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

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

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

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

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

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,

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

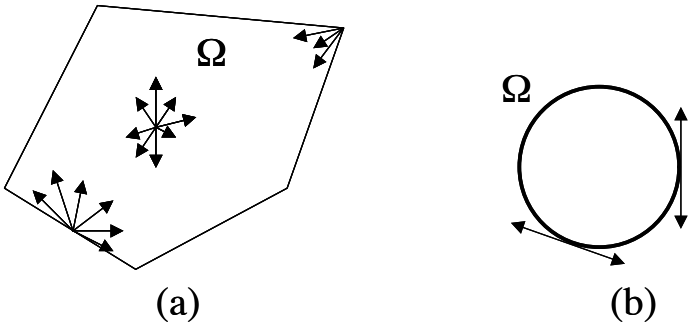


Figure 1: (a) Feasible directions in the interior and on the boundary of Ω. (b) Feasible directions when Ω (the circle itself, *not* the disc!) does not contain *any* line segment.

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 $0 < \|p\| < \delta$,

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

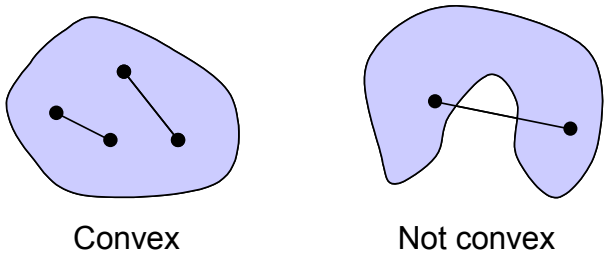


Figure 2: For a convex set, all straight line segments connecting two points are contained in the set.

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 $\mathbb{R}^2$
- $\{(x,y) \in \mathbb{R}^2;\ x^2 + 2y^2 \leq 2\}$
- $\{(x,y) \in \mathbb{R}^2;\ x^2 - 2y^2 \leq 2\}$
- $\{x \in \mathbb{R}^n; Ax \geq b,\, b \in \mathbb{R}^m$ 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.

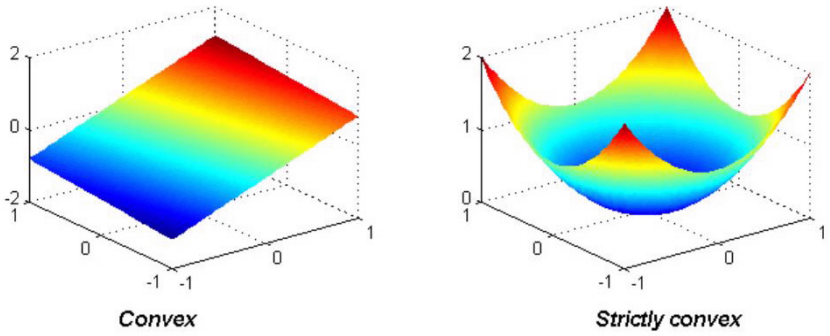


Figure 3: Simple examples of graphs of convex and strictly convex functions (should be used only as mental images!).

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,

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

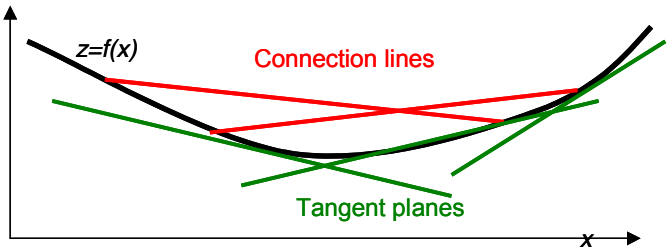


Figure 4: A useful mental image of a convex function: Connecting line segments above, and tangent planes below the graph!

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.

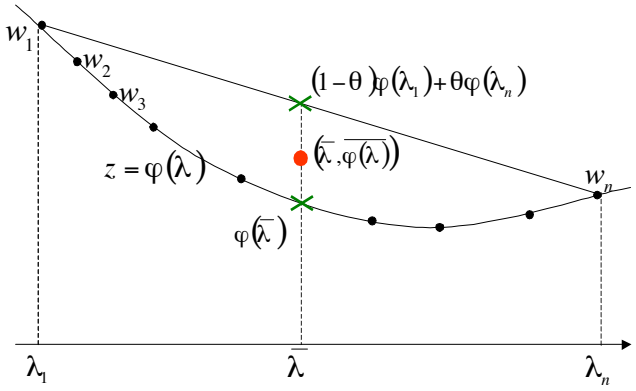


Figure 5: Think of the points as mass-particles and determine their center of gravity!

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:

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

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,

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

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

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

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.

Optimization Theory

Lower semi-continuity, compactness, and existence of solutions

Anton Evgrafov

Department of Mathematical Sciences, NTNU anton.evgrafov@math.ntnu.no

1 Reading

Chapter 1 in Nocedal and Wright, "Numerical optimization."

2 What is optimization?

Optimum (the neuter form of optimus) originates from the Latin, and translates to English as "the best." Therefore, "to optimize something (system/process/acitivity etc)" is normally understood as "to bring something to its best possible state." There are three important terms in this interpretation, which need further clarification:

- "To bring": a modeller needs to identify the parameters of the system (process, activity), which can be varied. These may be discrete or real-valued parameters, or even more general objects such as functions, geometric surfaces, or similar. In this course we will mostly deal with parameters assuming real values. It will be convenient to collect all such parameters in a vector $x \in \mathbb{R}^n$ of *optimization* or *decision variables*.
- "The best": in order to compare the states corresponding to various parameter values we need to introduce a total ordering on the set of parameters. Typically this is done by employing a real-valued *objective* or *cost function* $x \mapsto f(x) \in \mathbb{R}$ (sometimes $x \mapsto f(x) \in \mathbb{R} \cup \{+\infty\}$), with the convention that the "better" values of the parameters correspond to the smaller values of f. Thus to choose the best parameter values we need to find x corresponding to the smallest value of f.
- "Possible": not all combinations or values of the parameters are valid. Limited availability of physical resources (time, money, raw materials, labour, etc) or demand requirements may introduce upper and lower limits on the parameters. There might be technical/logical restrictions on the values or the relationships between various parameters. We will abstractly collect all the admissible values of the parameters for the problem under the consideration into a *feasible set* Ω.
  In most applications, the set Ω is defined as a solution set to a system of inequalities and equalities, which results from the list of all the restrictions on the parameters:

$$\Omega = \{\, x \in \mathbb{R}^n \mid g_i(x) \geq 0, i \in \mathcal{I},\ g_j(x) = 0, j \in \mathcal{E} \,\}, \tag{1}$$

and the function g_i, $i \in \mathcal{I}$ and g_j, $j \in \mathcal{E}$ will be referred to as the *inequality* and *equality constraints*[1]. Depending on whether the constraints are present, we classify the problem (2) as *constrained* or *unconstrained*.

To summarize, we will be concerned with solving problems of the type

$$\begin{aligned} &\text{minimize } f(x), \\ &\text{subject to } x \in \Omega, \end{aligned} \tag{2}$$

where Ω may be further described with inequality and equality constraints. One may generalize this framework in many ways; for example, instead of parameters in $\mathbb{R}^n$ we may consider other spaces with different topological and/or algebraic structures, such as for example spaces of matrices, functions, curves and surfaces, etc. Instead of inequality (equality) constraints of the type $g_i(x) \geq 0$ ($g_i(x) = 0$) one may instead demand $g(x) \in \mathcal{K}$, where $\mathcal{K}$ is a *cone*[2] (satisfying some technical requirements) in a suitable vector space. We will keep the problem formulation (2) as it provides plenty of the room for modelling, development of the theory, and efficient algorithms. Furthermore, this is the formulation considered in the textbook of the course.

Please note that people often use the expression *mathematical programming* interchangeably with optimization. The program refers to a "decision program" and not a computer program, as optimization/mathematical programming has a much longer history than computer programming.

3 What does it mean to solve the problem (2)?

One distinguishes between two most important types of solutions to (2).

Definition 1 (Global minimum). *A point $x^* \in \Omega$ is called the point of* global minimum*, if for every $x \in \Omega$ we have the inequality $f(x^*) \leq f(x)$.*

Geometrically, x^* is a point of global minimum if the graph $\{\, (x, f(x)) \mid x \in \Omega \,\}$ lies "above" the horizontal plane $\{\, (x, f(x^*)) \mid x \in \Omega, \alpha = f(x^*) \,\}$ *and* touches it at the point $(x^*, f(x^*))$ (but possibly at other points, too).

Points of global minimum may not exist even when the function is bounded from below:

Example 1. Consider a positive function $f(x) = \exp(-x^2)$. This function approaches zero arbitrarily close: for every $\epsilon > 0$ it suffices to take $|x| > [\log(\epsilon^{-1})]^{1/2}$ to get $0 < f(x) < \epsilon$. Therefore, the global minimum, if existed, must satisfy the inequality $f(x^*) < \epsilon$, for any $\epsilon > 0$. However, there is no $x^* \in \mathbb{R}$ such that $f(x^*) = 0$.

[1] In this course we will assume that both $\mathcal{I}$ and $\mathcal{E}$ are *finite* index sets.

[2] A cone C in a vector space is a set, which is invariant under multiplication with positive scalars; that is $\lambda C = C$, for every $\lambda > 0$. Examples include the zero cone $\{0\}$; the cone of vectors with non-negative components $\mathbb{R}^n_+$; or the cone of symmetric positive semi-definite matrices $\mathcal{S}^n_+$.

Unless we have information about the global behaviour of the function over the feasible set, global solutions, even when exist, are incredibly difficult to recognize. Indeed, assume that an oracle provides us with a globally optimal solution $x^* \in \Omega$, and our task is to verify her/his guess. Then, in accordance with the definition 1, we should compare $f(x^*)$ with the value $f(x)$, evaluated at *every other point* $x \in \Omega$, which is most often practically impossible. Instead, we will look for points, which can be characterized with the knowledge of the function only in the vicinity of a given point. For differentiable functions such an information will be available from the local Taylor series expansion of f and the constraints.

Definition 2 (Local minimum). *A point $x^* \in \Omega$ is called the point of* local minimum*, if it is a point of global minimum in the feasible set restricted to some neighbourhood of x^*. That is, if there is $\epsilon > 0$ such that for every $x \in \{\, y \in \Omega \mid \|y - x^*\| < \epsilon \,\}$ we have the inequality $f(x^*) \leq f(x)$. If the latter inequality is strict whenever $x \neq x^*$ in the vicinity of x^*, we say that x^* is the point of* strict *local minimum.*

4 Very briefly: "standard tricks" in optimization modelling

4.1 Auxiliary optimization variables

It is often convenient to introduce additional variables, which are not associated with the parameters of the system we are trying to model. One standard type of such auxiliary variables is a slack variable, which allows us to switch from inequality to equality constraints (and simple bounds):

$$g(x) \geq 0 \qquad \Longleftrightarrow \qquad g(x) - s = 0,\ s \geq 0.$$

Note that one may, in principle, replace s with s^2 and drop the restriction on the slack variable, but most often this is not such a good idea.

Another type of auxiliary variables appears when we move the objective function f into constraints instead:

$$\begin{cases} \min\limits_{x} f(x), \\ \text{s.t. } x \in \Omega, \end{cases} \qquad \Longleftrightarrow \qquad \begin{cases} \min\limits_{(x,z)} z, \\ \text{s.t. } (x,z) \in \{\, (\tilde{x},\tilde{z}) \in \Omega \times \mathbb{R} \mid \tilde{z} - f(\tilde{x}) \geq 0 \,\}, \end{cases}$$

This trick allows one to transform a problem of minimizing a piece-wise smooth objective function $f(x) = \max\{f_1(x), f_2(x), \ldots, f_k(x)\}$, where $f_1, \ldots, f_k$ are smooth functions, into a problem with smooth objective and constraints:

$$\begin{aligned} &\min_{(x,z)} z, \\ &\text{s.t. } x \in \Omega, \\ &\qquad z - f_1(x) \geq 0, \\ &\qquad \vdots \\ &\qquad z - f_k(x) \geq 0. \end{aligned}$$

Similar tricks allow one to deal with minimizing a variety of non-smooth functions such as $\|\cdot\|_1$ and $\|\cdot\|_\infty$-norms of vectors (provide the details utilizing the fact that $|x| = \max\{x, -x\}$).

4.2 Soft and hard constraints

In some applications, most notably financial, certain constraints may be violated at a cost. Such constraints are typically known as "soft" constrains (as opposed to the "hard" constraints, which must be satisfied no matter what). We can turn a "hard" inequality constraint $g(x) \geq 0$ into a soft constraint as follows. First, we introduce an artificial variable $s \geq 0$, which will measure how much the constraint g is violated, that is, we consider the constraints $g(x) + s \geq 0$, $s \geq 0$ instead. Second, we need to add the cost of violation, say $h(s)$[3], to the objective function. That is, instead of $f(x)$ we minimize $f(x) + h(s)$.

The idea of soft constraints is also utilized in penalty methods for constrained optimization, allowing one to transform the constrained problem into an unconstrained one, or the one with very simple constraints.

5 Very briefly: classification

- *Unconstrained optimization* refers to the situations when $\Omega = \mathbb{R}^n$ in (2); *constrained optimization* otherwise.
- *Linear programming/optimization*: the objective function and all the constraints are first order polynomials; *non-linear optimization* otherwise.
- *Quadratic programming*: the objective function is a second order polynomial and all the constraints are first order polynomials.
- *Convex programming/optimization*: the objective function and the feasible set Ω is convex; if the constraints are given explicitly, then all the inequality constraints are concave functions and the equality constraints are affine (first order polynomials).
- *Non-smooth/non-differentiable optimization*: normally refers to the situation, when the objective function $f(x)$ (or some of the constraints, though problems in this class are often unconstrained or involve only simple constraints, such as bounds on the variables) is not differentiable at least at some points. If all the functions involved in the problem are at least once differentiable, we deal with *differentiable* (sometimes *smooth*) *optimization*.
- *Semi-infinite programming*: the number of decision variables is finite, but the number of constraints is infinite.
- *Semi-definite programming*: optimization over spaces of symmetric matrices, restricted to be positive semi-definite.
- *Calculus of variations*: optimization over spaces of functions.

