Constrained Optimization In Finite and Infinite Dimensional Spaces

April 2, 2019

Abstract

Acknowledgements The sections on finite–dimensional optimization are taken from lecture notes by P. Spellucci, Einführung in die Optimierung, TU Darmstadt und P. Spellucci, Numerische Verfahren der nichtlinearen Optimierung, Birkhäuser Verlag. Further parts of these lecture notes are taken from the following books or lecture notes. G. Luenburger, Optimization in Vector Spaces, McGrawHill; F. Tröltzsch, Optimierung mit partiellen Differentialgleichungen, Vieweg; A. Bressan, Lecture notes on Optimal control and HJB equations, Penn State; S. Ulbrich, M. Ubrich, M. Hinze, R. Pinnau, Lecture notes to the Autumn school on Optimization with PDEs, Hamburg; Chapter 10 is taken mainly from notes of S. Steffensen, Numerical Optimization, RWTH Aachen, SS10.

Contents

1 Introduction

The theory of optimization presented is derived from a few simple, intuitive, geometric relations. The extension of these relations to infinite-dimensional spaces is the motivation for the mathematics of functional analysis which, in a sense, often enables us to extend our three-dimensional geometric insights to complex infinite–dimensional problems. This is the conceptual utility of functional analysis. On the other hand, these simple geometric relations have great practical utility as well because a vast assortment of problems can be analyzed from this point of view. We give a few examples.

The projection theorem. This theorem is one of the simplest results of optimization theory. In ordinary three-dimensional Euclidean space, it states that the shortest line from a point to a plane is the perpendicular from the point to the plane. This theorem has direct extensions to the infinite-dimensional Hilbert spaces and in the generalized form, this optimization principle forms the basis of all least–squares approximations, control and estimation procedures.

The Hahn–Banach Theorem. The theorem takes many forms and is one of the main theorems on which most of the following theory is based upon. One version of the theorem extends the projection theorem to problems having non–quadratic objectives. In this manner the simple geometric interpretation is preserved. Another form states that a given sphere and a point not in the sphere can be separated by a hyperplane.

Duality. There are several principles in optimization theory relying on the fact that certain facts can be represented by vectors as well as by hyperplanes (vectors in $\mathbb{R}^{1\times n}$). Many duality relations are based on a geometric relation illustrated below. Consider a linear subspace M, e.g. $\{x \in \mathbb{R}^2 : x_1 = 0\}$ and a point $x \in \mathbb{R}^2$. Then, the shortest distance from x to M in the Euclidean norm can also be expressed in the dual. Consider the dual to M, which is the orthogonal complement $M^T := \{x_2 = 0\}$. Then, the minimum distance from x to M coincides with the maximum of $\langle x, x^*\rangle$ taken over all $x^* \in M^T$ with $\|x^*\| \leq 1$. In 2d this can be seen by the fact that the scalar product can be expressed as $\cos\theta\|x\|\|x^*\|$ and its length is the projection of x to M^T if $\|x^*\| = 1$. In other words, minimizing the distance over vectors is maximizing over hyperplanes. This holds also true when considering minimizing the distance to a convex set which is equivalent to maximizing the distance of hyperplanes (between the point and the set) to this point.

Differentials. The most familiar optimization technique is setting the derivative to zero. The geometric interpretation is obvious for a one–dimensional

graph. At an extremum the tangent to the graph is horizontal. Similarly, in higher dimensions the tangent hyperplane is horizontal.

2 Preliminary discussion

2.1 Intent

The theory of constrained optimization in finite dimensional spaces is well known. The question which constraint qualifications to put, what the necessary and sufficient condition are, and so forth are treated in all details. An exhaustive reference is for example [13]. However, in most applications we are forced to consider continuous problems. Those problems are given in general infinite dimensional space. Naturally, one can ask the same question as in the finite dimensional space. An introduction to this theory is given in [10]. The intent of this lecture is to take something from both approaches and focus mainly on constraint qualifications and problem definitions, that can be translated in the infinite dimensional setting. Often the examples are given in the finite dimensional space or are related to partial differential equations. We given applications of the theorem and proof existence in the case of linear quadratic optimal control problems with elliptic or parabolic constraints. This follows closely[9]. We also consider abstract numerical methods for the optimization problems.

Further we discuss optimal control problems. This class of problems is very common in the applications and can be seen as optimization problems with additional structure information. The underlying theory for both is the same, but their are differences in the way of solving the problcems.

Further references are [14, 5, 11, 3].

2.2 Motivation

The theory of finite dimensional spaces can be extended to an infinite dimensional setting. First, we give a short formal calculation showing the strong relation between both approaches.

Consider $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^m$

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } h(x) = 0 \tag{2.1}$$

Assume sufficient constraint qualifications hold, e.g. $\nabla h(x)$ has full rank. Then the Karush-Kuhn-Tucker system is necessary and we have

$$\nabla f(x) - \nabla h(x)\lambda = 0 \tag{2.2}$$
$$h(x) = 0 \tag{2.3}$$

The Lagrange function is given by

$$L(x, \lambda) = f(x) - \lambda^T h(x) = f(x) - (\lambda, h(x))_m \tag{2.4}$$

If we read the equations carefully, we note that, $h(x) = 0$ is an equality in the space $\mathbb{R}^m$. Therefore, the following is true

$$h(x) = 0 \Leftrightarrow \forall \phi \in \mathbb{R}^m : (h(x), \phi)_m = (\phi, h(x))_m = 0 \tag{2.5}$$

This looks similar to a weak formulation of a partial differential equation in Hilbert spaces. Furthermore,

$$\nabla f(x) - \nabla h(x)\lambda = 0 \Leftrightarrow \forall \psi \in \mathbb{R}^n : (\nabla f(x), \psi)_n - (\nabla h(x)\lambda, \psi)_n = 0 \tag{2.6}$$
$$\Leftrightarrow \forall \psi \in \mathbb{R}^n : (\nabla f(x), \psi)_n - (\lambda, Dh(x)\psi)_m = 0 \tag{2.7}$$

Reformulation of the last equation yields

$$\forall \psi \in \mathbb{R}^n : \nabla_x L(x, \lambda)[\psi] = 0 \tag{2.8}$$

The equation (2.7) is the adjoint equation and (2.5) is the state equation. Solving both for (λ, x) gives (under additional assumptions) a solution for the optimization problem. Since all terms are scalar products in $\mathbb{R}^n$ or $\mathbb{R}^m$ we might generalize the above to abritrary Hilbert spaces. Furthermore, replacing the scalar product by a general duality product we can give a meaning to the equations even in Banach spaces.

We give an example of the problems, which we have in mind.

$$\min_{u \in H^1(\Omega)} \int_\Omega (u - u_d)^2 dx \tag{2.9}$$
$$\text{subject to } -\Delta u + u = 0 \text{ in } \Omega, \nabla u \cdot n = 0 \text{ on } \partial\Omega \tag{2.10}$$

We have to give a meaning to $\Delta u = 0$ in $H^1(\Omega)$:

$$\forall \phi \in H^1(\Omega) : (\nabla u, \nabla \phi)_2 + (u, \phi)_2 = (u, \phi)_1 = 0 \tag{2.11}$$

Now, compare the last equation with the state equation (2.5). At least formally we introduce the Lagrange function

$$L(u, \lambda) = f(u) - (\lambda, u)_1 = \int_\Omega (u - u_d)^2 dx - \int \nabla\lambda \cdot \nabla u dx - \int \lambda u dx \tag{2.12}$$

The adjoint equation is given by $\nabla_u L(u,\lambda)[\psi] = 0$:

$$\int_\Omega 2(u-u_d)\psi - \nabla\lambda\cdot\nabla\psi - \lambda\psi dx = 0 \tag{2.13}$$

This is the weak form of

$$2(u-u_d) = -\Delta\lambda + \lambda \text{ in } \Omega,\ \nabla\lambda\cdot n = 0 \text{ on } \partial\Omega \tag{2.14}$$

Naturally, the following questions arise.

1. What are the constraint qualifications?
2. Can we solve constrained optimization problems in general Banach spaces?
3. $\cdots$

2.3 Control and state variables

Difference between optimization and optimal control problems.

Common problems in optimization theory consider a split of the optimization variables. Usually, a state and a control are given. First, we illustrate this for a finite dimensional example. However, the motivation is the infinite dimensional setting:

$$\min_{u,y} J(u,y) \text{ subject to } -\Delta y = f + u,\ \nabla y\cdot n = 0 \tag{2.1}$$

In the finite dimensional setting we consider the problem

$$\min_{y,u\in\mathbb{R}^n} f(u,y) \text{ subject to } Ay = u \tag{2.2}$$

Of course, one can extend the following arguments to the case $u \in \mathbb{R}^m$ with $m < n$. If we assume, that $A \in \mathbb{R}^{n\times n}$ is invertible, then we can also study the reduced (unconstrained) problem

$$\min_{u\in\mathbb{R}^n} \tilde{f}(u) \tag{2.3}$$

Herein, $\tilde{f}(u) = f(u, y(u))$ and $y(u) = A^{-1}u$.

We reformulate the former problem for $x = (u, y)$ with $m = 2n$

$$\min_{x\in\mathbb{R}^m} f(x) \text{ subject to } (-Id, A)x = 0 \tag{2.4}$$

Since Id has rank n, the KKT equations are necessary

$$\begin{bmatrix} \nabla_u f(u,y) \\ \nabla_y f(u,y) \end{bmatrix} - \begin{bmatrix} -Id \\ A^T \end{bmatrix} \lambda = 0 \tag{2.5}$$

$$(-Id, A) \begin{bmatrix} u \\ y \end{bmatrix} = 0 \tag{2.6}$$

Rewriting this system we obtain

$$\text{state equation} \qquad Ay = u \tag{2.7}$$

$$\text{adjoint equation} \qquad \nabla_y f(u,y) - A^T \lambda = 0 \tag{2.8}$$

$$\text{gradient equation} \qquad \nabla_u f(u,y) + \lambda = 0 \tag{2.9}$$

Compared to the names and notation of the previous section, there is a slight abuse here. Before the gradient and adjoint equation could be separated. We called both adjoint equation. But, the above notation seems to be reasonable, if we consider the reduced functional. The minimization problem is unconstrained and therefore the necessary condition is

$$\nabla_u \tilde{f}(u) = 0 \tag{2.10}$$

Evaluating the gradient gives

$$D_u \tilde{f}(u) = D_u f(u,y) + D_y f(u,y) D_u y \text{ or} \tag{2.11}$$

$$\nabla_u \tilde{f}(u) = \nabla_u f + A^{-T} \nabla_y f(u,y) \tag{2.12}$$

In theory we can compute $D_u y$ and A^{-T} by differentiating $y = A^{-1}u$. However, there is a more clever way to do it. Solving the adjoint equation for λ and insert this we see

$$\nabla_u \tilde{f}(u) = \nabla_u f(u,y) + \lambda \tag{2.13}$$

and the gradient equation coincides with the necessary condition for the unconstrained optimization problem. So the optimality system coincides and by the adjoint equation, it is not necessary to compute $D_u y$. Similar considerations hold in the nonlinear case.

Summarizing, in the control and state case can be reduced to the previous discussion. Splitting of the equations yield three different equations closely related to the reduced functional.

2.4 Examples

ex01

Consider two domains $\Omega_r, \Omega_s \subset \mathbb{R}^2$ connected by the boundary Γ. We consider the problem

$$\partial_t u = D_r \Delta u \text{ in } \Omega_r \tag{2.1a}$$

$$\partial_t u = D_s \Delta u \text{ in } \Omega_s \tag{2.1b}$$

$$u = 0 \text{ on } (\partial\Omega_r \cup \partial\Omega_s) \cap \Gamma \tag{2.1c}$$

We measure the following quantity on Γ :

$$f_\Gamma(x) := k_r \partial_n u - k_s \partial_n u \; \forall x \in \Gamma \tag{2.2}$$

We assume that Γ is parametrized as follows

$$\Gamma := \{(x, y) : y = \gamma(x)\} \tag{2.3}$$

for some function $\gamma \in C^1(\mathbb{R})$. Then we would like to find the boundary Γ (or equivalently a parametrization γ), such that the functional J given below is minimized.

$$J(\gamma, u) = \int_0^T \int_\Gamma (f_\Gamma - \bar{f})^2 dS dt + \int_0^T \int_\Omega u^2 dx dt \tag{2.4}$$

subject to equations (2.1). we offer an interpretation for this problem. The equations (2.1) are the heat equations with for two different materials. The boundary Γ gives the structure of the final profile.

Example 2.1. *The integration of p-dimensional manifolds in the $\mathbb{R}^n$ can be computed as follows. Assume $f : \mathbb{R}^n \to \mathbb{R}$ and $\gamma : \mathbb{R}^p \to \mathbb{R}^n$ is a parametrization of the p-dimensional manifold M. Then we define the integral over M as*

$$\int_M f(x) dS := \int_{\mathbb{R}^p} f(\gamma(y)) \sqrt{\det(g_{ij})_{ij}} dx \tag{2.5}$$

wherein g_{ij} is the Gram matrix and is given by

$$g_{ij} = \partial_j \gamma \cdot \partial_i \gamma \tag{2.6}$$

The most important case is

$$M := \{(x, y, z) : z = \phi(x, y)\}. \tag{2.7}$$

In this case $\gamma : \mathbb{R}^2 \to \mathbb{R}^3$ *is given by* $\gamma(x, y) = (x, y, \phi(x, y))$ *and*

$$(g_{ij})_{ij} = \begin{bmatrix} 1 + \phi_x^2 & \phi_x \phi_y \\ \phi_y \phi_x & 1 + \phi_y^2 \end{bmatrix} \tag{2.8}$$

The transformation reads

$$\int_M f(x) dS = \int_{\mathbb{R}^2} f(x, y, \phi(x, y)) \sqrt{1 + \phi_x^2 + \phi_y^2} dx dy \tag{2.9}$$

2.5 The linear case - dual problems

Some remarks concerning linear optimization are included to show the close relation between both topics. A linear program in standard form is given by

$$\min c^T x \text{ subject to } Ax - b = 0 \tag{2.1}$$

Herein, slack variables may appear to reformulate the inequality constraints. We do not focus on that but refer to [13]. The problem is known as primal problem. Assuming constraint qualifications on A^T, the problem is a convex optimization problem, e.g. the objective function is convex and the constraints are affine linear. Therefore, the KKT system is necesary and sufficient for optimality. We introduce the Lagrange function

$$L(x, \lambda) = c^T x - \lambda^T (Ax - b) \tag{2.2}$$

Let us rewrite the Lagrange function in the following way

$$L(x, \lambda) = (c - A^T \lambda)^T x - b^T \lambda \tag{2.3}$$

In the latter formulation consider x as Lagrange multiplier and λ as optimization variable. Requiring that A fulfills a constraint qualification, we see, that the latter Lagrange function belongs to the following convex optimization problem

$$\min -b^T \lambda \text{ subject to } A^T \lambda = c \tag{2.4}$$

The problem (2.4) is also known as dual problem. If we rewrite the KKT system for (2.1) we obtain

$$c - A^T \lambda = 0 \tag{2.5}$$

$$Ax - b = 0 \tag{2.6}$$

By our previous notation the first equation is refered to as adjoint equation. The KKT system for (2.4) is given by

$$-b + Ax = 0 \tag{2.7}$$
$$-A^T\lambda + c = 0 \tag{2.8}$$

Assume that the $A \in \mathbb{R}^{m\times n}$. Let $rank(A^T) = m < n$, which implies that the LICQ condition is satisfied. Hence we have a unique multiplier λ for the primal KKT system. Obviously, LICQ is NOT satisfied for the dual problem. We would need $rank(A) = n$ but A has $m < n$ rows. Since the constraints are affine, the Abadie condition is satisfied. Therefore, there exists a multiplier x, s.t. the KKT for the dual is necessary. The multiplier might not be unique. No rank condition on ∇h is needed.

Conversly, if the problem is convex, i.e., f convex, h affin linear and the KKT system holds at (x^*, λ^*), then x^* is the global minimum.[1] If LICQ holds, then λ^* is unique and hence (x^*, λ^*) is unique.

We see, that in the linear the case the KKT system coincide. This means, that solving the dual problem is equivalent to solving the primal problem (if LICQ holds).

There are more results on this topic avaible, especially when including bound constraints. For sake of completness we state the main theorem of linear optimization. A proof can be found in [13].

Theorem 2.2. *Consider the problems*

$$\min c^T x \text{ subject to } Ax = b,\ x \geq 0 \tag{2.9}$$

and

$$\min -b^T y \text{ subject to } A^T y \leq c \tag{2.10}$$

If any of them has a finite minimum, then this holds also true for the other one. The optimal values coincide, i.e., $c^T x = -b^T y$. If the objective of any of both problem is unbounded, then the other is infeasible. If both problems are feasible, then there exists a optimal solution.

[1] Thm. 6.6 in [13]

3 Necessary and sufficient optimality conditions for unconstrained problems in finite dimensional spaces

n this section we are mainly concerned with the techniques for solving unconstrained optimization problems. Although it can be argued that such problems arise relatively infrequently in practice, the underlying ideas are so important that it is best to understand them first in their simplest setting.

The following is the setting of an unconstrained minimization:

Problem Setting:

Given $f: \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}; \quad f \in C^1(\mathcal{D}),\ \mathcal{D}$ open.

Aim: Find a (local) minimum x^* of f, i.e. there exists $\delta > 0$:
$f(x^*) \leq f(x)$ for all x with $\|x - x^*\| < \delta$.

The existence of a solution can only be guaranteed under certain requisite conditions (Counterexample: $n = 1,\ f(x) = \exp(x)$). A **sufficient condition for the existence** of at least one local minimum is the following:

There exists $x^0 \in \mathcal{D}: \ \mathcal{L}_f(x^0) = \{x \in \mathcal{D} : f(x) \leq f(x^0)\}$ compact (bounded and closed). i.e. the boundary of a region defined by a level surface is not also the boundary of $\mathcal{D}$. We will refer to such a set as a "level sphere" with respect to x_0.

Theorem 3.1. *If f is continuously differentiable in the neighborhood of x^* and x^* is a local minimum for f, then $f'(x^*) = 0$. (Necessary first-order condition)*

Theorem 3.2. *If f is $2k$-times continuously differentiable in the neighborhood of x^* and the following applies:*

$$f'(x^*) = 0, \ldots, \quad f^{(2k-1)}(x^*) = 0, \quad f^{(2k)}(x^*) > 0$$

then x^* is a strict local minimum of f, i.e. $f(x) > f(x^*)$ for all $x \neq x^*$ with $|x - x^*|$ sufficiently small. □

Sketch of Proof: Apply Taylor Series in x^* up to Order 1 with respect to $2k$. □

Theorem 3.3. *Let $f: \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}, \quad f \in C^1(\mathcal{D}),\ \mathcal{D}$ open, $x^* \in \mathcal{D}$. Suppose x^* is a local minimum of f, the following must necessarily hold*

$$\nabla f(x^*) = 0 \qquad \text{(necessary first-order condition)}$$

If $f \in C^2(\mathcal{D})$, then the following further holds:

$\nabla^2 f(x^*)$ *positive semi-definite* *(necessary second-order condition)*

□

Sketch of Proof: Taylor Series expansion up to first- and second- derivative. If $\nabla f(x^*) \neq 0$ then consider $x^* - \tau \nabla f(x^*)$ for small τ and in case $\nabla f(x^*) = 0$, but $\nabla^2 f(x^*)$ is not positive semi-definite consider $x^* - \tau z$, where z is a so called direction of negative curvature, i.e. $z^T \nabla^2 f(x^*) z < 0$ (for example, with z an eigenvector with the smallest algebraic eigenvalue).

Theorem 3.4. *Let $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$, $f \in C^2(\mathcal{D})$, $\mathcal{D}$ open, $x^* \in \mathcal{D}$. In case*

$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*)$ *positive definite*
(Sufficient second-order condition)

then x^* is a strict local minimum of f. □

Definition 3.5. *A symmetric real matrix A is called* **positive definite**, *if*

$$x^T A x > 0 \quad \text{for all } x \neq 0$$

and **positive semi-definite**, *if*

$$x^T A x \geq 0 \quad \text{for all } x.$$

□

Remark 3.6. *If $B \in \mathbb{R}^{m \times n}$ and Rank$(B) = n$, i.e. $Bx = 0$ only for $x = 0$, then $A = B^T B \in \mathbb{R}^{n \times n}$ is symmetric and positive definite.* □

We also need a practical algorithm that allows us to check positive definiteness of a matrix.

A sufficient condition, with which a local minimum of f is also a global minimum is the convexity of the function:

Definition 3.7. *$\mathcal{D} \subset \mathbb{R}^n$ is called* **convex**, *if $x \in \mathcal{D}$, $y \in \mathcal{D}$ then $[x, y] \subset \mathcal{D}$ also.*

D3 **Definition 3.8.** *$f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$, $\mathcal{D}$ convex, is said to be* **convex** *on $\mathcal{D}$, if for $x, y \in \mathcal{D}$*

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y) \quad \text{for } \lambda \in [0, 1] \tag{3.11}$$ 1

and **strict convex**, *if for $0 < \lambda < 1$ only "$>$" is true in (3.11).* □

S6 **Theorem 3.9.** *If $\mathcal{D} \neq \emptyset \subset \mathbb{R}^n$ is convex and f is convex on $\mathcal{D}$, then every local minimum is also a global minimum.* □

Sketch of Proof: Let x^* be a local minimum. Then there exists $\delta > 0$ such that

$$||y - x^*|| \ \leq \ \delta \ \Rightarrow \ f(x^*) \ \leq \ f(y) \ .$$

Suppose $x \in \mathcal{D}$ arbitrary. Consider $x^* + t(x - x^*)$. From the hypothesis

$$f(x^* + t(x - x^*)) \ \leq \ (1-t)f(x^*) + tf(x)$$

and

$$||(x^* + t(x - x^*)) - x^*|| \ \leq \ \delta \text{ in case } 0 \ < \ t \ < \ \delta/(||x|| + ||x^*||) \ .$$

This gives

$$f(x^*) \ \leq \ f(x^* + t(x - x^*)) \ \leq \ (1-t)f(x^*) + tf(x)$$

for such a t. A bit of manipulation and division by $t > 0$ gives the conclusion.

Theorem 3.10. *(Convexity Criteria) Let $\mathcal{D} \subset \mathbb{R}^n$ be convex, open ($\neq \emptyset$). If $f \in C^1(\mathcal{D})$, then*

1. *f is convex on $\mathcal{D} \Leftrightarrow f(y) \geq f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in \mathcal{D}$. If $f \in C^2(\mathcal{D})$, then in addition*
2. *f is convex on $\mathcal{D} \Longleftrightarrow \nabla^2 f(x)$ positive semi-definite on $\mathcal{D}$.*
3. *If $\nabla^2 f$ is positive definite on $\mathcal{D}$, then f is strictly convex on $\mathcal{D}$, i.e.*

$$\lambda f(x) + (1-\lambda)f(y) > f(\lambda x + (1-\lambda)y) \quad \text{for } 0 < \lambda < 1$$

and $x, y \in \mathcal{D}$ are arbitrary.

□

Corollary 3.11. *$f \in C^1(\mathcal{D})$, $\mathcal{D}$ is open convex, $x^* \in \mathcal{D}$, f is convex on $\mathcal{D}$, $\nabla f(x^*) = 0 \Rightarrow x^*$ is a minimum.*

From Theorem 3.3 and Theorem 3.10 the proof of the previous corollary follows easily. In theorem 3.9 the existence of a local minimum is assumed. The following criterion guarantees the **existence and uniqueness** of a local and at the same time a global minimum.

S8 **Theorem 3.12.** *Suppose $f \in C^2(\mathcal{D})$. Suppose $\mathcal{D}$ is open and convex ($\neq \emptyset$). If the following holds:*

(i) There exists $\alpha_0 \in \mathbb{R}$, such that $\mathcal{L}_f(\alpha_0) = \{x \in \mathcal{D} : \ f(x) \leq \alpha_0\}$ is bounded and closed or $\mathcal{D} = \mathbb{R}^n$.

(ii) $d^T \nabla^2 f(x) d \geq \alpha d^T d$ with $\alpha > 0$ for all $x \in \mathcal{D}$ and $d \in \mathbb{R}^n$,

then there exists exactly one strict local minimum of f on $\mathcal{D}$, which is also a global minimum. □

The first case we conclude that a continuous function attains its maximum on a closed and bounded, i.e., compact, set. Therefore, existence is guaranteed. In the second case we have that f is positive definite and uniformly convex. In the case $\mathcal{D} = \mathbb{R}^n$ we note that strict convexity is not(!) sufficient for f ti have a global minimum, see example 1 below.

Definition 3.13. (Extension of Definition 3.8) *f is called uniformly convex on the convex set $\mathcal{D} \in \mathbb{R}^n$, if there exists a $\gamma > 0$ with*

$$tf(x) + (1-t)f(y) \geq f(tx + (1-t)y) + t(1-t)\gamma \|x - y\|^2$$

for all $t \in [0,1]$ and $x,\ y \in \mathcal{D}$. (This is the case in the assumption of the theorem 3.12 above)

Example 3.14. *1. $f(x) = \exp(-x_1) + \exp(-x_2)$ is strictly convex on $\mathcal{D} = \mathbb{R}^2$. There exists* **no** *local minimum. (But: $\nabla^2 f(x) = \begin{pmatrix} \exp(-x_1) & 0 \\ 0 & \exp(-x_2) \end{pmatrix}$ is positive definite!)*

2. $f(x) = \frac{3}{2}((x_1)^2 + (x_2)^2) + \sin x_1 \cdot \sin x_2 + 3x_1 - 4x_2$.

$$\nabla^2 f(x) = \begin{pmatrix} 3 - \sin x_1 \sin x_2 & \cos x_1 \cos x_2 \\ \cos x_1 \cos x_2 & 3 - \sin x_1 \sin x_2 \end{pmatrix}$$

$\alpha = 1$ in (ii), theorem 3.12. $\mathcal{D} = \mathbb{R}^2$. Also there exists exactly one local critical point, which is the only local and at the same time the global minimum.

3. $f(x) = 1 + x_1 \ln x_1 + x_2 \ln x_2 + (x_1)^2 + (x_2)^2$
$\mathcal{D} = \{x : \ x_1 > 0,\ x_2 > 0\}$
$x^0 = (0.25, 0.25);\ f(x^0) = 0.4318528194$

$\mathcal{L}_f(f(x^0))$ is compact, since $f \to \infty$ for $x_1 \to \infty$ or $x_2 \to \infty$ and on

the boundary of $\mathcal{D}$ $f \geq 0.7148671412$.
(That is the minimum value of $1 + z \ln z + z^2, \quad z \geq 0$*).*

$$\nabla^2 f(x) = \begin{pmatrix} 2 + \frac{1}{x_1} & 0 \\ 0 & 2 + \frac{1}{x_2} \end{pmatrix}$$

indeed $\alpha = 2$ *in theorem 3.12.*

4 Necessary and sufficient optimality conditions for constrained problems in finite dimensional spaces

We consider the general nonlinear constrained optimization problem.

$$\left.\begin{array}{rcl} \min_{x \in \mathfrak{S}} & f(x) & \\ \mathfrak{S} & = & \{x \in \mathbb{R}^n : \ g(x) \geq 0, \ h(x) = 0\} \quad \mathfrak{S} \text{ "admissible set"} \end{array}\right\} \text{NLO}$$

i.e. for example,

$$g : \mathbb{R}^n \to \mathbb{R}^m; \ g(x) \geq 0 \text{ implies } g_i(x) \geq 0 \ \forall i = 1, \ldots, m$$

The problem will be studied under the following assumptions which are assumed to be valid throughout the remaining chapter.

A1 $\mathfrak{S} \neq \emptyset$.

The problem should be formulated such that there is at least an admissible point. This can be difficult to check in applications.

A2 f, g, h are defined on an open set $\mathcal{D} \subset \mathbb{R}^n$ and $\mathfrak{S} \subset \mathcal{D}$. Further, we assume that $\mathfrak{S}$ is a closed subset of $\mathbb{R}^n$.

If $\mathfrak{S}$ is closed, then $\mathfrak{S}$ is described by g and h only. Since $\mathfrak{S} \subset \mathcal{D}$ we can leave the admissible set without leaving the domain of definition for f, g and h. This is important for algorithms based on penalty functions. The requirement $\mathcal{D}$ allows to define the derivative f, g and h for all points $x \in \mathcal{D}$. In most cases we will consider $\mathcal{D} = \mathbb{R}^n$.

A3 $f, g, h \in C^1(\mathcal{D})$.

To derive sufficient conditions we sometimes require $f, g, h \in C^2(\mathcal{D})$.

In the following discussion we set

$$f, g, h \in C^1\ (\mathbb{R}^n), \quad g:\ \mathbb{R}^n \to \mathbb{R}^m, \quad h:\ \mathbb{R}^n \to \mathbb{R}^p$$

Some remarks are in order.

1. The previous notation implies to have m inequality and p equality constraints.

   Example 4.1. *(a)* $S = \{x \in \mathbb{R}^n : a_i \leq x_i \leq b_i,\ \forall i = 1, \ldots, n\}$

fig:feasible_set

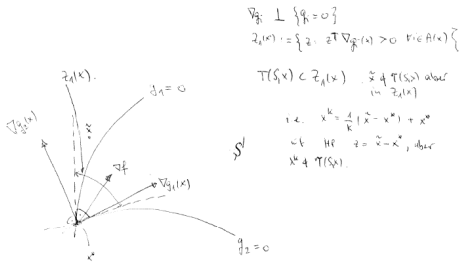


Figure 1: The feasible set.

   (b) Linear Programme:

$$\begin{aligned} f(x) &= c^T x,\ H \in \mathbb{R}^{p \times n} \\ h(x) &= Hx + b_0 \\ g(x) &= x. \end{aligned}$$

   (c) Quadratic Programme:

$$\begin{aligned} f(x) &= x^T Ax - b^T x; \\ h(x) &= Hx + h_0 \\ g(x) &= Gx + g_0. \end{aligned}$$

   (d) Lagrangian Multiplier Theorem: $g = 0$, $Dh(x)$ *has full rank.*

$$\min_x f(x) s.t. \quad h(x) = 0$$

:lagrangemultiplierthm

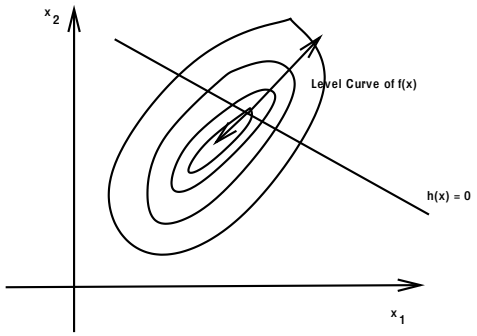


Figure 2: Constrained optimization using Lagrange Multiplier.

2. In the literature we also find the constraint

$$g(x) \leq 0$$

which can be reformulated in the above context.

3. If g_i is highly nonlinear it might be useful to reformulate the inequality constraint $g_i(x) \geq 0$ as equality constraint using slack variables x_{i+1} :

$$g_i(x) \geq 0 \Leftrightarrow g_i(x) - x_{i+1} = 0,\ x_{i+1} \geq 0.$$

This should be done only if the number of inequalities is small and if the equality constraint can be treated efficiently.

4. The following reformulations are **not** recommended since they effect the constraint qualification properties.

$$\begin{aligned} h_i(x) = 0 \quad &\Leftrightarrow \quad h_i(x) \geq 0 \quad \text{and} \quad h_i(x) \leq 0 \quad i = 1, \ldots, p \\ g_i(x) \geq 0 \quad &\Leftrightarrow \quad g_i(x) - (x_{n+i})^2 = 0, \quad i = 1, \ldots, m. \end{aligned}$$

($\nabla h_i(x)$ and $-\nabla h_i(x)$ are always linearly dependent, nonlinear constraints are harder to solve than linear ones.)

There are some difficulties when solving NLO numerically. Depending on the constraints the problem becomes more difficult. We distinguish the following cases.

a) Simplest case: $\mathfrak{S} = \{x \in \mathbb{R}^n :\ a \leq x \leq b\}, \quad a < b \in \mathbb{R}^n$ given ("box constraints")

b) Similar and efficiently to compute is the following case: h, g affin linear. The admissible set is a polyhedron. There are further special and important subcases:
$f(x) = c^T x \quad \mapsto$ linear programming
$f(x) = \frac{1}{2} x^T A x - b^T x \quad \mapsto$ quadratic programming.

c) h nonlinear, but ∇h has full column rank and no inequalities. Then we apply the implicit function theorem and consider an unconstrained problem.

d) f convex, h affin linear, g_i concave, $i = 1, \ldots, m$
$\mapsto$ "convex optimization"

Before discussing the general theory we have the following remark on the notation.

Remark 4.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be sufficiently smooth. We denote by*

$$\nabla f(x) = (\partial_{x_1} f(x), \partial_{x_2} f(x), \ldots, \partial_{x_n} f(x))^T$$

the gradient of f. This is a column vector. We denote by

$$(\nabla^2 f(x))_{ij} = (\partial_{x_j} \partial_{x_i} f)_{ij}$$

the (symmetric) Hessian matrix of f.

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be sufficiently differentiable. We denote by ∇g the transposed(!) Jacobi-matrix ($J_g(x)$) of $g = (g_1, \ldots, g_m)^T$, i.e.,

$$(\nabla g(x))_{ij} = (\partial_{x_i} g_j)_{i=1,\ldots,n; j=1,\ldots,m} = (\nabla g_1, \ldots, \nabla g_m)$$

Hence, the derivative $Dg(x) = \nabla g(x)^T$.

We often require $\nabla h(x)$ to have full column rank for $h : \mathbb{R}^n \to \mathbb{R}^p$. This is equivalent to assume that $\nabla h(x)$ has p linearly independent columns. Note that the jth column of $\nabla h(x)$ is $\nabla h_j(x)$. If h is linear, then $h = H$ with $H \in \mathbb{R}^{p \times n}$ and $\nabla h(x) = H^T \in \mathbb{R}^{n \times p}$

Example 4.3. *Consider $g(x) = Gx$, $G \in \mathbb{R}^{m \times n}$. Then $J_g(x) = G \in \mathbb{R}^{m \times n}$, and $\nabla g(x) = G^T \in \mathbb{R}^{n \times m}$*

4.1 Necessary optimality conditions

We discuss the characterization of minima to NLO under the assumptions $(A1) - (A3)$ of the previous section. We achieve to obtain a similar characterization as in the case of unconstrained optimization, i.e., necessary and

sufficient conditions involving derivatives of f and the constraints h and g. The most important result will be the Karush–Kuhn–Tucker necessary optimality conditions, see below. We do not discuss the existence of minimum in details. The existence is granted under various conditions. E.g., suppose that

$$\mathcal{L}_f(f(x^0)) \cap \mathfrak{S}$$

bounded for some $x^0 \in \mathfrak{S}$.

Definition 4.4. *$x^* \in \mathfrak{S}$ is a local solution (local minimum) to* NLO*, if there exists a neighborhood $U_\delta(x^*)$ such that for all $y \in U_\delta(x^*) \cap \mathfrak{S}$ we have $f(y) \geq f(x^*)$.*

Using this definition the most general necessary first–order conditions are due to Fritz John (1948).

S28 **Theorem 4.5.** *Let $x^* \in \mathfrak{S}$ be a local minimum of f on $\mathfrak{S}$.*

Then there exists multipliers $\lambda_0^*, \lambda_1^*, \ldots, \lambda_m^* \geq 0$ and $\mu_1^*, \ldots, \mu_p^* \in \mathbb{R}$ such that

$$\lambda_0^* \nabla f(x^*) - \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) - \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*) = 0$$
$$\lambda_i^* g_i(x^*) = 0 \qquad i = 1, \ldots, m$$

and such that $\|\lambda_0^*, \lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_p^*\| \neq 0$.

Note that the vector of multipliers is non–trivial due to the last condition. The assertion of this theorem is typically of no use in applications since the possibility $\lambda_0^* = 0$ is not excluded. If $\lambda_0^* = 0$, then the necessary condition does not involve the function f!

Example 4.6. *Example due to Kuhn–Tucker. Let $n = 2,\ m = 2,\ p = 0$*

$$f(x) = -x_1, \quad g_1(x) = x_2, \quad g_2(x) = (1 - x_1)^3 - x_2$$
$$x^* = \binom{1}{0}, \quad \lambda_0^* = 0, \quad \lambda_1^* = \alpha \in \mathbb{R}, \quad \lambda_2^* = -\alpha$$

The previous example shows that in general additional assumptions have to be imposed to guarantee $\lambda^* > 0$. These conditions are called **constraint qualifications**. We start with some general remarks and notations before proving the main theorem. The concept of cones is important for characterization of a local minimum. This is due to the following theorem, see also figure 4.1

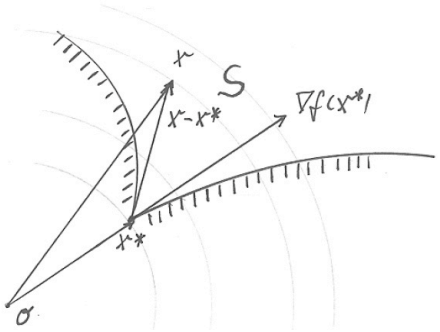
co-thm1 **Theorem 4.7.** *Let x^* be a local minimum. Let $x^k \in U_\delta(x^*) \cap \mathfrak{S}$ be a convergent sequence with limit x^* and such that $x^k \neq x^*$. Let z be any accumulation point of the sequence*

$$\frac{x^k - x^*}{\|x^k - x^*\|}.$$

Then,

$$\nabla f(x^*)^T z \geq 0. \tag{4.12}$$

eqn:unconstrainedneces



convergencexk

Figure 3: Illustration of theorem 4.7

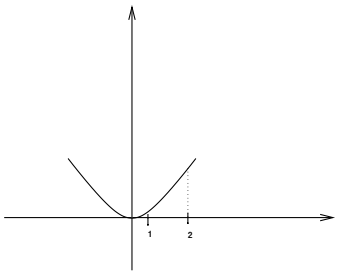
Proof. We have $\alpha^k = \|x^k - x^*\| \to 0$ due to the assumptions and $x^k = x^* + \alpha z^k$ for $z^k := (x^k - x^*)/\alpha^k$. Since z is an accumulation there exists a convergent subsequence $z^k \to z$. Since f is continuously differentiable we have $f(x^k) = f(x^*) + \alpha^k \nabla f(x^*)^T z^k + \alpha^k \epsilon^k$ for ϵ^k with the property $\epsilon^k \to 0$ for $k \to \infty$. Since x^* is a local minimum we have

$$\begin{aligned} f(x^*) \leq f(x^k) = f(x^*) + \alpha^k \left(\nabla f(x^*)^T z^k + \epsilon^k \right) \\ \implies \nabla f(x^*)^T z^k + \epsilon^k \geq 0 \; \forall k \end{aligned}$$

□

$\min_x f(x) = x^2$ subject to $1 \leq x \leq 2$. In this case the minimizer $x^* = 0$ and $f\prime(x^*) = 0$. But in the interval given the minimizer is $x^* = 1$ (sketch graph of function). Can you prove that $x^* = 1$ is really the constrained minimizer? The theorem says $f(x^*)z \geq 0$ for $z = \dfrac{x^k - x^*}{\|x^k - x^*\|}$, $1 \leq x^k \leq 2$, $x^k \to 1$ for $k \to \infty$. We guess $x^* = 1$, $x^k = 1 + \dfrac{1.5}{k}$. Thus $x^k \to 1$ as $k \to \infty$. Hence $z = \dfrac{1 + \frac{1.5}{k} - 1}{\|1 + \frac{1.5}{k} - 1\|} = 1$. We conclude $f'(1) \cdot 1 \geq 0 \Rightarrow 2 \geq 0$.

fig:para_constrain



Example 4.8. *(a)*

Figure 4: Minimising a quadratic function on an interval.

(b)

$$\min_{a \le x \le b} f(x)$$

fig:necessary_cond

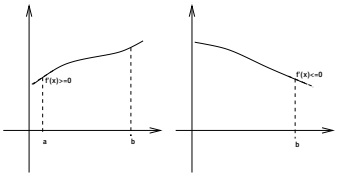


Figure 5: Necessary conditions for a constrained minimum.

$$f'(x^*) \cdot 1 \ge 0 \quad \textit{necessary condition - necessary if } x^* = a$$
$$f'(x^*) \cdot -1 \ge 0 \quad \textit{necessary condition - necessary if } x^* = b$$
$$f'(x^*) = 0 \quad \textit{necessary condition - necessary if } a < x^* < b$$

Note in the above theorem there will be a sequence $x^k \to x^*$ *(e.g.* $x^k = x^* + \frac{1}{k}$*) and an accumulation point* z *of the sequence* $\dfrac{x^k - x^*}{\|x^k - x^*\|}$.

$$\min_{(x_1,x_2)} x_1^2 + x_2^2 \quad \textit{subject to } 1 \le x_1 \le 2; \quad 1 \le x_2 \le 2.$$

Guess: $x^* = (1,1)^T$ *Question: Can you verify that it is a minimum?*

$$\nabla f(x^*)^T z \ge 0 \textit{ iff } f(x^*)^T \in \Pi'(\mathcal{S}, x^0)$$

Hence $\nabla f(x^*)^T \ge 0 \quad \forall z \in \Pi(\mathcal{S}, x^*)$. *If you guess* $x = (2,2)$ *you can verify that the condition does not hold.*

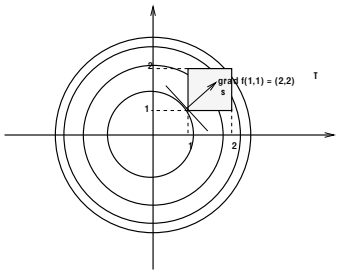


fig:paraboloid_constr

Figure 6: Necessary conditions for a constrained minimum.

Note that in the case $m = p = 0$ and for $x^k = x^* + \alpha^k \pm e_i$, $\alpha^k \to 0$ for any fixed i, we obtain $z^k \to z = \pm 1$ and $(\nabla f(x^*))_i = 0$. This yields the well–known first–order optimality condition $\nabla f(x^*) = 0$.

Equation (4.12) is the most general necessary condition which is known. However the condition can not be used for practical problem solving, for example, in numerical schemes.

Further note that if $\nabla f(x^*)^T z \geq 0$ for some z, then $\nabla f(x^*)^T(\tau z) \geq 0$ for $\tau \geq 0$. The set of all vectors τz belongs therefore to a cone. The cone consisting of all z satisfying the previous theorem and all τz for $\tau \in \mathbb{R}^+$ is the **tangential cone of $\mathfrak{S}$ at** x^*.

Definition 4.9. *$K \subset \mathbb{R}^n$ is a cone, iff for $z \in K$, we have $\tau z \in K$.*
$K' \subset \mathbb{R}^n$ is the dual cone to $K \subset \mathbb{R}^n$, iff

$$K' := \{y \in \mathbb{R}^n : y^T x \geq 0 \; \forall \; x \in K\}.$$

A cone and its dual is depicted in Figure 4.1.

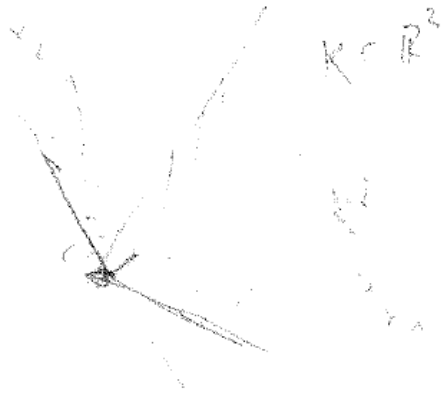
Definition 4.10. *Let $\mathcal{S} \subset \mathbb{R}^n$ and $y \in \mathcal{S}$. The closed* **tangential cone** *$\mathcal{T}(\mathcal{S}, y)$ is the cone consisting of all vectors which are given as differences $a - y$ for $a \in \mathcal{S}$ where a is sufficiently close to y and the accumulation points of such vectors, i.e.,*

$$\mathcal{T}(\mathcal{S}, y) := \cap_{k=1}^{\infty} \Pi_k$$

where

$$\Pi_k := \overline{\{\alpha(a - y) : \alpha \geq 0, a \in \mathcal{S}, \|a - y\| < \frac{1}{k}\}}.$$

The important observation is now that all points satisfying the previous theorem belong to the closed tangential cone on $\mathfrak{S}$ at x^* and vice versa. Hence, the tangential cone can be described by exactly those vectors z defined previously with $\alpha^k = \frac{1}{\|x^k - x^*\|}$.



cone

Theorem 4.11. *We have $z \in \mathcal{T}(\mathfrak{S}, x)$, if and only if there exists a sequence $\alpha^k > 0$ and a sequence $x^k \in \mathfrak{S}$ with $x^k \to x$ and such that*

$$\lim_k \alpha^k (x^k - x) = z.$$

Proof. Let $z \in \mathcal{T}(\mathfrak{S}, x)$. Then we have two cases. Either we have $z \in T_k$ for all k and hence $z = \alpha^k(x^k - x), \alpha^k \in \mathbb{R}^+, \|x^k - x\| \leq \frac{1}{k}$. Hence, $x^k \to x$. Or we have z as accumulation point of a sequence $z_k \in T_k$. Then, $z^k := \alpha^k(y^k - x), \alpha^k \in \mathbb{R}^+, \|y^k - x\| \leq \frac{1}{k}, \|z^k - z\| \leq \epsilon_k \to 0$. Hence, $y^k \to x$.

Let $x^k \to x$. Hence, for arbitrary p there exists k_0 such that $\|x^k - x\| \leq \frac{1}{p}$, $\forall k \geq k_0$ and hence $\alpha^k(x^k - x) \in T_p$ for all $k \geq k_0$. Since T_p is closed and since $\alpha^k(x^k - x) \to z$, we have $z \in T_p$. The previous arguments hold true for any p. Hence, $z \in \mathcal{T}(\mathfrak{S}, x)$. □

Applying the previous theorem to the sequence of theorem 4.7 with $\alpha^k = \frac{1}{\|x^k - x^*\|}$, we obtain the important necessary optimality condition

$$x^* \text{ local minimum of NLO} \implies \nabla f(x^*)^T z \geq 0 \; \forall z \in \mathcal{T}(\mathfrak{S}, x^*) \tag{4.13}$$

or equivalently using the definition of the dual cone

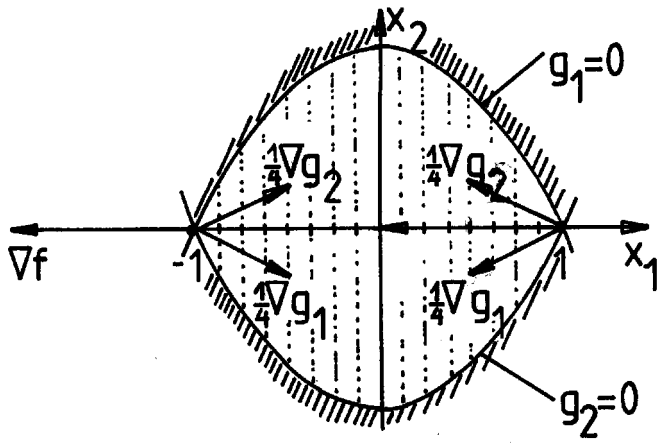
$$x^* \text{ local minimum of NLO} \implies \nabla f(x^*) \in \mathcal{T}(\mathfrak{S}, x^*)'. \tag{4.14}$$

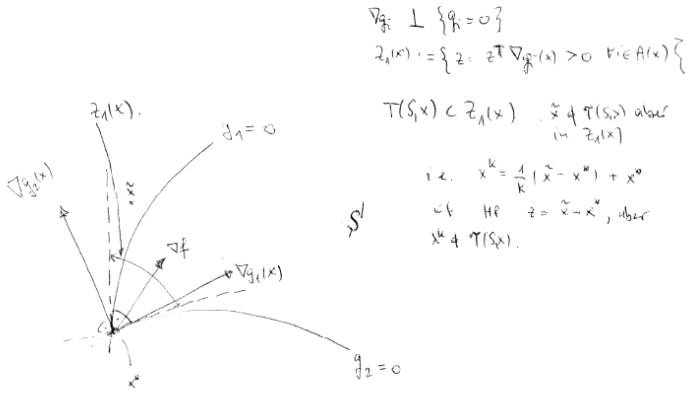
The previous characterization is unsatisfactory, since the tangential cone is in general not easy to construct. Further, the constraints g and h do not appear explicitly in the definition of the cone. The following constraint qualifications aim to describe the tangential cone in expressions of the Jacobi–matrices of the constraints. Under certain assumptions on g and h a com-

plete description of the cone is possible. These assumptions are called constraint qualifications. The cone Z_1 is obtained by the constraints g and h and is called **linearised cone**. The set $\mathcal{A}(x)$ is called the **active set**.

$$Z_1(x) := \{z \in \mathbb{R}^n : z^T \nabla g_i(x) \geq 0, i \in \mathcal{A}(x),\ z^T \nabla h_j(x) = 0, j = 1, \ldots, p\} \tag{4.15}$$ co-z1

$$\mathcal{A}(x) := \{i \in \{1, \ldots, m\} : g_i(x) = 0\} \tag{4.16}$$ co-a





The relation between $\mathcal{T}(\mathfrak{S}, x)$ and $Z_1(x)$ is seen below.

Theorem 4.12. *We have $\mathcal{T}(\mathfrak{S}, x) \subset Z_1(x)$ for all $x \in \mathfrak{S}$.*

Proof. For each $z \in \mathcal{T}(\mathfrak{S}, x)$ we have $x^k \in \mathfrak{S}, x^k \to x, \alpha^k(x^k - x) \to z$ for $\alpha^k \in \mathbb{R}^+$. Since $x^k \in \mathfrak{S}$ and since h_j is differentiable for each j we have

$$0 = h_j(x^k) = h_j(x) + (x^k - x)^T \nabla h_j(x) + \|x^k - x\| \epsilon_{k,j}, \epsilon_{k,j} \to 0.$$

Since $x \in \mathfrak{S}$ we have after multiplication with α^k

$$0 = \alpha^k (x^k - x)^T \nabla h_j(x) + \|\alpha^k (x^k - x)\| \epsilon_{k,j} \to z^T \nabla h_j(x).$$

Similarly, we have $z^T \nabla g_i(x) \geq 0$ for $i \in \mathcal{A}(x)$ since $g_i(x) = 0$. □

Unfortunately, we have $\mathcal{T}(\mathfrak{S}, x) \neq Z_1(x)$ without further assumptions. The weakest possible assumption is exactly assuming equality and yields Karush–Kuhn–Tucker theorem:

Theorem 4.13. *Let x^* be a local solution to* NLO *and assume (A1)-(A3). Assume the constraint qualification of Guignard*

$$Z_1(x^*)' = \mathcal{T}(\mathfrak{S}, x^*)'$$

holds.

Then there exists multipliers $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)$ with $\lambda_i^* \in \mathbb{R}_0^+$ and $\mu^* = (\mu_1^*, \ldots, \mu_p^*)$ with $\mu_i^* \in \mathbb{R}$ such that

co-kkt

$$\nabla f(x^*) - \nabla g(x^*)\lambda^* - \nabla h(x^*)\mu^* = 0, \tag{4.17a}$$
$$(\lambda^*)^T g(x^*) = 0 \tag{4.17b}$$
$$g(x^*) \geq 0 \tag{4.17c}$$
$$h(x^*) = 0 \tag{4.17d}$$

The proof is based on Farkas' Lemma which is a special case of the Hahn–Banach theorem. Farkas Lemma is as follows:

Lemma 4.14. *Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then we have either a solution r to $Ar = b$ with $r \geq 0$ or there exists a y with $A^T y \geq 0, b^T y < 0$*

From Figure 4.1 ist is clear that there are only to possibilities: Either the vector b is in the (positive) span of the columns of A or not. In the proof of the previous Theorem we have the first case.

Proof. Due to the previous discussion and Guignard assumptions we have $\nabla f(x^*) \in \mathcal{T}(\mathfrak{S}, x^*)' = Z_1(x^*)'$ or equivalently $z^T \nabla f(x^*) \geq 0$ for all

farkas

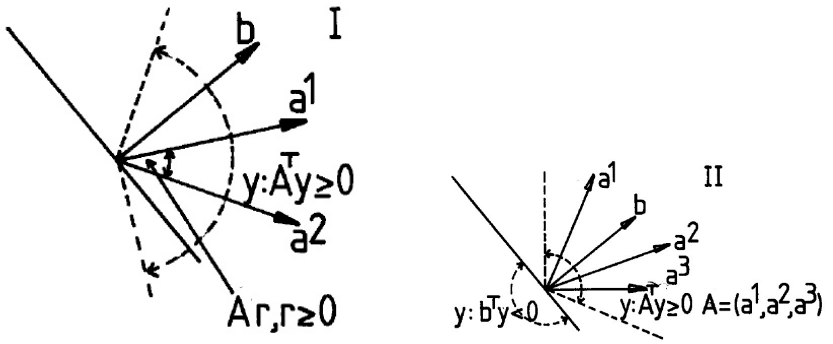


Figure 7: Illustration of Farkas' Lemma. Part (II) corresponds to the case where b belongs to the positive span of the columns of A and hence there exists $r \geq 0$ such that $Ar = b$. In part (I) the set $A^T y \geq 0$ and $b^T y < 0$ is non–empty and is depicted in the lower left part.

$z \in Z_1$. Define the complementary set $Z := \{y \in \mathbb{R}^n : y^T \nabla f(x^*) < 0\}$. Hence, $Z \cap Z_1(x^*) = \emptyset$. Rewriting the definition of $Z_1(x^*)$ we have

$$Z_1(x^*) = \left\{ z \in \mathbb{R}^n : \begin{pmatrix} \nabla g_i(x^*)^T z \geq 0, i \in \mathcal{A}(x^*) \\ \nabla h_j(x^*)^T z \geq 0, \forall j \\ -\nabla h_j(x^*)^T z \geq 0, \forall j \end{pmatrix} \right\}$$

We now apply Farkas Lemma with

$$A := \begin{pmatrix} \nabla g_{i \in \mathcal{A}(x^*)}(x^*) & \nabla h_{j:\forall j}(x^*) & -\nabla h_{j:\forall j}(x^*) \end{pmatrix} \in \mathbb{R}^{n \times |\mathcal{A}|+2m}$$

and

$$b = \nabla f(x^*).$$

It holds $A^T z \geq 0$ fpr $z \in \mathcal{Z}_1(x^*)$ by definition and for $z \in \mathcal{Z}_1(x^*)$ we have $\nabla f(x^*)^T z \geq 0$. Since we have $A^T z \geq 0 \implies b^T z \geq 0$, there exists a solution $0 \leq r \in \mathbb{R}^{|\mathcal{A}|+2m}$ with $Ar = b$ or written explicitly

$$\sum_{i=1}^{|\mathcal{A}|} r_i \nabla g_i(x^*) + \sum_{j=1}^{m} r_{j+|\mathcal{A}|} \nabla h_j(x^*) + \sum_{j=1}^{m} -r_{j+|\mathcal{A}|+m} \nabla h_j(x^*) = \nabla f(x^*).$$

Hence, the Karush–Kuhn–Tucker theorem holds true for the multipliers $\lambda_i^* \geq 0, i \in \mathcal{A}(x^*), \lambda_i^* = 0, i \notin \mathcal{A}(x^*)$ and $\mu_j^* = r_{j+|\mathcal{A}|} - r_{j+|\mathcal{A}|+m}$. □

Remark 4.15. *Condition (4.17) explicitly reads*

$$\nabla f(x^*) - \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) - \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*) = 0$$

and for $i = 1, \ldots, m$ $\lambda_i^* g_i(x^*) = 0$.

For $m = p = 0$ *we obtain again the well–known first–order condition* $\nabla f(x^*) = 0$.

