Peeter Joot
peeterjoot@protonmail.com

ECE1505H Convex Optimization. Lecture 2: Mathematical background. Taught by Prof. Stark Draper

Disclaimer Peeter's lecture notes from class. These may be incoherent and rough.

These are notes for the UofT course ECE1505H, Convex Optimization, taught by Prof. Stark Draper, covering ch. 1 [1] content.

Topics

- Calculus: Derivatives and Jacobians, Gradients, Hessians, approximation functions.
- Linear algebra, Matrices, decompositions, ...

1.1 Norms

Definition 1.1: Vector space

A set of elements (vectors) that is closed under vector addition and scaling.

This generalizes the directed arrow concept of vector space (fig. 1.1) that is familiar from geometry.

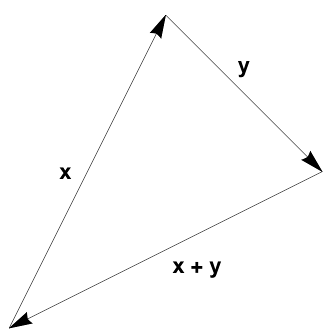

Figure 1.1: Vector addition.

Definition 1.2: Normed vector spaces

A vector space with a notion of lenght of any signle vector, the "norm".

Definition 1.3: Inner product space.

A normed vector space with a notion of a real angle vetween any pair of vectors.

This course has a focus on optimization in $\mathbb{R}^n$. Complex spaces in the context of this course can be considered with a mapping $\mathbb{C}^n \to \mathbb{R}^{2n}$.

Definition 1.4: Norm.

A norm is a function operating on a vector

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

that provides a mapping

$$\|\cdot\| : \mathbb{R}^n \to \mathbb{R},$$

where

- $\|\mathbf{x}\| \geq 0$
- $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$
- $\|t\mathbf{x}\| = |t|\,\|\mathbf{x}\|$
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. This is the triangle inequality.

Example: Euclidean norm

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{1.1}$$

Example: l_p-norms

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}. \tag{1.2}$$

For $p = 1$, this is

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|, \tag{1.3}$$

For $p = 2$, this is the Euclidean norm eq. (1.1). For $p = \infty$, this is

$$\|\mathbf{x}\|_\infty = \max_{i=1}^{n} |x_i|. \tag{1.4}$$

Definition 1.5: Unit ball

$$\{\mathbf{x} | \|\mathbf{x}\| \leq 1\}$$

The regions of the unit ball under the $l_1, l_2, and l_\infty$ norms are plotted in fig. 1.2.

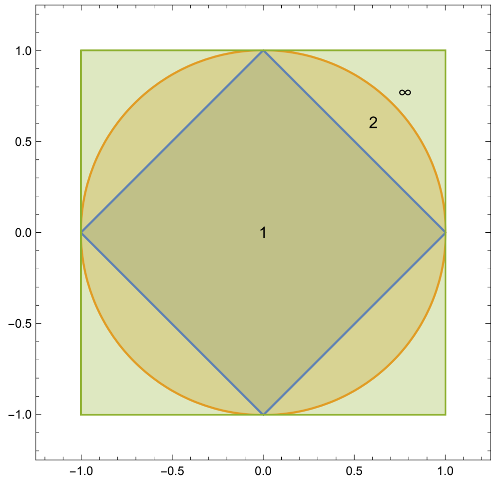

Figure 1.2: Some unit ball regions.

The l_2 norm is not only familiar, but can be "induced" by an inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathrm{T}} \mathbf{y} = \sum_{i=1}^{n} x_i y_i, \tag{1.5}$$

which is not true for all norms. The norm induced by this inner product is

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{y} \rangle} \tag{1.6}$$

Inner product spaces have a notion of angle (fig. 1.3) given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \|\mathbf{y}\| \cos\theta, \tag{1.7}$$

and always satify the Cauchy-Schwartz inequality

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 . \tag{1.8}$$

In an inner product space we say $\mathbf{x}$ and $\mathbf{y}$ are orthogonal vectors $\mathbf{x} \perp \mathbf{y}$ if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, as sketched in fig. 1.4.