[3] A typical example of $h(s)$ is Ms, where $M > 0$ is the cost of violating the constraint $g(x) \geq 0$ "per unit of violation."

6 Basic existence of solutions results

One of the weakest forms of continuity under which one may expect the problem (2) to admit solutions is as follows.

Definition 3 (Lower semi-continuity). *Consider a function $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ on a set $\Omega \subseteq \mathbb{R}^n$. We say that the function f is lower semi-continuous (l.s.c.) on Ω, if the lower level set $\Omega_\alpha = \{\, x \in \Omega \mid f(x) \leq \alpha \,\}$ is relatively closed*[4] *in Ω for any $\alpha \in \mathbb{R}$.*

Example 2. Lower semi-continuous functions appear naturally in the context of so-called min-max (inf-sup) problems, where we try to find a minimum of a function, which is defined through a maximization problem. For example, let

$$f(x) = \sup_{t \in \mathbb{R}} \{1 - \exp(-x^2 t^2)\}.$$

Then

$$f(x) = \begin{cases} 0, & \text{if } x = 0, \\ 1, & \text{otherwise,} \end{cases}$$

which is clearly discontinuous at $x = 0$, despite the fact that each individual function $1 - \exp(-x^2 t^2)$ is continuous in (x, t). Nevertheless, the function f is still l.s.c. on $\mathbb{R}$, because for any $\alpha \in \mathbb{R}$ we have

$$\Omega_\alpha = \begin{cases} \emptyset, & \alpha < 0, \\ \{0\}, & 0 \leq \alpha < 1, \\ \mathbb{R}, & 1 \leq \alpha, \end{cases}$$

all of which are closed sets in $\mathbb{R}$.

We now define the concept of a minimizing sequence. One can establish the existence of solutions to the problem (2) without appealing to minimizing sequences (as indeed we will). Nevertheless, minimizing sequences constitute an important and often utilized concept in their own right.

Definition 4 (Minimizing sequence). *Assume that $\Omega \neq \emptyset$ and let $f^* = \inf_{x \in \Omega} f(x)$.*[5] *By the definition of the infimum, there is a sequence of numbers $\{f_k\}_{k=1}^\infty$ in the set $\{\, f(x) \mid x \in \Omega \,\}$, such that $\lim_{k \to \infty} f_k = f^*$. By the definition of f_k, there is $x_k \in \Omega$ such that $f_k = f(x_k)$. We say that the sequence $\{x_k\}_{k=1}^\infty$ is a* minimizing sequence *for the problem* (2).

[4] Relative closedness in this notation means that if a sequence of points $\{x_k\}_{k=1}^\infty$ in Ω_α converges to a limit $x^* \in \Omega$, then $x^* \in \Omega_\alpha$. In other words, the set Ω_α contains all its limit points, which are also in Ω. Therefore closedness, relative to the whole space $\mathbb{R}^n$, coincides with the regular definition of closedness.

[5] We adopt the convention that every subset of the real line has an infimum, or the greatest lower bound, by letting the infimum of the set unbounded from below be equal to $-\infty$ and the infimum of the empty set be equal $+\infty$. We apply a similar convention to sup.

We know that the function is continuous iff the convergence of sequences is preserved by the function, that is, $x_k \to x^* \implies f(x_k) \to f(x^*)$. One can provide a similar characterization of lower semi-continuity as well, which will be useful for our purposes when applied to minimizing sequences.

Proposition 1. *A function $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ is l.s.c. on Ω iff for every sequence $\{x_k\}_{k=1}^{\infty}$ in Ω converging to $x^* \in \Omega$ it holds that $f(x^*) \leq \liminf_{k\to\infty} f(x_k)$*[6]

Proof. Suppose that f is l.s.c. on Ω. For the sake of contradiction assume that for some convergent sequence $\{x_k\}_{k=1}^{\infty}$ in Ω it holds that $\liminf_{k\to\infty} f(x_k) < f(x^*)$, where $\Omega \ni x^* = \lim_{k\to\infty} x_k$. In particular, we get that $\liminf_{k\to\infty} f(x_k) < +\infty$ and as a result there is $\alpha \in \mathbb{R}$, such that $\liminf_{k\to\infty} f(x_k) < \alpha < f(x^*)$. We will demonstrate that there there is a subsequence of x_k, which belongs to Ω_α, thus showing that Ω_α is not relatively closed (it does not contain one of its limit points, namely x^*); this is our desired contradiction with the lower semi-continuity of f. We proceed with yet another proof by contradiction. Indeed, suppose that Ω_α contains only finitely many elements x_k, that is, there is an index N such that for all $k \geq N$ it holds that $x_k \notin \Omega_\alpha$, or equivalently, $f(x_k) > \alpha$. Then $\inf_{k\geq n} f(x_k) \geq \alpha$ for every $n \geq N$ and as a result also $\lim_{n\to\infty} \inf_{k\geq n} f(x_k) \geq \alpha$. This contradicts with our choice of α: $\liminf_{k\to\infty} f(x_k) < \alpha < f(x^*)$.

We now prove the implication in the opposite direction. Suppose that f is *not* l.s.c. on Ω. Then there is $\alpha \in \mathbb{R}$ such that Ω_α is not relatively closed in Ω. That is, there is a sequence $\{x_k\}_{k=1}^{\infty}$ in Ω_α, with a limit $x^* \in \Omega \setminus \Omega_\alpha$. Clearly, we then have $\liminf_{k\to\infty} f(x_k) \leq \liminf_{k\to\infty} \alpha = \alpha < f(x^*)$, which concludes the proof. □

We now establish existence of solutions to the problem (2) without requiring any algebraic/geometric properties of the problem (such as convexity). The result is normally attributed to Weierstrass.

Theorem 1 (Existence of solutions). *Let Ω be a non-empty compact*[7] *set in $\mathbb{R}^n$ (or any other metric space, for that matter) and $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ be lower semi-continuous on Ω. Then the problem* (2) *admits at least one global minimum.*

Proof. Let $\{x_k\}_{k=1}^{\infty}$ be a minimizing sequence for the problem (2) (see Definition 4). Owing to the compactness of Ω it holds that for $\{x_k\}_{k=1}^{\infty}$ contains a subsequence, say $\{x'_k\}_{k=1}^{\infty}$, converging to some point $x^* \in \Omega$. Utilizing the definition of the infimum, Proposition 1, the fact that $\{f(x'_k)\}_{k=1}^{\infty}$ is a subsequence of the converging sequence $\{f(x_k)\}_{k=1}^{\infty}$, and the definition of the minimizing sequence we obtain the following string of inequalities:

$$\inf_{x\in\Omega} f(x) \leq f(x^*) \leq \liminf_{k\to\infty} f(x'_k) = \lim_{k\to\infty} f(x'_k) = \lim_{k\to\infty} f(x_k) = \inf_{x\in\Omega} f(x),$$

[6] Recall that for any sequence $\{\alpha_k\}_{k=1}^{\infty}$ of real numbers, $\liminf_{k\to\infty} \alpha_k = \lim_{n\to\infty} \inf_{k\geq n} \alpha_k$. lim inf (finite or infinite) exists for an arbitrary sequence, as the sequence $\beta_n = \inf_{k\geq n} \alpha_k$ is monotonically non-decreasing.

[7] Recall that a subset of $\mathbb{R}^n$ is compact iff it is closed and bounded (Heine–Borel Theorem). Further, from any sequence in a compact subset of a metric space we can extract a converging subsequence, which is utilized in the proof of Theorem 1.

which implies that $x^* \in \Omega$ is the point of global minimum for (2). □

We now give an alternative proof of Theorem 1, which does not appeal to the convergence of sequences.

Proof (Alternative proof of Theorem 1). Let $f^* = \inf_{x\in\Omega} f(x)$, and let Ω_α denote the lower-level set of f for any real number α. Every Ω_α is a closed set owing to the lower semi-continuity of f, which is non-empty for any $\alpha > f^*$ by the definition of inf. Therefore, for any finitely many numbers $\alpha_1 > f^*$, $\alpha_2 > f^*$,...,$\alpha_N > f^*$ it holds that

$$\cap_{i=1}^{N} \Omega_{\alpha_i} = \Omega_{\min_{i=1,\ldots,N} \alpha_i} \neq \emptyset,$$

since $\min_{i=1,\ldots,N} \alpha_i > f^*$. This is precisely the condition for the family of closed sets $\{ \Omega_\alpha \mid \alpha > f^* \}$ to have a *finite intersection property*. Owing to the compactness of Ω it holds that $\cap_{\alpha>f^*} \Omega_\alpha \neq \emptyset$. By construction, every point in $\cap_{\alpha>f^*} \Omega_\alpha = \Omega_{f^*}$ is a point of global minimum for (2); in particular, $f^* > -\infty$ as f cannot assume the value $-\infty$. □

When establishing the existence of solutions to optimization problems, one can trade the compactness of Ω for the growth of f at infinity, which guarantees that the minimizing sequences stay bounded.

Definition 5. *A function $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ is called* coercive *if $f(x) \to +\infty$ whenever $\|x\| \to +\infty$.*

Theorem 2. *Let Ω be a non-empty closed set in $\mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$ be a coercive lower semi-continuous function on Ω. Then the problem* (2) *admits at least one global minimum.*

Exercises

1. Prove Theorem 2.
2. Show that all the assumptions in Theorems 1 and 2 are essential. Indeed, Example 1 shows that either compactness or coercivity are essential for the existence of solutions. Demonstrate that the lower semi-continuity is also needed by constructing an instance of the problem (2) with a not lower semi-continuous objective function f, which does not attain its infimum on some non-empty compact feasible set Ω.
3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic polynomial $f(x) = 0.5x^T Gx + x^T d$, where $G \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, and $d \in \mathbb{R}^n$ is arbitrary. Show that f is coercive on $\mathbb{R}^n$. (Hint: expand x in terms of the eigenvectors of G to estimate $x^T Gx$ from below.)
4. Provide details on how one can transform the non-smooth optimization problem of minimizing the 1- or ∞-norm of a vector to a smooth minimization problem by introducing auxiliary variables and additional constraints.

5. * Example 2 is not incidental! Indeed, consider a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined as $f(x) = \sup\{ f_s(x) \mid s \in S \}$, where each function $f_s : \mathbb{R}^n \to \mathbb{R}$ is l.s.c., and S is an arbitrary index set (possibly uncountable). Show that f is l.s.c. on $\mathbb{R}^n$.
6. Show that the function $f : \Omega \to \mathbb{R}$ is l.s.c. on Ω if and only if it satisfies the following "ϵ/δ" definition of lower semi-continuity: for any x^* in Ω and any $\epsilon > 0$ there is $\delta > 0$ such that for any $x \in \Omega$, $\|x - x^*\| < \delta \implies f(x^*) < f(x) + \epsilon$. (You may use the equivalent characterization of lower semi-continuity given by Proposition 1).

Optimization Theory

Convergence of descent methods with backtracking (Armijo) linesearch

Anton Evgrafov

Department of Mathematical Sciences, NTNU anton.evgrafov@math.ntnu.no

Read: Section 3.1 in Nocedal and Wright, "Numerical optimization," in particular Algorithm 3.1, p. 37.

Consider the following iteration:

$$x_{k+1} = x_k + \alpha_k p_k, \qquad k = 0, 1, 2, \dots$$

where $B_k = B_k^T$,

$$B_k p_k = -\nabla f(x_k),$$

and α_k is selected using the backtracking (Armijo) linesearch with parameters $c, \rho \in (0, 1)$.

Theorem 1. *Suppose that*

1. $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable;*
2. *the set* $S := \{x \in \mathbb{R}^n \mid f(x) \le f(x_0)\}$ *is bounded;*
3. *the matrices* B_k *are uniformly positive definite and bounded, that is* $\exists m > 0, M > 0 : m \le \lambda_{\min}(B_k) \le \lambda_{\max}(B_k) \le M$*, where* $\lambda_{\min}$ *and* $\lambda_{\max}$ *are the smallest and the largest eigenvalues of* B_k*.*

Then the sequence $\{x_k\}$ *is bounded, and every its limit point* $\hat{x}$ *is a stationary point for* f*.*

Proof. Owing to the sufficient descent condition in the linesearch procedure the sequence $f(x_k)$, $k = 0, 1, 2, \dots$ is non-increasing; thus $x_k \in S$ for all k; in particular it is bounded and therefore has at least one limit point. The set S is closed because f is continuous, and thus is compact owing to the assumption 2 and Heine–Borel theorem. Therefore, the function f attains its minimum value on S (Weierstrass theorem) and thus is bounded from below on S. As a result, the non-increasing sequence $f(x_k)$ has a finite limit, and furthermore $\lim_{k\to\infty}[f(x_{k+1}) - f(x_k)] = 0$.

Owing to the sufficient descent condition it holds that

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\le c\alpha_k \nabla f(x_k)^T p_k = -c\alpha_k \nabla f(x_k)^T B_k^{-1} \nabla f(x_k) \\ &\le -c\alpha_k \lambda_{\max}(B_k^{-1}) \|\nabla f(x_k)\|^2 \\ &\le -cM^{-1}\alpha_k \|\nabla f(x_k)\|^2 \le 0. \end{aligned} \tag{1}$$

The sequence on the left converges to 0, meaning that the sequence on the right must also converge to zero. We will show that this implies that $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$.