For $m = 0$ *we obtain the classical Lagrange Multiplier theorem*

$$\nabla f(x^*) - \nabla h(x^*)\mu^* = 0.$$

Example 4.16. *Again we consider the example with the paraboloid above:*

$$\min_{(x_1,x_2)} x_1^2 + x_2^2 \quad \text{subject to } 1 \le x_1 \le 2; \quad 1 \le x_2 \le 2.$$

If x^* *is a local minimum then (with further conditions) it also satisfies the Karush-Kuhn-Tucker conditions:*

$$\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(x^*)$$

with

$$\begin{aligned} \lambda_i g_i(x^*) &= 0; \\ g_i(x^*) &\ge 0; \\ h_j(x^*) &= 0; \end{aligned}$$

where $\lambda_i \ge 0$. *To rewrite the optimization problem in this format we need to re-write the constraints,* $1 \le x_1 \le 2$, $1 \le x_2 \le 2$ *in the form:*

$$\begin{aligned} g_1(x_1, x_2) &= x_2 - 1; \\ g_2(x_1, x_2) &= 2 - x_1; \\ g_3(x_1, x_2) &= x_2 - 1; \\ g_4(x_1, x_2) &= 2 - x_2. \end{aligned}$$

Hence evaluating the system above for the specific problem given we ob-

tain:

$$\begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \lambda_3 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \lambda_4 \begin{pmatrix} 0 \\ -1 \end{pmatrix};$$
$$\lambda_1(x_1 - 1) = 0;$$
$$\lambda_2(2 - x_1) = 0;$$
$$\lambda_3(x_2 - 1) = 0;$$
$$\lambda_4(2 - x_2) = 0;$$
$$g_i(x) \geq 0.$$

We now verify the KKT for the case $x^* = (1, 1)$:

$$\begin{pmatrix} 2 \\ 2 \end{pmatrix} = \begin{pmatrix} \lambda_1 - \lambda_2 \\ \lambda_3 - \lambda_4 \end{pmatrix};$$
$$\lambda_2(2 - x_1) = 0;$$
$$\lambda_4(2 - x_2) = 0;$$

Hence $\lambda_2 = \lambda_4 = 0$, and $\lambda_1 = \lambda_3 = 2$. If we took $x^* = (1, 0)$ we obtain $\lambda_3 = \lambda_4 = 0$ but taking $\tilde{x} = (1, 2)$ gives $\lambda_1 = 2$ and $\lambda_4 = -4$, thus $\tilde{x}$ can never be a minimum since some λs are negative!

The condition of Guginard is the weakest possible condition in the following sense. Assume x^* is a local minimum and f is continuously differentiable. If and only if (4.17) holds in x^*, then $\mathcal{T}(\mathfrak{S}, x^*)' = Z_1(x^*)$. For a proof we refer to Gould und Tolle.

The remaining discussion focus on constraint qualifications which imply the Guginard condition. The other constraint qualification are typically simpler to verify. The most important constraint qualifications are summarized below.

Definition 4.17. *For $x \in \mathfrak{S}$ let $\mathcal{A} = \mathcal{A}(x) = \{i : \ g_i(x) = 0\}$ be the active set.*

$x^* \in \mathfrak{S}$ satisfies the **Mangasarian-Fromowitz-condition (MFCQ)**, *if*

$$\nabla h(x^*) \quad \text{has full column rank and} \quad \exists z \in \mathbb{R}^n : \nabla h(x^*)^T z = 0, \quad \nabla g_{\mathcal{A}}(x^*)^T z > 0$$

$x^* \in \mathfrak{S}$ satisfies the **regularity condition / linear independence constraint qualification (LICQ)**, *if*

$$(\nabla h(x^*), \nabla g_{\mathcal{A}}(x^*)) \text{ has full column rank.}$$

The **Slater condition** *for* NLO *is satisfied, if* $-g_i$ *is convex for all* $i = 1, \ldots, m$ *and* h_j *is affine linear for all* $j = 1, \ldots, p$ *and if there exists an* x^0 *such that*

$$h(x^0) = 0,\ g_L(x^0) \geq 0,\ g_{NL}(x^0) > 0,$$

where g_L *are all affine linear and* g_{NL} *are all other inequality constraints.*

For convenience we introduce the following additional notation. If $\mathcal{A} \subset \{1, \ldots, m\}$ is a set of indices, e.g., $\mathcal{A} = \{i_1, \ldots, i_s\}$, then we denote by $\nabla g_{\mathcal{A}}$ the matrix $(\nabla g_{i_1}, \ldots, \nabla g_{i_s})$. The Slater condition can only hold, if the admissible set $\mathfrak{S}$ is convex. The previous constraint qualifications all imply the condition of Guginard. The MFCQ is the weakest of the above given conditions and some conditions are related to each other. We state some important theorems on the conditions and their implications.

Theorem 4.18. *If* $x \in \mathfrak{S}$ *satisfies the MFCQ condition, then the condition of Guginard is satisfied.*

If $x \in \mathfrak{S}$ *satisfies the LICQ condition, then the condition of Guginard is satisfied.*

If NLO *satisfies the Slater condition, then the modified MFCQ condition holds. The modified MFCQ implies the Guignard condition.*

A consequence of the previous theorem and theorem 4.7 is the following result.

S29 **Theorem 4.19.** *Let* $x^* \in \mathfrak{S}$*. If* x^* *is a local minimum of* f *on* $\mathfrak{S}$ **and if** *either in* x^* *MFCQ or LICQ or if* NLO *satisfies the Slater condition, then there exists multipliers* $\lambda_1^*, \ldots, \lambda_m^* \geq 0, \quad \mu_1^*, \ldots, \mu_p^* \in \mathbb{R}$ *such that*

$$\nabla f(x^*) - \nabla g(x^*)\lambda^* - \nabla h(x^*)\mu^* = 0, \quad \lambda_i^* g_i(x^*) = 0, \quad i = 1, \ldots, m.$$

All constraint qualifications of the previous theorem are stronger than the Guignard condition. Hence, we can expect additional properties of the multipliers when requiring MFCQ or LICQ. Indeed, we have the following additional 'regularity results'.

Theorem 4.20. *Let* $x^* \in \mathfrak{S}$ *and let* x^* *be a local minimum of* f *on* $\mathfrak{S}$*.*

If and only if in x^* *MFCQ is satisfied, then the set of possible multipliers* (μ^*, λ^*) *in (4.17) is bounded (*Gauvin 1977*).*

If and only if in x^* *MFCQ is satisfied, then there exists neighborhoods* $U_\delta(0) \subset \mathbb{R}^p, V_\delta(0) \subset \mathbb{R}^m$ *and* $\gamma(x^*) > 0$ *such that the system*

$$h(x) = e^1, g(x) \geq e^2, \|x - x^*\| \leq \gamma(x^*) \left(\|e^1\| + \|(e^2)^+\| \right)$$

has a solution for $e^1 \in U_\delta(0), e^2 \in V_\delta(0)$ *(*Robinson 1976*).*

If in x^* *LICQ is satisfied, then* μ^* *and* λ^* *of (4.17) are uniquely defined.*

The condition MFCQ guarantees stability against perturbations of the constraints. All conditions only yield necessary optimality conditions. These conditions can also be formulated using the Lagrange function.

Other approaches use the Morse Lemma in order to study optimality conditions. We will see later that this corresponds to the LICQ condition. To this end we introduce the following definition and theorems for normal forms of order one and two.

Definition 4.21. *Let* U, V *be open sets in* $\mathbb{R}^n$ *and* $F : U \to V$ *bijective.* F *is called a* C^k*- diffeomorphism, iff* $F \in C^k(U; V)$ *and* $F^{-1} \in C^k(V, U)$*. If* F *and* F^{-1} *are continuous (k=0), then* F *is called a homeomorphism.*

If F is a homeomorphism and $\bar{x}$ a local minimum for f. Then, $\bar{y} = F(\bar{x})$ is a local minimum of $f(F^{-1}(y))$.

Theorem 4.22. *Let* $k \geq 1$ *and* $f \in C^k$*. Suppose that* $Df(\bar{x}) \neq 0$*. Then, there exists an open neighborhood* U *and* V *of* $\bar{x}$ *and* 0 *and a* C^k *diffeomorphism* $F : U \to V$ *with* $F(\bar{x}) = 0$ *such that*

$$f(F^{-1}(y)) = f(\bar{x}) + y_1.$$

Proof. W.o.l.g. we may assume that $\frac{\partial f}{\partial x_1}(\bar{x}) \neq 0$. Define $y = F(x)$ by

$$y_1 = f(x) - f(\bar{x}),\ y_j = x_j - \bar{x}_j.$$

Note that $detDF(\bar{x}) = \frac{\partial f}{\partial x_1}(\bar{x}) \neq 0$ and hence $DF(\bar{x})$ is non-singular. Due to the inverse function theorem F is locally invertible with $F^{-1} \in C^k$:

$$f(\bar{x}) + y_1 = f(x) = f(F^{-1}(y)).$$

The point $\bar{x}$ is not a local minimum of f since $\partial f \neq 0$ and in the new coordinates $F^{-1}(y)$ the function f is linear in y_1.

Theorem 4.23 (Morse Lemma)**.** *Let* $f \in C^2$*,* $\nabla f(\bar{x}) = 0$ *and* $\nabla^2 f(\bar{x})$ *non-singular. Suppose exactly* k *eigenvalues of* $\nabla^2 f(\bar{x})$ *are negative. Then, there exists an open neighborhood* U *and* V *of* $\bar{x}$ *and* 0 *and a* C^k *diffeomorphism* $F : U \to V$ *with* $F(\bar{x}) = 0$ *such that*

$$f(F^{-1}(y)) = f(\bar{x}) - \sum_{i=1}^{k} y_i^2 + \sum_{i=k+1}^{n} y_{j^2}$$

Theorem 4.24. *Let $k \geq 1$ and $h_i, g_j \in C^k$. Suppose that LICQ holds at $\bar{x} \in \mathfrak{S}$ Set $q = |A(\bar{x})|$. Then, there exists an open neighborhood U and V of $\bar{x}$ and 0 and a C^k diffeomorphism $F : U \to V$ such that*

$$F(\bar{x}) = 0, F(\mathfrak{S} \cap U) = \{(0)_m \times \mathbb{R}^q_+ \times \mathbb{R}^{n-m-q}\} \cap V$$

with $y = F(x)$ defined by

$$\begin{aligned} y_i &= h_i(x), i = 1, \ldots, m \\ y_{m+i} &= g_i(x), i \in |A(\bar{x})| \\ y_{q+m+i} &= \xi^T_{q+m+i}(x - \bar{x}) \end{aligned}$$

where $(\nabla h_i(\bar{x}), \nabla g_{i,i\in|A(\bar{x})|}\xi_j)$ are a base of $\mathbb{R}^n$.

The admissible set $\mathfrak{S}$ in y coordinates is described by linear equalities $y_i = 0$ and inequalities $y_i \geq 0, i = m+1, \ldots, m+q$. Minimizing f on $\mathfrak{S}$ is equivalent to minimize $f(F^{-1})$ as function of y. One can prove that 0 is a local minimum of $f(F^{-1}(y))$. The derivatives of $f(F^{-1})$ at 0 is the KKT system.

To close we wish to consolidate the ideas presented above by applying them to a simple example. We apply KKT theorem with constraint qualifications which give necessary conditions for a minimum conditions:

$$\nabla f(x^*) - \sum_{i=1}^{n} \lambda_i^* \nabla g_i(x^*) - \sum_{j=1}^{p} \mu_i^* \nabla h_j(x^*) = 0$$

$$\begin{aligned} \lambda_i g_i(x^*) &= 0; \\ g_i(x^*) &\geq 0; \\ h_j(x^*) &= 0. \end{aligned}$$

These are necessary for x^* to be a local minimum, if the constraint qualifications hold.

Example 4.25. *(a) $n = 2$, $m = 0$, $p = 1$, $f(x) = -x_1 - x_2$, $h_1(x) = \frac{1}{2}((x_1)^2 + (x_2)^2) - 1$.*

Thus solve $\min_x f(x)$ subject to $h_1(x) = 0$.

Constraint Qualification:

(i) MFCQ:

$$\nabla h(x) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

and $\text{rank}(\nabla h(x)) = 1$ *for* $x \in \mathcal{S}$ *which is the maximum column rank.*

(ii) LICQ:≡ MFCQ.

(iii) Slater: h_j *is affine linear if* $h_j(x) = Hx + b$ *which is not true since* h *is non-linear.*

Since MFCQ holds then KKT necessarily holds:

$$\begin{pmatrix} -1 \\ -1 \end{pmatrix} - \mu_1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0;$$
$$x_1^2 + x_2^2 = 0.$$

the solution of which are: $(x_1, x_2) = (-1, -1)$ *with* $\mu_1 = 1$ *and* $(x_1, x_2) = (1, 1)$ *with* $\mu_1 = -1$. *Sufficient conditions will rule out one of the solutions.*

b) $n = 2,\ m = 2,\ p = 0,\ f(x) = -x_1,\ g_1(x) = 1 - (x_1)^2 - x_2,\ g_2(x) = 1 + x_2 - (x_1)^2$.
$x^* = \binom{1}{0}$
Constraint Qualifications:

(i) MFCQ: $\exists z : z^T \nabla g_{\mathcal{A}(x^*)}(x^*) > 0$
$(p \equiv 0)\quad \mathcal{A}(x^*) = \{1, 2\}$ *since* $x^* = (1, 0)$ *we obtain* $g_1(x^*) = 0$, $g_2(x^*) = 0$.

$$\begin{aligned} \nabla g_{\mathcal{A}(x^*)}(x) &= \begin{pmatrix} \nabla g_1(x^*) & \nabla g_2(x^*) \end{pmatrix} \\ &= \begin{pmatrix} -2 & -2 \\ -1 & 1 \end{pmatrix} \end{aligned}$$

We need $z \in \mathbb{R}^2$ *such that*

$$z^T \begin{pmatrix} -2 & -2 \\ -1 & 1 \end{pmatrix} > 0$$

Take $z = (-1, -1)$.

(ii) LICQ: $\nabla g_{\mathcal{A}(x^*)}$ *must have full rank. If the minimum is not known, we have to check* $\nabla g_{\mathcal{A}(x^*)}$ *has full rank* $\forall x \in \mathcal{S}$.

$$\begin{aligned} \nabla g_1 &= \begin{pmatrix} -2x_1 \\ -1 \end{pmatrix} \\ \nabla g_2 &= \begin{pmatrix} -2x_1 \\ 1 \end{pmatrix}. \end{aligned}$$

If $\mathcal{A}(x^*) = \{1, 2\}$ *then*

$$\nabla g_{\mathcal{A}} = \begin{pmatrix} -2x_1 & -2x_1 \\ -1 & 1 \end{pmatrix}$$

has rank 2 for $x_1 \neq 0$. *If* $\mathcal{A}(x^*) = \{1\}$ *or* $\mathcal{A}(x^*) = \{2\}$ *then the vectors have full rank.*

(iii) Slater: Check that $-g_i$ *is convex:*

$$-\nabla^2 g_1 = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

4.2 Lagrange function and its relation to necessary optimality conditions

The optimality conditions can also be formulated using the Lagrange function.

Definition 4.26.

$$L(x, \lambda, \mu) := f(x) - \lambda^T g(x) - \mu^T h(x)$$

is called the associated **Lagrange function** *to* NLO.

A reformulation of the KKT equation (4.17) is hence given by

$$\begin{aligned} \nabla_x L(x^*, \lambda^*, \mu^*) &= 0 \\ h(x^*) &= 0 \\ \min(\lambda_i^*, g_i(x^*)) &= 0 \quad i = 1, \ldots, m \end{aligned}$$

Again, the previous set of conditions and suitable constraint qualifications are only necessary for a local minimum but not sufficient.

There is a reformulation of the solution to NLO as saddle point of the Lagrangian. In the following we assume that (x, λ, μ) vary independently. If we have a solution

$$(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p$$

such that

$$L(x^*, \lambda, \mu) \leq L(x^*, \lambda^*, \mu^*) \leq L(x, \lambda^*, \mu^*) \tag{4.18}$$ G32

for all $x \in \mathbb{R}^n,\ \lambda \geq 0 \in \mathbb{R}^m,\ \mu \in \mathbb{R}^p$. then, we also have a solution to NLO.

Theorem 4.27. *If (x^*, λ^*, μ^*) is a solution to (4.18), then x^* is a solution to* NLO.

Proof. Let $(\bar{x}, \bar{\lambda}, \bar{\mu})$ be a saddle point. Then, by definition we obtain

$$\begin{aligned} f(\bar{x}) - \sum \lambda_i g_i(\bar{x}) - \sum \mu_j h_j(\bar{x}) \leq \\ f(\bar{x}) - \sum \bar{\lambda}_i g_i(\bar{x}) - \sum \bar{\mu}_j h_j(\bar{x}) \leq \\ f(x) - \sum \bar{\lambda}_i g_i(x) - \sum \bar{\mu}_j h_j(x). \end{aligned}$$

for any $\lambda \in \mathbb{R}^+$ and μ and x. This implies that

$$\sum (\bar{\lambda}_i - \lambda_i) g_i(\bar{x}) + \sum (\bar{\mu}_j - \mu_j) h_j(\bar{x}) \leq 0.$$

For $\lambda_i = \bar{\lambda}_i$ and $\mu_j = \bar{\mu}_j$ except for $j = j_0$ where $\bar{\mu}_{j_0} = \alpha + \mu_{j_0}$ we obtain $\alpha h_{j_0}(\bar{x}) \leq 0$ for any $\alpha \in \mathbb{R}$. This implies $h_j(\bar{x}) = 0$. Furthermore, for $\lambda_i = \bar{\lambda}_i$ except for i_0 where $\lambda_{i_0} = \bar{\lambda}_{i_0} + 1$ we obtain $-g_{i_0} \leq 0$. For $\lambda = 0$ we have $\bar{\lambda}^T g(\bar{x}) \leq 0$ which together with $\bar{\lambda} \geq 0$ and $g(\bar{x}) \geq 0$ implies that $\lambda^T g(\bar{x}) = 0$.

$$\begin{aligned} f(\bar{x}) - \sum \lambda_i g_i(\bar{x}) - \sum \mu_j h_j(\bar{x}) \leq f(\bar{x}) \leq \\ f(x) - \sum \bar{\lambda}_i g_i(x) - \sum \bar{\mu}_j h_j(x) \leq f(x) \end{aligned}$$

if x is in the feasible set. Hence, $\bar{x}$ is the minimum.

Remark 4.28. *For many problems there is* **no** *solution (4.18) even so there is a solution to* NLO.

4.3 Sufficient optimality conditions

In this section sufficient conditions for x^* to be a local minimum will be discussed. The first theorem gives sufficient conditions under the additional and restrictive assumption that f **convex.**

S30 **Theorem 4.29.** *Let $-g_i,\ i = 1,\dots,m$ be convex, $h_j,\ j = 1,\dots,p$ affine linear and f convex on $\mathbb{R}^n$. Then, the multiplier rule (4.17) is also sufficient for global optimality of x^*. If f is strictly convex, then x^* is unique.*

Proof. First note that the set $\mathfrak{S}$ is convex, i.e., $x, x^* \in \mathfrak{S}$ we have $\lambda x + (1-\lambda)x^* \in \mathfrak{S}$ for all $\lambda \in [0,1]$. Let $x \in \mathfrak{S}$ and hence $g(x) \geq 0, h(x) = 0$. Since x^* satisfies (4.17) we have multipliers $\lambda_i^* \geq 0$ and hence due to the convexity of $f, -g$ and the linearity of h and the convexity of $\mathfrak{S}$:

$$\begin{aligned} f(x) &\geq f(x) - (\lambda^*)^T g(x) - (\mu^*)^T h(x) \\ &\geq f(x^*) - \nabla f(x^*)^T (x - x^*) + (\lambda^*)^T (-g(x^*) - \nabla g(x^*)^T (x - x^*)) \\ &\dots \quad -(\mu^*)^T (h(x^*) + \nabla h(x^*)^T (x - x^*)) \\ &= f(x^*) - (\nabla f(x^*) - \nabla g(x^*)\lambda^T - \nabla h(x^*)\mu^*)^T (x - x^*) \\ &= f(x^*) \end{aligned}$$

Since the previous holds for all $x \in \mathfrak{S}$, we have global optimality. □

Combining this results with the general KKT–theorem we obtain: x^* is a minimum, if and only if the multiplier rule (4.17) holds under the assumption that f is convex and NLO satisfies the Slater condition. Hence, in the convex case the multiplier rule is sufficient for optimality.

However, in general the condition f convex is not satisfied. The following theorem gives a sufficient condition in the general case. We need the following definition.

Let $x^* \in \mathfrak{S}$ and

$$N^* := (\nabla h_1(x^*), \dots, \nabla h_p(x^*), \nabla g_{i_1}(x^*), \dots, \nabla g_{i_l}(x^*)),$$

where $\mathcal{A} = \mathcal{A}(x^*)$ is the active set. We abbreviate the previous by $N^* = (\nabla h(x^*), \nabla g_{\mathcal{A}}(x^*)$. We call

$$\mathcal{Z}_1^0(x^*) := \{z :\ (N^*)^T z = 0\}$$

the linearised subspace at $\mathfrak{S}$ in x^*. If N^* has full column rank and n columns, then $\mathcal{Z}_1^0(x^*)$ is x^*. x^* is hence a corner of $\mathfrak{S}$.

S31 **Theorem 4.30.** *Let $f, g, h \in C^2(\mathcal{U}(x^*))$ and $x^* \in \mathfrak{S}$ be a local minimum of f on $\mathfrak{S}$. Let $N^* = (\nabla h(x^*), \nabla g_{\mathcal{A}}(x^*))$ have full column rank, i.e., LICQ is satisfied.*

Then, there exists uniquely defined multipliers $\lambda^* \geq 0$ $(\in \mathbb{R}^m)$ and $\mu^* \in \mathbb{R}^p$ such that

$$\left.\begin{aligned} \nabla_x L(x^*, \lambda^*, \mu^*) &= 0 \\ (\lambda^*)^T g(x^*) &= 0 \\ z^T \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) z &\geq \alpha z^T z \quad \textit{for all } z \in \mathcal{Z}_1^0(x^*) \textit{ mit } \alpha \geq 0. \end{aligned}\right\} \quad (4.19) \quad \text{G31}$$

holds.

Conversely, if LICQ is satisfied and (4.19) holds with $\alpha > 0$ and if $\lambda^* + g(x^*) > 0$, then x^* is a strict local minimum of f on $\mathfrak{SS}$.

Remark 4.31. *The condition $\lambda^* + g(x^*) > 0$ is called strict complementarity and implies that if $g_i(x^*) = 0$ we have $\lambda_i^* > 0$ and also $\lambda_i^* = 0$ implies $g_i(x^*) > 0$. There are also optimality conditions which hold true without the condition of strict complementarity .*

If x^* satisfies the multiplier rule and LICQ. Then the multipliers are unique and can be computed as follows

$$\nabla f(x^*) - N(x^*, \mathcal{A}) \begin{pmatrix} \mu^* \\ \lambda_{\mathcal{A}}^* \end{pmatrix} = 0, \qquad N(x^*, \mathcal{A}) = (\nabla h(x^*), \nabla g_{\mathcal{A}}(x^*))$$

Using a $QR-$decomposition of N we have

$$\begin{aligned} QN &= \begin{pmatrix} R \\ \hline 0 \end{pmatrix} \Longrightarrow \\ R \begin{pmatrix} \mu^* \\ \lambda_{\mathcal{A}}^* \end{pmatrix} &= \Big(Q \nabla f(x^*) \Big)_{i=1,\ldots,p+|\mathcal{A}|} \Big\} \text{ first } p + |\mathcal{A}| \text{ components} \end{aligned}$$

The condition (4.19) of theorem 4.30 for $\alpha = 0$ is also called **second–order sufficient condition**.

We offer the following motivation for the proof of (4.30). Denote by

$$\mathcal{Z}_1^0(x) := \{(\nabla h_1, \ldots, \nabla h_m, \nabla g_{i_1}, \ldots, \nabla g_{i_r})^T z = 0,\ i_j \in \mathcal{A}(x)\}.$$

Then a constraint qualification of second–order is given by

$$\forall z \in \mathcal{Z}_1(x)^0 \exists \chi \in C^1 : \chi(0) = x,\ \chi'(0) = z, h(\chi(t)) = 0, g_A(\chi(t)) = 0. \tag{4.20}$$

constraint qualificati

Theorem 4.32. *Let x^* be a local minimum, $\mathcal{Z}_1' = \mathcal{T}'$ and assume the second–order constraint qualification (4.20). Then, we have for all $z \in \mathcal{Z}_1^0(x^*)$*

$$z^T \left(\nabla^2 f(x^*) - \sum \lambda_i \nabla^2 g_i(x^*) - \sum_j \mu_j \nabla^2 h_j(x^*) \right) z \geq 0.$$

The latter condition is necessary. If we consider the unconstrained case we had x^* local minimum, then $z^T\nabla^2 f(x^*)z \geq 0$ is necessary. If $\nabla f(x^*) = 0$ and $z^T\nabla^2 f(x^*)z > 0$, then x^* is a local minimum. Hence, it is tempting to replace in the previous theorem $\geq$ by $>$ in order to obtain sufficient conditions. However, there is a counterexample in the constrained case. The reason being that the set of admissible directions for variations of $\nabla^2 f(x^*)$ is restricted to $\mathcal{Z}_1^0$ compared to $\mathbb{R}^n$ in the unconstrained case. Since we allow less directions to test with, the assumptions on $\nabla^2 f(x^*)$ have to be enforced.

We give some examples on the conditions.

a) $n = 2,\ m = 0,\ p = 1,\ f(x) = -x_1 - x_2,\ h_1(x) = \frac{1}{2}((x_1)^2 + (x_2)^2) - 1.$
$x^* = \binom{1}{1}$ satifies (4.19) with $\alpha > 0$: There is a strict local minimum at $x = \binom{-1}{-1}$. Further, $x^0 = \binom{-1}{-1}$ satisfies $\nabla_x L(x^0, \lambda, \mu) = 0$ with $\mu = 1$, but not (4.19), since x^0 is a maximum.

Next we give a detailed presentation: If the minimum is at $x^* = (1, 1)$,

$$0 = \nabla f(x^*) - \mu^* \nabla h(x^*) = \begin{pmatrix} -1 \\ -1 \end{pmatrix} - \mu^* \begin{pmatrix} 1 \\ 1 \end{pmatrix};$$

$\mu^* = -1$. Hence sufficient conditions are

$$\nabla^2 f(x^*) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}; \quad \nabla^2 h(x^*) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus

$$z^T \left(\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} - (-1) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) z = z^T \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} z = z^T z > 0; \quad \forall z \neq 0;$$

is satisfied with $\alpha = 1$. Hence x^* is a local minimum.

b) $n = 2,\ m = 2,\ p = 0,\ f(x) = -x_1,\ g_1(x) = 1 - (x_1)^2 - x_2,\ g_2(x) = 1 + x_2 - (x_1)^2.$
$x^* = \binom{1}{0},\ \lambda_1^* = \lambda_2^* = \frac{1}{4},\ \mathcal{A}(x^*) = \{1, 2\}.$ x^* is a corner and (4.19) is satisfied with $\alpha > 0$ ($\mathcal{Z}_1^0 = \{0\}$). This is a global minimum since it is a convexe problem with Slater condition. $x^{**} = \binom{-1}{0}$ is a maximum with $\lambda_1^{**} = \lambda_2^{**} = -\frac{1}{4}$.

Again we make a detailed presentation

The local minimum $x^* = (1, 0)$. We verify the sufficient conditions:

$$\begin{pmatrix} -1 \\ 0 \end{pmatrix} - \lambda_1^* \begin{pmatrix} -2 \\ -1 \end{pmatrix} - \lambda_2^* \begin{pmatrix} -2 \\ 1 \end{pmatrix} = 0;$$
$$g_1(x^*) = 0; \quad g_2(x^*) = 0; \quad \lambda_i g_i = 0; \quad \lambda_i + g_i > 0 \Rightarrow \lambda_i \geq 0.$$

which implies $\lambda_1^* = \lambda_2^* = \frac{1}{4}$ from the last equation.

$$z^T \left(\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} - \frac{1}{4} \begin{pmatrix} -2 & 0 \\ 0 & 0 \end{pmatrix} - \frac{1}{4} \begin{pmatrix} -2 & 0 \\ 0 & 0 \end{pmatrix} \right) z = z^T \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} z = \alpha z^T z > 0; \quad \alpha > 0; \; \forall z \in \mathcal{Z}_1^0(x^*)$$

where

$$\begin{aligned} \mathcal{Z}_1^0(x^*) &= \left\{ z \in \mathbb{R}^2 : z^T \begin{pmatrix} -2 & -2 \\ -1 & 1 \end{pmatrix} = 0 \right\}; \\ &= \left\{ z \in \mathbb{R}^2 : -2z_1 - z_2 = 0; -2z_1 + z_2 = 0 \Rightarrow z_1 = 0 \right\}; \\ &= \{(0, 0)\}; \end{aligned}$$

We need now to check that

$$z^T \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} z \geq \alpha z^T z$$

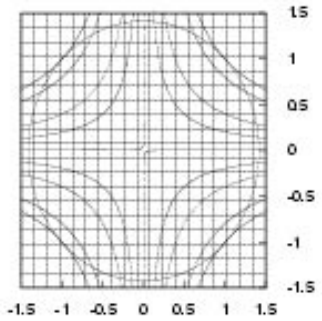
just for $z = (0, 0)$ for any α. Hence x^* is a strict local minimum.

c) $n = 2, \; m = 0, \; p = 1, \; f(x) = -10x_1x_2, \; h_1(x) = \frac{1}{2}((x_1)^2 + (x_2)^2) - 1.$ $\binom{1}{1}, \; \binom{-1}{-1}$ satisfy (4.19) with $\alpha = 20 : 2$ is a strict **local** minimum, see figure 4.3.

1. Linear or quadratic programming problems are a special case of convex optimization. The general linear programming problem is as follows

$$\begin{aligned} f(x) &= c^T x \\ g(x) &= G^T x + g^0 \quad (\geq 0) \\ h(x) &= H^T x + h^0 \quad (= 0) \end{aligned}$$

The standard form for linear programming is covered by the previous theory with



co-c

Figure 8: Contourlines for example (c)

$$\begin{aligned} f(x) &= c^T x \\ h(x) &= H^T x + h^0 \\ g(x) &= x \end{aligned}$$

The convex quadratic programming problem is

$$\begin{aligned} f(x) &= -b^T x + \tfrac{1}{2} x^T A x \qquad A \textbf{ positive semi–definite} \\ g(x) &= G^T x + g^0 \quad (\geq 0) \\ h(x) &= H^T x + h^0 \quad (= 0) \end{aligned}$$

All previous problems are convex optimization problems and special algorithms for their solution exist.

4.4 Lagrange function and its relation to sufficient optimality conditions

In case of convex optimization problems the saddle point condition (4.18) is also sufficient for existence of a local minimum.

S33 **Theorem 4.33.** *Let $f, -g_1, \ldots, -g_m : \mathbb{R}^n \to \mathbb{R}$ be konvex, $h_1, \ldots, h_p$ and affine linear. Let the Slater–condition be satisfied. Then x^* is a solution* NLO*, if and only if there exists $\lambda^* \in \mathbb{R}^m,\ \lambda^* \geq 0$ and $\mu^* \in \mathbb{R}^p$ such that*

$$L(x^*, \lambda, \mu) \leq L(x^*, \lambda^*, \mu^*) \leq L(x, \lambda^*, \mu^*)$$

for all $x \in \mathbb{R}^n$, for all $\lambda \geq 0$ and for all $\mu \in \mathbb{R}^p$.

There are important implications of the previous theorem. At first, the results can be used as a first numerical method for solving convex NLO: Assume f is uniformly convex on $\mathbb{R}^n$, i.e., $\lambda_{\min}(\nabla^2 f(x)) \geq \gamma > 0$ for all $x \in \mathbb{R}^n$. Then, for fixed $\lambda \geq 0,\ \mu \in \mathbb{R}^p$ solve the unconstrained minimization problem

$$x = x(\lambda, \mu) := argmin L(x, \lambda, \mu).$$

The problem has a unique solution provided that f is uniformly convex. Then, define

$$\Phi(\lambda, \mu) = L(x(\lambda, \mu), \lambda, \mu)$$

and solve the box–constrained problem

$$\Phi(\lambda, \mu) \stackrel{!}{=} \max, \quad \lambda \geq 0.$$

Second, the results can be used to derive the duality relations in linear programming. This will be done in the sequel for a linear programming problem in standard form:

$$\begin{aligned} f(x) &= c^T x \stackrel{!}{=} \min \\ Ax &= b\ , \\ x &\geq 0\ . \end{aligned}$$

The Lagrange function is $L = c^T x - \mu^T(Ax - b) - \lambda^T x$ for $\lambda_i \in R^+$ and $\mu \in \mathbb{R}$. The Slater condition ($-g$ convex, h affine linear and $\exists x_0 : h(x_0) = 0, g_L(x_0) \geq 0$) is satisfied if the feasible set $\{x : Ax = b, x \geq 0\}$ is not empty. Then, Theorem 4.25 applies:

$$\begin{aligned} c - A^T\mu - \lambda &= 0\ , \\ Ax - b &= 0\ , \\ x_i\lambda_i &= 0\ ,\ i\ =\ 1, \ldots, n \end{aligned}$$

This is also the optimality system for the problem

$$\begin{aligned} \hat{f}(\mu) &= \mu^T b \stackrel{!}{=} \max \\ c - A^T\mu &\geq 0. \end{aligned}$$

with multipliers $x_i \in \mathbb{R}^+$. Due to the saddle point condition $L(x^*, \lambda, \mu) \leq L(x^*, \lambda^*, \mu^*)$ we have

$$\begin{aligned}(\lambda^*, \mu^*) = &\text{ argmax }_{\lambda,\mu} c^T x^* - \mu^T A x^* + \mu^T b - \lambda^T x^* \\ = &\text{ argmax } (c - A^T \mu)^T x^* + \mu^T b - \lambda^T x^*\end{aligned}$$

and $x_i \geq 0, \lambda_i \geq 0$. Hence, $-\lambda^T x^* \to \max$ implies $\lambda = 0$. If (x, μ) is feasible we have $\mu^T b \leq c^T x$. Due to the KKT system we have that in the optimum

$$(\mu^*)^T b \;=\; c^T x^* \;.$$

4.5 Examples, discussion of nonlinear constraint qualifications

We recall the general nonlinear optimization problem in standard form. This is given by

$$\begin{gathered}\min f(x) \text{ subject to} \\ \mathcal{S} := \{x : g(x) \geq 0, h(x) = 0\}.\end{gathered}$$

Under the assumptions that $\mathcal{S} \neq \emptyset$,$f, g, h$ are defined on some open set $\mathcal{D}$ where $\mathcal{S}$ is a closed subset of $\mathcal{D}$ and f, g, h are at least twice continuously differentiable and *constraint qualifications*, we obtain the necessary (first oder) conditions

$$\begin{aligned}\nabla f(x^*) - \lambda^T \nabla g(x^*) - \mu^T \nabla h(x^*) &= 0 \\ \lambda &\geq 0 \\ \lambda^T g(x^*) &= 0 \\ h(x^*) &= 0 \\ g(x^*) &\geq 0\end{aligned}$$

and the sufficient (second order) condition

$$z^T \left(\nabla^2 f(x^*) - \sum_i \lambda_i^* \nabla^2 g_i(x^*) - \sum_j \mu_j \nabla^2 h_j(x^*) \right) \geq \alpha z^T z$$

for suitable z.

Remark 4.34. *Some authors use*

$$g(x) \leq 0$$

instead of the previous formulation. In this case the sign in front of the inequality multipliers λ has to be changed. The sign in front of the equality multiplier μ can be changed to a plus at costs, since there is no further restriction on μ.

In the case of inequality constraints only, we have that z is an admissible direction, iff $z^T \nabla g_i(x^*) = 0$ for all inequalities i.

1. Let $f(x) = -x_1 - x_2$ and $h(x) = \frac{1}{2}\left(x_1^2 + x_2^2\right) - 1$. We minimize $f(x)$ on a circle of radius one with center $x = (0, 0)$. The minimum is attained at $x^* = (1, 1)$. The multiplier $\mu = -1$ and the necessary conditions are satisfied. However, $\nabla^2 f(x^*)$ is the all zero matrix. But as expected the matrix
$$\nabla^2 f(x^*) - \mu \nabla^2 h(x^*)$$
is positive definite.

2. Let $f(x) = -x_1$ and $g_1(x) = 1 - x_1^2 - x_2$ and $g_2(x) = 1 - x_1^2 + x_2$. The problem as the minimum attained at $x^* = (1, 0)$ and $\lambda_i = \frac{1}{4}$. There is an additional extremum at $x = (-1, 0)$. In this case the necessary optimality conditions are not satisfied since $\lambda_i < 0$. For $x^* = (1, 0)$ we obtain
$$\nabla^2 f(x^*) - \lambda_1 \nabla^2 g_1(x^*) - \lambda_2 \nabla^2 g_2(x^*) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$
. This matrix is not(!) positive definite. However, the necessary condition states that only for certain directions z we need the positivity. In our case $z^T \nabla g_i(x^*) = 0$ implies $z = (0, 0)$.

3. Let $f(x) = -x_1$ and $g_1(x) = (1 - x_1)^3 - x_2$ and $g_2(x) = x_1$ and $g_3(x) = x_2$. Then, the minimum is attained at $x^* = (1, 0)$ and $\nabla f(x^*) = (-1, 0)^T$. The gradients of the constraints are
$$\nabla g_1(x^*) = (0, -1)^T,\ \nabla g_2(x^*) = (1, 0)^T, \nabla g_3(x^*) = (0, 1)^T.$$
Due to $\lambda_2 g_2((1, 0)) = 0$ we have $\lambda_2 = 0$. Hence, in the first equation of the necessary conditions we need to express $\nabla f(x^*)$ in terms of $\nabla g_1(x^*)$ and $\nabla g_2(x^*)$. This is impossible! The reason is that no(!) constraint qualification does not hold at the $x^* = (1, 0)$. The tangential cone $\mathcal{T}(M, y)$ is the set that contains all positive multiples of directions $a - y$ where $a \in U_\delta(y) \cap M$. In our example we have
$$\mathcal{T}(\mathcal{S}, (1, 0)) = \{(\alpha, 0) : \alpha \leq 0\}.$$

The linearized cone $Z_1(x)$ is given by linearization of the constraints

$$Z_1(y) = \{z : z^T \nabla g_i(y) \geq 0, \forall i : g_i(y) = 0 \text{ and} z^T \nabla h_j(y) = 0 \; \forall j\}.$$

In our case

$$Z_1((1,0)) = \{(\alpha, 0) : \alpha \in \mathbb{R}\}.$$

The dual cones are $\mathcal{T}' = \{y : y_2 \in \mathbb{R}, y_1 \leq 0\}$ and $Z_1' = \{y : y_2 \in \mathbb{R}, y_1 = 0\}$. The dual cone represents vectors having an angle less or equal to $\pm\pi$ to any point in the set. Since $\nabla f(x^*) \in \mathcal{T}'$ we need at least $Z_1' = \mathcal{T}'$. Here, even the weakest of all constraint qualification

$$Z_1' = \mathcal{T}'$$

is violated!

Further counterexamples are stated below. The general problem reads

$$\min f(x) \text{ subject to } g(x) \geq 0, \; h(x) = 0$$

The KKT system is

$$\begin{aligned} &\nabla f(x^*) - \lambda^T \nabla g(x^*) - \mu^T \nabla h(x^*) = 0 \\ &\lambda^* \geq 0 \\ &\lambda^T g(x^*) = 0 \end{aligned}$$

Example 4.35.

$$f(x) = -x_1, \; g(x) = \begin{bmatrix} (1-x_1)^3 - x_2 \\ x_1 \\ x_2 \end{bmatrix}$$

The gradients are

$$\nabla f(x^*) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \nabla g(x^*) = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

and the minimum is $x^* = (1, 0)$. Hence, $\lambda_2 = 0$ and we cannot write $\nabla f(x^*)$ as a linear combination of the remaining gradients. The multiplier rule is not valid. Note that ∇g_1 and ∇g_3 are linear dependent and therefore the MFCQ constraint qualification is violated.

Example 4.36.

$$f(x),\ h(x) = 1$$

We convert the equality constraint in two inequality constraints

$$f(x),\ h(x) - 1 \geq 0,\ 1 - h(x) \geq 0$$

Since $\nabla g(x) = (\nabla h(x), -\nabla h(x))$ *which cannot be column regular (LICQ) no matter how "good"* $h(x)$ *was. Hence, the constraint qualification is violated after reformulation.*

5 Differentiability for operators on Banach spaces

We introduce the basic definitions of differentiable operators and contraction principles used later for solving optimization problems.

5.1 Introduction

If not stated otherwise, we consider a Banach space X, a Banach space Y, a possible nonlinear transformation T defined on a subdomain $U \subset X$ and having range $R \subset Y$.

Definition 5.1. *Let* $U \subset X$ *be open.*

1. $F : U \to Y$ *is called (Frechet-)differentiable at* $u \in U$*, iff there exists a linear map* $A \in L(X, Y)$*, such that*

$$F(u + h) = F(u) + Ah + o(\|h\|), \|h\| \to 0$$

i.e.

$$\frac{1}{\|h\|}\|F(u + h) - F(u) - Ah\| \to 0, \quad \|h\| \to 0$$

2. $F : U \to Y$ *is called Gateaux-differentiable at* $u \in U$*, iff there exists a linear map* $A \in L(X, Y)$*, such that for* $t \in \mathbb{R}$

$$\frac{1}{t}(F(u + th) - F(u) - t\ Ah) \to 0$$

for all $h \in X$ *and* $t \to 0$*.*

We write $DF(u)$ for the Frechet derivative of F in u and $D_G F(u)$ for the Gateaux derivative. The Gateaux derivative is a directional derivative. The Frechet differentiability does not depend on the norm. We will call a function

differentiable, if it is Frechet differentiable. If F is Frechet differentiable, then it also has a Gateaux derivative. The converse is only true, if $u \to D_G F(u)$ is continuous, see Propositions below. Some authors also allow A to be a nonlinear map in the case of Gateaux differentials.

Example 5.2. *The constant map $F(u) = c$ is differentiable at any u and $DF(u) = 0$ for all $u \in X$.*

Let $B : X \times Y \to Z$ be a bilinear continuous map. Then B is differentiable at every point $(u, v) \in X \times Y$ and $DB(u, v)$ is the map $[h, k] \mapsto B(h, v) + B(v, k)$.

Let H be a Hilbert space with scalar product. Then $F : u \mapsto \|u\|^2 = (u, u)$ is differentiable at every point $u \in H$ and $DF(u)$ is the map $h \mapsto 2(u, h)$.

Let $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ with $f(x) = f(x_1, \ldots, x_n) : X \to Y$ continuously differentiable. Then $Df(x)$ is the Frechet differentiable and its derivative can be expressed as the matrix $(\frac{\partial f_i}{\partial x_j})_{ij}$.

We conclude with assumption and definitions of above:

1. The map $A \in L(X, Y)$ is uniquely determined.
2. F is differentiable at $u \in U \Rightarrow F$ is continuous
3. F is differentiable at $u \in U \Rightarrow F$ is Gateaux-differentiable at u with $DF(u) = D_G F(u)$
4. Let X, Y, Z be Banach spaces and $U \subset X$, $V \subset Y$ open; further let $F : U \to Y$ with $FU \subset V$ be differentiable at $u \in U$ and $G : V \to Z$ differentiable at $v = Fu$.

   Then $G \circ F : U \to Z$ is differentiable at $u \in U$ and

$$D(G \circ F)(u) = DG(v)DF(u)$$

   .

Exercise 5.3. *1. Prove that $f(u) = sin(u(1))$ is F-differentiable in $C(0, 1)$.*

2. Prove that $f(u) = \|u\|_H^2$ is F-differentiable in a Hilbert space H.

Proposition 5.4. *("Mean-Value" Theorem) Let $F : U \subset X \to Y$ be (Gateaux-)differentiable at any point of U.*

Then for every $u, v \in U$ with $[u, v] = conv(u, v) \subset U$

$$\|F(u) - F(v)\| \leq \|u - v\| \sup\{\|D_G F(w)\| : \quad w \in [u, v]\}$$

Proposition 5.5. *[2.3] Let $F : U \subset X \to Y$ be Gateaux-differentiable at any $u \in U$ and let $D_G F(u) : U \to L(X,Y)$ be continous at $u_0 \in U$.*

Then F is (Frechet-)differentiable at u_0 and $DF(u_0) = D_G F(u_0)$.

On a Banach space X consider the open set $U \subset L(X)$ of invertible linear continuous operators and the map $i : U \to U$ defined by $i(T) = T^{-1}$.

Then i is Frechet differentiable and $Di(T)H = T^{-1}HT^{-1}$. Motivation of the formula:

$$T\ Di(T)H = T((T+H)^{-1} - T^{-1}) = (Id + HT^{-1})^{-1} - Id = Di(Id)HT^{-1}$$
$$Di(T)H = T^{-1}Di(Id)HT^{-1}$$

Proof. For all $H \in U$ we have

$$\begin{aligned} 0 &= \frac{1}{\|H\|}\Big((T+H)^{-1} - T^{-1} - Di(T)\Big) \\ &= \frac{1}{\|H\|}\Big(((Id + HT^{-1})T)^{-1} - T^{-1}(Id + HT^{-1})\Big) \end{aligned}$$

Suppose the line–segment $[u,v] \in U$ and $F \in C^1(U;Y)$. The map $F \circ \gamma : [0,1] \to Y$ given by

$$F \circ \gamma(t) = F(tu + (1-t)v) : [0,1] \to Y$$

is in C^1 and the derivative is given by

$$(F \circ \gamma)'(t) = F'(tu + (1-t)v)[u - v]$$

Integrating from 0 to 1 we obtain the useful equivalence

$$F(v) - F(u) = \left[\int_0^1 F'(tu + (1-t)v)dt\right](u - v)$$

wherein F' takes values in $L(X;Y)$.

Higher order derivatives are introduced in the following definition.

Definition 5.6. *Let X, Y be Banach spaces and $U \subset X$ open. $F : U \to Y$ be differentiable with the derivative*

$$F' : U \to L(X,Y)$$

1. *If F' is differentiable at $u_0 \in U$ and we write*

$$DF'(u_0) = D^2F(u_0) = F''(u_0)$$

Here $F''(u_0) \in L(X, L(X,Y)) = L_2(X,Y)$ is the set of all bilinear continous maps from $X \times X$ to Y.

2. *Analogously we define*

$$F^{(n)}(u_0) = D^n F(u_0) \in L_n(X, Y)$$

We write

$$F \in C^n(U, Y)$$

if $F^{(n)}$ exists everywhere and the map $F : U \to L_n(X, Y)$ is continous.

Proposition 5.7. *Let $F : U \to Y$ be twice differentiable in $u \in U$.*
Then

$$F(u+h) = F(u) + F'(u)h + \frac{1}{2}F''(u)[h,h] + o(\|h\|^2), h \to 0$$

Furthermore $F''(u) \in L_2(X, Y)$ is symetric, i.e.

$$F''(u)[h,k] = F''(u)[k,h] \quad \forall h, k \in X$$

Definition 5.8. *Let X, Y, Z Banach spaces, $U \subset X$, $V \subset Y$ open and let $F : U \times V \to Z$ be given. Assume that for fixed $v_0 \in V$ the map $F(\cdot, v_0) : U \to Z, u \to F(u, v_0)$ is differentiable at $u_0 \in U$.*

Then we say that F is partially differentiable in (u_0, v_0) with respect to u and write

$$D_u F(u_0, v_0) = F_u(u_0, v_0)$$

for its partial derivative w.r.t. u.

Proposition 5.9. *(Partial Derivatives)*

1. *$F : U \times V \to Z$ is differentiable in $(u_0, v_0) \in U \times V$.*

   Then the partial derivatives w.r.t. u, v exists in (u_0, v_0) and

$$F_u(u_0, v_0)h = DF(u_0, v_0)(h, 0) \quad \forall h \in X$$

$$F_v(u_0, v_0)k = DF(u_0, v_0)(0, k) \quad \forall k \in Y$$

2. *If F_u, F_v exists and are continous on $U \times V$, then F is differentiable*
3. *If $F \in C^2(U \times V, Z)$, then $(F_u)_v = (F_v)_u$*

Theorem 5.10. *(Taylor's Formula)*
Let $F \in C^n(X, Y), u \in U, [u, u+h] \subset U$.
Then

$$F(u+h) = \sum_{j=0}^{n-1} \frac{1}{j!} F^{(j)}(u)[h, \ldots, h] + R$$

where

$$R = \frac{1}{(n-1)!} \int_0^1 (1-t)^{(n-1)} F^{(n)}(u+th) dt [h, \ldots, h]$$

Theorem 5.11. *Let X be a Banach space and let A be a bounded linear operator from X into X. If $\|A\| = a < 1$ then $(I - A)^{-1}$ exists and $\|(I - A)^{-1}\| \leq 1/(1 - a)$.*

Proof. Existence. Given an arbitrary $x \in X$, we show that there exists a unique $y \in X$, such that $(I - A)y = x$. Consider the sequence $x_n = \sum_{i=0}^{n} A^i/i!x$. Then x_n is a Cauchy sequence, since $\|A\| < 1$. Further note that $(I - A)\sum_{i=0}^{n} A^i/i!x = x - A^n/n!x \to x$. Therefore, the limit point y of x_n exists and has the desired properties. □

Theorem 5.12. *Let $f : X \to Y$ be continuously Frechet differentiable on X. Let $x^* \in X$. Then f is Lipshitz in a neighbourhood of x^*.*

Proof. Note that in the finite dimensional case we have $f'(x)$ is continuous by assumption and therefore $\|f(x)-f(y)\| \leq \|f'(x)\|\|x-y\| \leq M\|x-y\|$ by mean value theorem and the continuity of f' on a ball $B_r(x)$. The proof is analogously in infinite dimensions. □

Example 5.13. *Let $X = L^2([a, b])$ and $f, x \in X$. Consider the integral equation*

$$x(t) = f(t) + \lambda \int_a^b K(t, s)x(s)ds \text{ a.e.}$$

with the kernel $\int_a^b \int_a^b K(t, s)^2 dsdt = \beta^2 < \infty$. Then the integral defines a bounded linear operator on X, i.e.

$$T : L^2(a, b) \to L^2(a, b)$$

by $T(x)(t) = f(t) + \int_a^b K(t, s)x(s)ds$, since by Hölders inequality

$$\int_a^b \left(\int_a^b K(t, s)x(s)ds \right)^2 dt \leq \int_a^b \left(\int_a^b K(t, s)^2 ds \int_a^b x^2(s)ds \right) dt \leq \beta^2 \|x\|^2$$

Hence $T : L^2(a, b) \to L^2(a, b)$ and T is a contraction, iff $\lambda < 1/\beta$.

Example 5.14. *We show that the differential*

$$\delta f(x; h) = \int_0^1 g_x(x, t)h(t)dt$$

in Example 2 is a Frechet differential. We have

$$|f(x + h) - f(x) - \delta f(x; h)| = |\int_0^1 \{g(x + h, t) - g(x, t) - g_x(x, t)h(t)\}dt|$$

For a fixed t we have, by the one–dimensional mean value theorem,

$$g(x+h,t) - g(x,t) = g_x(\bar{x}(t),t)h(t)$$

where $\|x(t) - \bar{x}(t)\| \leq h(t)$. Given $\epsilon > 0$ the uniforma continuity of g_x in x and t implies that there is a $\delta > 0$ such that $\|h\| \leq \delta$,

$$\|g_x(x+h,t) - g_x(x,t)\| < \epsilon.$$

Therefore, we have

$$|f(x+h) - f(x) - \delta f(x;h)| = |\int_0^1 \left(g_x(\bar{x},t) - g_x(x,t)\right) h(t)dt| \leq \epsilon\|h\|$$

for $\|h\| \leq \delta$. The result follows.

Example 5.15. *Let $X = C^n[0,1]$ the space of continuous $n-$vector functions on $[0,1]$, let $Y = C^m[0,1]$ and define $T : X \to Y$ by*

$$T(x) = \int_0^t F(x(\tau),\tau)d\tau$$

where F has continuous partial derivatives with respect to its arguments .The Gateaux differentiable of T is easily seen to be

$$\delta T(x;h) = \int_0^t F_x(x(\tau),\tau)h(\tau)d\tau.$$

This is the Frechet differential and $\delta T(x;h)$ is continuous in the variable x.

Example 5.16. *Let H_0^1 where Ω is a bounded domain of $\mathbb{R}^n$ denote the usual Sobolev spaces with scalar product $(\cdot,\cdot)_{H^{1,2}}$ and usual norm. Let $n > 2$ and suppose f satisfies*

$$|f(x,s)| \leq a + b|s|^\sigma \text{ with } \sigma \leq \frac{n+2}{n-2} = 2^* - 1.$$

. By Sobolev embedding we have that H_0^1 is imbededded in 2^*. and hence $\|v\|_{L^{2^*}} \leq c\|v\|_{H^{1,2}}$. Now, we consider the map $fL^{2^*} \to L^q$ with $q = 2^*/\sigma$. We obtain that f is continuous, i.e., we conclude from $u \to u^*$ in L^{2^*} we conclude $f(u) \to f(u^*)$ in L^q. Furthermore, we have $f \in L^{2n/(n+2)}$ for all $u \in H_0^1$ and due to Hölder's inequality we further obtain $f(u)v \in L^1$ for all u, vßinH_0^1. The equality

$$(N(u),v)_{H^{1,2}} = \int_\Omega f(x,u(x))v(x)dx$$

for all $u, v \in H_0^1$ defines an operator. Since $H^{1,2}$ is a Hilbert space the operator is defined on $H_0^1 \to H_0^1$. The operator N is continuous, since

$$\|N(u) - N(v)\| = \sup\{\| \int_\Omega (f(x,u) - f(x,v))w dx\| : \|w\|_{H^{1,2}} \leq 1\}$$

$$\leq \sup\{\|f(u) - f(v)\|_{L^{2n/(n+2)}} \|w\|_{L^{2^*}} \leq c\|f(u) - f(v)\|_{L^{2n/(n+2)}}$$

If now $u_m \to u^*$ in H_0^1 then, $u_m \to u^*$ in L^{2^*} and due to the continuity of f the assertion follows. Then, define

$$F(x,s) = \int_0^s f(x,t)dt$$

and estimate using the assertion on f, F as

$$|F| \leq c + d|s|^{2^*}$$

Then, $F(\cdot, u(\cdot)) \in L^1$ for all $u \in H_0^1$ and it makes sense to consider $\phi : H_0^1 \to \mathbb{R}$ by setting

$$\phi(u) = \int_\Omega F(x, u(x))dx.$$

The functional ϕ can be obtained by composition according to the following diagram

$$\phi : \ H_0^1 \to^\alpha L^{2^*} \to^F L^{2n/(n+2)} \to^\beta L^1 \to^\phi \mathbb{R}.$$

We can proof that ϕ is differentiable with

$$\phi'(u)v = \int_\Omega f(x, u(x))v(x)dx$$

and hence for any f satisfying the above condition we obtain that ϕ is a C^1- functional on H_0^1 with gradient

$$\nabla\phi(u) = N(u).$$

5.2 Successive Approximations

In the classical formulation the method of successive approximation applies to equations of the type

$$x = T(x) \tag{5.1}$$

A solution is a fixed point for T and under certain coniditions this fixed point can be obtained by considering the sequence $x_{n+1} = Tx_n$. The most famous theorem on existence is Banachs Fixed Point Theorem.

Definition 5.17 (Contraction)**.** *Let S be a subset of a normed space X and let T be a mapping S to S. Then T is a contraction, iff there exists $0 \leq \alpha < 1$ such that*

$$\|T(x_1) - T(x_2)\| \leq \alpha \|x_1 - x_2\|$$

for all $x_1, x_2 \in S$.

