CHAPTER 5

Elements of Multivariable Calculus

1. Norms and Continuity

As we have seen the 2-norm gives us a measure of the magnitude of a vector v in $\mathbb{R}^n$, $\|v\|_2$. As such it also gives us a measure of the distance between to vectors $u,\ v \in \mathbb{R}^n$, $\|u - v\|_2$. Such measures of magnitude and distance are very useful tools for measuring model misfit as is the case in linear least squares problem. They are also essential for analyzing the behavior of sequences and functions on $\mathbb{R}^n$ as well as on the space of matrices $\mathbb{R}^{m\times n}$. For this reason, we formalize the notion of a norm to incorporate other measures of magnitude and distance.

DEFINITION 1.1. *[Vector Norm] A function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is a vector norm on $\mathbb{R}^n$ if*

(1) $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$ with equality if and only if $x = 0$,
(2) $\|\alpha x\| = |\alpha| \, \|x\|$ for all $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, and
(3) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$.

EXAMPLE 1.1. *Perhaps the most common examples of norms are the p-norms for $1 \leq p \leq \infty$. Given $1 \leq p < \infty$, the ℓ_p-norm on $\mathbb{R}^n$ is defined as*

$$\|x\|_p := \left[\sum_{j=1}^{n} |x_j|^p \right]^{1/p} .$$

For $p = \infty$, we define

$$\|x\|_\infty := \max \{ |x_i| \, | \, i = 1, 2, \ldots, n \} .$$

This choice of notation for the ∞-norm comes from the relation

$$\lim_{p \uparrow \infty} \|x\|_p = \|x\|_\infty \quad \forall \ x \in \mathbb{R}^n.$$

In applications, the most important of these norms are the $p = 1, 2, \infty$ norms as well as variations on these norms.

In finite dimensions all norms are said the *equivalent* in the sense that one can show that for any two norms $\|\cdot\|_{(a)}$ and $\|\cdot\|_{(b)}$ on $\mathbb{R}^n$ there exist positive constants α and β such that

$$\alpha \|x\|_a \leq \|x\|_b \leq \beta \|x\|_a \quad \forall \, x \in \mathbb{R}^n \ .$$

But we caution that in practice the numerical behavior of these norms differ greatly when the dimension is large.

Since norms can be used to measure the distance between vectors, they can be used to form a notions of continuity for functions mapping $\mathbb{R}^n$ to $\mathbb{R}^m$ that parallel those established for mappings from $\mathbb{R}$ to $\mathbb{R}$.

DEFINITION 1.2. *[Continuous Functions] Let $F : \mathbb{R}^n \to \mathbb{R}^n$.*

(1) F is said to be continuous at a point $\overline{x} \in \mathbb{R}^n$ if for all $\epsilon > 0$ there is a $\delta > 0$ such that

$$\|F(x) - F(\overline{x})\| \leq \epsilon \quad \text{whenever} \quad \|x - \overline{x}\| \leq \delta \ .$$

(2) F is said to be continuous on a set $S \subset \mathbb{R}^n$ if it is continuous at every point of S.
(3) The function F is said to be continuous relative to a set $S \subset \mathbb{R}^n$ if

$$\|F(x) - F(\overline{x})\| \leq \epsilon \quad \text{whenever} \quad \|x - \overline{x}\| \leq \delta \quad \text{and} \quad x \in S \ .$$

(4) The function F is said to be uniformly continuous on a set $S \subset \mathbb{R}^n$ if if for all $\epsilon > 0$ there is a $\delta > 0$ such that

$$\|F(x) - F(y)\| \leq \epsilon \quad \text{whenever} \quad \|x - y\| \leq \delta \quad \text{and} \quad x, y \in S.$$

Norms allow us to define certain topological notions that are very helpful in analizing the behavior of sequences and functions. Since we will make frequent use of these concepts, it is helpful to have certain notational conventions associated with norms. We list a few of these below:

the closed unit ball	$\mathbb{B} := \{x \mid \|x\| \leq 1\}$
the unit vectors	$\mathbb{S} := \{x \mid \|x\| = 1\}$
ϵ-ball about $\overline{x}$	$\overline{x} + \epsilon\mathbb{B} := \{x + \epsilon u \mid u \in \mathbb{B}\} = \{x \mid \|x - \overline{x}\| \leq \epsilon\}$

The unit ball associated with the 1, 2, and ∞ norms will be denoted by $\mathbb{B}_1$, $\mathbb{B}_2$, and $\mathbb{B}_\infty$, respectively.