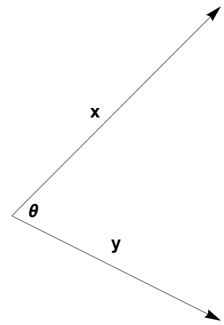

Figure 1.3: Inner product induced angle.

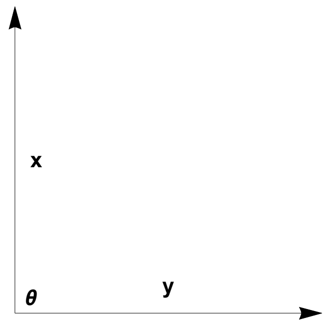

Figure 1.4: Orthogonality.

1.2 Dual norm

Definition 1.6: Dual norm

Let $\|\cdot\|$ be a norm in $\mathbb{R}^n$. The "dual" norm $\|\cdot\|_*$ is defined as

$$\|\mathbf{z}\|_* = \sup_{\mathbf{x}} \left\{ \mathbf{z}^T\mathbf{x} | \|\mathbf{x}\| \leq 1 \right\}.$$

where sup is roughly the "least upper bound".
This is a limit over the unit ball of $\|\cdot\|$.

l_2 dual .
Dual of the l_2 is the l_2 norm.
Proof:

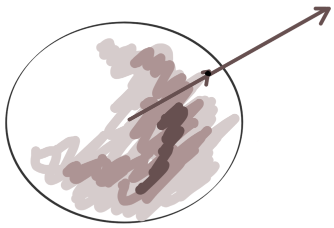

Figure 1.5: l_2 dual norm determination.

$$\begin{aligned}
\|\mathbf{z}\|_* &= \sup_{\mathbf{x}} \left\{ \mathbf{z}^{\mathrm{T}}\mathbf{x} \,|\, \|\mathbf{x}\|_2 \leq 1 \right\} \\
&= \sup_{\mathbf{x}} \{ \|\mathbf{z}\|_2 \|\mathbf{x}\|_2 \cos\theta \,|\, \|\mathbf{x}\|_2 \leq 1 \} \\
&\leq \sup_{\mathbf{x}} \{ \|\mathbf{z}\|_2 \|\mathbf{x}\|_2 \,|\, \|\mathbf{x}\|_2 \leq 1 \} \\
&\leq \cancel{\|\mathbf{z}\|_2} \left\| \frac{\mathbf{z}}{\cancel{\|\mathbf{z}\|_2}} \right\|_2 \\
&= \|\mathbf{z}\|_2 .
\end{aligned} \tag{1.9}$$

l_1 dual

. For l_1, the dual is the l_∞ norm. Proof:

$$\|\mathbf{z}\|_* = \sup_{\mathbf{x}} \left\{ \mathbf{z}^{\mathrm{T}}\mathbf{x} \,|\, \|\mathbf{x}\|_1 \leq 1 \right\}, \tag{1.10}$$

but

$$\begin{aligned}
\mathbf{z}^{\mathrm{T}}\mathbf{x} &= \sum_{i=1}^{n} z_i x_i \\
&\leq \left| \sum_{i=1}^{n} z_i x_i \right| \\
&\leq \sum_{i=1}^{n} |z_i x_i|,
\end{aligned} \tag{1.11}$$

so

$$\begin{aligned}
\|\mathbf{z}\|_* &= \sum_{i=1}^{n} |z_i||x_i| \\
&\leq \left(\max_{j=1}^{n} |z_j| \right) \sum_{i=1}^{n} |x_i| \\
&\leq \left(\max_{j=1}^{n} |z_j| \right) \\
&= \|\mathbf{z}\|_\infty .
\end{aligned} \tag{1.12}$$

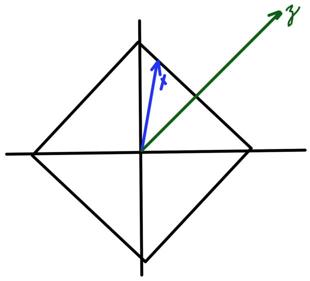

Figure 1.6: l_1 dual norm determination.

l_∞ dual .

