TMA 4180 Optimization Theory

Basic Mathematical Tools

H. E. Krogstad, IMF, spring 2008

1 INTRODUCTION

During the lectures we need some basic topics and concepts from mathematical analysis. This material is actually not so difficult, – if you happen to have seen it before. If this is the first time, experience has shown that even if it looks simple and obvious, it is necessary to spend some time digesting it.

Nevertheless, the note should be read somewhat relaxed. Not all details are included, nor are all proofs written out in detail. After all, this is not a course in mathematical analysis.

Among the central topics are the *Taylor Formula* in n dimensions, the general optimization setting, and above all, basic properties of convex sets and convex functions. A very short review about matrix norms and Hilbert space has also been included. The big optimization theorem in Hilbert space is the *Projection Theorem*. Its significance in modern technology and signal processing can hardly be over-emphasized, although it is often disguised under other fancy names.

The final result in the note is the *Implicit Function Theorem* which ensures the existence of solutions of implicit equations.

The abbreviation N&W refers to the textbook, J. Nocedal and S. Wright: *Numerical Optimization*, Springer. Note that page numbers in the first and second editions are different.

2 TERMINOLOGY AND BASICS

Vectors in $\mathbb{R}^n$ are, for simplicity, denoted by regular letters, x, y, z, $\cdots$, and $\|x\|$ is used for their length (norm),

$$\|x\| = \left(x_1^2 + x_2^2 + \cdots + x_n^2\right)^{1/2}. \tag{1}$$

Occasionally, $x_1, x_2, \cdots$ will also mean a sequence of vectors, but the meaning of the indices will then be clear from the context.

We are considering functions f from $\mathbb{R}^n$ to $\mathbb{R}$. Such a function will often be defined for all or most of $\mathbb{R}^n$, but we may only be considering f on a subset $\Omega \subset \mathbb{R}^n$. Since the definition domain of f typically extends Ω, it is in general not a problem to define the derivatives of f, $\frac{\partial f}{\partial x_i}$, also on the boundary of Ω. The gradient, ∇f, is a vector, and in mathematics (but not in N&W!) it is considered to be a *row vector*,

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right). \tag{2}$$

We shall follow this convention and write $\nabla f\,(x)\,p$ for $\nabla f\,(x) \cdot p$. There are, however, some situations where the direction d, defined by the gradient is needed, and then $d = \nabla f'$. In the lectures we use $'$ for transposing vectors and matrices.

A set $V \subset \mathbb{R}^n$ is *open* if all points in V may be surrounded by a ball in $\mathbb{R}^n$ belonging to V: For all $x_0 \in V$, there is an $r > 0$ such that

$$\{x \ ; \ \|x - x_0\| < r\} \subset V. \tag{3}$$

(This notation means "The collection of all x-s such that $\|x - x_0\| < r$").

It is convenient also to say that a set $V \subset \Omega$ is *open in* Ω if there is an open set $W \subset \mathbb{R}^n$ such that $V = W \cap \Omega$ (The mathematical term for this is a *relatively open* set). Let $\Omega = [0, 1] \subset \mathbb{R}$. The set $[0, 1/2)$ is not open in $\mathbb{R}$ (why?). However, as a subset of Ω, $[0, 1/2) \subset [0, 1]$, it is open in Ω, since $[0, 1/2) = (-1/2, 1/2) \cap [0, 1]$ (Think about this for a while!).

A *neighborhood* N of a point x is simply an open set containing x.

2.1 Sup and Inf – Max and Min

Consider a set S of real numbers. The *supremum* of the set, denoted

$$\sup S, \tag{4}$$

is the *smallest number that is equal to or larger than all members of the set.*

It is a very fundamental property of real numbers that the supremum always exists, although it may be infinite. If there is a member $x_0 \in S$ such that

$$x_0 = \sup S, \tag{5}$$

then x_0 is called a *maximum* and written

$$x_0 = \max S. \tag{6}$$

Sometimes such a maximum does not exists: Let

$$S = \left\{ 1 - \frac{1}{n} \, ; n = 1, 2, \cdots \right\}. \tag{7}$$

In this case, there is *no* maximum element in S. However, $\sup S = 1$, since no number less than 1 fits the definition. Nevertheless, 1 is *not* a maximum, since it is not a member of the set. This is the rule:

A supremum always exists, but may be $+\infty$. If a maximum exists, it is equal to the supremum.

For example,

$$\begin{aligned} \sup\{1, 2, 3\} &= \max\{1, 2, 3\} = 3, \\ \sup\{x; 0 < x < 3\} &= 3, \\ \sup\{1, 2, 3, \cdots\} &= \infty. \end{aligned} \tag{8}$$

The *infimum* of a set S, denoted

$$\inf S, \tag{9}$$

is the *largest number that is smaller than or equal to all members in the set.*

The *minimum* is defined accordingly, and the rule is the same.

We will only meet sup and inf in connection with real numbers, although this can be defined for other mathematical structures as well. As noted above, the existence of supremum and infimum is quite fundamental for real numbers!

A set S of real numbers is *bounded above* if $\sup S$ is finite ($\sup S < \infty$), and *bounded below* if $\inf S$ is finite ($-\infty < \inf S$). The set is *bounded* if both $\sup S$ and $\inf S$ are finite.

2.2 Convergence of Sequences

A *Cauchy sequence* $\{x_i\}_{i=1}^{\infty}$ of real numbers is a sequence where

$$\lim_{n\to\infty}\left(\sup_{m\geq n}|x_m - x_n|\right) = 0. \tag{10}$$

This definition is a bit tricky, but if *you* pick an $\varepsilon > 0$, I can always find an n_ε such that

$$|x_m - x_{n_\varepsilon}| < \varepsilon \tag{11}$$

for *all* x_m where $m > n_\varepsilon$.

Another very basic property of real numbers is that *all Cauchy sequences converge*, that is,

$$\lim_{n\to\infty} x_n = a \tag{12}$$

for a (unique) real number a.

A sequence $S = \{x_n\}_{n=1}^{\infty}$ is *monotonically increasing* if

$$x_1 \leq x_2 \leq x_3 \leq \cdots. \tag{13}$$

A monotonically increasing sequence is always convergent,

$$\lim_{n\to\infty} x_n = \sup S, \tag{14}$$

(it may diverge to $+\infty$). Thus, a monotonically increasing sequence that is *bounded above,* is always convergent (You should try to prove this by applying the definition of sup and the definition of a Cauchy sequence!).

Similar results also apply for *monotonically decreasing* sequences.

2.3 Compact Sets

A set S in $\mathbb{R}^n$ is *bounded* if

$$\sup_{x\in S}\|x\| < \infty. \tag{15}$$

A Cauchy sequence $S = \{x_n\}_{n=1}^{\infty} \subset \mathbb{R}^n$ is a sequence such that

$$\lim_{n\to\infty}\left(\sup_{m\geq n}\|x_m - x_n\|\right) = 0. \tag{16}$$

It is easy to see, by noting that every component of the vectors is a sequence of real numbers, that all Cauchy sequences in $\mathbb{R}^n$ converge.

A set C in $\mathbb{R}^n$ is *closed* if all Cauchy sequences that can be formed from elements in C converge to elements in C. This may be a bit difficult to grasp: Can you see why the interval $[0, 1]$ is closed, while $(0, 1)$ or $(0, 1]$ are not? What about $[0, \infty)$? Thus, a set is closed if it already contains all the limits of its Cauchy sequences. By adding these limits to an arbitrary set C, we *close* it, and write $\bar{C}$ for the *closure* of C. For example,

$$\overline{(0, 1)} = [0, 1]\,. \tag{17}$$

Consider a bounded sequence $S = \{x_n\}_{i=1}^{\infty}$ in $\mathbb{R}$, and assume for simplicity that

$$0 = \inf S \leq x_n \leq \sup S = 1. \tag{18}$$

Split the interval $[0\ ,\ 1]$ into half, say $[0\ ,\frac{1}{2})$ and $[\frac{1}{2}\ ,\ 1]$. Select one of these intervals containing *infinitely many elements* from S, and pick one $x_{n_1} \in S$ from the same interval. Repeat the operation by halving this interval and selecting another element x_{n_2}. Continue the same way. On step k, the interval I_k will have length 2^{-k} and all later elements x_{n_k}, $x_{n_{k+1}}$, $x_{n_{k+2}}$,$\cdots$ will be members of I_k. This makes the *sub-sequence* $\{x_{n_k}\}_{k=1}^{\infty} \subset S$ into a Cauchy sequence (why?), and hence it converges. A similar argument works for a sequence in $\mathbb{R}^n$.