A few basic topological notions are listed in the following definition. The most important of these for our purposes is *compactness*.

DEFINITION 1.3. *Let S be a subset of $\mathbb{R}^n$, and let $\|\cdot\|$ be a norm on $\mathbb{R}^n$.*

1. *The set S is said to be an open set if for every $\overline{x} \in S$ there is an $\epsilon > 0$ such that $\overline{x} + \epsilon\mathbb{B} \subset S$.*
2. *The set S is said to be a closed set if S contains every point $\overline{x} \in \mathbb{R}^n$ for which there is a sequence $\{x^k\} \subset S$ with $\lim_{k\to\infty} \|x^k - \overline{x}\| = 0$.*
3. *The set S is said to be a bounded set set if there is a $\beta > 0$ such that $S \subset \beta\mathbb{B}$.*
4. *The set S is said to be a compact set if it is both closed and bounded.*
5. *A point $\overline{x} \in \mathbb{R}^n$ is a cluster point of the set S if there is a sequence $\{x^k\} \subset S$ with $\lim_{k\to\infty} \|x^k - \overline{x}\| = 0$.*
6. *A point $\overline{x} \in \mathbb{R}^n$ is said to be a boundary point of the set S if for all $\epsilon > 0$, $(\overline{x} + \epsilon\mathbb{B}) \cap S \neq \emptyset$ while $(\overline{x} + \epsilon\mathbb{B}) \not\subset S$, i.e., every ϵ ball about $\overline{x}$ contains points that are in S and points that are not in S.*

The importance of the notion of compactness in optimization is illustrated in following basic theorems from analysis that we make extensive use of, but do not prove.

THEOREM 1.1. *[Compactness implies Uniform Continuity] Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous function on an open set $S \subset \mathbb{R}^n$. Then F is uniformly continuous on every compact subset of S.*

THEOREM 1.2. *[Weierstrass Compactness Theorem] A set $D \subset \mathbb{R}^n$ is compact if and only if every infinite sequence in D has a cluster point in D.*

THEOREM 1.3. *[Weierstrass Extreme Value Theorem] Every continuous function on a compact set attains its extreme values on that set. That is, there are points in the set at which both the infimum and the supremum of the function relative to the set are attained.*

We will also have need of a norm on the space of matrices. First note that the space of matrices $\mathbb{R}^{m\times n}$ is itself a vector space since it is closed with respect to addition and real scalar multiplication with both operations being distributive and commutative and $\mathbb{R}^{m\times n}$ contains the zero matrix. In addition, we can embed $\mathbb{R}^{m\times n}$ in $\mathbb{R}^{mn}$ by stacking one column on top of another to get a long vector of length mn. This process of stacking the columns is denoted by the vec operator (column vec): given $A \in \mathbb{R}^{m\times n}$,

$$\operatorname{vec}(A) = \begin{pmatrix} A_{\cdot 1} \\ A_{\cdot 2} \\ \vdots \\ A_{\cdot n} \end{pmatrix} \in \mathbb{R}^{mn} \ .$$

EXAMPLE 1.2.

$$\operatorname{vec}\begin{bmatrix} 1 & 2 & -3 \\ 0 & -1 & 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2 \\ -1 \\ -3 \\ 4 \end{bmatrix}$$

Using the vec operation, we define an inner product on $\mathbb{R}^{m\times n}$ by taking the inner product of these vectors of length mn. Given $A, B \in \mathbb{R}^{m\times n}$ we write this inner product as $\langle A, B\rangle$. It is easy to show that this inner product obeys the formula

$$\langle A, B\rangle = \operatorname{vec}(A)^T \operatorname{vec}(B) = \operatorname{tr}\left(A^T B\right).$$