Definition 5.18 (Contraction Mapping Theorem)**.** *If T is a contraction on a closed subset S of a Banach space, there exists a unique $x_0 \in S$ satisfying $x_0 = T(x_0)$. Furthermore, x_0 can be obtained as limit of the iteration $x_{n+1} = Tx_n$ for an arbitrary $x_1 \in S$.*

Proof. Since T is a contraction, we have $\|x_{n+1} - x_n\| \leq \alpha^{n-1}\|x_2 - x_1\|$ and therefore $\|x_{n+p} - x_n\| \leq (\alpha^{n+p-2} + \cdots + \alpha^{n-1})\|x_2 - x_1\| \leq \alpha^{n-1} \sum_{i=0}^{p-1} \alpha^k = \alpha^{n-1} \frac{1-\alpha^p}{1-\alpha} \|x_2 - x_1\| \leq \alpha^{n-1}\|x_2 - x_1\|$. Hence, x_n is a Cauchy sequence. Since X is a Bananch space, we conclude that $x_0 = \lim x_n$ exists and $x_0 \in S$ since S is closed. By $\|x_0 - T(x_0)\| = \|x_0 - x_n\| + \alpha\|x_{n-1} - x_0\|$ we see, that x_0 is the fixed point. Uniquenuess is due to the contraction property of T. $\square$

Example 5.19. *Consider $A \in \mathbb{R}^{n \times n}$ strictly diagonal dominant, i.e., for all i $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$. Consider the equation $Ax = b$. We scale the ith equation by $1/|a_{ii}|$, i.e.,*

$$\tilde{b}_i = b_i / a_{ii},\ (\tilde{A})_{ij} = A_{ij}/a_{ii}\ \forall j.$$

In the fixed point form the equation reads

$$x = (I - \tilde{A})x + \tilde{b}$$

where $\|ca_{ij}\| < 1$ for all i, j. To define appropriate Banach spaces we need to introduce a norm. Since on finite dimensional spaces all norms are equivalent, we choose the following

$$\|x\| = \max_i |x_i|$$

and the corresponding matrix norm is

$$\|A\| = \max_i \sum_j |a_{ij}|$$

The corresponding matrix norm (operator norm) is defined by

$$\|A\| = \max |Ax|/|x|$$

Hence for the mapping $T(x) = (I - A)x + b$ we have

$$\|T(x) - T(y)\| \leq \|I - \tilde{A}\|\|x - y\| = \max_i \sum_{j \neq i} |\tilde{a}_{ij}|\|x - y\| \leq \alpha \|x - y\|$$

and therefore T is a contraction with $\alpha < 1$.

The above iteration converges. Since $\max_i \sum_{j \neq i} |a_{ij}|$ might be close to 1, the convergence rate is poor.

This method is a gradient descent method for the problem

$$\min_x \frac{1}{2} x^T A x - b^T x =: f(x)$$

with A symmetric positive definite. Since $\nabla f = Ax - b$ the gradient decsent is given by

$$x^+ = (I - A)x + b.$$

The main drawback of contraction mappings is, that the convergence is only linear. In many applications faster techniques are considered. This can be obtained by Newton's method.

newtons method

Theorem 5.20. *Let X and Y be Banach spaces and let P be a mapping from X to Y. Assume further that P is twice Frechet differentiable and that $\|P''\| \leq K$. There exists a point $x_1 \in X$ s.t. $p_1 = P'(x_1)$ has a bounded inverse with $\|p_1^{-1}\| \leq \beta_1$ and $\|p_1^{-1}[P(x_1)]\| \leq \eta_1$. Further let $h_1 = \beta_1 \eta_1 K$ satisfy $h_1 < 1/2$.*

Then the sequence $x_{n+1} = x_n - p_n^{-1}[P(x_n)]$ exists for all $n > 1$ and converges to a solution of $P(x) = 0$.

Proof. We prove that for a point x_1 satisfying all assumptions, the point $x_2 := x_1 - p_1^{-1} P(x_1)$ satisfies the same assumptions with new constants p_2, η_2 and β_2.

x_2 is well-defined due to the assumption on x_1. Further, $\|x_2 - x_1\| \leq \eta_1$. Since P is twice Frechet differentiable we can apply the mean value theorem to p_1 and since p_1^{-1} is bounded we obtain $\|p_1^{-1}(p_1 - p_2)\| \leq \beta_1 \sup_x \|P''(x)\|\|x_2 - x_1\| = \beta_1 K \eta_1 = h_1$. Since P'' is globally bounded by K. Consider the operator $H := I - p_1^{-1}(p_1 - p_2) = p_1^{-1} p_2$ By the above H is invertible (Neumann series) and has a bounded inverse $\|H^{-1}\| \leq \frac{1}{1 - h_1}$. Since $p_2 = p_1 H$ we have the existence of p_2^{-1} and the bound of its norm by $\|p_2^{-1}\| \leq \beta_1/(1 - h_1) =: \beta_2$. Now, we have to estimate $\|p_2^{-1} P(x_2)\|$. We consider $T_1(x) = x - p_1^{-1} P(x)$ which has the properties $T_1(x_1) = x_2$ and $T'(x_1) = 0$. Further, T is twice Frechet differentiable.

$$p_1^{-1} P(x_2) = x_2 - T_1(x_2) - T_1'(x_1)(x_2 - x_1) = T_1(x_1) - T_1(x_2) - T_1'(x_1)(x_2 - x_1)$$

By the Frechet differentiability we can estimate the last terms by the second derivative $T'' = p_1^{-1}P''$

$$\|p_1^{-1}P(x_2)\| \leq \frac{1}{2}\sup_x \|T''(x)\|\|x_2 - x_1\|^2 \leq \frac{1}{2}\beta_1 K\eta_1^2$$

Finally,

$$\|p_2^{-1}P(x_2)\| = \|H^{-1}p_1^{-1}P(x_2)\| \leq \frac{1}{2}h_1\eta_1/(1-h_1) =: \eta_2 < \frac{1}{2}\eta_1$$

and $h_2 = \beta_2\eta_2 K \leq (\beta_1\eta_1 K)h_1/(1-h_1)^2 1/2 < 1/2$. Since $\eta_{n+1} < 1/2\eta_n \leq (1/2)^{n-1}\eta_1$ we obtain that x_n is a Cauchy sequence with limit x. $\|x_{n+k} - x_n\| = \|x_{n+k} - x_{n+k-1} + x_{n+k-1} \pm \cdots - x_n\| \leq \eta_n \sum_{i=0}^{k-1}(1/2)^i \leq 2\eta_n$.

Now, $p_n(x_{n+1} - x_n) + P(x_n) = 0$ for all n. p_n is bounded, since $\|p_n\| \leq \|p_n - p_1 + p_1\| \leq \|p_1\| + K\|x_n - x_1\|$ and the sequence x_n is converging. Hence, in the previous equation $\|P(x_n)\| \to 0$ and by continuity $P(x) = 0$. □To prove the quadratic convergence property one need to assume that the inverse of the derivative at the stationary point exists and that the following terms are bounded: P^{-1}, P'', P'''.

5.3 Pseudo-Inverse Operators

For the definition and the convergence of augmented Lagrangian methods we introduce pseudo-inverse operators and discuss some properties.

Assume that $A \in L(X,Y)$ is a linear operator and let $y \in Y$ be given. Then we can consider the optimization problem

$$\min_{x \in X} \|Ax - y\|_Y \tag{5.2}$$

There might be more than one solution to the above problem. Therefore, we define the set $S := \{\tilde{x} \in X : \tilde{x} = \text{argmin}\|Ax - y\|\}$ for a given operator A and a point $y \in Y$. The pseudo-inverse is defined as the mapping

$$A^{\#} : Y \to X \tag{5.3}$$

which maps a given $y \in Y$ to the minimum norm element $x \in S$.

Example 5.21. *Consider the finite dimensional case with $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. Hence, $A \in \mathbb{R}^{m\times n}$. For L^2-approximation we usually have $rank(A) = n << m$. Then $A^TA \in \mathbb{R}^{n\times n}$ is invertible and the above minimization problem has a unique solution x which satisfies the set of normal equations*

$$A^TAx = A^Ty \tag{5.4}$$

If now $n >> m$ and the matrix A has rank(A)=m we can conclude as follows: $AA^T \in \mathbb{R}^{m\times m}$ is invertible and the minimum of (5.2) satisfies

$$A^T Ax = A^T y = A^T(AA^T)(AA^T)^{-1}y \tag{5.5}$$

Therefore, $x = A^T(AA^T)^{-1}y =: A^{\#}y$ is the unique solution.

In abritrary Hilbert spaces (needed by definition of the minimization problem in terms of a scalar product) we define the pseudo-inverse operator as

$$A^{\#} := A^*(AA^*)^{-1} \tag{5.6}$$

Usually we later on assume that the operator A is surjective (corresponds to the rank(A)=m assumption in the above example). We prove the invertability of AA^* and hence $A^{\#}$ is well-defined. Usually, we have to given the precise range and null spaces to define the operators properly. But we are particular interested in operators A that are surjective and hence $R(A) = Y$ is closed. Therefore [2], $N(A)^T = R(A^*)$ and vice versa.

Theorem 5.22. *Let $A \in L(X, Y)$ be surjective. Then AA^* is bijective.*

Proof. We show that $N(AA^*) = \{0\}$. Assume $AA^*y = 0$ for some $y \in Y$. Then $0 =< y, AA^*y >= \|A^*y\|$ and hence $A^*y = 0$. But due to $N(A^*) = R(A)^T = Y^T = \{0\}$ we obtain $y = 0$.

We show that AA^* is onto Y. First, $R(AA^*) \subset R(A) = Y$. It remains to show $R(A) \subset R(AA^*)$. Let $y = Ax$. Exists $z \in Y$ with $A^*z = x$, i.e. $x \in R(A^*) = N(A)^T$. Now decompose $x = x_{N(A)} + x_{N(A)^T}$. Then $\exists z$ such that $A^*z = x_{N(A)^T}$ and further $AA^*z + 0 = AA^*z + Ax_{N(A)} = Ax = y$. This finishes the proof. □

Theorem 5.23. *Let $D \subset X$ be open and non-empty. Let $h : D \to Y$ be continuous Frechet differentiable on D and let $x_0 \in D$. Assume that $Dh(x_0)(\cdot)$ is surjective. Then $Dh(\cdot)()$ is surjective in some $\delta-$neighbourhood of x_0.*

Proof. Idea is to define the following operators (well-defined by the previous discussion):

$$T() := Dh(x_0)\left(Dh(x_0)^*\right)() \tag{5.7a}$$

$$S() := \left(Dh(x)\left(Dh(x)^*\right) - Dh(x_0)\left(Dh(x_0)^*\right)\right)() \tag{5.7b}$$

and to prove by the pertubation theorem that $T + S$ is bijective. This in turn implies that $Dh(x)()$ is bijective. First, see that

$$\|S\| \leq M_1 M_2 \|h(x) - h(x_0)\| \tag{5.8}$$

wherein $M_1 := \max_{x \in B_r(x_0)} \|h(x)\|$ and M_2 is the Lipshitz constant on $B_r(x_0)$ for $Dh(\cdot)()$. Hence

$$\|ST^{-1}\| \leq M_1 M_2 \|x - x_0\| \|T^{-1}\| < 1 \tag{5.9}$$

For more details refer to [11] page 32. □

The last result is given without proof. A proof can be found for example in [11] page 33-35.

Theorem 5.24. *Let $D \subset X$ be open and non-empty. Let $h : D \to Y$ be continuous Frechet differentiable on D and let $x_0 \in D$. Assume that $Dh(x_0)(\cdot)$ is surjective. Then $Dh(\cdot)^{\#}$ is Lipshitz in some $\delta-$neighbourhood of x_0.*

6 Controllability involving ordinary differential equations

This section is mainly taken from [6] but it is also present in many textbooks on controllability and $H^\infty-$calculus.

The problem is stated as follows. We are interested in the controllability of linear ordinary differential equations. We have two times T_0 and T_1 where $T_0 < T_1$ and a matrix $A : (T_0, T_1) \to \mathbb{R}^{n\times n}$ such that $L^\infty((T_0, T_1); \mathbb{R}^{n\times n})$ as well as a matrix $B \in L^\infty((T_0, T_1); \mathbb{R}^{n\times m})$. We define the time–varying linear control system as

$$x'(t) = A(t)x + B(t)u(t). \tag{6.1}$$ coron01

The control is $u \in \mathbb{R}^m$, the state is $x \in \mathbb{R}^n$. For given $x^0 \in \mathbb{R}^n$ and $u \in L^1(T_0, T_1; \mathbb{R}^m)$ the Cauchy problem

$$(6.1) \text{ and } x(T_0) = x^0 \tag{6.2}$$ coron02

has a unique solution $x(t)$ on (T_0, T_1) and such that

$$x \in C^0(T_0, T_1; \mathbb{R}^n).$$

Definition 6.1. *The linear time–varying control system (6.1) is controllable if, for every (x_0, x_1), there exists $u \in L^\infty(T_0, T_1; \mathbb{R}^m)$ such that the solution $x \in C^0(T_0, T_1; \mathbb{R}^n)$ of the Cauchy problem (6.2) satisfies $x(T_1) = x_1$.*

Next, we derive necessary and sufficient conditions such that (6.1) is controllable. To this end we define the resolvent of a homogenous time–varying system

$$x'(t) = A(t)x(t). \tag{6.3}$$ coron03

The resolvent $R(t, \tau)$ is the solution to the homogenous system in the time interval (τ, t).

Definition 6.2. *The resolvent $R : [T_0, T_1]^2 \to \mathbb{R}^{n\times n}$ of the time–varying system*

$$x'(t) = A(t)x(t)$$

is the map

$$R : (t, \tau) \to R(t, \tau)$$

such that for every $\tau \in [T_0, T_1]$ the map $t \to R(t, \tau)$ is the solution to the Cauchyproblem

$$M'(t) = A(t)M(t),\ M(\tau) = Id. \tag{6.4}$$

Note that in the case of $A(t) = A$, the resolvent is given by

$$R(t,\tau) = \exp(A(t-\tau)) = \sum_{k=0}^{\infty} \frac{(t-\tau)^k}{k!} A^k$$

If A is dependent on t and if $A(t)A(s) = A(s)A(t)$ for every s, t then the resolvent is given by

$$R(t,\tau) = \exp(\int_{\tau}^{t} A(s)ds).$$

The solution to the Cauchy problem (6.3) and $x(\tau) = x_0$ is under the assumptions as before given by

$$x(t) = R(t,\tau)x_0 = \exp(\int_{\tau}^{t} A(s)ds)x_0.$$

In general, the following results hold true.

Proposition 6.3. *The resolvent R satisfies*

- $R \in C^0([T_0, T_1]^2; \mathbb{R}^{n\times n})$
- $R(\tau,\tau) = Id.$
- $R(t,\tau)R(\tau,\eta) = R(t,\eta).$
- *If $A \in C^0([T_0, T_1]; \mathbb{R}^{n\times n})$, then $R \in C^1([T_0, T_1]^2; \mathbb{R}^{n\times n})$ and we have*

$$R_t(t,\tau) = A(t)R(t,\tau),\ R_\tau(t,\tau) = -R(t,\tau)A(\tau).$$

Furthermore, R is invertible and it's inverse is the resolvent with interchanged arguments.

$$R(t,\tau)R(\tau,t) = Id.$$

The proof is immediate. The previous proposition is obtained from the definition and the derivative with respect to τ follows from the formula for the inverse of R. The regularity properties follow from the Picard–Lindelöf Theorem.

Proposition 6.4. *The solution of the Cauchy problem (6.2) is given by*

$$x(t) = R(t,T_0)x^0 + \int_{T_0}^{t} R(t,s)B(s)u(s)ds \tag{6.5}$$ coron05

for all $t \in [T_0, T_1]$.

Indeed, a simple computation shows that

$$x'(t) = A(t)R(t,T_0)x^0 + R(t,t)B(t)u(t) + \int_{T_0}^{t} A(t)R(t,s)B(s)u(s)ds$$
$$= A(t)\left(R(t,T_0)x^0 + \int_{T_0}^{t} R(t,s)B(s)u(s)ds\right) + B(t)u(t).$$

Definition 6.5. *The controllability Gramian of the control system (6.1) is the symmetric $n \times n$ matrix*

$$\mathcal{C} = \int_{T_0}^{T_1} R(T_1,\tau)B(\tau)B^T(\tau)R(T_1,\tau)^T d\tau. \tag{6.6}$$ coron04

Clearly $\mathcal{C}$ is symmetric. Furthermore, it is a non–negative matrix, since

$$x^T\mathcal{C}x = \int_{T_0}^{T_1} x^T R(T_1,\tau)B(\tau)B^T(\tau)R(T_1,\tau)^T x d\tau = \int_{T_0}^{T_1} \|xR(T_1,\tau)B(\tau)\|^2 d\tau \geq 0.$$

Theorem 6.6. *The linear time varying control system (6.1) is controllable if and only if its controllability Gramian $\mathcal{C}$ is invertible.*

Proof. Assume $\mathcal{C}$ is invertible. Set for a.e. $\tau \in (T_0, T_1)$

$$u(\tau) = B^T(\tau)R(T_1,\tau)^T\mathcal{C}^{-1}(x_1 - R(T_1,T_0)x_0).$$

Then, the solution $x(\tau)$ to (6.1) with the previous conrol is given by (6.5). Hence,

$$x(T_1) = R(T_1,T_0)x^0 + \int_{T_0}^{t} R(T_1,s)B(s)u(s)ds$$
$$= R(T_1,T_0)x^0 + \int_{T_0}^{t} R(T_1,s)B(s)B^T(s)R(T_1,s)^T\mathcal{C}^{-1}(x_1 - R(T_1,T_0)x_0)ds$$
$$= R(T_1,T_0)x^0 + \mathcal{C}\,\mathcal{C}^{-1}(x_1 - R(T_1,T_0)x_0) = x_1.$$

Assume $\mathcal{C}$ is not invertible. We show that the system is not controllable. If $\mathcal{C}$ is not invertible, then there exists $y \neq 0$ such that $\mathcal{C}y = 0$. Therefore, $y^T\mathcal{C}y = 0$ which is

$$0 = \int_{T_0}^{T_1} y^T R(T_1,s)B(s)B^T(s)R(T_1,s)^T y ds$$
$$= \int_{T_0}^{T_1} \|B^T(s)R(T_1,s)^T y\|^2 ds$$
$$\implies y^T B(s)R(T_1,s) = 0 \ \text{ a.e. } s \in (T_0,T_1).$$

Let $u \in L^1(T_0, T_1; \mathbb{R}^m)$, $x(T_0) = 0$ and let $x \in C^0(T_0, T_1; \mathbb{R}^n)$ be the solution to (6.1). Then,

$$y^T x(T_1) = \int_{T_0}^{T_1} y^T R(T_1, s) B(s) u(s) ds = 0.$$

Since $y \neq 0$, there exists x^1 such that $y^T x^1 \neq 0$, e.g., $x^1 = y$. Hence, whatever the control u is , it can never reach $x^1 = x(T_1)$, since $y^T x(T_1) = 0$. This contradicts the controllability. □The exact control u defined in the proof has also a different interpretation. It is the optimal control with minimal L^2-norm.

Proposition 6.7. *Let $(x^0, x^1) \in \mathbb{R}^n \times \mathbb{R}^n$ and let $u \in L^2(T_0, T_1; \mathbb{R}^m)$ be such that the solution of the Cauchy problem (6.2) satisfies $x(T_1) = x^1$. Then,*

$$\int_{T_0}^{T_1} \|\bar{u}(t)\|^2 dt \leq \int_{T_0}^{T_1} \|u(t)\|^2 dt. \tag{6.7}$$

coron06

Proof. Let $v = u - \bar{u}$. Then, $\bar{x}$ and x being the solutions of the Cauchy problems (6.2) with control u and $\bar{u}$, respectively. The terminal states are $\bar{x}(T_1) = x(T_1) = x^1$. Then,

$$\begin{aligned} \int_{T_0}^{T_1} R(T_1, t) B(t) v(t) dt &= \int_{T_0}^{T_1} R(T_1, t) B(t) u(t) dt - \int_{T_0}^{T_1} R(T_1, t) B(t) \bar{u}(t) dt \\ &= \big(x_1 - R(T_1, T_0) x(T_0)\big) - \big(\bar{x}_1 - R(T_1, T_0) x(T_0)\big) \\ &= 0 \end{aligned}$$

Due to the parallelogram equality we have for $\bar{u}$ and u in L^2-norm

$$\|u\|^2_{L^2(T_0,T_1)} = \|\bar{u}\|^2_{L^2(T_0,T_1)} + \|v\|^2_{L^2(T_0,T_1)} + 2 \int_{T_0}^{T_1} \bar{u}^T v dt.$$

The scalar product is evaulated as follows

$$\int_{T_0}^{T_1} \bar{u}^T v dt = (x^1 - R(T_1, T_0) x^0)^T \mathcal{C}^{-1} \int_{T_0}^{T_1} R(T_1, s) B(s) v(s) ds = 0$$

Hence, the result follows from the parallelogram equality. □

Note, that (6.7) holds true with equality if and only if $\bar{u} = u$ a.e. in $t \in (T_0, T_1)$. The necessary and sufficient condition for controllability requires the compuation of the matrix $\mathcal{C}$ which is quite difficult, since it involves the full evolution of the pde. Next, we discuss the so–called Kalman rank condition for controllability. It is also a necessary and sufficient condition for controllability in the case of linear time invariant systems. However, it is only necessary in the case of time variant systems.

Theorem 6.8. *The time invariant linear control systems*

$$x'(t) = Ax(t) + Bu(t), \tag{6.8}$$

is controllable on $[T_0, T_1]$ *if and only if*

$$rank\left(B, AB, A^2B, \ldots, A^{n-1}B\right) = n. \tag{6.9}$$ coron07

We need a few basic facts from linear algebra to prove the result. Assume we have a linear operator $A \in \mathbb{R}^{n\times m}$ with full rank, $rank(A) = n$. If the $rank(A) < n$, then there exists $x \in \mathbb{R}^m \neq 0$ with $Ax = 0$. If $rank(A) < n$, then the image of A is a linear subspace in $\mathbb{R}^n$. Therefore, there exists a vector $y \in \mathbb{R}^n$ that is not in the image of A. Since $\mathbb{R}^n$ is a Hilbert space we can compute the orthogonal projection of y on the image of A. Hence, there exists $\bar{y}^T Ax = 0$ for all $x \in \mathbb{R}^n$.

Proof. Since A does not dependent on time we have $R(t,\tau) = \exp((t-\tau)A)$ for all $\tau, t \in [T_0, T_1]$. Hence,

$$\mathcal{C} = \int_{T_0}^{T_1} \exp(T_1 - \tau)ABB^T \exp(T_1 - \tau)Ad\tau.$$

Let us first assume that the time–invariant system is not controllable. Then, $\mathcal{C}$ is not invertible and there exists $y \neq 0$ and $y \in \mathbb{R}^n$ such that

$$0 = \mathcal{C}y \implies 0 = y^T\mathcal{C}y = \int_{T_0}^{T_1} y^T \exp(T_1 - \tau)ABB^T \exp(T_1 - \tau)Ayd\tau.$$

Therefore for all τ,

$$B^T \exp(T_1 - \tau)Ay = 0.$$

If we denote by

$$k(\tau) = y^T \exp(T_1 - \tau)A^T B,\ \tau \in [T_0, T_1]$$

we observe $k = 0$. Differentiating with respect to τ we observe

$$k^{(i)}(\tau) = (-1)^i y^T A^i B = 0,\ i \in \mathbb{N},\ \tau \in [T_0, T_1].$$

Hence, there exists $y \neq 0$ such that $y^T A^i B = 0, i =, \ldots, n-1$ and therefore (6.9) does not hold true.

To prove the converse assume that (6.9) does not hold true. Hence, there exists $y \neq 0$ with $y \in \mathbb{R}^n$ and $y^T A^i B = 0$ for $i = 0, \ldots, n-1$. The Caley-Hamilton Theorem states that for $p(\lambda) = det(A - \lambda Id)$ we have $p(A) = 0$.

$p(A) = \sum_{k=0}^{n} a_k A^k$ is a polynomial in A of degree n. Due to our assumption we and since $p(A) = 0$ we have $y^T A^n B = 0$. Furthermore, $p(A^{n+m}) = p(A^n)A^m = 0$ for all $m \geq 1$. Therefore, $k^i(\tau) = (-1)^i y^T \exp(T_1 - \tau)A^i B = 0$ for $\tau = T_1$ and all $i \in \mathbb{N}$. k is the exponential function that is real analytic: it therefore allows an expansion as infinite Taylor series. Choosing as point for the expansion T_1 we observe $k \equiv 0$. Since $k \equiv 0$, we have $y^T \mathcal{C} y = 0$ for $y \neq 0$. If $0 \leq \|y^T \mathcal{C} z\| \leq \left(y^T \mathcal{C} y\right)^{\frac{1}{2}} \left(z^T \mathcal{C} z\right)^{\frac{1}{2}} = 0$ for all $z \in \mathbb{R}^n$. Then, $\mathcal{C} y = 0$. Therefore, $\mathcal{C}$ is not invertible. Hence, the system is not controllable. □Next, we discuss time–varying linear control systems. We assume that A and B are of class C^∞ on $[T_0, T_1]$. Let us define a sequence of maps $B_i : C^\infty(T_0, T_1; \mathbb{R}^{m \times n})$ in the following way

$$\begin{aligned} B_0(t) &= B(t), \\ B_i(t) &= B_{i-1}(t)' - A(t)B_{i-1}(t). \end{aligned}$$

Theorem 6.9. *Assume that for some $\bar{t} \in [T_0, T_1]$*

$$\textit{rank } \left\{B_i(\bar{t}),\ i \in \mathbb{N}\right\} = n.$$

Then, the linear control system $x' = A(t)x + B(t)u$ is controllable.

Proof. We assume that the linear control system $x' = A(t)x + B(t)u$ is not controllable. Then, there exists $\mathcal{C} y = 0$ and $y \neq 0$. Hence,

$$\begin{aligned} 0 = y^T \mathcal{C} y &= \int_{T_0}^{T_1} \|B(\tau)^T R(T_1, \tau)^T y\|^2 d\tau \\ &\implies B(\tau)^T R(T_1, \tau)^T y = 0 \ \forall \tau \\ 0 &= B(\tau)^T \left(R(T_1, \bar{t}) R(\bar{t}, \tau)\right)^T y \\ &= B(\tau)^T R(\bar{t}, \tau)^T R(T_1, \bar{t})^T y \\ &= z^T R(\bar{t}, \tau) B(\tau) = K(\tau) \end{aligned}$$

with $z \neq 0$, since R is invertible. Furthermore, $K^{(i)}(\tau) = z^T R(\bar{t}, \tau) B_i(\tau) = 0$ for all $i \in \mathbb{N}$ and all τ. Since R is invertible, there exists $w = R(\bar{t}, \tau) z \neq 0$ such that $w^T B_i(\tau) = 0$ and therefore the rank of $(B_i(\bar{t}))$ is not maximal. □

7 Optimal control problems involving ordinary differential equations

This section is mainly taken from [4], but it is also present in many textbooks on optimal control theory. This part is the classical first extension of finite–dimensional optimization problems to the infinite dimensional spaces.

The problem is stated as follows. We are interested in the evolution of a deterministic system described by an ordinary differential equation with $x \in \mathbb{R}^n$. We assume that the evolution can be influenced by an external input u. All functions are time–dependent, i.e., $x = x(t)$ and $u = u(t)$. We assume that an appropriate model is then provided by the controlled system. We do not write the dependence on t, if it is obvious.

$$x'(t) = f(x, u), x(0) = x_0 \tag{7.1}$$ 6.1

We furthermore assume that the control $u : [0, T] \to U \subset \mathbb{R}^m$ takes values inside a given set U. This set is not necessarily bounded. The set of admissible controls is denoted by $\mathcal{U}$ and contains all measurable functions u such that $u(t) \in U$ for a.e. t. To guarantee local existence and uniqueness of solutions $x(t)$ to (7.1) it is natural to assume

A $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$is Lipschitz wrt to x and continuous wrt to u. (7.2)

For any initial condition local existence for any control u is then given by the Theorem of Picard-Lindelöf. We will denote by

$$t \to x(t, t_0, x_0, u)$$

the solution to the Cauchy problem (7.1) with initial condition

$$x(t_0) = x_0 \tag{7.3}$$ 6.2

Since $u \in U$ we obtain a family of trajectories $x(t; t_0, x_0, u)$.

Example 7.1. *Call $x(t)$ the position of a boat in a river and let $v(x) \in \mathbb{R}^2$ be the velocity of the water at the point x. Assume that the boat is powered by an engine giving at most speed ρ and a steering wheel such that it can move in any direction u, $\|u\| \leq \rho$. Then, the evolution of the boat can be modelled by*

$$x'(t) = v(x) + u, \|u\|_2 \leq \rho.$$

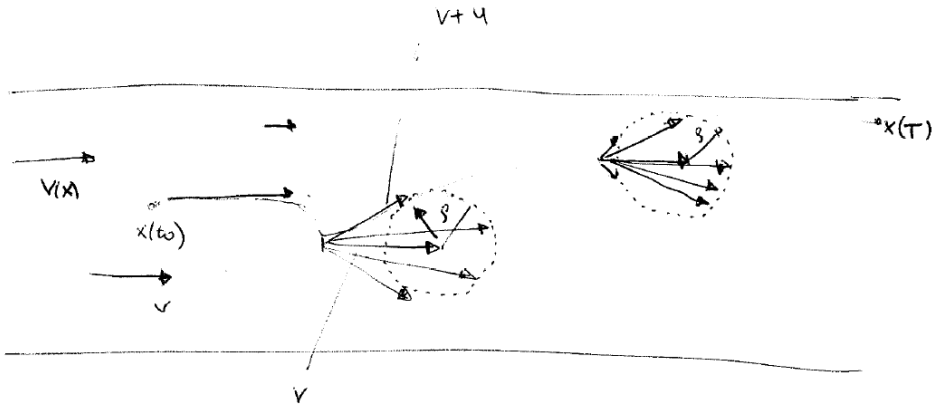


Figure 9: Field of the ode and sample trajectory. Figure 6.1

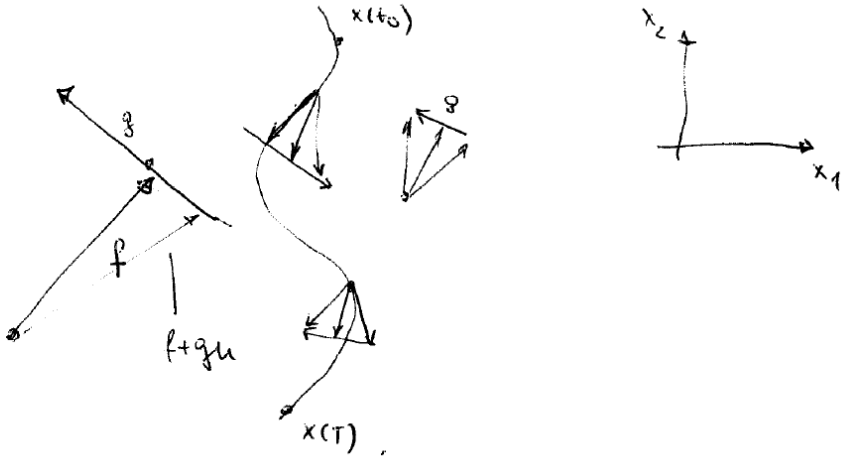


Figure 10: Field of the ode with x-dependent shift and sample trajectory. f is the flow field and $x' = f(x) + g(x)u$. Figure 6.2

If set for example $\rho = 1$ and assume an x-dependent shift whenever a control is applied we obtain the very important class of control systems

$$x'(t) = f(x) + g(x)u,\ u \in [-1, 1]$$

Example 7.2. *Consider a mechanism, such as a crane or trolley, of mass m which moves along a horizontal track without friction. If $x(t)$ represents the position at time t, we assume the motion of the trolley is governed by*

$$m\ddot{x}(t) = u(t), \quad t > 0, \tag{7.4}$$

eqn:trolley

where $u(t)$ is an external controlling force that is applied to the trolley.

Assume initial position and velocity of trolley are given by $x(0) = x_0$, $x'(0) = y_0$, respectively. We wish to choose a function u (which is naturally enough called a control function) to bring the trolley to rest at the origin in minimum time.

Physical Restrictions: controlling force must be bounded in magnitude: $|u(t)| \leq M$.

Take $m = M = 1$, and rewrite equation (7.4):

$$x_1'(t) = x_2(t); \quad x_2'(t) = u(t)$$

where now $x_1(t)$ and $x_2(t)$ are now the position and velocity of the body at time t:

or

$$x'(t) = Ax(t) + bu(t), \quad x(0) = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \tag{7.5}$$

eqn:trolley1

where $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$

Control Problem:

- *Find a function u, subject to (7.5), which brings the solution of (7.5), $x(t)$ to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ in minimum time t.*
- *Any control that steers us to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ in minimum time is called an optimal control, e.g., a period of maximum acceleration ($u = +1$) and then maximum braking ($u = -1$).*

Example 7.3. *A control surface on an aircraft is to be kept at rest at a fixed position. A wind gust displaces the surface from the desired position. Assume that if nothing were done, the control surface would behave as a damped harmonic oscillator, i.e. if θ measures deviation from a desired position, then free motion of surface satisfies:*

$$\ddot{\theta} + a\dot{\theta} + \omega 2\theta = 0; \tag{7.6}$$

$$\theta(0) = \theta_0 \text{ displacement due to wind gust;} \tag{7.7}$$

$$\dot{\theta}(0) = \dot{\theta}_0 \text{ velocity imparted to surface by gust.} \tag{7.8}$$

On an aircraft the oscillation of the control surface can not be permitted, and so we need to design a servo-mechanism to apply a restoring torque and bring surface back to rest in minimum time.

$$\ddot{\theta} + a\dot{\theta} + \omega 2\theta = u(t); \tag{7.9}$$

$$\theta(0) = \theta_0 \quad \dot{\theta(0)} = \dot{\theta}_0. \tag{7.10}$$

where $u(t)$ represents restoring torque at time t. Assume $|u(t)| \leq C$, where C is a constant, which can be normalised to 1.

Problem: Find u such that the system will be brought to $\theta = 0$, $\dot{\theta} = 0$ in minimum time.

Assume $\theta > 0$ and $\dot{\theta} > 0$ then torque must be directed to negative θ and should be as large as possible: $u(t) = -1$.

If we let $u(t) = -1$ for too long we might overshoot the terminal conditions $\theta = 0$, $\dot{\theta} = 0$. Thus at some point we need $u(t) = +1$ inorder to brake i.e. we are led to controls that take on (only) values (± 1); such controls are called bang-bang controls.

Note: Setting $x_1 = \theta$ and $x_2 = \dot{\theta}$ we have

$$x_1 = \theta; \quad \dot{x}_1(0) = \dot{\theta}_0 \tag{7.11}$$

$$\dot{x}_2 = -ax_2 - \omega 2x_1 + u; \quad x_2(0) = \dot{\theta}_0 \tag{7.12}$$

$$\dot{x} = Ax + bu, \quad x(0) = \begin{bmatrix} \theta_0 \\ \dot{\theta}_0 \end{bmatrix} \tag{7.13}$$

where $A = \begin{bmatrix} 0 & 1 \\ \omega 2 & -a \end{bmatrix}$, $b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and u is chosen with $|u(t)| \leq 1$ and to minimize $J(u) = \int_0^T dt$ where T is any time for which $x_1(T) = 0$ and $x_2(T) = 0$.

Example 7.4. *Let $x(t)$ be the amount of steel produced by a mill at time t. The amount produced at time t is to be allocated to one of the two uses: production of consumer products; or investment.*

Note: Steel located to investment is used to increase productive capacity - using steel to produce new steel mills, transport facilities, or whatever.

Let $u(t)$, where $0 \leq u(t) \leq 1$, denote the fraction of steel produced at time t that is allocated to investment. Then $1 - u(t)$ is for consumption.

The assumption that re-invested steel is used to increase the productive capacity could be written as:

$$\frac{dx}{dt} = ku(t)x(t),$$

where $x(0) = C$, the initial endowment, k is an appropriate constant (i.e. rate of increase in production is proprtional to amount located to investment).

Problem: Choose $u(t)$ so as to maximise the total consumption over a fixed $T > 0$, i.e. maximise

$$J(u) = \int_0^T (1 - u(t))x(t)\, dt$$

Next, we introduce the notion of the reachable or attainable set at time T starting from x_0 at time t_0. It consists of all points $x \in \mathbb{R}^n$ which can be connected through a trajectory to x_0 during the time $[t_0, T]$ using some control $u \in U$.

$$R(T) := \{x(T; t_0, x_0, u),\ u \in U\} \tag{7.14}$$ 6.4

The control is called an **open loop** control if it is a function $u = u(t)$ and a **closed loop** or feedback control if it is of the type $u = u(x)$. There is a variety of literature on the study of control systems (7.1), e.g.,

1. Starting from a given state x_0 describe the set $R(T)$ and study its properties. This is concerned with the dynamics of the ODE.

2. For each initial state x_0 find a control $u(\cdot)$ that steers the system towards the origin, so that for $t \to \infty$ we obtain $x(t, 0, u) \to 0$. This is called stabilization. Alternatively, one can steer the system at rest ($x' = 0$), i.e., to states $\hat{x}$ such that $f(\hat{x}, u) = 0$.

   Preferably, a control satisfying stabilization should be of the form $u = u(x)$. Then, the evolution of the trajectory is

   $$x'(t) = f(x, u(x)).$$

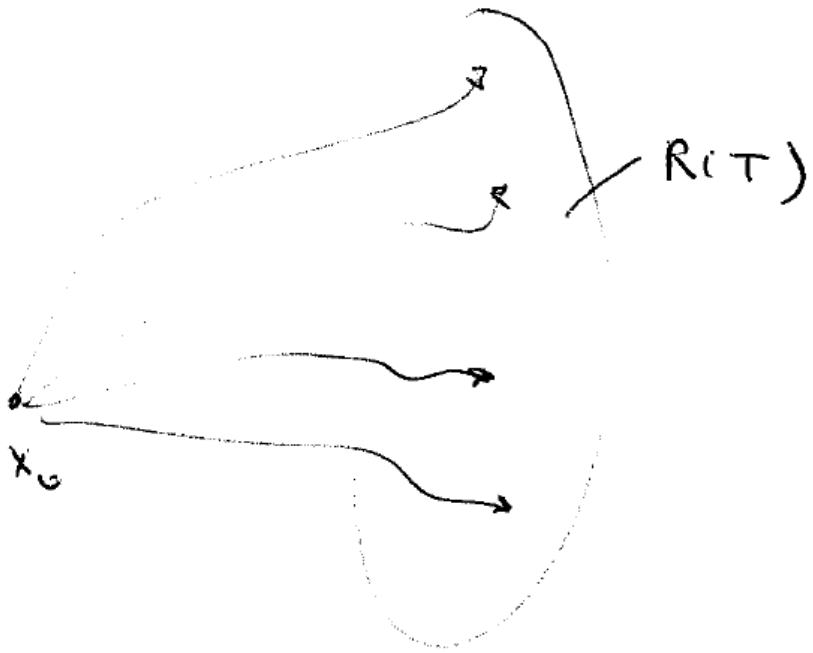


Figure 11: Reachable set $R(T)$ Figure 6.3

In the engineering community $u(x)$ is also called a control law or feedback controller. It gives for every state x a control u such that finally the system will be at rest. No further optimization is needed. Explicit control laws $u(x)$ can be computed in special cases only, e.g., linear controlled systems. This is dealt with in $H^\infty-$calculus.

3. Find a control $u \in U$ being opimal with respect to a certain cost function. For example, given the initial condition (7.3), one may seek to minimize the cost

$$J(u) = \int_{t_0}^{T} h(x(t), u(t))dt + \phi(x(T))$$

over all control functions $u \in U$ and with $x = x(t, t_0, x_0, u)$. The running cost h and the terminal cost ϕ are assumed to be continuous functions. This is a problem of optimal control and can be solved using Pontryagin's Maximum Principle.

Clearly, by using the term 'minimal value' of the cost functional we mean the analogon of the finite dimensional problem, i.e.,

$$J(u^*) \leq J(u), \quad \forall u \in \Omega[t_0, T]$$

thus there exists $\delta > 0$ such that $J(u^*) \leq J(u)$, $\forall u \in B_\delta(u^*) \cap \Omega[t_0, T]$.

Hence we require a norm for piecewise continuous functions, for example:

$$\|u\|_\infty := \sup_{t \in \cup_{k=0}^{N}(\theta_k, \theta_{k+1})} \|u\|$$

with $t_0 = \theta_0 < \theta_1 < \ldots < \theta_N < \theta_{N+1} = T$ being a partition for u. If we assume that the controls are continuously differentiable between two successive discontinuities $[\theta_k, \theta_{k+1}]$, $k = 0, \ldots, N$ another norm would be:

$$\|u\|_{1,\infty} := \sup_{t \in \cup_{k=0}^{N}(\theta_k, \theta_{k+1})} \|u(t)\| + \sup_{t \in \cup_{k=0}^{N}(\theta_k, \theta_{k+1})} \|u'(t)\|$$

7.1 Pontryagin Maximum Principle

We discuss Pontryagin's maximum principle on a particular control problem. This problem is called Mayer's problem.

$$x' = f(x, u),\ u \in U,\ t \in [0, T],\ x(0) = x_0 \tag{7.15}$$ 7.1

$$\max_{u \in U} \psi(x(T, u)) \tag{7.16}$$ 7.2

There is no running cost involved and we only measure the terminal payoff over all admissible controls. We seek now necessary optimality conditions. To this end assume that $t \to u^*$ is an optimal control function and x^* the corresponding optimal trajectory, i.e., $x^* = x(t; 0, x_0, u^*)$.

In order to obtain necessary optimality conditions we proceed as in the finite dimensional case. We study perturbations of the optimal control and the behaviour of the cost functional on these perturbations. Let $x(t)$ be a solution to

$$x'(t) = g(t, x) \tag{7.17}$$ 7.3

where g is measurable wrt t and continuosly differentiable wrt to x. Later g will be given as $g(x, t) = f(x, u(t))$ for some given control u.

Consider a family $\epsilon \to x_\epsilon$ of 'nearby' solutios $x_\epsilon(t)$ as depicted in Figure 12. These canbe obtained for example by using different starting points or due to different control functions at $t = s$. We have $x'_\epsilon = g(t, x_\epsilon)$.

At a fixed time s we are interested in the dependence of this family of solutions on ϵ. We study

$$\lim_{\epsilon \to 0} \frac{x_\epsilon(s) - x(s)}{\epsilon} = v(s).$$

If we assume that this limit exists at some time s, then it exists for any time t and the function $t \to v(t)$ given by

$$\lim_{\epsilon \to 0} \frac{x_\epsilon(t) - x(t)}{\epsilon} = v(t).$$

is well–defined. In fact, we have for any ϵ and g Lipschitz in x due to Gronwall's inequality

$$\|x_\epsilon(t) - x(t)\| \leq \exp(L|t - s|)\|x_\epsilon(s) - x(s)\|$$

and hence the previous limit exists provided that the limit at $x = s$ exists.

Furthermore, the *linearized tangent vector* $v(t)$ satisfies the linearized evolution equation

$$v'(t) = A(t)v(t) \tag{7.18}$$ 7.4

$$A(t) = D_x g(t, x(t)) \tag{7.19}$$ 7.5

This holds true, since we have $x_\epsilon(t) = \int_s^t g(\tau, x_\epsilon(\tau))d\tau + x_\epsilon(s)$ and

$$\begin{aligned} v(t) &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_s^t D_x g(\tau, x(\tau))(x_\epsilon(\tau) - x(\tau))d\tau + O(\|x_\epsilon - x\|^2) + v(s) \\ &= \lim_{\epsilon \to 0} \int_s^t D_x g(\tau, x(\tau))v(\tau)d\tau + O(\|\epsilon v\|^2) + v(s). \end{aligned}$$

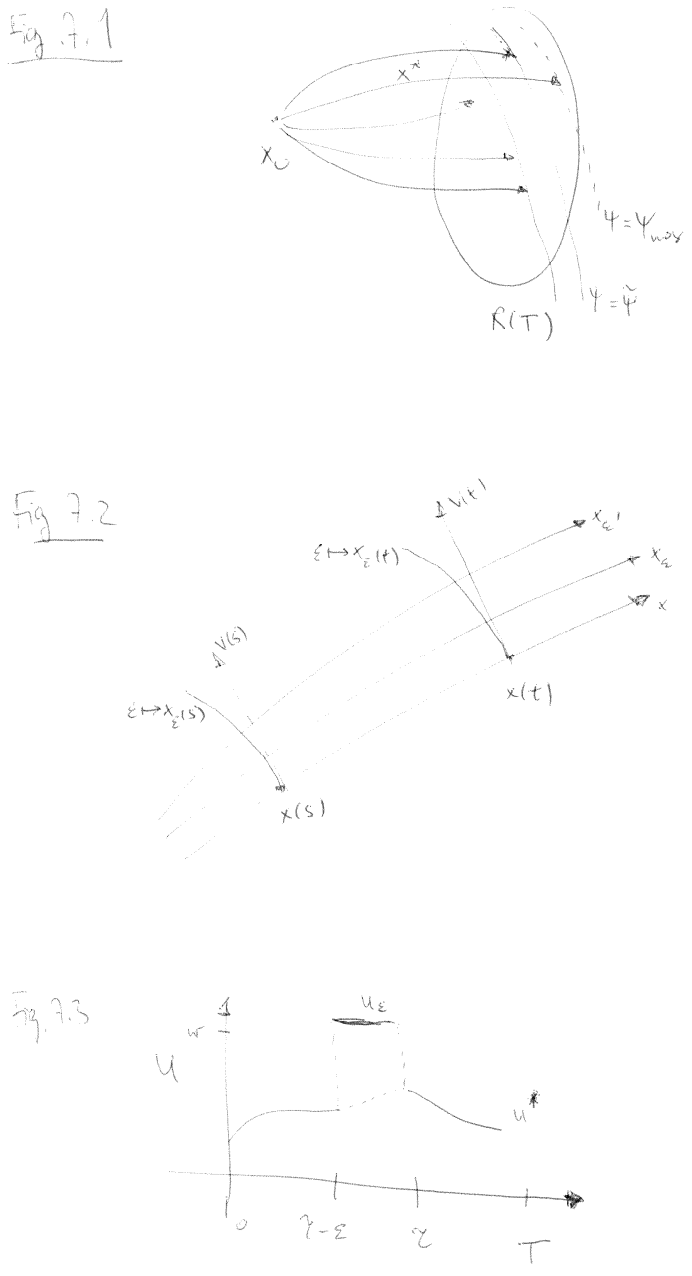


Figure 12: Optimal trajectory and isolines of ψ..Optimal control and needle perturbation on interval of length ϵ.

Figure 7.3 Figure 7.1

The equations (7.18,7.19) are the sensitivity equations, since they describe variations of the solution. A different system is the *adjoint equation* given by

$$p'(t) = -p(t)A(t) \Leftrightarrow \bar{p}' = -A^T(t)\bar{p} \tag{7.20}$$ 7.6

where $p \in \mathbb{R}^{1\times n}$ and $p^T = \bar{p}$. The matrix $A = D_x g(t, x(t))$. The adjoint system will be derived later formally using the Lagrange calculus. A solution $p(t)$ to the adjoint system and a solution $v(t) \in \mathbb{R}^{n\times 1}$ to the system (7.18) satisfy

$$\frac{d}{dt}\left(\bar{p}\cdot v\right) = p'(t)v(t) + p(t)v'(t) = 0. \tag{7.21}$$ 7.7

Therefore, the product of adjoint and sensitivity is constant over time.

The previous calculations hold true for any perturbation of the trajectory x and only require $v(s)$ to exist at some time s. In order to derive necessary optimality conditions for u^* we study particular perturbations of the control u^* and the corresponding perturbed trajectories. If u^* is the optimal control, then any perturbation cannot increase the optimal payoff $\psi(x(T; u^*))$. Fix a time $\tau \in [0, T]$ and a control value $\omega \in U$. For $\epsilon > 0$ small, we consider the needle variation $u_\epsilon \in U$ given by

$$u_\epsilon(t) = \begin{cases} \omega & t \in [\tau - \epsilon, \tau] \\ u^*(t) & \text{else} \end{cases} \tag{7.22}$$ 7.8

Call $t \to x_\epsilon(t) \equiv x(t, 0, x_0, u_\epsilon)$ the perturbed trajectory. We compute the terminal point $x_\epsilon(T)$ and study the variation of ψ. ψ shall not increase due to this perturbation. We have $x_\epsilon(\tau - \epsilon) = x^*(\tau - \epsilon)$, since the pertubation starts acting at $\tau - \epsilon$. On a small time interval $[\tau - \epsilon, \tau]$ we consider the first–order Taylor expansion as

$$x_\epsilon(\tau - \epsilon) = x_\epsilon(\tau) - \epsilon x'_\epsilon(\tau) + O(\epsilon^2)$$

or

$$x_\epsilon(\tau) = x_\epsilon(\tau - \epsilon) + \epsilon x'_\epsilon(\tau) + O(\epsilon^2)$$

and

$$x^*(\tau) = x^*(\tau - \epsilon) + \epsilon x^*(\tau) + O(\epsilon^2).$$

Therefore, $x^*(\tau) = x_\epsilon + O(\epsilon)$. Since $(x^*)' = f(x^*, u^*)$ and $x'_\epsilon = f(x_\epsilon, \omega) = f(x^*, \omega) + f_x(\cdot)O(\epsilon)$. If u^* is continuous in τ, then $v(\tau)$ exists, since we obtain

$$v(\tau) = \lim \frac{x_\epsilon(\tau) - x^*(\tau)}{\epsilon} = f(x^*(\tau), \omega) - f(x^*(\tau), u^*(\tau)) \qquad (7.23) \quad \boxed{7.9}$$

If v exists at time τ, then due to the previous computations it exists for $t \geq \tau$ with

$$A(t) = D_x f(x^*(t), u^*(t)).$$

Since for $t \geq \tau$, $u_\epsilon \equiv u^*$ the evolution of the tangent vector $v(t)$ is given by the linear equation

$$v'(t) = A(t)v(t), t \in [\tau, T] \qquad (7.24) \quad \boxed{7.10}$$

By maximility of ψ at x^* we have $\psi(x_\epsilon(T)) \leq \psi(x^*(T))$ and using the mean value theorem $0 \geq \frac{1}{\epsilon}\left(\psi(x_\epsilon(T)) - \psi(x^*(T))\right) = \nabla\psi(\xi)\left(\psi(x_\epsilon(T)) \leq \psi(x^*(T))\right) \rightarrow \nabla\psi(x^*(T))v(T) \leq 0$.

$$\nabla\psi(x^*(T))v(T) \leq 0 \qquad (7.25) \quad \boxed{7.11}$$

For every time τ where u^* is continuous and every admissible value $\omega \in U$ we generate the vector

$$v(\tau) = f(x^*(\tau), \omega) - f(x^*(\tau), u^*(\tau))$$

and propagate it forward in time until $t = T$ by solving the linearized equation (7.24). Then, the inequality (7.25) is necessary for optimality. This requires to propagate the infinitely many vectors $v(\tau)$ forward, we can use the adjoint equation (7.21) to propagate ψ backwards in time, since we have $\frac{d}{dt}\hat{p} \cdot v = 0$. We solve

$$p'(t) = -p(t)A(t),\ p(T) = \nabla\psi(\psi^*(T)) \qquad (7.26) \quad \boxed{7.12}$$

and hence

$$0 \geq p(T)v(T) = p(\tau)v(\tau) = p(\tau)\left(f(x^*(\tau), \omega) - f(x^*(\tau), u^*(\tau))\right)\ \forall\omega, \tau.$$

This yields

$$\max_{\omega \in U}\left(p(\tau)f(x^*(\tau), \omega)\right) = p(\tau)f(x^*(\tau), u^*(\tau)) = p(\tau)x^{*\,\prime}(\tau) \qquad (7.27) \quad \boxed{7.13}$$

and therefore for every time $\tau \in (0, T)$ where u^* is continuous, the speed $x'(\tau)$ corresponding to the optimal control is the one with the inner product with p being as large as possible. The result can be extended to any τ being a Lebesgue point of u^* and f only Lipschitz in x.

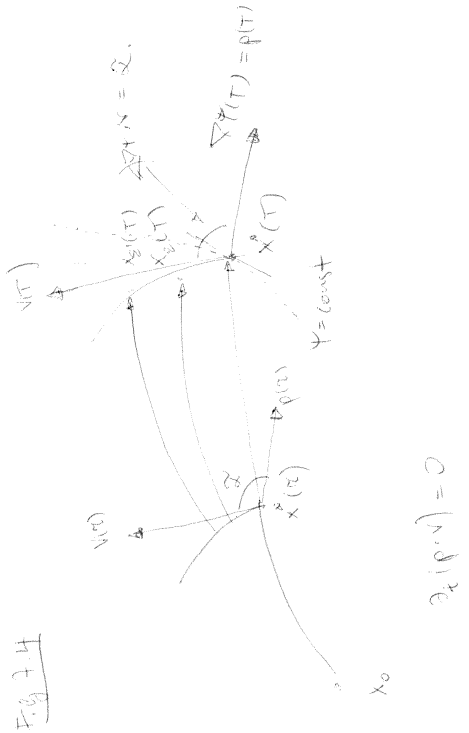


Figure 13: Trajectories for different controls u_ϵ and vectors $v(\cdot)$ and $p(\cdot)$ with $p(T) := \nabla\psi(x^*(T))$. Figure 7.4

Theorem 7.1

Theorem 7.5. *[Pontryagin Maximum Principle for Mayer's problem] Consider the problem $x' = f(x,u)$, $u(t) \in U$ und $x(0) = x_0$. Let u^* be an optimal control for the maximization problem*

$$\max_{u\in U} \psi(x(T)).$$

Define the vector $p(t)$ as the solution to the adoint system $p'(t) = -p(t)D_x f(x^*(t), u^*(t))$ with terminal condition $p(T) = \nabla\psi(x^*(T))$.

Then for almost every $\tau \in (0,T)$ the following necessary first–order optimality condition holds

$$\max_{\omega\in U} \big(p(\tau)f(x^*(\tau),\omega)\big) = p(\tau)f(x^*(\tau),u^*(\tau))$$

Pontryagin's maximum principle allows for the following computation of an optimal control: At first solve the pointwise maximization problem (7.27) for u and fixed values x and p. We obtain a highly nonlinear, possibly mulit-valued function $u(x,p)$ given by

$$\bar{u}(x,p) = \operatorname{argmax}_\omega \big(pf(x,\omega)\big).$$

7.15

Continue by solving the two-point boundary value problem

$$x' = f(x,\bar{u}(x,p)), x(0) = x_0, \tag{7.28a}$$
$$p = -pD_x f(x,u(x,p)), p(T) = \nabla\psi(x(T)) \tag{7.28b}$$

A two–point boundary value problem (7.28) can be solved by the shooting method. Guess an initial value $p(0) = p_0$ solve the forward problem in x and p. Check, whether we reach the final state $p(T)$ and increase or decrease the initial value depending whether we overshoot or not. Typically, adding conditions on $\hat{u}$ yields so–called bang–bang controls, i.e., $\hat{u}$ either attains the maximial or the minimial value in U and any time $\tau \in (0,T)$.