A closed set with the property that all bounded sequences have convergent subsequences, is called *compact* (this is a mathematical term, not really related to the everyday meaning of the word).

By an easy adaptation of the argument above, *we have now proved that all bounded and closed sets in* $\mathbb{R}^n$ *are compact.*

Of course, as long as the set above is bounded, $\{x_{n_k}\}_{k=1}^{\infty}$ will be convergent, but the limit may not belong to the set, unless it is closed.

If you know the Hilbert space l^2 (or see below) consisting of all infinite-dimensional vectors $x = \{\alpha_1, \alpha_2, \cdots\}$ such that $\|x\|^2 = \sum_{i=1}^{\infty} |\alpha_i|^2 < \infty$, you will probably also know that the unit ball, $B = \{x\ ;\ \|x\| \leq 1\}$ is bounded (obvious) and closed (not so obvious). All unit vectors $\{e_i\}_{i=1}^{\infty}$ in an orthogonal basis will belong to B. However, $\|e_i - e_j\|^2 = \|e_i\|^2 + \|-e_j\|^2 = 2$,whenever $i \neq j$. We have *no* convergent subsequences in this case, and B is *not* compact! This rather surprising example occurs because l^2 has infinite dimension.

2.4 $\mathcal{O}()$ and $o()$ statements

It is convenient to write that the size of $f(x)$ *is of the order of* $g(x)$ *when* $x \to a$ in the short form

$$f(x) = \mathcal{O}(g(x)),\ x \to a. \tag{19}$$

Mathematically, this means that there exists two finite numbers, m and M such that

$$mg(x) \leq f(x) \leq Mg(x) \tag{20}$$

when $x \to a$. In practice, we often use the notation to mean

$$|f(x)| \leq Mg(x) \tag{21}$$

and assume that lower bound, not very much smaller than $Mg(x)$ can be found. For example,

$$\log(1+x) - x = \mathcal{O}\left(x^2\right)$$

when $x \to 0$.

The other symbol, $o()$, is slightly more precise: We say that $f(x) = o(g(x))$ when $x \to a$ if

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 0. \tag{22}$$

2.5 The Taylor Formula

You should all be familiar with the *Taylor series* of a function g of one variable,

$$g(t) = g(t_0) + g'(t_0)(t - t_0) + \frac{g''(t_0)}{2!}(t - t_0)^2 + \frac{g'''(t_0)}{3!}(t - t_0)^3 + \cdots . \tag{23}$$

The *Taylor Formula* is not a series, but a quite useful *finite identity*. In essence, the Taylor Formula gives an expression for the error between the function and its Taylor series truncated after a finite number of terms.

We shall not dwell with the derivation of the formula, which follows by successive partial integrations of the expression

$$g(t) = g(0) + \int_0^t g'(s)\, ds, \tag{24}$$

and the *Integral Mean Value Theorem*,

$$\int_0^t f(s)\varphi(s)\, ds = f(\xi t)\int_0^t \varphi(s)\, ds,\ \varphi \geq 0,\ f \text{ continuous}, \xi\ \in (0,1).$$

The formulae below state for simplicity the results around $t = 0$, but any point is equally good. The simplest and very useful form of Taylor Formula is also known as the *Secant Formula*:

If the derivative g' exists for all values between 0 and t, there is a $\xi \in (0,1)$ such that

$$g(t) = g(0) + g'(\xi t)\, t. \tag{25}$$

This is an identity. However, since we do *not* know the value of ξ, which in general depends on t, we can not use it for computing $g(t)$! Nevertheless, the argument ξt is at least somewhere in the open interval between 0 and t.

If g' is continuous at $t = 0$, we may write

$$\begin{aligned} g(t) &= g(0) + g'(\xi t)\, t \\ &= g(0) + g'(0)\, t + \left[g'(\xi t) - g'(0)\right] t \\ &= g(0) + g'(0)\, t + o(t). \end{aligned} \tag{26}$$

Moreover, if g'' exists between 0 and t, we have the second order formula,

$$g(t) = g(0) + g'(0)\, t + g''(\xi t)\frac{t^2}{2!} \tag{27}$$

(Try to prove this using the Integral Mean Value Theorem and assuming that g'' is continuous! Be sure to use $s - t$ for the integral of ds).

Hence, if g'' is bounded,

$$g(t) = g(0) + g'(0)\, t + \mathcal{O}\left(t^2\right) \tag{28}$$

The general form of Taylor Formula, around 0 and with sufficiently smooth functions, reads

$$g(t) = \sum_{j=0}^{N} \frac{g^{(j)}(0)}{j!} t^j + R_N(t), \tag{29}$$

$$R_N(t) = \int_0^t \frac{g^{(N+1)}(s)}{N!}(t - s)^N\, ds = \frac{g^{(N+1)}(\xi t)}{(N+1)!} t^{N+1},\ \ \xi \in (0,1). \tag{30}$$

2.6 The n-dimensional Taylor Formula

The n-dimensional Taylor formula will be quite important to us, and the derivation is based on the one-dimensional formula above.

Let $f : \mathbb{R}^n \to \mathbb{R}$, and assume that ∇f exists around $x = 0$. Let us write $g(s) = f(sx)$. Then

$$g'(\xi) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\xi x) \frac{d(sx_i)}{ds}(\xi) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\xi x) x_i = \nabla f(\xi x) x, \tag{31}$$

and we obtain

$$\begin{aligned} f(x) &= g(1) \\ &= g(0) + g'(\xi) \cdot 1 \\ &= f(0) + \nabla f(\xi x) x, \ \xi \in (0,1), \end{aligned} \tag{32}$$

which is the n-dimensional analogue of the Secant Formula. Note that the point ξx is somewhere on the line segment between 0 and x, and that the *same* ξ applies to all components of x (but again, ξ is an unknown function of x).

As above, if ∇f is continuous at $x = 0$,

$$\begin{aligned} f(x) &= f(0) + \nabla f(0) x + (\nabla f(\xi x) - \nabla f(0)) x \\ &= f(0) + \nabla f(0) x + o(\|x\|). \end{aligned} \tag{33}$$

At this point we make an important digression. If a relation

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + o(\|x - x_0\|) \tag{34}$$

holds at x_0, we say that f is *differentiable at* x_0. The linear function

$$T_{x_0}(x) \triangleq f(x_0) + \nabla f(x_0)(x - x_0), \tag{35}$$

is then called the *tangent plane* of f at x_0. Thus, for a differentiable function,

$$f(x) = T_{x_0}(x) + o(\|x - x_0\|). \tag{36}$$

Contrary what is stated in the first edition of N&W (and numerous other non-mathematical textbooks!), it is *not* sufficient that all partial derivatives exist at x_0 (Think about this for a while: The components of ∇f contain only partial derivatives of f along the coordinate axis. Find a function on $\mathbb{R}^2$ where $\nabla f(0) = 0$ but which, nevertheless, is not differentiable at $x = 0$. E.g., consider the function defined as $\sin 2\theta$ in polar coordinates)

The next term of the n-dimensional Taylor Formula is derived similarly:

$$g''(\xi) = \frac{d}{ds} \sum_{i=1}^{n} \frac{\partial f(sx)}{\partial x_i} x_i \bigg|_{s=\xi} = \sum_{i,j=1}^{n} \left(\frac{\partial^2 f(sx)}{\partial x_i \partial x_j} \right) \bigg|_{s=\xi} x_j x_i = x' H(\xi x) x. \tag{37}$$

The matrix H is called the *Hess matrix of* f, or the *Hessian*,

$$H(x) = \nabla^2 f(x) = \left\{ \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right\}_{i,j=1}^{n}. \tag{38}$$

Yes, Optimization Theory uses sometimes the unfortunate notation $\nabla^2 f(x)$, which is *not* the familiar Laplacian used in Physics and PDE theory!

From the above, the second order Taylor formula may now be written

$$f(x) = f(0) + \nabla f(0)\, x + \frac{1}{2} x' \nabla^2 f(\xi x)\, x, \quad \xi \in (0,1). \tag{39}$$

Higher order terms get increasingly more complicated and are seldom used.

By truncating the n-dimensional Taylor series after the second term, we end up with what is called a *quadratic function*, or a quadratic form,

$$q(x) = a + b'x + \frac{1}{2} x' A x. \tag{40}$$

By considering quadratic functions we may analyze many important algorithms in optimization theory analytically, and one very important case occurs if A is *positive definite*. The function q is then *convex* (see below) and $\min q(x)$ is obtained for the unique vector

$$x^* = -A^{-1} b. \tag{41}$$

We shall, from time to time, use the notation "$A > 0$" to mean that the matrix A is positive definite (NB! This does not mean that all $a_{ij} > 0$!). Similarly, "$A \geq 0$" means that A is positive semidefinite.