Suppose that this is not true; then, for some subsequence of indices k' and some $\epsilon > 0$ we must have that $\|\nabla f(x_{k'})\| \geq \epsilon$. From (1) it then follows that $\lim_{k' \to \infty} \alpha_{k'} = 0$. In particular, it means that the step $\alpha_{k'}/\rho$ was not acceptable to the linesearch procedure for all large k', that is

$$f(x_{k'} + \alpha_{k'} \rho^{-1} p_{k'}) > f(x_{k'}) + c\alpha_{k'} \rho^{-1} \nabla f(x_{k'})^T p_{k'}. \tag{2}$$

The sequence of directions $p_k = -B_k^{-1} \nabla f(x_k)$ is bounded. Indeed, by our assumption 3 the norms $\|B_k^{-1}\| = \lambda_{\min}^{-1}(B_k) \leq m^{-1}$. Furthermore, the continuous function $x \mapsto \|\nabla f(x)\|$ attains its maximum over the compact set S, and thus $\|\nabla f(x_k)\|$ is bounded by this maximum value, for all k. As a result, we may assume that for some subsequence of k', say k'', it holds that $\lim_{k'' \to \infty} x_{k''} = \hat{x}$ and $\lim_{k'' \to \infty} p_{k''} = \hat{p}$. Rearranging the terms in (2) we get

$$\begin{aligned} 0 &\leq \lim_{k'' \to \infty} \frac{f(x_{k''} + \alpha_{k''} \rho^{-1} p_{k''}) - f(x_{k''})}{\alpha_{k''} \rho^{-1}} - c \nabla f(x_{k''})^T p_{k''} \\ &= (1 - c) \nabla f(\hat{x})^T \hat{p}, \end{aligned} \tag{3}$$

and therefore $\nabla f(\hat{x})^T \hat{p} \geq 0$ as $0 < c < 1$. On the other hand,

$$\begin{aligned} \nabla f(\hat{x})^T \hat{p} &= \lim_{k'' \to \infty} \nabla f(x_{k''})^T p_{k''} = - \lim_{k'' \to \infty} \nabla f(x_{k''})^T B_{k''}^{-1} \nabla f(x_{k''}) \\ &\leq -M^{-1} \epsilon^2 < 0. \end{aligned} \tag{4}$$

However, equations (3) and (4) contradict each other. This must mean that our assumption that $\|\nabla f(x_{k'})\| \geq \epsilon$ over some subsequence k' is wrong and in fact

$$\lim_{k \to \infty} \|\nabla f(x_k)\| = 0. \tag{5}$$

Finally, let $\hat{x}$ be an arbitrary limit point of $\{x_k\}$, that is, $\hat{x} = \lim_{k''' \to \infty} x_{k'''}$ for some subsequence k'''. Owing to the continuity of the function $x \mapsto \|\nabla f(x)\|$ (assumption 1) and (5) it holds that $\|\nabla f(\hat{x})\| = 0$, as we claimed. □

Exercise: Using this theorem, show that the steepest descent algorithm with Armijo (backtracking) linesearch converges to the minimim of Rosenbrock function from any starting point.

Introduction to optimality conditions:
Optimality conditions for optimization over convex sets*

Anton Evgrafov

Department of Mathematical Sciences, NTNU anton.evgrafov@math.ntnu.no

1 Special case: optimization over convex sets

Consider a constrained optimization problem

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & f(x), \\ \text{subject to} \quad & x \in \Omega, \end{aligned} \tag{1}$$

where the *feasible set* $\Omega \subset \mathbb{R}^n$ is assumed to be non-empty and closed, and the *objective function* $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on Ω. Additionally, all the results in this note apart from Proposition 1, case 1, and Proposition 3, case 1, apply when Ω is *convex*.

Example 1. We have already encountered problems of the type (1) with a closed and convex Ω in this course:

- the trust-region subproblem

$$\begin{aligned} \underset{p\in\mathbb{R}^n}{\text{minimize}} \quad & m_k(p) = f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p, \\ \text{subject to} \quad & \|p\| \le \Delta_k, \end{aligned}$$

  is of this class since the ball $\Omega = \{\, p \in \mathbb{R}^n \mid \|p\| \le \Delta_k \,\}$ is convex and closed;
- the problem leading to DFP/BFGS-type quasi-Newton update formulas:

$$\begin{aligned} \underset{B\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad & f(B) = \frac{1}{2}\|B - B_k\|^2, \\ \text{subject to} \quad & B - B^T = 0, \\ & Bs_k = y_k, \end{aligned}$$

  where $B_k \in \mathbb{R}^{n\times n}$, $s_k, y_k \in \mathbb{R}^n$ are given, and $\|\cdot\|$ is typically a scaled Frobenius norm of the matrix. (Convince yourself that the set $\{\, B \in \mathbb{R}^{n\times n} \mid B^T = B, Bs_k = y_k \,\}$ is convex and closed. More generally, check that the solution set of any system of linear equations and non-strict inequalities is convex and closed.)

* Section 1 of this note is based on Section 4.4 in "Introduction to continuous optimization" by N. Andréasson, AE, M. Patriksson, E. Gustavsson, M. Önnheim: Studentlitteratur (2013), 2nd ed.

In the theory of unconstrained optimization we have started with the first order necessary optimality conditions, which could be succinctly stated as $\nabla f(x) = 0$ (see Theorem 2.2 in N&W). This statement expresses the fact that at any point of local minimum $x \in \mathbb{R}^n$, there should be no direction $p \in \mathbb{R}^n$, along which we could move and decrease the function (*direction of descent*). In the absence of constraints we are allowed to move along any direction, and thus we could always take the *steepest descent direction* $p = -\nabla f(x)$, unless the gradient vanishes at x.

The situation is drastically different in the presence of constraints. Indeed, if we take x to be on the boundary of Ω, that is $x \in \Omega \setminus \text{interior}(\Omega)$, then it may happen that we cannot take even the smallest step along some directions $p \in \mathbb{R}^n$ without leaving Ω, that is, $\forall$"small"$\delta > 0 : x + \delta p \notin \Omega$. Therefore these directions may still be directions of descent for f and yet not prevent $x \in \Omega$ from being a point of local minimum for f over Ω.

Definition 1 (Feasible direction). *Direction $p \in \mathbb{R}^n$ is a* feasible direction *at $x \in \Omega \subset \mathbb{R}^n$ if small steps along p do not take us outside of Ω.*

Formally, $p \in \mathbb{R}^n$ is a feasible direction at $x \in \Omega \subset \mathbb{R}^n$ if there is $\hat{\delta} > 0$ such that for all $0 < \delta < \hat{\delta}$ the inclusion $x + \delta p \in \Omega$ holds.

Exercise 1 (Feasible directions for linear constraints). Suppose that all equality and inequality constraints are *linear*, that is, $c_i(x) = a_i^T x + b_i$, $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, $i \in \mathcal{I} \cup \mathcal{J}$, see eq. (12.1) in N&W. Show that the set of feasible directions $p \in \mathbb{R}^n$ at $x \in \Omega$ is

$$\{\, p \in \mathbb{R}^n \mid a_i^T p = 0, i \in \mathcal{E}, \quad a_i^T p \geq 0, i \in \mathcal{A}(x) \cap \mathcal{I} \,\},$$

where $\mathcal{A}(x)$ is the set of *active* or *binding* constraints at x, see Definition 12.1 in N&W.

With this definition we are ready to prove the following version of the first order optimality conditions for constrained problems.

Proposition 1 (First order necessary optimality conditions). *Consider a set $\Omega \subset \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that $x^* \in \Omega$ is a point of local minimum for f over Ω, and further assume that f is continuously differentiable around x^*.*

1. *Then for every feasible direction $p \in \mathbb{R}^n$ at x^* it holds that $\nabla f^T(x^*)p \geq 0$.*
2. *If, additionally, Ω is convex then*

$$\nabla f(x^*)^T(x - x^*) \geq 0, \quad \forall x \in \Omega. \tag{2}$$

Proof. 1. The proof is along the lines of Theorem 2.2 in N&W. For the sake of contradiction, let p be a feasible direction at x^* but $\nabla f(x^*)^T p < 0$. Owing to the continuity of $\nabla f(\cdot)$ around x^*, the same inequality holds at all nearby points, that is, $\nabla f(x^* + \delta p)^T p < 0$, for all $0 \leq \delta < \bar{\delta}$. Furthermore, the point $x_\delta = x^* + \delta p$ is feasible, for all $0 < \delta < \hat{\delta}$, owing to Definition 1.

Utilizing the first order Taylor series expansion we conclude that for every $0 < \delta < \min\{\bar{\delta}, \hat{\delta}\}$ we have the inclusion $x_\delta \in \Omega$ and the strict inequality

$$f(x_\delta) = f(x^*) + \nabla f(x^* + t_\delta \delta p)^T p < f(x^*),$$

where $0 < t_\delta < 1$. Since $x_\delta \to x^*$ as $\delta \to 0$ the point x^* cannot be a point of local minimum for f over Ω.

2. It suffices to show that $p_x = x - x^*$, $x \in \Omega$, is a feasible direction at x^*. Indeed, $\forall 0 < \lambda < 0$ we have

$$x^* + \lambda p_x = x^* + \lambda(x - x^*) = \lambda x + (1 - \lambda)x^* \in \Omega,$$

owing to the convexity of Ω.

□

As before, we will refer to the points satisfying the first order necessary conditions as the stationary points.

For a given point $\bar{x} \in \Omega$ let us consider the following linearization the original problem (1):

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad & m(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}), \\ \text{subject to} \quad & x \in \Omega. \end{aligned} \tag{3}$$

One can formulate the equivalent characterization of local optimality over convex sets in terms of the linearized problem (3).

Corollary 1. *Under the assumptions of Proposition 1, case 2, the locally optimal solution x^* to* (1) *must be a globally optimal solution to the linearized problem* (3)*, where the linearization point is $\bar{x} = x^*$.*

Proof. For any $x \in \Omega$ we have

$$m(x) = f(x^*) + \underbrace{\nabla f(x^*)^T (x - x^*)}_{\geq 0, \text{ see (2)}} \geq f(x^*) = m(x^*).$$

□

In fact the problem (3) has been utilized for algorithmic purposes. Indeed, since the linearization point $\bar{x}$ is feasible, it holds that any solution $\bar{x}^*$ to this convex optimization problem, whenever exists (for example when Ω is bounded), satisfies $m(\bar{x}^*) \leq m(\bar{x}) = f(\bar{x})$. Thus if we were able to find a solution to (3) such that $m(\bar{x}^*) < f(\bar{x})$ then the direction $p = \bar{x}^* - \bar{x}$ satisfies $\nabla f(\bar{x})^T p < 0$ and thus is a feasible descent direction for f at $\bar{x}$. One may then perform linesearch along such a direction and compute a better feasible point than $\bar{x}$. Such a step is the basis of some early first order algorithms for constrained optimization (Frank–Wolfe or conditional gradient algorithm). For the algorithm based on these ideas to be efficient the subproblem (3) should be easier to solve than the original problem (1). This is for example the case when Ω is a solution set for a system of linear equations and inequalities.

For convex objective functions the condition (2) is also sufficient for global optimality over Ω.

Proposition 2 (Necessary and sufficient optimality conditions for convex problems). *Suppose that $\Omega \subset \mathbb{R}^n$ is a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable on Ω. Then $x^* \in \Omega$ is a point of global minimum for f over Ω* if and only if (2) *holds.*

Proof. The necessity of (2) has been established in Proposition 1, thus is only remains to show sufficiency. For differentiable convex functions we can write, for any $x \in \Omega$,

$$f(x) \geq f(x^*) + \underbrace{\nabla f(x^*)^T (x - x^*)}_{\geq 0} \geq f(x^*),$$

where the first inequality is established in Proposition 3, "Basic tools" note. Thus f attains its smallest value over Ω at $x^* \in \Omega$. □

The *variational inequality* (2) may still be difficult to check in practical situations as it involves infinitely many inequalities, one for every $x \in \Omega$. It is possible to convert it into a system of equations, which may be easier to verify. In order to do this we need a concept of Euclidean projection.

Proposition 3. *Let $\Omega \subset \mathbb{R}^n$ be a non-empty and closed set, and $z \in \mathbb{R}^n$ be an arbitrary point. We consider the optimization problem*

$$\begin{aligned} &\underset{x \in \mathbb{R}^n}{\text{minimize}} && \frac{1}{2}\|x - z\|_2^2, \\ &\text{subject to} && x \in \Omega. \end{aligned} \tag{4}$$

1. *There is at least one globally optimal solution to* (4).
2. *If, additionally, the set Ω is convex, then such a solution is unique.*

Proof. 1. The triangle inequality implies that $\|x - z\|_2 \geq \|x\|_2 - \|z\|_2$, and thus the objective function in (4) is coercive. Therefore, the claim follows from Theorem 2 in the Note on "Lower semi-continuity, compactness, and existence of solutions."

2. Suppose that $x_1 \in \Omega$ and $x_2 \in \Omega$ are both closest to z; then for every $0 < \lambda < 1$ we have the inequality

$$\|z - [\lambda x_1 + (1 - \lambda)x_2]\|_2^2 \geq \|z - x_1\|_2^2 = \|z - x_2\|_2^2,$$

since $\lambda x_1 + (1 - \lambda)x_2 \in \Omega$. On the other hand,

$$\begin{aligned} &\|z - [\lambda x_1 + (1 - \lambda)x_2]\|_2^2 \\ &= \lambda^2 \|z - x_1\|_2^2 + 2\lambda(1 - \lambda)(z - x_1)^T (z - x_2) + (1 - \lambda)^2 \|z - x_2\|_2^2 \\ &\leq \lambda^2 \|z - x_1\|_2^2 + 2\lambda(1 - \lambda)\|z - x_1\|_2 \|z - x_2\|_2 + (1 - \lambda)^2 \|z - x_2\|_2^2 \\ &= \|z - x_1\|^2 = \|z - x_2\|^2, \end{aligned}$$

where the equality sign in the Cauchy–Schwarz inequality is only possible when $z - x_1 = \alpha(z - x_2)$ for some $\alpha \geq 0$. In view of $\|z - x_1\| = \|z - x_2\|$ either $\alpha = 1$ or $\|z - x_1\| = \|z - x_2\| = 0$. In either case, $x_1 = x_2$, as claimed. □

For a non-empty, closed, and convex Ω we will write $\pi_\Omega[z]$ to denote the unique solution of the problem (4).

Proposition 4 (First order necessary optimality conditions and gradient projection). *Consider a convex set $\Omega \subset \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that $x^* \in \Omega$ is a point of local minimum for f over Ω, and further assume that f is continuously differentiable around x^*. Then for every $\alpha > 0$ it holds that*

$$\pi_\Omega[x^* - \alpha \nabla f(x^*)] = x^*, \tag{5}$$

that is, x^* is a fixed point of the map $x \mapsto \pi_\Omega[x - \alpha \nabla f(x)]$.

