C6.2/B2. Continuous Optimization

Mathematical Background
(brief review)

Optimization draws on a number of key results in analysis and linear algebra. We briefly summarize some useful notions here. For more details, you may consult **Burden, R.L., & Faires, J.D.**, *Numerical Analysis*, 6th edition or later, Brooks/Cole Publishing.

Single valued functions and their derivatives

All the functions $f : \mathbb{R}^n \mapsto \mathbb{R}$ in this course are assumed to be smooth.

- The function $l : \mathbb{R}^n \mapsto \mathbb{R}$ is a **linear function** iff it is of the form

$$l(x) = d + g^T x \equiv d + \sum_{i=1}^{n} g_i x_i, \quad \text{where} \quad g = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},$$

  and $d \in \mathbb{R}$ and $g \in \mathbb{R}^n$ are known.

- The function $q(x) : \mathbb{R}^n \mapsto \mathbb{R}$ is a **quadratic function** iff it is of the form

$$q(x) = d + g^T x + \frac{1}{2} x^T H x = d + \sum_{i=1}^{n} g_i x_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} x_i x_j, \text{ where } H = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{pmatrix}.$$

  may be taken to be constant and symmetric. Although a quadratic function is strictly nonlinear, its properties are such that it is treated separately. Thus the term 'nonlinear function' often refers to a function which is not linear *or* quadratic.

- For the function $f : \mathbb{R}^n \mapsto \mathbb{R}$, the **vector of first partial derivatives** or **gradient vector** is

$$g(x) \equiv \nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} (x),$$

  where ∇ denotes the gradient operator $(\partial/\partial x_1 \;\; \partial/\partial x_2 \;\; \dots \;\; \partial/\partial x_n)^T$.

- For the function $f : \mathbb{R}^n \mapsto \mathbb{R}$, the **matrix of second partial derivatives** or **Hessian matrix**

$$H(x) \equiv \nabla [g(x)]^T = \nabla [\nabla f(x)]^T = \nabla \nabla^T f(x) = \nabla^2 f(x),$$

where

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} (x).$$

Note that $\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$, for all $i, j \in \{1, \ldots, n\}$, whenever $f \in \mathcal{C}^2(\mathbb{R}^n)$ (i.e., f is twice continuously differentiable, and so the Hessian exists and is continuous).

Properties of quadratic functions A quadratic function $q(x) = d + g^T x + \frac{1}{2} x^T H x$ has the following properties

- $\nabla q = g + Hx$.
- $\nabla^2 q = H$.

Vector valued functions and their derivatives

All the vector valued functions $r : \mathbb{R}^n \mapsto \mathbb{R}^m$ in this course are assumed to be smooth.
The Jacobian matrix of first partial derivatives of a function $r : \mathbb{R}^n \mapsto \mathbb{R}^m$ is

$$J(x) = r(x)\nabla^T = \begin{pmatrix} \frac{\partial r_1}{\partial x_1} & \frac{\partial r_1}{\partial x_2} & \cdots & \frac{\partial r_1}{\partial x_n} \\ \frac{\partial r_2}{\partial x_1} & \frac{\partial r_2}{\partial x_2} & \cdots & \frac{\partial r_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial r_m}{\partial x_1} & \frac{\partial r_m}{\partial x_2} & \cdots & \frac{\partial r_m}{\partial x_n} \end{pmatrix} (x).$$

Note that the Hessian matrix for a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ may be interpreted as being the Jacobian matrix of ∇f.

Taylor expansions

Numerical methods for solving nonlinear equation and optimization problems are frequently based on Taylor expansions. The following expansions are particularly important.

The first-order Taylor expansion of $f : \mathbb{R}^n \mapsto \mathbb{R}$ around $x \in \mathbb{R}^n$ is

$$f(x + h) = f(x) + \nabla f(x)^T h + \|h\|_2 z(h),$$

where $z(h) \to 0$ as $h \to 0$. This yields the following linear approximation to f which interpolates its value and gradient at x,

$$l(h) = f(x) + \nabla f(x)^T h.$$

We also have the alternative expression for the first-order Taylor expansion

$$f(x + h) = f(x) + \nabla f(\xi)^T h,$$

where $\xi \in \mathbb{R}^n$ is a point on the line segment determined by x and $x + h$.