7.2 Extensions to Pontryagin's maximum principle – Running costs

In conection with the control system (7.15) more general optimization problems can be considered. We discuss the following extension involving not only a terminal payoff but also a runnig cost depending on time, state $x(t;u)$ and/or control $u(t)$.

$$\max_{u\in U} \psi(x(T;u)) - \int_0^T h(t,x(t;u),u(t))dt$$

This problem can be reduced to Mayer's problem by setting

$$x'_{n+1}(t) := h(t, x(t), u(t)),\ x_{n+1}(0) = 0$$

and solving

$$\max_{u \in U} \psi(x(T; u)) - x_{n+1}(T; u).$$

7.3 Extensions to Pontryagin's maximum principle – Terminal constraints

In connection with the control problem (7.15) and (7.16) we additionally ask the trajectory to satisfy terminal constraints. We are interested in solving

$$\max_{u \in U} \psi(x(T; u)) \tag{7.29}$$ 8.6

subject to

$$x' = f(t, x, u),\ u(t) \in U,\ t \in (0, T) \tag{7.30}$$ 8.7

and initial and terminal constraints

$$x(0) = x_0,\ g_i(x(T)) = 0,\ i = 1, \ldots, m. \tag{7.31}$$ 8.8

The idea is to treat this problem as Mayer's problem with additional algebraic equality constraints. We are interested in *necessary* first–order conditions. A first idea on how to derive these conditions is given by comparing with the finite–dimensional case. The most simple to prove theorem on necessary optimality conditions is due to Fritz John (Theorem 4.5). Specialized to our case it reads

Theorem 7.6. *Let $x^* \in \mathfrak{S} := \{x \in \mathbb{R}^n : h_i(x) = 0,\ i = 1, \ldots, m\}$ be a local maximizer of the function $\psi : \mathbb{R}^n \to \mathbb{R}$ on x^*. Then, there exists multipliers $\lambda_i^* \in \mathbb{R}$ and $\lambda_0 \geq 0$ such that $\|\lambda\| \neq 0$ and*

$$\lambda_0 \nabla \psi(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0.$$

There is also a geometric interpretation of this condition: Assume that at a point $x^* = x(T) \in \mathfrak{S}$ the $m+1$ gradients $\nabla\psi(x^*), \nabla h_1(x^*), \ldots \nabla h_m(x^*)$ are linearly independent. Then, the tangent space to $\mathfrak{S}$ at this point consists of all vectors orthogonal to the span of $\mathfrak{S}$ and is given by

$$T_S := \{v \in \mathbb{R}^n : \nabla h_i(x^*) \cdot v = 0\}. \tag{7.32}$$ 8.1

The set T_S gives a local set of directions with the idea that if we stay locally in this set we do not violate the constraints. This has to be considered in relation with the variation of the functional value. Hence, the set $\mathfrak{S}^+ := \{x \in \mathfrak{S} : \psi(x) \geq \psi(x^*)\}$ consists of all feasible points x where we can improve the functional value but still stay in the feasible region. We look for another description of this set. The tangent cone to $\mathfrak{S}^+$ (or set of profitable directions) is given by (see Figure 14).

$$T_{S^+} := \{v \in \mathbb{R}^n : \nabla\psi(x^*) \cdot v \geq 0,\ \nabla h_i(x^*) \cdot v = 0\}. \tag{7.33}$$ 8.2

Indeed, $\nabla\psi(x^*)$ points in the direction of increasing values of ψ. The set T_{S^+} is a cone.

The set of feasible directions is charaterized by a hyperplane with normals h_i, i.e, $v \cdot \nabla h_i(x^*) = 0$. In order to have v pointing in the same direction as $\nabla\psi$ we need to have $\nabla\psi \cdot v \geq 0$. We look for a characterization of vectors $p \in \mathbb{R}^n$ satisfying

$$p \cdot v \geq 0,\ \forall v \in T_{S^+}$$

Note that if there is only one vector $\tilde{v} \in T_{S^+}$, then $p = \tilde{v}$ satisfies this condition and we can increase the functional value of ψ by going along $\tilde{v}$. In general, the following result is true.

Lemma 7.7. *Assume that $\nabla\psi(x^*), \nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ are linearly independent (Slater condition).*

A vector $p \in \mathbb{R}^n$ satisfies

$$p \cdot v \geq 0 \text{ for all } v \in T_{S^+} \tag{7.34}$$ 8.3

if and only if it can be written as a linear combination

$$p = +\lambda_0 \nabla\psi(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*). \tag{7.35}$$ 8.4

with $\lambda_0 \geq 0$.

Concluding from (7.35) to (7.34) is obtained by multpyling equation (7.35) by $v \in T_{S^+}$ and obtain $pv = \lambda_0 \nabla\psi(x^*)v \geq 0$, if $\lambda_0 > 0$. For the conserve direction we refer to [**?**]. However, there is the following interpretation: Consider a single equality constraint. Given $\nabla\psi(x^*)$ we can increase the value of ψ by moving along vectors such that $\nabla\psi(x^*)v \geq 0$. On the other hand, every vector has to be othorgonal to $\nabla h(x^*)$ to guarantee that the new point is feasible. Hence, $p \neq 0$ can be written as linear combination

with positive λ_0 of $\nabla\psi$ and ∇h. For a formal proof define $w_1 = \nabla\psi(x^*)$ and $w_2, \ldots, w_{m+1}$ as $w_i = \nabla h_{i-1}(x^*)$. Due to the assertion these vectors are linearly independent and can be extended to a base oof $\mathbb{R}^n$. Denote by v_i the dual base, i.e., $v_i w_j = \delta_{ij}$. Then, $p = \sum \lambda_i w_i$ and $v \in T_{S^+}$ is given by $v = \sum c_i v_i$. Since $w_1 v \geq 0$ for all $v \in T_{S^+}$ we have $c_1 \geq 0$. Since $w_i v = 0$ for $i = 2, \ldots, m+1$ we have $c_i = 0$. Hence,

$$pv = \lambda_1 c_1 + \sum_{i>m+1} \lambda_i c_i$$

Since $pv \geq 0$ for all $v \in T_{S^+}$, this is only true if $\lambda_1 \geq 0$ and $\lambda_i = 0$.

Next, we apply this result to the control problem. To this end we let $x^* = x(t; u^*)$ denote the optimal trajectory. As before, we consider the needle variation for given $\tau \in (0, T)$ and $\omega \in U$ as

$$u_\epsilon(t) := \begin{cases} \omega & t \in [\tau - \epsilon, \tau] \\ u^*(t) & \text{else} \end{cases} \tag{7.36}$$

8.5

As before, we define the first–order variation of the terminal point as

$$v^{\tau,\omega} := \lim_{\epsilon \to 0} \frac{x(T; u^\epsilon) - x(T; u^*)}{\epsilon}$$

We denote by Γ the smallest convex cone containing all vectors $v^{\tau,\omega}$. This is the cone of feasible directions, i.e., directions in which we can move the terminal point $x(T; u^*)$. This cone is due to the fact, that the constraints do not act on the control u (corresponding to the previous vector x) but on the state x coupled to u by the sytem dynamics. In the cone Γ the sensitivity of the system's dynamics with respect to the control is measured, see Figure 14.

Figure 14: Feasible set $\mathfrak{S}$, tangent set and cone for a single equality constraint ∇h_i. The cone Γ and it's relation to the perturbation of the control u^*.

Figure 8.2

Figure 8.1

This yields the geometric (and due to the previous Lemma also the analytic) version of Mayer's probem with terminal constraints. The cone of feasible directions is Γ. Γ depends on u^*, x^* and T. The cone of profitable directions is T_{S^+}. This cone depends on $x^*(T)$ and $\nabla\psi(x^*(T))$ and $\nabla h_i(x^*(T))$. If both cones are separated, then the trajectory x^* corresponding to the control u^* is optimal. If two cones are separate, then there exists

a separating hyperplane denote by a vector $p = p(T)$. This is the theory of finitely many constraints, since at time T we have to discuss the optimality conditions only at $x^*(T)$.

Theorem 7.8. *Let $x^*(t) := x^*(t; u^*(t))$ be the optimal trajectory for the problem (7.29)-(7.31).*

Then the cones Γ and T_{S^+} are weakly separated, i.e., there exists a non-zero vector $p(T)$ such that

$$p(T) \cdot v \geq 0 \; \forall v \in T_{S^+} \tag{7.37}$$ 8.9

and

$$p(T) \cdot v \leq 0, \forall v \in \Gamma \tag{7.38}$$ 8.10

$p(T)$ is the normal of a hyperplane separating Γ and T_{S^+}. This is necessary for optimality since we cannot improve ψ **and** stay feasible. All feasible directions satisfy $p(T) \cdot v \leq 0$, those which improve ψ satisfy $p(T) \cdot v \geq 0$. The difference to Fritz John's theorem is due to the fact that we have to treat *two* feasibility constraints: we need to stay feasible wrt ot the terminal constraints h_i described in the set T_S and we need to stay feasible wrt to variations of the trajectory described by Γ. If we are optimal both sets have to be separate, i.e., we cannot move in any of these sets without leaving the other one, see Figure 15.

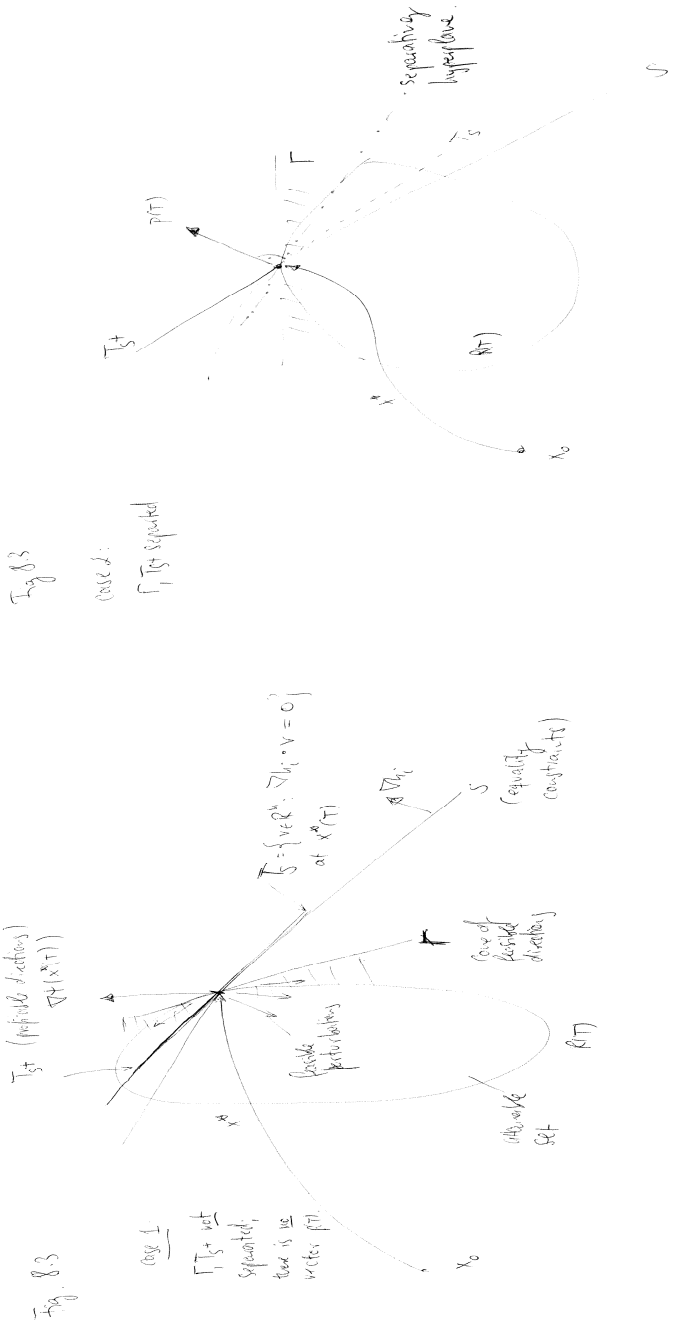


Figure 15: The cone Γ and it's relation to the perturbation of the control u^*.

Figure 8.3

Using the previous Lemma we can restate this theorem as follows.

Theorem 7.9. *Let $x^*(t) := x^*(t; u^*(t))$ be the optimal trajectory for the problem (7.29)-(7.31). Then, there exists a non–zero vector function $t \to p(t)$ such that*

$$p(T) = +\lambda_0 \nabla \psi(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*(T)),\ \lambda_0 \geq 0 \tag{7.39}$$

$$p' = -p(t) D_x f(t, x^*(t), u^*(t)),\ t \in (0, T) \tag{7.40}$$

$$p(\tau) f(\tau, x^*(\tau), u^*(\tau)) = \max_{\omega \in U} \{p(\tau) f(\tau, x^*(\tau), w)\} \tag{7.41}$$

Proof. By the previous Lemma the equations (7.37) and (7.39) are equivalent. Every tangent vector $v^{\tau,\omega} \in \Gamma$ satisfies the linearized evolution equation and if p satisfies the adjoint equation, the product $p(t)v^{\tau,\omega}(t)$ is constant. Equation (7.38), then states $0 \geq p(T)v^{\tau,\omega}(T) = p(t)v^{\tau,\omega}(t)$ and using the definition of $v^{\tau,\omega}$ we obtain (7.41).

Some further remarks are in order. If the function f has sufficient regularity, and **if the maximization problem has a unique solution inside** U, then the last condition can be rewritten as

$$\partial_u \left(p(\tau) f(x^*(\tau), u^*(\tau)) \right) = 0.$$

Introducing now the Hamiltonian operator $\mathcal{H}$ as

$$\mathcal{H}(t, x, u, p) = p f(t, x, u)$$

we obtain another representation of the optimality system. This is typically used in applications. We have

$$\begin{aligned} x' = f(x, u) &\equiv \nabla_p \mathcal{H} \\ p' = -p D_x f(x, u) &\equiv -\nabla_x \mathcal{H} \\ p f(x, u) = \max_{\omega \in U} \{p f(x, w)\} &\equiv \nabla_u \mathcal{H} \end{aligned}$$

and

$$\begin{aligned} x(t_0) &= x_0 \\ p(T) &= +\lambda_0 \nabla \psi(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*(T)),\ \lambda_0 \geq 0 \end{aligned}$$

If we additionally have running costs we obtain by the same trick as before the following result:

Theorem 7.10. *Consider the problem:*

$$\max J(u) = \int_{t_0}^{T} h(t, x(t), u(t))\, dt$$

subject to:

$$x'(t) = f(t, x(t), u(t)) \quad t_0, x(t_0) \text{ - fixed} \tag{7.42}$$

for $u \in C[t_0, T]^m$ with fixed end-point $t_0 < T$, where h and f are continuous in $(t, x, u)$and have continuous first partial derivatives with respect to x and $u\forall(t, x, u) \in [t_0, T] \times \mathbb{R}^n \times \mathbb{R}^m$. Suppose $u^* \in C[t_0, T]^m$ is a (local) minimizer for the problem, and let $x^* \in C^1[t_0, T]^n$ denote the corresponding state. Then there exists a vector function $p^* \in C^1[t_0, T]^n$ such that (x^*, u^*, p^*) satisfies the system:

$$\mathcal{H} = -h(t, x, u) + pf(t, x, u) \tag{7.43}$$

$$x' = f(t, x, u) = \mathcal{H}_\lambda; \quad x(t_0) = x_0; \tag{7.44}$$

$$p' = h_x(t, x(t), u(t)) - p(t)f_x(t, x(t), u(t)) = -\mathcal{H}_x; \quad \lambda(T) = 0 \tag{7.45}$$

eqn:adjoint

$$0 = -h_u(t, x(t), u(t)) + p(t)f_u(t, x(t), u(t)) = \mathcal{H}_u \tag{7.46}$$

eqn:min

for $t_0 \le t \le T$.

From the previous more general theorem we had the condition that u has to maximize $pf(t, x, w)+h(t, x, w)$ Hence, we additionally conclude that the condition

$$\mathcal{H}_{uu} \le 0$$

is also necessary for optimality.

Example 7.11. *Consider the optimal control problem*

$$\min J(u) := \int_0^1 [\frac{1}{2}u(t)^2 - x(t)]\, dt$$

subject to: $x'(t) = 2[1 - u(t)]$; $x(0) = 1$.

Solution: Form the Hamiltonian function

$$\mathcal{H}(x, u, \lambda) = \frac{1}{2}u^2 - x + 2\lambda(1 - u).$$

Candidate solutions (u^*, x^*, λ^*) are those satisfying the Euler-Lagrange equation:

$$x' = \mathcal{H}_\lambda = 2(1 - u^*); \quad x^*(0) = 1 \tag{7.47}$$

$$\lambda' = -\mathcal{H}_x = 1; \quad \lambda^*(1) = 0 \tag{7.48}$$

$$\mathcal{H}_u = 0 = u^*(t) - 2\lambda^*(t). \tag{7.49}$$

The adjoint equation yields: $\lambda^* = t - 1$, *and from optimality condition, we get:* $u^*(t) = 2(t-1)$.

Note: u^* *is a candidate minimum for the problem since* $H_{uu} = 1 > 0$ *for each* $0 \leq t \leq 1$. *Thus*

$$x^{*\prime}(t) = 6 - 4t; \quad x^*(0) = 1 \tag{7.50}$$

$$\Rightarrow x^*(t) = -2t^2 + 6t + 1 \tag{7.51}$$

$$u^*(t) = 2(t-1); \tag{7.52}$$

$$\lambda^*(t) = t - 1. \tag{7.53}$$

One can verify that $\mathcal{H}$ *is constant along the optimal trajectory,* $\mathcal{H}(t, x^*(t), u^*(t), \lambda^*(t)) = -5$.

Finally, to illustrate the optimality of u^* *consider modified controls* $v(t, \eta) := u^*(t) + \eta w(t)$, *and their associated responses* $y(t; \eta)$. *The perturbed cost function reads:*

$$J(v(t;\eta)) := \int_0^1 [\frac{1}{2}u^*(t) + \eta w(t)]^2 - y(t;\eta)]\, dt$$

subject to $y'(t;\eta) = 2[1 - u^*(t) - \eta w(t)]; \quad y(0) = 1$.

Note in this case minimum is always attained at $\eta = 0$.

In the above context we can also give a sufficient condition for optimality.

Theorem 7.12. *Consider the problem:*

$$\max J(u) = \int_{t_0}^{T} h(t, x(t), u(t))\, dt$$

subject to:

$$x'(t) = f(t, x(t), u(t)) \quad t_0, x(t_0) \text{ - fixed} \tag{7.54}$$

for $u \in C[t_0, T]^m$ *with fixed end-point* $t_0 < T$, *where* g *and* f *are continuous in* (t, x, u)*and have continuous first partial derivatives with respect to* x *and* $u \forall (t, x, u) \in [t_0, T] \times \mathbb{R}^n \times \mathbb{R}^m$. *Suppose that* f *and* h *are strictly jointly convex in* x *and* u. *Suppose* $u^* \in C[t_0, T]^m$, $x^* \in C^1[t_0, T]^n$ *and* $p^* \in C^1[t_0, T]^n$ *satisfy the following system.*

$$x' = f(t, x, u); \quad x(t_0) = x_0; \tag{7.55}$$

$$p' = h_x(t, x(t), u(t)) - p(t) f_x(t, x(t), u(t)); \quad \lambda(T) = 0 \tag{7.56}$$

$$0 = -h_u(t, x(t), u(t)) + p(t) f_u(t, x(t), u(t)) \tag{7.57}$$

Suppose further that

$$p \geq 0 \; \forall t. \tag{7.58}$$

Then, u^* *is a strict global minimizer for the problem.*

7.4 Extensions to Pontryagin's maximum principle – Lagrange Minimization Problem and Problems of the Calculus of Variations

We consider a control problem with fixed terminal point and running costs.

$$\min_x \int_0^T L(t, x(t), u(t))dt \text{ or } \max_x \int_0^T -Ldt \tag{7.59}$$ 8.14

subject to

$$x' = f(t, x, u),\ u(t) \in U \tag{7.60}$$ 8.15

and the initial and terminal constraints

$$x(0) = x_0,\ x(T) = x_T. \tag{7.61}$$ 8.16

This problem can be restated in terms of the previous theorem by introducing x_{n+1} satisfying

$$x'_{n+1} = L(t, x, u),\ x_{n+1}(0) = 0$$

and without a terminal constraint on x_{n+1}. Hence, the adjoint vector $p = (p_1, \ldots, p_n, p_{n+1})$ is governed by the adjoint equation

$$p' = -p \begin{pmatrix} D_x f & 0 \\ D_x L & 0 \end{pmatrix}.$$

Therefore, the adjoint function p_{n+1} to x_{n+1} is constant in time. The constraints are given by

$$h_i = x_i(T) - x_{i,T}$$

for $i = 1, \ldots, n$ and

$$\psi(x(T)) = x_{n+1}(T).$$

A direct application of the previous theorem yields

$$p_{n+1}(T) = \lambda_0,\ p_i(T) = +\lambda_i\ 1$$

and the following dynamics for $i = 1, \ldots, n$

$$p'_i(t) = -\sum_j p_j \partial_{x_i} f_j - p_{n+1} \partial_{x_i} L$$

. Since p has to be non–zero, the only requirement is either $\lambda_i = 0$ and $p_{n+1} = \lambda_0$ is strictly positive or some λ_i is non-zero. Summarizing, we proved

Theorem 7.13. *Let $t \to x^*(t;u^*)$ be an optimal trajectory corresponding to the optimal control u^*. Then, there exists a constant $\lambda_0 \geq 0$ and a row vector $p^T \in \mathbb{R}^n$ (not both equal to zero) such that*

$$p' = -pD_x f(t,x^*,u^*) - \lambda_0 D_x L(t,x^*,u^*) \tag{7.62}$$

$$p(\tau)f(\tau,x^*(\tau),u^*(\tau)) + \lambda_0 D_x L(t,x^*,u^*) = \max_{\omega\in U}\{p(\tau)f(\tau,x^*(\tau),w) + \lambda_0 D_x L(t,x^*,\omega)\} \tag{7.63}$$

There is no terminal condition on $p(T)$.

The previous theorem can also be used to derive the Euler–Lagrange equations for ODE control theory. To this end consider the problem

$$\min_{x,x'} \int_0^T L(t,x(t),x'(t))dt \tag{7.64}$$ 8.19

over all absolutely continuous functions $x : [0,T] \to \mathbb{R}^n$ subject to

$$x(0) = x_0,\ x(T) = x_b \tag{7.65}$$ 8.20

Since x is absolutely continuous it has weak derivatives a.e. in $W^{1,1}$.

We rewrite this as the problem (7.59) subject to

$$x'(t) = u(t),\ u(t) \in \mathbb{R}^n \tag{7.66}$$ 8.21

We assume that L is smooth and that x^* is an optimal solution. By the previous theorem there exists a constant $(\lambda_0) =: \lambda \geq 0$ and a row vector $p(t)$ (not both equal to zero) such that

$$p' = -\lambda \partial_x L(t,x^*,x'^*) \tag{7.67}$$ 8.22

$$p(x'^*) + \lambda L(t,x^*,x'^*) = \min_{\omega}\{p\omega + \lambda L(t,x^*,\omega)\} \tag{7.68}$$ 8.23

If $\lambda \equiv 0$, then $p(t) \neq 0$ and by (7.68) x'^* has to be the minimum over $\mathbb{R}^n$ which is a contradiction. Hence, we have $\lambda > 0$. We normalize p, λ such that $\lambda = 1$ (possible due to (7.67).

From (7.68) we obtain a necessary condition for ω to be a minimizer is $\partial_\omega \left(p\omega + L(t,x,\omega)\right) = 0$ and hence

$$p(t) = -\partial_{x'} L(t,x^*(t),x'^*(t)).$$

Formally, this is also obtained from (7.67) by

$$\frac{dp}{dt} = -\frac{\partial}{\partial x} L(t, x^*(t), x'^*(t)) \implies$$
$$p(t) = -\frac{\partial}{\partial x} \int^t L(\cdot) d\tau = -\frac{\partial}{\frac{\partial x}{dt}} \frac{d}{dt} \int^t L(\cdot) d\tau = -\partial_{x'} L(t, x^*(t), x'^*(t))$$

Combining (7.67) with the equation for p we obtain the Euler–Lagrange equations.

$$\frac{d}{dt}\left(\frac{\partial}{\partial x'} L(t, x^*(t), x'^*)\right) = \frac{\partial}{\partial x} L(t, x^*(t), x'^*(t)) \qquad (7.69) \quad \boxed{8.24}$$

Euler–Lagrange's equation can also be derived directly from (7.64). Consider a perturbation $x_\epsilon = x + \epsilon\eta$ with $\eta(0) = \eta(1) = 0$. Necessary for optimality is

$$0 = \frac{d}{d\epsilon} \int L(t, x_\epsilon, x'_\epsilon) dt = \int \partial_x L\eta + \partial_{x'} L\eta' dt = \int (\partial_x L - \frac{d}{dt}\partial_{x'} L)\eta dt \ \forall \eta$$

7.5 Application: Linear Time–Varying Systems and Linear Quadratic Regulators

Consider a state of linear time-varying (LTV) system,

$$x'(t) = A(t)x(t) + B(t)u(t),$$

with $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$, from an initial state $x(t_0) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$, from an initial state $x(t_0) \neq 0$ to a terminal state $x(T) \approx 0$, T given, using "acceptable" levels of control $u(t)$, and not exceeding acceptable levels of the state $x(t)$ on the path. To simplify the notation we use the transposed vector $p \in \mathbb{R}^n$.

The performance index is defined by:

$$J(u) = \int_{t_0}^{T} \frac{1}{2}[u(t)^T Q(t)u(t) + x(t)^T R(t)x(t)]\ dt + \frac{1}{2}x(T)^T S_f x(T)$$

where $S_f \geq 0$, $R(t) \geq 0$, $Q(t) > 0$ are symmetric.

Euler-Lagrange Equations:

$$\begin{aligned} x' &= \mathcal{H}_p(t, x(t), u(t), p(t)); \quad x(t_0) = x_0; \\ p' &= -\mathcal{H}_x(t, x(t), u(t), p(t)); \quad p(T) = S_f x(T); \\ 0 &= \mathcal{H}_u(t, x(t), u(t), p(t)) \end{aligned}$$

where

$$\mathcal{H}_x(t, x(t), u(t), p(t)) = \frac{1}{2}u(t)^T Q(t)u(t) + \frac{1}{2}x(t)^T R(t)x(t) + p^T[A(t)x(t) + B(t)u(t)]$$

Hence

$$u^*(t) = -Q(t)^{-1}B(t)^T p^*(t) \tag{7.70}$$ eqn:ustar

which in turn gives:

$$\begin{bmatrix} x^*(t) \\ p^*(t) \end{bmatrix} = \begin{bmatrix} x^*(t) \\ p^*(t) \end{bmatrix} \begin{bmatrix} A & -BQ^{-1}BT \\ -R & -A^T p^*(t) \end{bmatrix}; \tag{7.71}$$ eqn:adj_system

$$x^*(t_0) = x_0 \tag{7.72}$$

$$p^*(T) = S_f x^*(T). \tag{7.73}$$

Sweep Method:

- Determine $p(t_0)$ such that can be integrated in time as an initial value problem: i.e. sweep $p^*(T) = S_f x^*(T)$ backward to the initial time:

$$p^*(t_0) = S(t_0)x^*(t_0).$$

- For intermediate times, substitute $p^*(t) = S(t)x^*(t)$ into (7.71) to obtain a matrix Riccati equation:

$$S' = -SA - A^T S + SBQ^{-1}B^T S - R; \tag{7.74}$$

$$S(T) = S_f \tag{7.75}$$

  Note: $S(t)$ is a symmetric matrix at each $t_0 \leq t \leq T$ since S_f is symmetric.

- Integrate from T to t_0 to obtain:

$$p^*(t_0) = S(t_0)x^*(t_0)$$

- Once $p^*(t_0)$ is known, $x^*(t)$ and $p^*(t)$ are found by forward integration of (7.71) from $x^*(t_0)$ and $p^*(t_0)$, respectively, to obtain $u^*(t)$, $t_0 \leq t \leq T$ from (7.70).

- Alternatively, use the entire trajectory $S(t)$, $t_0 \leq t \leq T$ to determine the continuous feedback law for optimal control:

$$u^*(t) = -[Q(t)^{-1}B(t)^T S(t)]x^*(t)$$

Remark 7.14. *The approach can be extended to Linear Quadratic Regulator (LQR) problems with mixed state/control terms in the integral cost:*

$$J(u) := \int_{t_0}^{T} \frac{1}{2} \begin{bmatrix} u(t) \\ x(t) \end{bmatrix}^T \begin{bmatrix} Q(t) & P(t) \\ P(t)^T & R \end{bmatrix} \begin{bmatrix} u(t) \\ x(t) \end{bmatrix} \, dt + \frac{1}{2} x(T)^T S_f x(T)$$

Thus the matrix Riccati Equation takes the form:

$$S' = -S(A - BQ^{-1}P^T) - (A - BQ^{-1}P^T)^T S + SBQ^{-1}B^T S + PQ^{-1}P^T - R,$$

The state/adjoint equations take the form

$$\begin{bmatrix} x^*(t) \\ p^*(t) \end{bmatrix} = \begin{bmatrix} A - BQ^{-1}P^T & -BQ^{-1}B^T \\ -R + PQ^{-1}P^T & -(A - BQ^{-1}P^T)^T \end{bmatrix} \begin{bmatrix} x^*(t) \\ p^*(t) \end{bmatrix}$$

and the control is given by:

$$\begin{aligned} u^*(t) &= -Q(t)^{-1}[P(t)^T x^*(t) + B(t)^T p^*(t)] \\ &= -Q(t)^{-1}[P(t)^T + B(t)^T S(t)] x^*(t). \end{aligned}$$

7.6 Dynamic Programming For Pontryagin's Maximum Principle

We consider again a control system of the form as in (7.1)

$$x' = f(x, u), u(t) \in U \tag{7.76}$$ 9.1

We now assume that the set U of admissible controls is compact while f is continuous, uniformly bounded and Lipschitz in x. Further, we are given initial data $x(s) = y \in \mathbb{R}^n$. Under these assumptions there exists for every control $u(\cdot) \in U$ and initial data a unique solution. We seek an admissible control function $u^* : [s, T] \to U$ which minimizes the general problem with running costs, i.e.,

$$J(s, y, u) = \int_s^T h(x(t), u(t))dt + g(x(T)) \tag{7.77}$$ 9.4

for bounded, Lipschitz functions (in x) g and h.

Dynamic programming is a technique to solve the minimization problem $\min_u J$. The problem is a special case of Mayer's problem and necessary conditions have been given in the previous section. The idea of dynamic

programming is to study the mininimization problem by looking at the **value function**.

$$V(s,y) = \min_{u(\cdot)\in U} J(s,y,u) \tag{7.78}$$ 9.8

We are looking at a family of minimization problems for different initial values $x(s) = y$. We study how the costs vary as a function of the initial data. The orginal problem (7.1), resp. (7.76) is then a subproblem solved by evaluating $V(t_0, x_0)$.

The idea of dynamic programming is to start studying $V(T, y)$ which is supposed to be simple to study, since it is stated at terminal time. From the behavior of $V(T, y)$ one proceeds backwards in time and studies $V(T-\tau, y)$. A control acting on $(T-\tau, T)$ connects certain states $y(T-\tau)$ to states $\bar{y}(T)$. For the latter from the previous study it is clear how the optimal control u at time T has to be. This way we can go backwards and determine optimal controls for small time intervals, see Figure 16. A main point will be that the solution to the ode (7.76) is a semigroup

Hence, we are interested in how the minimum cost varies as a function of the initial data. Preliminary we note that the following Lemma holds true.

Lemma 9.1 **Lemma 7.15.** *Let the functions f, g, h be bounded and Lipschitz in x. Then, the value function V defined by (7.78),(7.77) is bounded and Lipschitz continuous in y and s.*

Proof is immediate since J is Lipschitz in s and y and min is continuous.

9.2 **Theorem 7.16** (Dynamic Programming Principle)**.** *For every $\tau \in [s, T]$ and $y \in \mathbb{R}^n$ one has*

$$V(s,y) = \inf_{u(\cdot)\in U} \left\{ \int_s^\tau h(x(t;s,y,u), u(t))dt + V(\tau, x(\tau;s,y,u)) \right\} \tag{7.79}$$ 9.11

In other words (16) the optimization problem on the time interval $[s, T]$ can be split into two separate problems:

1. As a first step, we sole the optimization problem on the subinterval $[\tau, T]$ with running cost h and terminal costs g. In this way we obtain the value function $V(\tau, \cdot)$ at time τ.

2. As a second step we solve the optimization problem on the sub-interval $[s, \tau]$ with running cost h and terminal costs $V(\tau, \cdot)$ determined by the first step.

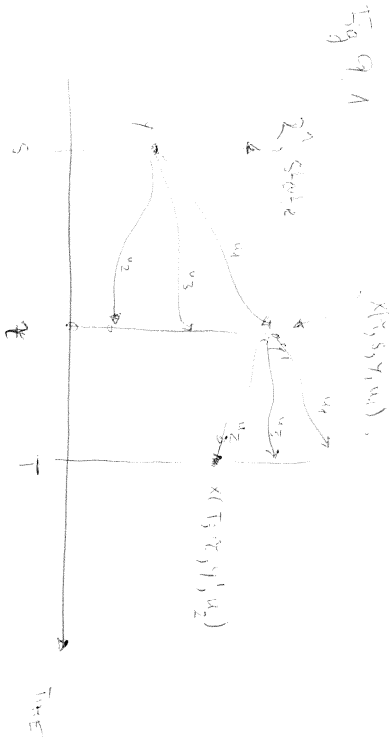


Figure 16: Dynamic Programming Principle.

Figure 9.1

At the initial time s given by (7.79) the value $V(s,y)$ is the solution to minimiziation problem min(7.77) subject to (7.76). The importance of the previous theorem is the practical application: Giving a time–discretization of $[s,T]$ in suitable time cells τ_i one can apply the second step on each interval except for the last. Using e.g. piecewise constant controls on each time interval one obtains a numerical representation of the value functional. We refer to [4], page 37.

7.7 Hamilton Jacobi Bellmann Equation For Pontryagin's Maximum Principle

The purpose of the HJB equation is to provide a characterization of an optimal control u. So far, we only have necessary conditions for optimality involving state, adjoint and control function as well as a scheme by dynamic programming on how to compute an optimal control by successively computing the value functional.

Here, in contrast the focus is on sufficient conditions for a control to be optimal. Furthermore, the value function (7.78) will be characterized as solution to a partial differential equation, namely the HJB equation. The idea is to consider $V(s,y)$ as function of two parameters and try to find a characterizing equation. The main theorem is as follows.

10.1 **Theorem 7.17** (HJB-Equation)**.** *Assume that U is compact, let f,g,h be bounded and Lipschitz with respect to x. Consider the cntrol problem (7.76), the value function V defined by (7.78) and (7.77).*

Then, V is the unique, viscosity solution of the HJB-equation

$$-\big(V_t + H(x,\nabla V)\big) = 0 \tag{7.80}$$ 10.1

with terminal condition

$$V(T,x) = g(x)\ x \in \mathbb{R}^n \tag{7.81}$$ 10.2

and Hamiltonian

$$H(x,p) = \min_{\omega\in U}\big\{f(x,\omega)p + h(x,\omega)\big\} \tag{7.82}$$ 10.3

We refer to [4] for the definition of viscosity solutions and the proof. In fact, dynamic programming is the method of characteristics applied to the partial differential equation, see below, and the lines of Figure 16 are the characteristics of (7.80). The function H is really a Hamiltonian and we later erive the corresponding equations. We conclude with some important

calculations showing the relation between Pontryagin's maximum principle, dynamic programming and HJB-equation. The precise relation is as follows:

The trajectories which satisfy Pontryagin's maximum principle provide characteristic curves for the HJB equation of dynamic programming.

All the following calculations assume that the involved functions are sufficiently smooth. At first, we start from the HJB equation. Call $p = \nabla V$ such that $p_i = \partial_{x_i} V$. Then,

$$\frac{\partial^2}{\partial_{x_j}\partial_{x_i}} V = \frac{\partial^2}{\partial_{x_i}\partial_{x_j}} V = \partial_{x_i} p_j = \partial_{x_j} p_i.$$

Therefore, differentiating (7.80) with respect to x_i we obtain

$$\partial_t p_i + \partial_{x_i} H + \sum_j \partial_{p_j} H \partial_{x_i} p_j = \partial_t p_i + \partial_{x_i} H + \sum_j \partial_{p_j} H \partial_{x_j} p_i = 0$$

If $t \to x(t)$ is any smooth curve the total derivative of p_i along x is compute as

$$\frac{d}{dt} p_i(t, x(t)) = \partial_t p_i + \sum_j \partial_{x_j} p_i x_j'.$$

Combining with the previous equation we obtain

$$\frac{d}{dt} p_i(t, x(t)) = -\partial_{x_i} H + \sum_j \partial_{x_j} p_i \left(x_j' - \partial_{p_j} H \right).$$

Hence, if $x_j' = \partial_{p_j} H$ (characteristic equation for a Hamiltonian) the last term vanishes and the derivative of the generalized impuls is given by the spatial derivative of the Hamiltonian. The previous calculations show that we can obtain a solution to (7.80,7.81) by solving the characteristic system of the PDE given by the set of ODE's parameterized by $\bar{x}$

10.20

$$x_i' = \partial_{p_i} H(x,p),\ p_i' = -\partial_{x_i} H(x,p), \tag{7.83a}$$

$$x_i(T) = \bar{x}_i,\ p_i(T) = \partial_{x_i} V(T,x) = \partial_{x_i} g(\bar{x}) \tag{7.83b}$$

Obviously, the previous system is obtained for solutions $V(t,x)$.

Conversly, we recover V by setting

$$V(t, x(t,\bar{x})) = g(\bar{x}) + \int_t^T \left(H(x(s), p(s) - p(s)\partial_p H(x(s), p(s)) \right) ds,\ \nabla_x V(t, x(t,\bar{x})) = p(t;\bar{x})$$

where $x(s), p(s)$ are the solutions to the system of ODEs above. Note, that

$$d_t V(t, x(t,\bar{x})) = -H(x(t), p(t)) - p(t)\partial_p H(x(t), p(t))$$

and by definition

$$d_t V(t, x(t, \bar{x})) = V_t(t, x(t, \bar{x})) + x'(t) \nabla_x V(t, x(t, \bar{x}))$$

Therefore, we obtain for V defined as above and $x(t), p(t)$ solutions to the ODEs (7.83) we obtain the PDE again. Hence, (7.83) is equivalent to (7.80,7.81).

It remains to show that Pontryagin's maximum principle also yields (7.83). This will be discussed in the simpler situation of $h \equiv 0$ and $U = \mathbb{R}^n$. Pontryagin's maximum principle (7.5) applied here as minimum principle then reads

$$x' = f(x, u), x(s) = x_0,\ \bar{p}' = -\bar{p} D_x f(x, u), \bar{p}(T) = -\nabla g(x(T)).$$

and

$$\max\{\bar{p} f(x, \omega)\} = \bar{p} f(x, u)$$

or written in terms of $p = -\bar{p}$

$$x' = f(x, u), x(s) = x_0,\ p' = -p D_x f(x, u), p(T) = \nabla g(x(T)).$$

and

$$\min_{\omega}\{p f(x, \omega)\} = p f(x, u)$$

.

The HJB equation is equivalent to (7.83) for an abitrary point $\bar{x}$ for trajectories p and x. Hence, it remains to show that the Hamiltonian H given by (7.82) satisfies

$$\nabla_p H(x, p) = f(x),\ \nabla_x H(x, p) = p D_x f \text{ or } H(x, p) = p f(x, u).$$

Due to equation (7.82) we define the optimal trajectory $u = u(x, p)$ by

$$H(x, p) = p f(x, u) := \min f(x, \omega p).$$

Then, at a minimum $p D_u f(x, u) = 0$ and hence

$$D_p H(x, p) = f(x, u) + p D_u f(x, u) \partial_p u = f(x, u)$$

and similarly

$$D_x H(x, p) = p D_x f(x, u).$$

This shows that HJB is equivalent to Pontryagin's maximum principle. We summarize the results in Figure 17.

Figure 17: Relations of HJB, dynamic programming and Pontryagin's maximum principle

Figure 9.4

8 Necessary optimality conditions in infinite dimensional spaces

8.1 Equality constraints

We want to derive the necessary optimality conditions for a general minimization problem. The Lagrange multiplier theorem is based on the inverse function theorem. This theory is a *local theory* in the sense that no convexity assumption on the functional or the describing set is required. In contrast we present below global results under the assumption that certain sets are convex. In the latter theory the Lagrange multiplier is even more meaningfull in the following sense: A general minimization problem $\min_X f$ over some set $H : X \to Z$, $H(x) = 0$, can either be viewed in the *primal space* X by studying the contonourlines of f (local approach) or by studying the constraint space Z (global approach). The Lagrange multiplier is an element of the dual to the constraint space Z^*. It therefore is a linear functional on Z or geometrically a hyperplane. Having a hyperplane to separate certain sets we can obtain global results for convex functions lying above or below this hyperplane.

The more general theory is the local approach which we discuss at first.

Definition 8.1 (Regular point)**.** *Let T be a continuously Fréchet differentiable mapping from an open set D in a Banach space X into a Banach space Y. If $x_0 \in D$ is such that $T'(x_0)$ maps X onto Y, then the point x_0 is called* regular point *of the transformation T.*

An easy example is $T : \mathbb{R}^n \to \mathbb{R}^m$. If at x_0 the Jacobian of T has rank m, then x_0 is a regular point. Consider the operator $T := \Delta : H_0^1(\Omega) \to H_0^1(\Omega)'$ for an open set $\Omega \subset \mathbb{R}^n$ with sufficiently smooth boundary. Since $\forall f \in H_0^1(\Omega)' \exists! u \in H_0^1(\Omega)$ such that $\Delta u = f$, we obtain, that T' is surjective.

Theorem 8.2 (Inverse Function Theorem)**.** *Let x_0 be a regular point of a mapping $T : X \to Y$, where X and Y are Banach spaces. Then there is a neighborhood $N(y_0)$ for the point $y_0 = T(x_0)$ and a constant K such that the equation $T(x) = y$ has a solution for every $y \in N(y_0)$ and the solution satisfies $\|x - x_0\|_X \leq K\|y - y_0\|$.*

A proof can be found in standard textbooks, e.g. [10, 2, 16]

Derivation of necesary conditions for equality constraint optimization problems is the content of the next theorems.

Lemma 8.3. *Let $f : X \to \mathbb{R}$ and $H : X \to Y$, wherein X and Y are Banach spaces. Assume that f, H are continuously Fréchet differentiable in an open*

set containing x_0 and that x_0 is a regular point of H. Furthermore, assume that x_0 is a local minimum of the problem $\min f(x)$ *subject to* $H(x) = 0$.

Then $f'(x_0)h = 0$ for all h satisfying $H'(x_0)h = 0$.

Proof. Consider the transformation $T : X \to \mathbb{R} \times Y$ given by $T(x) = (f(x), H(x))$. Assume there exists an h such that $H'(x_0)h = 0$ and $f'(x_0)h \neq 0$. The Fréchet derivative of T at x_0 is $T'(x_0) = (f'(x_0), H'(x_0)) : X \to \mathbb{R} \times Y$ and x_0 is a regular point of T. Since x_0 is a regular point of H and $f'(x_0) \neq 0$ is a linear mapping from X to $\mathbb{R}$. Applying the inverse function theorem to T, we obtain that there $\exists x$ such that $\|x - x_0\| \leq \epsilon$ and such that $T(x) = (f(x_0) - \delta, 0)$ with $\epsilon, \delta > 0$. This contradicts the assumption on a local minimum in x_0. □

There is a geometric interpretation of this result. The set of vectors h, such that $H'(x_0) = 0$ is the tangent space. We translate the tangent space to the point x_0. The above theorem states, that f is stationary at x_0 with respect to variations in the tangent plane.

Theorem 8.4 (Lagrange Multiplier Theorem)**.** *Let $f : X \to \mathbb{R}$ and $H : X \to Y$, wherein X and Y are Banach spaces. Assume that x_0 is the minimizer of f subject to $H = 0$. Assume that f, H are continuously Fréchet differentiable in an open set containing x_0 and that x_0 is a regular point of H.*

Then there exists $y_0^* \in Y^*$ such that the Lagrange function

$$L(x, y^*) : X \times Y^* \to \mathbb{R}.\ L(x, y^*) = f(x) + < y^*, H(x) >_{Y^*,Y} \tag{8.1}$$

is stationary at (x_0, y_0^*), i.e.,

$$f'(x_0) - < y_0^*, H'(x_0) >= 0 \tag{8.2}$$

$$H(x_0) = 0 \tag{8.3}$$

Note, that $H(x_0) = 0$ is an equality in X. The first equation has to be understand in the sense of linear mappings $L(X; \mathbb{R}) = X^*$, i.e.,

$$\forall h \in X. f'(x_0)h - < y_0^*, H'(x_0)h >= 0\ (\in \mathbb{R}) \tag{8.4}$$

Further, we denote the application of a linear operator A on an element x by Ax. The application of a functional $f \in X^*$ on a element in $x \in X$ is denoted by $< f, x >$. If $A : X \to \mathbb{R}$, then by notation we have $< A, x >= Ax$.

Proof. From the previous lemma, we have $f'(x_0) \in L(X; \mathbb{R}) = X^*$ is orthogonal on the null space of $H'(x_0)$. The range of $H'(x_0)$ is closed. If a linear operator A has a closed range, then the range of the adjoint operator

A^* is the orthogonal complement of the nullspace of A. A proof of this statement can be found in [10, 2]. Therefore,

$$f'(x_0) \in N(H'(x_0))^T = \mathcal{R}[H'(x_0)^*]. \tag{8.5}$$

and there exists $y_0^* \in Y^*$ such that

$$f'(x_0) - H'(x_0)^* y_0^* = 0 \ (\in X^*) \tag{8.6}$$

Now we can rewrite this equation by using the definition of the adjoint $< A^* y^*, x >=< y^*, Ax >$ to obtain

$$f'(x_0) - y_0^* H'(x_0) = 0 \ (\in X^*) \tag{8.7}$$

□

Due to our choice of the notation the following lines are equivalent

$$< f'(x_0), x >_{X^*,X} - < y_0^*, H'(x_0)x >_{Y^*,Y}= \tag{8.8}$$
$$f'(x_0)x - < H'(x_0)^* y^*, x >_{X^*,X}= \tag{8.9}$$
$$f'(x_0)x - (H'(x_0)^* y^*)x \tag{8.10}$$

If X is a Hilbert space we can use the Riesz representation theorem, see [2, 16], to reformulate as follows:

$$< x^*, x >= (\bar{x}, x) \tag{8.11}$$

where $(,)$ denotes the scalar product on X and $\bar{x}$ is the image of x^* under the isometric, conjugate linear isomorphism $J : X^* \to X$.

We offer a further interpretation of the Lagrange Multiplier Theorem. Consider the finite dimensional case. Then the theorem states, that the gradient f' at the optimum is a linear combination of the gradients of the constraints h_i'. In case of two constraints, the gradient $f'(x_0)$ is contained in the plane spanned by $h_1'(x_0)$ and $h_2'(x_0)$.

Remark 8.5. *Formally, the KKT systems in the finite and infinite dimensional case are equal. However, one has to notice that the assumptions put in the infinite dimensional case are much more restrictive. In the case of equality constraints we conclude for the infinite dimensional case. The operator of the equality constraints h' has to be surjective* **and** *has to be a bounded linear operator. The latter condition is automatically satisfied in the finite dimensional case assuming that h is sufficiently regular. This is replaced by continuously Frechet differentiable in the infinite dimensional*

case. Both are not just a change of the notation, but actually a stronger assumption. This can also be seen in the proof of the Lagrange multiplier theorem, where we need to use the "closed range theorem", to deduce the existence of a Lagrange multiplier y_0^* such that

$$f'(x_0) - H'(x_0)^* y_0^* = 0.$$

This is automatically satisfied in the finite dimensional case, since there the adjoint is given by the transposed gradient.

Example 8.6. Linear quadratic problems

Consider the problem

$$\min_{(y,u)\in Y\times U} \frac{1}{2}\|Q(y-y_d)\|_H^2 + \frac{\alpha}{2}\|u\|_U^2$$

subject to

$$Ay + Bu = 0$$

and under the assumptions.

- *$\alpha > 0$ and $A \in L(Y,Z)$ has a bounded inverse, $B \in L(U,Z)$ and there exists a feasible point.*

The problem fits in the previouse theorem for $X = Y \times U$. Then,

$$f(x) = \frac{1}{2}\|Q(y-y_d)\|_H^2 + \frac{\alpha}{2}\|u\|_U^2$$

is continuously Frechet–differentiable. The same is true for the constraint mapping $H(y,u) = Ay + Bu : Y \times U \to Z$ since it is linear in y and u. It remains to check for the regularity condition at a point $x = (y,u) \in X$. The derivative of H at y_0, u_0 is given by

$$DH(y_0,u_0)[y,u] = Ay + Bu \in Z.$$

We have to check whether or not for any $z \in Z$ there exists $(\bar{y},\bar{u}) \in Y \times U$ such that $A\bar{y} + B\bar{u} = z$. In fact, since A has an inverse we set $\bar{u} = 0$ and $\bar{y} = A^{-1}z$ independent of y_0 and u_0. Applying the previous theorem at a loca minimizer $x_0 = (y_0,u_0)$ we obtain the necessary first–order optimality conditions for $z^* \in Z^*$ and all $\tilde{y},\tilde{u} \in Y \times U$.

The functional is $J := \frac{1}{2}\big(Q(y-y_d), Q(y-y_d)\big) + \frac{\alpha}{2}(u,u)$ and $J_y\tilde{y} = (Q(y-y_d), Q\tilde{y}) = (Q^*Q(y-y_d),\tilde{y})$, $J_u()\tilde{u} = \alpha(u,\tilde{u})$.

$$\begin{aligned} Ay + Bu &= 0 \\ (Q^*Q(y-y_d),\tilde{y}) + \alpha(u,\tilde{u}) - &< z^*, A\tilde{y} + B\tilde{u} >= 0 \end{aligned}$$

We can reformulate this by taking all variations we obtain

$$\begin{aligned} Ay + Bu &= 0 \\ Q^*Q(y - y_d) - A^*z^* &= 0 \\ \alpha u - B^*z^* &= 0 \end{aligned}$$

The Lagrange function is

$$L(y, u, \lambda) = \frac{1}{2}\|Q(y - y_d)\|_H^2 + \frac{\alpha}{2}\|u\|_U^2 - < \lambda, Ay + Bu >_{Z^*,Z} .$$

Example 8.7. Distributed control of elliptic equations *Consider the problem*

$$\min_{(y,u)\in Y\times U} \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2$$

subject to

$$-\Delta y = u,\ y = 0 \text{ in } \Omega, \quad \text{on } \partial\Omega$$

and under the assumptions. Using classical theory on weak solutions to elliptic equations we observe that the followig operators are well–defined under the asumption that Ω is sufficiently regular.

- $Y = H_0^1(\Omega), U = L^2(\Omega)$
- $< Ay, v >_{Y^*,Y} = \int_\Omega \nabla y \cdot \nabla v dx \in L(Y, Y^*)$
- $< Bu, v >_{Y^*,Y} = -\int_\Omega uv dx \in L(U, Y*)$
- $Y = Y^*$ *since* Y *is a Hilbert space.*
- *For any* $u \in L^2(\Omega)$ *there exists a unique solution* $y \in Y$ *to* $< Ay + Bu, v >= 0$ *for all* $v \in V$. *Furthermore, the norm of the solution can be bounded by* u.
- *Derivative of* c *wrt to* y *has bounded inverse*

Again, we have to check, if the constraint mapping is surjective. The constraint mapping is

$$H(x) = Ay + Bu : Y \times U \to Y^*$$

and given by

$$< H(y, u), v >= \int_\Omega \nabla y \nabla v - uv dx$$

Since the equation $-\Delta y = y^*$ admits a unique solution for any $y^* \in Y^*$ and since H is a linear and bounded operator we obtain as before that any point $\bar{y}, \bar{u}$ is a regular point of the mapping $DH \in L(Y \times U; Y^*)$.

The optimality condition at y_*, u_* reads for a multiplier $\lambda_* \in Y \equiv Y^*$ and any $v \in Y$.

$$Ay_* + Bu_* = 0 \tag{8.12}$$

$$(y_* - y_d, v)_{L^2} + \int_\Omega \nabla v \nabla \lambda_* dx = 0 \tag{8.13}$$

$$\int_\Omega (\alpha u_* - \lambda_*) v dx = 0 \tag{8.14}$$

The strong form is therefore

$$-\Delta y_* = u \text{ in } \Omega,\ y_* = 0 \text{ on } \partial\Omega \tag{8.15}$$

$$-\Delta \lambda_* = (y_* - y_d) \text{ in } \Omega,\ \lambda_* = 0 \text{ on } \partial\Omega \tag{8.16}$$

$$\alpha u_* + \lambda_* = 0 \text{ a. e. in } \Omega. \tag{8.17}$$

The Lagrange function is

$$L(y, u, \lambda) = \frac{1}{2}\|y - y_d\|^2_{L^2} + \frac{\alpha}{2}\|u\|^2_{L^2} - \int_\Omega \nabla y \cdot \nabla v - uv dx.$$

Example 8.8. Application to pde's: Stokes problem

We give an example of the Lagrange multiplier rule omitting technical details. Consider a bounded set $\Omega \subset \mathbb{R}^n$ with smooth boundary and the space $X = H^1_0(\Omega)$. Let

$$f(u) = \int_\Omega \frac{1}{2}|\nabla u|^2 - \int_\Omega \phi u dx \tag{8.18}$$

with the constraint $H(u) := divu$, i.e. $H(u) = 0 \in L^2(\Omega)$. Hence in the setting above $Y = L^2(\Omega)$. We introduce as in the Lagrange multiplier theorem the Lagrange function

$$L(x, y^*) := f(x) + < y^*, H(x) >_{Y^*,Y} = \int_\Omega \frac{1}{2}|\nabla u|^2 - \int_\Omega \phi u dx + \int_\Omega q div u dx$$

since Y^* is isomorphic to Y and with $x \equiv u, y^* \equiv q$. Assume, that all necessary operations (i.e. differentiation, constraint qualification) are fulfilled we obtain for $v \in X$ $f'(u)v = \int_\Omega \left((\nabla u \nabla) v - fv dx\right.$ and $H'(u)v = divv$. Hence,

$$\forall v \in H^1_0(\Omega) : \int \nabla u \nabla v - fv dx + \int_\Omega q div v dx = 0, divu = 0$$

This is obviously the strong formulation of the equation

$$-\Delta u + \nabla q = f, divu = 0, u|_{\partial\Omega} = 0$$

It remains to show, that $H'(u) = div$ *is a linear, bounded operator, i.e.* $\in L(H_0^1(\Omega); L^2(\Omega))$*, and is surjective. For* $v \in H_0^1(\Omega)$ *we claim* $\int_\Omega (divv)^2 dx = \int_\Omega (\sum_i \partial_{x_i} v_i)^2 dx \leq \int_\Omega (\sum_i |\nabla v|)^2 dx \leq n^2 \int_\Omega |\nabla v|^2 + |v|^2 dx \leq c\|v\|_{H_0^1(\Omega)}$*. Further, for* $w \in L^2(\Omega)$ *the functions* $v_i = \frac{1}{n}\int_0^{x_i} w(x_1, \ldots, x_i = y, \ldots, x_n)dy$ *have the property* $divv = w$ *and* $v_i \in L^2(\Omega)$ *since* Ω *is bounded. Indeed,* $\int_\Omega (\int_0^{x_i} w()dy)^2 dx$ *is by Jensens' inequality and Fubini's theorem less than:* $\int_\Omega \int_0^{x_i} w()^2 dydx \leq \int_\Omega \int_0^c w()^2 dydx \leq \int_\Omega \int_\Omega w^2 dzdx \leq \mu(\Omega)|w|_{L^2(\Omega)}$*. Therefore,* H' *is surjective.*

Example 8.9. Lagrange multiplier approach for elliptic problems

We consider the same problem as above, but formulate this in a different way.

$$\min J(y,u) = \int_\Omega (y - y_d)^2 + \alpha u^2 dx$$

subject to

$$-\Delta y = f + u \text{ in } \Omega \text{ and } y = 0 \text{ on } \partial\Omega$$

Let $y_d, f \in L^2(\Omega)$ *be given functions. To be in the setting of the previous section, we define the following operators and spaces.* $J : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ *and* $H(y,u) : H_0^1(\Omega) \times L^2(\Omega) \to H^{-1}(\Omega)$ *defined by* $H(y,u)\phi = \int_\Omega \nabla y \nabla\phi - (f+u)\phi dx$ *for all* $\phi \in H_0^1(\Omega)$*. Hence,* $H(y,u)$ *defines a continuous, affin-linear functional on* $H_0^1(\Omega)$*. In the setting of the Lagrange multiplier theorem we have* $X := H_0^1(\Omega) \times L^2(\Omega)$ *and* $Y := H^{-1}(\Omega)$*. To apply the theorem we need to prove that* J, H *are Frechét differentiable and that there exists a regular point* (y_0, u_0)*. The existence of a minimizer is given by the discussion of the previous paragraph. Clearly,* J *is differentiable. Further,* H *is a affine linear operator in* y *and* u*, i.e.,* $H = \tilde{H} + f$ *where* $\tilde{H}$ *is a linear operator. It is differentiable, if the operator is bounded. Since* $|H(y,u)\phi| \leq c(\|y\| + \|u\|)\|\phi\|$ *this holds true. Further, we have to prove, that* (y_0, u_0) *obtained as solution of the previous paragraph is a regular point, i.e.,* $H'(y_0, u_0) \in L(H_0^1(\Omega) \times L^2(\Omega); H^{-1}(\Omega))$ *is surjective. Consider* $h \in H^{-1}(\Omega)$*. We need to find* $(y,u) \in H_0^1(\Omega) \times L^2(\Omega)$ *s.t.* $H'(y_0,u_0)(y,u) = h \Leftrightarrow H'(y_0,u_0)(y,u)\phi = h(\phi)$ *for all* $\phi \in H_0^1(\Omega)$*. Since* H *is a linear operator* $H'(y_0,u_0)(y,u) = H(y,u)$*. Then applying the definition of* H *and using that* $H_0^1(\Omega)$ *is a Hilbert space*

$$\int_\Omega \nabla y \nabla\phi + (f+u)\phi dx = \int_\Omega \nabla\tilde{h}\nabla\phi + \tilde{h}\phi dx \ \forall \phi \in H_0^1(\Omega)$$

wherein $\tilde{h}$ denotes the element in $H_0^1(\Omega)$ such that $h(\phi) = (\tilde{h}, \phi)_{H_0^1(\Omega)}$ This element exists by Riesz representation theorem. Hence, we need to find (y, u) such that the following equation is satisfied $\int_\Omega \nabla(y - \tilde{h})\nabla\phi + (f + (u - \tilde{h}))\phi dx = 0$. We set $y = y_0 - \tilde{h}$ and $u = u_0 - \tilde{h}$. This proves the surjectivity of $H'(y_0, u_0) = H(\cdot)$. Hence, we can apply the Lagrange multiplier theorem and obtain: There exists $p^* \in (H^{-1}(\Omega))' = H_0^1(\Omega)$:

$$J'(y_0, u_0)(y, u) - < p^*, H'(y_0, u_0)(y, u) >_{H_0^1, H^{-1}} = 0 \ \forall (y, u) \in X$$
$$H(y_0, u_0) = 0 \ \in X^*$$

Written in more detail

$$\int_\Omega (y_0 - y_d)y + 2\alpha u_0 u dx - \int_\Omega \nabla p^* \nabla y + u p^* dx = 0 \ \forall (y, u) \in H_0^1(\Omega) \times L^2(\Omega)$$

which implies

$$\Delta p^* = -(y_0 - y_d) \in H_0^1(\Omega) \tag{8.19}$$
$$2\alpha u_0 - p^* = 0 \in L^2(\Omega) \tag{8.20}$$

and by the second equation

$$-\Delta y_0 = f + u_0 \in H_0^1(\Omega) \tag{8.21}$$

This gives the adjoint equation and the optimality system. However, the proof of existence of an optimal solution was given by the considerations of the previous paragraph.