2.7 Matrix Norms

Positive definite matrices lead to what is called *matrix* (or *skew*) *norms* on $\mathbb{R}^n$. The matrix norms are important in the analysis of the Steepest Descent Method, and above all, in the derivation of the Conjugate Gradient Method.

Assume that A is a symmetric positive definite $n \times n$ matrix with eigenvalues

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n, \tag{42}$$

and a corresponding set of orthogonal and normalized eigenvectors $\{e_i\}_{i=1}^n$. Any vector $x \in \mathbb{R}^n$ may be expanded into a series of the form

$$x = \sum_{i=1}^{n} \alpha_i e_i, \tag{43}$$

and hence,

$$Ax = \sum_{i=1}^{n} \alpha_i A e_i = \sum_{i=1}^{n} \alpha_i \lambda_i e_i, \tag{44}$$

and

$$x' A x = \sum_{i=1}^{n} \alpha_i^2 \lambda_i. \tag{45}$$

The *A-norm* is defined

$$\|x\|_A \triangleq \left(x' A x\right)^{1/2}. \tag{46}$$

Since

$$\lambda_1 \|x\|^2 = \lambda_1 \sum_{i=1}^{n} \alpha_i^2 \leq x' A x \leq \lambda_n \sum_{i=1}^{n} \alpha_i^2 = \lambda_n \|x\|^2, \tag{47}$$

we observe that

$$\lambda_1^{1/2} \|x\| \leq \|x\|_A \leq \lambda_n^{1/2} \|x\|, \tag{48}$$

and the norms $\|x\| = \|x\|_2$ and $\|x\|_A$ are *equivalent* (as are any pair of norms in $\mathbb{R}^n$). The verifications of the norm properties are left for the reader:

$$\begin{array}{ll} \text{(i)} & x = 0 \Longleftrightarrow \|x\|_A = 0, \\ \text{(ii)} & \|\alpha x\|_A = |\alpha| \, \|x\|_A \,, \\ \text{(iii)} & \|x + y\|_A \leq \|x\|_A + \|y\|_A \,. \end{array} \tag{49}$$

In fact, $\mathbb{R}^n$ even becomes a *Hilbert space* in this setting if we define a corresponding inner product $\langle \ , \ \rangle_A$ as

$$\langle y, x \rangle_A \stackrel{\Delta}{=} y' A x. \tag{50}$$

It is customary to say that x and y are A-*conjugate* (or A-orthogonal) if $\langle y, x \rangle_A = 0$.

2.8 Basic Facts About Hilbert Space

A *Hilbert space* H is a linear space, and for our applications, consisting of vectors or functions. In case you have never heard about a Hilbert space, use what you know about $\mathbb{R}^n$.

It is first of all a *linear space,* so that if $x, y \in H$ and $\alpha, \beta \in \mathbb{R}$, also $\alpha x + \beta y$ has a meaning and is an element of H (We will not need complex spaces).

Furthermore, it has a scalar product $\langle \ , \ \rangle$ with its usual properties,

$$\begin{array}{ll} \text{(i)} & \langle x, y \rangle = \langle y, x \rangle \in \mathbb{R}, \\ \text{(ii)} & \langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle \,. \end{array} \tag{51}$$

We say that two elements x and y are *orthogonal* if $\langle x, y \rangle = 0$.

The scalar product defines a *norm*,

$$\|x\| = \langle x, x \rangle^{1/2}, \tag{52}$$

and makes H into a normed space (The final, and a little more subtle property which completes the definition of a Hilbert space, is that it is complete with respect to the norm, *i.e.* it is also what is called a *Banach space*).

A Hilbert space may be finite dimensional, like $\mathbb{R}^n$, or infinite dimensional, like $l^2(\mathbb{N})$ (This space consists of all infinitely dimensional vectors $x = \{x_i\}_{i=1}^{\infty}$, where $\sum_{i=1}^{\infty} |x_i|^2 < \infty$).

Important properties of any Hilbert space include

- **The Schwarz' Inequality:** $|\langle x, y \rangle| \leq \|x\| \, \|y\|$
- **The Pythagorean Formula:** *If* $\langle x, y \rangle = 0$*, then* $\|x + y\|^2 = \|x\|^2 + \|y\|^2$

However, the really big theorem in Hilbert spaces related to optimization theory is the *Projection Theorem*:

The Projection Theorem: *If* H_0 *is a closed subspace of* H *and* $x \in H$*, then* $\min_{y \in H_0} \|x - y\|$ *is obtained for a unique vector* $y_0 \in H_0$*, where*

- y_0 *is orthogonal to the error* $e = x - y_0$*, that is,* $\langle y_0, e \rangle = 0$*,*

- y_0 *is the best approximation to* x *in* H_0.

The theorem is often stated by saying that any vector in H may be written in a unique way as

$$x = y_0 + e, \tag{53}$$

where $y_0 \in H_0$, and y_0 and e are orthogonal.

The projection theorem is by far the most important *practical* result about Hilbert spaces. It forms the basis of everyday control theory and signal processing algorithms (*e.g.*, dynamic positioning, noise reduction and optimal filtering).

Our Hilbert spaces will have sets of orthogonal vectors of norm one, $\{e_i\}$, such that any $x \in H$ may be written as a *series,*

$$\begin{aligned} x &= \sum_i \alpha_i e_i, \\ \alpha_i &= \langle x, e_i \rangle \,, \; i = 1, 2, \cdots \end{aligned} \tag{54}$$

The set $\{e_i\}$ is called a *basis*. Note also that

$$\|x\|^2 = \sum_i \alpha_i^2. \tag{55}$$

If H_n is the subspace spanned by $e_1, \cdots, e_n$, that is

$$H_n = \operatorname{span}\{e_1, \cdots, e_n\} = \left\{ y \; ; \; y = \sum_{i=1}^{n} \beta_i e_i \; , \{\beta_i\} \in \mathbb{R}^n \right\}, \tag{56}$$

then the series of any $x \in H$, truncated at term n, is the best approximation to x in H_n,

$$\sum_{i=1}^{n} \alpha_i e_i = \arg \min_{y \in H_n} \|x - y\| \,. \tag{57}$$

If you ever need some Hilbert space theory, the above will probably cover it.

3 THE OPTIMIZATION SETTING

Since there is no need to repeat a result for maxima if we have proved it for minima, *we shall only consider minima in this course.* That is, we consider the problem

$$\min_{x \in \Omega} f(x) \,. \tag{58}$$

where Ω is called the *feasible domain.*

The definitions of *local*, *global*, and *strict* minima should be known to the readers, but we repeat them here for completeness.

- x^* is a *local minimum* if there is a neighborhood N of x^* such that $f(x^*) \leq f(x)$ for all $x \in N$.
- x^* is a *global minimum* if $f(x^*) \leq f(x)$ for all $x \in \Omega$.

- A local minimum x^* is *strict* (or an isolated) local minimum if there is an N such that $f(x^*) < f(x)$ for all $x \in N$, $x \neq x^*$.

It is convenient to use the notation

$$x^* = \arg \min_{x \in \Omega \subset \mathbb{R}^n} f(x) \tag{59}$$

for a solution x^* of (58). If there is only one minimum, which is then both global and strict, we say it is *unique*.

3.1 The Existence Theorem for Minima

As we saw for some trivial cases above, a minimum does not necessarily exist. So what about a criterion for existence? The following result is fundamental:

Assume that f is a continuous function defined on a closed and bounded set $\Omega \subset \mathbb{R}^n$. Then there exists $x^* \in \Omega$ such that

$$f(x^*) = \min_{x \in \Omega} f(x). \tag{60}$$

This theorem, which states that the minimum (and not only an infimum) really exists, is the most basic existence theorem for minima that we have. A parallel version exists for maxima.

Because of this result, we always prefer that the domain we are taking the minimum or maximum over is bounded and closed (Later in the text, when we consider a domain Ω, think of it as closed).

Let us look at the proof. We first establish that f is *bounded below* over Ω, that is, $\inf_{x \in \Omega} f(x)$ is finite. Assume the opposite. Then there are $x_n \in \Omega$ such that $f(x_n) < -n$, $n = 1, 2, 3 \cdots$. Hence $\lim_{n \to \infty} f(x_n) = -\infty$. At the same time, since Ω was bounded and closed, there are convergent subsequences, say $\lim_{k \to \infty} x_{n_k} = x_0 \in \Omega$. But $\lim_{k \to \infty} f(x_{n_k}) = -\infty \neq f(x_0)$; thus contradicting that f is continuous, and hence finite at x_0.

