Numerical Optimization

Graduate Course

Unconstrained Smooth Optimization

Part I: Basic theories and differentials

Wen Huang

School of Mathematical Sciences
Xiamen University

Compiled on February 14, 2022

Preliminaries and Basic Theories

A descent method

Input: Initial iterate x_0;
Initial descent direction p_0 and set $k \leftarrow 0$;
while not accurate enough **do**
Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$ with an appropriate step size α_k;
Find a descent direction p_{k+1};
$k \leftarrow k + 1$;
end while

What is the definition of a descent direction?

Definition 1

Let $f \in C^1$ and $p \in \mathbb{R}^n$. p is a descent direction of f at x if for all sufficiently small $\alpha > 0$

$$f(x + \alpha p) < f(x)$$

Theorem 2

Let $f \in C^1$ and $p \in \mathbb{R}^n$. If

$$-p^T \nabla f(x_k) = \|p\|_2 \|\nabla f(x_k)\|_2 \cos\theta_k > 0$$

then p is a descent direction at x.

Note that a direction p is descent for $f \in C^1$ if the angle between p and $-\nabla f(x_k)$ is acute.

To prove this theorem, we need the Taylor's theorem of multiple variables.

Taylor's Theorem

Theorem 3 (Taylor's Theorem)

Let f be a real-value function $f : \mathbb{R}^n \to \mathbb{R}$.

- *If $f \in C^1$ and $p \in \mathbb{R}^n$ then for some $0 < \tau < 1$*

$$f(x+p) = f(x) + \nabla f(x+\tau p)^T p$$

- *If $f \in C^2$ and $p \in \mathbb{R}^n$ then for some $0 < \tau < 1$*

$$f(x+p) = f(x) + \nabla f(x)^T p + 0.5 p^T \nabla^2 f(x+\tau p) p$$

and

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+\tau p) p \, d\tau$$

The proofs are omitted here.

Minimizers

Definition 4

The point $x^* \in \mathbb{R}^n$ is a global minimizer if $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$.

Definition 5

The point $x^* \in \mathbb{R}^n$ is a local minimizer if $f(x^*) \leq f(x)$ for all $x \in \mathcal{N}_{x^*}$ where $\mathcal{N}_{x^*}$ is a neighborhood of x^*. Further, x^* is

- a strict local minimizer if $f(x^*) < f(x)$ for all $x \in \mathcal{N}_{x^*}$; and
- an isolated local minimizer if x^* is the only local minimizer in $\mathcal{N}_{x^*}$.

Definition 6

The point $x^* \in \mathbb{R}^n$ is a stationary point if $\nabla f(x^*) = 0$.

Theorem 7 (First order necessary condition)

Suppose $f \in C^1$ in a neighborhood of x^*. If x^* is a local minimizer then $\nabla f(x^*) = 0$, i.e., x^* is a stationary point.

Theorem 8 (Second order necessary condition)

Suppose $f \in C^2$ in a neighborhood of x^*. If x^* is a local minimizer then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

Theorem 9 (Second order sufficient condition)

Suppose $f \in C^2$ *in a neighborhood of* x^*. *If* $\nabla f(x^*) = 0$ *and* $\nabla^2 f(x^*)$ *is positive definite then* x^* *is a strict local minimizer.*

Definition 10

A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is convex if for any two points, x and y, in the domain we have for any $0 \leq \alpha \leq 1$

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

Also, $f(x)$ is concave if $-f(x)$ is convex.

Theorem 11 (Sufficient condition of global minimizer)

If f is convex then any local minimizer is a global minimizer. If, in addition, $f \in C^1$ then any stationary point is a global minimizer.

Rates of Convergence

Desire convergence properties

- Global convergence (to a stationary point from any initial point)
- Fast convergence speed
  - $x_k \to x_*$, let $e_k = \|x_k - x_*\|$;
  - $f_k \to f_*$, let $e_k = |f_k - f_*|$;

- Q-sublinear $\frac{e_{k+1}}{e_k} \to 1$
- Q-linear $\limsup \frac{e_{k+1}}{e_k} < \delta < 1$
- Q-superlinear $\frac{e_{k+1}}{e_k} \to 0$
- Q-quadratic $\limsup \frac{e_{k+1}}{e_k^2} < C$

Rates of Convergence

Desire convergence properties

- Global convergence (to a stationary point from any initial point)
- Fast convergence speed
  - $x_k \to x_*$, let $e_k = \|x_k - x_*\|$;
  - $f_k \to f_*$, let $e_k = |f_k - f_*|$;