Proof. Owing to Proposition 1 we know that (2) holds at x^*. Also, $x^* = \pi_\Omega[z]$ if and only if

$$\nabla_x \left[\frac{1}{2} \|x - z\|_2^2 \right] \Big|_{x=x^*}^T (x - x^*) = (x^* - z)^T (x - x^*) \geq 0, \quad \forall x \in \Omega, \tag{6}$$

owing to the convexity of the objective function in (4) and Proposition 2. It only remains to substitute $z = x^* - \alpha \nabla f(x^*)$ into (6) to conclude the proof. □

Exercise 2. When we discussed the solution of the trust-region problem, we have not established the necessity of the conditions (4.8) in N&W for the optimality, only their sufficiency. Utilizing Proposition 4 and the explicit characterization of the projection onto the trust-region show that every locally optimal solution to the trust-region problem must satisfy the first order conditions (4.8a)–(4.8b) in N&W for some non-negative scalar λ.

Exercise 3. Prove the following finite-dimensional geometric version of Hahn–Banach theorem: every non-empty closed convex set $\Omega \subset \mathbb{R}^n$ and a point $z \in \mathbb{R}^n \setminus \Omega$ may be separated with a hyperplane. That is, there is $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, such that

$$a^T z < b \leq a^T x, \quad \forall x \in \Omega.$$

Thus a is a normal for the said hyperplane. Hint: consider the optimality conditions (2) for the projection problem (4).

Proposition 4 may in rare circumstances be utilized for algorithmic purposes as well. Suppose that Ω is such that $\pi_\Omega[\cdot]$ is easy to evaluate. Further assume that it is possible to select $\alpha > 0$ such that the function $F(x) := \pi_\Omega[x - \alpha \nabla f(x)]$ is a *contraction*, that is, $\exists 0 < \gamma < 1$ such that $\forall x, y \in \Omega$ it holds that $\|F(x) - F(y)\| \leq \gamma \|x - y\|$. Then the Banach fixed point theorem tells us that the iteration $x_{k+1} = F(x_k)$ converges to the unique fixed point of $F(\cdot)$. By the construction of $F(\cdot)$, this fixed point satisfies the necessary optimality conditions (5).

Finally, let us consider yet another equivalent characterization of optimality over convex sets.

Definition 2 (Normal cone). *For a convex set Ω and $\bar{x} \in \Omega$ let us consider the set*

$$N_\Omega[\bar{x}] = \{\, q \in \mathbb{R}^n \mid q^T(x - \bar{x}) \le 0, \forall x \in \Omega \,\}, \tag{7}$$

that is, a set of directions forming an angle of at least $\pi/2$ with feasible directions $p = x - \bar{x}$ at $\bar{x}$. Often one defines $N_\Omega[\bar{x}] = \emptyset$ for $\bar{x} \notin \Omega$. $N_\Omega[\bar{x}]$ is called the normal cone for Ω at $\bar{x}$.

Exercise 4. Show that for any $\bar{x} \in \Omega$ the set $N_\Omega[\bar{x}]$ is non-empty, convex, closed, and is a cone. The latter property is defined as follows: $q \in N_\Omega[\bar{x}] \implies \alpha q \in N_\Omega[\bar{x}]$, for any $\alpha > 0$.

Proposition 5. *Consider a convex set $\Omega \subset \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that $x^* \in \Omega$ is a point of local minimum for f over Ω, and further assume that f is continuously differentiable around x^*. Then $-\nabla f(x^*) \in N_\Omega[x^*]$.*

Proof. $-\nabla f(x^*)^T(x - x^*) \le 0$, $\forall x \in \Omega$ by Proposition 1, case 1. □

2 Even more special case: optimization over hyperplanes

Consider now the situation when the set Ω is defined by the linear equality constraints only, i.e.

$$\Omega = \{\, x \in \mathbb{R}^n \mid a_i^T x = b_i, i \in \mathcal{E} \,\}. \tag{8}$$

In particular Ω is closed and convex, and therefore for a point $x^* \in \Omega$ to be a point of local minimum for a continuously differentiable function f it must satisfy the variational inequality (2).

Let $L = \{\, p \in \mathbb{R}^n \mid p = x - x^*, x \in \Omega \,\} = \Omega - x^*$, and $\hat{L} = \{\, p \in \mathbb{R}^n \mid a_i^T p = 0 \,\}$. Owing to the linearity of the constraints, we have the inclusions $\Omega - x^* \subset \hat{L}$ and $x^* + \hat{L} \subset \Omega$; thus $L = \hat{L}$. In particular, L is a linear subspace of $\mathbb{R}^n$. Therefore if $p \in L$ then also $-p \in L$, and thus both $\nabla f(x^*)^T p \ge 0$ and $\nabla f(x^*)^T(-p) \ge 0$. We conclude that $\nabla f(x^*)^T p = 0$ for all $p \in L$.

Let us now form a matrix A with rows a_i^T, $i \in \mathcal{E}$. Then L is precisely the nullspace of A, and we have shown that $\nabla f(x^*)$ is orthogonal to all vectors in $\mathrm{null}(A)$. Thus $\nabla f(x^*) \in \mathrm{null}(A)^\perp = \mathrm{range}(A^T)$ (basic linear algebra results). We have established the following characterization of points of local minimum.

Proposition 6 (First order necessary optimality conditions for linear equality constraints). *Consider a set Ω given by* (8) *and an arbitrary continuously differentiable function f. Then $x^* \in \Omega$ is a point of local minimum for f over Ω only if $\nabla f(x^*) \in span[a_i, i \in \mathcal{E}]$. That is, there is a vector of* Lagrange multipliers *$\lambda \in \mathbb{R}^{|\mathcal{E}|}$, where $|\mathcal{E}|$ is the number of equality constraints, such that*

$$\nabla f(x^*) = \sum_{i \in \mathcal{E}} \lambda_i a_i = \sum_{i \in \mathcal{E}} \lambda_i \nabla c_i. \tag{9}$$

We will see that the characterization of the type (9) can be further generalized to include non-linear equality and inequality constraints under some technical regularity conditions; such a generalization is known as the Karush–Kuhn–Tucker (KKT) theorem. Equation (9) is simply a specialization of KKT conditions for linear equality constraints, in the same way as the conditions (4.8a)–(4.8b) in N&W is the specialization of the KKT conditions for the trust-region subproblem with one inequality constraint $\|p\| \leq \Delta_k$.

Example 2. Using Proposition 6 we can derive DFP (or BFGS) quasi-Newton algorithm. To simplify the notation we define $C = B - B_k$ and until the end of the derivation we will omit the quasi-Newton iteration index k. Remember that $B_k = B_k^T$. Consider an arbitrary symmetric and positive definite matrix W and the optimization problem

$$\begin{aligned}
\underset{C \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad & f(C) = \frac{1}{2}\text{trace}[WCW^TC^T] = \frac{1}{2}\sum_{\alpha=1}^{n}[CWCW^T]_{\alpha\alpha}, \\
\text{subject to} \quad & g_{ij}(C) = C_{ij} - C_{ji} = 0, \qquad i,j = 1,\ldots,n, \\
& h_i(C) = \sum_{j=1}^{n} C_{ij}s_j - \bar{y}_i = 0, \qquad i = 1,\ldots,n,
\end{aligned}$$

where $\bar{y}_i = y_i - \sum_{j=1}^{n}[B_k]_{ij}s_j$. Thus we have a minimization problem in n^2 optimization variables with $n^2 + n$ linear equality constraints. As a result, Proposition 6 applies. Let us denote the Lagrange multipliers corresponding to the matrix symmetry constraints with η_{ij}, $i,j = 1,\ldots,n$, and the ones corresponding to the secant equation with μ_i, $i = 1,\ldots,n$. It remains to calculate the derivatives of the objective function and the constraints with respect to B_{ij} (gradients of the constraints w.r.t. optimization variables result in the vectors a_i in the notation of Proposition 6).

$$
\begin{aligned}
[WCW^TC^T]_{\alpha\beta} &= \sum_{\gamma,\delta,\epsilon=1}^{n} W_{\alpha\gamma}C_{\gamma\delta}W_{\epsilon\delta}C_{\beta\epsilon}, \\
\frac{1}{2}\mathrm{trace}[WCW^TC^T] &= \frac{1}{2}\sum_{\alpha,\gamma,\delta,\epsilon=1}^{n} W_{\alpha\gamma}C_{\gamma\delta}W_{\epsilon\delta}C_{\alpha\epsilon}, \\
\frac{\partial f}{\partial C_{ij}} &= \underbrace{\frac{1}{2}\sum_{\alpha,\epsilon=1}^{n} W_{\alpha i}W_{\epsilon j}C_{\alpha\epsilon} + \frac{1}{2}\sum_{\gamma,\delta=1}^{n} W_{i\gamma}C_{\gamma\delta}W_{j\delta}}_{\text{remember: } W = W^T} \\
&= \sum_{\gamma,\delta=1}^{n} W_{i\gamma}C_{\gamma\delta}W_{\delta j} = [WCW]_{ij}, \\
\frac{\partial g_{k\ell}}{\partial C_{ij}} &= \begin{cases} 1, & \text{if } (i,j) = (k,\ell), \\ -1, & \text{if } (i,j) = (\ell,k), \\ 0, & \text{otherwise,} \end{cases} \\
\frac{\partial h_k}{\partial C_{ij}} &= \begin{cases} s_j, & \text{if } k = i, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\sum_{k,\ell=1}^{n} \eta_{k\ell}\frac{\partial g_{k\ell}}{\partial C_{ij}} &= \eta_{ij} - \eta_{ji}, \\
\sum_{k=1}^{n} \mu_k \frac{\partial h_k}{\partial C_{ij}} &= \mu_i s_j,
\end{aligned}
$$

and the feasibility of C and optimality conditions (9) are expressed as

$$
\begin{aligned}
C_{ij} - C_{ji} &= 0, && i,j = 1,\dots,n, \\
\sum_{j=1}^{n} C_{ij}s_j &= \bar{y}_i, && i = 1,\dots,n, \\
[WCW]_{ij} &= \eta_{ij} - \eta_{ji} + \mu_i s_j, && i,j = 1,\dots,n.
\end{aligned}
$$

In matrix-vector notation we write:

$$
\begin{aligned}
C - C^T &= 0, \\
Cs &= \bar{y}, \\
WCW &= \eta - \eta^T + \mu s^T,
\end{aligned}
$$

where we have introduced a matrix η with elements η_{ij} and a vector μ with elements μ_i. Thus we end up with a system of $2n^2 + n$ linear algebraic equations in $2n^2 + n$ unknowns C, η, μ - it can easily be solved numerically, but in this case we can also solve it analytically with some manipulations.

Since W is positive definite (non-singular) symmetric, we can solve the last equation for C and then transpose it:

$$\begin{aligned} C &= W^{-1}(\eta - \eta^T + \mu s^T)W^{-1}, \\ C^T &= W^{-1}(\eta^T - \eta + s\mu^T)W^{-1}. \end{aligned}$$

Remebering that $C - C^T = 0$ we get

$$\begin{aligned} W^{-1}(2\eta - 2\eta^T + \mu s^T - s\mu^T)W^{-1} &= 0, \\ \eta - \eta^T &= \frac{1}{2}[s\mu^T - \mu s^T]. \end{aligned}$$

We can substitute the latter equation back into the expression for C to obtain

$$C = \frac{1}{2}W^{-1}(s\mu^T + \mu s^T)W^{-1},$$

We now use the secant equation:

$$\begin{aligned} \bar{y} = Cs &= \frac{1}{2}W^{-1}(s\mu^T + \mu s^T)W^{-1}s, \\ 2W\bar{y} &= (s\mu^T + \mu s^T)W^{-1}s, \\ \mu(s^TW^{-1}s) &= 2W\bar{y} - s\mu^TW^{-1}s, \\ \mu &= [2W\bar{y} - s\mu^TW^{-1}s]/(s^TW^{-1}s), \end{aligned}$$

which gives us μ if we know $\mu^TW^{-1}s = s^TW^{-1}\mu$. To obtain the latter quantity we proceed further:

$$\begin{aligned} W^{-1}\mu &= [2\bar{y} - W^{-1}s\mu^TW^{-1}s]/(s^TW^{-1}s), \\ s^TW^{-1}\mu &= 2s^T\bar{y}/(s^TW^{-1}s) - \mu^TW^{-1}s, \\ s^TW^{-1}\mu &= s^T\bar{y}/(s^TW^{-1}s). \end{aligned}$$

Finally, the last expression is used to obtain the final expression for μ:

$$\begin{aligned} \mu &= 2W\bar{y}/(s^TW^{-1}s) - ss^T\bar{y}/(s^TW^{-1}s)^2 \\ &= \frac{2}{s^TW^{-1}s}W\bar{y} - \frac{s^T\bar{y}}{(s^TW^{-1}s)^2}s \end{aligned}$$

Therefore, the final expression for C is:

$$\begin{aligned} C &= \frac{1}{2}W^{-1}\left[\frac{2}{s^TW^{-1}s}s\bar{y}^TW - \frac{s^T\bar{y}}{(s^TW^{-1}s)^2}ss^T + \frac{2}{s^TW^{-1}s}W\bar{y}s^T - \frac{s^T\bar{y}}{(s^TW^{-1}s)^2}ss^T\right]W^{-1} \\ &= \frac{1}{s^TW^{-1}s}W^{-1}s\bar{y}^T + \frac{1}{s^TW^{-1}s}\bar{y}s^TW^{-1} - \frac{s^T\bar{y}}{(s^TW^{-1}s)^2}W^{-1}ss^TW^{-1}. \end{aligned}$$

Note that this expression is valid for an arbitrary symmetric positive definite matrix W, and it defines the only stationary point for our optimization problem.

With some more linear algebra one can show that the optimization problem is convex (this is the only place where positive definiteness, and not just non-singularity and symmetry of W is utilized) so that this is the globally optimal solution to the problem thanks to Proposition 2.