Since f is bounded below, we know that there is an $a \in \mathbb{R}$ such that

$$a = \inf_{x \in \Omega} f(x). \tag{61}$$

Since a is the largest number that is less or equal to $f(x)$ for all $x \in \Omega$, we also know that for any n, there must be an $x_n \in \Omega$ such that

$$f(x_n) < a + \frac{1}{n} \tag{62}$$

(think about it!).

We thus obtain, as above, a sequence $\{x_n\}$ that has a convergent subsequence $\{x_{n_k}\}_{k=1}^{\infty}$,

$$\lim_{k \to \infty} x_{n_k} = x_0 \in \Omega. \tag{63}$$

Since f is continuous, we also have

$$f(x_{n_k}) \underset{k \to \infty}{\longrightarrow} f(x_0). \tag{64}$$

On the other hand,

$$a \leq f(x_{n_k}) < a + \frac{1}{n_k}. \tag{65}$$

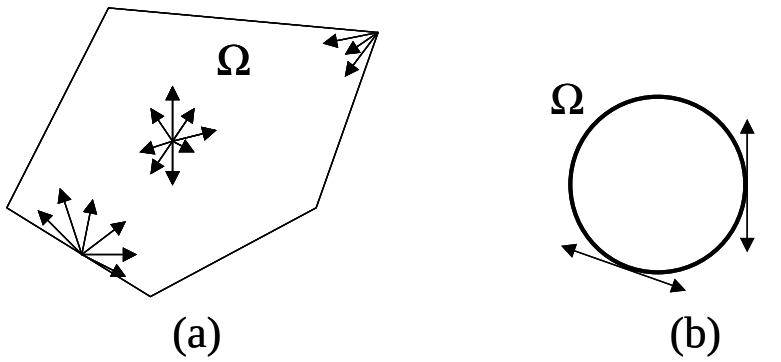

Figure 1: (a) Feasible directions in the interior and on the boundary of Ω. (b) Feasible directions when Ω (the circle itself, *not* the disc!) does not contain *any* line segment.

Hence

$$f(x_0) = a. \tag{66}$$

But this means that

$$f(x_0) = a = \inf_{x\in\Omega} f(x) = \min_{x\in\Omega} f(x), \tag{67}$$

which is exactly what we set out to prove!

3.2 The Directional Derivative and Feasible Directions

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ as above. The *directional derivative* of f at x in the direction $d \neq 0$ is defined as

$$\delta f(x, d) = \lim_{\varepsilon\to 0+} \frac{f(x+\varepsilon d) - f(x)}{\varepsilon}. \tag{68}$$

Assume that ∇f is continuous around x. Then, from Taylor's Formula,

$$\delta f(x, d) = \lim_{\varepsilon\to 0+} \frac{f(x+\varepsilon d) - f(x)}{\varepsilon} = \lim_{\varepsilon\to 0+} \frac{\nabla f(x+\xi\varepsilon d)\cdot(\varepsilon d)}{\varepsilon} = \nabla f(x)\, d, \tag{69}$$

which is the important formula for applications. The notation $\delta f(x, d)$ contains both a point x and a direction d out from x. Note that the definition does not require that $\|d\| = 1$ and that the answer depends on $\|d\|$. The directional derivative exists where ordinary derivatives don't, like for $f(x) = |x|$ at the origin (What *is* $\delta|x|(0, d)$?).

If we consider a domain $\Omega \subset \mathbb{R}^n$ and $x \in \Omega$, a *feasible direction* out from x is a vector d pointing into Ω, as illustrated in Fig. 1 (a). Note that the length of d is of no importance for the existence, since $x + \varepsilon d$ will be in Ω when ε is small enough. At an *interior point* (surrounded by a ball in $\mathbb{R}^n$ that is also in Ω), *all* directions will be feasible.

It will later be convenient also to consider *limiting feasible directions*, as shown in Fig. 1(b): A direction d is feasible if there exists a continuous curve $\gamma(t) \in \Omega$, where $\gamma(0) = x$, so that

$$\frac{d}{\|d\|} = \lim_{t\to 0+} \frac{\gamma(t) - x}{\|\gamma(t) - x\|}. \tag{70}$$

3.3 First and Second Order Conditions for Minima

First order conditions deal with *first* derivatives.

The following result is basic: *If* $\delta f(x,d) < 0$*, there is an* ε_0 *such that*

$$f(x+\varepsilon d) < f(x) \text{ for all } \varepsilon \in (0,\varepsilon_0). \tag{71}$$

In particular, *such a point can not be a minimum*! The proof is simple: Since $\delta f(x,d) < 0$, also

$$\frac{f(x+\varepsilon d) - f(x)}{\varepsilon} < 0 \text{ for all } \varepsilon \in (0,\varepsilon_0) \tag{72}$$

when ε_0 is small enough.

Corollary 1: *If* x^* *is a local minimum for* $f(x)$ *where directional derivatives exist, then* $\delta f(x^*,d) \geq 0$ *for all feasible directions.*

Otherwise, we can walk out from x^* in a direction d where $\delta f(x^*,d) < 0$.

Corollary 2: *If* x^* *is a local minimum for* $f(x)$*, and* ∇f *is continuous around* x^**, then* $\nabla f(x^*)\, d \geq 0$ *for all feasible directions.*

Yes, in that case, $\delta f(x^*,d)$ is simply equal to $\nabla f(x^*)\, d$.

Corollary 3 (N&W, Thm. 2.2): *If* x^* *is an interior local minimum for* $f(x)$ *where* ∇f *exists, then* $\nabla f(x^*) = 0$.

Assume that, *e.g.* $\frac{\partial f}{\partial x_j}(x^*) \neq 0$. Then one of the directional derivatives (in the x_j or $-x_j$-direction) are negative.

Corollaries 1–3 state necessary conditions; they will not guarantee that x^* is really a minimum (Think of $f(x) = x^3$ at $x = 0$).

The *second order* conditions bring in the Hessian, and the first result is Thm. 2.3 in N&W:

If x^* *is an interior local minimum and* $\nabla^2 f$ *is continuous around* x^**, then* $\nabla^2 f(x^*)$ *is positive semidefinite* $(\nabla^2 f(x^*) \geq 0)$.

The argument is again by contradiction: Assume that $d'\nabla^2 f(x^*)\, d = a < 0$ for some $d \neq 0$. Since $\nabla f(x^*)\, d = 0$ (Corollary 3), it follows from Taylor Formula that

$$\frac{f(x^*+\varepsilon d) - f(x^*)}{\varepsilon^2} = \frac{1}{2} d'\nabla^2 f(x^* + \xi\varepsilon d)\, d \underset{\varepsilon\to 0}{\longrightarrow} \frac{1}{2} a < 0. \tag{73}$$

Thus, there is an ε_0 such that

$$f(x^*+\varepsilon d) < f(x^*) \tag{74}$$

for all $\varepsilon \in (0,\varepsilon_0)$, and x^* can not be a minimum.

However, contrary to the first order conditions, the slightly stronger property that $\nabla^2 f(x^*)$ is positive definite, $\nabla^2 f(x^*) > 0$, and not only semidefinite, gives a sufficient condition for a strict local minimum:

Assume that $\nabla^2 f$ *is continuous around* x^**,* $\nabla f(x^*) = 0$*, and* $\nabla^2 f(x^*) > 0$*,then* x^* *is a strict local minimum.*

Since $\nabla^2 f$ is continuous and $\nabla^2 f(x^*) > 0$, it will even be positive definite in a neighborhood of x^*, say $\|x - x^*\| < \delta$ (The eigenvalues are continuous functions of the matrix elements, which in turn are continuous functions of x). Then, for $0 < \|p\| < \delta$,

$$\begin{aligned} f(x^*+p) - f(x^*) &= \nabla f(x^*)\cdot p + \frac{1}{2} p'\nabla^2 f(x^*+\xi p)\, p \\ &= 0 + \frac{1}{2} p'\nabla^2 f(x^*+\xi p)\, p > 0. \end{aligned} \tag{75}$$

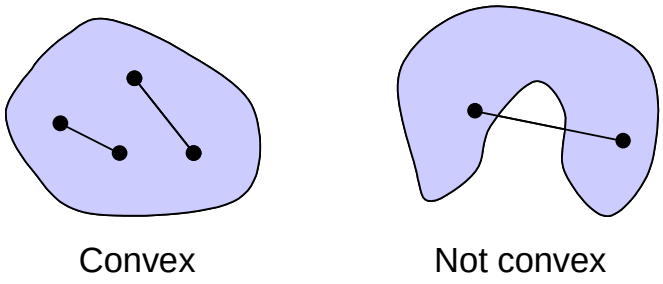

Figure 2: For a convex set, all straight line segments connecting two points are contained in the set.