The previous examples can be summarized in the following theorem.

labell-5 **Theorem 8.10.** *Let U, H be given Hilbert spaces and let U_{ad} be a given non-empty, closed and convex set in U. Let $y_d \in H$ and $\lambda \geq 0$ be given.*

Assume that $S : U \to H$ is a linear and bounded operator. An element $u^* \in U_{ad}$ solves the problem

$$\min_{u \in U_{ad}} f(u) := \frac{1}{2}\|Su - y_d\|_H^2 + \frac{\lambda}{2}\|u\|_U^2$$

if the following variational inequality is satisfied

$$\left(S^*\left(Su^* - y_d\right) + \lambda u^*, u - u^*\right) \geq 0, \forall u \in U_{ad}.$$

Proof. Using the result of the exercise below we have that $2f(u) = G(F(u)) + K(u)$ is Frechet-differentiable with the functions $F(u) = Su : U \to H$, $K(u) = \lambda\|u\|_U^2$ and $G(z) = \|z - y_d\|_H^2$. We have that

$$DK(u)v = \lambda(2u, v)_U,\ DG(z)\eta = (2(z - y_d), \eta)_H$$

and therefore

$$Df(u)h = (Su - y_d, Sh)_H + \lambda(u, h)_U = (S^*(Su - y_d) + \lambda u, h)_U.$$

The variational inequality is obtained as a consequence of Theorem 9.11. This proves the result. □

Exercise 8.11. *1. Show that the following holds true: Let U, V, W be Banach spaces and $F : U \to V$ and $G : V \to W$ be Frechet–differentiable maps. Then $E(u) = G(F(u)) : U \to W$ is Frechet-differentiable with derivative*

$$DE(u) = DG(F(u)) \circ DF(u).$$

2. Prove the variational inequality in the following setting: Let U be a Hilbert space and $C \subset U$ a convex set. Assume that $\bar{u}$ is the minimum of $J : U \to \mathbb{R}$ on C. Then, the variational inequality

$$(DJ(\bar{u}), u - \bar{u}) \geq 0$$

is satisfied for all $u \in C$ by considering the function $\tilde{J} : \mathbb{R} \to \mathbb{R}$ given by $\tilde{J}(t) = J(\bar{u} + t(u - \bar{u}))$.

Remark 8.12. *The operator S in the linear quadratic case of the first example is given by $S = -A^{-1}B$. Therefore, the general assumptions in the previous examples have always been that A has a bounded inverse in order to have S as bounded linear operator.*

8.2 Inequality constraints

Next, we briefly discuss inequality constraints. We consider the problem

$$\min f(x) \text{ subject to } G(x) \leq 0 \tag{8.22}$$

where $f : X \to \mathbb{R}$ and $G : X \to Y$ and X, Y are Banach spaces.

To understand the definition of the problem, we introduce the following deifnitions.

Definition 8.13. *A set P in a vector space is a cone, iff $\forall x \in P : \alpha x \in P$ forall $\alpha \in \mathbb{R}$ and $\alpha \geq 0$.*

A set P in a vector space is convex, iff $\forall x, y \in P, \alpha \in [0,1]$ the point $\alpha x + (1-\alpha)y \in P$.

Let P be a subset of a normed space X. Then x is in the interior of P ($intP$), iff there is $\epsilon > 0$ such that $\forall y : \|y - x\| \leq \epsilon$ we have $y \in P$.

Definition 8.14 (Inequalities)**.** *In a general Banach space Y we call $x \geq y$, iff $x - y \in P$ and P is a (positive) convex cone.*

The negative convex cone N is defined by $N := -P$.

The following example might illustrate the definition.

$$X = \mathbb{R}^n, P = \{x \in \mathbb{R}^n : x_i \geq 0\} \tag{8.23}$$

is a positive convex cone. Further, the set of all non-negative continuous functions is a convex cone in the set $C(\mathbb{R})$. Since a cone is a set we have that the interior and the closure of P are well-defined. Also, $0 \in P$. Hence we write $x > 0$, iff $x \in intP$. If we assume a normed space we conclude that $\|x\| > 0$ if $x \in intP$. For the KKT theorem stated below it is essential that P posses an interior point. Nevertheless, this is not granted for every Banach space.

Example 8.15. *Let $X = L^1([0,1]), \|\cdot\|$ and P is taken as the subset of a nonnegative functions on the interval $[0,1]$, i.e., $P := \{f \in L^1([0,1]) : f(x) \geq 0,\ a.e.\}$. We claim, that forall $f \in P$ and forall $\epsilon > 0$ and $g \in L^1([0,1])$ such that $g \notin P$ and $\|f - g\| \leq \epsilon$. Therefore, f is not an interior point. Wlog assume $\epsilon < \|f\|$. Then there exists $a < b, a, b \in [0,1]$ such that $\int_a^b |f|dx \leq \epsilon/2$. Indeed, assume converse ($\forall a < b : \int_a^b |f|dx > \epsilon/2$) and let $\epsilon/2k = \|f\|$. Then $\|f\| = \sum_{i=0}^{k} \int_{i/(k+1)}^{(i+1)/(k+1)} |f|dx > \sum_{i=0}^{k} \epsilon$ is a contradiction. Hence, define $g = -f$ forall $x \in [a,b]$ and $g = f$ else. We obtain $\|f - g\| \leq \epsilon$ and $g \notin P$.*

In a convex positive cone we have the common relations $x \geq y, y \geq z \implies x \geq z$. For $a, b \in P$ and P convex cone, we have $a + b \in P$. Since $2\frac{1}{2}(a+b) \in P$. Therefore, $x - z = x - y + y - z \in P$.

Example 8.16. *The inequality constraint is well–posed in $X = C^0([0,1])$. This space has interior points and is equipped with the sup–norm.*

Using the result on equality constrained optimization, we can conjecture that the following is true:

$$0 = f'(x_0) - < y_0^*, G'(x_0) >= 0 \tag{8.24}$$

for

$$y^* \geq 0 \tag{8.25}$$

and wherein

$$< y^*, G(x_0) >= 0. \tag{8.26}$$

Since we need to satisfy an inequality we only allow positive y^*. This corresponds to the fact, that we do not have the full plane spanned by $G'(x_0)$, but a cone. Further, we need to impose the above relation only for those inequality constraints which are active at the solution. Therefore, we get the second condition, which implies (in the finite dimensional case) $y_i^* = 0$ if $g(x_0) > 0$. The following theorem can be proven.

Definition 8.17 (Regular point for the inequality constraints)**.** *Let X, Z be Banach spaces. Assume Z has a (positive) convex cone with nonempty interior. Let G be mapping from X to Z which is Gateaux differentiable. A point $x \in X$ is called regular for the inequality constraint, iff $G(x) \leq 0$ and if there exists $h \in X$ such that $G(x) + G'(x)h < 0$.*

KKT-I **Theorem 8.18** (Generalized Karush-Kuhn-Tucker Theorem)**.** *Let X, Z be Banach spaces where Z has a (positive) convex cone $P \neq \emptyset$ with interior points. Let $f : X \to \mathbb{R}$ and $G : X \to Z$ be Fréchet differentiable mappings with continuous derivatives. Suppose x_0 minimizes $f(x)$ subject to $G(x) \leq 0$ and x_0 is a regular point for the inequality constraint. Then there exists a $z_0^* \in Z^*$ with*

$$z_0^* \geq 0 \tag{8.27}$$

$$< z_0^*, G(x_0) >_{Z^*,Z}= 0 \tag{8.28}$$

$$f'(x_0) + < z_0^*, G'(x_0) >_{Z^*,Z}= 0 \tag{8.29}$$

Proof. Consider the space $W = \mathbb{R} \times Z$ and define the two sets

$$A = \{(r, z) : r \geq f'(x_0)h, z \geq G(x_0) + G'(x_0)h, \text{ for some } h \in X\} \tag{8.30}$$

$$B = \{(r, z) : r \leq 0, z \leq 0\} \tag{8.31}$$

We are going to prove the theorem along the following claims

1. A, B are convex. Clear.
2. A, B are cones. $A = \big(f'(x_0)h, G(x_0) + G'(x_0)h\big) + \tilde{A}$ and $\tilde{A}$ is a cone.

3. $B \neq \emptyset$ and contains interior points. Since P has an interior point.

4. Set A does not contain any interior point of set B. Assume the converse $\exists(r, z) \in A$ with $r < 0, z < 0$. Then by definition of A: $f'(x_0)h < 0$ and $G(x_0) + G'(x_0)h < 0$. Now, there exists a open sphere $B(G(x_0) + G'(x_0)h; r)$ such that $\forall G \in B : G \leq 0$ and hence B is contained in the negative cone $N := -P$ in Z. Further, for $0 < \alpha < 1$ the sphere $\tilde{B}(\alpha[G(x_0) + G'(x_0)h], \alpha r)$ is contained in $-P$. Since x_0 is a regular point $(1 - \alpha)G(x_0) \leq 0$ and P is convex, we conclude that $(1 - \alpha)G(x_0) + \alpha[G(x_0) + G'(x_0)h] < 0 \in N$. Due to

$$\|G(x_0 + \alpha h) - G(x_0) - \alpha G'(x_0)h\| = o(\alpha)$$

   Hence, $G(x_0 + \alpha h) < 0$ for α sufficiently small. Analogously, for $f(x_0 + \alpha h) < f(x_0)$. This contradicts the x_0 optimality.

5. According to a variant of the Hahn-Banach theorem there exists a hyperplane separating A and B. Hence there are r_0, z_0^*, δ :

$$r_0 r + < z_0^*, z > \geq \delta \; \forall (r, z) \in A$$
$$r_0 r + < z_0^*, z > \leq \delta \; \forall (r, z) \in B$$

   Since $(0, 0)$ contains in A and B, we have $\delta = 0$. By the properties of $B : r_0, z_0^* \geq 0$. Further, $r_0 \neq 0$ since $G(x_0) + G'(x_0)h < 0$. Hence, we normalize $r_0 = 1$. Rewriting the separation condition for A gives

$$f'(x_0)h + < G(x_0) + G'(x_0)h, z_0^* > \geq 0 \; \forall h \in X$$

   For $h = 0$ we obtain $< z_0^*, G(x_0) > \geq 0, z_0^* \geq 0, G(x_0) \leq 0 \implies$

$$< z_0^*, G(x_0) > = 0$$

   Since f', G' are Fréchet differentiable we conclude for $\pm h$:

$$f'(x_0) + < z_0^*, G'(x_0) > = 0$$

□

Definition 8.19 (Sublinear functionals)**.** *A real-valued function p defined on a normed vector space X is a sublinear functional, iff $p(x+y) \leq p(x)+p(y)$ and $p(\alpha x) = \alpha p(x)$ for all $x, y \in X$ and $\alpha \in \mathbb{R}$.*

Theorem 8.20 (Hahn-Banach)**.** *Let X be a real normed vector space and p a continuous sublinear functional on X. Let f be a bounded linear functional defined on a subspace M of X with $f(m) \leq p(m)\ \forall m \in M$. Then there exists a bounded linear functional F extending f from M to X and such that $F(x) \leq p(x)$ on X.*

A proof can be found in [2]. To obtain a geometric interpretation of Hahn-Banach's theorem we recall that $|x| = p(x)$ is a sub–linear functional on X. We assume that $0 \notin M$. Then, for a subspace M we define the linear functional with $f(m) = -1$ for all $m \in M$. Then, the Banach theorem tells us that this functional can be extended to X by $F \leq p$. Since $p(0) = 0$ and the set $M \subset K = \{x \in X : F(x) = -1\}$ is a hyerplane, we have separated $x = 0$ from M by K.

The variant used in the proof above is known as the Eidelheit Separation Theorem.

Theorem 8.21. *Let K_1, K_2 be convex sets in the normed vector space X such that K_1 has interior points and K_2 does not contain any interior point of K_1. Then there is a closed hyperplane H (i.e. there exists a linear functional x^* on X such that $H = \{x :< x^*, x >= c\}$.) separating K_1 and K_2. I.e., there exists a linear functional x^* on X such that*

$$\sup_{x \in K_1} < x^*, x > \leq c \leq \inf_{x \in K_2} < x^*, x >$$

A proof can be found in [10]. We give the proof in a slightly different setting.

Proof. Consider $K = K_1 - K_2$. Then $0 \notin K$ since K_2 contains no interior point of K_1. Now, additionaly we assume that K is a **ball** with center x_0, radius $r = 1$ and consider the shifted ball $\tilde{K} = K - x_0$. Hence, $-x_0$ is not in the interior of $\tilde{K}$. Further, $-x_0$ spans a one-dimensional subspace $M = \{-\alpha x_0 : \alpha \in \mathbb{R}\}$. We consider the functional $f(\alpha(-x_0)) = \alpha\|x_0\|$ on M. f is bounded and linear on M with norm 1. Hence (with $p(m) = \|m\|$) we apply Hahn-Banachs theorem and obtain the functional F on X with norm 1. For all points x in the interior of $\tilde{K}$ we conclude

$$F(x) \leq \|F\|\|x\| < 1\|x_0\| = f(-x_0) = F(-x_0)$$

We decompose $x \in \tilde{K}$ as follows

$$x = x_1 - x_2 + (-x_0),\ x_i \in K_i$$

and obtain by the linearity of F

$$F(x_1) - F(x_2) + F(-x_0) < F(-x_0)$$

Hence, $x^* = F$ and $c = F(-x_0)$. In the general proof we have to replace the $p(\cdot)$ by the Minkowski functional. □

Next, we discuss further constraint qualifications. We note that the qualification used in the previous theorem is not sufficient, since e.g. there are no interior points in $L^p(\mathbb{R}^n)$. The other famous condition is due to Zowe et. al. and dates back to 1979, and Robinson (1976). The presentation of other possible constraint qualifications is similar to the finite–dimensional case. We start to transfer the steps in the finite–dimensional case to infinite–dimensions. The proofs given before are more or less direct proofs. Here, we proceed similar to the presentation given in the previous section on finite space dimensions or in the book of Spellucci.

Consider the following problem

$$\min_{x \in X} f(x) \text{ subject to } g(x) \in K, x \in C \tag{8.32}$$ PG

under the assumptions **A** that

1. X, Z are Banach spaces
2. $f : X \to \mathbb{R}, g : X \to Z$ continuously Frechet differentiable
3. $C \subset X$ is non–empty, closed and convex (in order to guarantee existence)
4. $K \subset Z$ is a closed, convex cone (as extension to inequality constraints in finite dimensions), i.e.,

$$\forall \lambda \geq 0, z \in K \implies \lambda z \in K.$$

5. The feasible set $X_{ad} = \{x \in X : g(x) \in K, x \in C.\}$ is non-empty.

Exactly, as in the finite–dimensional case we define the tangenet cone of X_{ad} at a point $x \in X_{ad}$ by

$$T_{X_{ad},x} := \{s \in X : \exists \eta_k \in \mathbb{R}, \eta_k > 0, x_k \in X_{ad} : \lim_k x_k = x, \lim_k \eta_k(x_k - x) = s\} \tag{8.33}$$ tangent cone

and we proof completly analogously the following theorem.

Theorem 8.22. *Let the assumptions* **A** *hold true. Then, any local solution* x^* *of (8.32) satisfies the optimality condition*

$$x^* \in X_{ad}, < f'(x^*), s >_{X^*,X} \geq 0 \; \forall s \in T_{X_{ad},x^*}.$$

Proof. By assumption $x^* \in X_{ad}$. Furthermore, for any s by definition of T there exists a sequence η_k and x_k approximating s and x^*. Since x^* is a local minimum and since η_k is positive we have

$$0 \leq \eta_k \left(f(x_k) - f(x^*) \right) = \eta_k < f'(x^*), x_k - x^* > + \eta_k o(\|x_k - x^*\|.$$

Taking the limit on both sides we obtain the desired inequality. □Of courese, this optimality condition cannot be used in applications, since the tangent cone is too complicated to work with. As in the finite–dimensional case we seek for characterizations of the tangent cone in terms of the describing the equality and inequality constraints. Comparing with the assumptions of our main theorem we investigate at first the following cone (this is *not*) the most general. It corresponds to the LICQ condition. The linearized cone at a point $x \in X_{ad}$ is defined as

$$L_{X_{ad},x,g} = \{\eta s : \eta > 0, s \in X, g(x) + g'(x)s \in K, x + s \in C\} \qquad (8.34) \quad \boxed{\texttt{LICQ-I}}$$

The following result is then easy consequence of the previous theorem.

Theorem 8.23. *Let the assumptions* **A** *hold true. Let* x^* *be a local solution to (8.32). Assume additionally that*

$$L_{X_{ad},x^*,g} \subset T_{X_{ad},x^*} \qquad (8.35) \quad \boxed{\texttt{ACQ}}$$

Then, the following necessary optimality conditions holds true.

$$x^* \in X_{ad}, < f'(x^*), s >_{X^*,X} \geq 0 \; \forall s \in L_{X_{ad},g,x^*}.$$

The question is now whether or not (8.35) holds true. Comparing with (8.18) we observe that it is reasonable to require that there exists $x \in K, x \in L_{X_{ad},x,g}$ satisfying a strict inequality constraint (this requires that K has interior points, see computations below).

Lemma 8.24. *Let the assumptions* **A** *hold true. Let* g *be affine linear. Then, (8.35) is satisfied.*

Proof. Let $s \in L_{X_{ad},g,x}$ be given. Then, there exists $\eta > 0$ such that $K \in g(x) + \eta s g'(x) = g(x + \eta s)$ and $x + \eta s \in C$. By convexity of K we ave

$g(x+\eta/ks) \in K$ and hence $x+\eta/ks \in X_{ad}$ and similarly $x+\eta/ks \in C$. Hence, we conlude that $x_k = x+\eta_k s$ and $\eta_k = \eta/s$ satisfies the assumptions on $T_{X_{ad},x}$. □

Another more general condition is due to Robinson:

$$x^* \in X_{ad},\ g(x^*) + g'(x^*)(C - x^*) \in \text{ interioir } K \tag{8.36}$$ RR

The condition is understood in the following sense: there exists $c \in C$ the point $g(x^*) + g'(x^*)(c - x^*)$ is belongs to the interior of K. In particular, this requires that there exists an interioir point in K. If $C \equiv X$, then since $x^* \in C$ the condition can be rewritten as

$$x^* \in X_{ad},\ g(x^*) + g'(x^*)s < 0 \tag{8.37}$$ RR-2

being exactly the definition of a regular point in the case of inequality constraints.

Theorem 8.25. *The condition (8.36) implies (8.35)*

Summarizing, the previous results we obtain the following theorem (Zowe et. al.)

Theorem 8.26. *Let the assumptions* **A** *hold true. Then, for any local solution x^* of (8.32), where condition (8.36) holds true, the following optimality condiiton are necessary. There exists $z^* \in Z^*$ such that*

$$g(x^*) \in K, x^* \in C \tag{8.38}$$
$$z^* \in \{\bar{z} \in Z^* :< \bar{z}, z >\leq 0\ \forall z \in K\} \tag{8.39}$$
$$< z^*, g(x^*) >= 0 \tag{8.40}$$
$$< f'(x^*) + g'(x^*)z^*, x - x^* >\geq 0\ \forall x \in C \tag{8.41}$$

We apply the previous result problems in PDE–constrained optimization. Consider the following situation. Let $\Omega \subset \mathbb{R}^n$ be sufficiently regular. Let Y, U be Banach spaces. $U_{ad} \subset U$ a closed convex set in U and $K_Y \subset \tilde{Y}$ a closed convex cone in a Banach space $\tilde{Y}$. Assume that $Y \subset \tilde{Y}$.[2]

The PDE is described by $c(y,u) = 0$ where $c : Y \times U \to X$ is continuosly Frechet differentiable. We consdier the problem

$$\min f(y,u) \text{ subject to } c(y,u) = 0,\ y \in K_y, u \in U_{ad}$$

[2]Example. $Y = H^1(\Omega)$, $\tilde{Y} = C(\bar{\Omega}), n \leq 3$ due to Sobolev embedding.

Hence, the constraint operator is

$$H : Y \times U \in X \times \tilde{Y} \text{ s.t. } H(y,u) = (c(y,u), y) \in K := \{0\} \times K_Y$$

and the feasible set is

$$C := Y \times U_{ad}.$$

Robinson's regularity condition at a feasible point $x = (y,u)$ reads therefore

$$\begin{pmatrix} 0 \\ K_y \end{pmatrix} \in \text{ interior } \left(\begin{pmatrix} 0 \\ y \end{pmatrix} + \begin{pmatrix} c_y(y,u) & c_u(y,u) \\ I_{Y,\tilde{Y}} & 0 \end{pmatrix} \begin{pmatrix} Y \\ U_{ad} - u \end{pmatrix} \right).$$

We give another condition implying Robinsons's regularity condition.

Lemma 8.27. *Let $(y,u) \in C$. If $c_y(y,u) \in L(Y;X)$ is surjective and if there exists $\tilde{u} \in U_{ad}$ and $\tilde{y} \in$ interior K_y such that*

$$c_y(y,u)(\tilde{y} - y) + c_u(y,u)(\tilde{u} - u) = 0,$$

then Robinson's regularity condition is satisfied.

Note that the previous is exactly the condition given by the multiplier theorem for inequality constraints applied to $c(y,u)$.

Proof. Denote by $B_{Y,\epsilon}$ the ball in Y with radius ϵ centered at the point 0. Since $\tilde{y}$ belongs to the interior of K_Y there exists $\epsilon > 0$, s.t., $\tilde{y} + B_{\tilde{Y},2\epsilon} \subset K_Y$. Since $B_{Y,\epsilon}$ is open in Y and c_y is a linear and bounded operator, we obtain by the open mapping theorem that $c_y(y,u)B_{Y,\epsilon}$ is an open set in X. Since Y is a Banach space $\tilde{y} - y + B_{Y,\epsilon} \subset Y$. Finally, we have that $\tilde{u} \in U_{ad}$, $c_y(y,u)(\tilde{y} - y) + c_u(y,u)(\tilde{u} - u) = 0$, $K_Y \subset \tilde{Y}$ and $Y \subset \tilde{Y}$.

We start with Robinson's regularity condition.

$$\begin{pmatrix} 0 \\ K_y \end{pmatrix} \in \text{ interior } \left(\begin{pmatrix} 0 \\ y \end{pmatrix} + \begin{pmatrix} c_y(y,u) & c_u(y,u) \\ I_{Y,\tilde{Y}} & 0 \end{pmatrix} \begin{pmatrix} Y \\ U_{ad} - u \end{pmatrix} \right)$$

rewritten as

$$\begin{pmatrix} 0 \\ y \end{pmatrix} + \begin{pmatrix} c_y(y,u) & c_u(y,u) \\ I_{Y,\tilde{Y}} & 0 \end{pmatrix} \begin{pmatrix} Y \\ U_{ad} - u \end{pmatrix} - \begin{pmatrix} 0 \\ K_y \end{pmatrix}$$

$$\supset$$

$$\begin{pmatrix} 0 \\ y \end{pmatrix} + \begin{pmatrix} c_y(y,u) & c_u(y,u) \\ I_{Y,\tilde{Y}} & 0 \end{pmatrix} \begin{pmatrix} \tilde{y} - y + B_Y(\epsilon) \\ \tilde{u} - u \end{pmatrix} - \begin{pmatrix} 0 \\ \tilde{y} + B_{\tilde{Y},2\epsilon} \end{pmatrix}$$

since $\tilde{y} - y + B_Y(\epsilon) \subset Y$ and $\tilde{u} - u \in U_{ad}$ and $\tilde{y} + B_{\tilde{Y},2\epsilon}$ and then

$$\begin{aligned} &\begin{pmatrix} c_y(y,u) & c_u(y,u) \\ I_{Y,\tilde{Y}} & 0 \end{pmatrix} \begin{pmatrix} \tilde{y} - y + B_Y(\epsilon) \\ \tilde{u} - u \end{pmatrix} - \begin{pmatrix} 0 \\ -y + \tilde{y} + B_{\tilde{Y},2\epsilon} \end{pmatrix} \\ = &\begin{pmatrix} c_y(y,u) \\ I_{Y,\tilde{Y}} \end{pmatrix} B_{Y,\epsilon} + \begin{pmatrix} c_y(y,u)(\tilde{y} - y) + c_u(y,u)(\tilde{u} - u) \\ (\tilde{y} - y) \end{pmatrix} - \begin{pmatrix} 0 \\ -y + \tilde{y} + B_{\tilde{Y},2\epsilon} \end{pmatrix} \\ = &\begin{pmatrix} c_y(y,u) \\ I_{Y,\tilde{Y}} \end{pmatrix} B_{Y,\epsilon} + \begin{pmatrix} 0 \\ B_{\tilde{Y},2\epsilon} \end{pmatrix} \supset \begin{pmatrix} c_y(y,u) B_{Y,\epsilon} \\ B_{\tilde{Y},2\epsilon} \end{pmatrix} \end{aligned}$$

since $Y \subset \tilde{Y}$ and $c_y(y,u)B_{Y,\epsilon}$ is open in X containing the zero. Hence, the set on the right hand side is an open neighborhood in $X \times \tilde{Y}$. Hence, there exists an interior and Robinson's condition is satisified. □ The previous computations can be applied to treat the following elliptic problem with state constraints. Consider the problem on a domain Ω with smooth boundary.

$$\min \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \tag{8.42}$$

$$\text{subject to} \tag{8.43}$$

$$-\Delta y + y = \gamma u \ \text{ in } \Omega \tag{8.44}$$

$$\partial_v y = 0 \ \text{ on } \partial\Omega \tag{8.45}$$

$$y \geq 0 \ \text{ on } \partial\Omega \tag{8.46}$$

Let $n \leq 3$ and we know that there exists a solution operator

$$S : u \in U = L^2 \to y \in H^1 \cap C(\bar{\Omega})$$

mapping to the space of continuous functions. We rewrite the problem using an operator $A : Y \to X$ defined through the bilinear form

$$a(y,v) = \int \nabla y \nabla v dx + (y,v)_{L^2}$$

and

$$Bu = \gamma u : U \to X.$$

We set $U_{ad} = U$ and the spaces $Y = H^1$ and $X = (H^1)^*$, $\tilde{Y} \supset Y$ and $K_y = \{y \in \tilde{Y} : y \geq 0\}$. We introduce the following operator $c(y,u) : Y \times U \to X$ as $c(y,u) = Ay + Bu$.

Choosing $\tilde{Y} = H^1$ and $X = Y^*$ is possible, but the cone $K_y \subset \tilde{Y}$ has no interior point and therefore Robinson's regularity condition does not hold. However, the solution operator has higher regularity since it maps into the space $H^1 \cap C(\bar{\Omega})$. The operator $B : U \to X$ is an operator on L^2 we can choose $X = L^2$. We now choose

$$Y = \{y \in H^1 \cap C(\bar{\Omega}) : Ay \in L^2\}.$$

This is a Banach space with norm $\|\cdot\|_{H^1} + \|\cdot\|_C + \|A\cdot\|_{L^2}$ and the operator A is then well–defined as operator $A : Y \to X = L^2$. Furthermore, it is bounded and surjective. Since $c_y(y,u) = A$ we have hence satisfied the first assumption of the previous Lemma. Setting now $\tilde{Y} = C(\bar{\Omega})$. Then, $K_Y \subset \tilde{Y}$ and K_Y has an interior point. Last, we have to check, that there exists point such that for $\tilde{y} \in$ interior of K_Y there exists $\tilde{u} \in U_{ad} = L^2$ such that

$$0 = A(\tilde{y} - y) + B(\tilde{u} - u)$$

Using $\tilde{y} = y + \gamma$ and $\tilde{u} = u$ gives the result for any feasible point $(y,u) \in Y \times U_{ad}$. Hence, the condition is satisfied and we can state the necessary first order conditions at the optimal point $(\bar{y}, \bar{u})$ as follows.

$$A\bar{y} + B\bar{u} = 0, \bar{y} \geq 0 \tag{8.47}$$

$$(\bar{y} - y_d, v)_{L^2} + (\bar{p}, Av)_{L^2} + < \bar{q}, v >_{C^*,C} = 0 \tag{8.48}$$

$$\bar{q} \in K_Y^*, \ < \bar{q}, \bar{y} >_{C^*,C} = 0 \tag{8.49}$$

$$(\alpha\bar{u} + \gamma\bar{p}, u - \bar{u}) \geq 0 \ \forall u \in U \tag{8.50}$$

We also give the strong form. The dual C^* consists of regular Borel measures (Radon measures) and the set K_Y^* is the set of nonpositive functionals on C^* which can be identified as measures $-\mu_\Omega, -\mu_{\partial\Omega}$, respectively, i.e., we have

$$q \in K_Y^*, < q, v >_{C^*,C} = -\int_\Omega v d\mu_\Omega - \int_{\partial\Omega} v d\mu_{\partial\Omega}.$$

Therefore, we obtain

$$-\Delta\bar{y} + \bar{y} = \gamma\bar{u}, \partial_\nu\bar{y} = 0, \bar{y} \geq 0 \tag{8.51}$$

$$-\Delta\bar{p} + \bar{p} = -(\bar{y} - y_d) + \mu_\Omega, \partial_\nu\bar{p} = \mu_{\partial\Omega} \tag{8.52}$$

$$\mu_\Omega, \ \mu_{\partial\Omega} \geq 0 \tag{8.53}$$

$$\int_\Omega \bar{y} d\mu_\Omega - \int_{\partial\Omega} \bar{y} d\mu_{\partial\Omega} (\bar{y} - y_d, v)_{L^2} + (\bar{p}, Av)_{L^2} + < \bar{q}, v >_{C^*,C} = 0 \tag{8.54}$$

$$\alpha\bar{u} + \gamma\bar{p} = 0 \tag{8.55}$$

8.3 Examples of PDE constrained optimization problems

We consider parabolic optimal control problems in one spatial dimension. As an example one may consider heat conductivity problems. The PDE is the linear heat equation with Robin–type boundary conditions. A theoretical of this problem can be either done using weak solutions in anisotropic Sobolev spaces or using classical theory based on Green's function. We consider the later in this paragraph. The theory of anisotropic Sobolev spaces leads to operator equations similar to those discussed for elliptic problems. The setting is now as follows.

Let Ω be a domain and y_0 an initial temperature distribution. The optimal control should drive the a temperature $y(x,t)$ in time time T to some desired temperature y_d. Clearly, we assume that the object can only be heated through the boundary. This leads to

$$\min \frac{1}{2}\int_\Omega (y(x,T)-y_d)^2 dx + \frac{\lambda}{2}\int_0^T \int_\Omega u(x,t)^2 ds(x)dt \tag{8.56}$$

labell-2

under the constraints

$$y_t - \Delta y = 0 \in (0,T)\times\Omega =: Q \tag{8.57a}$$
$$\nabla y n(x) + \alpha y = \beta u, \text{ on } (0,T)\times\partial\Omega =: \Gamma \tag{8.57b}$$
$$y(x,0) = y_0(x) \in \Omega \tag{8.57c}$$

We also might add pointwise constraints on the control of the type

$$0 \le u(x,t) \le 1 \text{ on } \Gamma \tag{8.58}$$

labell-3

The set of admissible controls is defined by

$$U_{ad} = \{u \in L^2(\Gamma) : u \text{ satisfies (8.58)}\}.$$

At first, we derive a formal optimality system and prove later on that the formal computations are feasible. The Lagrange function corresponding to (8.56) is given by

$$L(y,u,p) = J(y,u) - \int_Q (y_t - \Delta y)p dxdt - \int_\Gamma (\nabla y n + \alpha y - \beta u)p ds dt.$$

We expect the necessary optimality conditions to be

$$D_y L(\bar y, \bar u, \bar p)y = 0, D_u L(\bar y, \bar u, \bar p)(u-\bar u) \ge 0$$

for all variations y with $y(0) = 0$ and all $u \in U_{ad}$. We linearize L with respect to y and obtain

$$D_y L(\bar{y}, \bar{u}, \bar{p})y = \int_\Omega (\bar{y}(x,T) - y_d)y(T)dx - \int_Q (y_t - \Delta y)pdxdt \tag{8.59}$$

$$- \int_\Gamma (\nabla y n + \alpha y)pdst = \tag{8.60}$$

$$\int_\Omega (\bar{y}(x,T) - y_d)y(T)dx - \int_Q (-p_t - \Delta p)ydxdt - \int_\Omega y(T)p(T) \tag{8.61}$$

$$+ \int_\Gamma \nabla y n p dst + \int_\Gamma y \nabla p n dst - \int_\Gamma (\nabla y n + \alpha y)pdst \tag{8.62}$$

If we choose $y \in C_0^\infty(Q)$, then boundary terms disappear and we obtain in L^2 the following identity

$$p_t + \Delta p = 0.$$

If we choose test functions such that $y(T) \neq 0$ we obtain as equailty in L^2 that

$$p(T) = \bar{y}(T) - y_d \in \Omega$$

Finally, if we allow functions y to vary also on the boundary of the domain we obtain

$$\nabla p n + \alpha p = 0.$$

These equations are the adjoint equations for the problem as derived formally. It remains to derive the gradient equation by computing the derivative of L with respect to u.

$$D_u L(\bar{y}, \bar{u}, \bar{p})(u - \bar{u}) = \lambda \int_\Gamma \bar{u}(u - \bar{u})dst + \int_\Gamma \beta p(u - \bar{u})dst \tag{8.63}$$

and hence the variational inequality

$$\int_\Gamma (\lambda \bar{u} + \beta p)(u - \bar{u})dst \geq 0 \forall u \in U_{ad}. \tag{8.64}$$

The previous computations are valid only in the case where the regularity of y and p is such that y_t and p_t is well-defined. A setting in which this is possible is given in the case of one–dimensional problem using Green's function.

Let us assume that $\Omega = (0, 1)$ and the control is prescribed on the right boundary $x = 1$ only. It is atime-dependent control $u(t)$. At the boundary $x = 0$ we prescirbe Neumann conditions. Furthermore, we assume

an initial condition of $y_0 = 0$. The numbers $T > 0$, $\lambda \geq 0$ and $\beta \geq 0$ with $\beta \in L^\infty(0,T)$. Furthermore, we assume that $f, y_d \in L^2(0,1)$. We look for optimal controls u in $U_{ad} \subset L^2(0,T)$.

In order to discuss rigorous results for the optimization problem (8.56) we consider the general initial-value problem for the one-dimensional heat equation given by

$$y_t - y_{xx} = f \tag{8.65}$$
$$y_x(0,t) = 0 \tag{8.66}$$
$$y_x(1,t) + \alpha y(1,t) = u(t) \tag{8.67}$$
$$y(x,0) = y_0(x) \tag{8.68}$$

It can be shown that the equation has the following solution

$$y(x,t) = \int_0^1 G(x,\xi,t) y_0(\xi) d\xi + \int_0^t \int_0^1 G(x,\xi,t-s) f(\xi,s) d\xi ds + \int_0^t G(x,1,t-s) u(s) ds \tag{8.69}$$

where G is the Green's function. The precise type of the Green's function depends on the boundary conditions as well as the domain. It can be shown that in the previous case we have

$$\alpha = 0: \qquad G(x,\xi,t) = 1 + 2\sum_{n=1}^{\infty} \cos(n\pi x)\cos(nx\pi)\exp(-n^2\pi^2 t) \tag{8.70}$$

$$\alpha > 0: \qquad G(x,\xi,t) = \sum_{n=1}^{\infty} \frac{1}{N_n} \cos(\mu_n x)\cos(\mu_n \xi)\exp(-\mu_n^2 t) \tag{8.71}$$

Here, $\mu_n \geq 0$ are the solutions to the equation $\mu \tan(\mu) = \alpha$ and we have $N_n = \frac{1}{2} + \sin(2\mu_n)/(4\mu_n)$.

Exercise 8.28. *Show that G is non–negative, symmetric in x and ξ and singular at $x = \xi$ as well as $t = 0$.*

One has the following regularity result. Provided the assumptions on f, y_0 are satisfied we have that $y \in L^2(\Omega)$ if $u \in L^2(0,T)$. Hence, the solution is not a classical, but a generalized solution to the one-dimensional heat equation.

1. In the case of control at the boundary we set in the previous formula $f = y_0 = 0$. Then, the problem reads: minimize (8.56) subject to $y_y - y_{xx} = 0, y_x(0,t) = 0, y_x(1,t) + \alpha y(1,t) = u, y(x,0) = 0$.

In the cost functional only the terminal costs $y(x,T)$ appear and they can be rewritten as

$$y(x,T) = \int_0^T G(x,1,T-s)\beta(s)u(s)ds = (Su)(x)$$

for the integral operator S mapping $L^2(0,T)$ into $L^2(0,1)$. It can be shown that S is in fact a bounded linear operator.

2. In the case of distributed control the function f is the control and $u = y_0 = 0$. The problem reads minimize

$$\int_Q \frac{1}{2}(y(x,t) - y_d)^2 dxdt + \frac{\lambda}{2}u^2 dxdt$$

subject to $y_t - y_{xx} = f, y_x(0,t) = 0, y_x(1,t) + \alpha y(1,t) = 0, y(x,0) = 0$ where we now assume that y_d is in $L^2(Q)$. We replace again y in the cost functional by it's representation as Green's function and obtain

$$y(x,t) = \int_0^t \int_0^1 G(x,\xi,t-s)f(\xi,s)d\xi ds = (Sf)(x,t).$$

Again, S is a linear (and even bounded) operator from $L^2(Q)$ to $L^2(Q)$.

We discuss briefly existence results and optimality conditions. Using the operator representation we obtain as formulation in the first case:

$$\min_{u\ inU_{ad}} f(u) = \frac{1}{2}\|Su - y_d\|_{L^2(0,1)} + \frac{\lambda}{2}\|u\|_{L^2(0,T)} \tag{8.72}$$

labell-4

We have existence due to Theorem 9.16 and the necessary optimality cnodition at a minimum $\bar{u}$

$$\left(S^*S(\bar{u} - y_d) + \lambda\bar{u}, u - \bar{u}\right) \geq 0. \tag{8.73}$$

In order to obtain the first-order system we introduce the adjoint operator S^* and discuss its properties. The adjoint is defined by the equality in the Hilbert space $U = L^2(0,1)$ and $V = L^2(0,T)$ with $S : V \to U$

$$(v, Su)_U = (S^*v, u)_V$$

and we obtain by the following calculation for any $u, v \in U$ that

$$(v, Su)_U = \int_0^1 v(x) \int_0^T G(x, 1, T-s)\beta(s)u(s)dsdx \tag{8.74}$$

$$= \int_0^1 \int_0^T G(x, 1, T-s)\beta(s)v(x)u(s)dsdx \tag{8.75}$$

$$= \int_0^T u(s) \int_0^1 u(s)\beta(s)G(x, 1, T-s)v(x)dxds \tag{8.76}$$

$$= \int_0^T (S^*v)(s)u(s)ds = (S^*v, u)_V \tag{8.77}$$

$$S^*v(s) = \beta(s) \int_0^1 G(\xi, 1, T-s)v(\xi)d\xi \tag{8.78}$$

Lemma 8.29. *We have $S^*v(t) = \beta(t)p(1,t)$ where p is the solution to the parabolic problem*

$$-p_t - p_{xx} = 0,\ p_x(0,t) = 0,\ p_x(1,t) + \alpha p(1,t) = 0,\ p(x,T) = v(x) \tag{8.79}$$

Proof. From the definition of S^* and due to the symmetry of Green's function with respect to the first two arguments we obtain

$$(S^*v)(t) = \beta(t) \int_0^1 G(1, \xi, T-t)v(\xi)d\xi \tag{8.80}$$

This defines a function

$$p(t) = \int_0^1 G(1, \xi, T-t)v(\xi)d\xi.$$

In the definition of S^* we now transform time by using $\tau = T - t$. We thus obtain

$$\tilde{y}(x, \tau) = p(x, T-\tau) = \int_0^1 G(1, \xi, \tau)v(\xi)d\xi.$$

Hence, $\tilde{y}$ satisfies the equation

$$\tilde{y}_\tau - \tilde{y}_{xx} = 0,\ \tilde{y}_x(0,\tau) = 0,\ \tilde{y}_x(1,\tau) + \alpha\tilde{y}(1,\tau) = 0,\ \tilde{y}(x,0) = v(x) \tag{8.81}$$

Now, using again the substiution we recover p as $p(x,t) = \tilde{y}(x, T-t)$. Note that provided $\tilde{y}$ has sufficient regularity we have $p_t = -\tilde{y}_t$. □

Combining the previous lemma with the necessary variational inequality gives the following result.

Theorem 8.30. *A control $\bar{u} \in U_{ad}$ with corresponding state $\bar{y} = S\bar{u}$ is optimal, if and only if the following inequality holds for $\bar{p}$ satisfying (8.79) with $v = (\bar{y}(T) - y_d)$.*

$$\int_0^T \big(\beta(t)\bar{p}(1,t) + \lambda\bar{u}(t)\big)\big(u(t) - \bar{u}(t)\big)\,dt \geq 0,\ \forall u \in U_{ad}. \tag{8.82}$$

labell-8

We want to obtain a pointwise result we first prove the following result.

Lemma 8.31. *Let Ω be an open set. Let $u, v, \phi, u_a, u_b \in U = L^2(\Omega), \Omega \subset \mathbb{R}$ and $U_{ad} = \{u \in L^2(\Omega) : u_a(x) \leq u(x) \leq u_b(x)\ a.e.\ \}$ satisfy*

$$\int_\Omega \phi(x)\big(v(x) - u(x)\big)\,dx \geq 0,\ \forall v(x) \in U_{ad}.$$

Then, we have

$$\phi(x)\big(v - u(x)\big) \geq 0$$

for all $v \in [u_a(x), u_b(x)]$ and a.e. in $x \in \Omega$.

Proof. Since $z \in L^2(0,1)$ we have that almost everywhere in x the following limit exists

$$\lim_{\rho\to 0} \frac{1}{\|B_\rho(x)\|} \int_{B_\rho(x)} \int_{B_\rho(x)} \phi(\xi)d\xi = \phi(x).$$

Assuming ρ is sufficiently small we have that $B_\rho \in \Omega$. A similar equality holds for u, u_a, u_b. Consider now the set of points where the equality holds for u and ϕ (the complement of this set has measure zero). Choose such a point x_0 and $w \in [u_a(x_0), u_b(x_0)] \in \mathbb{R}$. Define $v(x) = w$ in $B_\rho(x_0)$ and $v(x) = u(x)$ for $x \in \Omega\backslash B_\rho(x_0)$. The variational inequality then yields

$$0 \leq \frac{1}{\|B_\rho(x)\|} \int_\Omega \phi(x)(v(x) - u(x))dx = \tag{8.83}$$

$$\frac{1}{\|B_\rho(x)\|} \int_{B_\rho(x)} \phi(x)(w - u(x))dx \tag{8.84}$$

$$\to_{\rho\to 0} \tag{8.85}$$

$$0 \leq \phi(x_0)(w - u(x_0)) \tag{8.86}$$

This proves the lemma. □

From the previous lemma we also obtain after simple rearrangment:

$$\phi(x)v \geq \phi(x)u(x)\ \forall v \in [u_a(x), u_b(x)], a.e.\ in\ x. \tag{8.87}$$

and hence a.e. in x

$$\min_{v\in[u_a(x),u_b(x)]} \phi(x)v = \phi(x)u(x). \tag{8.88}$$

This equality implies the following for a.e. x

$$\phi(x) > 0 : u(x) = u_a(x) \quad \phi(x) = 0 : u(x) \in [u_a(x), u_b(x)] \tag{8.89}$$
$$\phi(x) < 0 : u(x) = u_b(x) \tag{8.90}$$

labell-7 **Lemma 8.32.** *Let Ω be an open set. Let $u, v, \phi, u_a, u_b \in U = L^2(\Omega), \Omega \subset \mathbb{R}$ and $U_{ad} = \{u \in L^2(\Omega) : u_a(x) \le u(x) \le u_b(x)$ a.e. $\}$ satisfy*

$$\int_\Omega \phi(x)\big(v(x) - u(x)\big)\,dx \ge 0,\ \forall v(x) \in U_{ad}.$$

Then, we have

$$\phi(x) > 0 : u(x) = u_a(x) \tag{8.91}$$
$$\phi(x) = 0 : u(x) \in [u_a(x), u_b(x)] \tag{8.92}$$
$$\phi(x) < 0 : u(x) = u_b(x) \tag{8.93}$$

We apply lemma 8.32 to the parabolic problem and obtain

labell-9 **Lemma 8.33.** *Let $U_{ad} = \{u \in L^2(0,T) : u_a(t) \le u(t) \le u_b(t) a.e. int\}$ for some functions $u_a, u_b \in L^2(0,T)$. Then, the variational inequality*

$$\int_0^T \big(\beta(t)\bar{p}(1,t) + \lambda\bar{u}(t)\big)\big(u(t) - \bar{u}(t)\big)\,dt \ge 0,\ \forall u \in U_{ad}$$

holds true, if and only if a.e. in t

$$\bar{u}(t) = P_{[u_a(t),u_b(t)]}(-\beta(t)/\lambda(t)p(1,t)),\ \lambda > 0$$

and using the Heavi-side function H

$$\bar{u}(t) = u_a(x)H(\beta(t)p(1,t)) + u_b(x)H(-\beta(t)p(1,t)),\ \lambda = 0$$

Proof. This is a direct consequence of lemma 8.32 with $\phi(t) = \beta(t)p(1,t) + \lambda\bar{u}(t)$. Assume first $\lambda 0$ then we have

$$(\frac{\beta(t)p(1,t)}{\lambda} + \bar{u}(t))v \ge (\frac{\beta(t)p(1,t)}{\lambda} + \bar{u}(t))\bar{u}(t).$$

Hence, if $-\frac{\beta(t)p(1,t)}{\lambda} \leq u_a(t)$, then $0 \leq u_a + \frac{\beta(t)p(1,t)}{\lambda} \leq u + \frac{\beta(t)p(1,t)}{\lambda}$ and therefore $u(t) = u_a(t)$. Similarly, we have $u(t) = u_b(t)$ if $-\frac{\beta(t)p(1,t)}{\lambda} \geq u_b(t)$. In the case $\frac{\beta(t)}{\lambda}p(1,t) + \bar{u}(t) = 0$ we have $\bar{u}(t) = -\frac{\beta(t)}{\lambda}p(1,t)$. This proves the result. Except for the last case the discussion is similar in the case $\lambda = 0$. In the case $\lambda = 0$ $u(t)$ is not defined if $\beta p = 0$. The converse directions follows from integration of the given inequalities. □

In the case of distributed control we obtain the following result.

Theorem 8.34. *A control $\bar{u} \in U_{ad} = \{u \in L^2(Q) : u_a \leq u \leq u_b a.e. in x, t\}$ for some functions $u_a, u_b \in L^2(0,T)$ is optimal for the problem*

$$\min \int_Q \frac{1}{2}(y(x,t) - y_d)^2 dxdt + \frac{\lambda}{2}u^2 dxdt$$

subject to $y_t - y_{xx} = u, y_x(0,t) = 0, y_x(1,t) = 0, y(x,0) = 0$, if and only if

$$\big(\bar{p}(x,t) + \lambda\bar{u}(x,t)\big)\big(v - \bar{u}(x,t)\big) \geq 0,\ v \in [u_a(x,t), u_b(x,t)].$$

where $\bar{p}$ satisfies the following set of equations

$$-p_t = p_{xx} + y - y_d,\ p_x(0,t) = 0, p_x(1,t) = 0, p(x,T) = 0. \tag{8.94}$$

Exercise 8.35. *Prove the previous theorem.*

Finally, we discuss the case $\lambda = 0$ in more detail. For simplicity we set $U_{ad} = \{-1 \leq u \leq 1\}$ and $\beta = 1$ and consider the problem

$$\min \frac{1}{2}\int_0^1 \big(y(x,T) - y_d\big)^2 dx \tag{8.95}$$

$$y_t - y_{xx} = 0, y_x(0,t) = 0, y_x(1,t) = u(t) - \alpha y(1,t), y(x,0) = 0, u \in U_{ad} \tag{8.96}$$

Due to Lemma 8.33 we have

$$\bar{u}(t) = \begin{pmatrix} 1 & p(1,t) < 0 \\ -1 & p(1,t) \end{pmatrix}. \tag{8.97}$$

However, there is no information at points where $p(1,t) = 0$. It can be shown that the set of points where $p(1,t) = 0$ is of measure zero. This are the switching points and outside these points the control is either one or minus one. Therefore, we speak of a bang-bang control.

Theorem 8.36 (Bang-bang principle)**.** *Let $\bar{u}$ be the optimal control for the boundary control problem (8.95) and $\bar{y}$ the corresponding state. Assume that*

$$\|\bar{y} - y_d\|_{L^2(0,1)} > 0.$$

Then $p(1,t)$ has only a countable number of zeros. The only accumulation point is $t = T$. Further, $|u(t)| = 1$ a.e. in t.

A proof can be found in Troeltzsch, p78.

8.4 Necessary optimality conditions in the convex case – sensitivity, duality

As explained above the theory of Lagrange multipliers allows for a particular nice geometric interpretation in the case of convex optimization. We recall the definition of a positive vector and cone.

Let P be a convex cone in the vector space X. Then for $x, y \in X$ we write $x \geq y$ if $x - y \in P$. The cone $N = -P$ is the negative conce in X. We easily verify that $x \geq y, y \geq z$ implies $x \geq z$ and since $0 \in P$ we obtain $x \geq x$. In the normed space X we write $x > 0$ if x is an interior point of the positive cone. As before, we will require to have an interior point in order to apply a separating hyperplane argument.

Given a normed space X and a positive cone P we denote by P^* the corresponding positive cone in the dual space X^* defined by all linear functionals being positive on P:

$$P^* = \{x^* \in X^* :< x, x^* > \geq 0 \ \forall x \in P\} \tag{8.98}$$ luen2

Example. In the space $\mathbb{R}^n$ the positive functionals are given by the row–vectors with non–negative components. In the space $C^0(0,1)$ the positive cone are all continuous, non–negative functions. The dual cone consists of all functions of bounded variation which are non–decreasing.

Remark 8.37. *If f is a bounded linear functional on $X = C^0(0,1)$. Then there exists a function of v of bounded variation on $(0,1)$ such that for all $x \in X$*

$$f(x) = \int_0^1 x(t)dv(t).$$

and such that the norm of f is the total variation of v. Conversly every function defines a bounded linear functional on X.

This result does not hold for more than one space dimension. The total variation of a function v is defined by $TV(v) = \sup_{t_i} \sum_i |v(t_i) - v_{t_{i-1}}|$. Any

function v of bounded variation is continuous to the right. The characterization is not unique. One has to normalize such that e.g. $v(a) = 0$. The norm is TV. For any function v of bounded variation and x continuous the integral is well-defined as Stieltjes-Integral and computed as

$$\lim_{0<t_j<1} \sum_j x(t_j)|v(t_j) - v(t_{j-1})|.$$

Lemma 8.38. *Let the positive cone P in the normed space X be closed. If $x \in X$ satisfies $< x, x^* > \geq 0$ for all $x^* \geq 0$.*

Proof. Assume $x \notin P$. Then, by the separating hyperplane theorem, there is an $x^* \in X^*$ such that $< x, x^* > < < p, x^* >$ for all $p \in P$. Since P is a cone, $< p, x^* >$ can never be negative because then $< x, x^* >< \alpha p, x^* >$ for some $\alpha > 0$. Thus, $x^* \in P^*$. Since $\inf_{p \in P} < p, x^* >= 0$ we obtain $< x, x^* >< 0$ and a contradiction. □

Remark 8.39. *A hyperplane H in a linear vector space X can be characterized by a linear functional x^* in the following sense: $H = \{x \in X :< x, x^* >= c\}$. Hyperplanes including the origin, play a prominent role. If H does contain the origin, then there exists aunique linear functional x^* such that $H = \{x \in \mathbb{R} :< x, x^* >= 1.\}$. Any hyperplane is closed, if f is a continuous linear functional. Hence, we almost always are interested in closed hyperplanes.*

Hahn-Banach's Theorem can be expressed in terms of hyperplanes: Let K be a convex set with non-empty interior. Suppose that V is a linear subspace of X which does not contain any interior point of K. Then, there exists a closed hyperplane H containing V but not containing interior points of $K : < v, x^* >= c$ for all $v \in V$ and $< k, x^* >< c$ for all $k \in int(K)$. This can be furher specialised to Eidelheit's separation theorem, see above.

We give a general definition of convexity in normed spaces.

Definition 8.40. *Let X be a vector space and let Z be a vector space having a positive cone P. A mapping $G : X \to Z$ is said to be convex, if the domain Ω of G is a convex and if*

$$G(\alpha x + (1 - \alpha)y) \leq \alpha G(x) + (1 - \alpha)G(y)$$

for all $x, y \in \Omega$ and $\alpha \in (0, 1)$.