We can now select W satisfying the equation $Wy = s$, so that $W^{-1}s = y$. This results in

$$C = \frac{1}{s^T y} y\bar{y}^T + \frac{1}{s^T y}\bar{y}y^T - \frac{s^T\bar{y}}{(s^T y)^2} yy^T.$$

We now recall that $B = B_k + C$, $s = s_k$, $y = y_k$, $\bar{y} = y_k - B_k s_k$, $B_k^T = B_k$:

$$\begin{aligned} B &= B_k + \frac{1}{s_k^T y_k} y_k[y_k - B_k s_k]^T + \frac{1}{s_k^T y_k}[y_k - B_k s_k]y^T - \frac{s_k^T[y_k - B_k s_k]}{(s_k^T y_k)^2} y_k y_k^T \\ &= B_k + \frac{1}{s_k^T y_k} y_k y_k^T - \frac{1}{s_k^T y_k} y_k s_k^T B_k - \frac{1}{s_k^T y_k} B_k s_k y_k^T + \frac{s_k^T B_k s_k}{(s_k^T y_k)^2} y_k y_k^T \\ &= \left[I - \frac{1}{s_k^T y_k} y_k s_k^T\right] B_k \left[I - \frac{1}{s_k^T y_k} s_k y_k^T\right] + \frac{1}{s_k^T y_k} y_k y_k^T, \end{aligned}$$

which is precisely formula (6.13) for DFP Hessian update in N&W.

Representation theorem for polyhedral sets*

Anton Evgrafov

Department of Mathematical Sciences, NTNU anton.evgrafov@math.ntnu.no

Consider the following linear programming problem in the standard form:

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & c^T x, \\ \text{subject to} \quad & Ax = b, \\ & x \geq 0, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^N$, $b \in \mathbb{R}^m$. The existence of solutions for a feasible and bounded problem (1) relies upon the representation of the feasible set $\Omega = \{\, x \in \mathbb{R}^n \mid Ax = b, x \geq 0 \,\}$ as a sum $\Omega = P + C$, P is a convex, closed, and bounded set and C is a closed convex cone.

Before we begin, we reformulate Ω in terms of inequalities only:

$$\Omega = \{\, x \in \mathbb{R}^n \mid \tilde{A}x \leq \tilde{b} \,\}, \tag{2}$$

where

$$\tilde{A} = \begin{pmatrix} A \\ -A \\ -I \end{pmatrix}, \qquad \tilde{b} = \begin{pmatrix} b \\ -b \\ 0 \end{pmatrix}. \tag{3}$$

Note that the matrix $\tilde{A} \in \mathbb{R}^{(2m+n)\times n}$ always has rank n due to the presence of the identity matrix in the last block-row. The representation theorem applies to all matrices $\tilde{A} \in \mathbb{R}^{\ell\times n}$ with rank n (full column rank in particular $\ell \geq n$), not only matrices of the form (3).

For every $x \in \Omega$ we will write $\bar{\bar{A}}_x$ and $\bar{\bar{b}}_x$ to denote those rows of $\tilde{A}$ and the corresponding components of $\tilde{b}$, where the inequalities are active (binding) at x. The rest of the rows of $\tilde{A}$/components of $\tilde{b}$ will be denoted with $\overset{<}{A}_x$ and $\overset{<}{b}_x$. Thus $\bar{\bar{A}}_x x = \bar{\bar{b}}_x$ and $\overset{<}{A}_x x < \overset{<}{b}_x$.

Consider all points $v_i \in \Omega$ such that $\operatorname{rank} \bar{\bar{A}}_{v_i} = n$; thus $v_i = \bar{\bar{A}}_{v_i}^{-1} \bar{\bar{b}}_{v_i}$. Note that the number of such points is not larger than the number of ways of selecting n rows out of ℓ possibilities, that is $\ell!/(n!(\ell-n)!)$, and in principle could be 0. For a given $\tilde{A}$ and $\tilde{b}$ we will denote this number with N. Let

$$\begin{aligned} P &= \left\{ \sum_{i=1}^{N} \lambda_i v_i \mid \lambda_i \geq 0, \sum_{i=1}^{N} \lambda_i = 1 \right\}, \\ C &= \{\, d \in \mathbb{R}^n \mid \tilde{A}d \leq 0 \,\}. \end{aligned} \tag{4}$$

* Based on Section 3.2.3 in "Introduction to continuous optimization" by N. Andréasson, AE, M. Patriksson, E. Gustavsson, M. Önnheim: Studentlitteratur (2013), 2nd ed.

Theorem 1 (Representation theorem). *Consider a matrix $\tilde{A} \in \mathbb{R}^{\ell \times n}$ and a vector $\tilde{b} \in \mathbb{R}^{\ell}$ defining the set* (2)*, and the sets P and C defined in* (4)*. Suppose that* $\operatorname{rank} \tilde{A} = n$*. If P is non-empty then $\Omega = P + C$.*

Proof. The inclusion $P + C \subset \Omega$ is easy to verify. The other inclusion is proved is by induction in $\operatorname{rank} \bar{\tilde{A}}_x$, $x \in \Omega$.

First, consider the points in $x \in \Omega$ with $\operatorname{rank} \bar{\tilde{A}}_x = n$. These are precisely the points v_i defining the non-empty set P. Thus $x = v_i + 0$, for some $i = 1, \dots, N$. Note that $0 \in C$, thus $x \in P + C$.

Now assume that the representation holds for all $x \in \Omega$ such that $k \leq \operatorname{rank} \bar{\tilde{A}}_x \leq n$. We will show that the representation holds also for points $x \in \Omega$ with $\operatorname{rank} \bar{\tilde{A}}_x = k - 1$.

Let $x \in \Omega$ be such a point. Since $\operatorname{rank} \bar{\tilde{A}}_x < n$ there is $0 \neq z \in \operatorname{null} \bar{\tilde{A}}_x$. Consider a perturbed point $x + \lambda z$, $\lambda \in \mathbb{R}$. Since $\check{\tilde{A}}_x x < \check{\tilde{b}}_x$ and $\bar{\tilde{A}}_x z = 0$, it holds that $x + \lambda z \in \Omega$ for all small λ.

Let $\lambda^+ = \sup\{\, \lambda \in \mathbb{R} : x + \lambda z \in \Omega \,\}$ and $\lambda^- = \sup\{\, \lambda \in \mathbb{R} : x - \lambda z \in \Omega \,\}$. If $\lambda^+ = +\infty$ then

$$\tilde{A} z = \lim_{\lambda \to +\infty} \lambda^{-1} \tilde{A}[x + \lambda z] \leq \lim_{\lambda \to +\infty} \lambda^{-1} \tilde{b} = 0.$$

and therefore $z \in C$. Similarly, if $\lambda^- = +\infty$ then $-z \in C$.

Case 1: Suppose that $\lambda^- = \lambda^+ = +\infty$; then $0 \neq z \in C \cap [-C] = \operatorname{null} \tilde{A}$, which contradicts the assumption $\operatorname{rank} \tilde{A} = n$.

Case 2: Suppose $\lambda^+ < \infty$ but $\lambda^- = +\infty$. Consider the point $x^+ = x + \lambda^+ z$. Then $x^+ \in \Omega$ since Ω is closed. We claim that $\operatorname{rank} \bar{\tilde{A}}_{x^+} \geq k$. Indeed, $\bar{\tilde{A}}_x$ is a submatrix of $\bar{\tilde{A}}_{x^+}$ (recall, $\bar{\tilde{A}}_x z = 0$) and thus $\operatorname{rank} \bar{\tilde{A}}_{x^+} \geq k-1$. If $\operatorname{rank} \bar{\tilde{A}}_{x^+} = k-1$ then the additional rows in $\bar{\tilde{A}}_{x^+}$ (in relation to $\bar{\tilde{A}}_x$) may be expressed as linear combinations of rows in $\bar{\tilde{A}}_x$. Therefore, $z \in \operatorname{null} \bar{\tilde{A}}_{x^+}$ and $x^+ + \lambda z \in \Omega$, for all small λ. This contradicts the selection of λ^+, which was such that $x + \lambda z \notin \Omega$, $\lambda > \lambda^+$. It remains to utilize the induction hypothesis for x^+, that is $x^+ = x + \lambda z \in P + C$, and as a result $x \in P + (C + \lambda^+(-z)) = P + C$, since in this case $-z \in C$.

Case 3: Suppose $\lambda^+ = +\infty$ but $\lambda^- < \infty$. This case is completely symmetric with *Case 2*.

Case 4: Suppose that $\lambda^+ < \infty$ and $\lambda^- < \infty$. In this case the induction hypothesis applies to both x^+ and x^-. Therefore

$$x = \frac{\lambda^+}{\lambda^+ + \lambda^-} x^- + \frac{\lambda^-}{\lambda^+ + \lambda^-} x^+ \in \frac{\lambda^+}{\lambda^+ + \lambda^-}(P + C) + \frac{\lambda^-}{\lambda^+ + \lambda^-}(P + C) \subset P + C,$$

where the last inclusion is owing to the convexity of P, C. □

Proposition 1 (Existence of extreme poitns; see Theorem 13.2 in N&W). *Suppose that Ω given by* (2) *is non-empty and* $\operatorname{rank} \tilde{A} = n$*. Then the set P defined in* (4) *is non-empty.*

Proof. Take any $x \in \Omega \neq \emptyset$. If $\operatorname{rank} \bar{\bar{A}}_x = n$ we are done; otherwise we proceed as in the proof of Theorem 1 and define λ^+, λ^-. If $\lambda^+ < \infty$ we then go to the point x^+; otherwise $\lambda^- < \infty$ and then we go to the point x^-. In any case, $\operatorname{rank} \bar{\bar{A}}_{x^+} > \operatorname{rank} \bar{\bar{A}}_x$ or $\operatorname{rank} \bar{\bar{A}}_{x^-} > \operatorname{rank} \bar{\bar{A}}_x$. Repeating this procedure, we eventually reach a point $x \in \Omega$ where $\operatorname{rank} \bar{\bar{A}}_x = n$. □

Thm

KKT conditions are sufficient under convexity

Assume:
- f - convex
- c_i - concave, $i \in I$
- c_i - affine (linear), $i \in \mathcal{E}$
- x^* - KKT point for (P):

$$(P) \quad \min f(x) \quad \text{s.t.} \begin{cases} c_i(x) \geq 0, & i \in I \\ c_i(x) = 0, & i \in \mathcal{E} \end{cases} =: \Omega$$

Then x^* - global optimum for (P)

Proof: We will show that $f(x) \geq f(x^*), \forall x \in \Omega$

KKT - conditions for (P):

$$\nabla f(x^*) - \sum_{i \in \mathcal{E} \cup I} \lambda_i \nabla c_i(x^*) = 0$$

$$c_i(x^*) \geq 0, \quad i \in I$$

$$c_i(x^*) = 0, \quad i \in \mathcal{E}$$

$$\lambda_i \geq 0, \quad i \in I$$

$$\underbrace{\lambda_i \cdot c_i(x^*) = 0}_{\text{complementarity condition}}$$

Take $\forall x \in \Omega$

$f(x) \geq f(x^*) + \nabla f^T(x^*)[x - x^*]$ (f - convex, propositions in basic tools)

$$\nabla f(x^*)^T [x - x^*] = \quad (*)$$

$$= \sum_{i \in \mathcal{E} \cup I} \lambda_i \nabla c_i^T(x^*)[x - x^*]$$

- $\forall i \in \mathcal{E}$: $c_i(x) = c_i(x^*) = 0$, c_i - affine

$\Rightarrow \nabla c_i^T(x^*)[x - x^*] = c_i(x) - c_i(x^*) = 0$

- $\forall i \in I$: $c_i(x^*) > 0 \Rightarrow \lambda_i = 0$ (complementarity)

- $\forall i \in I$: $c_i(x^*) = 0 \Rightarrow$

$$0 \leq c_i(x) = c_i(x) - c_i(x^*) \leq \nabla c_i^T(x^*)[x - x^*]$$

(c_i - concave)

$\lambda_i \geq 0 \quad i \in I \Rightarrow \lambda_i \nabla c_i^T(x^*)[x - x^*] \geq 0.$

$(*) \Rightarrow \nabla f(x^*)^T [x - x^*] \geq 0.$

$\Rightarrow f(x) \geq f(x^*) \quad \forall \Omega$

∎

0 Preface

These lecture notes contain additional material for the optimization course. Section 1 gives a short introduction to variational calculus. A more detailed introduction to variational calculus and optimal control of ordinary differential equations will be given in a half-course this autumn at NTNU.

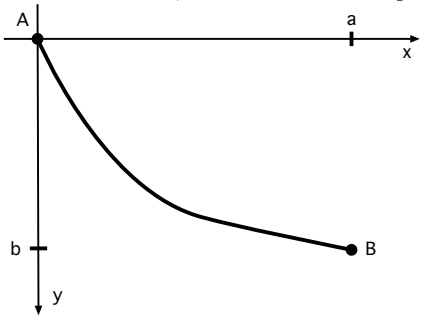
Section 2 describes some basic facts of optimal control of partial differential equations, where the PDE is first discretised by a finite difference or finite element approximation. These kind of problems will be discussed in detail in a new advanced course in optimization starting in January 2015.

Dietmar Hömberg
NTNU Trondheim and Technische Universität Berlin

1 Introduction to variational calculus

1.1 Examples, Introduction

Example 1.1 *Brachistochrone curve (Johann Bernoulli, 1696)*
Find the fastet path on which a point-like body moves from point A to point B under the influence of gravitation without friction. For convenience, take A as the origin of the coordinate system.



Path is described by the curve

$$\vec{r}(X) = \begin{pmatrix} x \\ y(x) \end{pmatrix} .$$

Arc length from $x = 0$ to $x = a$ is given as

$$L = s(a) = \int_0^a |\vec{r}(x)| dx = \int_0^a \sqrt{1 + (y')^2} dx \, .$$

For later use we note

$$\frac{ds}{dx} = \sqrt{1 + [y'(x)]^2} \, . \tag{1}$$

Enery conservation yield an expression for the velocity

$$\frac{1}{2} m v^2 = m \cdot gh = m \cdot gy$$

hence $v = \sqrt{2gy}$.