Thus, x^* is a *strict* local minimum.

Simple counter-examples show that only $\nabla^2 f(x^*) \geq 0$ is not sufficient: Check $f(x,y) = x^2 - y^3$.

To sum up, the possible minima of $f(x)$ are at points x_0 where $\delta f(x_0, d) \geq 0$ for all feasible directions. In particular, if $\nabla f(x)$ exists and is continuous, possible candidates are

- interior points where $\nabla f(x) = 0$,
- points on the boundary where $\nabla f(x)\, d \geq 0$ for all feasible directions.

4 BASIC CONVEXITY

Convexity is one of the most important concepts in optimization. Although the results here are all quite simple and obvious, they are nevertheless very powerful.

4.1 Convex Sets

A *convex set* Ω in $\mathbb{R}^n$ is a set having the following property:

If $x, y \in \Omega$, *then* $\theta x + (1-\theta)\, y \in \Omega$ *for all* $\theta \in (0,1)$.

The concept can be generalized to all kind of sets (functions, matrices, stochastic variables, etc.), where a combination of the form $\theta x + (1-\theta)\, y$ makes sense.

It is convenient, but not of much practical use, to define the *empty set as convex.*

Note that a convex set has to be connected, and can not consist of isolated subsets.

Determine which of the following sets are convex:

- The space $\mathbb{R}^2$
- $\{(x,y) \in \mathbb{R}^2;\ x^2 + 2y^2 \leq 2\}$
- $\{(x,y) \in \mathbb{R}^2;\ x^2 - 2y^2 \leq 2\}$
- $\{x \in \mathbb{R}^n; Ax \geq b,\ b \in \mathbb{R}^m \text{ and } \ A \in \mathbb{R}^{m \times n}\}$

One basic theorem about convex sets is the following:

Theorem 1: *If* $\Omega_1, \cdots, \Omega_N \subset R^n$ *are convex sets, then*

$$\Omega_1 \cap \cdots \cap \Omega_n = \bigcap_{i=1}^{N} \Omega_i \tag{76}$$

is convex.

Proof: Choose two points $x, y \in \cap_{i=1}^{N} \Omega_i$. Then $\theta x + (1-\theta) y \in \Omega_i$ for $i = 1, \cdots, N$, that is, $\theta x + (1-\theta) y \in \cap_{i=1}^{N} \Omega_i$.

Thus, intersections of convex sets are convex!

4.2 Convex Functions

A real-valued function f is *convex on the convex set* Ω if for all $x_1, x_2 \in \Omega$,

$$f(\theta x_1 + (1-\theta) x_2) \leq \theta f(x_1) + (1-\theta) f(x_2), \ \theta \in (0,1). \tag{77}$$

Consider the graph of f in $\Omega \times \mathbb{R}$ and the *connecting line segment* from $(x_1, f(x_1))$ to $(x_2, f(x_2))$, consisting of the following points in $\mathbb{R}^{n+1}$:

$$\begin{gathered} \theta x_1 + (1-\theta) x_2, \\ \theta f(x_1) + (1-\theta) f(x_2) \ , \ \theta \in (0,1). \end{gathered}$$

The function is convex if all such line segments lie *on or above the graph.* Note that a linear function, say

$$f(x) = b'x + a, \tag{78}$$

is convex according to this definition, since in that particular case, Eqn. 77 will always be an equality.

When the inequality in Eqn. 77 is *strict*, that is, we have "$<$" instead of "$\leq$", then we say that the function is *strictly convex.* A linear function is convex, but *not* strictly convex.

Note that a convex function may not be continuous: Let $\Omega = [0, \infty)$ and f be the function

$$f(x) = \begin{cases} 1, & x = 0, \\ 0, & x > 0. \end{cases} \tag{79}$$

Show that f is convex. This example is a bit strange, and *we shall only consider continuous convex functions in the following.*

Proposition 1: *If* f *and* g *are convex, and* $\alpha, \beta \geq 0$, *then* $\alpha f + \beta g$ *is convex* (*on the common convex domain where both* f *and* g *are defined*).

Idea of proof: Show that $\alpha f + \beta g$ satisfies the definition in Eqn. 77.

What is the conclusion in Proposition 1 if at least one of the functions are strictly convex and α, $\beta > 0$? Can Proposition 1 be generalized?

Proposition 2: *If* f *is convex, then the set*

$$\mathcal{C} = \{x; f(x) \leq c\} \tag{80}$$

is convex.

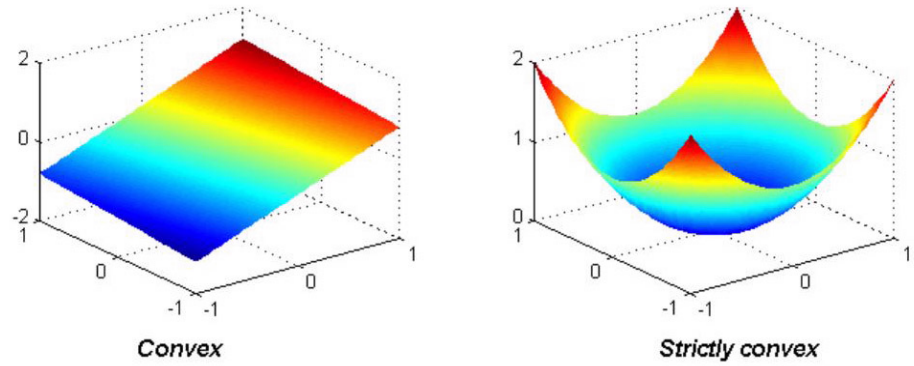

Figure 3: Simple examples of graphs of convex and strictly convex functions (should be used only as mental images!).

Proof: Assume that $x_1, x_2 \in \mathcal{C}$. Then

$$\begin{aligned} f(\theta x_1 + (1-\theta) x_2) &\leq \theta f(x_1) + (1-\theta) f(x_2) \\ &\leq \theta c + (1-\theta) c = c. \end{aligned} \tag{81}$$

This proposition has an important corollary for sets defined by several inequalities:

Corollary 1: *Assume that the functions* $f_1,\ f_2,\ \cdots, f_m$, *are convex. Then the set*

$$\Omega = \{x\ ;\ f_1(x) \leq c_1,\ f_2(x) \leq c_2, \cdots, f_m(x) \leq c_m\} \tag{82}$$

is convex.

Try to show that the maximum of a collection of convex functions, $g(x) = \max_i \{f_i(x)\}$, is also convex.

We recall that differentiable functions had *tangent planes*

$$T_{x_0}(x) = f(x_0) + \nabla f(x_0)(x - x_0), \tag{83}$$

and

$$f(x) - T_{x_0}(x) = o(\|x - x_0\|). \tag{84}$$

Proposition 3: *A differentiable function on the convex set* Ω *is convex if and only if its graph lies above its tangent planes.*

Proof: Let us start by assuming that f is convex and $x_0 \in \Omega$. Then

$$\begin{aligned} \nabla f(x_0)(x - x_0) = \delta f(x_0; x - x_0) &= \lim_{\varepsilon \to 0} \frac{f(x_0 + \varepsilon(x - x_0)) - f(x_0)}{\varepsilon} \\ &\leq \lim_{\varepsilon \to 0} \frac{[(1-\varepsilon) f(x_0) + \varepsilon f(x)] - f(x_0)}{\varepsilon} \\ &= f(x) - f(x_0). \end{aligned} \tag{85}$$

Thus,

$$f(x) \geq f(x_0) + \nabla f(x_0)(x - x_0) = T_{x_0}(x). \tag{86}$$

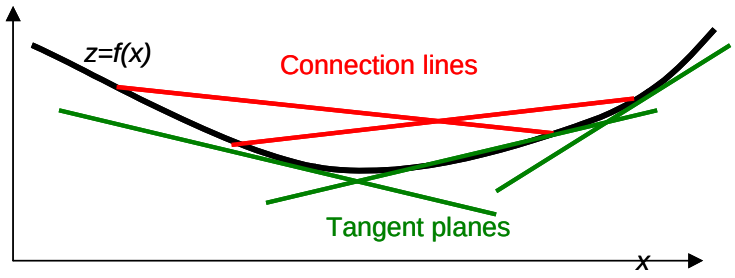

Figure 4: A useful mental image of a convex function: Connecting line segments above, and tangent planes below the graph!