This is known as the *Frobenius inner product*. It generates a corresponding norm, called the *Frobenius norm*, by setting

$$\|A\|_F := \|\mathrm{vec}(A)\|_2 = \sqrt{\langle A,\, A\rangle}.$$

Note that for a given $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$ we have

$$\|Ax\|_2^2 = \sum_{i=1}^m (A_{i\cdot} \bullet x)^2 \le \sum_{i=1}^m (\|A_{i\cdot}\|_2 \|x\|_2)^2 = \|x\|_2^2 \sum_{i=1}^m \|A_{i\cdot}\|_2^2 = \|A\|_F^2 \|x\|_2^2 ,$$

and so

$$\|Ax\|_2 \le \|A\|_F \|x\|_2 . \tag{61}$$

This relationship between the Frobenius norm and the 2-norm is very important and is used extensively in our development. In particular, this implies that for any two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times k}$ we have

$$\|AB\|_F \le \|A\|_F \|B\|_F \ .$$

2. Differentiation

In this section we use our understanding of differentiability for mappings from $\mathbb{R}$ to $\mathbb{R}$ to build a theory of differentiation for mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$. Let F be a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ which we denote by $F : \mathbb{R}^n \to \mathbb{R}^m$. Let the component functions of F be denoted by $F_i : \mathbb{R}^n \to \mathbb{R}$:

$$F(x) = \begin{pmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_m(x) \end{pmatrix} .$$

Example 2.1.

$$F(x) = F\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3x_1^2 + x_1x_2x_3 \\ 2\cos(x_1)\sin(x_2x_3) \\ \ln[\exp(x_1^2 + x_2^2 + x_3^2)] \\ 1/\sqrt{1 + (x_2x_3)^2} \end{pmatrix} .$$

In this case, $n = 3$, $m = 4$, *and*

$$F_1(x) = 3x_1^2 + x_1x_2x_3, \quad F_2(x) = 2\cos(x_1)\sin(x_2x_3), \quad F_3(x) = \ln[\exp(x_1^2 + x_2^2 + x_3^2)], \quad F_4(x) = 1/\sqrt{1 + (x_2x_3)^2} .$$

The first step in understanding the differentiability of mappings on $\mathbb{R}^n$ is to study their one dimensional properties. For this, consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and let x and d be elements of $\mathbb{R}^n$. We define the *directional derivative* of f in the direction d, when it exits, to be the one sided limit

$$f'(x; d) := \lim_{t\downarrow 0} \frac{f(x + td) - f(x)}{t}.$$

Example 2.2. *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be given by* $f(x_1, x_2) := x_1 |x_2|$, *and let* $x = (1, 0)^T$ *and* $d = (2, 2)$. *Then,*

$$f'(x; d) = \lim_{t\downarrow 0} \frac{f(x + td) - f(x)}{t} = \lim_{t\downarrow 0} \frac{(1 + 2t)\ |0 + 2t| - 1\ |0|}{t} = \lim_{t\downarrow 0} \frac{2(1 + 2t)t}{t} = 2 ,$$

while, for $d = -(2, 2)^T$,

$$f'(x; d) = \lim_{t\downarrow 0} \frac{f(x + td) - f(x)}{t} = \lim_{t\downarrow 0} \frac{(1 - 2t)\ |0 - 2t| - 1\ |0|}{t} = \lim_{t\downarrow 0} \frac{2(1 - 2t)t}{t} = 2 .$$

In general, we have

$$f'((1, 0); (d_1, d_2)) = \lim_{t\downarrow 0} \frac{(1 + d_1t)\ |d_2t|}{t} = |d_2|.$$

For technical reasons, we allow this limit to take the values $\pm\infty$. For example, if $f(x) = x^{1/3}$, then

$$f'(0; 1) = \lim_{t\downarrow 0} t^{-2/3} = +\infty \quad \text{and} \quad f'(0; -1) = \lim_{t\downarrow 0} -t^{-2/3} = -\infty.$$

This example as well as the one given in Example 2.2 show that the directional derivative $f'(x;d)$ is not necessarily either continuous or smooth in the d argument even if it exists for all choices of d. However, the directional derivative is always *positively homogeneous* in the sense that, given $\lambda \geq 0$, we have

$$f'(x;\lambda d) = \lim_{t\downarrow 0} \frac{f(x+\lambda t d) - f(x)}{t} \lambda \lim_{t\downarrow 0} \frac{f(x+\lambda t d) - f(x)}{\lambda t} = \lambda f'(x;d) \ .$$

The directional derivative idea can be extended to functions F mapping $\mathbb{R}^n$ into $\mathbb{R}^m$ by defining it componentwise: if the limit

$$F'(x;d) := \lim_{t\downarrow 0} \frac{F(x+td) - F(x)}{t} = \begin{pmatrix} \lim_{t\downarrow 0} \frac{F_1(x+td)-F_1(x)}{t} \\ \lim_{t\downarrow 0} \frac{F_2(x+td)-F_2(x)}{t} \\ \vdots \\ \lim_{t\downarrow 0} \frac{F_m(x+td)-F_m(x)}{t} \end{pmatrix}$$

exists, it is called the directional derivative of F at x in the direction d.