On the other hand $v = \dfrac{ds}{dt}$, where $s(t)$ is the arc length. Then we obtain

$$\frac{dt}{ds} = \frac{1}{v} \, . \tag{2}$$

The run-time of a mass point can be computed from

$$T = \int_0^T dt \overset{(2)}{=} \int_0^L \frac{1}{v} ds = \int_0^L \frac{1}{\sqrt{2gy(t(s))}} ds = \int_0^a \sqrt{\frac{1 + y'(x)^2}{2gy(x)}} dx \, .$$

Here the last manipulation has been done with the substitution $x = t(s)$ utilising (1).

Hence, the run-time is given as

$$T(y) = \int_0^a \sqrt{\frac{1+y'(x)^2}{2gy(x)}}dx\,, \tag{3}$$

and we seek a function y, which is at least continuously differentiable and satisfies the boundary conditions

$$y(0) = 0\,,\ \ y(a) = b$$

such that a minimal run-time is attained.

For a more rigorous formulation of the problem we recall the definition of the $\mathbb{R}$- vector space

$$V = \{f : [0,a] \to \mathbb{R} \Big|\ f \text{ is twice cont. differentiable}\} \ = C^2[0,a]\,.$$

To include the boundary conditions we introduce the subset

$$D = \{f \in C^2[0,a] \Big|\ f(0) = 0\,,\ f(a) = b\}\,.$$

Definition 1.1 *Let V be an $\mathbb{R}$-vector space and $D \subset V$. A mapping $I : D \to \mathbb{R}$, which assigns to each vector (or function) $f \in D$ a real value $I(f)$, is called functional.*

In the case of the Brachistochrone curve, $T(y)$ is the fuctional assigning to a given curve $y \in D$ the run-time$T(y) \in \mathbb{R}$ (or more precisely $\mathbb{R}_+$).

Example 1.2 *(Functionals)*

1. *$V = \mathbb{R}^n$, $I : V \to \mathbb{R}$, $I(v) = |v|$*
2. *V vector space of real-valued polynomials p, $I(p) =$ degree of p*
3. *point evaluation*
   V vector space of all functions $f : \mathbb{R} \to \mathbb{R}$, $\alpha \in \mathbb{R}$ fixed,

   $$I : V \to \mathbb{R}\,,\ I(f) = f(\alpha)$$

4. *V vector space of all continuous functions $f : [a,b] \to \mathbb{R}$, $p \geq 1$ fixed,*

   $$I : V \to \mathbb{R}\,,\ I(f) = \left(\int_0^b |f(x)|^p dx\right)^{\frac{1}{p}}$$

   is the so-called p-norm. For $p = \infty$ one obtains

   $$I : V \to \mathbb{R}\,,\ I(f) = \max_{x\in[a,b]} |f(x)|\,.$$

5. arc length $V = C^1[a,b]$,

$$I : V \to \mathbb{R}\,,\ I(f) = \int_0^b \sqrt{1 + f'(x)^2}dx\,.$$

In the following we always assume V to be a vector space of functions. Let I be a functional and $D \subset V$, then we consider the variational problem

$$\text{(P)} \qquad \min_{f\in D} I(f)\,.$$

Note that problem (P) not necessarily has a solution. Weierstraß has shown in 1870 that problem (P) in the case of

$$I(f) = \int_{-1}^1 (xf'(x))^2 dx$$

with $D = \left\{ f \in C^1[-1,1] \,\middle|\, f(-1) = 0\,,\ f(1) = 1 \right\}$ has no solution.

Mathematically, this problem could be solved later by D. Hilbert by introducing a new class of function spaces. Here, we only focus on techniques that allow to compute a solution to the variational problem if it exists. To this end we use the method of Gateaux variations, going back to L. Euler. To this end we assume y^* to be a solution to our problem (P), then we take a direction $v \in V$, such that

$$y_\varepsilon := y^* + \varepsilon v \quad \in D$$

for all $\varepsilon \in]-\varepsilon_0, \varepsilon_0[$, i.e. in a neighbourhood of 0. Then, the real-valued function

$$h : \mathbb{R} \to \mathbb{R}\,,\ h(\varepsilon) = I(y^* + \varepsilon v)$$

exhibits a local minimum in $\varepsilon = 0$. If h is differentiable, there holds

$$0 = \frac{d}{d\varepsilon} h(\varepsilon) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\Big[I(y^* + \varepsilon h) - I(y^*)\Big]\,.$$

Definition 1.2 *For a functional $I : D \subset V \to \mathbb{R}$, $y \in D$ and $v \in V$ with $y + \varepsilon v \in D$ for all $\varepsilon \in]-\varepsilon_0, \varepsilon_0[$,*

$$\delta I(y; v) := \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\Big[I(y + \varepsilon v) - I(y)\Big] = \frac{d}{d\varepsilon} I(y + \varepsilon v)\Big|_{\varepsilon=0}$$

is called first variation oder Gateaux variation of I at y in direction v, if this limit exists.

Example 1.3

1. Directional derivative in $\mathbb{R}^n$. $V = \mathbb{R}^n$, $I : \mathbb{R}^n \to \mathbb{R}$

$$\delta I(y; v) = \partial_v I(y) = \operatorname{grad} I \cdot v$$

2. $V = C([a,b],\ I(f) = \int_0^b f^2 dx,$

then there holds for $g \in V$

$$\begin{aligned}\delta I(f;g) &= \frac{d}{d\varepsilon}\int_a^b (f+\varepsilon g)^2 dx\Big|_{\varepsilon=0} = \frac{d}{d\varepsilon}\int_a^b (f^2 + 2\varepsilon fg + \varepsilon g^2)dx\Big|_{\varepsilon=0} \\ &= \frac{d}{d\varepsilon}\left(\int_a^b f^2 dx + 2\varepsilon\int_a^b fg\,dx + \varepsilon^2\int_a^b g^2 dx\right)\Big|_{\varepsilon=0} = 2\int_a^b f(x)g(x)dx\,.\end{aligned}$$

The above considerations show that we obtain the following condition for the solution of our variational problem:

Theorem 1.1 *Let* y^* *be a solution of* (P) *and* $v \in V$ *an admissible direction, i.e., there exists* $\varepsilon_0 > 0$*, such that* $v^* + \varepsilon v \in D$ *for all* $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$*, then there holds*

$$\delta I(y^*; v) = 0$$

if the Gateaux variation exists.

Remark:
Theorem 1.1 only provides a necessary optimality condition, hence one has to check case by case, if the obtained function indeed is the desired minimum.

1.2 The Euler-Lagrange differential equation

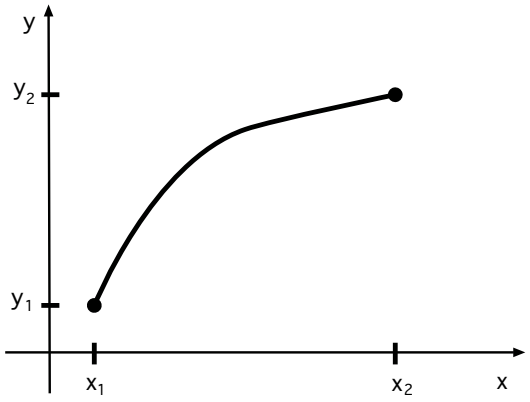
Now we consider the following typical problem:

Let (x_0, y_0) and (x_1, y_2) be two points in $\mathbb{R}^2$ and

$$D = \left\{ y \in C^2[x_0, x_1] \,\middle|\, y(x_0) = y_1\,,\ y(x_1) = y_1 \right\}.$$

Moreover, we assume that $F(x, y, y')$ is continuously differentiable in all arguments and introduce the problem

$$\text{(P)} \qquad \min_{y\in D} I(y) := \int_{x_0}^{x_1} F(x, y(x), y'(x))dx \,...$$



An important tool for our considerations will be the following fundamental lemma of the calculus of variations:

Lemma 1.1 *If for* $f : [a, b] \to \mathbb{R}$ *there holds*

$$\int_a^b f(x)g(x)dx = 0 \quad \textit{for all } g \in C^2[a, b] \quad \textit{with } g(a) = g(b) = 0$$

then $f = 0$.

Proof: Assume $f(x_0) \neq 0$ for an x_0, for example, let $f(x_0) > 0$. f is continuous, hence there exists $\varepsilon_0 > 0$, such that $f(x) > 0$ for all $x \in]x_0 - \varepsilon_0,\, x_0 + \varepsilon_0[$. Let g be a function satisfying $g(x) > 0$ for all $x \in]x_0 - \varepsilon_0,\, x_0 + \varepsilon_0[$ and $g(x) = 0$ for $|x - x_0| \geq \varepsilon_0$, then

$$\int_a^b f(x)g(x)dx = \int_{x_0-\varepsilon}^{x_0+\varepsilon} f(x)g(x)dx > 0$$

contradicting the precondition. ∎

The solution of the variational problem (P) can be characterised as the solution to an ordinary differential equation:

Theorem 1.2 *Each solution* y^* *to* (P) *necessarily solves the Euler-Langrange equation to problem* (P),

$$\frac{\partial F}{\partial y}(x, y, y') - \frac{d}{dx}\frac{\partial F}{\partial y'}(x, y, y') = 0$$

Proof: Let $g \in C^2[x_0, x_1]$ mit $g(x_0) = g(x_1) = 0$, then

$$y_\varepsilon(x) = y^*(x) + \varepsilon g(x) \in D$$

for all $\varepsilon > 0$ sufficiently small, i.e., y_ε is admissible. moreover, we have

$$y'_\varepsilon(x) = y'(x) + \varepsilon g'(x)$$

In view of the assumptions on F, the Gateaux variation exists and with Theorem 1.1, we can infer

$$\begin{aligned} 0 = \partial I(y^*; g) &= \frac{d}{d\varepsilon}\int_{x_0}^{x_1} F(x, y_s(x), y'_\varepsilon(x))dx\Big|_{\varepsilon=0} \\ &= \int_{x_0}^{x_1} F_y(x, y^*, y^{*'}g(x)dx + \int_{x_0}^{x_1} F_{y'}(x, y^*, y^{*'})g'(x)dx\,. \end{aligned}$$

Integration by parts in the second summand yields

$$\begin{aligned} \int_{x_0}^{x_1} F_{y'}(x, y^*), y^{*'}g'(x)dx &= \underbrace{F_{y'}(x, y^*, y^{*'})g(x)\Big|_{x_0}^{x_1}}_{=0} \\ &\quad - \int_{x_0}^{x_1} \frac{d}{dx}F_{y'}(x, y^*, y^{*'}g(x)dx\,. \end{aligned}$$

All in all, we obtain

$$\int_{x_0}^{x_1} \left[F_y(x, y^*, y^{*'}) - \frac{d}{dx} F_{y'}(x, y^*, y^{*'}) \right] g(x) dx = 0$$

for all $g(x)$ with $g(x_0) = g(x_1) = 0$.

Applying the fundamental lemma concludes the proof.

Please note: The second term in the Euler-Lagrange equation written down explicitly is

$$\frac{d}{dx} F_{y'}(x, y, y') = F_{y'x} + F_{y'y} y' + F_{y'y'} y'' \,.$$

Hence, the Euler-Lagrange equation is a differential equation of second order.

Example 1.4 *(The shortest connection between two points in a plain is a line segment.) Let D as before and $y(x)$ the curve connecting these two points. Then the arc length is given as*

$$I(y) = \int_{x_0}^{x_1} \sqrt{1 + y'(x)^2} dx \,.$$

To illustrate the theorem we again derive the Euler-Lagrange equation from the first variation of the functional.

To this end, we consider an admissible perturbation

$$y_\varepsilon = y + \varepsilon g \,, \; y'_\varepsilon = y' + \varepsilon g' \quad \textit{and} \quad \frac{d}{d\varepsilon} y'_\varepsilon = g'$$

and use Theorem 1.1 to obtain

$$0 = \partial I(y, g) = \frac{d}{d\varepsilon} \int_{x_0}^{x_1} \sqrt{1 + y'_\varepsilon(x)^2} dx \Big|_{\varepsilon \in 0} = \int_{x_0}^{x_1} \frac{y'(x)}{\sqrt{1 + y'(x)}} g'(x) dx$$

$$= \frac{2y'(x)}{\sqrt{1 + y'(x)}} g(x) \Big|_{x_0}^{x_1} - \int_{x_0}^{x_1} \left(\frac{d}{dx} \frac{y'(x)}{\sqrt{1 + y'(x)}} \right) g(x) dx = 0 \,.$$

Using the variational lemma leads to

$$\frac{d}{dx} \frac{y'(x)}{\sqrt{1 + y'(x)}} = 0 \,,$$

a differential equation of second order. We can conclude

$$\frac{y'}{\sqrt{1 + y'(x)}} = c \,.$$

With $c < 1$ we obtain $y^2 = c^2 + c^2 y'^2$ or $y'^2(1 - c^2) = c^2$, hence

$$y' = \pm \sqrt{\frac{c^2}{1 - c^2}} = \text{ const } ,$$

and $y = ax + b$ is a linear function. Now we can compute a, b from initial and end conditions, i.e., we obtain

$$y = y_0 + (y_1 - y_0) \frac{x - x_0}{x_1 - x_0} \,.$$

Example 1.5 *(Brachistochrone curve)*
According to Example 1.1 there holds

$$T(y) = \int_0^a \sqrt{\frac{1+y'(x)^2}{2gy(x)}}dx\,.$$

We may neglect constant factors for optimisation, hence we define

$$F(y,y') = \sqrt{\frac{1+y'^2}{y}}\,.$$

Then we obtain the following Euler-Lagrange equation

$$0 = \frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'} = \frac{\partial F}{\partial y} - F_{y'y}y'(x) - F_{y'y'}y''(x)\,.$$

Multiplying the equation with y' yields

$$0 = F_y y' - F_{y'y}y'^2 - F_{y'y'}y''y' = \frac{d}{dx}\left(F - y'F_{y'}\right).$$