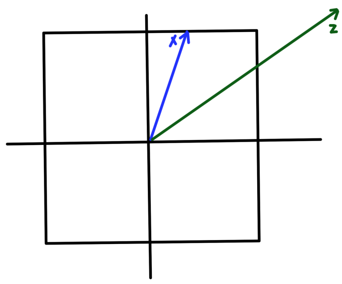

Figure 1.7: l_∞ dual norm determination.

$$\|\mathbf{z}\|_* = \sup_{\mathbf{x}} \left\{ \mathbf{z}^T \mathbf{x} \middle| \|\mathbf{x}\|_\infty \leq 1 \right\}. \tag{1.13}$$

Here

$$\begin{aligned} \mathbf{z}^T \mathbf{x} &= \sum_{i=1}^{n} z_i x_i \\ &\leq \sum_{i=1}^{n} |z_i||x_i| \\ &\leq \left(\max_j |x_j| \right) \sum_{i=1}^{n} |z_i| \\ &= \|\mathbf{x}\|_\infty \sum_{i=1}^{n} |z_i|. \end{aligned} \tag{1.14}$$

So

$$\|\mathbf{z}\|_* \leq \sum_{i=1}^{n} |z_i| = \|\mathbf{z}\|_1 . \tag{1.15}$$

Statement from the lecture: I'm not sure where this fits:

$$x_i^* = \begin{cases} +1 & z_i \geq 0 \\ -1 & z_i \leq 0 \end{cases} \tag{1.16}$$

1.3 Multivariable Taylor approximation

The Taylor series expansion for a scalar function $g : \mathbb{R} \rightarrow \mathbb{R}$ about the origin is just

$$g(t) = g(0) + t g'(0) + \frac{t^2}{2} g''(0) + \cdots \tag{1.17}$$

In particular

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(0) + \cdots \tag{1.18}$$

Now consider $g(t) = f(\mathbf{x} + \mathbf{a}t)$, where $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $g(0) = f(\mathbf{x})$, and $g(1) = f(\mathbf{x} + \mathbf{a})$. This trick, from [2] allows for a direct expansion of the multivariable Taylor series of a scalar function

$$f(\mathbf{x} + \mathbf{a}) = f(\mathbf{x}) + \left.\frac{df(\mathbf{x} + \mathbf{a}t)}{dt}\right|_{t=0} + \frac{1}{2} \left.\frac{d^2 f(\mathbf{x} + \mathbf{a}t)}{dt^2}\right|_{t=0} + \cdots \tag{1.19}$$

The first order term is

$$\begin{aligned} \left.\frac{df(\mathbf{x} + \mathbf{a}t)}{dt}\right|_{t=0} &= \sum_{i=1}^n \frac{d(x_i + a_i t)}{dt} \left.\frac{\partial f(\mathbf{x} + \mathbf{a}t)}{\partial (x_i + a_i t)}\right|_{t=0} \\ &= \sum_{i=1}^n a_i \frac{\partial f(\mathbf{x})}{\partial x_i} \\ &= \mathbf{a} \cdot \boldsymbol{\nabla} f. \end{aligned} \tag{1.20}$$

Similarily, for the second order term

$$\begin{aligned} \left.\frac{d^2 f(\mathbf{x} + \mathbf{a}t)}{dt^2}\right|_{t=0} &= \left.\left(\frac{d}{dt} \left(\sum_{i=1}^n a_i \frac{\partial f(\mathbf{x} + \mathbf{a}t)}{\partial (x_i + a_i t)} \right) \right)\right|_{t=0} \\ &= \left.\left(\sum_{j=1}^n \frac{d(x_j + a_j t)}{dt} \sum_{i=1}^n a_i \frac{\partial^2 f(\mathbf{x} + \mathbf{a}t)}{\partial (x_j + a_j t) \partial (x_i + a_i t)} \right)\right|_{t=0} \\ &= \sum_{i,j=1}^n a_i a_j \frac{\partial^2 f}{\partial x_i \partial x_j} \\ &= (\mathbf{a} \cdot \boldsymbol{\nabla})^2 f. \end{aligned} \tag{1.21}$$