These elementary ideas lead to the following notions of differentiability.

Definition 2.1. *[Differentiable Functions] Let $f : \mathbb{R}^n \to \mathbb{R}$ and $F : \mathbb{R}^n \to \mathbb{R}^m$.*

(1) If $f'(x;d) = \lim_{t\to 0} \frac{f(x+\lambda t d)-f(x)}{t}$, then we say that f is differentiable in the direction d, in which case $f'(x;-d) = -f'(x;d)$.

(2) Let e_j $j = 1,\ldots,n$ denote the unit coordinate vectors. If f is differentiable in the direction e_j, we say that the partial derivative of f with respect to the component x_j exists and write

$$\frac{\partial f(x)}{\partial x_j} := f'(x;e_j).$$

In particular, we have

$$f(x+te_j) = f(x) + t\frac{\partial f(x)}{\partial x_j} + o(t), \quad \text{where } \lim_{t\to 0} \frac{o(t)}{t} = 0.$$

Note that $\frac{\partial f(\cdot)}{\partial x_j} : \mathbb{R}^n \to \mathbb{R}$.

(3) We say that f is (Fréchet) differentiable at $x \in \mathbb{R}^n$ if there is a vector $g \in \mathbb{R}^n$ such that

$$\lim_{y\to x} \frac{|f(y) - f(x) - g^T(y-x)|}{\|y-x\|} = 0.$$

If such a vector g exists, we write $g = \nabla f(x)$ and call $\nabla f(x)$ the gradient of f at x. In particular, the differentiability of f at x is equivalent to the following statement:

$$f(y) = f(x) + \nabla f(x)^T(y-x) + o(\|y-x\|)$$

for all y near x, where $\lim_{y\to x} \frac{o(\|y-x\|)}{\|y-x\|} = 0$.

(4) We say that F is (Fréchet) differentiable at $x \in \mathbb{R}^n$ if there is a matrix $J \in \mathbb{R}^{m\times n}$ such that

$$\lim_{y\to x} \frac{\|F(y) - F(x) - J(y-x)\|}{\|y-x\|} = 0.$$

If such a matrix J exists, we write $J = \nabla F(x)$ and call $\nabla F(x)$ the Jacobian of F at x. In particular, the differentiability of f at x is equivalent to the following statement:

$$F(y) = F(x) + \nabla F(x)^T(y-x) + o(\|y-x\|)$$

for all y near x, where $\lim_{y\to x} \frac{o(\|y-x\|)}{\|y-x\|} = 0$.

Remark 2.1. *Note that there is an inconsistency here in the use of the ∇ notation when $F : \mathbb{R}^n \to \mathbb{R}^m$ with $m = 1$. The inconsistency arises due to the presense of g^T in Part (3) of Definition 2.1 and the absence of a transpose in Part (4) of this definition. For this reason, we must take extra care in in interpreting this notation in this case.*

Remark 2.2. *[Little-o Notation] In these notes we use the notation $o(t)$ to represent any element of a function class for which $\lim_{t\to 0} \frac{o(t)}{t} = 0$. In particular, this implies that for all $\alpha \in \mathbb{R}$*

$$\alpha o(t) = o(t), \ o(t) + o(t) = o(t), \quad \text{and} \quad t^s o(t^r) = o(t^{r+s}).$$

Several observations about these notions of differentiability are in order. First, the existence of the directional derivative $f'(x;d)$ nor the differentiability of f at x in the direction d requires the continuity of the function at that point. Second, the existence of $f'(x;d)$ in all directions d does imply the continuity of the mapping $d \mapsto f'(x;d)$. Therefore, the directional derivative, although useful, is a very weak object to describe the local variational properties of a function. On the other hand, differentiability is a very powerful statement. A few consequences of differentiability are listed in the following theorem.

THEOREM 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $F : \mathbb{R}^n \to \mathbb{R}^m$.*

(1) If f is differentiable at x, then

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

and $f'(x;d) = \nabla f(x)^T d$ for all $d \in \mathbb{R}^n$.

(2) If F is differentiable at x, then

$$(\nabla F(x))_{ij} = \frac{\partial F_i(x)}{\partial x_j} \quad i = 1, \ldots, m \quad \text{and} \quad j = 1, 2, \ldots, n.$$

(3) If F is differentiable at a point x, then it is necessarily continuous at x.