For the opposite, assume that the graph of f lies above its tangent planes. Consider two arbitrary points x_1 and x_2 in Ω and a point x_θ on the line segment between them, $x_\theta = \theta x_1 + (1-\theta)\, x_2$. Then

$$\begin{aligned} f(x_1) &\geq f(x_\theta) + \nabla f(x_\theta)(x_1 - x_\theta), \\ f(x_2) &\geq f(x_\theta) + \nabla f(x_\theta)(x_2 - x_\theta). \end{aligned} \tag{87}$$

Multiply the first equation by θ and the last by $(1-\theta)$ and show that this implies that

$$\theta f(x_1) + (1-\theta) f(x_2) \geq f(x_\theta), \tag{88}$$

which is exactly the property that shows that f is convex.

The rule to remember is therefore:

The graph of a (differentiable) convex function lies above all its tangent planes and below the line segments between arbitrary points on the graph.

The following proposition assumes that the second order derivatives of f, that is, the *Hessian* $\nabla^2 f$, exists in Ω. We leave out the proof, which is not difficult:

Proposition 4: *A smooth function f defined on a convex set Ω is convex if and only if $\nabla^2 f$ is positive semi-definite in Ω. Moreover, f will be strictly convex if $\nabla^2 f$ is positive definite.*

The opposite of convex is *concave*. The definition should be obvious. Most functions occurring in practice are either convex and concave locally, but not for their whole domain of definition.

All results above have counterparts for concave functions.

4.3 The Main Theorem Connecting Convexity and Optimization

The results about minimization of convex functions defined on convex sets are simple, but very powerful:

Theorem 2: *Let f be a convex function defined on the convex set Ω. If f has minima in Ω, these are global minima and the set of minima,*

$$\Gamma = \left\{ y \; ; \; f(y) = \min_{x \in \Omega} f(x) \right\} \tag{89}$$

is convex.

Note 1: Let $\Omega = \mathbb{R}$ and $f(x) = e^x$. In this case the convex function $f(x)$ defined on the convex set $\mathbb{R}$ has *no* minima.

Note 2: Note that Γ itself is convex: All minima are collected at one place. There are no isolated local minima here and there!

Proof: Assume that x_0 is a minimum which is *not* a global minimum. We then know there is a $y \in \Omega$ where $f(y) < f(x_0)$. The line segment going from $(x_0, f(x_0))$ to $(y, f(y))$ is therefore sloping downward. However, because f is convex,

$$f(\theta x_0 + (1-\theta) y) \leq \theta f(x_0) + (1-\theta) f(y) < f(x_0), \tag{90}$$

for all $\theta \in [0, 1)$. Hence, x_0 can *not* be a local minimum, but a global minimum!

Assume that $f(x_0) = c$. Then

$$\begin{aligned} \Gamma &= \left\{ y \ ; \ f(y) = \min_{x \in \Omega} f(x) \right\} \\ &= \{ y \ ; \ f(y) = c \} \\ &= \{ y \ ; \ f(y) \leq c \}, \end{aligned} \tag{91}$$

is convex by Proposition 2.

Corollary 1: *Assume that f is a convex function on the convex set Ω and assume that the directional derivatives exist at x_0. Then x_0 belongs to the set of global minima of $f(x)$ in Ω if and only if $\delta f(x_0, d) \geq 0$ for all feasible directions.*

Proof: We already know that $\delta f(x_0, d)$ would be nonnegative if x_0 is a (global) minimum, so assume that x_0 is not a global minimum. Then $f(y) < f(x_0)$ for some $y \in \Omega$, and $d = y - x_0$ is a feasible direction (why?). But this implies that

$$\begin{aligned} \delta f(x_0, y - x_0) &= \lim_{\varepsilon \to 0+} \frac{f(x_0 + \varepsilon(y - x_0)) \ - f(x_0)}{\varepsilon} \\ &\leq \lim_{\varepsilon \to 0+} \frac{\varepsilon f(y) + (1-\varepsilon) f(x_0) - f(x_0)}{\varepsilon} = f(y) - f(x_0) < 0. \end{aligned} \tag{92}$$

Corollary 2: *Assume, that f is a differentiable convex function on the convex set Ω and that $\nabla f(x_0) = 0$. Then x_0 belongs to the set of global minima of $f(x)$ in Ω.*

Proof: Here $\delta f(x_0, d) = \nabla f(x_0)\, d = 0$ (which is larger or equal to 0!).

Note that if f is convex on the convex set Ω, and $\delta f(x, y - x)$ exists for all $x, y \in \Omega$, then inequality (92) may be written

$$f(y) \geq f(x) + \delta f(x, y - x).$$

Life is easy when the functions are convex, and one usually puts quite some effort either into formulating the problem so that it is convex, or tries to prove that for the problem at hand!

4.4 JENSEN'S INEQUALITY AND APPLICATIONS

Jensen's Inequality is a classic result in mathematical analysis where convexity plays an essential role. The inequality may be extended to a double-inequality which is equally simple to derive.

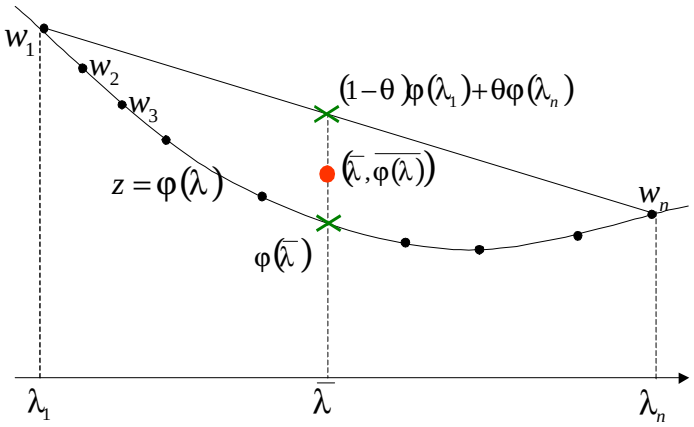

Figure 5: Think of the points as mass-particles and determine their center of gravity!

The inequality is among the few statements in mathematics where the proof is easier to remember than the result itself!

Let φ be a convex function, $\varphi : \mathbb{R} \to \mathbb{R}$. We first consider the discrete case where $\lambda_1 \leq \cdots \leq \lambda_n$, and $\{w_i\}_{i=1}^n$ are positive numbers. *Jensen's double-inequality* then goes as follows:

$$\varphi\left(\bar{\lambda}\right) \leq \overline{\varphi\left(\lambda\right)} \leq \left(1-\theta\right)\varphi\left(\lambda_1\right) + \theta\varphi\left(\lambda_n\right), \tag{93}$$

where

$$\begin{aligned} \bar{\lambda} &= \frac{\sum_{i=1}^n w_i\lambda_i}{\sum_{i=1}^n w_i}, \\ \overline{\varphi\left(\lambda\right)} &= \frac{\sum_{i=1}^n w_i\varphi\left(\lambda_i\right)}{\sum_{i=1}^n w_i}, \\ \theta &= \frac{\bar{\lambda}-\lambda_1}{\lambda_n-\lambda_1}. \end{aligned} \tag{94}$$

The name "Jensen's double inequality" is not very common, but suitable since there are two (non-trivial) inequalities involved.

The proof may be read *directly* out from Fig. 5, thinking in pure mechanical terms: The *center of gravity* for the n mass points at $\{\lambda_i, \varphi\left(\lambda_i\right)\}_{i=1}^n$ with weights $\{w_i\}_{i=1}^n$, is located at $\left(\bar{\lambda}, \overline{\varphi\left(\lambda\right)}\right)$. Because of the convexity of φ, the ordinate $\overline{\varphi\left(\lambda\right)}$ has to be somewhere between $\varphi\left(\bar{\lambda}\right)$ and $l\left(\bar{\lambda}\right)$, that is, the point corresponding to $\bar{\lambda}$ on the line segment joining $\left(\lambda_1, \varphi\left(\lambda_1\right)\right)$ and $\left(\lambda_n, \varphi\left(\lambda_n\right)\right)$.

That is all!

It is the *left* part of the double inequality that traditionally is called Jensen's Inequality.