Note that by definition the set

$$\{x \in X : x \in \Omega, G(x) \leq z\}$$

is convex, provided that G and Ω are convex.

We will now consider the optimization problem

$$\min_{x\in X} f(x) \text{ subject to } x \in \Omega, G(x) \leq 0 \tag{8.99}$$ 1:1

under the assumptions (**A**) that

- X, Z are normed spaces, $G : X \to Z$, Z contains a positive cone.
- Ω is convex and G is convex and f is real-valued, convex

The analysis of the preceeding section is based on the following function. Let

$$\Gamma = \{z \in Z : \exists x \in \Omega : G(x) \leq z\}$$

and define

$$w(z) := \inf\{f(x) : x \in \Omega, G(x) \leq z\}$$

The original problem is of course the solution to $w(0)$. The set Γ is convex. We study hence variations of the feasible set and it's relation to the minimal functional value. We hvae the following properties: At first, w is convex, since the appearing sets in the following equation get smaller and smaller. $w(\alpha z_1+(1-\alpha)z_2) = \inf\{f(x) : x \in \Omega : G(x) \leq \alpha z_1+(1-\alpha)z_2\} \leq \inf\{f(x) : x = \alpha x_1 + (1-\alpha)x_2, x_i \in \Omega : G(x_i) \leq z_i\} \leq\leq \alpha \inf\{f(x) : x_1 \in \Omega : G(x_1) \leq z_1\} + (1-\alpha)\inf\{f(x) : x_2 \in \Omega : G(x_2) \leq z_2\}$ Second, the functional is decreasing in z, i.e., $z_1 \geq z_2 \implies w(z_1) \geq w(z_2)$. This also due to the infimum. Hence, a typical plot of $w(z)$ against z shows a decreasing convex function. Therefore, any hyperplane tangential at a point z_0 *lies below* the function w. Hence, if we would move the coordinate system such that the tangent at z_0 is horizontal, we observe that w is *minimized* at z_0. The shift of the tangent is viewed as adding a linear function to the tangent and w respectively. Another way, of saying this is as follows: Adding an appropriate linear functional $< z, z^* >$ to $w(z)$ the resulting combination

$$w(z)+ < z, z^* >$$

is minimized z_0. We are now interested to apply this procedure at $z_0 = 0$. The functional (or graphically the linear function) is z^* and will be later on the Lagrange multiplier. It is the element $1, z^*$ in $\mathbb{R} \times Z^*$. This idea results in the following theorem.

Theorem 8.41. *Assume (A) holds. Assume that there exists a point $x_1 \in \Omega$ such that $G(x_1) < 0$ (i.e., interior point of P must exists). Let $\mu_0 = \inf f(x)$*

subject to $x \in \Omega$ *and* $G(x) \leq 0$. *Assume* $|\mu_0| < \infty$. *Then, there is an element* $z^* \in Z^*$ *and* $z^* \geq 0$ *such that*

$$\mu_0 = \inf_{x \in \Omega} f(x) + < G(x), z^* > .$$

If the infimum is achieved at $x_0 \in \Omega$ *with* $G(x_0) \leq 0$, *then* $< G(x_0), z^* >= 0$.

A proof can be found in [**?**], pages 218. The convexity is explained by the reasoning on the position of the hyperplane above. The interior point condition is severe. It is used to obtain the existence of a non–vertical $z^* \neq \infty$ hyperplane. The case of $H(x) = 0$ excluded by the previous theorem. Note that the trick to set $G_1 = H$ and $G_2 = -H$ and $G = (G_1, G_2)$ does not(!) work, since there is no point satisfying $G(x_1) < 0$. There is a version of this theorem with $H : X \to Y$ being affine linear and Y being finite–dimensional.

Lemma 8.42. *Assume the setting of the previous theorem. Assume* x_0 *achieves the constrained minimum. Then, there is a* z_0^* *such that the Lagrangian*

$$L(x, z^*) = f(x) + < G(x), z^* >$$

has a saddle point at (x_0, z_0^*).

Proof. A saddle point at x_0 and z_0^* is characterized by

$$L(x_0, z^*) \leq L(x_0, z_0^*) \leq L(x, z_0^*)$$

for all $x \in \Omega$ and $z^* \geq 0$. We take z_0^* from the previous theorem. From

$$\mu_0 = \inf_{x \in \Omega} \{f(x) + < G(x), z_0^* >\}$$

we obtain the second inequality. Since x_0 achieves the minimum we have $< G(x_0, z_0^* >= 0$, $x_0 \in \Omega, G(x_0 <= 0$

$$L(x_0, z^*) - L(x_0, z_0^*) = f(x_0) + < G(x_0), z^* > -f(x_0) =< G(x_0, z^* > \leq 0$$

□

Convexity also ensures sufficient conditions for optimality. The details are given in [10] chapter 8.5. The main result is as follows

Theorem 8.43. *Assume(A) holds and additionally that* P *is closed. Suppose there exists* $z_0^* \in Z^*$, $z_0^* \geq 0$ *and an* $x_0 \in \Omega$ *such that the Lagrangian* $L(x, z^*) = f(x) + < G(x), z^* >$ *possess a saddle point at* x_0, z_0^* *among* $x \in \Omega, z^* \in Z^*$. *Then,* x_0 *solves the minimization problem* $\min_{x \in \Omega} f(x)$ *subject to* $G(x) \leq 0$.

8.4.1 Sensitivity in the convex case

The Lagrange function is also used in order to study sensitivity and duality relations of the optimization problem. The interpretation is as follows. Consider the minimization problem

$$\min f(x) \text{subject to } G(x) \leq z_0$$

and assume that due to the previous theorem z_0^* is the supporting hyperplane at z_0. Since w is convex this hyperplane also serves as a lower bound for w.

Theorem 8.44. *Let f and G be convex and suppose $x_i, i = 0, 1$ is a solution to the problem*

$$\min f(x), x \in \Omega, G(x) \leq z_i.$$

Suppose z_i^* are the corresponding Lagrange multipliers.

Then,

$$< z_1 - z_0, z_1^* > \leq f(x_0) - f(x_1) \leq < z_1 - z_0, z_0^* > .$$

Proof. The Lagrange multipliers z_0^* makes

$$f(x_0) + < G(x_0) - z_0, z_0^* > \leq f(x) + < G(x) - z_0, z_0^* >$$

for all $x \in \Omega$. Since x_0 attains the minimum we have $< G(x) - z_0, z_0^* > = 0$ and therefore the first inequality follows, if $x = x_1$. Since then $< G(x_1), z^* > \leq < z_1, z^* >$. □

Hence, the right–hand side of the equation yields

$$f(x) - f(x_0) \geq < z_0 - z, z_0^* >$$

where x solves the minimization problem $\min f(x) : G(x) \leq z$. Hence, $f(x) = \inf\{f : x \in \Omega, G(x) \leq z\}$ and therefore the statement can be rewritten as

$$w(z) - w(z_0) \geq < z_0 - z, z_0^* > = - < z - z_0, z_0^* > .$$

If w is differentiable at z_0 we obtain

$$w'(z) = -z_0^*$$

and therefore, *the Lagrange multiplier is the negative of the first–order sensitivity of the optimal objective with respect to the constraints.* Therefore, in economics the Lagrange multipliers are called hidden or shadow costs. If w

is differentiable, then we obtain from the previous inequality for $z = z_0 + \epsilon h$ that $|w(z) - w(z_0)| \leq \epsilon < h, z_0^* >$. This justifies the gradient of w'. Note that in many textbooks one can find the admissibility condition $G(x) \geq 0$. In this case the Lagrangian is given by $L = f(x) - < G(x), z >$ and similarly for w.

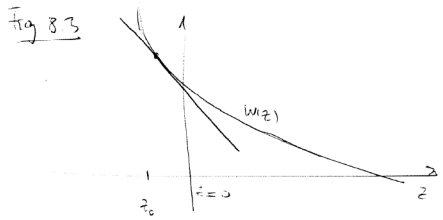


Figure 18: Sensitivitiy

8.4.2 Duality in the convex case

Consider again the convex problem min $f(x)$ subject to $x \in \Omega$ and $G(x) \leq 0$. Again, we consider the plot of the convex function

$$w(z) := \inf\{f(x) : G(x) \leq 0, x \in \Omega\}.$$

We consider the intersection at $z = 0$ of hyerplanes attached to different points z_i. The observation is as follows. The point of intersection μ_0 lies above(!) all other points of intersection. This hyperplane of course is given by the Lagrange multiplier. Next, we express this fact in terms of duality principle in terms of the multiplier. We define the dual functional

$$\phi(z^*) = \inf_{x \in \Omega} \{f(x) + < G(x), z^* >\}$$

on the positive cone P in Z^*. The function ϕ may well be unbounded. However, there is an interpretation of this function in terms of w. Recall, that $\Gamma = \{z \in Z : \exists x \in \Omega : G(x) \leq z\}$.

Lemma 8.45. *The dual function ϕ is concave and can be expressed as*

$$\phi(z^*) = \inf_{z \in \Gamma} \{w(z) + < z, z^* >\}.$$

Proof. Let $\eta, \psi \in Z^*$ and $\alpha \in (0,1)$ $\phi(\alpha\eta + (1-\alpha)\psi) = \inf_{z\in\Gamma}\{\alpha w + \alpha < z, \eta > + (1-\alpha)w + (1-\alpha) < z, \psi >\} \geq \inf_{z\in\Gamma}\{\alpha w + \alpha < z, \eta >$

$\} + \inf\{(1-\alpha)w + (1-\alpha) < z, \psi >\}$ and hence ψ is concave. For any $z^* \geq 0$ and any $z \in \Gamma$ we estimate $\phi(z^*) = \inf_{x\in\Omega}(f(x)+ < G(x), z^* >) \leq \inf_{x\in\Omega, G(x)\leq z}(f(x)+ < z, z^* >) = w(z)+ < z, z^* >$. On the other hand for any fixed $x_1 \in \Omega$ with $z_1 = G(x_1)$ we have $f(x_1)+ < G(x_1), z^* > \geq \inf\{f(x)+ < z_1, z^* >: G(x) \leq z_1, x \in \Omega\} = w(z_1)+ < z_1, z^* >$. Since this holds for any x_1 such that there exists a z_1 with $G(x_1) \leq z_1$ we take the infimum on both sides and obtain $\phi(z^*) \geq \inf_{z\in\Gamma}(w(z)+ < z, z^* >)$. □

We offer the following additional interpretation of this result. The element $(1, z^*) \in \mathbb{R} \times Z^*$ determines a family of hyperplanes consisting of points (r, z) such that

$$r+ < z, z^* >= k$$

for some $k \in \mathbb{R}$ constant. If we consider the hyperplane with $k = \phi(z^*)$ the previous Lemma tells that this hyperplanes supports the set (w, Γ) (the region above w). An applied result of this fact is the Lagrange duality.

Theorem 8.46. *Assume (A). Suppose there exists an x_1 suc that $G(x_1) < 0$ and that $\mu_0 = \inf\{f : G(x) \leq 0, x \in \Omega\}$ is finite. Then,*

$$\inf_{G(x)\leq 0, x\in\Omega} f(x) = \max_{z^*\geq 0} \phi(z^*).$$

The maximum is achieved for some z^*.

Since w is decreasing an equivalent reformulation using $w(0) \leq w(z)$ for all $z \leq 0$ is

$$\min_{z\leq 0} w(z) = \max_{z^*\geq 0} \phi(z^*).$$

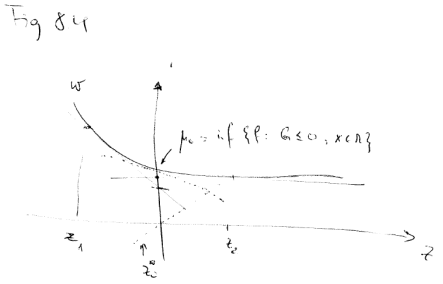


Figure 19: Duality

8.5 Descent directions for cost functionals and PDE constrained problems

In the case of PDE constrained problems the previous discussions allow for a numerical algorithm for efficient computation of the optimal solution. This algorithm can be stated in infinite–dimensional setting and is applicable in the following situation.

Let U, Y be Banach spaces. Let W' be the dual of a reflexive Banach space W. Let the objective functions be

$$J : Y \times U \to \mathbb{R} \tag{8.100}$$

and the state operator be

$$E : Y \times U \to W' \tag{8.101}$$

The controls are $u \in U$, the states are $y \in Y$. We want to solve the following problem[3]

$$\min_{y \in Y, u \in U} J(y, u) \text{ subject to } E(y, u) = 0 \tag{8.102}$$

The crucial assertion for the following derivation to be true is that *For all $u \in U$ the state equation $E(y, u) = 0$ posses a uniqe solution $y = y(u)$. .*

Example 8.47. *The classical example is of course*

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

and

$$-\Delta y = u \text{ in } \Omega, \; y = 0 \quad \text{on } \partial\Omega$$

with $U = L^2$ and $Y = H^1$ and $W = (H^1)^* = H^1$.

8.5.1 Definitions and Notation

From the previous chapters we recall that the following definitions

Definition 8.48 (Dual space)**.** *Let Y be a Banach space. Then the dual space is defined as*

$$Y' = L(Y; \mathbb{R}) \tag{8.103}$$

[3] $E = 0$ in W' means $E[w] = 0 \forall w \in W$.

For notation: the evaluation of a functional $y' \in Y'$ at a point $y \in Y$ is denoted by

$$y'[y] \in \mathbb{R} \tag{8.104}$$

For equalities $y_1' = y_2'$ means $y_1'[y] = y_2'[y] \forall y \in Y$.

Definition 8.49 (Adjoint operator)**.** *Let X, Y be Banach spaces and $T : X \to Y$ be an operator. Then the adjoint operator is defined by*

$$T' : Y' \to X', \quad y' \to (T'y')[x] := y'[Tx] \forall x \in X \tag{8.105}$$

Example 8.50. *We obmit the parenthese for T', i.e. $T'y' = T'(y')$. Derivatives are* **not** *be denoted by $'$.*

Definition 8.51 (Frechet derivative)**.** *Let X, Y be Banach spaces. Let $T : X \to Y$ then the operator $\forall x \in X : \frac{dT(x)}{dx} \in L(X; Y)$ is the Frechet derivative, iff $\forall x, x_0 \in X$*

$$T(x) = T(x_0) + \frac{dT(x_0)}{dx}(x - x_0) + o(\|x - x_0\|^2) \tag{8.106}$$

Example 8.52. *In the case $T : X \to \mathbb{R}$ we have $\frac{dT(x)}{dx} \in L(X; \mathbb{R}) = X'$. Thus by notation $\frac{dT(x_0)}{dx}(x) = \frac{dT(x_0)}{dx}[x]$.*

Theorem 8.53 (Chain rule in speacial case)**.** *Let U, Y be Banach spaces. Let $y : U \to Y$ and $\tilde{L} : Y \to \mathbb{R}$ be differentiable operators. Define the operator $L = L \circ y$ by*

$$u \to L(y(u)) : U \to \mathbb{R} \tag{8.107}$$

Then the derivative $\frac{dL(y(u))}{du} \in U'$ is given by $\forall \tilde{u} \in U, u \in U$

$$\frac{dL(y(u))}{du}[\tilde{u}] = \frac{dy(u)}{du}' \Big[\frac{dL(y(u))}{dy}[\tilde{u}]\Big] \tag{8.108}$$

or in short notation

$$L_u(y(u)) = y_u'(u) L_y(y(u)) \tag{8.109}$$

The derivative of y is given by $\frac{dy(u)}{du} \in L(U; \mathbb{R}) = U'$, i.e. let denote $\tilde{y} := \frac{dy(u)}{du} \in L(U; Y)$ for arbitrary, fixed $u \in U$. We have $\frac{dL(y(u))}{du} \in L(U; \mathbb{R}) = U'$ and by chain rule $\frac{dL(y(u))}{du}[\tilde{u}] = \frac{dL(y(u))}{dy}[\tilde{y}(\tilde{u})]$. Thus the structure is $A[B(\tilde{u})]$,

where $A \in Y', B \in L(U;Y), \tilde{u} \in U$. Since U, Y are Banach spaces we can define the operator $B' : Y' \to U'$ and have $(B'y')[u] = y'[Bu]$. With $y' = A, u = \tilde{u}$ we get $(B'A)[u] = A[B(\tilde{u})]$. Inserting all definitions we derive the above.

Hence, we can compute

Theorem 8.54 (Derivative of a duality product)**.** *Let Y be a Banach space. Let W be a reflexive Banach space, i.e. $W'' = W$. Let $E : Y \to W'$ be a differentiable operator. Then the following holds $\forall \tilde{y} \in Y$ and fixed $w \in W$*

$$\frac{d\{E(y)[w]\}}{dy}(\tilde{y}) = \Big(\frac{dE(y)}{dy}' w\Big)[\tilde{y}] \tag{8.110}$$

We note $\frac{dE(y)}{dy} \in L(Y;W')$ for fixed y and hence $\frac{dE(y)}{dy}' : W'' = W \to Y'$. Since W is reflexiv and with $B := \frac{dE(y)}{dy}$ we have for all fixed $w = w'' \in W, W''$ and for all $\tilde{y} \in Y$ $(B'w'')[\tilde{y}] = w''(B\tilde{y}) = (B\tilde{y})[w]$. By appling the definition of the derivative to the operator $y \to E(y)[w]$ and using the above calculations, we obtain $\frac{d\{E(y)[w]\}}{dy}(\tilde{y}) = \Big(\frac{dE(y)}{dy}(\tilde{y})\Big)[w] = (\frac{dE(y)}{dy}' w)[\tilde{y}]$.

8.5.2 Algorithm for computing the descent direction

Let U, Y be Banach spaces. Let W' be the dual of a reflexive Banach space W. Let the objective functions be

$$J : Y \times U \to \mathbb{R} \tag{8.111}$$

and the state operator be

$$E : Y \times U \to W' \tag{8.112}$$

The controls are $u \in U$, the states are $y \in Y$.

We assume, that forall $u \in U$ the state equation $E(y, u) = 0$ posses a uniqe solution $y = y(u)$. Then the above control problem is equivalent to the reduced control problem, namely

$$\min_{u \in U} j(u) = J(y(u), u) \tag{8.113}$$

$j(u)$ is the reduced objective function.

We assume J, E Frechet differentiable and $u \to y(u)$ is Frechet differentiable.

We introduce a Lagrange multiplies $w \in W$ for the state equation and define the Lagrange function $L : Y \times U \times W \to \mathbb{R}$ by

$$L(y, u, w) = J(y, u) + E(y, u)[w] \tag{8.114}$$

We obtain

$$L(y(u), u, w) = J(y(u), u) = j(u) \quad \forall u \in U, w \in W \tag{8.115}$$

since $E(y(u), u) = 0 \quad \forall u \in U$.

Hence we get $\forall u \in U, w \in W$

$$\frac{dj(u)}{du}[\tilde{u}] = \left(\frac{dy(u)}{du}' \frac{dL(y(u), u, w)}{dy} + \frac{dL(y(u), u, w)}{du} \right)[\tilde{u}] \forall \tilde{u} \in U \tag{8.116}$$

The idea is now to choose $w \in W$, s.t.

$$\frac{dL(y(u), u, w)}{dy}[\tilde{y}] = 0 \quad \forall \tilde{y} \in Y \tag{8.117}$$

This is the **adjoint equation** and w its solution $w = w(u)$ is the **adjoint state**. Expressing L_y in terms of J and E we obtain

$$0 = \frac{dL(y(u), u, w)}{dy}[\tilde{y}] = \left(\frac{dJ(y(u), u)}{dy} + \frac{dE(y(u), u)}{dy}' w \right)[\tilde{y}] \quad \forall \tilde{y} \in Y \tag{8.118}$$

If we assume, that E_y is continuously invertible the adjoint state $w = w(u)$ is uniquely determined.

Inserting the adjoint state $w(u)$ in j_u we obtain

$$\begin{bmatrix} \frac{dj(u)}{du}[\tilde{u}] & = & \left(\frac{dy(u)}{du}' \frac{dL(y(u),u,w(u))}{dy} + \frac{dL(y(u),u,w(u))}{du} \right)[\tilde{u}] \\ & = & \frac{dL(y(u),u,w(u))}{du}[\tilde{u}] \\ & = & \left(\frac{dJ(y(u),u)}{du} + \frac{dE(y(u),u)}{du}' w \right)[\tilde{u}] \\ \forall \tilde{u} \in U & & \end{bmatrix} \tag{8.119}$$

Algorithm for computing the gradient $\frac{dj(u)}{du}$ for given $u \in U$.

1. Compute $y = y(u) \in Y$ by solving the state equation

$$E(y, u) = 0 \tag{8.120}$$

2. Compute the adjoint state $w = w(u) \in W$ by solving the adjoint equation

$$\frac{d}{dy}E(y,u)'w = -\frac{d}{dy}J(y,u) \tag{8.121}$$

3. Compute $\frac{d}{du}j(u)$ by the equation

$$\frac{d}{du}j(u) = \frac{d}{du}J(y,u) + \frac{d}{du}E(y,u)'w \tag{8.122}$$

For the given control problem we obtain the following steps:

1. Solve $-\Delta y = u,\ y = 0$ for y given u.
2. Solve $-\Delta p = (y - y_d),\ p = 0$ for p given y
3. Obtain the gradient of the reduced cost functional as $u + \lambda p$.

9 Existence of Minimizers in infinite space dimensions

9.1 General case

We prove the existence of minimizers in infinite space dimensions. It turns out that this can be achieved by a variational approach. This introduction can also be found in [14, 7]. In the following sections we will then assume that there exists at least a local minimizer and focus on constraint qualifications. Another approach focus on montone operators that extend the well-known montone functions in $\mathbb{R}$.

At first we present a general theory before latter treating at first unconstrained problems by means of variational theory and second constrained problems.

We consider the following abstract formulation

$$\min J(u) \text{ subject to } u \in C \tag{9.1}$$

where $J : C \subset U \to \mathbb{R}$ and wherein U is a Banach or Hilbert space.

We may extend J to U by setting $J(U) = +\infty$ for all $u \notin C$. We distinguish between local and global minimizer.

Definition 9.1 (Minimzers)**.** *A point $u \in U$ is called a local minimizer, if there exists an open set $V \subset U$ such that $u \in U$ and $J(u) \leq J(v) \forall v \in V$ and a global minimizer, if $J(u) \leq J(v) \forall v \in U$.*

As we shall see later, local minimizers of smooth functionals can be characterized by conditions on the first and second derivative. For global optima, there is in general no other condition than the above. Of course, each global minimizer is also a local one, but in general not vice versa.

In order to obtain existence of solutions for a general optimization problem, two basic properties are needed: compactness and lower semicontinuity. We recall the definition of the latter.

Definition 9.2. *Let (U,T) be a topological space and let $J : U \to \mathbb{R}$ be a given functional. Then J is lower semicontinuous at $u \in U$ if*

$$J(u) \leq \sup_{V \in T} \inf_{v \in V} J(v) \tag{9.2}$$

If U is a metric space this definition is equivalent to a charaterization by sequences. The functional J is lower semicontinuous at $u \in U$ if

$$J(u) \leq \liminf J(u_k) \tag{9.3}$$

for all sequences u_k converging to u.

To prove existence of a minimizer lower semicontinuity and compactness is sufficient. Note that compactness in infinite dimensional spaces is a stronger assumption than in finite dimensional spaces. In particular, the conclusion "bounded and closed implies compact" does not hold in infinite dimensional spaces. We repeat the definition of compactness.

Definition 9.3 (Compactness)**.** *A topological space (X,T) is compact, iff every open cover of X has a finite cover.*

If X is a metric space and topological compact, then there exists a convergent subsequence. In a metric space "sequentially compact" and "topological compact" are equivalent.

Theorem 9.4 (Existence)**.** *Let $J : U \to \mathbb{R}$ be lower semicontinuous and let the level set*

$$\{u \in U : J(u) \leq M\} \tag{9.4}$$

be non-empty and compact for some $M \in \mathbb{R}$. Then there exists a global minimum of of the problem

$$\min_{u \in U} J(u) \tag{9.5}$$

Proof. Let $\alpha := \inf J(u)$. Then there exists a sequence u_k such that $J(u_k) \to \alpha$. For k sufficiently large we have $J(u_k) \leq M$ and hence u_k is contained in a compact set. Then there exists a subsequence also denoted by u_k such that $u_k \to u^* \in U$. Note that $J(u^*) > -\infty$ since $J : C \subset U \to (-\infty, \infty)$ by definition, i.e., evaluations of J for elements $u \in C$ yields finite values. Due to the lower semicontinuity we obtain

$$\alpha \leq J(u^*) \leq \liminf J(u_k) = \alpha \tag{9.6}$$

Therefore, u^* is a local minimizer. □Usually, the compactness of the above set is not guranteed. One can at most hope, to obtain a bounded set. The following theorem yields a weak convergent subsequence for bounded sets, for proofs see [2, 16].

Compactness Theorem

Theorem 9.5. *Let U be a Hilbert space and let u_k be a bounded sequence in U. Then there exists a weak convergent subsequence (u_{k_l}), ie.,*

$$< v, u_{k_l} > \to < v, u^* > \ \forall v \in U \tag{9.7}$$

for some $u^* \in U$.

Exercise 9.6.

1. *Show that every strongly convergent sequence is weakly convergent.*
2. *Prove that $\|A\|_L(X,Y) = \sup_{\|x\|_X=1} \|Au\|_Y$ defines a norm on the space of all linear operators on Banach spaces X, Y.*
3. *Given a Hilbert space X. Prove that $x_n \to x$ weakly and $y_n \to y$ strong implies $(x_n, y_n) \to (x, y)$.*

Note that this theorem is a special case of the theorem of Eberlein-Smulyan.

Eberlein-Smulyan

Theorem 9.7 (Eberlein-Smulyan)**.** *A space X is reflexiv (ie. there exists a linear, isometric and continuous mapping I such that $I(X) = X''$ where X'' is the bi-dual), iff the closed unit ball in X is weakly sequentially compact.*

There is a similar result for the dual space X'.

Theorem 9.8. *Let X be separabel (ie. there exists a dense, countable subset in X), then the closed unit ball in X' is weakly-$*$ sequentially compact.*

We finally prove the weak version of the previous theorem. Note that we need to require J to be weak lower semicontinuous, ie.,

$$u_k \to u \text{ weakly } \implies J(u) \leq \liminf J(u_k) \tag{9.8}$$

We will later see, under which additional assumptions on J this condition holds.

Existence theorem

Theorem 9.9 (Existence)**.** *Let U be a Hilbert space. Let $J : U \to \mathbb{R}$ be weak lower semicontinuous (wlsc) and let the level set*

$$\{u \in U : J(u) \leq M\} \tag{9.9}$$

be non-empty and bounded for some $M \in \mathbb{R}$. Then there exists a global minimum of of the problem

$$\min_{u \in U} J(u) \tag{9.10}$$

Proof. Let $\alpha := \inf J(u)$. Then there exists a sequence u_k such that $J(u_k) \to \alpha$. For k sufficiently large we have $J(u_k) \leq M$ and hence u_k is bounded. Hence there exists a weakly convergent subsequences denoted by u_k such that $u_k \to u^* \in U$ weakly. Since J is wlsc we have

$$\alpha \leq J(u^*) \leq \liminf J(u_k) = \alpha \tag{9.11}$$

Therefore, u^* is a local minimizer. □

The condition $\{u \in U : J(u) \leq M\}$ is a bounded set can be enforced by coercivity.

coercivity theorem

Theorem 9.10 (Coercivity)**.** *Let U be a Hilbert space and $J : U \to \mathbb{R}$ satisfy*

$$\frac{J(u)}{\|u\|} \to \infty \text{ as } \|u\| \to \infty \tag{9.12}$$

Then the set

$$\{u \in U : J(u) \leq M\} \tag{9.13}$$

is non-empty and bounded for M sufficiently large.

Proof. For arbitrary $u_0 \in U$ we have $J(u_0) < \infty$ and hence for $M \geq J(u_0)$ the set $\mathcal{M} := \{u \in U : J(u) \leq M\}$ is non-empty. Now, let $\mathcal{M}$ be given with a sufficiently large and arbitrary M and assume that $\mathcal{M}$ is

unbounded. Then there exists $u_k \in \mathcal{M}$ with $\|u_k\| \to \infty$ and $J(u_k) \leq M$. Dividing this inequality by $\|u_k\|$ and let $k \to \infty$, we obtain a contradiction to the coercivity assumption. □

Reviewing the existence results of the previous section, we observe that there are only two main ingredients: We need boundedness of the sequence to conclude the existence of a weakly convergent subsequence and we need weak lower semicontinuity to exchange exchange application of the functional with the limit. Typically, in applications the admissible set C is not the full Hilbert space. We can either set now $J(u) = +\infty$ for $u \notin C$ but then have to proof the coercivity or we may proceed as follows: *If C is additionally bounded, closed and convex, then we trivially obtain the boundedness of the sequence and that the limit exists in C. All the previous arguments can then be applied analogously.* This gives the following existence result.

tence for bounded sets

Theorem 9.11 (Existence on bounded, closed, convex subsets)**.** *Let U be a Hilbert space and C be a closed, convex and bounded subset of U. Let $J : U \to \mathbb{R}$ be a weakly lower semicontinuous function. Then, there exists a minimum to*

$$\min_{u \in U_{ad}} J(u).$$

Finally, one can skip the assumption on boundedness of U and replace it by the coercivity. The proof is then as before with the difference that the limit is now in U_{ad} since it is still and closed and convex.

nce for unbounded sets

Theorem 9.12 (Existence on closed, convex subsets)**.** *Let U be a Hilbert space and C be a closed and convex subset of U. Let $J : U \to \mathbb{R}$ be a weakly lower semicontinuous and coercive function. Then, there exists a minimum to*

$$\min_{u \in U_{ad}} J(u).$$

Next, we consider first necessary conditions which be extended later on. In this general setting we can prove the following Let U be a Banach space and $C \subset U$ a convex set. Let $J : C \to \mathbb{R}$ be Gateaux differentiable. Let $\bar{u} \in C$ a solution to

$$\min_{u \in C} J(u) \tag{9.14}$$

Then the variational inequality

$$J'(\bar{u})(u - \bar{u}) \geq 0 \; \forall u \in C \tag{9.15}$$

is satisfied.

Proof. Let $u \in C$ be arbitrary. Then consider $u(t) = \bar{u} + t(u - \bar{u})$ and since C is convex: $u(t) \in C$ for all $t \in [0, 1]$. Obviously

$$J(\bar{u}) \leq J(u(t)) \tag{9.16}$$

and therefore

$$\frac{1}{t}\left(J(\bar{u} + t(u - \bar{u})) - J(\bar{u})\right) \geq 0 \tag{9.17}$$

For $t \to 0$ we obtain the desired result. □

Note that C convex is a rather strict restriction on the constraints. If we for example consider an equality constrained problem, $\min J(u)$ subject to $h(u) = 0$, then h has to be linear to satisfy this restriction. In the following sections we will therefore consider more general constraint qualifications which also gives some information, if C is not convex and which are also related to the structure of h. Nevertheless, the important case of box constraints is covered by this theorem.

Also, comparing with the finite dimensional case (see below), there exists a more general result which can be found for example in [5] pp. 16.

Definition 9.13. *Let C be a closed subset of the Banach space U. For $u \in C$ we define the tangential cone $T_C(u)$ at u by*

$$T_C(u) := \{v \in U : \exists \epsilon > 0 \forall 0 \leq t \leq \epsilon \exists w(t) \in C : \|u + tv - w(t)\| = o(t)\} \tag{9.18}$$

If u belongs to the interior of C, then the tangential cone is $T_C(U) = U$. Further, we have a similar theorem to the above of necessary first order conditions if C is closed.

Theorem 9.14. *Let $J : U \to \mathbb{R}$ be continuously Frechet differentiable and let u^* be a local minimizer of $\min_{u \in C} J(u)$. If C is closed, then*

$$J'(u^*)v \geq 0 \ \forall v \in T_C(u^*) \tag{9.19}$$

Proof. Obviously, $T_C \neq \emptyset$. Let $v \in T_C(u^*)$ be given. Then for each $t \in [0, \epsilon]$ we have due to the continuity of J

$$J(w(t)) = J(u^* + vt) + o(t) = J(u^*) + tvJ'(u^*) + o(t) \tag{9.20}$$

Since $J(u^*)$ is a local minimum we have

$$0 \leq tvJ'(u^*) + o(t) \tag{9.21}$$

for t sufficiently small. This implies that $vJ'(u^*) \geq 0$, since for any fixed constant $\tilde{c} > 0 : t\tilde{c} \neq o(t)$. □

The above is the most general result for necessary first order conditions. Since we later on consider equality and inequality constraints of the type $h(x) = 0, g(x) \leq 0$, i.e., $C := \{x : h(x) = 0, g(x) \leq 0\}$ we would like to characterize the set $T_C(u^*)$ in terms of h and g. In general, this is not possible. We need to impose **further assumptions** to obtain a **relation** between the gradient at the minimum **and** the functions h and g respectively. This will also give are more precise characterisation of the minimum (i.e. KKT theorem and existence of Lagrange multipliers).

9.2 Examples of PDE constrained optimization problems

The previous theory can be applied to several problems in PDE constrained optimization. We give some examples below. A more general treatment of unconstrained problems by variational methods is given in the following subsection.

ocp pde 1

The first example is tracking–type problem for an elliptic partial differential equation. We consider the problem

$$\min \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\lambda}{2}\|u\|_{L^2}^2 \tag{9.22a}$$

$$\text{subject to } -\Delta y = \beta u \text{ in } \Omega \tag{9.22b}$$

$$y = 0 \text{ on } \partial\Omega \tag{9.22c}$$

$$u_a \leq u \leq u_b \; a.e. x \in \Omega \tag{9.22d}$$

To avoid technical difficulities will make the following assumption on the domain $\Omega \in \mathbb{R}^n$. It is assumed that Ω is a bounded, Lipschitz domain. Furthermore, we assume that $\lambda \geq 0$ and $\beta \in L^\infty$ and that the other appearing functions are sufficiently regular, e.g., $u_a \leq u_b \in L^2$ and $y_d \in L^2$. Then, we define the set of suitable control functions u to be in Hilbertspace with suitable compactness properties and choose

$$U_{ad} = \{u \in L^2 : u_a \leq u \leq u_b \text{ a.e. } x \in \Omega\}.$$

Then, U_{ad} is a non–empty, convex subset and closed subset of L^2, since it is defined pointwise a.e. in Ω.

Exercise 9.15. *Prove that U_{ad} is closed and convex.*

Due to standard elliptic existence theory we obtain that for every $u \in U_{ad}$ a unique solution $y \in H_0^1(\Omega)$. Therefore, a reasonable state space is $Y = H_0^1$

(where we skipp the dependence of Ω whenever the intention is clear). To apply the existence result of the previous section we itnroduce a mapping

$$G(u) = y : \ L^2 \to H_0^1$$

where for a given control u, the state y is defined as the solution to

$$-\Delta y = \beta u \text{ in } \Omega y = 0 \text{ on } \partial\Omega \tag{9.23}$$

We obtain that G is linear and bounded, since due to standard theory , we have

$$\|G(u)\|_{H^1} = \|y\|_{H^1} \leq c\|u\|_{L^2}.$$

We furtherforemore introduce the operator

$$S = I_{L^2,H^1}G : \ L^2 \to L^2$$

where I is the imbedding from $H_0^1 \to L^2$ and I is linear and bounded. Introducing S has the advantage that the adjoint operator S^* is an operator from $L^2 \to L^2$. Introducing, S we can reformulate the optimization problem as follows

$$\min_{u \in U_{ad}} \frac{1}{2}\|Su - y_d\|_{L^2}^2 + \frac{\lambda}{2}\|u\|_{L^2}^2 \tag{9.24}$$

Using this reformulation we can apply the existence results and obtain the following result.

e result hilbert space

Theorem 9.16. *Let $\{U, \|\cdot\|_U\}$ and $\{H, \|\cdot\|_H\}$ be real vector spaces and U_{ad} a non–empty, closed and convex set $U_{ad} \subset U, y_d \in H$ and $\lambda \geq 0$. Let $S : U \to H$ be a linear and continuous operator. Then, the following quadratic problem admits*

$$\min_{u \in U_{ad}} \frac{1}{2}\|Su - y_d\|_H^2 + \frac{\lambda}{2}\|u\|_U^2 \tag{9.25}$$

an optimal solution u^* and its unique if $\lambda > 0$.

Exercise 9.17. *In the setting of Theorem 9.16 prove that $f(u) = \|Su - y_d\|_H^2 + \frac{\lambda}{2}\|u\|_U^2$ is strict convex if $\lambda > 0$.*

We can either proof this directly or by applying the previous results. The direct proof is along the following lines. Since $j(u) := \frac{1}{2}\|Su - y_d\|_H^2 + \frac{\lambda}{2}\|u\|_U^2 \geq 0$ there exists the infimum of all functional values $j := \inf_{u \in U_{ad}} j(u)$. Hence, there exists a sequence $u_n \in U_{ad}$ such that for $n \to \infty$ we obtain

$$j(u_n) \to j.$$

Now, since we are in the infinite-dimensional space, the bounded and closed set U_{ad} is *not* necessarily compact. However, it is weak sequentially compact, i.e., amy bounded sequence has a weakly convergent subsequence, see Theorem 9.5. Since U_{ad} is also convex, the limit u^* is in U_{ad}, see Theorem 9.7. Since S is a continuous operator, so is j. However, this is *not* sufficient to conclude $j(u_n) = j(u)$. However, j is a composition of norms and therefore weak lower semicontinuous (alternatively: j continous and convex implies weakly lower semicontinuity) . Hence, we obain for a subsequence

$$j(u^*) \leq \liminf j(u_n) = j.$$

Since j is the infimum on all values we obtain that u^* is the minimizer of j. Since j is convex as soon as $\lambda > 0$ we obtain uniqueness.

Proof. Consider the case $\lambda > 0$. The function $j(u) := \frac{1}{2}\|Su - y_d\|_H^2 + \frac{\lambda}{2}\|u\|_U^2$ is weak lower semicontinuous. For $u \notin U_{ad}$ we set $j(u) = +\infty$. For $\lambda > 0$ the function j satifsfies

$$j(u)/\|u\| \to \infty$$

as $\|u\| \to \infty$ and hence the set $\{u \in U : j(u) \leq M\}$ is bounded form some due to Theorem 9.10. Furthermore, it is non empty, since $U_{ad} \subset U$. Then, due Theorem 9.9 there exists a minimum $u \in U$. Since for $u \notin U_{ad}$ we have $j(u) = +\infty$, we obtain that $u \in U_{ad}$. Uniqueness is due to the fact that j is convex if $\lambda > 0$.

Alternatively, we could apply Theorem (9.11) to deduce the existence. □

This allows now for the following result on the PDE constrained problem. Let $U = H = L^2$. Then, $U_{ad} = \{u \in L^2 : u_a \leq u \leq u_b\}$ is a closed, convex and bounded subset of U. Hence, by applying Theorem 9.16 we obtain

Lemma 9.18. *The problem (9.22) admits an optimal solution $u^* \in U_{ad}$ under the given assumptions on Ω, β, u_a and u_b. The solution is unique if $\lambda > 0$.*

and by applying theorem (9.12) we have

Lemma 9.19. *The problem (9.22) admits an unique optimal solution $u^* \in U_{ad}$ under the given assumptions on Ω, β and for $u_a = -\infty$ and $u_b = +\infty$ and if $\lambda > 0$.*

The previous results obviously extend to further PDE constrained problems. Consider for example Robin boundary data instead of the Dirchlet

ocp pde 2

data:

$$\min \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\lambda}{2}\|u\|_{L^2}^2 \tag{9.26a}$$

$$\text{subject to } -\Delta y = \beta u \text{ in } \Omega \tag{9.26b}$$

$$\partial_n y = \alpha(y_a - y) \text{ on } \partial\Omega \tag{9.26c}$$

$$u_a \leq u \leq u_b \; a.e.x \in \Omega \tag{9.26d}$$

We assume here that $y_a \in L^2(\partial\Omega)$ and $\alpha \in L^\infty(\partial\Omega)$ and $\int_{\partial\Omega} \alpha ds > 0$. Standard PDE theory guarantees the existence and uniqueness of $y \in H^1$ for any fixed $u \in L^2$ and $y_a \in L^2$. We denote by y_u the solution to $(u, y_a \equiv 0)$ and by y_0 the solution to $(u \equiv 0, y_a)$. Then, any solution y can be written as $y = y_u + y_0$. The solution operator $G : u \to y_u$ is linear and continuous as a mapping from L^2 to H^1 and we again apply the embedding from H^1 to L^2 to define the operator $S = Id_{L^2,H^1}G$. The full solution is hence written as

$$Su + y_0 : L^2 \to L^2$$

and the problem can be reformulated as

$$\min_{u \in U_{ad}} \frac{1}{2}\|Su - y_d + y_0\|_{L^2}^2 + \frac{\lambda}{2}\|u\|_{L^2}^2. \tag{9.27}$$

ocp pde 3

Obviously, the previously established results also cover this case and we obtain existence and for $\lambda > 0$ of an optimal control u^*.

Next, consider for example Robin boundary data with boundary control

ocp pde 2

$$\min \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\lambda}{2}\|u\|_{L^2(\partial\Omega)}^2 \tag{9.28a}$$

$$\text{subject to } -\Delta y = 0 \text{ in } \Omega \tag{9.28b}$$

$$\partial_n y = \alpha(u - y) \text{ on } \partial\Omega \tag{9.28c}$$

$$u_a \leq u \leq u_b \; a.e.x \in \partial\Omega \tag{9.28d}$$

We assume here that $\alpha \in L^\infty(\partial\Omega)$ and $\int_{\partial\Omega} \alpha ds > 0$ to guarantee uniqueness of the solution. Standard PDE theory guarantees the existence and uniqueness of $y \in H^1$ for any fixed $u \in L^2(\partial\Omega)$. Hence, the solution operator $G : u \to y : L^2(\partial\Omega) \to H^1$ is linear and continuous as a mapping from L^2 to H^1 and we again apply the embedding from H^1 to L^2 to define the operator $S = Id_{L^2,H^1}G$. The full solution is hence written as

$$Su : L^2(\partial\Omega) \to L^2(\Omega)$$

and the problem can be reformulated as

$$\min_{u\in U_{ad}} \frac{1}{2}\|Su - y_d\|^2_{L^2} + \frac{\lambda}{2}\|u\|^2_{L^2(\partial\Omega)}. \tag{9.29}$$

ocp pde 3

Obviously, the previously established results also cover this case and we obtain existence and for $\lambda > 0$ of an optimal control u^*.

Exercise 9.20. *1. Given a smooth bounded domain $\Omega \subset \mathbb{R}^n$, $b, y_d \in L^2(\Omega)$ and $d \in L^2(\partial\Omega)$. We assume that d is a boundary value of a function $y \in H^1(\Omega)$. Derive the necessary optimality conditions for the problem*

$$\min \int_\Omega (y - y_d)^2 dx$$

subject to

$$-\Delta y = u + b$$

and

$$y = d$$

on $\partial\Omega$ with the box constraints $-1 \le u \le 1$ a.e. in x.

2. Derive formally the necessary conditions for the problem

$$\min \int_\Omega (y - y_d)^2 + \lambda u^2 dx + \int_{\partial\Omega} dydSx$$

subject to

$$-\Delta y + y = u + b$$

and

$$y = d$$

on $\partial\Omega$ with the box constraints $-1 \le u \le 1$ a.e. in x.

9.3 Variational problems or unconstrained energy minimization

Applications of the above theory are given in the book by Evans [7]. In particular we are interested in minimization problems related to partial differential equations. The problems are unconstrained in the sense that we minimize a cost functional depending on ∇u and/or u without further PDE constraints. The functional we investigate is given in Definition 9.21.

2:def **Definition 9.21.** *Let $U \subset \mathbb{R}^n$ open, bounded domain with smooth boundary. Let $L : \mathbb{R}^n \times \mathbb{R} \times \overline{U} \to \mathbb{R}, L = L(p, z, x)$ be a given function. Then we define the functional*

$$I(u) = \int_U L(\nabla u(x), u(x), x)dx,$$

where $u : \overline{U} \to \mathbb{R}$.

In the next few lines, we see that there is a pde problem behind such minimization problems. Sei u ein Minimierer von I und $v \in C_0^\infty(U)$ eine beliebige Testfunktion. Weiterhin erfüllt $u = g$ auf ∂U. Die Euler-Lagrange Gleichung ergibt sich dann aus der Bemerkung, daß die Funktion

$$i(\tau) := I(u + \tau v), \tau \in \mathbb{R}$$

ein Minimum für $\tau = 0$ hat, d.h. es gilt

$$i'(0) = 0.$$

Differentiation (L, u hinreichend glatt) der Funktion $i(\tau)$ nach τ, Auswertung im Punkt $\tau = 0$ und partielle Integration ergibt dann die zugehörige PDE

$$0 = \int_U \Big[-\sum_{i=1}^n (L_{p_i}(Du, u, x))_{x_i} + L_z(Du, u, x) \Big] v dx$$

und damit (obige Gleichung gilt $\forall v$) die klassische nichtlineare PDE zweiter Ordnung

$$\begin{aligned} 0 &= -\sum_{i=1}^n (L_{p_i}(Du, u, x))_{x_i} + L_z(Du, u, x), \quad \forall x \in U \\ u &= g, \quad \forall x \in \partial U \end{aligned}$$

9.3.1 Beispiele

1. $L(p, z, x) = 1/2|p|^2$ führt zu $\Delta u = 0$.
2. $L(p, z, x) = 1/2 \sum\limits_{i,j=1}^n a^{ij}(x)p_i p_j - zf(x)$ führt zu $-\sum\limits_{i,j=1}^n (a^{ij}(x)u_{x_i})_{x_j} = f$.
3. $L(p, z, x) = \sqrt{(1 + |p|^2)}$ führt zu $div(\frac{Du}{(1+|Du|^2)^{1/2}}) = 0$, der sogenannten Minimalflächen Gleichung.

Bisher war $u : U \to \mathbb{R}$ gesucht. Die Funktionale lassen sich jedoch leicht auf Systeme erweitern , d.h. gesucht werden Funktionen $u : U\mathbb{R}^m$. Damit ergibt sich ∇u zu einer $m \times n$ Matrix und die Funktion L hat den Definitionsbereich

$$L : \mathbb{R}^{m\times n} \times \mathbb{R}^m \times U \to \mathbb{R}$$

Die gleiche Idee wie oben ergibt dann ein System von m Differentialgleichungen in U, welches gekoppelt und nichtlinear ist. Die Struktur ist wie folgt

$$-\sum_{i=1}^{n}\left(L_{p_i^k}(Du,u,x)\right)_{x_i} + L_{z^k}(Du,u,x) = 0, \quad k = 1,\dots,m$$

Dabei bezeichnet $L_{p_i^k}$ die Ableitung nach der Komponente (k,i) der Funktion L.

Bestimmte Lagrange Funktionen sind interessant zu studieren. Die sogenannten Null-Lagrange Funktionen.

Definition 9.22. *Die Funktion L heißt Null-Lagrange Funktion, wenn das System der Euler-Lagrange Gleichungen für alle glatten Funktionen gelöst werden.*

Daraus ergibt sich der folgende Satz.

Theorem 9.23. *(Null-Lagrange und Funktionale)*
Sei L eine Null-Lagrange Funktion und seien $u, \hat{u}$ zwei $C^2(\overline{U};\mathbb{R}^m)$ Funktionen mit $u \equiv \hat{u}$ auf ∂U. Dann gilt $I(u) = I(\hat{u})$.

Proof. Betrachten der Ableitung der Funktion $i(\tau) = I(\tau u + (1-\tau)\hat{u})$ ergibt $i'(\tau) = 0$.

Theorem 9.24. *(Beispiel einer Null-Lagrange Funktion)*
Die Determinantenfunktionen $L(P) = detP$ ist ein Null-Lagrange.

Proof. Der Beweis benutzt das Lemma $\sum_{i=1}^{n}(cofDu)^k_{i,x_i} = 0, \quad k = 1,\dots,n.$[4]

[4]Evans, siehe Seite 440ff

9.3.2 Coercivity

To apply existence and uniqueness theorems of the previous discussion, we introduce some growth criterias for L.

Für das Funktional $I(u)$ der Lagrange Funktion L, definiert man ein starkes Wachstum des Funktionales für große Argumente u durch folgende Koerzivitätsbedingung

Definition 9.25. *Sei $1 < q < \infty$ und gelte weiterhin*

$$\exists \alpha > 0, \beta \geq 0 : L(p, z, x) \geq \alpha |p|^q - \beta \quad \forall p \in \mathbb{R}^n, z \in \mathbb{R}, x \in U$$

Dann erfüllt I die Koerzivitätsbedingung.

Example 9.26. *Es gilt*

$$|Dw| \to \infty \implies I(w) \geq \alpha |Dw|^q - \beta |U| \to \infty,$$

d.h. I koerziv impliziert, daß

$$\inf I(w) \geq -\beta |U| > -\infty$$

gilt.

Die letzte Ungleichung kann auch noch sinnvoll definiert werden, wenn die Funktion u nicht glatt ist, sondern lediglich aus $W^{1,p}$. Somit definiert man für die Kanditaten der Minimierung die Menge $\mathcal{A}$ durch

$$\mathcal{A} = \{w \in W^{1,p}(U) : w = g \quad \forall x \in \partial U\}$$

$w = g$ auf ∂U ist zu verstehen im Spursinn, d.h. die Abbildung $T : W^{1,p}(U) \to L^p(\partial U)$ soll stetig sein. Dies ist erfüllt, wenn $\partial U \in C^1$ gilt. Zum Beweis dieser Aussage zeige man dies zuerst für glatte Funktionen und approximiere dann Sobolevfunktionen durch glatte Funktionen[5]

9.3.3 Weak lower semicontinuity

Besides some coercivity we also needed weak lower semicontinuity. We introduce the notation again and focus on the Sobolev spaces necessary in the problem description.

Bei dem Existenzbeweis wird eine Minimalfolge in $\mathcal{A}$ gewählt, von der gezeigt werden kann, daß sie in $W^{1,p}$ beschränkt ist. Somit hat diese eine schwach konvergente Teilfolge.

[5] Evans

Nun möchte man gerne, daß gilt $u_m \rightharpoonup u$ impliziert $I(u_m) \to I(u)$, um auf ein Minimum zu schliessen.

An Voraussetzungen hat man jedoch nur $u_m \rightharpoonup u, L^p$ und $Du_n \rightharpoonup Du, L^p$. Durch die Sobolevschen Einbettungssätze $W^{1,p}(U) \subset\subset L^q(U)$ kann die Konvergenz $u_m \rightharpoonup u$ noch verstärkt werden, doch über die Gradienten ergibt sich keine Aussage. Die obige Implikation ist aber zu stark zu fordern, es reicht auch den Begriff der Unterhalbstetigkeit einzuführen.

Definition 9.27. *I heißt schwach folgenunterhalbstetig auf $W^{1,p}(U)$, wenn für alle Folgen u_m mit*

$$u_m \rightharpoonup u, W^{1,p}$$

gilt

$$I(u) \leq \liminf I(u_m)$$

9.3.4 Convexity

We obtain that convexity can gurantee the weak lower semicontinuity. This is an important fact, since convexity is much easier to check, then wlsc.

Für hinreichend glatte Funktionale I erfüllt L eine Konvexitätsbedingung. Im folgenden Satz ergibt sich ein Zusammenhang zwischen der Unterhalbstetigkeit.

9.3.5 Konvexität

Definition 9.28. *(Konvexität)*
Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt konvex, wenn gilt

$$f(\tau x + (1-\tau)y) \leq \tau f(x) + (1-\tau)f(y) \quad \forall x, y \in \mathbb{R}^n, 0 \leq \tau \leq 1$$

Theorem 9.29.

1. *Sei f konvex, dann existiert für alle $x \in \mathbb{R}^n$ ein $r \in \mathbb{R}^n$ mit*
$$f(y) \geq f(x) + r(y-x) \quad \forall y \in \mathbb{R}^n$$
2. *Sei $f \in C^1(\mathbb{R}^n)$ und konvex, dann gilt $r = Df(x)$*
3. *Sei $f \in C^2(\mathbb{R}^n)$. Dann ist äquivalent $D^2 f(x) \geq 0 \Leftrightarrow f$ konvex.*

Theorem 9.30. *Sei U beschränkt, L beschränkt, glatt und $p \to L(p, z, x)$ konvex für alle z, x. Dann ist I schwach (folgen)unterhalbstetig in $W^{1,p}(U)$.*

Proof. [6] Es werden verschiedene Sätze aus der Lebesguetheorie verwendet. Zuerst gibt man sich eine Folge $u_k \subset W^{1,p}(U)$ vor, die schwach konvergent ist. Hieraus schließt man die Beschränktheit von Du_k und u_k in $L^p(U)$. Problem: Zu untersuchen ist der Grenzwert

$$\int_U L(Du_k, u_k, x)dx,$$

d.h. ein Grenzwert in beiden Argumenten der Funktion L.

Nun wird die Konvergenz von u_k in $L^p(U)$ verstärkt durch (1) kompakte Einbettung von $W^{1,p}$ in $L^p(U)$, (2) aus starker Konvergenz in L^p folgt Konvergenz f.ü. (bzgl. $|\cdot|$ in $\mathbb{R}$, (3) auf geeigneten Teilmengen folgt gleichmässige Konvergenz.

Die Konvergenz des Gradienten kann nicht verstärkt werden, daher benutzt man die Konvexität von L in der Variablen des Gradienten, um diesen günstiger darzustellen. Es ergibt sich das neue Problem

$$\int_U L(Du, u_k, x) + D_pL(Du, u_k, x)(Du_k - Du)dx$$

Die Konvergenzen für u_k sichern den Grenzübergang, denn der erste Term konvergiert nach dem Satz von Lebesgue (L, U beschränkt) und der zweite Term konvergiert (gegen Null), da das Produkt einer gleichmässig konvergenten Folge und einer schwach konvergenten Folge konvergiert (siehe Zettel).

Insgesamt ergibt sich dann die Behauptung. Wichtig war, daß der Gradient nur noch linear auftritt und dann die schwache Konvergenz ausgenutzt werden kann.

9.3.6 Existence and Uniqueness

Finally, we combine our findings and prove existence and uniquness. We again state the proof, although it is similar to the general discussion before.

Theorem 9.31. *(Existenz)*
Sei L koerziv und konvex in der Variablen p. Sei weiterhin $\mathcal{A}$ nicht leer. Dann gibt es mindestens eine Funktion $u \in \mathcal{A}$ mit

$$I(u) = \min_{w \in \mathcal{A}} I(w)$$

[6] Evans, siehe Seite 446ff

Proof. [7] Nach obigen Bemerkungen gilt $\inf I > -\infty, Iw.slsc.$ Das Ziel ist aus der Minimalfolge eine schwach konvergente Teilfolge auszuwählen. Hinreichend dafür ist die Beschränktheit der Minimalfolge in $W^{1,q}$. Anschließend muß noch $u \in \mathcal{A}$ gezeigt werden, d.h. das Minimum liegt auch Definitionsbereich.

Existenz der Minimalfolge

1. Sei $\inf I < \infty$, denn sonst existiert kein Minimum
2. Aus $\inf I > -\infty$ folgt $\exists u_k \subset \mathcal{A}$ mit $I(u_k) \to \inf I$.

Beschränktheit der Minimalfolge

3. Aus $I(w) \geq \alpha \|Dw\|^q - \beta$ (Koerzivität) und (1) folgt $\|Dw\|_{L^q}$ ist beschränkt für alle $w \in \mathcal{A}$
4. U beschränkt und damit liefert Poincaresche Ungleichung: $\|u\|_{L^q}$ ist beschränkt.
5. $\implies u_k \subset W^{1,q}$ beschränkt, d.h. es existiert eine schwach konvergente Teilfolge $u_k \rightharpoonup u$.