Higher order derivatives are obtained by applying these notions of differentiability to the derivatives themselves. For example, to compute the second derivative, the derivative needs to exist at all points near the point at which the second derivative needs to be computed so that the necessary limit is well defined. From the above, we know that the partial derivative $\frac{\partial F_i(x)}{\partial x_j}$, when it exists, is a mapping from $\mathbb{R}^n$ to $\mathbb{R}$. Therefore, it is possible to consider the partial derivatives of these partial derivatives. For such partial derivatives we use the notation

$$\frac{\partial^2 F_i(x)}{\partial x_j \partial x_k} := \frac{\partial \left(\frac{\partial F_i(x)}{\partial x_k} \right)}{\partial x_j}. \tag{62}$$

The second derivative of $f : \mathbb{R}^n \to \mathbb{R}$ is the derivative of the mapping $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, and we write $\nabla(\nabla f(x)) =: \nabla^2 f(x)$. We call $\nabla^2 f(x)$ the *Hessian* of f at x. By (62), we have

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

We have the following key property of the Hessian.

THEOREM 2.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that all of the second partials $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$, $ij = 1, 2, \ldots, n$ exist and are continuous near $x \in \mathbb{R}^n$. Then $\nabla^2 f(x)$ is a real symmetric matrix, i.e., $\nabla^2 f(x) = \nabla^2 f(x)^T$.*

The partial derivative representations of the gradient, Hessian, and Jacobian matrices is a convenient tool for computing these objects. For example, if we have

$$f(x) := 3x_1^2 + x_1 x_2 x_3,$$

then

$$\nabla f(x) = \begin{pmatrix} 6x_1 + x_2 x_3 \\ x_1 x_3 \\ x_1 x_2 \end{pmatrix} \quad \text{and} \quad \nabla^2 f(x) = \begin{bmatrix} 6 & x_3 & x_2 \\ x_3 & 0 & x_1 \\ x_2 & x_1 & 0 \end{bmatrix}.$$

However, the partial derivatives are not the only tool for computing derivatives. In many cases, it is easier to compute the gradient, Hessian, and/or Jacobian directly from the definition using the little-o notation.

3. The Delta Method for Computing Derivatives

Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be differentiable at a point x if there is a vector $g \in \mathbb{R}^n$ such that

$$f(x + \Delta x) = f(x) + g^T \Delta x + o(\|\Delta x\|). \tag{63}$$

Hence, if we can write $f(x + \Delta x)$ in this form, then $g = \nabla f(x)$. To see how to use this idea, consider the least squares objective function

$$f(x) = \tfrac{1}{2} \|Ax - b\|_2^2, \qquad \text{where } A \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^m.$$

Then

$$\begin{aligned} f(x + \Delta x) &= \tfrac{1}{2} \|A(x + \Delta x) - b\|_2^2 \\ &= \tfrac{1}{2} \|(Ax - b) + A\Delta x\|_2^2 \\ &= \tfrac{1}{2} \|Ax - b\|_2^2 + (Ax - b)^T A\Delta x + \tfrac{1}{2} \|A\Delta x\|_2^2 \\ &= f(x) + (A^T(Ax - b))^T \Delta x + \tfrac{1}{2} \|A\Delta x\|_2^2 . \end{aligned} \tag{64}$$

In this expression, $\frac{1}{2} \|A\Delta x\|_2^2 = o(\|\Delta x\|_2)$ since

$$\frac{\frac{1}{2} \|A\Delta x\|_2^2}{\|\Delta x\|_2} = \tfrac{1}{2} \|A\Delta x\|_2 \left\| A \frac{\Delta x}{\|\Delta x\|_2} \right\|_2 \to 0 \quad \text{as} \quad \|\Delta x\|_2 \to 0 \ .$$

Therefore, by (63), the expression (64) tells us that

$$\nabla f(x) = A^T(Ax - b).$$

This approach to computing the derivative of a function is called the *delta method*. In a similar manner it can be used to compute the Hessian of f by applying the approach to ∇f:

$$\nabla f(x + \Delta x) = A^T(A(x + \Delta x) - b) = A^T(Ax - b) + A^T A\Delta x = \nabla f(x) + A^T A\Delta x,$$

and, hence, $\nabla^2 f(x) = A^T A$.