The complete Taylor expansion of a scalar function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is therefore

$$f(\mathbf{x} + \mathbf{a}) = f(\mathbf{x}) + \mathbf{a} \cdot \boldsymbol{\nabla} f + \frac{1}{2} (\mathbf{a} \cdot \boldsymbol{\nabla})^2 f + \cdots, \tag{1.22}$$

so the Taylor expansion has an exponential structure

$$f(\mathbf{x}+\mathbf{a}) = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{a}\cdot\boldsymbol{\nabla})^k f = e^{\mathbf{a}\cdot\boldsymbol{\nabla}} f. \tag{1.23}$$

Should an approximation of a vector valued function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be desired it is only required to form a matrix of the components

$$\mathbf{f}(\mathbf{x}+\mathbf{a}) = \mathbf{f}(\mathbf{x}) + [\mathbf{a}\cdot\boldsymbol{\nabla} f_i]_i + \frac{1}{2}[(\mathbf{a}\cdot\boldsymbol{\nabla})^2 f_i]_i + \cdots, \tag{1.24}$$

where $[.]_i$ denotes a column vector over the rows $i \in [1, m]$, and f_i are the coordinates of $\mathbf{f}$.

1.4 The Jacobian matrix

In [1] the Jacobian $D\mathbf{f}$ of a function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is defined in terms of the limit of the l_2 norm ratio

$$\frac{\|\mathbf{f}(\mathbf{z}) - \mathbf{f}(\mathbf{x}) - (D\mathbf{f})(\mathbf{z}-\mathbf{x})\|_2}{\|\mathbf{z}-\mathbf{x}\|_2}, \tag{1.25}$$

with the statement that the function $\mathbf{f}$ has a derivative if this limit exists. Here the Jacobian $D\mathbf{f} \in \mathbb{R}^{m\times n}$ must be matrix valued.

Let $\mathbf{z} = \mathbf{x} + \mathbf{a}$, so the first order expansion of eq. (1.24) is

$$\mathbf{f}(\mathbf{z}) = \mathbf{f}(\mathbf{x}) + [(\mathbf{z}-\mathbf{x})\cdot\boldsymbol{\nabla} f_i]_i. \tag{1.26}$$

With the (unproven) assumption that this Taylor expansion satisfies the norm limit criteria of eq. (1.25), it is possible to extract the structure of the Jacobian by comparison

$$\begin{aligned}(D\mathbf{f})(\mathbf{z}-\mathbf{x}) &= \left[(\mathbf{z}-\mathbf{x})\cdot\boldsymbol{\nabla} f_i\right]_i \\ &= \left[\textstyle\sum_{j=1}^n (z_j - x_j)\frac{\partial f_i}{\partial x_j}\right]_i \\ &= \left[\frac{\partial f_i}{\partial x_j}\right]_{ij}(\mathbf{z}-\mathbf{x}),\end{aligned} \tag{1.27}$$

so

$$\boxed{(D\mathbf{f})_{ij} = \frac{\partial f_i}{\partial x_j}} \tag{1.28}$$

Written out explictly as a matrix the Jacobian is

$$D\mathbf{f} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{\nabla} f_1)^{\mathrm{T}} \\ (\boldsymbol{\nabla} f_2)^{\mathrm{T}} \\ \vdots \\ (\boldsymbol{\nabla} f_m)^{\mathrm{T}} \end{bmatrix}. \tag{1.29}$$

In particular, when the function is scalar valued

$$Df = (\nabla f)^{\mathrm{T}}. \tag{1.30}$$

With this notation, the first Taylor expansion, in terms of the Jacobian matrix is

$$\boxed{\mathbf{f}(\mathbf{z}) \approx \mathbf{f}(\mathbf{x}) + (D\mathbf{f})\,(\mathbf{z} - \mathbf{x})\,.} \tag{1.31}$$