Also try to write the inequality in the case when w is a positive *function* of λ, and derive the following inequality for a real stochastic variable:

$$\exp\left(\mathbb{E}X\right) \leq \mathbb{E}\left(\exp\left(X\right)\right) \tag{95}$$

(Hint: The mass density is now the probability density $w\left(\lambda\right)$ for the variable, and recall that $\mathbb{E}X = \int_{-\infty}^{\infty} \lambda w\left(\lambda\right) d\lambda$).

A lot of inequalities are derived from the left hand side of Jensen's double-inequality. However, the *Kantorovitch Inequality,* discussed next is an exception, since it is based on the *right* hand part of the inequality.

4.4.1 Application 1: Kantorovitch Inequality

The Kantorovitch Inequality goes as follows:

If A is a positive definite matrix with eigenvalues $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_n$, then

$$\frac{\|x\|_A^2 \|x\|_{A^{-1}}^2}{\|x\|^4} \leq \frac{1}{4}\frac{(\lambda_1+\lambda_n)^2}{\lambda_1\lambda_n}. \tag{96}$$

Since the inequality is invariant with respect to the norm of x, we shall assume that $x = \sum_{i=1}^n \alpha_i e_i$, and set $w_i = \alpha_i^2$ so that

$$\sum_{i=1}^n w_i = \|x\|^2 = 1. \tag{97}$$

Since we are on the positive real axis, the function $\varphi(\lambda) = \frac{1}{\lambda}$ is convex, and

$$\begin{aligned} \|x\|_A^2 &= x'Ax = \sum_{i=1}^n \lambda_i w_i = \bar{\lambda}, \\ \|x\|_{A^{-1}}^2 &= x'A^{-1}x = \sum_{i=1}^n \frac{1}{\lambda_i} w_i = \overline{\varphi(\lambda)}. \end{aligned} \tag{98}$$

Thus, by applying the RHS of Jensen's double-inequality,

$$\begin{aligned} \|x\|_A^2 \|x\|_{A^{-1}}^2 &= \bar{\lambda}\overline{\varphi(\lambda)} \\ &\leq \bar{\lambda}\left[(1-\theta)\frac{1}{\lambda_1} + \theta\frac{1}{\lambda_n}\right] \\ &= \bar{\lambda}\left[\left(1 - \frac{\bar{\lambda}-\lambda_1}{\lambda_n-\lambda_1}\right)\frac{1}{\lambda_1} + \frac{\bar{\lambda}-\lambda_1}{\lambda_n-\lambda_1}\frac{1}{\lambda_n}\right]. \end{aligned} \tag{99}$$

The right hand side is a second order polynomial in $\bar{\lambda}$ with a maximum value,

$$\frac{1}{4}\frac{(\lambda_1+\lambda_n)^2}{\lambda_1\lambda_n}, \tag{100}$$

attained for $\bar{\lambda} = (\lambda_1+\lambda_2)/2$ (Check it!). This proves the inequality.

Show that the inequality can not, in general, be improved by considering A equal to the 2×2 unit matrix.

4.4.2 Application 2: The Convergence of the Steepest Descent Method

It will in general be reasonable to assume that f has the form

$$f(x) = f(x^*) + \nabla f(x^*)(x-x^*) + \frac{1}{2}(x-x^*)'\nabla^2 f(x^*)(x-x^*) + \cdots \tag{101}$$

near a local minimum x^*. The convergence can therefore be studied in terms of the *Test problem*

$$\min_x f(x), \tag{102}$$

where

$$f(x) = b'x + \frac{1}{2}x'Ax,\ A > 0.$$

We know that the *gradient direction* $g = (\nabla f)'$ in this case is equal to $b + Ax$, and the Hessian $\nabla^2 f$ is equal to A. The problem has a unique solution for $b + Ax = 0$, that is, $x^* = -A^{-1}b$.

At a certain point x_k, the steepest descent is along the direction $-g_k = -(b + Ax_k)$. We therefore have to solve the one-dimensional sub-problem

$$\alpha_k = \arg\min_{\alpha} f(x_k - \alpha g_k).$$

It is easy to see that the minimum is attained at a point

$$x_{k+1} = x_k - \alpha_k g_k, \tag{103}$$

where the level curves (contours) of f are parallel to g_k, that is,

$$\nabla f(x_{k+1}) \cdot g_k = 0, \tag{104}$$

or $g'_{k+1}g_k = 0$. This gives us the equation

$$\begin{aligned}[b + A(x_k - \alpha_k g_k)]' g_k = \\ (g_k - \alpha_k A g_k)' g_k = 0,\end{aligned} \tag{105}$$

or

$$\alpha_k = \frac{g'_k g_k}{g'_k A g_k} = \frac{\|g_k\|}{\|g_k\|_A}. \tag{106}$$

The algorithm, which at the same time is an *iterative method* for the system $Ax = -b$, goes as follows:

Given x_1 **and** $g_1 = b + Ax_1$.

for $k = 1$ **until** ***convergence*** **do**

$\alpha_k = \frac{g'_k g_k}{g'_k (Ag_k)}$

$x_{k+1} = x_k - \alpha_k g_k$

$g_{k+1} = g_k - \alpha_k (Ag_k)$

end

In order to get an estimate of the error on step k, we note that

$$A^{-1}g_k = A^{-1}(b + Ax_k) = -x^* + x_k. \tag{107}$$

Hence,

$$\|x_k - x^*\|_A^2 = \left(A^{-1}g_k\right)' A \left(A^{-1}g_k\right) = \|g_k\|_{A^{-1}}^2, \tag{108}$$

and

$$\frac{\|x_{k+1} - x^*\|_A^2}{\|x_k - x^*\|_A^2} = \frac{\|g_{k+1}\|_{A^{-1}}^2}{\|g_k\|_{A^{-1}}^2}. \tag{109}$$

Let us look at $\|g_{k+1}\|^2_{A^{-1}}$ on the right hand side:

$$\begin{aligned}
\|g_{k+1}\|^2_{A^{-1}} &= g'_{k+1}A^{-1}\left(g_k - \alpha_k\left(Ag_k\right)\right) \\
&= g'_{k+1}A^{-1}g_k - \alpha_k g'_{k+1}g_k \\
&= g'_{k+1}A^{-1}g_k \\
&= \left(g_k - \alpha_k\left(Ag_k\right)\right)'A^{-1}g_k \\
&= g_kA^{-1}g_k - \frac{\left(g'_kg_k\right)^2}{g'_k\left(Ag_k\right)} \\
&= \|g_k\|^2_{A^{-1}} - \frac{\|g_k\|^4}{\|g_k\|^2_A}.
\end{aligned} \tag{110}$$

Thus,

$$\begin{aligned}
\frac{\|x_{k+1} - x^*\|^2_A}{\|x_k - x^*\|^2_A} &= \frac{\|g_{k+1}\|^2_{A^{-1}}}{\|g_k\|^2_{A^{-1}}} \\
&= 1 - \frac{\|g_k\|^4}{\|g_k\|^2_{A^{-1}}\|g_k\|^2_A} \\
&\leq 1 - \frac{4\lambda_1\lambda_n}{\left(\lambda_1 + \lambda_n\right)^2} \\
&= \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2 = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2,
\end{aligned} \tag{111}$$

where Kantorovitch Inequality was applied for the inequality in the middle. We recognize $\kappa = \lambda_n/\lambda_1$ as the *condition number* for the Hessian A.

If the condition number of the Hessian is large, the convergence of the steepest descent method may be very slow!

5 THE IMPLICIT FUNCTION THEOREM

The *Implicit Function Theorem* is a classical result in mathematical analysis. This means that all mathematicians know it, but can't really recall where they learnt it. The theorem may be stated in different ways, and it is not so simple to see the connection between the formulation in, *e.g.* N&W (Theorem A1, p. 585) and Luenberger (s. 462–3). In this short note we first state the theorem and try to explain why it is reasonable. Then we give a short proof based on *Taylor's Formula* and *Banach's Fixed-point Theorem*.

An *implicit function* is a function defined in terms of an equation, say

$$x^2 + y^2 - 1 = 0. \tag{112}$$

Given a general equation $h(x, y) = 0$, it is natural to ask whether it is possible to write this as $y = f(x)$. For Eqn. 112, it works well locally around a solution (x_0, y_0), except for the points $(-1, 0)$ and $(1, 0)$. In more difficult situations it may not be so obvious, and then the Implicit Function Theorem is valuable.