Hence, we can infer $F - y'F_{y'} = \text{const}$, i.e.,

$$\sqrt{\frac{1+y'^2}{y}} - \frac{y'^2}{y\sqrt{\frac{1+y'^2}{y}}} = \sqrt{\frac{1+y'^2}{y}} - \frac{y'^2}{\sqrt{y(1+y'^2)}} = c\,.$$

After multiplication with $\sqrt{y(1+y'^2)}$ we get

$$1 + y'^2 - y'^2 = c\sqrt{y(1+y'^2)}.$$

After squaring,

$$y(1+y'^2) = \frac{1}{c^2}$$

and thus $y' = \sqrt{\dfrac{1-yc^2}{yc^2}}$.
Separation of variables gives

$$\int dx = \int \sqrt{\frac{yc^2}{1-yc^2}}dy.$$

We substitute

$$y(t) = \frac{1}{c^2}\sin^2 t \quad , \text{ i.e., } \frac{dy}{dt} = \frac{2}{c^2}\sin t\cos t$$

with $t \in \left[0, \frac{\pi}{2}\right]$. *Then, we obtain*

$$\begin{aligned}\int \sqrt{\frac{yc^2}{1-yc^2}}dy \;&=\; \int \sqrt{\frac{\sin^2 t}{1-\sin^2 t}}\,\frac{2}{c^2} \quad \sin t\cos dt = \frac{2}{c^2}\int \sin^2 t\, dt \\ &\overset{(*)}{=}\; \frac{2}{c^2}\int \left(\frac{1}{2}-\frac{1}{2}\cos(2t)\right)dt = \frac{1}{c^2}t - \frac{1}{2c^2}\sin(2t) + \tilde{c} \\ &=\; \frac{1}{2c^2}(2z-\sin(2t)) + \tilde{c} = x(t)\,.\end{aligned}$$

Since $x(0) = 0$ *we have* $\tilde{c} = 0$. *For* y *we obtain*

$$y(t) = \frac{1}{c^2}\sin^2 t \overset{(*)}{=} \frac{1}{2c^2}(1-\cos(2t))\,.$$

Altogether, we obtain the solution curve

$$(x(t), y(t)) = \frac{1}{2c^2}\Big(2t - \sin(2t),\, 1-\cos(2t)\Big)\,,$$

i.e., a cycloid. In our computations we have used the addition formula for cosine, which implies

$$\cos(2t) = \cos^2 t - \sin^2 t = 1 - 2\sin^2 t,$$

and thus

$$(*) \qquad \sin^2 t = \frac{1}{2} - \frac{1}{2}\cos(2t)\,.$$

1.3 Natural boundary conditions

Example 1.6 *(tension-compression bar)*
Consider a bar Ω *of length* l *with constant cross section* A. *Let* $u(x)$ *denote the displacement in* x*-direction. The bar is clamped on the left, i.e.,*

$$u(0) = 0\,.$$

On the right-hand side it is stress free. We recall Hooke's law

$$\sigma = E\varepsilon = Eu_x,$$

where ε *is the strain and* E *the modulus of elasticity. The stable equilibrium of the bar corresponds to its energetic minimum. We define the overall elastic potential of the bar as*

$$\Pi(u) = W_f(u) - W_a(u)\,.$$

Here, the strain energy is given as

$$W_f(u) = \frac{1}{2}\int_\Omega \sigma\varepsilon dV = \frac{1}{2}A\int_0^l Eu_x^2\,dx$$

and the potential energy of applied forces is defined by

$$W_a(u) = \int_\Omega pudV = A\int_0^l p(x)u(x)\,dx\,,$$

with a given force per length $p(x)$.

Let $D = \{u \in C^2[0,l] \text{ with } u(0) = 0\}$ and $h \in D$ then $u^\varepsilon = u + \varepsilon h \in D$ for all $\varepsilon > 0$. We search for a displacement function u^* as the solution to

$$\text{(P)} \qquad \min_{u\in D}\Pi(u)$$

According to Theorem 1.1 the necessary condition $\delta\Pi(u^*;h) = 0$ for all admissible h holds, i.e.,

$$\begin{aligned}
0 \quad &= \quad \delta\Pi(u,h) = \frac{d}{d\varepsilon}\left(\frac{A}{2}\int_0^l Eu_x^{\varepsilon^2}dx - A\int_0^l pu^\varepsilon dx\right)\Bigg|_{\varepsilon=0}\\
&= \quad \frac{d}{d\varepsilon}\left(\frac{A}{2}\int_0^l E(u_x+\varepsilon h_x)^2dx - A\int_0^l p(u+\varepsilon h)dx\right)\Bigg|_{\varepsilon=0}\\
&= \quad A\left(\int_0^l E(u_x+\varepsilon h_x)h_x dx - A\int_0^l ph\,dx\right)\Bigg|_{\varepsilon=0}\\
&= \quad A\int_0^l Eu_x h_x\,dx - A\int_0^l ph\,dx\\
&\overset{\text{integration by parts}}{=} \quad A\int_0^l \left(-(Eu_x)_x - p\right)h(x)dx + AEu_x h\Big|_0^l\,.
\end{aligned}$$

Now, we demand in addition that h also vanishes at $x = l$, in other words $h(l) = h(0) = 0$. Then

$$\int_0^l \big(-(Eu_x)_x - p\big)h\,dx = 0 \tag{4}$$

for all $h \in C^2[0,l]$ with $h(0) = h(l) = 0$ and with Lemma 5.1 we can infer

$$-(Eu_x)_x = p\,,\ x \in (0,l).$$

This also implies

$$AEu_x h\Big|_0^l = AEu_x(l)h(l) = 0 \tag{5}$$

for all $h \in D$.

Now we choose $h(x) = \frac{1}{l}x$, then $h \in C^2[0,l]$, $h(0) = 0$ und $h(l) = 1$, and thus $h \in D$. We obtain $AEu_x(l) = 0$ and hence

$$u_x(l) = 0\,. \tag{6}$$

Remark: We have recovered boundary condition (6) from the variational formulation. Such a condition is called a natural boundary condition or Neumann condition. The condition $u(0) = 0$ has to be prescribed in the variational space D, hence it is called an essential or a Dirichlet boundary condition.

All in all, we obtain the boundary value problem

$$-\frac{d}{dx}\left(E\frac{du}{dx}\right) = p(x) \quad \text{in } (0,l) \tag{7a}$$

$$u(0) = 0 \tag{7b}$$

$$u_x(l) = 0\,. \tag{7c}$$

Applying Hooke's law (7c), we see that

$$\sigma(u)\Big|_{x=l} = Eu_x\Big|_{x=l} = 0,$$

hence indeed the right-hand side of the bar is stress free.

2 An introduction to PDE-constrained control

1.4. Basic concepts for the finite-dimensional case

Some fundamental concepts of optimal control theory can easily be explained by considering optimization problems in Euclidean space with finitely many equality constraints. A little detour into finite-dimensional optimization has

the advantage that the basic ideas will not be complicated by technical details from partial differential equations or functional analysis.

1.4.1. Finite-dimensional optimal control problems. Suppose that $J = J(y,u)$, $J : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, denotes a cost functional to be minimized, and that an $n \times n$ matrix A, an $n \times m$ matrix B, and a nonempty set $U_{ad} \subset \mathbb{R}^m$ are given (where "ad" stands for "admissible"). We consider the *optimization problem*

$$\boxed{\begin{array}{ll} \min J(y,u) & \\ A\,y = B\,u, & u \in U_{ad}. \end{array}} \tag{1.1}$$

We seek vectors y and u minimizing the cost functional J subject to the constraints $A\,y = B\,u$ and $u \in U_{ad}$. In this connection, we introduce the following convention: Unless specified otherwise, throughout this book vectors will always be regarded as *column* vectors.

Example. Often quadratic cost functionals are used, for instance

$$J(y,u) = |y - y_d|^2 + \lambda\,|u|^2,$$

where $|\cdot|$ denotes the Euclidean norm. ◇

As it stands, (1.1) is a standard optimization problem in which the unknowns y and u play similar roles. But this situation changes if we make the additional assumption that the matrix A has an inverse A^{-1}. Indeed, we can then solve for y in (1.1), obtaining

$$y = A^{-1}B\,u, \tag{1.2}$$

and for any $u \in \mathbb{R}^m$ there is a uniquely determined solution $y \in \mathbb{R}^n$; that is, we may choose (i.e. "control") u in an arbitrary way to produce the associated y as a dependent quantity. We therefore call u the control vector or, for short, the *control*, and y the associated state vector or *state*. In this way, (1.1) becomes a finite-dimensional optimal control problem.

Next, we introduce the *solution matrix* of our control system

$$S : \mathbb{R}^m \to \mathbb{R}^n, \quad S = A^{-1}B\,.$$

Then $y = S\,u$, and, owing to (1.2), we can eliminate y from J to obtain the *reduced cost functional* f,

$$J(y,u) = J(S\,u,u) =: f(u).$$

For instance, for the quadratic function in the above example we get $f(u) = |Su - y_d|^2 + \lambda|u|^2$. The problem (1.1) then becomes the nonlinear optimization

problem

$$\min f(u), \quad u \in U_{ad}. \tag{1.3}$$

In this *reduced problem* only the control u appears as an unknown.

In the following sections, we will discuss some basic ideas that will be repeatedly encountered in similar forms in the optimal control of partial differential equations.

1.4.2. Existence of optimal controls.

Definition. *A vector $\bar{u} \in U_{ad}$ is called an* optimal control *for problem* (1.1) *if $f(\bar{u}) \leq f(u)$ for all $u \in U_{ad}$; then $\bar{y} := S\bar{u}$ is called the* optimal state *associated with $\bar{u}$.*

Optimal or locally optimal quantities will be indicated by overlining, as in $\bar{u}$.

Theorem 1.1. *Suppose that J is continuous on $\mathbb{R}^n \times U_{ad}$ and that the set U_{ad} is nonempty, bounded, and closed. If the matrix A is invertible, then* (1.1) *has at least one solution.*

Proof: Obviously, the continuity of J implies that f is also continuous on U_{ad}. Moreover, as a bounded and closed set in a finite-dimensional space, U_{ad} is compact. By the well-known Weierstrass theorem, f attains its minimum in U_{ad}. Hence, there is some $\bar{u} \in U_{ad}$ such that $f(\bar{u}) = \min\limits_{u \in U_{ad}} f(u)$. □

This proof becomes more complicated in the case of optimal control problems for partial differential equations, since bounded and closed sets need not be compact in (infinite-dimensional) function spaces.

1.4.3. First-order necessary optimality conditions. In this section, we investigate what conditions the optimal vectors $\bar{u}$ and $\bar{y}$ must satisfy. We do this in the hope that we will be able to extract enough information from these conditions to determine $\bar{u}$ and $\bar{y}$. Usually, this will have to be done using numerical methods.

Notation. We use the following notation for the derivatives of functions $f : \mathbb{R}^m \to \mathbb{R}$:

$$\begin{aligned} D_i &= \frac{\partial}{\partial x_i}, \quad D_x = \frac{\partial}{\partial x} && \text{(partial derivatives)} \\ f'(x) &= \big(D_1 f(x), \ldots, D_m f(x)\big) && \text{(derivative)} \\ \nabla f(u) &= f'(u)^\top && \text{(gradient)} \end{aligned}$$

where $^\top$ stands for transposition. For functions $f = f(x, y) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, we denote by $D_x f$ the row vector of partial derivatives of f with respect to

$x_1, ..., x_m$, and by $\nabla_x f$ the corresponding column vector. The expressions $D_y f$ and $\nabla_y f$ are defined in a similar way. Moreover,

$$\big(u\,,\,v\big)_{\mathbb{R}^m} = u \cdot v = \sum_{i=1}^{m} u_i\, v_i$$

denotes the standard Euclidean scalar product in $\mathbb{R}^m$. For the sake of convenience, we will use both kinds of notation for the scalar product between vectors. The application of $f'(u)$ to a column vector $h \in \mathbb{R}^m$, denoted by $f'(u)\,h$, coincides with the directional derivative of f in the direction h,

$$f'(u)\,h = \big(\nabla f(u)\,,\,h\big)_{\mathbb{R}^m} = \nabla f(u) \cdot h.$$

We now make the additional assumption that the cost functional J is continuously differentiable with respect to y and u; that is, the partial derivatives $D_y J(y,u)$ and $D_u J(y,u)$ with respect to y and u are continuous in (y,u). Then, by virtue of the chain rule, $f(u) = J(Su,u)$ is continuously differentiable.

Example. Suppose that $f(u) = \frac{1}{2}\,|Su - y_d|^2 + \frac{\lambda}{2}|u|^2$. Then it follows that

$$\begin{aligned} \nabla f(u) &= S^\top(Su - y_d) + \lambda\, u, \quad f'(u) = \big(S^\top(Su - y_d) + \lambda\, u\big)^\top, \\ f'(u)\,h &= \big(S^\top(Su - y_d) + \lambda\, u\,,\,h\big)_{\mathbb{R}^m}. \end{aligned}$$ ◇

Theorem 1.2. *Let U_{ad} be convex. Then any optimal control $\bar{u}$ for* (1.1) *satisfies the* variational inequality

$$f'(\bar{u})(u - \bar{u}) \geq 0 \qquad \forall u \in U_{ad}. \tag{1.4}$$

This simple yet fundamental result is a special case of Lemma 2.21 on page 63. It reflects the observation that f cannot decrease in any direction at a minimum point.

Invoking the chain rule and the rules for total differentials, we can determine the derivative f' in (1.4), which is given by $f' = D_y J\, S + D_u J$. We find that

$$\begin{aligned} f'(\bar{u})\,h &= D_y J(S\bar{u},\bar{u})\,Sh + D_u J(S\bar{u},\bar{u})\,h \\ &= \big(\nabla_y J(\bar{y},\bar{u})\,,\,A^{-1}B\,h\big)_{\mathbb{R}^n} + \big(\nabla_u J(\bar{y},\bar{u})\,,\,h\big)_{\mathbb{R}^m} \\ &= \big(B^\top(A^\top)^{-1}\nabla_y J(\bar{y},\bar{u}) + \nabla_u J(\bar{y},\bar{u})\,,\,h\big)_{\mathbb{R}^m}. \end{aligned} \tag{1.5}$$

Hence, the variational inequality (1.4) takes the somewhat clumsy form

$$\big(B^\top(A^\top)^{-1}\nabla_y J(\bar{y},\bar{u}) + \nabla_u J(\bar{y},\bar{u})\,,\,u - \bar{u}\big)_{\mathbb{R}^m} \geq 0 \qquad \forall\, u \in U_{ad}. \tag{1.6}$$

It can be considerably simplified by introducing the adjoint state, a simple trick that is of utmost importance in optimal control theory.