   $u \in \mathcal{A}$
6. Satz von Mazur über schwach konvergente Folgen und die Abgeschlossenheit der Menge $\mathcal{A}$ lieferen die Behauptung

   u ist Minimum
7. Da I w.slsc ist, folgt $I(u) \leq \liminf I(u_k)$ und nach Wahl der u_k gilt $I(u_k) \to \inf I$, d.h. $I(u) = \min I$. □

Um Eindeutigkeit der Lösung zu zeigen, sind weitere Voraussetzungen an L nötig.

Theorem 9.32. *(Eindeutigkeit)*

Sei $L = L(p,x)$ *und* $\exists \theta > 0 : \sum_{i,j=1}^{n} L_{p_i p_j}(p,x)\xi_i\xi_j \geq \theta|\xi|^2 \quad \forall x, p.$ *Dann ist das Minimum eindeutig.*

[7] Evans, siehe Seite 448ff

Proof. [8] Annahme es existieren zwei Minima u, w. Dann ist das Ziel, zu zeigen, daß mit $v = (u + w)/2$ gilt $I(v) < (I(u) + I(w))/2$, was ein Widerspruch ist.

Um diese Ungleichung zu beweisen, entwickelt man L nach p um q in eine Taylroreihe bis zum zweiten Glied und schätzt dieses mit der zweiten Bedingung ab. Geschickte Wahl der Größen p, q und Integration liefern dann die Behauptung (zuerst erhält man $Du = Dw$, aber mit Poincarescher Ungleichung folgt daraus $u = w$).

Die zweite Bedingung entspricht der Koerzivitätsbedingung elliptischer Differentialgleichungen, die dort ebenfalls Eindeutigkeit garantiert.

9.3.7 Summary

Hinreichend zur Lösung des Minimierungsproblemes über $W^{1,p}(U)$ ist also

I koerziv und I w.slsc	
$\Longrightarrow$	$\exists u : \min I = I(u)$

bzw. mit stärkeren, aber einfacher zu überprüfenden Voraussetzungen an L dann

L glatt, koerziv und L konvex	
$\Longrightarrow$	$\exists u : \min I = I(u)$
und L gleichmässig konvex	
$\Longrightarrow$	$\exists! u : \min I = I(u)$

Example 9.33. *We will give some examples that exactly fit to the theory above. Of course, most of them are already included in the discussion of the previous section. They are combined here to see some more simple applications.*

We consider the problem with $\Omega \subset \mathbb{R}^n$ bounded:

$$J(u) = \int_\Omega \Big(|\nabla u(x)|^2 + u^2(x) - \phi(x)u(x) \Big)\, dx. \tag{9.30}$$

We consider the problem stated in $U = H_0^1(\Omega)$ and assume $\phi \in L^2(\Omega)$ to be given. First, we check, if $\{u \in U : J(u) \leq M\}$ is non-empty and compact. Using Hölder inequality and the assumption we obtain

$$\int_\Omega \|\nabla u\|^2 dx \leq J(u) \leq M \tag{9.31}$$

[8] Evans, siehe Seite 449ff

Using the Poincare inequality we further see, that $J(u) \leq M$ *implies* $u \in H_0^1(\Omega)$ *and* $\|u\|_{H_0^1} \leq M$. *Hence the above set is weakly (sequentially) compact. Since* $u = 0$ *satisfies the constraint, we further notice that the set if not empty. Next, we have to show, that* J *is weakly sequentially lower semicontinuous. Assume* $u_k \to u$ *weak in* $H_0^1(\Omega)$. *Since weak convergent sequences are bounded, we obtain* ∇u_k *and* u_k *are uniformly bounded in* $L^2(\Omega)$. *Now, we need some embedding theorems from the theory of Sobolev spaces. If* Ω *has smooth boundary and* $n > 2$, *then* $H_0^1(\Omega) \subset\subset L^2(\Omega)$. *Using the embedding we obtain that* u_k *converges in* $L^2(\Omega)$ *to* u. *(I.e. compact operator maps bounded sequences to convergent (sub-)sequences). Then we have* $\int u_k^2 - \phi u_k dx \to \int u^2 - \phi u dx$. *Now, we need to consider the convergence of the gradients* ∇u_k. *We note that*

$$|\nabla u_k|^2 = \textstyle\sum_i (\partial_{x_i} u_k)^2 = \tag{9.32a}$$

$$\textstyle\sum_i \left(\left(\partial_{x_i} u\right)^2 + 2\partial_{x_i} u(\partial_{x_i} u_k - \partial_{x_i} u) + (\partial_{x_i} u_k - \partial_{x_i} u)^2 \right) \geq \tag{9.32b}$$

$$\|\nabla u\|^2 + 2\nabla u(\nabla u_k - \nabla u) \tag{9.32c}$$

$$\implies \qquad \textstyle\int_\Omega |\nabla u_k|^2 - |\nabla u|^2 dx \geq 2 \int_\Omega \nabla u(\nabla u_k - \nabla u) dx \to 0 \tag{9.32d}$$

Indeed, $\nabla u_k - \nabla u$ *is weakly convergent by assumption and* $\nabla u \in L^2(\Omega)$. *Therefore, the above is just the definition of weakly convergent sequences. Finally, we proved*

$$J(u_k) - J(u) \geq 0 \tag{9.33}$$

if $u_k \to u$ *weakly in* $H_0^1(\Omega)$. *Note that we have used the convexity property of* J *in the argument* ∇u. *Now, we can apply the above theory and deduce that there exists a minimum of the problem* $\min J(u)$. *We assume (without proof) that* J *is Frechet differentiable. Then the we can conclude that the variational inequality (with* $C = U$*) is satisfied in the minimum:*

$$J'(u^*) = 0$$

Computing the derivative we obtain the (strong form) of the elliptic differential equation

$$-\Delta u^* + u^* = \phi/2 \in \Omega, u = 0 \text{ on } \partial\Omega \tag{9.34}$$

10 Numerical methods for Unconstrained Optimization In Finite Space Dimensions

In the following we take $f: \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}; \quad f \in C^1(\mathcal{D})$ (In most cases it is implicitly assumed that also $f \in C^{2,1}(\mathcal{D})$, i.e., $f \in C^2(\mathcal{D})$ and for every compact subset $\mathcal{D}_1 \subset \mathcal{D}$ there exists $L \geq 0$, such that

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \text{ for all } x, y \in \mathcal{D}.$$

Also $C^{2,1}(\mathcal{D}) \supset C^3(\mathcal{D})$.)

Normally a scheme determines a sequence $\{x^k\}$ with

(i) $x^k \in \mathcal{L}_f(f(x^0)) = \{x \in \mathcal{D}: \ f(x) \leq f(x^0)\}$

(ii) $\nabla f(x^k) \to 0$

(iii) $x^k - x^{k+1} \to 0$

(There is also a scheme, which in addition enforces the necessary second-order condition "$\nabla^2 f(x^*)$ positive semi-definite" for every accumulation point of $\{x^k\}$.)

The conditions (i), (ii), (iii) are enforced during construction of the scheme.

The existence of accumulation points, the convergence of the whole sequence and the minimality of the limit x^* follows only from additional assumptions on f. It has to be known that $x^0 \in \mathcal{D}$.

The assumption: $\mathcal{L}_f(f(x^0)) = \{x \in \mathcal{D}: \ f(x) \leq f(x^0)\}$ is compact guarantees the existence of the accumulation points and the existence of convergent subsequences with $\nabla f(x^*) = 0$.

10.1 Numerical Schemes for Unconstrained Minimization, $n = 1$

In the following we simplify the problem by taking $\mathcal{D} = \mathbb{R}$. We would like to find the (one) minimum of f through a systematic search on the graph of f, only by applying the function values. An analog for the Bisection method for finding a root (zero) is in this case the Tri-section:

Only the continuity of f is needed. This is favourable if the f-values are inaccurately known ("'corrupted Data or Data with noise"'). The Trisection partitions the interval into three parts, which means two values are evaluated and the length of the total interval is reduced to 2/3 per step.

Consider $f(x)$ (for some x, $f'(x)$ may not exist). Suppose we want to solve

$$\min f(x) \text{ such that } a \leq x \leq b. \tag{10.35}$$

eqn:1dprob

Let $x^* \in [a, b]$ denote the optimal solution (minimizer). We evaluate $f(x)$ at x_1 and x_2 (assume $x_1 < x_2$ on $[a, b]$) to determine the sub-interval in which x^* lies. Such a subinterval is called an **interval of uncertainty.**

There are three cases which we need to consider:

1. Case 1: $f(x_1) < f(x_2)$

   Since $f(x)$ is increasing before x reaches x_2, then $x^* \in [a, x_2)$ as illustrated in the figure.

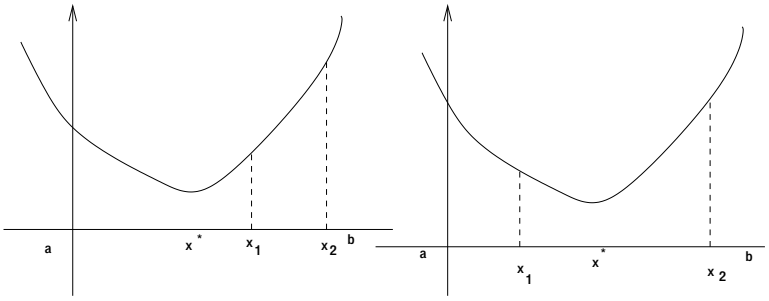


Figure 20: Case 1

2. Case 2: $f(x_1) = f(x_2)$

   For some part of the interval $[x_1, x_2]$, $f(x)$ decreases and the optimal solution $x^* < x_2$ i.e. $x^* \in [a, x_2)$.

3. Case 3: $f(x_1) > f(x_2)$

   For some part of the interval $[x_1, x_2]$, $f(x)$ decreases and the optimal solution $x^* \notin [a, x_1)$ i.e. $x^* \in (x_1, b]$.

The following is the algorithm that can be applied to search for a minimum in one-dimension:

1: Let $I_0 = [a, b]$ be the initial interval of uncertainty.

2: Determine x_1^0 and x_2^0 and evaluate $f(x)$ on these points.

3: While the length of I_k is **not** sufficiently small do

- Determine case 1 - 3 above and reduce I_k.

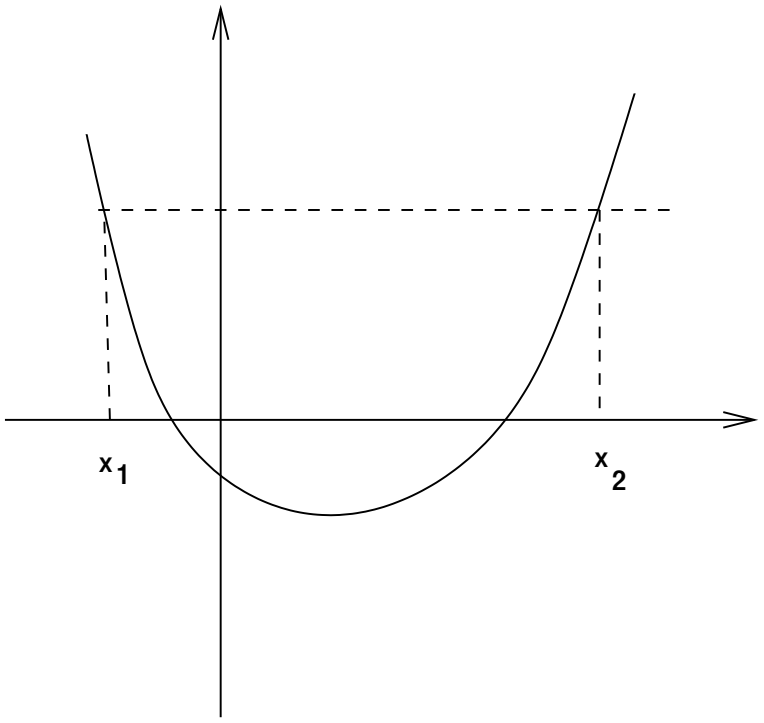


Figure 21: Case 2

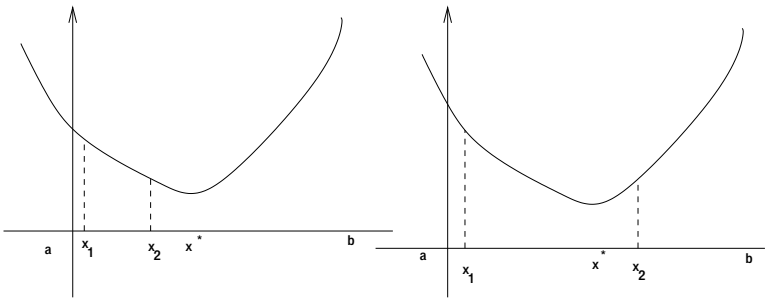
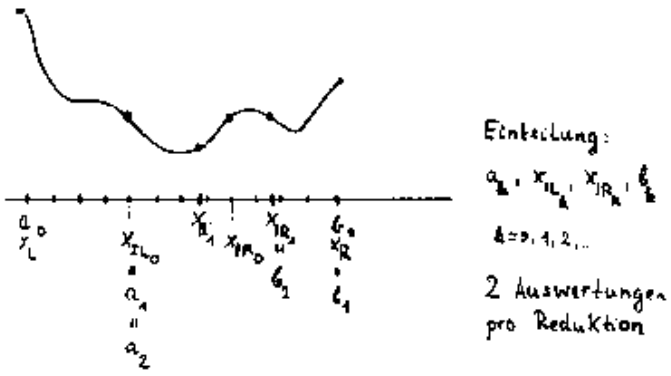


Figure 22: Case 3

- Evaluate $f(x)$ at the two-endpoints of I_k.

To select the endpoints x_1^k and x_2^k one can apply the Golden Section Search described below.

To select the endpoints x_1^k and x_2^k one needs a judicious way of operating. A common example is the Golden Section Search.

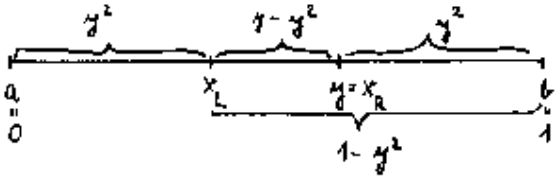


Golden Section Search:

We need a cleverer way of partitioning the interval such that only one test point per interval reduction is needed and the interval reduction is as large as possible. This gives the partition based on the " Principle of a Golden Section Search":

$$\frac{\text{Total Interval}}{\text{Bigger sub-interval}} = \frac{\text{bigger sub-interval}}{\text{smaller sub-interval}}$$

$$\frac{1}{\varrho} = \frac{\varrho}{1-\varrho}, \quad \varrho > 0 \Rightarrow \varrho = \frac{1}{2}(\sqrt{5}-1) \approx 0.618$$



$$\Longrightarrow \quad \frac{\varrho}{(\varrho)^2} = \frac{(\varrho)^2}{\varrho-(\varrho)^2} \quad , \quad \frac{1-(\varrho)^2}{1-\varrho} = \frac{1-\varrho}{\varrho-(\varrho)^2}$$

Here the sub-interval $\overline{0,\varrho}$ is divided by ϱ^2 and the sub-interval $\overline{\varrho^2,1}$ is divided by $1-\varrho$ using the principle of the Golden Section Search.

Problem: Given $[a,b]$. Sought: a minimum x^* of f on $[a,b]$.
Assumption: $f \in C[a,b]$.

Search using the principle of the golden section:

$\varrho := \frac{1}{2}(\sqrt{5}-1)$; $\quad a^{(0)} := a$; $\quad b^{(0)} := b$; $\quad l^{(0)} := (b^{(0)} - a^{(0)})\varrho$
$x_R^{(0)} := b - \varrho l^{(0)}$; $\quad x_L^{(0)} := a + \varrho l^{(0)}$;
Compute $f(x_R^{(0)})$, $f(x_L^{(0)})$.
$k = 0, 1, 2, ...$
$l^{(k+1)} := \varrho l^{(k)}$
In case $f(x_R^{(k)}) > f(x_L^{(k)})$, then
$b^{(k+1)} := x_R^{(k)}$; $\quad a^{(k+1)} := a^{(k)}$;
$x_L^{(k+1)} := a^{(k+1)} + \varrho l^{(k+1)}$; $\quad x_R^{(k+1)} := x_L^{(k)}$
Compute $f(x_L^{(k+1)})$
otherwise
$a^{(k+1)} := x_L^{(k)}$; $\quad b^{(k+1)} := b^{(k)}$; $\quad x_L^{(k+1)} := x_R^{(k)}$;
$x_R^{(k+1)} := b^{(k+1)} - \varrho l^{(k+1)}$
Compute $f(x_R^{(k+1)})$.

Theorem 10.1. *Let $f : [a,b] \to \mathbb{R}$ be strict quasi-convex on $[a,b]$. Then x^* is a minimum of f $x^* \in [a^{(i)}, b^{(i)}]$ for all i and*

$$l^{(i)} = \varrho(b^{(i)} - a^{(i)}) = (\varrho)^i l^{(0)} \quad \text{for all } i,$$

also $a^{(i)} \to x^*$, $b^{(i)} \to x^*$, i.e. the rate of convergence is linear . □

Remark 10.2. • *We would like to note that with the Golden Section Search:*

$$\begin{aligned} x_1 &= b - \rho(b-a); \\ x_2 &= a + \rho(b-a); \end{aligned}$$

i.e. we move a distance of $\rho(b-a)$ from both end points of interval. Thus evaluating $f(x_1)$ and $f(x_2)$ reduces the interval of uncertainty to length $\rho(b-a)$:

if $f(x_1) \le f(x_2)$, $x^* \in [a, x_2)$ which gives $x_3 = x_2 - \rho(x_2 - a)$ and $x_4 = a + \rho(x_2 - a) = a + \rho^2(b-a) = a + (1-\rho)(b-a) = x_1$. Similarly, for $f(x_1) > f(x_2)$ we have $x^* \in (x_1, b]$ in which case $x_3 = x_2$ and $x_4 = x_1 + \rho(b - x_1)$. In general f must only be computed once at each iteration step.

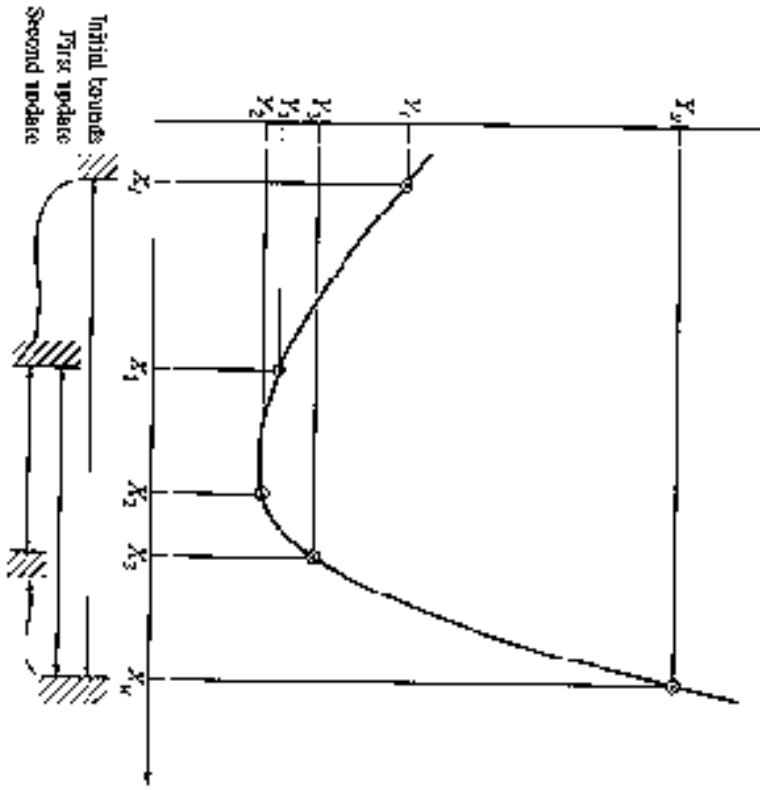
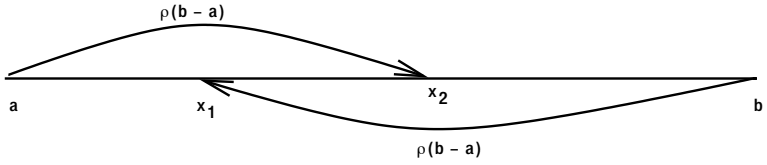


Figure 23: Applying the Golden Section Search



- *Also* $l_2 = \rho l_1 = \rho^2(b-a)$. *In general* $l_k = \rho l_{k-1} = \rho^k(b-a)$. *If we want a final interval of uncertainty to have length* $\ll \epsilon$, *we perform* k *iterations of Golden Section Search where*

$$\rho^k(b-a) < \epsilon$$

Criticism: this is a simple, reliable but expensive algorithm
(too many function values have to be computed if high accuracy is demanded)

But: Faster schemes demand better differentiability properties of f. A helpful tool is the polynomial interpolation:

Here we are interested in polynomials of degree 2 or 3, since in this case a minimum of the polynomials can be determined simply.

10.2 Numerical Schemes for Unconstrained Minimization, $n > 1$

Remark 10.3. *Most usual minimizing schemes normally apply the gradients of the objective function to compute directions, in which x^k will change. Assembling the formulae for the gradients can be a very daunting and expensive process.*

1. *Using formulae for numerical differentiation, for example,*

$$\frac{\partial f}{\partial x_i}(y) = \frac{f(y+\tau e^i) - f(y-\tau e^i)}{2\tau} + \Omega((\tau)^2).$$

   In this case one has to choose the discretisation step-width τ carefully, especially depending on the evaluation accuracy in f, inorder to achieve reasonable results. If the function values of f are themselves results from another algorithm (for example, a Finite Element computation program or a differential equation solver) then it is not clear if such an algorithm produces in the least a differentiable function of the ptimization parameter. An example is a differential equation solver with step-width control. Hence it is recommended that numerical differentiation should be avoided whenever possible. In principle we can also achieve here high accuracy, if the function values themselves are proved to be of high accuracy.

2. *Automatic Differentiation - Application of automatic systems (Precompiler) which generate automatically a programme for differentiation. This approach can be taken, if the evaluation of f is independently undertaken by a separate procedure. Such an approach is called "automatic Differentiation" and there are such programs available. (jakef, ADIFOR, ADOL-C, TAMC etc.) The program code for*

the function is given and the output is also a program code for the function, the gradient and eventually even a Hessian matrix. The so called Modelling Systems which allow an approximative formulation of the Optimization problem, and this is internally transformed into an evaluation program for the corresponding functions, do also contain an option for automatic differentiation (AMPL, GAMS).

3. *Application of Formula Manipulators (Computer Algebra Packages) which after inputting a formula for f generate the formula for ∇f. (MATHEMATICA, MAPLE, AXIOM, DERIVE, MUPAD, etc.)*

10.2.1 Line-Search Methods

The methods are used to determine a point x^* such that $\nabla f(x^*) = 0$. In general one only obtains a point x^* which is not necessarily the minimum. For all currently used methods we require at least a descent in the functional values. The general form of a descent method is

$$x^{k+1} = x^k - \sigma_k d^k \tag{10.36}$$

and we construct methods generating a sequence x^k such that $x^k \to x^*$ and $f(x^*) \leq f(x^k)$. For a proof of convergence it is sufficient to have

$$f(x^k) - f(x^{k+1}) \geq \psi(\|\nabla f(x^k)\|) \tag{10.37}$$

descent suff

with a function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ and of the form $\psi = ct^\beta$ and f bounded from below: $f^k = f(x^k)$ is strictly decreasing and bounded from below. Hence, $f = \lim f^k$ exists and therefore $0 \geq \lim \psi(\|\nabla f^k\|) \geq 0$. The construction of suitable σ_k and d^k, such that (10.37) holds is given below.

Definition 10.4. *Let $x \in \mathcal{S}, f \in C^1$. Then $-d$ is a descent direction, if $\nabla f(x)^T d \geq 0$.*

Definition 10.5. *Let $x \in \mathcal{S}, f \in C^1, -d$ is a descent direction. σ satisfies the principle of sufficient decrease if*

$$f(x) - f(x - \sigma d) \geq c_1 \sigma \nabla f(x)^T d$$

and if

$$\sigma \geq c_2 \nabla f(x)^T d / \|d\|^2$$

for constants $c_{1,2}$ independent of d, x, σ.

In the following we assume that

A1. $D = \mathbb{R}^n, f \in C^2$ bounded from below $, x^0 \in D : \mathcal{L}_f(x^0)$ is compact. (10.38) ass1

Under the previous assumptions is M_2 well-defined.

$$M_2 = \max\{\|\nabla^2 f\| : y \in \mathcal{L}_f(x^0)\}.$$

σ satisfying the principle of sufficient decrease can be found using the following theorem.

Theorem 10.6. *Let $x \in \mathcal{S}, f \in C^1$. Let $d \in \mathbb{R}^n$ with $\nabla f(x)^T d > 0$ and $\delta \in (0,1)$. Then, there exists τ such that*

1. $f(x - \sigma d) < f(x) - \delta\sigma\nabla f(x)^T d \; \forall \sigma \in (0, \tau)$
2. $f(x - \tau d) = f(x) - \delta\tau\nabla f(x)^T d$
3. $\tau \geq \rho := \frac{2(1-\delta)}{M_2} \frac{\nabla f(x)^T d}{\|d\|_2^2}$.
4. $-\frac{d}{d\sigma} f(x - \sigma d) = \nabla f(x - \sigma d)^T d > \delta\nabla f(x)^T d \; \forall \sigma \in (0, \frac{\rho}{2})$.

The previous theorem states that there exists σ sufficiently large ($\geq \frac{\rho}{2}$) satisfying the principle of sufficient decrease. Several possibilities to obtain σ are known.

1. **Exact Line Search** corresponding to property III.

   In the case of the exact line search approach we consider the following problem:

   $$\min_{\sigma > 0} f(x^k + \sigma d^k).$$

   To solve the problem the minimizer for

   $$\phi(\sigma) = f(x^k + \sigma d^k), \quad \sigma > 0$$

   has to be found. The process of finding such a minimizer is too expensive and the exact line search approach is not prefered. However, if f is convex then we evaluate

   $$\sigma = \frac{\nabla f(x)^T d}{d^T \nabla^2 f(x - \theta\sigma d) d}, \; \theta \in (0,1) \quad (10.39)$$ line search convex cas

2. **Inexact Line Search:**

   In this approach steps that are neither too long nor too short are picked. The methods also assist in picking a "useful" initial guess for each step length inorder to ensure fast asymptotic convergence.

   The following is a summary of a basic line search algorithm with backtracking:

   1: Given $\sigma_{init} > 0$ (e.g. $\sigma_{init} = 1$);
   2: Let $\sigma_0 = \sigma_{init}$;
   3: While $f(x^k) \geq f(x^k + \sigma_l d^k)$ do
      - set $\sigma_{l+1} = \tau\sigma_l$, where $\tau \in (0,1)$ (e.g. $\tau = \frac{1}{2}$)
      - $l \leftarrow l + 1$
   4: $\sigma_k = \sigma_l$.

   Backtracking prevents steps from getting too small since $\tau^i\sigma_{init}$, $i = 0, 1, \ldots$ is accepted. Unfortunately, there is no mechanism for preventing steps taken from being too large relative to the decrease in f. A remedy is Armijo's rule.

3. **Goldstein – Armijo**

   This is an inexact line–search with backtracking. Let $(x, d) \in \mathcal{L}_f(x^0) \times \mathbb{R}^n$ with $\nabla f(x)^T d > 0$. Assume parameters δ, β with $0 < \delta, \beta < 1$ and $c_{3,4}$ positive and $c_3 < c_4$. Set

$$\sigma_0 \in \left(c_3 \frac{\nabla f(x)^T d}{\|d\|^2}, c_4 \frac{\nabla f(x)^T d}{\|d\|^2} \right)$$

   and determine

$$k = \min\{j : f(x) - f(x - \beta^j \sigma_0 d) \geq \delta\beta^j \sigma_0 \nabla f(x)^T d\}$$

   and obtain

$$\sigma = \sigma_0 \beta^k.$$

   Theorem 10.7. *The Goldstein–Armijo rule yields always an admissible stepsize σ satisfying the principle of sufficient decrease.*

fig:armijo

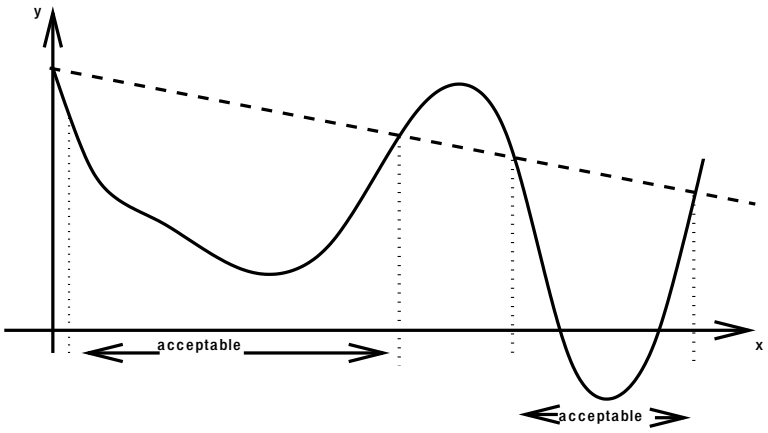


Figure 24: The Armijo condition: $\phi(\sigma) = f(x^k + \sigma d^k)$ on the y axis, the x-axis is σ, the dotted straight line $l(\sigma) = f(x^k) + c_1\sigma\langle d^k, \nabla f(x^k)\rangle$.

If f is convex, $c_{3,4}$ should be chosen such that (10.39) is not(!) excluded. This amounts to set

$$c_3 \leq \frac{1}{\sup \|\nabla^2 f\|},\ c_4 \geq \sup \|\nabla^{-2} f\|.$$

Next, we discuss how to **determine the admissible direction** d^k. The stepsize σ is chosen independent of d. From the principle of sufficient decrease we have

$$f(x^k) - f(x^k - \sigma d^k) \geq c_1 c_2 (\nabla f(x)^T d)^2 / \|d\|^2 \tag{10.40}$$

since $\sigma \geq c_2 \nabla f(x)^T d / \|d\|^2$. If f is bounded from below we obtain $\nabla f(x^k)^T d^k / \|d^k\| \to 0$. For d^k suitable we want to conclude that this implies $\nabla f(x^k) \to 0$. For $g^k = \nabla f(x^k)$ and $\beta^k = (g^k)^T d^k / (\|d^k\| \|g^k\|) = cos(d^k, g^k)$ we have

$$f(x^k) - f(x^k - \sigma d^k) \geq c_1 c_2 (\beta^k)^2 \|g^k\|^2 \tag{10.41}$$

and hence

$$\liminf \|g^k\| \to 0$$

if $\sum (\beta^k)^2 \to \infty$. This has to be satisfied by all directions d^k. On example is the SOR–Newton: d^k is a priori given, normally $d^k = \pm e^{(k \bmod n)+1}$ (coordinate direction) or $d^k = \pm v^{(k \bmod n)+1}$, $\{v^j\}$ **approximate** eigenvector system of $\nabla^2 f(x^k)$, is tried first in the course of computing. These directions are not uniformly downhill direction. Nevertheless with additional

assumption of f being uniformly convex the convergence of the corresponding descent methods can be proved.

General descent directions satisfying the previous assumptions on β^k can be found using gradient directions:

Definition 10.8. *Let $d \in \mathbb{R}^n$. Then d is called downhill direction if, for suitable constants $c_{5,6} > 0$ and independent of x, d we have*

$$c_5\|\nabla f(x)\| \geq \|d\| \geq \frac{1}{c_5}\|\nabla f(x)\|,\ \nabla f(x)^T d \geq c_6\|\nabla f(x)\|\|d\|.$$

From the previous definition it follows

$$\nabla f(x)^T d \geq c_7\|d\|^2,\ \nabla f(x)^T d \geq c_8\|\nabla f(x)\|^2.$$

For downhill directions we have

$$f(x^k) - f(x^k - \sigma d^k) \geq c_1 c_2 c_6^2 \|\nabla f(x^k)\|^2$$

and therefore $\nabla f(x^k) \to 0$.

Theorem 10.9. *Let (A1) hold. Let $x^k = x^{k-1} - \sigma_{k-1} d^{k-1}$ with d^k downhill direction, σ_k satisfying the principle of sufficient decrease. Then, $f(x^k)$ is monotone decreasing, $\nabla f(x^k) \to 0$, if σ_k is bounded, $x^{k+1} - x^k \to 0$, if x^k has a accumulation point x^*, then $\nabla f(x^*) = 0$ and if there are only finitely many points with $\nabla f(x^*) = 0$, then there is precisesly one accumulation point.*

Proof. $f(x^k) - f(x^k - \sigma d^k) \geq c_1 c_2 c_6^2 \|\nabla f(x^k)\|^2 \geq 0$ yields that $f(x^k)$ is monotone decreasing with $x^k \in \mathcal{L}_f(x^0)$. Since $\mathcal{L}_f(x^0)$ is compact $f(x^k) - f(x^{k+1}) \to 0$ and there exists a sequence $x^{k_l} \to x^*$. For every accumulation point of x^k we obtain from $\nabla f(x^*) = 0$ since $f(x^k) - f(x^{k+1}) \to 0$ implies $\|\nabla f(x^k)\|^2 \to 0$. Since $\|x^{k+1} - x^k\| \leq \sigma_k \|d^k\| \leq \sigma_k c_5 \|\nabla f(x^k)\|$ we have for σ bounded, $x^{k+1} - x^k \to 0$. □

Definition 10.10. *The sequence $A^k \in R^{n\times n}$ is called uniformly positive definite, if A^k is symmetric and all eigenvalues of A_k are larger than a constant $\rho > 0$.*

Theorem 10.11. *Let (A1) hold. Let A_k be a uniformly positive definite sequence of matrices with $\lim \|A_k\| \leq c_2$. Then*

$$d^k = A_k^{-1} \nabla f(x^k)$$

is a downhill direction and we have

$$\rho x^T x \leq x^T A_k x \leq c x^T x$$

and $\frac{1}{c} \geq \|A_k^{-1}\| \geq \frac{1}{\rho}$.

We give a few examples of downhill search directions.

1. **Gradient Search Direction/ Steepest Gradient Direction.** In this approach we set $d^k = -\nabla f(x^k)$. This can be also be justified by applying the Taylor Theorem:

   $$f(x^k+\sigma d) = f(x^k)+\sigma\langle \nabla f(x^k), d\rangle + \frac{1}{2}\sigma^2 d^T \nabla^2 f(x^k+td)d \quad \text{for some } t \in (0,\sigma).$$

   The rate of change of f along d at x^k is the coefficient of σ: $\langle \nabla f(x^k), d\rangle$. If we take d to be a unit vector i.e. $\|d\| = 1$ we can evaluate the d of most rapid decrease by solving the following:

   $$\min_d d^T \nabla f(x^k) \quad \text{subject to } \|d\| = 1.$$

   Since $\langle \nabla f(x^k), d\rangle = \|\nabla f(x^k)\|\|d\| \cos\theta$ and $\|d\| = 1$, we obtain

   $$\langle \nabla f(x^k), d\rangle = \|\nabla f(x^k)\| \cos\theta$$

   which is a minimum if $\cos\theta = -1$ at $\theta = \pi$. In conclusion we obtain:

   $$\begin{aligned} \langle \nabla f(x^k), d\rangle &= -\|\nabla f(x^k)\| \\ \langle \frac{-\nabla f(x^k)}{\|\nabla f(x^k)\|}, d\rangle &= 1. \end{aligned}$$

   Since d is a unit vector we have

   $$d = \frac{-\nabla f(x^k)}{\|\nabla f(x^k)\|},$$

   in general one can take

   $$d = -\nabla f(x^k).$$

2. **Newton's method**

   Here d^k is the solution (or approximate solution) of

   $$A_k d^k = -\nabla f(x^k)$$

fig:steepest

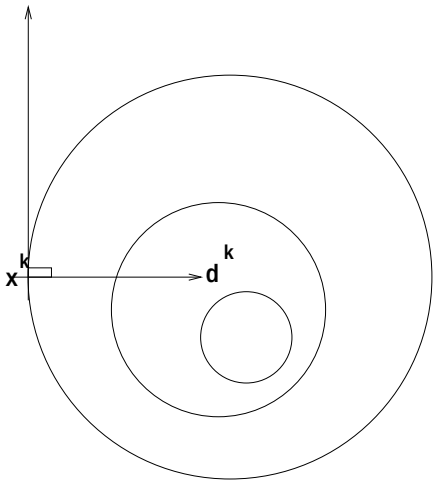


Figure 25: Gradient Search Direction

where $\{A_k\}$ is symmetric and positive definite, the matrix sequence must satisfy the following requirements

$$A_{k+1}(x^{k+1} - x^k) \;=\; \nabla f(x^{k+1}) - \nabla f(x^k) \quad \text{Secant relation}$$

and A_{k+1} can be computed recursively from A_k, $x^{k+1}-x^k$, $\nabla f(x^{k+1})-\nabla f(x^k)$ and eventually from data that was computed much earlier. For some of these constructions one can under additional assumptions prove that the matrix sequence $\{A_k\}$ converges to the Hessian matrix of f in the minimum.

In case the reader wonders where the idea comes from, again a version of Taylor's Theorem is used:

$$\nabla f(x+d) = \nabla f(x) + \nabla^2 f(x)d + \int_0^1 [\nabla^2 f(x+td) - \nabla^2 f(x)]d\; dt$$

where

$$\int_0^1 [\nabla^2 f(x+td) - \nabla^2 f(x)]d\; dt$$

is $o(\|d\|)$ since we assume that ∇f is continuous. Setting $x = x^k$ and $d = x^{k+1} - x^k$, we obtain

$$\nabla f(x^{k+1}) = \nabla f(x^k) + \nabla^2 f(x^{k+1})(x^{k+1} - x^k) + o(\|x^{k+1} - x^k\|)$$

if x^{k+1} and x^k are close to x^*. This result implies that

$$\nabla^2 f(x^{k+1})(x^{k+1} - x^k) \approx \nabla f(x^{k+1}) - \nabla f(x^k) \qquad (10.42)$$ eqn:sdirection

Hence a reasonable choice for A_k is to choose it in such a way that equation (10.42) is mimicked.

With the Newton's method itself we choose

$$A_k \;=\; \nabla^2 f(x^k)$$

The application of exact Hessian matrix is only for uniformly convex f sensible. In other cases A_k must be modified inorder to preserve the positive definiteness and in the process the fast rate of convergence is lost.

Note also that the Newton Method can be derived directly from the Taylor Theorem by performing the following steps: use

$$f(x^k + d) \approx f(x^k) + \langle \nabla f(x^k), d\rangle + \frac{1}{2} d^T \nabla^2 f(x^k) d = m_k(d).$$

Assume $\nabla^2 f(x^k) > 0$, we obtain a Newton direction by finding d that minimizes $m_k(d)$. Setting $\nabla m_k(d) = 0$ gives

$$\nabla^2 f(x^k) d + \nabla f(x^k) = 0.$$

Hence $d^k = -(\nabla^2 f(x^k))\nabla f(x^k)$ which can be rewritten as $\nabla^2 f(x^k) d^k = -\nabla f(x^k)$.

3. **Quasi–Newton methods, DFP, SP1, BFGS** Construction:

   d^k is the solution of $A_k d^k = -\nabla f(x^k)$ (A_k symm. pos. def.)

   With

   $$s^i := x^{i+1} - x^i, \qquad y^i := \nabla f(x^{i+1}) - \nabla f(x^i)$$

   the following requirements are expected:

   $A_i d^i = -\nabla f(x^i),\ x^{i+1} = x^i + \sigma_i d^i$
   $A_{i+1} s^i = y^i$ Secant relation
   "$A_{i+1} - A_i$ small " in the sense of a suitable norm which is also an additional requirement eventually other than symmetry on A_{i+1}.

In one-dimension A_{i+1} is exactly the slope of the secant and in more dimensions its meaning is clear from the Taylor series:

$$y^i \;=\; \int_0^1 \nabla^2 f(x^i + \tau s^i)\mathrm{d}\tau s^i \;.$$

A_{i+1} has in the direction s^i the same properties of the mapping like a local mean value of the Hessian matrix of f. This provides the so called Quasi-Newton method of minimum variation. The most famous are the
Davidon-Fletcher-Powell **DFP** 1959/1963 (the oldest method of this type)

$$A_{i+1} = \left(I - \frac{y^i(s^i)^T}{(s^i)^T y^i}\right) A_i \left(I - \frac{s^i(y^i)^T}{(s^i)^T y^i}\right) + \frac{y^i(y^i)^T}{(y^i)^T s^i} \;^{9}$$

The formula can be derived if the minimal principle

$$||(H_i)^{-1/2}(A_{i+1} - A_i)(H_i)^{-1/2}||_F \;=\; \min$$

with supplementary conditions like symmetry and secant relation is used. $||.||_F$ denotes the Frobenius norm of a matrix (the square root of the square sum of all elements). Hence

$$H_i \;=\; \int_0^1 \nabla^2 f(x^i + \tau s^i)\mathrm{d}\tau \;.$$

This method is not specially favourable since it is sensitive to deviations $\sigma_i - \bar{\sigma}_i$, $\bar{\sigma}_i$ = the optimal step-widths. In any case $(y^i)^T s^i > 0$ otherwise σ_i loses definiteness.

Broyden-Fletcher-Goldfarb-Shanno 1970 **BFGS**

$$A_{i+1} = A_i - \frac{A_i s^i (s^i)^T A_i}{(s^i)^T A_i s^i} + \frac{y^i(y^i)^T}{(y^i)^T s^i}$$

Practically, the inverse of A_i is updated:

$$\begin{aligned} H_i &:= (A_i)^{-1} \\ H_{i+1} &:= I - \rho^i y^i (s^i)^T ()H_i(I - \rho^i y^i (s^i)^T) + \rho^i s^i (s^i)^T \\ d^k &= -H^k \nabla f(x^k). \end{aligned}$$

[9] $xy^T = n \times n$-matrix with components $x_i y_j$ i=row, j=column

where $\rho^i = \dfrac{1}{(y^i)^T s^i}$ In this case the minimal principle is

$$||(H_i)^{1/2}(A_{i+1}^{-1} - A_i^{-1})(H_i)^{1/2}||_F \;=\; \min$$

with the same supplementary condition: $(y^i)^T s^i > 0$. The motivation of BFGS is as follows: If we consider a linear quadratic optimization problem with positive definite, symmetric matrix A of the form $f(x) = \frac{1}{2}x^T Ax - b^T x + \gamma$ the gradient is $\nabla f(x) = Ax - b$. Hence, we have

$$A(x^n - x^{n-1}, x^{n-1} - x^{n-2}, \ldots, x^1 - x^0) = (y^n, y^{n-1}, \ldots, y^1) \quad (10.43)$$

for $y^i = \nabla f(x^i) - \nabla f(x^{i-1})$. If $s^i = x^i - x^{i-1}$ are linearly independent, then A is uniquely defined by the equations $As^i = y^i$. Hence, A can be obtained from $n+1$ pairs $(x^i, \nabla f(x^i))$. The idea of second-order methods is to construct an approximation A_i of A such that

$$A_i s^i = y^i$$

holds and herewith reconstruct A at the local minimum. The only rank–one modification to go from A_i to A_{i+1} is

$$A_{i+1} = A_i + (z^i(z^i)^T)\beta$$

This is SR1. However, A_{i+1} is not necessarily positive definite. Therefore, a rank-2 modification is necessary. We have

$$\tilde{A}_i = A_i - \frac{A_i s^i (s^i)^T A_i}{(s^i)^T A_i s^i}$$

satisfies $\tilde{A}_i s^i = 0$ and $\tilde{A}_i$ is positive semi-definite. An additional rank-1 modification under the assumption $(y^i)^T s^i > 0$ yields the BFGS formula:

$$A_{i+1} = \tilde{A}_i + \frac{y^i (y^i)^T}{(y^i)^T s^i}.$$

In summary we re-iterate the strategy: impose symmetry (symmetry of Hessian), difference between A_{i+1} and A_i must have low rank, A_0 is chosen by the user.

SR1: Broyden 1967

$$\begin{aligned} A_{i+1} &= A_i + \frac{(y^i - A_i s^i)(y^i - A_i s^i)^T}{(y^i - A_i s^i)^T s^i} \\ &= A_i + \frac{(H_i - A_i) s_i s_i^T (H_i - A_i)}{s_i^T (H_i - A_i) s_i} \end{aligned}$$

with

$$H_i \;=\; \int_0^1 \nabla^2 f(x_i + \tau s_i)\mathrm{d}\tau \;.$$

SR1 often converges faster than BFGS, **but** A_{i+1} can not be guaranteed to be pos.def. or can even be not defined! With some modifications of the formulas one can get around this problem.

Originally this scheme was developed from the **quadratic form** of f (in practice it is for this case not interesting). Also

$$f(x) = \tfrac{1}{2}x^T Ax - b^T x, \quad A = A^T \text{pos. def.}$$

Then

$$\nabla f(x) = Ax - b \quad \text{thus} \quad As^i = y^i$$

For DFP, BFGS the following applies in this case: in case σ_j is optimally chosen, then

$$A_k s^j = y^j \qquad j = k-1, \ldots, 0.$$

It is then

$$A_k^{-1} A s^j = s^j,\ j = k-1, \ldots, 0$$

and if $k = n$ then $A_n = A$. Generally $x^N = x^*$ with $N \leq n$. For the SR1 scheme one does not need this requirement in order to run the scheme through inorder to obtain the same result.

10.2.2 Trust–Region Methods

In line–search and Quasi–Newton methods the descent direction and the stepwidth are determined separately. The idea of trust–region methods is to obtain both the stepwidth and the descent direction in a single step. This is only possible, if the function is sufficiently easy to evaluate. Therefore, in the trust region concept we apply a quadratic (or linear) approximation model for f at the beginning of the $k-$th step.

$$f(x) \;\approx \varphi_k(x) \;=\; f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T A_k (x - x^k)$$

as well as (already available) "'trust region radius"' $\tilde{\Delta}_k$, i.e. we expect the quadratic model of f to be sufficiently accurately described in the sphere

$$||x - x^k||_p \;\leq\; \tilde{\Delta}_k \;.$$

fig:trust_region

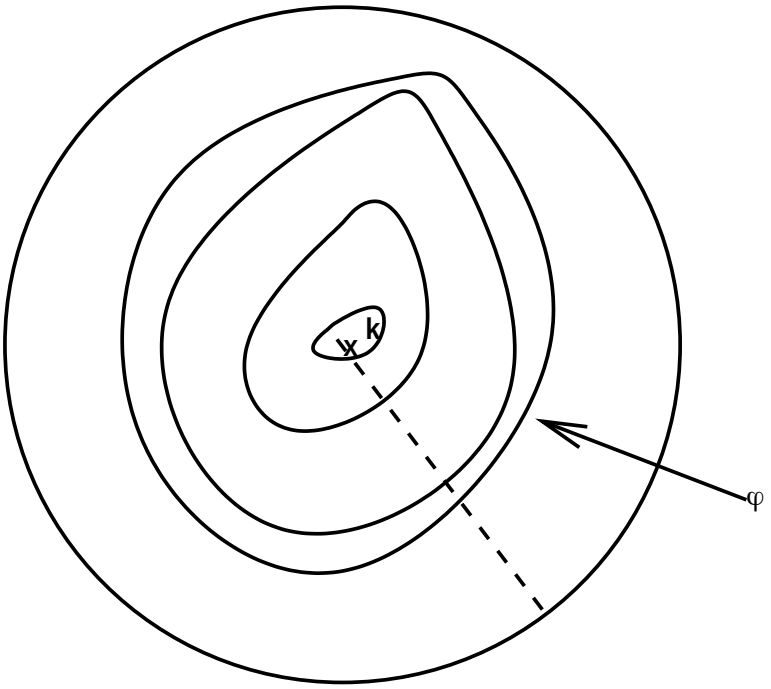


Figure 26: Trust Region: The dotted line is the trust region radius Δ_k, the outer circle is the trust region, $\phi_k(x)$ is the approximation of $f(x)$, and x^k is the current iterate.

The matrix A_k should be symmetric and is typically chosen as a Quasi–Newton approximation for $\nabla^2 f(x^k)$. The idea is now to minimize $\varphi_k(x)$ in the ball $\|x - x^k\| \leq \tilde{\Delta}_k$. Once, we obtain the minimum $\tilde{x}$ we verify, if there has been a sufficient decrease in the $f-$value. If this is the case the new iterate will be $\tilde{x}$. If not, we decrease the trust–region radius $\tilde{\Delta}_k$ and minimize $\varphi_k(x)$ again. The norm $\|.\|$ can be chosen arbitrarily, for example, the Euclidean norm, or the weighted Euclidean norm or the maximum norm.

We now consider the minimization of the model function. The constrained alternative problem is

$$\min_x \varphi_k(x) \;\text{ subject to } ||x - x^k||_p \;\leq\; \tilde{\Delta}_k \;. \tag{10.44}$$

trqp

In fact the global solution of the problem is considered. The problem 10.44 has to be solved for every iteration k and admits a unique solution that can be calculated efficiently. In the case of the Euclidean norm ($p = 2$) the solution problem can be characterised as:

Theorem 10.12. *Let A_k spd. $d^k = x - x^k$ is a unique global solution of the problem 10.44 (with the Euclidean vector norm) if:*

1 $\lambda_k \geq 0,\ ||d^k|| \leq \tilde{\Delta}_k, \lambda_k(\tilde{\Delta}_k - ||d^k||) = 0$

2 $(A_k + \lambda_k I)d^k = -\nabla f(x^k)$

3 $A_k + \lambda_k I$ *is positive semi-definite.*

Details can be found in Sorensen, SIAM J. Numer. Anal. 19, 1982, 409–426. The idea is as follows: Reformulate the constraint as $0 \leq \frac{1}{2}\left(\tilde{\Delta}_k^2 - \|d_k\|_2^2\right)$ and consider the Lagrangian

$$L(d^k, \lambda_k) = f(x^k) + \nabla f(x^k)^T d^k + \frac{1}{2}(d^k)^T A_k d^k - \frac{1}{2}\lambda_k \tilde{\Delta}_k + \frac{1}{2}\lambda_k \|d^k\|_2^2.$$

Provided that a constraint qualification applies the necessary first order optimality system reads

$$\begin{aligned} \nabla f(x^k) + A_k d^k + \lambda_k d^k &= 0 \\ \lambda_k \geq 0, \|d^k\|_2 &\leq \tilde{\Delta}_k, \\ \lambda_k \left(\tilde{\Delta}_k^2 - \|d^k\|_2^2\right) &= 0 \end{aligned}$$

□

Remark 10.13. *Trust Region for Solving* $\min f(x)$*: $k = 1, 2, \ldots$*

(a) Given x^k build $\varphi_k(x)$ as a "model function" of the form:

$$\varphi_k(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k) A_k (x - x^k)$$

with A_k spd (e.g. BFGS, DFP, $\nabla^2 f(x^k)$).

(b) Given $\tilde{\Delta}_k$, solve the problem 10.44 for x^*.

(c) Compute $\dfrac{f(x^k) - f(x^*)}{f(x^k) - \varphi_k(x^*)}$

if $\rho \ll 1$

- *φ_k was not a good model!*
- *$\tilde{\Delta}_k = \tau \tilde{\Delta}_k$ (e.g. $\tau = 1/2$), go to (b).*

– *if* $\rho \approx 1$
$x^{k+1} = x^*$, $\tilde{\Delta}_{k+1} = \tilde{\Delta}_k$ *go to (a).*

In step (b) we have to efficiently solve equation 10.44. We know the solution (x^*, λ_k) needs to satisfy

$$\begin{aligned} (A_k + \lambda_k Id)(x^* - x^k) &= -\nabla f(x^k); \\ \lambda_k \geq 0,\ \lambda_k(\Delta_k - \|x^* - x^k\|) &= 0 \end{aligned}$$

We proceed as follows:

- $A_k(x^* - x^k) = -\Delta f(x^k)$ solve for x^*.
- if $\|x^* - x^k\|_2 \leq \Delta_k$ it implies $\lambda_k := 0$ and have our solution (x^*, λ_k).
- if $\|x^* - x^k\|_2 > \Delta_k$: we define

$$x^*(\lambda) := x^k + (A_k + \lambda Id)^{-1}(-\nabla f(x^k))$$

and solve the nonlinear problem

$$\Delta_k \|x^*(\lambda) - \lambda_k\|_2 = 0$$

for $\lambda \in \mathbb{R},\ \lambda > 0$ using the Bisection Method.

Lower bound $\lambda_{\min} = 0$ and upper bound $\lambda_{\max} = \|\nabla f(x^k)\|_2/\Delta_k$.

If the norm is Euclidean and the A_k is positive semi-definite, then the solution to the previous optimality system reduces to solving a linear system and a nonlinear equation: We have to solve

$$(A_k + \lambda_k I)d^k = -\nabla f(x^k) \tag{10.45}$$ tr-leq

with either $\lambda_k = 0$ and $\|d^k\|_2 \leq \tilde{\Delta}_k$ or $\lambda_k > 0$ and $\|d^k\|_2 = \tilde{\Delta}_k$. For fixed λ_k let us denote by $d^k(\lambda_k)$ the solution to (10.45) (see below for details on its computation). We start by solving $d^k(0)$. If $\|d^k(0)\|_2 \leq \tilde{\Delta}_k$ we are done. Otherwise, we solve the nonlinear equation

$$\|d^k(\lambda)\|_2 - \tilde{\Delta}_k = 0$$

by a bisection method. The solution λ can be bracketed by the lower bound $\lambda_{\min} = 0$ and $\lambda_{\max} = \|\nabla f(x^k)\|_2/\tilde{\Delta}_k$. The bound $\lambda_{\max}$ is an upper bound, since

$$0 = \|d^k(\lambda)\|_2 - \tilde{\Delta}_k \leq \|(A + \lambda)^{-1}\|_2 \|\nabla f(x^k)\|_2 - \tilde{\Delta}_k \leq \frac{\|\nabla f(x^k)\|_2}{\lambda} - \tilde{\Delta}_k.$$

The later inequality holds true since the A_k is positive definite and therefore the eigenvalues of $A_k + \lambda$ are larger than λ. It remains to discuss the computation of $d^k(\lambda)$ as solution to (10.45) for fixed λ. The idea is to transform A_k to tri–diagonal form by Householder transformation which is possible since A_k is assumed to be symmetric and positive definite: We have

$$W_k A_k W_k^T = T_k$$

and $W_k(A_k + \lambda)W_k^T = T_k + \lambda_k$. Hence, in every solution step we compute only a Cholesky decomposition of $T_k + \lambda_k$.

Finally, we have either $\lambda_k = 0$, if A_k is positive definite and

$$A_k d^k \;=\; -\nabla f(x^k) \quad \text{with} \quad ||d^k||_u \;\leq \tilde{\Delta}_k \;,$$

or $\lambda_k \;>\; 0$ and λ_k such that

$$||d^k||_u \;=\; \tilde{\Delta}_k \;.$$

Then set

$$\tilde{x}^{k+1} \;=\; x^k + d^k$$

and test whether $\tilde{\Delta}_k$ has been reasonably chosen, that is the reduction in f using φ_k is somehow achieved. We define the test variables

$$\varrho_k \;\stackrel{def}{=}\; \frac{f(x^k) - f(\tilde{x}^{k+1})}{f(x^k) - \varphi_k(\tilde{x}^{k+1})} \;.$$

If $\varrho_k \;\leq \varepsilon$ where $0 < \varepsilon \ll 1$, then the step is ignored, take half a step, by halving $\tilde{\Delta}_k$ and repeat the computation of d^k. If $\varepsilon \;\leq \varrho_k \;\leq 1 - \eta_1$ mit $0 < \eta_1 < 1 - \varepsilon$, then the step is accepted and set

$$x^{k+1} \;=\; \tilde{x}^{k+1} \;,\; \tilde{\Delta}_{k+1} \;=\; \Delta_k \;=\; \tilde{\Delta}_k \;.$$

If it was the case that $\varrho_k \;>\; 1 - \eta_1$ then the step is also accepted, but we possibly increase $\tilde{\Delta}_{k+1}$:

$$x^{k+1} \;=\; \tilde{x}^{k+1} \;,\; \Delta_k \;=\; \tilde{\Delta}_k \;,\; \tilde{\Delta}_{k+1} \;= \min\{2\Delta_k, \Delta_{\max}\} \;,$$

where $\Delta_{\max}$ is specified by the user. The essential difference with the first concept is the fact that here the correction direction $(x^{k+1} - x^k)/||x^{k+1} - x^k||$ changes with Δ_k. In addition the assumptions that are imposed on A_k are weaker. The effort of doing the algebra per step is frequently higher.