Gradient The gradient provides a linear approximation of a function about a point $\mathbf{x}_0 \in \mathbb{R}^n$.

$$\begin{aligned} F(\mathbf{x}) &\approx F(\mathbf{x}_0) + \nabla F(\mathbf{x}_0)^{\mathrm{T}}\,(\mathbf{x} - \mathbf{x}_0)\,. \\ &= F(\mathbf{x}_0) + \left\langle \nabla F(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \right\rangle, \end{aligned} \tag{1.32}$$

or

$$F(\mathbf{x} + \Delta\mathbf{x}) = F(\mathbf{x}) + \left\langle \nabla F(\mathbf{x}), \Delta\mathbf{x} \right\rangle. \tag{1.33}$$

This can be thought of as the definition of the gradient in an inner product space. It will be possible to find the structure of the gradient by considering a pertubation of a function about a point.

When g is a scalar function, the chain rule can be expressed in terms of the gradient

$$\nabla(g(F(\mathbf{x}))) = (DF)^{\mathrm{T}}\Big|_{\mathbf{x}} \nabla g|_{F(\mathbf{x})}. \tag{1.34}$$

Example 1:

$$\begin{aligned} F &: \mathbb{R}^n \to \mathbb{R} \\ g &: \mathbb{R} \to \mathbb{R}, \end{aligned} \tag{1.35}$$

and let

$$h(\mathbf{x}) = g(F(\mathbf{x})), \tag{1.36}$$

for $\mathbf{x} \in \mathbb{R}^n$, then

$$\nabla h(\mathbf{x}) = g'(F(\mathbf{x}))\nabla F(\mathbf{x}). \tag{1.37}$$

Example 1.1: Quadratic form

$$F(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} P \mathbf{x} = \sum_{i,j=1}^{n} x_i x_j P_{ij}, \tag{1.38}$$

We want to show that

$$\nabla F(\mathbf{x}) = \left(P + P^{\mathrm{T}}\right)\mathbf{x}. \tag{1.39}$$

Consider the k-th derivative

$$
\begin{aligned}
\frac{\partial}{\partial x_k} F(\mathbf{x}) &= \frac{\partial}{\partial x_k} \left(P_{kk} x_k^2 + \sum_{i \neq k} x_i x_k \left(P_{ik} + P_{ki} \right) \right) \\
&= 2 P_{kk} x_k + 2 \sum_{i \neq k} x_i \frac{\left(P_{ik} + P_{ki} \right)}{2} \\
&= \sum_i^n x_i \frac{\left(P_{ik} + P_{ki} \right)}{2} \\
&= \sum_i^n \left(P_{ik} + P_{ki} \right) x_i,
\end{aligned} \tag{1.40}
$$

which proves eq. (1.39).

Symmetric matrices Let S^n be the set of symmetric matrices

$$
S^n = \left\{ P \in \mathbb{R}^{n \times n} | P = P^{\mathrm{T}} \right\}, \tag{1.41}
$$

then

$$
\nabla (\mathbf{x}^{\mathrm{T}} P \mathbf{x}) = 2 P \mathbf{x}. \tag{1.42}
$$

1.5 Chain rule

The gradients or Jacobians for compositions of functions can also be calculated

Theorem 1.1: Chain rule

Given functions

$$
\begin{aligned}
F &: \mathbb{R}^n \to \mathbb{R}^m \\
g &: \mathbb{R}^m \to \mathbb{R}^p,
\end{aligned} \tag{1.43}
$$

$$
D(g(F(\mathbf{x}))) = Dg|_{F(\mathbf{x})} DF|_{\mathbf{x}}. \tag{1.44}
$$

Scalar valued composition To illustrate this, first consider a scalar valued composition

$$
\begin{aligned}
F &: \mathbb{R}^n \to \mathbb{R}^n \\
g &: \mathbb{R}^n \to \mathbb{R},
\end{aligned} \tag{1.45}
$$