1.4.4. Adjoint state and reduced gradient. As motivation, let us assume that the use of the inverse matrix A^{-1} is too costly for numerical calculations. This is usually the case for realistic optimal control problems. Then, a numerical method that avoids the explicit calculation of A^{-1} (e.g., the conjugate gradient method) must be used for the solution of the linear system $Ay = b$. The same applies for $A^\top$. We therefore replace the term $(A^\top)^{-1}\nabla_y J(\bar{y}, \bar{u})$ in (1.6) by a new variable $\bar{p}$,

$$\bar{p} := (A^\top)^{-1}\nabla_y J(\bar{y}, \bar{u}).$$

The quantity $\bar{p}$ corresponding to the pair $(\bar{y}, \bar{u})$ can be determined by solving the linear system

$$A^\top \bar{p} = \nabla_y J(\bar{y}, \bar{u}). \tag{1.7}$$

Definition. *The equation* (1.7) *is called the* adjoint equation, *and its solution* $\bar{p}$ *is called the* adjoint state *associated with* $(\bar{y}, \bar{u})$.

Example. In the case of the quadratic function $J(y, u) = \frac{1}{2}|y - y_d|^2 + \frac{\lambda}{2}|u|^2$, we obtain the adjoint equation

$$A^\top \bar{p} = \bar{y} - y_d,$$

since $\nabla_y J(y, u) = y - y_d$. ◇

The introduction of the adjoint state has two advantages: the first-order necessary optimality conditions simplify, and the use of the inverse matrix $(A^\top)^{-1}$ is avoided. Also, the form of the gradient of f simplifies. Indeed, with $\bar{y} = S\bar{u}$, it follows from (1.5) that

$$\nabla f(\bar{u}) = B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}).$$

The vector $\nabla f(\bar{u})$ is referred to as the *reduced gradient*. Moreover, since $\bar{y} = S\bar{u}$, the directional derivative $f'(\bar{u})\,h$ at an arbitrary point $\bar{u}$ is given by

$$f'(\bar{u})\,h = \big(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\,,\, h\big)_{\mathbb{R}^m}.$$

The two expressions above involving the adjoint state $\bar{p}$ do not depend on whether $\bar{u}$ is optimal or not. We will encounter them repeatedly in control problems for partial differential equations. Moreover, the use of the adjoint state $\bar{p}$ also simplifies Theorem 1.2:

Theorem 1.3. *Suppose that the matrix* A *is invertible, and let* $\bar{u}$ *be an optimal control for* (1.1) *with associated state* $\bar{y}$. *Then the adjoint equation* (1.7) *has a unique solution* $\bar{p}$ *such that*

$$\big(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\,,\, u - \bar{u}\big)_{\mathbb{R}^m} \geq 0 \quad \forall\, u \in U_{ad}. \tag{1.8}$$

The assertion follows directly from the variational inequality (1.6) and the definition of $\bar{p}$. In summary, we have derived the following *optimality system* for the unknown vectors $\bar{y}$, $\bar{u}$, and $\bar{p}$, which can be used to determine the optimal control:

$$\boxed{\begin{aligned} &A\,y = B\,u, \quad u \in U_{ad} \\ &A^\top p = \nabla_y J(y,u) \\ &\big(B^\top p + \nabla_u J(y,u)\,,\, v-u\big)_{\mathbb{R}^m} \geq 0 \quad \forall\, v \in U_{ad}. \end{aligned}} \tag{1.9}$$

Every solution $(\bar{y},\bar{u})$ to the optimal control problem (1.1) must, together with $\bar{p}$, satisfy this system.

No restrictions on u. In this case, $U_{ad} = \mathbb{R}^m$. Then $u - \bar{u}$ may attain any value $h \in \mathbb{R}^m$, and thus the variational inequality (1.8) reduces to the equation

$$B^\top \bar{p} + \nabla_u J(\bar{y},\bar{u}) = 0.$$

Example. Suppose that

$$J(y,u) = \frac{1}{2}\,|C\,y - y_d|^2 + \frac{\lambda}{2}\,|u|^2,$$

with a given $n \times n$ matrix C. Then, obviously,

$$\nabla_y J(y,u) = C^\top (C\,y - y_d), \qquad \nabla_u J(y,u) = \lambda u.$$

The optimality system becomes

$$\begin{aligned} &A\,y = B\,u, \quad u \in U_{ad} \\ &A^\top p = C^\top (Cy - y_d) \\ &\big(B^\top p + \lambda\,u\,,\, v-u\big)_{\mathbb{R}^m} \geq 0 \quad \forall\, v \in U_{ad}. \end{aligned}$$

If $U_{ad} = \mathbb{R}^m$, then $B^\top \bar{p} + \lambda \bar{u} = 0$. In the case where $\lambda > 0$, we can solve for $\bar{u}$ to obtain

$$\bar{u} = -\frac{1}{\lambda}\,B^\top \bar{p}. \tag{1.10}$$

Substitution in the two other relations yields the optimality system

$$\boxed{\begin{aligned} &A\,y = -\frac{1}{\lambda} B\,B^\top p \\ &A^\top p = C^\top (Cy - y_d), \end{aligned}}$$

which is a linear system for the unknowns $\bar{y}$ and $\bar{p}$. Once $\bar{y}$ and $\bar{p}$ have been recovered from it, the optimal control $\bar{u}$ can be determined from (1.10). ◇

Remark. We have chosen a linear equation in (1.1) for the sake of simplicity. The fully nonlinear problem

$$\min J(y,u), \quad T(y,u) = 0, \quad u \in U_{ad} \tag{1.11}$$

will be discussed in Exercise 2.1 on page 116.

1.4.5. Lagrangians. By using the Lagrangian function from basic calculus, the optimality system can also be formulated as a *Lagrange multiplier rule.*

Definition. *The function*

$$L : \mathbb{R}^{2n+m} \to \mathbb{R}, \quad L(y,u,p) := J(y,u) - \big(A\,y - B\,u\,,\,p\big)_{\mathbb{R}^n},$$

is called the Lagrangian function *or* Lagrangian.

Using L, we can formally eliminate the equality constraints from (1.1), while retaining the seemingly simpler restriction $u \in U_{ad}$ in explicit form. Upon comparison, we find that the second and third conditions in the optimality system are equivalent to

$$\begin{aligned} &\nabla_y L(\bar{y},\bar{u},\bar{p}) = 0 \\ &\big(\nabla_u L(\bar{y},\bar{u},\bar{p})\,,\,u-\bar{u}\big)_{\mathbb{R}^m} \geq 0 \quad \forall\, u \in U_{ad}. \end{aligned}$$

Conclusion. *The adjoint equation* (1.7) *is equivalent to* $\nabla_y L(\bar{y},\bar{u},\bar{p}) = 0$ *and thus can be recovered by differentiating the Lagrangian with respect to* y*. Similarly, the variational inequality follows from differentiation of* L *with respect to* u*.*

Consequently, $(\bar{y},\bar{u})$ is a solution to the necessary optimality conditions of the following minimization problem without equality constraints:

$$\min_{y,u} L(y,u,p), \qquad u \in U_{ad},\ y \in \mathbb{R}^n. \tag{1.12}$$

By the way, this does not imply that $(\bar{y},\bar{u})$ can always be determined numerically as a solution to (1.12). In fact, the "right" $\bar{p}$ is usually not known, and (1.12) may not be solvable or could even lead to wrong solutions. The vector $\bar{p} \in \mathbb{R}^n$ also plays the role of a *Lagrange multiplier.* It corresponds to the equation $Ay - Bu = 0$.

We remark that the above conclusion remains valid for the fully nonlinear problem (1.11), provided that the Lagrangian is defined by $L(y,u,p) := J(y,u) - \big(T(y,u)\,,\,p\big)_{\mathbb{R}^n}$.

1.4.6. Discussion of the variational inequality. In later chapters the admissible set U_{ad} will be defined by upper and lower bounds, so-called *box constraints*. We assume this here too, i.e.,

$$U_{ad} = \{u \in \mathbb{R}^m : u_a \le u \le u_b\}. \tag{1.13}$$

Here, $u_a \le u_b$ are given vectors in $\mathbb{R}^m$, where the inequalities are to be understood componentwise, that is, $u_{a,i} \le u_i \le u_{b,i}$ for $i = 1, \dots, m$. Rewriting the variational inequality (1.8) as

$$\left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \,,\, \bar{u}\right)_{\mathbb{R}^m} \le \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \,,\, u\right)_{\mathbb{R}^m} \quad \forall\, u \in U_{ad},$$

we find that $\bar{u}$ solves the linear optimization problem

$$\min_{u \in U_{ad}} \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u}) \,,\, u\right)_{\mathbb{R}^m} = \min_{u \in U_{ad}} \sum_{i=1}^{m} \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_i u_i.$$

If U_{ad} is given as in (1.13), then it follows from the fact that the u_i are independent from each other that

$$\left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_i \bar{u}_i = \min_{u_{a,i} \le u_i \le u_{b,i}} \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_i u_i$$

for $i = 1, \dots, m$. Hence, we must have

$$\bar{u}_i = \begin{cases} u_{b,i} & \text{if} \quad \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_i < 0 \\ u_{a,i} & \text{if} \quad \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_i > 0. \end{cases} \tag{1.14}$$

No direct information can be recovered from the variational inequality for the components that satisfy $\left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_i = 0$. However, in many cases useful information can still be extracted simply from the fact that this equation holds.

1.4.7. Formulation as a Karush–Kuhn–Tucker system. Up to now, the Lagrangian L has only been used to eliminate the conditions in equation form. The same can be done with the inequality constraints induced by U_{ad}. To this end, we introduce the quantities

$$\begin{aligned} \mu_a &:= \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_+ \\ \mu_b &:= \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_-. \end{aligned} \tag{1.15}$$

We have $\mu_{a,i} = \left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_i$ if the right-hand side is positive, and $\mu_{a,i} = 0$ otherwise; likewise, $\mu_{b,i} = \left|\left(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\right)_i\right|$ for a negative right-hand side, and $\mu_{b,i} = 0$ otherwise. Invoking (1.14), we deduce the relations

$$\begin{array}{rcrcrcl} \mu_a & \ge & 0, & u_a - \bar{u} & \le & 0, & \left(u_a - \bar{u} \,,\, \mu_a\right)_{\mathbb{R}^m} = 0, \\ \mu_b & \ge & 0, & \bar{u} - u_b & \le & 0, & \left(\bar{u} - u_b \,,\, \mu_b\right)_{\mathbb{R}^m} = 0. \end{array}$$

In optimization theory, these are usually referred to as *complementary slackness conditions* or *complementarity conditions.*

The inequalities hold trivially, so that only the equations have to be verified. We confine ourselves to showing the first orthogonality condition: in view of (1.14), the strict inequality $u_{a,i} < \bar{u}_i$ can only be valid if $\big(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\big)_i \leq 0$. By definition, this implies that $\mu_{a,i} = 0$, hence $(u_{a,i} - \bar{u}_i)\, \mu_{a,i} = 0$. If $\mu_{a,i} > 0$, then, owing to the definition of μ_a, also $\big(B^\top \bar{p} + \nabla_u J(\bar{y}, \bar{u})\big)_i > 0$, and from (1.14) we conclude that $u_{a,i} = \bar{u}_i$. Again, it follows that $(u_{a,i} - \bar{u}_i)\, \mu_{a,i} = 0$. Summation over i then yields $(u_a - \bar{u}\,,\, \mu_a)_{\mathbb{R}^m} = 0$.

Note that (1.15) implies that $\mu_a - \mu_b = \nabla_u J(\bar{y}, \bar{u}) + B^\top \bar{p}$, so that

$$\nabla_u J(\bar{y}, \bar{u}) + B^\top \bar{p} - \mu_a + \mu_b = 0. \tag{1.16}$$

We now introduce an extended Lagrangian $\mathcal{L}$ by adding the inequality constraints in the following way:

$$\begin{aligned} \mathcal{L}(y, u, p, \mu_a, \mu_b) \;:=\; & J(y,u) - \big(A\,y - B\,u\,,\, p\big)_{\mathbb{R}^n} + \big(u_a - u\,,\, \mu_a\big)_{\mathbb{R}^m} \\ & + \big(u - u_b\,,\, \mu_b\big)_{\mathbb{R}^m}. \end{aligned}$$

Then (1.16) can be expressed in the form

$$\nabla_u \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0.$$

Moreover, the adjoint equation is equivalent to the equation

$$\nabla_y \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0,$$

since $\nabla_y L = \nabla_y \mathcal{L}$. Hence, μ_a and μ_b are the Lagrange multipliers corresponding to the inequality constraints $u_a - u \leq 0$ and $u - u_b \leq 0$. The optimality conditions can therefore by rewritten in the following alternative form.

Theorem 1.4. *Suppose that A is invertible, U_{ad} is given by* (1.13)*, and $\bar{u}$ is an optimal control for* (1.1) *with associated state $\bar{y}$. Then there exist Lagrange multipliers $\bar{p} \in \mathbb{R}^n$ and $\mu_i \in \mathbb{R}^m$, $i = 1, 2$, such that the following conditions hold:*

$$\boxed{\begin{aligned} & \nabla_y \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0 \\ & \nabla_u \mathcal{L}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b) = 0 \\ & \mu_a \geq 0, \quad \mu_b \geq 0 \\ & \big(u_a - \bar{u}\,,\, \mu_a\big)_{\mathbb{R}^m} = \big(\bar{u} - u_b\,,\, \mu_b\big)_{\mathbb{R}^m} = 0. \end{aligned}}$$