Let us now apply the delta method to compute the gradient and Hessian of the quadratic function

$$f(x) := \tfrac{1}{2} x^T Hx + g^T x, \quad \text{where } \ H \in \mathcal{S}^n,\ g \in \mathbb{R}^n.$$

Then

$$\begin{aligned} f(x + \Delta x) &= \tfrac{1}{2}(x + \Delta x)^T H(x + \Delta x) + g^T(x + \Delta x) \\ &= \tfrac{1}{2} x^T Hx + g^T x + (Hx + g)^T \Delta x + \tfrac{1}{2} \Delta x^T H\Delta x \\ &= f(x) + (Hx + g)^T \Delta x + \tfrac{1}{2} \Delta x^T H\Delta x, \end{aligned}$$

where $\frac{1}{2} \Delta x^T H\Delta x = o(\|\Delta x\|_2)$ since

$$\frac{\frac{1}{2} \Delta x^T H\Delta x}{\|\Delta x\|_2} = \tfrac{1}{2} \Delta x^T H \frac{\Delta x}{\|\Delta x\|_2} \to 0 \ .$$

Therefore, by (63), we must have

$$\nabla f(x) = Hx + g \ .$$

Again, we compute the Hessian by applying the delta method to the gradient:

$$\nabla f(x + \Delta x) = H(x + \Delta x) + g = (Hx + g) + H\Delta x = \nabla f(x) + H\Delta x,$$

and so

$$\nabla^2 f(x) = H \ .$$

4. Differential Calculus

There many further tools for computing derivatives that do not require a direct appeal to either the partial derivatives or the delta method. These tools allow us to compute new derivatives from derivatives that are already known based on a *calculus* of differentiation. We are familiar with this differential calculus for functions mapping $\mathbb{R}$ to $\mathbb{R}$. Here we show how a few of these calculus rules extend to mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$. The most elementary of these are the facts that the derivative of the scalar multiple of a function equals the scalar multiple of the derivative and the derivative of a sum is the sum of derivatives: given $F : \mathbb{R}^n \to \mathbb{R}^m$, $G : \mathbb{R}^n \to \mathbb{R}^m$ and $\alpha \in \mathbb{R}$,

$$\nabla(\alpha F) = \alpha \nabla F \quad \text{and} \quad \nabla(F + G) = \nabla F + \nabla G \ .$$

These rules are themselves derivable from the much more powerful *chain rule*.

THEOREM 4.1. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and $H : \mathbb{R}^m \to \mathbb{R}^k$ be such that F is differentiable at x and H is differentiable at $F(x)$. The $G := H \circ F$ is differentiable at x with*

$$\nabla G(x) = \nabla H(F(x)) \circ \nabla F(x) \ .$$

REMARK 4.1. *As noted in Remark 2.1, one must take special care in the interpretation of this chain rule when $k = 1$ due to the presence of an additional transpose. In this case,*

$$\nabla G(x) = \nabla F(x)^T \nabla H(F(x)) \ .$$

For example, let $F : \mathbb{R}^n \to \mathbb{R}^m$ and consider the function

$$f(x) := \tfrac{1}{2} \|F(x)\|_2^2 = (\tfrac{1}{2} \|\cdot\|_2^2) \circ F(x),$$

that is, we are composing half the 2-norm squared with F. Since $\nabla(\frac{1}{2} \|\cdot\|_2^2)(y) = y$, we have

$$\nabla f(x) = \nabla F(x)^T F(x) \ .$$

This chain rule computation can be verified using the delta method:

$$\begin{aligned}
f(x + \Delta x) &= \tfrac{1}{2} \|F(x + \Delta x)\|_2^2 \\
&= \tfrac{1}{2} \|F(x) + \nabla F(x)\Delta x + o(\|\Delta x\|_2)\|_2^2 \\
&= \tfrac{1}{2} \|F(x) + \nabla F(x)\Delta x\|_2^2 + (F(x) + \nabla F(x)\Delta x)^T (o(\|\Delta x\|_2)) + \tfrac{1}{2} \|o(\|\Delta x\|_2)\|_2^2 \\
&= \tfrac{1}{2} \|F(x) + \nabla F(x)\Delta x\|_2 + o(\|\Delta x\|_2) \\
&= \tfrac{1}{2} \|F(x)\|_2^2 + (\nabla F(x)^T F(x))^T \Delta x + \tfrac{1}{2} \|\nabla F(x)\Delta x\|_2^2 + o(\|\Delta x\|_2) \\
&= f(x) + (\nabla F(x)^T F(x))^T \Delta x + o(\|\Delta x\|_2),
\end{aligned}$$

where $\lim_{t \to 0} \frac{o(t)}{t} = 0$ and we have used this notation as described in Remark 4.1. Hence, again $\nabla f(x) = \nabla F(x)^T F(x)$.

