Optimization and Numerical Analysis: Nonlinear programming without constraints

Robert Gower

October 9, 2018

Table of Contents

The Problem: Nonlinear programming

Minimize a nonlinear differentiable function $f : x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}$

$$x^* = \arg\min_{x \in \mathbb{R}^n} f(x). \tag{1}$$

Warning: This problem is often impossible. First check there exists a minimum. Even linear programming does not always have a maximum!
Develop iterative methods $x^1, \ldots, x^k, \ldots,$ such that

$$\lim_{k \to \infty} x^k = x^*.$$

Template method

$$x^{k+1} = x^k + s_k d^k,$$

where $s_k > 0$ is a *step size* and $d^k \in \mathbb{R}^n$ is *search direction*. Satisfy the *descent condition*

$$f(x^{k+1}) < f(x^k).$$

Local and Global Minima

Definition of Local Minima

The point $x^* \in \mathbb{R}^n$ is a *local minima* of $f(x)$ if there exists $r > 0$ such that

$$f(x^*) \leq f(x), \quad \forall \|x - x^*\|_2 < r. \tag{2}$$

Definition of Global Minima

The point $x^* \in \mathbb{R}^n$ is a *global minima* of $f(x)$ if

$$f(x^*) \leq f(x), \quad \forall x. \tag{3}$$

In general finding global minima impossible. Finding local is good enough.

Multivariate Calculus

For a differentiable function $f : x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}$, we refer to $\nabla f(x)$ as the gradient evaluated at x defined by

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^\top .$$

Note that $\nabla f(x)$ is a column-vector. For any $F : x \in \mathbb{R}^n \to F(x) = [f_1(x), \ldots, f_n(x)]^\top \in \mathbb{R}^n$ define the *Jacobian matrix* by

$$\nabla F(x) \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial f_1(z)}{\partial x_1} & \frac{\partial f_2(z)}{\partial x_1} & \frac{\partial f_3(z)}{\partial x_1} & \cdots & \frac{\partial f_n(z)}{\partial x_1} \\ \frac{\partial f_2(z)}{\partial x_1} & \frac{\partial f_2(z)}{\partial x_2} & \frac{\partial f_3(z)}{\partial x_2} & \cdots & \frac{\partial f_n(z)}{\partial x_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(z)}{\partial x_n} & \frac{\partial f_2(z)}{\partial x_n} & \frac{\partial f_3(z)}{\partial x_n} & \cdots & \frac{\partial f_n(z)}{\partial x_n} \end{bmatrix}$$

$$= \left[\nabla f_1(x), \nabla f_2(x), \nabla f_3(x), \ldots, \nabla f_2(x)\right]$$

Multivariate Calculus

The gradient is useful because of 1st order Taylor expansion

$$f(x^0 + d) = f(x^0) + \nabla f(x^0)^\top d + \epsilon(d)\|d\|_2, \tag{4}$$

where $\epsilon(d)$ is a real valued such that

$$\lim_{d \to 0} \epsilon(d) = 0. \tag{5}$$

Example (The $\epsilon(d)$ function)

If $f(x) = \|x\|_2^2$ or $f(x) = x^\top Ax$ what is $\epsilon(d)$? Name three functions that satisfy (5).

The Hessian Matrix

If $f \in C^2$, we refer to $\nabla^2 f(x)$ as the Hessian matrix:

$$\nabla^2 f(x) \quad \stackrel{\text{def}}{=} \quad \begin{bmatrix} \frac{\partial^2 f(z)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(z)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(z)}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f(z)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(z)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(z)}{\partial x_2 \partial x_2} & \frac{\partial^2 f(z)}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 f(z)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(z)}{\partial x_n \partial x_1} & \frac{\partial^2 f(z)}{\partial x_n \partial x_2} & \frac{\partial^2 f(z)}{\partial x_n \partial x_3} & \cdots & \frac{\partial^2 f(z)}{\partial x_n \partial x_n} \end{bmatrix}$$

If $f \in C^2$ then

$$\frac{\partial^2 f(z)}{\partial x_i \partial x_j} = \frac{\partial^2 f(z)}{\partial x_j \partial x_i}, \ \forall i, j \in \{1, \ldots, n\}, \quad \Leftrightarrow \quad \nabla^2 f(x) = \nabla^2 f(x)^\top.$$

Hessian matrix useful for 2nd order Taylor expansion.

$$f(x^0 + d) = f(x^0) + \nabla f(x^0)^\top d + \frac{1}{2} d^\top \nabla^2 f(x^0) d + \epsilon(d) \|d\|_2^2. \qquad (6)$$

Example

If $f(x) = x^3$ or $f(x) = x^\top A x$ what is $\epsilon(d)$?

The Product-rule

The vector valued version of the product rule

- For any function $F(x) : \mathbb{R}^n \to \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{n\times n}$ we have
$$\nabla(F(x)^\top A) = \nabla F(x)^\top A. \tag{7}$$
- For any two vector valued functions F_1 and F_2 we have that
$$\nabla(F_1(x)^\top F_2(x)) = \nabla F_1(x) F_2(x) + \nabla F_2(x) F_1(x). \tag{8}$$

Example

Let $f(x) = \frac{1}{2}x^\top Ax$, where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. Calculate the gradient and the Hessian of $f(x)$.

The Product-rule

The vector valued version of the product rule

- For any function $F(x) : \mathbb{R}^n \to \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{n\times n}$ we have
$$\nabla(F(x)^\top A) = \nabla F(x)^\top A. \tag{7}$$
- For any two vector valued functions F_1 and F_2 we have that
$$\nabla(F_1(x)^\top F_2(x)) = \nabla F_1(x) F_2(x) + \nabla F_2(x) F_1(x). \tag{8}$$

Example

Let $f(x) = \frac{1}{2} x^\top A x$, where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. Calculate the gradient and the Hessian of $f(x)$.

Let $F_1(x) = A^\top x$ and $F_2(x) = x$ then
$\nabla f(x) = \frac{1}{2}\nabla(A^\top x)x + \frac{1}{2}\nabla(x)A^\top x = \frac{1}{2}(A + A^\top)x$ since
$\nabla(A^\top x) = A\nabla(x) = A$. Differentiating again
$\nabla(\nabla f(x)) = \nabla(Ax) = \nabla(A)x + \nabla(x)A = A.$