- Q-sublinear $\frac{e_{k+1}}{e_k} \to 1$
- Q-linear $\limsup \frac{e_{k+1}}{e_k} < \delta < 1$
- Q-superlinear $\frac{e_{k+1}}{e_k} \to 0$
- Q-quadratic $\limsup \frac{e_{k+1}}{e_k^2} < C$

Suppose $e_k \le \epsilon_k$

- ϵ_k Q-sublinear $\Rightarrow$ R-sublinear
- ϵ_k Q-linear $\Rightarrow$ R-linear
- ϵ_k Q-superlinear $\Rightarrow$ R-superlinear
- ϵ_k Q-quadratic $\Rightarrow$ R-Quadratic

Gradient and Hessian

Definition 12

If $f : \mathbb{R}^n \to \mathbb{R} : x \mapsto f(x)$ then

- the gradient of $f \in C^1$, denoted, $\nabla f(x) \in \mathbb{R}^n$, is the vector with i-th element,
$$\frac{\partial f}{\partial x_i}(x)$$
- the Hessian of $f \in C^2$ denoted, $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$, is the symmetric matrix with i, j-element,
$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

- Example: $f(x) = \frac{1}{2}x^T Ax$, where $A = A^T$

- Example: $f(x) = \frac{1}{2}x^T Ax$, where $A = A^T$
    - By definition

Gradient and Hessian

Examples

- Example: $f(x) = \frac{1}{2}x^T Ax$, where $A = A^T$
  - By definition
  - By directional derivative
    - $\mathrm{D}f(x)[d] = \nabla f(x)^T d$, $\forall d \in \mathbb{R}^n$;
    - $\mathrm{D}(\nabla f(x))[d] = \nabla^2 f(x)d$, $\forall d \in \mathbb{R}^n$;

Derivation formulas

If F is linear, then

$$\mathrm{D}F(x)[z] = F(z).$$

Chain rule: If $\mathrm{range}(F) \subseteq \mathrm{dom}(G)$, then

$$\mathrm{D}(G \circ F)(x)[z] = \mathrm{D}G(F(x))[\mathrm{D}F(x)[z]].$$

Product rule: Let $\bullet$ denote a bilinear operator on the ranges of F and G, then

$$\mathrm{D}(F \bullet G)(x)[z] = (\mathrm{D}F(x)[z]) \bullet G(x) + F(x) \bullet (\mathrm{D}G(x)[z]).$$

Gradient and Hessian

Example 13

Compute the gradient and the action of Hessian of the function $f(x) = \frac{1}{2}\|(xx^T - I_n)A\|_F^2$.

$$\min_{x \in \mathcal{E}} f(x)$$

- In previous examples, $\mathcal{E} = \mathbb{R}^n$;
- What if $\mathcal{E}$ is the set of real matrices $\mathbb{R}^{m \times n}$, the set of complex matrices $\mathbb{C}^{m \times n}$, or even real/complex tensors;

Gradient and Hessian

Matrix function

Function on matrices: $f : \mathbb{R}^{m\times n} \to \mathbb{R}$

- Gradient:

$$\nabla f(X) = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{m1}} \end{bmatrix}$$

- $\mathrm{D}f(X)[V] = \mathrm{trace}(\nabla f(X)^T V)$, $\forall V \in \mathbb{R}^{m\times n}$;

Example 14

Compute the gradient of

$$f : \mathbb{R}^{m\times n} \to \mathbb{R} : X \mapsto \log\det(I_m + XX^T).$$

(Hint: $\mathrm{D}(\log\det(Y))[V] = \mathrm{trace}(Y^{-1}V)$ for a symmetric positive definite matrix Y and symmetric matrix V.)

- Action of the Hessian

$$\nabla^2 f(X)[V] = \mathrm{D}(\nabla f(X))[V];$$

Example 15

Compute the action of the Hessian of

$$f : \mathbb{R}^{m\times n} \to \mathbb{R} : X \mapsto \log\det(I_m + XX^T).$$

- No matter whether the function is a vector function or a real matrix function, we have

$$\mathrm{D}f(x)[v] = \sum_{i_1, i_2, \ldots, i_s} (\nabla f(x))_{i_1, i_2, \ldots, i_s} v_{i_1, i_2, \ldots, i_s},$$

  where $s = 1$ for a vector and $s = 2$ for a matrix;
- Such idea can be extent to tensor or complex numbers or product of multiple spaces;

Example 16

Compute the gradient of

$$f : \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k} \to \mathbb{R} : (X, Y) \mapsto \mathrm{trace}(X^T A Y),$$

where $A \in \mathbb{R}^{m\times n}$;

References I