and let

$$\begin{aligned} h(\mathbf{x}) &= g(F(\mathbf{x})) \\ &= g\left(\begin{bmatrix} F_1(\mathbf{x}) \\ F_2(\mathbf{x}) \\ \vdots \\ F_n(\mathbf{x}) \end{bmatrix}\right) \end{aligned} \tag{1.46}$$

for $\mathbf{x} \in \mathbb{R}^n$, then

$$\frac{\partial h(\mathbf{x})}{\partial x_k} = \frac{\partial g}{\partial F_1}\frac{\partial F_1}{\partial x_k} + \frac{\partial g}{\partial F_2}\frac{\partial F_2}{\partial x_k} + \cdots \tag{1.47}$$

With

$$DF(\mathbf{x}) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_n} \\ \vdots & & \vdots & \\ \frac{\partial F_n}{\partial x_1} & \frac{\partial F_n}{\partial x_2} & \cdots & \frac{\partial F_n}{\partial x_n} \end{bmatrix} \tag{1.48}$$

the gradient $\nabla g = (Dg)^{\mathrm{T}}$ is

$$\nabla h(\mathbf{x}) = (DF)^{\mathrm{T}}\Big|_{\mathbf{x}} \nabla g(F(\mathbf{x})), \tag{1.49}$$

or

$$D(g(F(\mathbf{x}))) = Dg|_{F(\mathbf{x})} DF|_{\mathbf{x}}. \tag{1.50}$$

Affine functions An important example are affine functions of $\mathbf{x}$

$$\begin{aligned} F &: \mathbb{R}^n \to \mathbb{R}^n \\ g &: \mathbb{R}^n \to \mathbb{R}, \end{aligned} \tag{1.51}$$

$$F(\mathbf{x}) = A\mathbf{x} + \mathbf{b}, \tag{1.52}$$

where A is an $n \times n$ matrix and $\mathbf{b}$ is an $n \times 1$ column vector.
Given a function

$$h(\mathbf{x}) = g(F(\mathbf{x})) = g(A\mathbf{x} + \mathbf{b}). \tag{1.53}$$

$$\begin{aligned} F(\mathbf{x}) &= A\mathbf{x} + \mathbf{b} \\ &= \begin{bmatrix} \sum_{i=1}^n a_{1i}x_i \\ \sum_{i=1}^n a_{2i}x_i \\ \vdots \\ \sum_{i=1}^n a_{ni}x_i \vdots \end{bmatrix} + \mathbf{b}, \end{aligned} \tag{1.54}$$

so

$$DF(\mathbf{x}) = A, \tag{1.55}$$

and

$$\nabla(g(F(\mathbf{x}))) = (A\mathbf{x})^{\mathrm{T}} \nabla g|_{F(\mathbf{x})}. \tag{1.56}$$

General case The proof of the general case can be essentially be performed by example, provided that example is sufficiently non-trivial, such as a non-square case such as $n = 4, m = 3, p = 2$

$$F(\mathbf{x}) = \begin{bmatrix} F_1(\mathbf{x}) \\ F_2(\mathbf{x}) \\ F_3(\mathbf{x}) \end{bmatrix}, \tag{1.57}$$

and

$$g(\mathbf{y}) = \begin{bmatrix} g_1(\mathbf{y}) \\ g_2(\mathbf{y}) \end{bmatrix}. \tag{1.58}$$

For such a function

$$\frac{\partial g(F(\mathbf{x}))}{\partial x_1} = \begin{bmatrix} \partial g_1(F(\mathbf{x}))/\partial x_1 \\ \partial g_2(F(\mathbf{x}))/\partial x_1 \end{bmatrix}, \tag{1.59}$$

so

$$\begin{aligned} Dg(F(\mathbf{x})) &= \begin{bmatrix} \partial g_1(F(\mathbf{x}))/\partial x_1 & \partial g_1(F(\mathbf{x}))/\partial x_2 & \cdots & \partial g_1(F(\mathbf{x}))/\partial x_4 \\ \partial g_2(F(\mathbf{x}))/\partial x_1 & \partial g_2(F(\mathbf{x}))/\partial x_2 & \cdots & \partial g_2(F(\mathbf{x}))/\partial x_4 \end{bmatrix} \\ &= \begin{bmatrix} D(g_1(F(\mathbf{x}))) \\ D(g_2(F(\mathbf{x}))) \end{bmatrix}. \end{aligned} \tag{1.60}$$