The second-order Taylor expansion of $f : \mathbb{R}^n \mapsto \mathbb{R}$ around $x \in \mathbb{R}^n$ is

$$f(x + h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T [\nabla^2 f(x)] h + \|h\|_2^2 z(h),$$

where $z(h) \to 0$ as $h \to 0$. This yields the following quadratic approximation to f which interpolates its value, gradient and Hessian at x, namely,

$$q(h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T [\nabla^2 f(x)] h.$$

Alternatively, the second-order Taylor expansion of f around x can be expressed as

$$f(x+h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T [\nabla^2 f(\xi)] h,$$

where $\xi \in \mathbb{R}^n$ is a point on the line segment determined by x and $x+h$.

The first order Taylor expansion of $\nabla f : \mathbb{R}^n \mapsto \mathbb{R}^n$ around $x \in \mathbb{R}^n$ is

$$\nabla f(x+h) = \nabla f(x) + \nabla^2 f(x) h + \|h\|_2 z(h),$$

where $z(h) \to 0$ as $h \to 0$. This yields the following linear approximation to ∇f which interpolates its value and Jacobian at x, namely,

$$l(h) = \nabla f(x) + \nabla^2 f(x) h.$$

Note that now we only have the following integral alternative expression for the Taylor expansion (as the function ∇f is vector-valued),

$$\nabla f(x+h) = \nabla f(x) + \int_0^1 \nabla^2 f(x+th) h dt.$$

The first order Taylor expansion of $r : \mathbb{R}^n \mapsto \mathbb{R}^m$ about $x \in \mathbb{R}^n$ is

$$r(x+h) = r(x) + J(x) h + \|h\|_2 z(h),$$

where $z(h) \to 0$ as $h \to 0$. This yields the following linear approximation to r which interpolates its value and Jacobian at x, namely,

$$l(h) = r(x) + J(x) h.$$

Note that now we only have the following integral alternative expression for the Taylor expansion (as the function r is vector-valued),

$$r(x+h) = r(x) + \int_0^1 J(x+th) h dt.$$

Linear algebra

- **Linear independence and bases.**

  The set of vectors $\{x_i\}_{i=1}^m \subset \mathbb{R}^n$ is **linearly independent** iff $\sum_{i=1}^m \alpha_i x_i = 0 \Rightarrow \alpha_i = 0$, $i = 1, \ldots, m$.

  A set of n linearly independent vectors $\{x_i\}_{i=1}^n$ in $\mathbb{R}^n$ forms a **basis** for $\mathbb{R}^n$ and any vector $x \in \mathbb{R}^n$ can be expressed as $x = \sum_{i=1}^n \alpha_i x_i$.

- **Matrix definiteness.**

  The matrix A is **positive (negative) definite** $\iff$ $x^T A x > 0$ ($x^T A x < 0$) $\forall\, x \in \mathbb{R}^n$, $x \neq 0$.

  The matrix A is **positive (negative) semi-definite** $\iff$ $x^T A x \geq 0$ ($x^T A x \leq 0$) $\forall\, x \in \mathbb{R}^n$.

  A matrix which is not positive/negative definite or positive/negative semi-definite is **indefinite**.

- **Eigenvalues and eigenvectors.**

  If the matrix H is symmetric then there exists an orthogonal matrix Q and diagonal matrix Λ such that $H = Q \Lambda Q^T$.

  - The entries $\lambda_1, \ldots, \lambda_n$ of Λ are the **eigenvalues** of H.

– The columns (vectors) $q_1, \ldots, q_n$ of Q are the **eigenvectors** of H.

Any vector $x \in \mathbb{R}^n$ can be expressed as $x = \sum_{i=1}^{n} \alpha_i q_i$, where $\alpha_i = x^T q_i$. Also $H = \sum_{i=1}^{n} \lambda_i q_i q_i^T$.

If λ is an eigenvalue of a nonsingular matrix H then $1/\lambda$ is an eigenvalue of H^{-1} so $H^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} q_i q_i^T$.