5. The Mean Value Theorem

Given $f : \mathbb{R}^n \to \mathbb{R}$, the defining formula for the derivative,

$$f(y) = f(x) + \nabla f(x)(y - x) + o(\|y - x\|),$$

is a powerful tool for understanding the local behavior of the function f near x. If we drop the little-o term from the right hand side, we obtain the *first-order Taylor expansion* of f at x. This is called a *first-order approximation* to f at x due to the fact that the power of $\|y - x\|$ in the *error term* $o(\|y - x\|)$ is 1. Higher order approximations to f can be obtained using higher order derivatives. But before turning to these approximations, we make a closer study of the first-order expansion. Im particular, we wish the extend the *Mean Value Theorem* to functions of many variables.

THEOREM 5.1. *[1-Dimensional Mean Value Theorem]*
Let $\phi| : \mathbb{R} \to \mathbb{R}$ be $k+1$ times differentiable on an open interval $(a, b) \subset \mathbb{R}$. Then, for every $x, y \in (a, b)$ with $x \neq y$, there exists a $z \in (a, b)$ strictly between x and y such that

$$\phi(y) = \phi(x) + \phi'(x)(y - x) + \cdots + \frac{1}{k!}\phi^{(k)}(x)(y - x)^k + \frac{1}{(k + 1)!}\phi^{(k+1)}(z)(y - x)^{(k+1)} \ .$$

We use this results to easily obtain the following mean value theorem for function mapping $\mathbb{R}^n$ to $\mathbb{R}$.

THEOREM 5.2. *[n-Dimensional Mean Value Theorem]*
Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on an open set containing the two points $x, y \in \mathbb{R}^n$ with $x \neq y$. Define the closed and open line segments connecting x and y by

$$[x, y] := \{(1-\lambda)x + \lambda y \,|\, 0 \leq \lambda \leq y\} \quad \text{and} \quad (x, y) := \{(1-\lambda)x + \lambda y \,|\, 0 < \lambda < y\},$$

respectively. Then there exists a $z, w \in (x, y)$ such that

$$f(y) = f(x) + \nabla f(z)^T (y - x) \quad \text{and} \quad f(y) = f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(z)(y - x).$$

PROOF. Define the function $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(t) := f(x + t(y - x))$. Since f is differentiable, so is ϕ and the chain rule tells us that

$$\phi'(t) = \nabla f(x + t(y - x))^T (y - x) \quad \text{and} \quad \phi'(t) = (y - x)^T \nabla^2 f(x + t(y - x))(y - x).$$

By applying the Mean Value Theorem 5.1 to ϕ we obtain the existence of $t, s \in (0, 1)$ such that

$$f(y) = \phi(1) = \phi(0) + \phi'(t)(1 - 0) = f(x) + \nabla f(x + t(y - x))^T (y - x)$$

and

$$f(y) = \phi(1) = \phi(0) + \phi'(0)(1 - 0) + \tfrac{1}{2}\phi''(s)(1 - 0)^2 = f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(x + s(y - x))(y - x).$$

By setting $z := x + t(y - x)$ and $w := x + s(y - x)$ we obtain the result. □

In a similar manner we can apply the Fundamental Theorem of Calculus to such functions.

THEOREM 5.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on an open set containing the two points $x, y \in \mathbb{R}^n$ with $x \neq y$. Then*

$$f(y) = f(x) + \int_0^1 \nabla f(x + t(y - x))^T (y - x)\, dt\ .$$

PROOF. Apply the Fundamental Theorem of Calculus to the function ϕ defined in the proof of Theorem 5.2. □

Unfortunately, the Mean Value Theorem does not extend to general differentiable function mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ for $m > 1$. Nonetheless, we have the following approximate result.