This reduces the problem to the composition of a scalar and vector function, such as

$$\begin{aligned} D(g_1(F(\mathbf{x}))) &= \sum_{i=1}^{3} \sum_{j=1}^{4} \left.\frac{\partial g_1}{\partial y_i}\right|_{y_i = F_i(\mathbf{x})} \frac{\partial F_i}{\partial x_j} \\ &= \left(\begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \frac{\partial g_1}{\partial y_3} \end{bmatrix} \right) \Big|_{\mathbf{y}=F(\mathbf{x})} \begin{bmatrix} \frac{\partial F_i}{\partial x_j} \end{bmatrix}_{ij} \\ &= Dg_1|_{F(\mathbf{x})} DF(\mathbf{x}). \end{aligned} \tag{1.61}$$

The total Jacobian is

$$Dg(F(\mathbf{x})) = \begin{bmatrix} Dg_1|_{F(\mathbf{x})} DF(\mathbf{x}) \\ Dg_2|_{F(\mathbf{x})} DF(\mathbf{x}) \end{bmatrix}, \tag{1.62}$$

which can be factored as

$$D(g(F(\mathbf{x}))) = Dg|_{F(\mathbf{x})} DF(\mathbf{x}).. \tag{1.63}$$

1.6 The Hessian matrix

For scalar valued functions, the text expresses the second order expansion of a function in terms of the Jacobian and Hessian matrices

$$f(\mathbf{z}) \approx f(\mathbf{x}) + (Df)(\mathbf{z} - \mathbf{x}) + \frac{1}{2}(\mathbf{z} - \mathbf{x})^{\mathrm{T}} (\boldsymbol{\nabla}^2 f)(\mathbf{z} - \mathbf{x}). \tag{1.64}$$

Because $\boldsymbol{\nabla}^2$ is the usual notation for a Laplacian operator, this $\boldsymbol{\nabla}^2 f \in \mathbb{R}^{n \times n}$ notation for the Hessian matrix is not ideal in my opinion. Ignoring that notational objection for this class, the structure of the Hessian matrix can be extracted by comparison with the coordinate expansion

$$\mathbf{a}^\mathrm{T}(\boldsymbol{\nabla}^2 f)\mathbf{a} = \sum_{r,s=1}^{n} a_r a_s \frac{\partial^2 f}{\partial x_r \partial x_s} \tag{1.65}$$

so

$$\boxed{(\boldsymbol{\nabla}^2 f)_{ij} = \frac{\partial^2 f_i}{\partial x_i \partial x_j}.} \tag{1.66}$$

In explicit matrix form the Hessian is

$$\boldsymbol{\nabla}^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}. \tag{1.67}$$

Is there a similar nice matrix structure for the Hessian of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$?

Example 1.2: Second order scalar function

Given

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\mathrm{T} P \mathbf{x} + \mathbf{q}^\mathrm{T}\mathbf{x} + \mathbf{c}, \tag{1.68}$$

where P is a symmetric matrix $P = P^\mathrm{T}$, then

$$\begin{aligned} \boldsymbol{\nabla} F &= \frac{1}{2}\left(P + P^\mathrm{T}\right)\mathbf{x} + \mathbf{q} \\ &= P\mathbf{x} + \mathbf{q}, \end{aligned} \tag{1.69}$$

and

$$\boldsymbol{\nabla}^2 F = P. \tag{1.70}$$

Bibliography

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 1, 1.4

[2] D. Hestenes. *New Foundations for Classical Mechanics*. Kluwer Academic Publishers, 1999. 1.3