The Implicit Function Theorem tells us that if we have an equation $h(x, y) = 0$ and a solution (x_0, y_0), $h(x_0, y_0) = 0$, then there exists (if the conditions of the theorem are valid) a neighborhood

$\mathcal{N}$ around x_0 such that we may write

$$\begin{aligned} y &= f(x), \\ h\left(x, f\left(x\right)\right) &= 0, \text{ for all } x \in \mathcal{N}. \end{aligned} \tag{113}$$

The theorem guarantees that f exists, but does not solve the equation for us, and does not say in a simple way how large $\mathcal{N}$ is.

Consider the implicit function equation

$$x^2 - y^2 = 0 \tag{114}$$

to see that we only find solutions in a neighborhood of a known solution, and that we, in this particular case, will have problems at the origin.

We are going to present a somewhat simplified version of the theorem which, however, is general enough to show the essentials.

Let

$$h(x, y) = 0 \tag{115}$$

be an equation involving the m-dimensional vector y and the n-dimensional vector x. Assume that h is m-dimensional, such that there is hope that a solution with respect to y exists. We thus have m *nonlinear scalar equations* for the m unknown components of y.

Assume we know at least one solution (x_0, y_0) of Eqn. 115, and by moving the origin to (x_0, y_0), we may assume that this solution is the origin, $h(0, 0) = 0$. Let the matrix B be the *Jacobian* of h with respect to y at $(0, 0)$:

$$B = \frac{\partial h}{\partial y}(0) = \{\frac{\partial h_i}{\partial y_j}(0)\}. \tag{116}$$

The Implicit Function Theorem may then be stated as follows:

Assume that h is a differentiable function with continuous derivatives both in x and y. If the matrix B is non-singular, there is a neighborhood $\mathcal{N}$ around $x = 0$, where we can write $y = f(x)$ for a differentiable function f such that

$$h(x, f(x)) \equiv 0, \; x \in \mathcal{N}. \tag{117}$$

The theorem is not unreasonable: Consider the Taylor expansion of h around $(0, 0)$:

$$\begin{aligned} h(x, y) &= h(0, 0) + Ax + By + o\left(\|x\|, \|y\|\right) \\ &= Ax + By + o\left(\|x\|, \|y\|\right). \end{aligned} \tag{118}$$

The matrix A is the Jacobian of h with respect to x, and B is the matrix above. To the first order, we thus have to solve the equation

$$Ax + By = 0, \tag{119}$$

with respect to y, and if B is non-singular, this is simply

$$y = -B^{-1}Ax. \tag{120}$$

The full proof of the Implicit Function Theorem is technical, and it is perfectly OK to stop the reading here!

For the brave, we start by stating Taylor's Formula to first order for a *vector valued* function $y = g(x)$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$:

$$\begin{aligned} g(x) &= g(x_0) + \nabla g(x_\theta)(x - x_0), \\ x_\theta &= \theta x_0 + (I - \theta)x. \end{aligned} \tag{121}$$

Note that since g has m components, $\nabla g(x_\theta)$ is an $m \times n$ matrix (the Jacobian),

$$\nabla g(x_\theta) = \begin{bmatrix} \nabla g_1(x_{\theta_1}) \\ \nabla g_2(x_{\theta_2}) \\ \vdots \\ \nabla g_m(x_{\theta_m}) \end{bmatrix}, \tag{122}$$

and θ is a matrix, $\theta = \operatorname{diag}\{\theta_1, \cdots, \theta_m\}$. We shall assume that all gradients are continuous as well, and hence

$$\begin{aligned} g(x) &= g(x_0) + \nabla g(x_0)(x - x_0) + (\nabla g(x_\theta) - \nabla g(x_0))(x - x_0) \\ &= g(x_0) + \nabla g(x_0)(x - x_0) + a(x, x_0)(x - x_0) \end{aligned} \tag{123}$$

where $a(x, x_0) \underset{x \to x_0}{\longrightarrow} 0$.

Put

$$\Phi(x, y) = h(x, y) - Ax - By, \tag{124}$$

where, as above, $A = \partial h/\partial x(0)$ and $B = \partial h/\partial y(0)$. From Taylor's Formula,

$$\Phi(x, y) = a(x, y)x + b(x, y)y, \tag{125}$$

where both a and b tend to 0 when $x, y \to 0$. Thus, for any positive ε, there are neighborhoods

$$\begin{aligned} B(x, r_x) &= \{x; \|x\| < r_x\}, \\ B(y, r_y) &= \{y; \|y\| < r_y\}, \end{aligned} \tag{126}$$

such that

$$\begin{array}{ll} (i) & \|\Phi(x, y)\| \le \varepsilon \|x\| + \varepsilon \|y\|, \ x \in B(x, r_x), \ y \in B(y, r_y), \\ (ii) & \|\Phi(x_1, y_1) - \Phi(x_2, y_2)\| \le \varepsilon \|x_1 - x_2\| + \varepsilon \|y_1 - y_2\|, \\ & x_1, x_2 \in B(x, r_x), \ y_1, y_2 \in B(y, r_y). \end{array} \tag{127}$$

We now define the non-linear mapping $y \to T(y)$ as

$$y \to T(y) \stackrel{\Delta}{=} -B^{-1}Ax - B^{-1}\Phi(x, y), \tag{128}$$

and will show that this mapping is a *contraction* on $B(y, r_y)$ for all $x \in B(x, r_x)$. *This is the core of the proof.*

Choose ε so small that $\varepsilon + ||B^{-1}||\varepsilon < 1$. Then, find r_x and r_y such that (i) and (ii) hold, and also ensure that r_x is so small that

$$r_x < \frac{\varepsilon}{||B^{-1}A|| + ||B^{-1}||\varepsilon} r_y. \tag{129}$$

Let $y \in B(y, r_y)$ and $x \in B(x, r_x)$. Then,

$$\begin{aligned} ||T(y)|| &= || - B^{-1}Ax - B^{-1}\Phi(x, y)|| \\ &\le ||B^{-1}A||||x|| + ||B^{-1}|| \left(\varepsilon \|x\| + \varepsilon \|y\|\right) \\ &\le (||B^{-1}A|| + ||B^{-1}||\varepsilon) r_x + ||B^{-1}||\varepsilon r_y \\ &\le \varepsilon r_y + ||B^{-1}||\varepsilon r_y \le r_y. \end{aligned} \tag{130}$$

Thus $T\left(B(y,r_y)\right) \subset B(y,r_y)$. Moreover,

$$\begin{aligned} \|T(y_1) - T(y_2)\| &= ||B^{-1}\left(\Phi(x,y_1) - \Phi(x,y_2)\right)|| \\ &\leq \varepsilon\, ||B^{-1}||\, \|y_1 - y_2\| \\ &< (1-\varepsilon)\, \|y_1 - y_2\|\,. \end{aligned} \tag{131}$$

The *Banach Fixed-point Theorem* now guarantee solutions $y_0 \in B(y,r_y)$ fulfilling

$$y_0 = T\left(y_0\right) = -B^{-1}Ax - B^{-1}\Phi(x,y_0), \tag{132}$$

or

$$Ax + By_0 + \Phi(x,y_0) = h(x,y_0) = 0 \tag{133}$$

for all $x \in B(x,r_x)$!

This proves the existence of the function $x \to f(x) = y_0$ in the theorem for all $x \in B(x,r_x)$.

The continuity is simple:

$$y_2 - y_1 = -B^{-1}A(x_2 - x_1) - B^{-1}\left(\Phi(x_2,y_2) - \Phi(x_1,y_1)\right), \tag{134}$$

giving

$$\|y_2 - y_1\| \leq \left\|B^{-1}A\right\| \|x_2 - x_1\| + \left\|B^{-1}\right\| \left(\varepsilon \|x_2 - x_1\| + \varepsilon \|y_2 - y_1\|\right), \tag{135}$$

and hence

$$\|y_2 - y_1\| \leq \frac{\left\|B^{-1}A\right\| + \left\|B^{-1}\right\| \varepsilon}{1 - \|B^{-1}\|\,\varepsilon} \|x_2 - x_1\|\,.$$

Differentiability of f in the origin follows from the definition and (ii) above. Proof of the differentiability in other neighboring locations is simply to move the origin there and repeat the proof.

Luenberger gives a more complete and precise version of the theorem. The smoothness of f depends on the smoothness of h.

A final word: *Remember the theorem by recalling the equation*

$$Ax + By = 0, \tag{136}$$

where $A = \partial h/\partial x(0)$ *and* $B = \partial h/\partial y(0)$.

6 REFERENCES

Luenberger, D. G.: *Linear and Nonlinear Programming*, 2nd ed., Addison Westley, 1984.