Vector norms

The Euclidean (also called l_2) measure of the magnitude of the vector $x = (x_1 \, \ldots \, x_n)^T \in \mathbb{R}^n$ is the value

$$\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}.$$

This is an example of a **vector norm**.
A **norm** on the space of vectors $\mathbb{R}^n$ is a function, $\|\cdot\| : \mathbb{R}^n \mapsto \mathbb{R}$, such that for all vectors x, $y \in \mathbb{R}^n$ and scalars $\alpha \in \mathbb{R}$,

i) $\|x\| \geq 0$;

ii) $\|x\| = 0 \iff x = 0$;

iii) $\|\alpha x\| = |\alpha| \cdot \|x\|$;

iv) $\|x + y\| \leq \|x\| + \|y\|$.

The most commonly-used vector norms are referred to as the l_p-**norms** (or simply as the **p-norms**), namely,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}},$$

and so, in particular,

$$\begin{aligned} \|x\|_1 &= |x_1| + |x_2| + \cdots + |x_n| \\ \|x\|_2 &= \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \equiv \sqrt{x^T x} \\ \|x\|_\infty &= \max\{|x_1|, |x_2|, \ldots, |x_n|\}. \end{aligned}$$

Matrix norms

When $y = Ax$, the magnitude of y clearly depends on the magnitudes of A and x. In order to estimate this, without computing y explicitly, it is necessary to have a measure of the magnitude of A. This is achieved by using a **matrix norm**.
A **norm** on the space of square matrices $\mathbb{R}^{n \times n}$ is a function, $\|\cdot\| : \mathbb{R}^{n \times n} \mapsto \mathbb{R}$, such that for all matrices $A, B \in \mathbb{R}^{n \times n}$ and scalars $\alpha \in \mathbb{R}$,

i) $\|A\| \geq 0$;

ii) $\|A\| = 0 \iff A = 0$;

iii) $\|\alpha A\| = |\alpha| \|A\|$;

iv) $\|A + B\| \leq \|A\| + \|B\|$;

v) $\|AB\| \leq \|A\| \|B\|$.

The most commonly-used matrix norms are **p-norms**. These are given by the corresponding vector p-norms according to the definition

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \quad \text{or equivalently,} \quad \|A\|_p = \max_{\|x\|_p = 1} \|Ax\|_p,$$

and so, in particular,

$$\begin{aligned} \|A\|_1 &= \max_{\|x\|_1=1} \|Ax\|_1 &\equiv& \max_j \{\textstyle\sum_{i=1}^n |a_{ij}|\} \\ \|A\|_2 &= \max_{\|x\|_2=1} \|Ax\|_2 &\equiv& \sqrt{\max_i \lambda_i(A^TA)} \\ \|A\|_\infty &= \max_{\|x\|_\infty=1} \|Ax\|_\infty &\equiv& \max_i \left\{\textstyle\sum_{j=1}^n |a_{ij}|\right\}. \end{aligned}$$

where $\lambda_i(A^TA)$, for $i = 1, \ldots, n$, are the eigenvalues of A^TA. Note that although the matrix 2-norm has useful *theoretical* properties it may be too difficult to compute in *practice*.

Two particularly important properties of the matrix p-norms (which follow directly from their definition) are that for all vectors x,

$$\|Ax\|_p \leq \|A\|_p \|x\|_p$$

and, given any $A \in \mathbb{R}^{n\times n}$, there exists $x \neq 0$ such that

$$\|Ax\|_p = \|A\|_p \|x\|_p.$$

When referring to (p-) norms in general, it is convenient to drop the subscript.

The sequence of matrices $\{A^{(k)}\}_{n=1}^{\infty}$ in $\mathbb{R}^{n\times n}$ is said to **converge** to A with respect to the norm $\|\cdot\|$ if, given any $\epsilon > 0$, there exists an integer $K(\epsilon)$ such that

$$\|A^{(k)} - A\| < \epsilon \quad \text{for all} \quad k \geq K(\epsilon).$$

If the matrix A satisfies $\|A\| < 1$ for some norm $\|\cdot\|$, then $A^k \to 0$ as $k \to \infty$.