THEOREM 5.4. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable on an open set containing the two points $x, y \in \mathbb{R}^n$ with $x \neq y$. Then*

$$\|F(y) - F(x)\|_2 \leq \left[\max_{z \in [x,y]} \|F'(z)\|_F\right] \|y - x\|_2\ . \tag{65}$$

PROOF. By the Fundamental Theorem of Calculus, we have

$$F(y) - F(x) = \begin{pmatrix} \int_0^1 \nabla F_1(x + t(y - x))^T (y - x)\, dt \\ \vdots \\ \int_0^1 \nabla F_m(x + t(y - x))^T (y - x)\, dt \end{pmatrix} = \int_0^1 \nabla F(x + t(y - x))(y - x)\, dt\ .$$

Therefore,

$$\begin{aligned} \|F(y) - F(x)\|_2 &= \left\| \int_0^1 \nabla F(x + t(y - x))(y - x)\, dt \right\|_2 \\ &\leq \int_0^1 \|\nabla F(x + t(y - x))(y - x)\|_2\, dt \\ &\leq \int_0^1 \|\nabla F(x + t(y - x))\|_F \|y - x\|_2\, dt \\ &\leq \left[\max_{z \in [x,y]} \|F'(z)\|_F\right] \|y - x\|_2\ . \end{aligned}$$

□

The bound (65) is very useful in many applications. But it can be simplified in cases where ∇F is known to be continuous since in this case the Weierstrass extreme value theorem says that, for every $\beta > 0$,

$$\max_{z \in \beta\mathbb{B}} \|F'(z)\|_F =: K < \infty \ .$$

Hence, by Theorem 5.4,

$$\|F(x) - F(y)\|_2 \leq K \|x - y\|_2 \quad \forall\ x, y \in \beta\mathbb{B} \ .$$

This kind of inequality if extremely useful and leads to the following notion of continuity.

DEFINITION 5.1. *[Lipschitz Continuity]*
We say that $F : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous on a set $S \subset \mathbb{R}^n$ if there exists a constant $K > 0$ such that

$$\|F(x) - F(y)\| \leq K \|x - y\| \quad \forall\ x, y \in S \ .$$

The constant K is called the modulus of Lipschitz continuity for F over S, and depends on the choice of norms for $\mathbb{R}^n$ and $\mathbb{R}^m$.

As one application of Lipschitz continuity, we give the following lemma concerning the accuracy of the first-order Taylor approximation of a function.

LEMMA 5.1. *[Quadratic Bound Lemma]*
Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be such that ∇F is Lipschitz continuous on the set $S \subset \mathbb{R}^n$. If $x, y \in S$ are such that $[x, y] \subset S$, then

$$\|F(y) - (F(x) + \nabla F(x)(y - x))\|_2 \leq \frac{K}{2} \|y - x\|_2^2 ,$$

where K is the modulus of Lipschitz continuity for ∇F on S.

PROOF. Observe that

$$\begin{aligned} F(y) - F(x) - \nabla F(x)(y - x) &= \int_0^1 \nabla F(x + t(y - x))(y - x)dt - \nabla F(x)(y - x) \\ &= \int_0^1 [\nabla F(x + t(y - x)) - \nabla F(x)](y - x)dt. \end{aligned}$$

Hence

$$\begin{aligned} \|F(y) - (F(x) + \nabla F(x)(y - x))\|_2 &= \left\| \int_0^1 [\nabla F(x + t(y - x)) - \nabla F(x)](y - x)dt \right\|_2 \\ &\leq \int_0^1 \|(\nabla F(x + t(y - x) - \nabla F(x))(y - x)\|_2 \, dt \\ &\leq \int_0^1 \|\nabla F(x + t(y - x)) - \nabla F(x)\|_F \, \|y - x\|_2 \, dt \\ &\leq \int_0^1 Kt \, \|y - x\|_2^2 \, dt \\ &= \tfrac{K}{2} \|y - x\|_2^2 . \end{aligned}$$

□

The Mean Value Theorem also allows to obtain the following *second order* approximation.

THEOREM 5.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose that $\nabla^2 f(x)$ exists. Then*

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + o(\|y - x\|^2). \tag{66}$$

PROOF. The mean value theorem tells us the for every $y \in x + \epsilon\mathbb{B}$ there is a $z \in (x, y)$ such that

$$\begin{aligned} f(y) &= f(x) + \nabla f(z)^T (y - x) \\ &= f(x) + (\nabla f(x) + \nabla^2 f(x)(y - x) + o(\|y - x\|))^T (y - x) \\ &= f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + o(\|y - x\|^2). \end{aligned}$$

□

If we drop the $o(\|y - x\|^2)$ in the equation (66), we obtain the *second-order Taylor approximation* to f at x. This is a second-order approximation since the power of $\|y - x\|$ in the little-o term is 2, i.e., $o(\|y - x\|^2)$.