definition (9): (Index Set)

Let W_a be the set of all words containing $the\ letter\ a$, W_b be the set of all words containing the letter b and $similarly\ for\ W_c\ to\ W_z$. The subscripts $a, b, c, \cdots, z\ are\ knows\ as$ indices. Then the set $I = \{a, b, c, \cdots, z\}$ is called the index set.

Theorem (3):

1: Let $f(X)$ be continuously differentiable function on an $open\ set\ D \subset R^n$.

2: Assume that the sequence generated by Algorithm (2) satisfies

$$f(X_{k+1}) \le f(X_k)\ and\ g(X)^T d_k \le 0, where\ g\ is\ the\ gradient\ vector.$$

3: Let $\hat{X} \epsilon D$ be an accumulation point of $\{X_k\}$ and K_1 be an index set with

$$K_1 = \left\{k : \lim_{k \to \infty} X_k = X\right\}.$$

4: Assume that there exists $M > 0$ such that $\|d_k\| < M\ for\ all\ k \epsilon\ K_1$.

Then:

1: If $\widehat{d_k}$ is any accumulation $point\ of\ \{d_k\}$, we have $g(\hat{X})^T \hat{d} = 0$.

2: If $f(X)\ is\ twice\ continuously\ differentiable\ function\ on\ D$,

$$we\ have\ \hat{d}^T G(\hat{X})\hat{d} \ge 0.$$

Definition (10): (Continuous Function)

Let $f: D \subset R^n \rightarrow R$. $Let\ X_0 \epsilon D$. We say that f is continuous function at X_0 if for every $\varepsilon > 0$, there exists $\delta = \delta(X_0, \varepsilon)$ such that if $X \epsilon D$ with $\|X - X_0\| < \delta$ implies $\|f(X) - f(X_0)\| < \varepsilon$.

Definition (11): (Uniformly Continuous)

Let $f: D \subset R^n \rightarrow R$. We say that f is uniformly continuous if $for\ all\ \varepsilon > 0$, there exists $\delta = \delta(\varepsilon)$ such that if $X, Y \epsilon D$ with $\|X - Y\| < \delta$ implies $\|f(X) - f(Y)\| < \varepsilon$.

In other words, f is uniformly continuous if it is continuous at each point $X_0 \epsilon D$ and the δ corresponding to each ε in the definition of continuity at X_0 can be the same for all $X_0 \epsilon D$.

For example,

$f(x) = x$ *is uniformly continuous function on real numbers*.

$f(x) = x^2, x \epsilon [-M, M], M > 0$ *is uniformly continuous , while*

$f(x) = x^2$ *is not uniformly continuous on the set of real numbers* .

Note (12):

Each uniformly continuous function is continuous.

Definition (12): (Lipschitz Continuity)

Let $f: D \subset R^n \rightarrow R$. We say that f is Lipschitz continuous function if there exists $M > 0$ such that $\|f(X) - f(Y)\| \leq M\|X - Y\|$ $for\ all\ X, Y \epsilon D$.

Note (13):

Every Lipschitz continuous function is uniformly continuous function.

Definition (13): (Level Set)

A set where the function takes a given constant value.

Theorem (4):

1: $\boldsymbol{Let\ g(X)\ be\ uniformly\ continuous\ on\ the\ level\ set}$

$\boldsymbol{L = \{X \epsilon R^n : f(X) \leq f(X_0)}$.

2: $\boldsymbol{Let\ the\ angle\ \theta_k\ between\ -g(X_k)\ and}$

$\boldsymbol{the\ direction\ d_k\ generated}$

$\boldsymbol{by\ Algorithm\ (2) is\ uniformly\ bounded\ a\ way\ from\ 90^{\circ}}$

$\boldsymbol{, i.e. satisfies\ \theta_k \leq \frac{\pi}{2} - \mu\ for\ some\ \mu > 0}$.

$\boldsymbol{Then\ g(X_k) = 0\ for\ some\ k; or\ f(X_k) \to -\infty ; or\ g(X_k) \to 0}$.

$\boldsymbol{where\ g\ is\ the\ gradient\ vector}$.

<u>Lemma (1):</u>

1: $\boldsymbol{Let\ f(X) be\ twice\ continuously\ differentiable\ in\ the\ neighborhood\ of\ the\ minimizer\ X^*.}$

2: $\boldsymbol{Assume\ that\ there\ exists\ \varepsilon > 0\ and\ M > m > 0,\ such\ that\ m\|Y\|^2 \leq Y^T G(X) Y \leq M\|Y\|^2\ , for\ all\ Y \epsilon R^n\ holds\ when\ \|X - X^*\| < \varepsilon.}$

$\boldsymbol{Then}$

1: $\boldsymbol{\frac{1}{2} m\|X - X^*\|^2 \leq f(X) - f(X^*) \leq \frac{1}{2} M\|X - X^*\|^2.}$

2: $\boldsymbol{\|g(X)\| \geq m\|X - X^*\|.}$

Where

$\boldsymbol{g(X)\ and\ G(X)\ are\ the\ gradient\ and\ Hessian\ matrix\ of\ f\ at\ X\ respectively.}$

Theorem (5):

1: Let the sequence $\{X_k\}$ generated by Algorithm (2) converges to the minimizer X^* $of\ f(X)$.

2: Let $f(X)$ be twice continuously differentiable in a neighborhood of X^*.

3: If there exists $\varepsilon > 0\ and\ M > m > 0\ such\ that$

$$m\|Y\|^2 \le Y^T G(X) Y \le M\|Y\|^2\ , for\ all\ Y \epsilon R^n\ holds\ when$$

$$\|X - X^*\| < \varepsilon.$$

Then $the\ sequence$ $\{X_k\}$, $at\ least$, converges linearly to X^*.

Where $G(X)$ is the Hessian $matrix\ of\ f\ at\ X$.