Remark 10.14. *We can compare the descent in φ with the descent in f for the following reason: Assume $\lambda_k = 0$, then $d_k = -A_k^{-1}\nabla f(x^k)$ and the decrease in φ is*

$$\varphi_k(x^k) - \varphi_k(x^k + d^k) = \frac{1}{2}(A_k^{-1}\nabla f(x^k))^T \nabla f(x^k) = -\frac{1}{2}(d^k)^T \nabla f(x^k) > 0.$$

Recall the sufficient decrease condition on f,

$$f(x) - f(x + d) \geq - \text{ const } d^T \nabla f(x).$$

Hence, we can compare the descent in f with some fraction of the descent in φ_k. A similar argument holds true for $\lambda_k > 0$.

The following is the general convergence theorem:

Theorem 10.15. *Let f be twice continuously differentiable on the open set $\mathcal{D}$ and bounded from below. Let the matrix sequence $\{A_k\}$ be bounded. Then every accumulation point x^* of $\{x^k\}$ satisfies the following condition*

$$\nabla f(x^*) \;=\; 0 \;.$$

If the level sphere $\mathcal{L} = \{x : f(x) \leq f(x^0)\}$ is compact, then every infinite subsequence of this sequence has the same accumulation point. If in such an accumulation point $\nabla^2 f(x^*)$ is positive definite, the whole sequence converges to this accumulation point. If

$$A_k \;=\; \nabla^2 f(x^k)$$

then the accumulation point satisfies the second order condition that $\nabla^2 f(x^*)$ is positive semidefinite.

Note that here it is not necessary to explicitly construct directions with negative curvature. Proof: by Schultz, Byrd und Schnabel: A family of trust-region-based algorithms for unconstrained minimization with strong global convergence properties. SIAM J. Numer. Anal. 22, (1985), 47–67. □

11 Numerical Methods for Constrained Minimization Problems In Finite Space Dimensions

11.1 Method for Quadratic Programming Problems

co-qp

We call the following special type of NLO a quadratic programming problem:

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \tag{11.46a}$$

$$h(x) = H^T x + h^0 \tag{11.46b}$$

$$g(x) = G^T x + g^0 \tag{11.46c}$$

for matrices $H \in \mathbb{R}^{n\times p}, G \in \mathbb{R}^{n\times m}$ and vectors $h^0 \in \mathbb{R}^p$ and $g^0 \in \mathbb{R}^m$, respectively. In the following we additionally assume that A is symmetric.

$$\textbf{Assumption}. \qquad A^T = A \tag{11.47}$$

Problems of the type (11.46) typically appear as subproblems for general nonlinear constrained optimization methods. In the case A positive semidefinite the problem is a convex problem. In this case special methods exist for efficiently solving (11.46). There are methods which solve (11.46) in polynomial time. In the case A not positive definite the problem is known to be NP–hard. However, there exists methods which generate feasible points satisfying the first and second–order necessary conditions.

11.1.1 The Primal Projection Method

We present a method for computing a local minimizer of (11.46) in the case A symmetric positive definite. We refer to the remarks below for changes and modifications in the case A symmetric but indefinite. The method is called primal method, since it approximates only x^*. The methods requires a feasible initial value $x^0 \in \mathfrak{S}$.

Before stating the general algorithm we discuss the simpler case $m = 0$, i.e., an equality constrained quadratic programming problem. In the case A symmetric positive definite and H of full column rank (i.e., $rank(H) = p$) we have a convex optimization problem and the KKT–conditions are also sufficient for x^* to be a global minimizer.

Theorem 11.1. *Let $f(x) = \frac{1}{2}x^T Ax - b^T x + c$ with A symmetric, positive definite and $\mathfrak{S} = \{x \in \mathbb{R}^n : H^T x + h^0 = 0\}$. Let H have full column rank. Let $x^0 \in \mathbb{R}^n$ be an arbitrary vector.*

Then, x^* is a global minimum of NLO*, if and only if the following set of linear equations are satisfied*

$$\begin{pmatrix} A & H \\ H^T & 0 \end{pmatrix} \begin{pmatrix} x^0 - x^* \\ \mu^* \end{pmatrix} = \begin{pmatrix} \nabla f(x^*) \\ H^T x^0 + h^0 \end{pmatrix} \tag{11.48}$$

Proof. The KKT–conditions (4.17) are sufficient for global optimality since NLO is a convex optimization problem. Since $\nabla f(x^0) = Ax^0 - b$ we obtain from (4.17)

$$\begin{aligned} Ax^* - b - H\mu^* = 0,\ H^T x^* + h^0 = 0 \\ \Leftrightarrow A(x^0 - x^*) + H\mu^* = Ax^0 - b,\ H^T x^* + h^0 = 0 \\ \Leftrightarrow A(x^0 - x^*) + H\mu^* = \nabla f(x^0), H^T(x^0 - x^*) = H^T x^0 + h^0. \end{aligned}$$

This finishes the proof. □

The assertion of the theorem is depicted in figure 27.

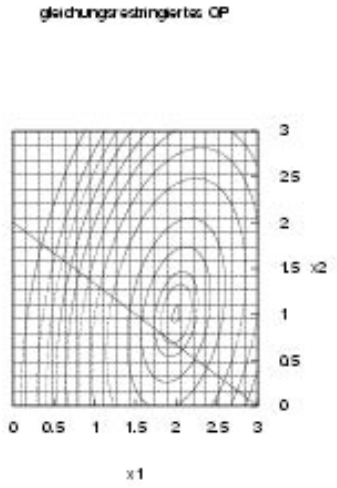


Figure 27: Graphical solution to an equality constraint quadratic programming problem. co-figqp

Hence, the minimum and its corresponding (unique) Lagrange multiplier μ^* are obtained by solving the linear system (11.48). The solution of this linear system should be performed by the following method. This method

relies on the fact that A is positive definite. Denote by $h(x^0) = H^T x0 + h^0 \in \mathbb{R}^p$ Then, apply a QR decomposition of H to obtain

$$QH = \begin{pmatrix} R \\ 0 \end{pmatrix},\ QAQ^T =: B = \begin{pmatrix} B_{11} & B_{21}^T \\ B_{21} & B_{22} \end{pmatrix},\ b := Q\nabla f(x^0) = (b_1, b_2)^T$$

where

$$R \in \mathbb{R}^{p\times p},\ B_{22} \in \mathbb{R}^{n-p\times n-p} \text{ and } b_2 \in \mathbb{R}^{n-p}.$$

Hence, (11.48) is equivalent to

$$QAQ^TQ(x^0 - x^*) + \begin{pmatrix} R\mu^* \\ 0 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$
$$H^TQ^TQ(x^0 - x^*) = (R^T, 0)Q(x^0 - x^*) = h(x^0)$$

If we write $Q(x^0 - x^*) = (s_1, s_2)^T$ with $s_1 \in \mathbb{R}^p$ we can obtain s_1 by

$$s_1 = R^{-T}h(x^0).$$

Hence, we obtain

$$s_2 = B_{22}^{-1}\left(b_2 - B_{21}s_1\right)$$

and

$$\mu^* = R^{-1}\left(b_1 - B_{11}s_1 + B_{21}^T s_2\right)$$

Finally, x^* is obtained by solving $x^* = x^0 + Q^T(s_1, s_2)^T$.

Next, we turn to the case of inequality constraints, i.e., $m > 0$. The solution is found iteratively by considering a sequence of *equality constraint* problems of the previous type. We proceed as follows: At any iterate x^k we fix the active constraints $A_k := \{i \in \{1, \ldots, m\} : g_i(x^k) = 0\}$ and solve the equality constraint problem

$$\min \frac{1}{2}x^TAx - b^Tx + c \text{ subject to } h(x^k) = 0,\ g_i(x^k) = 0\ \forall i \in A(x^k) \quad (11.49)$$

using the previously introduced method. The solution will be denoted by $x^* := x^k - d^k$. According to the sign of the Lagrange multiplier μ_i^k for $i \in A(x^k)$ we can decide if the index i belongs to the correct active set $A(x^*)$. If $\mu_i^k < 0$, then i does *not* belong to the active set for the inequality constraint problem. We hence deactivate the corresponding constraint

before continuing with the next iterate x^{k+1}. We introduce the following notation:

$$\begin{aligned} \mathcal{A}(x^k) &:= \{i \in \{1,\ldots,m\} : g_i(x^k) = 0\}, \\ G_{\mathfrak{B}} &:= (G_{ij})_{i\in\{1,\ldots,n\}, j\in\mathfrak{B}}, \\ N_{\mathfrak{B}} &:= (H, G_{\mathfrak{B}}) \in \mathbb{R}^{n\times p+|\mathfrak{B}|}. \end{aligned}$$

The details are as follows.

Algorithm

Let $x^0 \in \mathfrak{S}$ be given. For $k = 0, 1, 2, \ldots$:

1. Compute the indices of the active set $\mathfrak{B} := \mathcal{A}(x^k)$.

2. Solve the linear system with $d^k \in \mathbb{R}^n, \mu^k \in \mathbb{R}^p$ and $\lambda^k_{\mathfrak{B}} \in \mathbb{R}^{|\mathfrak{B}|}$.
$$\begin{pmatrix} A & N_{\mathfrak{B}} \\ N^T_{\mathfrak{B}} & 0 \end{pmatrix} \begin{pmatrix} d^k \\ \mu^k \\ \lambda^k_{\mathfrak{B}} \end{pmatrix} = \begin{pmatrix} \nabla f(x^k) \\ 0 \end{pmatrix}$$
Note that $x^k - d^k$ is the global minimum of f on the set
$$\mathfrak{F}_{\mathfrak{B}} := \{x \in \mathbb{R}^n : H^T x + h^0 = 0,\ G^T_{\mathfrak{B}} x + g^0 = 0\}.$$

3a. If $d^k = 0$ and $\lambda^k_{\mathfrak{B}} \geq 0$, then $x^k =: x^*$ is the minimum of f on $\mathfrak{S}$ and the KKT–conditions are satisfied for $(x^*, \mu^k \equiv \mu^*, (\lambda^*_{\mathfrak{B}}, 0))$.

3b. If $d^k = 0$ and $\lambda^k_i < 0$ for some $i \in \mathfrak{B}$, then deactivate the constraint $i := argmin\{i \in \mathfrak{B} : \lambda^k_i < 0\}$, i.e.,
$$\mathfrak{B} := \mathfrak{B}\backslash\{i\},$$
and go to (2).

3c. If $d^k \neq 0$, then obtain the optimal stepwidth σ_k such that $x^k - \sigma^*_k d^k \in \mathfrak{S}$ by [10]
$$\begin{aligned} \sigma^*_k &= \min\{\frac{(G^T x^k + g^0)_i}{(d^k)^T \nabla g_i} : \forall\ i \notin \mathfrak{B} \text{ and } (d^k)^T \nabla g_i > 0\} \\ \sigma_k &= \min\{1, \sigma^*_k\} \end{aligned}$$
and update
$$x^{k+1} = x^k - \sigma_k d^k$$
and continue with step (1).

[10] Let $\min\{\emptyset\} = +\infty$

We give an example of the behavior of the algorithm in figure 11.1.1. We have in $x^0 : d^0 \neq 0$ and $\mathfrak{B} = \{0\}$ and $\sigma_0 < \sigma_0^*$. In x^1 we have $\mathfrak{B} = \{1\}$ and $d^1 = 0$, but $\lambda_1^1 < 0$. Hence, we inactivate and obtain x^2 for $\mathfrak{B} = \emptyset$. Finally, in x^2 we have $\mathfrak{B} = \{2\}, d^2 \neq 0, \sigma_2 < \sigma_2^*$ and move to $x^3 \equiv x^*$.

co-figuqp

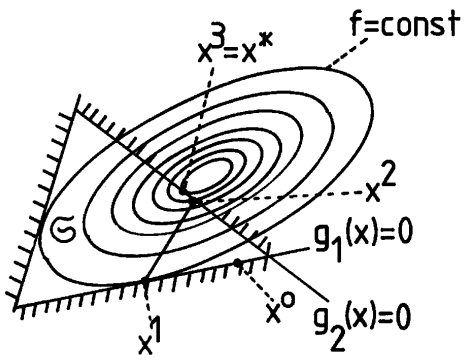


Figure 28: Iterates generated by the primal projection method.

The following theorem guarantees the convergence of the algorithm in a finite number of steps.

S39 **Theorem 11.2.** *Let* $f(x) = \frac{1}{2}x^T Ax - b^T x$ *and* $A = A^T$ *positive definite. Let* $g(x) = G^T x + g^0, \quad h(x) = H^T x + h^0, \quad \mathfrak{S} = \{x \in \mathbb{R}^n : \ g(x) \geq 0, \ h(x) = 0\} \neq \emptyset$. *Assume that*

$$\left.\begin{array}{c} \text{For all } x \in \mathfrak{S} \text{ let } N_{\mathcal{A}(x)} = (H, G_{\mathcal{A}(x)}) = (H, g^{i_1}, \ldots, g^{i_l}) \\ \text{with } \mathcal{A}(x) = \{i_1, \ldots, i_l\} \text{ have full column rank } p + l. \end{array}\right\} \quad (11.50)$$ Vor

Then, the previously introduced algorithm converges to $x^* = argmin\ \{f(x) : x \in \mathfrak{S}\}$ *in a finite number of steps.*

Proof. We have due to the definitions

$$f(x^k) > f(x^{k+1}) \text{ since } \partial_\sigma f(x^k - \sigma d^k) < 0 \ \forall \sigma \in [0, 1).$$

We have to distinguish the following cases

1. $\sigma_k^* \geq 1$. Then, $x^{k+1} = argmin\{f(x) : x \in \mathfrak{F}\}$ and $x \in \mathfrak{F}_{\mathfrak{B}}$ and $\mathcal{A}_{k+1} = \mathcal{A}_k$.
2. $\sigma_k^* < 1$ and $|\mathcal{A}_k| \leq |\mathcal{A}_{k+1}|$.

3. If $\sigma_k^* < 1$ and $|\mathcal{A}_k| = |\mathcal{A}_{k+1}|$, then $d^k = 0$.

4. If $|\mathcal{A}_k| < |\mathcal{A}_{k+1}|$, then $\sigma_k^* < 1$. Since $N_{\mathcal{A}(x)}$ has full column rank, this case can only happen at least n times. Hence, for all i there exists a k such that $i \leq k \leq i + n + 1$ and $x^k = argmin\{f(x) : x \in \mathfrak{F}_{\mathfrak{B}}, \mathfrak{B} \in \mathcal{P}(1, \ldots, m)\}$. Further, $f(x^{k+j}) < f(x^k)$ for all $j \geq 1$.

Since there are only a finite number of subsets of $\mathcal{P}(1, \ldots, m)\}$ the assertion follows. This finishes the proof. □

Some remarks are in order.

1. There are algorithms for solving the convex quadratic programming problem in polynomial time. However, these algorithms need to compute the exact solution to linear systems of the size n.

2. In the previous algorithm we allow at most one constraint to be deactivated in every step. There are alternative methods using multiple deactivations. However, these methods can run into trouble as the following example shows. Assume that x^0 is such that $\nabla f(x^0)$ is a linear combination of all gradients of the active constraints with negative factors. Deactivation of all(!) constraints yields hence a descent direction pointing out of the feasible set and the iterate x^0 does not change any more. Deactivation of only one active constraints yields an admissible descent direction.

3. In case of positive definite A one can also use the characterization of the optimal value as saddle point of the Lagrangian. The method of Goldfarb and Idnani is based on this approach .

4. The case of the indefinite quadratic programming problem. The previously introduced algorithm can still be used in the case of A symmetric and indefinite, if only equality constraints are present. The algorithm applies since it relies on the projected Hessian of $f(x) - \lambda^T h(x)$ only. Under the assumption that NLO admits a unique strict local minimizer the algorithm can still be used.

   In the case of inequality constraints the problem is more severe. Consider for example the case $n = 2, f(x) = \frac{1}{2}\left(x_1^2 + x_2^2\right)$ and $g(x) = (x_2, 1 - x_1 - x_2, 1 + x_1 - x_2)^T$. If we start at $x^0 = (\alpha, 0)$ for any $\alpha \in (-1, 1)$, then we obtain by the previous algorithm $x^1 = (0, 0)$ and $\nabla f(x^1) = 0$. Hence, even if we deactivate the constraints, we cannot decrease along $\nabla f(x^1)$. Further, $\nabla^2 f(x^1)$ is positive definite on $Z_1(x^1)$, but not on $Z_1^+(x^1)$. A remedy would be to allow a descent in direction

of negative(!) curvature of f, here for example $d^1 := -\nabla g_1(x^1)$. Using these ideas in the previous algorithm this implies to change the algorithm such that the projected Hessian B_{22} has at most one negative eigenvalue.

11.2 Trust-Region Methods

We consider the problem of minimizing a nonlinear objective function that depends on real variables with no restrictions on the values of the variables, i.e.

$$\min_{x \in \mathbb{R}^n} \quad f(x) \tag{11.51}$$

p_unconstrained

with $f : \mathbb{R}^n \to \mathbb{R}$. Moreover we assume that f is twice continuously differentiable.

Before we start to consider algorithms to solve (11.51), we begin with a brief review of some basic optimality conditions for (11.51).

-min-unconstrained_min

Definition 11.3.

1. *We call a vector x^* a* global minimizer (solution) *of (11.51) if*

$$f(x^*) \leq f(x) \qquad \text{for all} \qquad x \in \mathbb{R}^n.$$

2. *We call a vector x^* a* local minimizer (solution) *of (11.51) if there exists an $\varepsilon > 0$ such that*

$$f(x^*) \leq f(x) \qquad \text{for all} \qquad x \in \mathcal{B}_\varepsilon(x^*),$$

where $\mathcal{B}_\varepsilon(x^*)$ denotes the ε-Ball at x^*.

3. *We call a vector x^* a* strict (or isolated) local minimizer (solution) *of (11.51) if there exists a $\varepsilon > 0$ such that*

$$f(x^*) < f(x) \qquad \text{for all} \qquad x \in \mathcal{B}_\varepsilon(x^*) \setminus \{x^*\}.$$

Theorem 11.4 (Existence of Minimizers)**.** *Let f be continuous and assume there exists an $x_0 \in \mathbb{R}^n$, such that the level set $\mathcal{N}_0 := \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_0)\}$ is compact. Then there exists a global minimizer of (11.51).*

nec_unconstrained

Theorem 11.5 (Necessary Optimality Conditions)**.**

1. *If x^* is a local minimizer of (11.51) and f is continuously differentiable in an open neighbourhood of x^*, then x^* satisfies the* first order necessary optimality condition *: $\nabla f(x^*) = 0$.*

2. *If x^* is a local minimizer of (11.51) and f is twice continuously differentiable in an open neighbourhood of x^*, then x^* satisfies the* second order necessary optimality condition*: $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.*

We will call a point x^* that satisfies the first order necessary condition a *stationary point* of (11.51). According to Theorem 11.5 any local minimizer of must be a stationary point.

Theorem 11.6 (Sufficient Optimality Conditions)**.** *Suppose that $\nabla^2 f$ is continuous in an open neighbourhood of x^* satisfying $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of f.*

The necessary and sufficient conditions are used to recognize and identify local minimizers. The necessary conditions help us to identify the set of candidates for a local minimizer, whereas the sufficient conditions gives a guarantee that a candidate is in fact a strict local minimizer. However, the sufficient conditions are not necessary, i.e. a point may be a strict local minimizer, and yet might fail to satisfy the sufficient conditions (consider for example the strict global minimizer $x^* = 0$ of $f(x) = x^4$).

Finally, if f is a convex function, then any local minimizer of (11.51) is also a global minimizer.

Theorem 11.7. *When f is convex, any local minimizer of x^* is a global minimizer of f. If in addition f is differentiable, then any stationary point x^* is a global minimizer of f.*

11.2.1 Introduction

There exist a variety of algorithms to solve such unconstrained optimization problems involving a smooth objective function (see e.g. the monographs [**?**], [**?**], [**?**], [**?**]). All methods have in common, that one starts at an initial iterate x_0 and generates a sequence of successive iterates (x_k) using information about the function f at previous iterates. In general, they use this information to find a new iterate x_{k+1} such that $f(x_{k+1}) < f(x_k)$.

The methods differ however in the way move from one iterate to the next and in the choice which information they use. There exist two fundamental strategies: the *line search* and the *trust-region* strategy.

In the line search strategy one chooses a *search direction* s_k and searches along this direction from the current iterate x_k for a new iterate x_{k+1}. A *step length* α along the search direction is then determined by approximately

solving the one-dimensional optimization problem

$$\min_{\alpha>0} \; f(x_k + \alpha s_k).$$

In this chapter we mainly consider the second strategy, the so-called trust region method (as it is described in [**?**],[**?**]). Here one uses the information about f in the current iterate to construct a *model function* m_k that locally approximates f in the current iterate x_k. The model function m_k is constructed such that it is simpler than the nonlinear function f (e.g. quadratic) and thus is easier to solve than the original problem. However, since m_k is only a good approximation of the original objective function in a neighborhood of x_k one successively (approximately) minimizes m_k in some region around x_k (in which we trust our model function m_k to be sufficiently close to f). In other words, in the trust region method we find a candidate step s by (approximately) solving the subproblem

$$\min_{s\in\mathbb{R}^n} \; m_k(s) \qquad \text{subject to} \quad \|s\| \le \Delta. \tag{11.52}$$

tr-sub*

If the candidate s_k (the solution to (11.52)) does not produce sufficient decrease in f, then we deduce that the trust region is too large and we reduce the *trust-region radius* Δ and re-solve (11.52). On the other hand, if we could achieve a good reduction in f, then we have more trust in our model m_k and might enlarge the trust region for the computation of the next step.

Hence, in contrast to the line search strategy, here we determine the direction and the length of the next step simultaneously by (approximately) minimizing the model function m_k inside a trust-region. In general, the direction of the step varies for different sizes of the trust-region radius Δ.

The size of the trust-region is thus a critical factor for the effectiveness of each step. If the region is too small, the algorithm misses an opportunity to take a substancial step towards the minimizer of (11.51). However, if it is too large, then minimizing the model m_k might be misleading, as the solution of (11.52) might be far away from the minimizer of f in the chosen region. Therefore the trust-region in turn is chosen according to the performance of the algorithm for the previous iterations. As identifier one often uses the quotient of the so-called *predicted reduction* $m_k(0) - m_k(s_k)$ and the *actual reduction* $f(x_k) - f(x_k + s_k)$.

11.2.2 Outline of the Algorithm

Let $x_k \in \mathbb{R}^n$ be the current iterate and as mentioned before, we assume that f is twice continuously differentiable. In order to derive an explicit

algorithm from the ideas mentioned above, we have to decide about the model function m_k and trust-region that we want to use and we have to find a suitable update of our trust-region:

1. **The Trust-region Model:**
   A common approach to approximate a nonlinear function in a current iterate x_k uses Taylor-series expansion of f around x_k. Although a linear model might also be of interest, here we will focus on a quadratic model of f, i.e. our model function is of the form

$$m_k(s) := f_k + g_k^T s + \frac{1}{2} s^T B_k s \,, \tag{11.53}$$

   model-function

   where $f_k = f(x_k)$, $g_k = \nabla f(x_k)$ and B_k is some symmetric approximation to the Hessian matrix $\nabla^2 f(x_k)$. Since

$$f(x_k + s) = f_k + g_k^T s + \frac{1}{2} s^T \nabla^2 f(x_k + \theta s) s$$

   for some scalar $\theta \in (0, 1)$ and

$$f(x_k + s) = m_k(s) + \frac{1}{2} s^T (\nabla^2 f(x_k + \theta s) - B_k) s = m_k(s) + O(\|s\|^2)$$

   the approximation error is small if $\|s\|$ is small. Using the exact Hessian matrix, i.e. setting $B_k = \nabla^2 f(x_k)$ yields an approximation error of the order $O(\|s\|^3)$.

2. **The Trust-region:**
   Although our model function $m_k(s)$ is a good approximation of f in a neighbourhood of our current iterate x_k it might be unbounded from below, if B_k is indefinite. We therefore introduce the trust-region constraint

$$\|s\| \leq \Delta_k \,,$$

   where $\Delta_k > 0$ denotes a 'suitable' scalar trust-region radius. In theory, the trust-region subproblem (11.52) does not depend on which norm $\|\cdot\|$ we use, in practice however our choice might make a difference (e.g. if the ∞-norm is used, the feasible region is simply a rectangular box $\rightarrow$ (11.52) yields a box-constrained problem). By the equivalence of norms, there exists constants $\kappa_l \geq \kappa_s > 0$ (dependent on the norm we use), such that

$$\kappa_s \|\cdot\| \leq \|\cdot\|_2 \leq \kappa_l \|\cdot\|.$$

In particular we have $\kappa_l = 1$ and $\kappa_s = n^{-\frac{1}{2}}$ for the l_1 norm and $\kappa_l = n$ and $\kappa_s = 1$ for the l_∞ norm in $\mathbb{R}^n$.

As a general form of (11.52) we obtain the
Trust-region Subproblem:

$$\min_{s \in \mathbb{R}^n} \; m_k(s) = f_k + g_k^T s + \frac{1}{2} s^T B_k s \qquad \text{subject to} \quad \|s\| \leq \Delta_k. \tag{11.54}$$

tr-sub

3. **The Update of the Trust-region Radius:**
Having solved (11.54) we want to accept the trial step s_k and set $x_{k+1} = x_k + s_k$ only if the predicted model decrease $m_k(0) - m_k(s_k)$ (or at least a reasonable fraction of it) is realised by the actual decrease $f_k - f(x_k + s_k)$. We measure this by computing the ratio

$$\rho_k = \frac{ared_k}{pred_k} := \frac{f_k - f(x_k + s_k)}{m_k(0) - m_k(s_k)}, \tag{11.55}$$

tr-rho

where we call the numerator *actual reduction* and the denuminator *predicted reduction*. Note that $m_k(0) = f_k$.

If ρ_k is negative, the new objective value $f(x_k + s_k)$ is greater than the objective value in the current iterate, so the step s_k should be rejected. Moreover, since the model is not accurate, we reduce the trust-region radius to encourage a more suitable step at the next iteration.

On the otherhand, if ρ_k is close to (or even larger than) 1, then we have good reason, that we can trust our model and that the next step might benefit from an increase in the trust-region radius. Hence we expand the trust-region.

We now summarize these steps an describe the process of the basic trust-region method by the following algorithm.

Algorithm 1: Basic Trust-Region Algorithm

Choose an initial vector x_0, an initial trust-region radius Δ_0 and update parameters $0 < \eta_s \leq \eta_v < 1$, $0 < \gamma_d < 1 \leq \gamma_i$;;

for $k = 0, 1, 2, \ldots$ **do**

tr-alg1 **if** $g_k = 0$ **then**
  STOP.

tr-alg2 Choose symmetric approximation B_k of the Hessian matrix $\nabla^2 f(x_k)$.;

tr-alg3 Compute (approximately) the solution s_k of

$$\min_{s \in \mathbb{R}^n} \; m_k(s) = f_k + g_k^T s + \frac{1}{2} s^T B_k s \qquad \text{subject to} \quad \|s\| \leq \Delta_k.$$

;

tr-alg4 Compute ρ_k given by (11.55).;

tr-alg5 Update x_k and the trust-region radius Δ_k: ;

if $\rho_k \geq \eta_v$ **then**
  Set $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \gamma_i \Delta_k$
else if $\rho_k \geq \eta_s$ **then**
  Set $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \Delta_k$
else
  Set $x_{k+1} = x_k$ and $\Delta_{k+1} = \gamma_d \Delta_k$

tr-alg

Here, reasonable values for the parameters η and γ are for example $\eta_v = 0.9$, $\eta_s = 0.1$ and $\gamma_i = 2$,$\gamma_d = 0.5$. In the following we will call a step s_k very successful if $\rho_k \geq \eta_v$, we call it successful if $\rho_k \geq \eta_s$ and finally we call it unsuccessful if $\rho_k < \eta_s$.

What remains for us to clarify is how to solve the trust-region subproblems in step 3 of Algorithm 1.

11.2.3 The Trust-Region Subproblem

11.2.4 Characterization Exact Solutions

In this section we consider the exact solutions of the trust-region subproblem (11.54) using the l_2-norm, i.e. $\|\cdot\| = \|\cdot\|_2$. The following theorem gives us a precise characterization of the exact solutions of (11.54).

thm-tr-kkt

Theorem 11.8. *[Charakterization of Exact Solutions] The vector s^* is a global solution of the trust-region subproblem*

$$\min_{s \in \mathbb{R}^n} \; m(s) = f + g^T s + \frac{1}{2} s^T B s \qquad \textit{subject to} \quad \|s\|_2 \leq \Delta\,,$$

with B_k being symmetric, if and only if s^* is feasible and there exists a scalar $\lambda \geq 0$ such that the following conditions are satisfied:

$$\begin{aligned} (B + \lambda I)s^* &= -g, && (11.56)\\ \lambda(\Delta - \|s^*\|) &= 0, && (11.57)\\ (B + \lambda I) &\quad \textit{is positive semidefinite.} && (11.58) \end{aligned}$$

If (11.56) and (11.57) are satisfied and $B + \lambda I$ is positive definite, s^* is unique.

In order to prove this result we will make use of the following lemma.

tr-lem-exact

Lemma 11.9. *Let m be the quadratic function defined by*

$$m(s) = g^T s + \frac{1}{2}\, s^T B s,$$

where B is any symmetric matrix. Then

1. *m attains a minimum if and only if B is positive semidefinite and g is in the range of B;*
2. *m has a unique minimizer if and only if B is positive definite;*
3. *if B is positive semidefinite, then every s satisfying $Bs = -g$ is a global minimizer of m.*

Proof. First let g be in the range of B, then there exists a vector s such that $Bs = -g$. Hence, for all $v \in \mathbb{R}^n$ we have

$$\begin{aligned} m(s+v) &= g^T(s+v) + \frac{1}{2}\,(s+v)^T B(s+v)\\ &= g^T s + \frac{1}{2}\, s^T B s + g^T v + \frac{1}{2}\, v^T B v + (Bs)^T v\\ &= m(s) + \frac{1}{2}\, v^T B v \geq m(s), \qquad (11.59) \end{aligned}$$

since B is supposed to be positive semidefinite. Thus s is a minimum of m. For the opposite direction, assume that s is a minimizer of m. Then the necessary conditions (Definition 11.5) have to be satisfied, i.e.

$$\nabla m(s) = g + Bs = 0 \qquad \text{and} \qquad \nabla^2 m(s) = B \text{ is positive semidefinite,}$$

which proves the first part.

Concerning the second part, if B is positive definite, we get a strict inequality in (11.59). Hence, s is the unique minimum of m. To prove the opposite direction, it remains to show that B is positive definite. However, assume there exists a vector $v \neq 0$ such that $v^T B v = 0$, thus $Bv = 0$. Then (11.59) yields $m(s + v) = m(s)$ in contradiction to the uniqueness of s.

Finally, the third part follows from the proof of the first part. □

Next, we will prove Theorem 11.8.

Proof. (of Theorem 11.8) Assume that there exists a $\lambda \geq 0$ such that the conditions (11.56)-(11.58) are satisfied, then by Lemma 11.9 s is a global minimizer of the function

$$\tilde{m}(s) := g^T(s) + \frac{1}{2}\, s^T(B + \lambda I)s = m(s) + \frac{\lambda}{2} s^T s\,. \tag{11.60}$$

thm-tr-exact-prf1

Hence,

$$m(v) \geq m(s) + \frac{\lambda}{2}(s^T s - v^T v) \tag{11.61}$$

thm-tr-exact-prf2

for all $v \in \mathbb{R}^n$. By (11.57) it follows that $\lambda(\Delta^2 - s^T s) = 0$ and thus

$$m(v) \geq m(s) + \frac{\lambda}{2}(\Delta^2 - v^T v)$$

which implies that $m(v) \geq m(s)$ for all $v \in \mathbb{R}^n$ that satisfy $\|v\| \leq \Delta$. Hence, s is a global minimizer of the trust-region subproblem.

For the converse, assume that s is a global solution of the trust-region subproblem. We will show that there is a $\lambda \geq 0$ such that (11.56)-(11.58) are satisfied. First, consider the case $\|s\| < \Delta$ then by Lemma 11.9 the conditions are satisfied for $\lambda = 0$.

We therefore assume that $\|s\| = \Delta$. Then any $\lambda \in \mathbb{R}$ satisfies (11.57). Suppose there exist no $\lambda \geq 0$ such that (11.56) is satisfied. Then $y = \nabla m(s) = g + Bs \neq 0$ and $y^T s = (Bs + g)^T s \neq -\lambda s$ for any $\lambda \geq 0$. We therefore have $\alpha := \angle(y, s) \neq \pi$, thus

$$\cos(\alpha) = \frac{y^T s}{\|y\|\,\|s\|} > -1.$$

Define $v = -\left(\frac{y}{\|y\|} + \frac{s}{\|s\|}\right)$, then v is a descent direction of m in s since

$$\nabla m(s)^T v = y^T v = -\left(\frac{y^T y}{\|y\|} + \frac{y^T s}{\|s\|}\right) = -\|y\|\,(1 + \cos\alpha) < 0.$$

Moreover,

$$\left[\frac{d}{dt}\frac{1}{2}\|s + tv\|^2\right]_{t=0} = s^T v = -\left(\frac{s^T y}{\|y\|} + \frac{s^T s}{\|s\|}\right) = -\|s\|\,(\cos\alpha + 1) < 0.$$

Thus, for small $t > 0$, we have $\|s + tv\| < \|s\| = \Delta$. Hence, we obtain a contradiction to the optimality of s such that there has to exist a $\lambda \geq 0$ such that (11.56) and (11.57) are satisfied. To prove that (11.58) holds, we show that

$$w^T(B + \lambda I)w \geq 0 \qquad \forall w \in \mathbb{R}^n \qquad \text{with} \quad w^T s < 0.$$

Let $t = -2\frac{w^T s}{\|w\|^2} > 0$. Then

$$\|s + tw\|^2 = \|s\|^2 + 2\,tw^T s + t^2\|w\|^2 = \|s\|^2 \leq \Delta^2$$

and therefore

$$\begin{aligned} 0 \;&\leq\; m(s + tw) - m(s) = ty^T w + \frac{t^2}{2}\,w^T B w = -t\lambda\, s^T w + \frac{t^2}{2}\,w^T B w \\ &=\; \frac{t^2}{2}\,\lambda\|w\|^2 + \frac{t^2}{2}\,w^T B w = \frac{t^2}{2}\;w^T(B + \lambda I)w \end{aligned}$$

The inequality is independent of the sign of w and by reasons of continuity it remains true for vectors w with $w^T s = 0$, hence $(B + \lambda I)$ is positive semidefinite.

Finally, the uniqueness of s in the case that $(B + \lambda I)$ is positive definite follows from Lemma 11.9 and thus a strict inequality in (11.61).

□

11.2.5 Calculating Nearly Exact Solutions

Theorem 11.8 suggests a method how the trust-region subproblems in step 3 of Algorithm 1 can be solved. First, if B is positive semidefinite and the solution s_k of $B_k s = -g_k$ satisfies $\|s_k\| \leq \Delta_k$, then s_k solves (11.54). This can simply be checked by evaluating if B_k can be decomposed by Cholesky factors, using these factors to solve the linear system and evaluating the norm of that solution.

However, if the solution s_k of the linear system satisfies $\|s_k\| > \Delta_k$ or B is singular or indefinite, then we are searching for a solution (s, λ) of the nonlinear system of equations

$$(B_k + \lambda I)s = -g_k \qquad \text{and} \qquad \|s\| = \Delta_k, \tag{11.62}$$

tr-nonlinearsystem

where, following Therorem 11.8, we require in addition that $(B_k+\lambda I)$ is positive semidefinite. In the following, we are therefore searching for a method to solve (11.62).

Since B is supposed to be symmetric (as it is the exact or approximated Hessian of f), there exist an orthogonal matrix Q and a diagonal matrix Λ such that $B = Q\Lambda Q^T$, where Q consists of (orthonormal) eigenvectors of B and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ are the eigenvalues of B. It clearly holds that $(B + \lambda I) = Q(\Lambda + \lambda I)Q^T$. Next, we define the following function

$$\phi(\lambda) := \|s(\lambda)\|^2 = \| - Q(\Lambda + \lambda I)^{-1}Q^T g\|^2 = \sum_{j=1}^{n} \left(\frac{q_j^T g}{\lambda_j + \lambda} \right)^2 \tag{11.63}$$

tr-exact-fct1

where q_j denotes the jth column of Q. This function is well-defined for $\lambda \neq \lambda_j$ and if $\lambda > -\lambda_1$, then $\phi(\lambda)$ is strictly monotone decreasing, i.e. $\phi(\lambda) \to 0$ for $\lambda \to +\infty$ $(\lambda > -\lambda_1)$. Moreover, since $g \neq 0$, $\phi(\lambda) \to +\infty$ for $\lambda \to -\lambda_j$. Hence, there exist a unique solution λ^* of $\phi(\lambda) = \Delta^2$ in the interval $(-\lambda_1, +\infty)$. Thus, the idea to solve the trust-region subproblem that directly comes into one's mind is to find the root $\lambda^* > -\lambda_1$ of

$$\psi(\lambda) := \phi(\lambda) - \Delta^2.$$

However, since ϕ is very unpleasant function that is highly nonlinear and has many poles, Newton's method (as a standard method for finding a root of a nonlinear function) will be unreliable or slow. We therefore transform the problem and use Newton's method to solve instead

$$\tilde{\psi}(\lambda) := \frac{1}{\|s(\lambda)\|} - \frac{1}{\Delta}.$$

The derivative of $\tilde{\psi}(\lambda)$ is

$$\tilde{\psi}'(\lambda) = -\frac{s(\lambda)^T \nabla_\lambda s(\lambda)}{\|s(\lambda)\|^3}.$$

Furthermore, differentiating the nonlinear system of equations $(B+\lambda I)s(\lambda) = -g$ we obtain an expression for $\nabla_\lambda s(\lambda)$, as

$$(B + \lambda I)\nabla_\lambda s(\lambda) + s(\lambda) = 0 \qquad \Rightarrow \qquad \nabla_\lambda s(\lambda) = -(B + \lambda I)^{-1} s(\lambda).$$

However, given the factorization $(B + \lambda I) = L(\lambda)L(\lambda)^T$ we can directly replace $s(\lambda)^T \nabla_\lambda s(\lambda)$ by

$$\begin{aligned} -s(\lambda)^T \nabla_\lambda s(\lambda) &= s(\lambda)^T((B + \lambda I)^{-1} s(\lambda)) \\ &= s(\lambda)^T (L(\lambda)^{-T} L(\lambda)^{-1}) s(\lambda) \\ &= (L(\lambda)^{-1} s(\lambda))^T (L(\lambda)^{-1} s(\lambda)) \\ &= \|w(\lambda)\|^2, \end{aligned}$$

where $L(\lambda)w(\lambda) = s(\lambda)$. Applying Newton's method to $\tilde{\psi}(\lambda)$ then yields the following algorithm to calculate nearly exact solutions the trust-region subproblem

Algorithm 2: Nearly Exact - Subproblem Algorithm

Given an initial $\lambda^(0) > -\lambda_1$ and $\Delta > 0$;

for $\ell = 0, 1, 2, \ldots$ **do**

tr-sub-alg1 Factorize $B + \lambda^(\ell) I = LL^T$.;

tr-sub-alg2 Solve $LL^T s = -g$.;

tr-sub-alg3 Solve $Lw = s$.;

tr-sub-alg3 Set

$$\lambda_{\ell+1} = \lambda_\ell + \left(\frac{\|s\| - \Delta}{\Delta} \right) \frac{\|s\|^2}{\|w\|^2}.$$

tr-sub-alg

11.2.6 Convergence

In the previous section we discussed exact and nearly exact solutions of the trust-region subproblem (11.54). However, computing the (nearly) exact minimum of the subproblem might be quite expensive, especially in the case of large-scale problems. Fortunately, as we will see next, in order to guarantee global convergence it is sufficient to require that we achieve as much reduction in the model as we would from a step in the direction of steepest descent. This reference solution s_k^C is called the Cauchy point. In contrast to the previous section, here we consider again the general trust-region subproblem, i.e. we use a general norm $\|\cdot\|$ for the definiton of the trust-region.

11.2.7 The Cauchy Point

Definition 11.10 (Cauchy Point)**.** *Let $g_k = \nabla f(x^k)$, then the direction of steepest descent is $-g_k$ and we define the Cauchy point $s_k^C := -\alpha_k^C g_k$, where*

$$\alpha_k^C := arg\min_{\alpha>0} m_k(-\alpha g_k) \quad subject\ to \quad \alpha\|g_k\| \leq \Delta_k. \tag{11.64}$$

cauchy-problem

Hence, computing the Cauchy point results in a simple scalar optimization problem, since we minimize the quadratic function $\tilde{m}(\alpha) := m_k(-\alpha g_k)$ in the interval $0 < \alpha \leq \frac{\Delta}{\|g_k\|}$. Moreover, we can guarantee a reasonable reduction in the model at the Cauchy point.

thm-cauchy

Theorem 11.11. *Let s_k^C be the Cauchy point of (11.54), then*

$$f_k - m_k(s_k^C) \;\geq\; \frac{1}{2}\|g_k\|_2 \min\left(\frac{\|g_k\|_2}{1+\|B_k\|_2}, \kappa_s\Delta_k\right) \tag{11.65}$$

thm-cauchy-1

Proof. If $g_k = 0$ then (11.65) is clearly satisfied. Hence, suppose $g_k \neq 0$.

1. If the curvature of m_k along the direction g_k is not strictly positive, i.e. $g_k^T B_k g_k \leq 0$, then $\tilde{m}(\alpha) = m_k(-\alpha g_k)$ is not bounded from below and thus the solution of (11.64) is $\alpha^* = \frac{\Delta_k}{\|g_k\|}$. Moreover, for all $\alpha \geq 0$

$$m_k(-\alpha g_k) = f_k - \alpha\|g_k\|_2^2 + \frac{\alpha^2}{2} g_k^T B_k g_k \leq f_k - \alpha\|g_k\|_2^2.$$

   Hence,

$$\begin{aligned} f_k - m_k(s_k^C) &= f_k - m_k(-\alpha^* g_k) \\ &\geq \alpha^*\|g_k\|_2^2 \\ &= \frac{\Delta}{\|g_k\|}\|g_k\|_2^2 \geq \kappa_s\ \Delta_k\|g_k\|_2 \\ &\geq \frac{1}{2}\ \kappa_s\ \Delta_k\ \|g_k\|_2 \end{aligned}$$

2. If on the other hand $g_k^T B_k g_k > 0$, then $\tilde{m}(\alpha)$ is strictly convex. Thus, if the minimizer α^* of (11.64) lies inside the interval $(0, \frac{\Delta_k}{\|g_k\|}]$, then the first order necessary conditions for (11.64) reveal that

$$\alpha^* = \frac{\|g_k\|_2^2}{g_k^T B_k g_k} \tag{11.66}$$

thm-cauchy-2

and therefore

$$f_k - m_k(s_k^C) = \frac{\|g_k\|_2^4}{g_k^T B_k g_k} - \frac{1}{2}\frac{\|g_k\|_2^4}{g_k^T B_k g_k} = \frac{1}{2}\frac{\|g_k\|_2^4}{g_k^T B_k g_k} \geq \frac{1}{2}\frac{\|g_k\|_2^2}{1+\|B_k\|_2},$$

where we have used

$$g_k^T B_k g_k \leq \|g_k\|_2^2 \, \|B_k\|_2 \leq \|g_k\|_2^2 \, (1+\|B_k\|_2).$$

3. If $g_k^T B_k g_k > 0$ and the minimizer of $\tilde{m}(\alpha)$ lies outside the interval $(0, \frac{\Delta_k}{\|g_k\|}]$, then

$$\frac{\|g_k\|_2^2}{g_k^T B_k g_k} \geq \frac{\Delta_k}{\|g_k\|}$$

thus

$$\begin{aligned} f_k - m_k(s_k^C) &= \frac{\Delta_k}{\|g_k\|}\|g_k\|_2^2 - \frac{1}{2}\frac{\Delta_k^2}{\|g_k\|^2} \, g_k^T B_k g_k \\ &\geq \frac{1}{2}\frac{\Delta_k}{\|g_k\|} \, \|g_k\|_2^2 \\ &\geq \frac{1}{2} \, \kappa_s \, \Delta_k \, \|g_k\|_2 \end{aligned}$$

This yields (11.65). □

Furthermore, any step s_k that is at least as well suitable as the Cauchy point to reduce $m_k(s)$ inside our trust-region $\|s\| \leq \Delta_k$ will also satisfy the inequality (11.65).

cor-cauchy1 **Corollary 11.12.** *If s_k is an improvement on the Cauchy point s_k^C within the trust-region $\|s_k\| \leq \Delta_k$, then*

$$f_k - m_k(s_k) \geq \frac{1}{2}\|g_k\|_2 \min\left(\frac{\|g_k\|_2}{1+\|B_k\|_2}, \kappa_s \Delta_k\right) \qquad (11.67)$$ cor-cauchy1-1

Furthermore, for any vector s_k that satisfies $f_k - m_k(s_k) \geq \beta(f_k - m_k(s_k^C))$ we have

$$f_k - m_k(s_k) \geq \frac{\beta}{2}\|g_k\|_2 \min\left(\frac{\|g_k\|_2}{1+\|B_k\|_2}, \kappa_s \Delta_k\right).$$

11.2.8 Convergence Analysis

Next, we will use these results to analyse the convergence behaviour of our trust-region algorithm (Algorithm 1).

We start our analysis with an error estimate of our model function m_k, i.e. how much can the objective function f and our model function m_k vary (inside the trust-region). Since we are using a second order approximation, we expect that the error will be of second order in terms of the norm of s_k.

lem-tr1 **Lemma 11.13.** *Suppose that f is twice continuously differentiable and that there exists a constant $C_H \geq 1$ such that the true Hessian satisfies $\|\nabla^2 f(x)\|_2 \leq C_H$ for all x. Moreover, suppose that the model Hessian satisfies $\|B_k\|_2 \leq C_B$ for some positive parameter C_B. Then*

$$|f(x_k + s_k) - m_k(s_k)| \leq \frac{1}{2}\,\kappa_l^2\,(C_H + C_B)\,\Delta_k^2, \tag{11.68}$$ lem-tr1-1

for all k and s_k with $\|s_k\| \leq \Delta_k$.

Proof. From the generalized mean-value theorem it follows that there exists some $\xi_k \in [x_k, x_k + s_k]$ such that

$$f(x_k + s_k) = f_k + g_k^T s_k + \frac{1}{2}\, s_k^T\, \nabla^2 f(\xi_k) s_k.$$

Hence,

$$\begin{aligned} |f(x_k + s_k) - m_k(s_k)| &= \frac{1}{2}\,|s_k^T\,\nabla^2 f(\xi_k) s_k - s_k^T\, B_k s_k| \;\leq\; \frac{1}{2}\,(|s_k^T\,\nabla^2 f(\xi_k) s_k| + |s_k^T\, B_k s_k|) \\ &\leq \frac{1}{2}\,(C_H + C_B)\,\|s_k\|_2^2 \;\leq\; \frac{1}{2}\,(C_H + C_B)\,\kappa_l^2\,\Delta_k^2. \end{aligned}$$

□

The next result states the fact that we can always achieve good progress, if our current iterate x_k is not optimal yet (i.e. $g_k \neq 0$) and the trust-region is sufficiently small.

lem-tr2 **Lemma 11.14.** *Suppose that there exist two constants $C_H \geq 1$ and $C_B \geq 0$ such that the Hessians of the objective and the model function satisfy $\|\nabla^2 f(x_k)\|_2 \leq C_H$ and $\|B_k\|_2 \leq C_B$ and let $C = \frac{1}{2}\,(C_H + C_B)\,\kappa_l^2$. Suppose furthermore that $g_k \neq 0$ and that*

$$\Delta_k \leq \|g_k\|_2 \;\min\; \left(\frac{1}{\kappa_s(C_H + C_B)}, \frac{\kappa_s(1-\eta_v)}{2\,C}\right). \tag{11.69}$$ bound-tr-rad

Then iteration k is very successful and

$$\Delta_{k+1} \geq \Delta_k.$$

Proof. By assumption, we have that $1 + \|B_k\|_2 \leq C_H + C_B$. Therefore, it follows by (11.69)

$$\kappa_s \Delta_k \leq \frac{\|g_k\|_2}{C_H + C_B} \leq \frac{\|g_k\|_2}{1 + \|B_k\|_2}.$$

Corollary 11.12 then yields

$$f_k - m_k(s_k) \;\geq\; \frac{1}{2}\,\|g_k\|_2 \,\min\left(\frac{\|g_k\|_2}{1 + \|B_k\|_2},\, \kappa_s \Delta_k\right) \;=\; \frac{1}{2}\,\|g_k\|_2\,\kappa_s \Delta_k$$

However, Lemma 11.13 and again (11.69) then gives

$$|\rho_k - 1| = \left|\frac{f(x_k + s_k) - m_k(s_k)}{f_k - m_k(s_k)}\right| \;\leq\; 2\frac{C\Delta_k^2}{\kappa_s \|g_k\|_2 \Delta_k} \;=\; 2\frac{C}{\kappa_s}\frac{\Delta_k}{\|g_k\|_2} \;\leq\; 1 - \eta_v.$$

Therefore, $\rho_k \geq \eta_v$ and the iteration is very successful. □

Next we use this result to show that the trust-region radius Δ_k will not shrink to zero if the sequence (g_k) is bounded away from zero.

lem-tr3

Lemma 11.15. *Suppose that there exist two constants $C_H \geq 1$ and $C_B \geq 0$ such that the Hessians of the objective and the model function satisfy $\|\nabla^2 f(x_k)\|_2 \leq C_H$ and $\|B_k\|_2 \leq C_B$ and let $C = \frac{1}{2}\,(C_H + C_B)\,\kappa_l^2$. Suppose furthermore that there exists an $\epsilon > 0$ such that $\|g_k\|_2 \geq \epsilon$ for all k. Then*

$$\Delta_k \geq c_\epsilon \;:=\; \epsilon\,\gamma_d \,\min\left(\frac{1}{\kappa_s(C_H + C_B)}, \frac{\kappa_s(1 - \eta_v)}{2\,C}\right) \tag{11.70}$$

bound-tr-rad2

for all k.

Proof. Assume that the sequence (Δ_k) is not bounded from below and that the iteration k is the first one where the trust-region radius falls below the bound of (11.70), i.e.

$$\Delta_{k+1} < c_\epsilon. \tag{11.71}$$

lem-tr3-1

Since by assumption $\Delta_k \geq c_\epsilon > \Delta_{k+1}$, the previous iteration must have been unsuccessful and therefore $\Delta_{k+1} = \gamma_d\,\Delta_k$. However, if we substitute Δ_{k+1} in (11.71) we get

$$\Delta_k < \epsilon \,\min\left(\frac{1}{\kappa_s(C_H + C_B)}, \frac{\kappa_s(1 - \eta_v)}{2\,C}\right) \;\leq\; \|g_k\| \,\min\left(\frac{1}{\kappa_s(C_H + C_B)}, \frac{\kappa_s(1 - \eta_v)}{2\,C}\right).$$

This yields a contradiction, since by Lemma 11.14 this iteration then must have been very successful. □

The result of Lemma 11.15 now enables us to deduce that if there are only a finitely many successful iterations, then the iterates for sufficiently large k must be first-order optimal.

lem-tr4

Lemma 11.16. *Suppose that f is twice continuously differentiable and that both the true and the model Hessians remain bounded for all k. Suppose furthermore that there are only finitely many successful iterations. Then $x_k = x^*$ for sufficiently large k and $\nabla f(x^*) = 0$, i.e. Algorithm 1 terminates after finitely many iterations.*

Proof. If the algorithm produces only finitely many successful iterations, then $x_{k_0+j} = x_{k_0+1} = x^*$ for all $j > 0$, where k_0 denotes the last successful iteration. Thus $g_{k_0+j} = g_{k_0+1}$. Moreover, since all subsequent iterations are unsuccessful, the sequence of trust-region radius (Δ_k) converges to zero. However, if $\|g_{k_0+1}\| = \varepsilon > 0$, then this yields a contradiction to Lemma 11.15. Hence $\|g_{k_0+1}\| = 0$. □

Finally we are now in a position to prove our global convergence results.

thm-tr-convergence

Theorem 11.17. *Suppose that $f \in C^2$, and that both the true and model Hessians remain bounded for all k. Then either*

1. *Algorithm 1 terminates after finitely many iterations with a stationary point x^*, i.e.*

$$g(x_l) = 0 \quad \textit{for some} \quad l > 0 \qquad\qquad \textit{or}$$

2. *the objective function is unbounded from below, i.e.*

$$\lim_{k\to\infty} f(x_k) = -\infty \qquad\qquad \textit{or}$$

3. *Algorithm 1 produces an infinite sequence of iterates (x_k) that satisfies*

$$\lim_{k\to\infty} g_k = 0.$$

Proof. If the number of successful iterations is finite then Lemma 11.16 gives the result. We therefore assume that Algorithm 1 produces an infinite number of successful iterations. Let $\mathcal{S}$ be the index set of successful iterations. Furthermore, suppose, that

$$\|g_k\|_2 \geq \epsilon \tag{11.72}$$

thm-tr-convergence1

for some $\epsilon > 0$ and all k. Now consider a successful iteration k, then by Corollary 11.12 and Lemma 11.15 it then follows that

$$f_k - f_{k+1} \geq \eta_s \left(f_k - m_k(s_k)\right) \geq \delta := \frac{1}{2}\,\eta_s\,\epsilon\,\min\left(\frac{\epsilon}{1+C_B}, \kappa_s c_\epsilon\right),$$

with c_ϵ as in Lemma 11.15. Hence, the sum of all successful iterations from 0 to k is

$$f_0 - f_{k+1} = \sum_{\substack{j=0 \\ j \,\in\, \mathcal{S}}}^{k} (f_j - f_{j+1}) \geq \sigma_k\,\delta,$$

where $\sigma_k > 0$ denotes the number of successful iterations up to iteration k. Thus, since δ is a positive constant and

$$\lim_{k\to\infty} \sigma_k = +\,\infty,$$

it follows that (11.72) can only be true if f is unbounded from below.

Conversely, if f is bounded from below, (11.72) cannot be true and there exist a subsequence $(x_k)_{k\in\mathcal{K}}$ with

$$\lim_{\substack{k\to\infty \\ k \,\in\, \mathcal{K}}} \|g(x_k)\| = 0.$$

Suppose there exists another subsequence of successful iterates, that we will index by $(t_i) \subseteq \mathcal{S}$, such that

$$\|g_{t_i}\| \geq 2\epsilon > 0$$

for all i. Moreover define a subsequence $(\ell_i) \subseteq \mathcal{K}$, that consists of the first successful iterations $\ell_i > t_i$ with $\|g_{\ell_i}\| < \epsilon$. Hence we have

$$\|g_k\| \geq \epsilon \qquad \text{for} \quad t_i \leq k < \ell_i \qquad \text{and} \quad \|g_{\ell_i}\| < \epsilon\,. \tag{11.73}$$

thm-tr-convergence-sta

Next, let $\mathcal{J} := \{k \in \mathcal{S} \;:\; t_i \leq k < \ell_i\}$,

then as before it follows that

$$f_k - f_{k+1} \geq \eta_s(f_k - m_k(s_k)) \geq \eta_s \frac{1}{2}\epsilon \min\left(\frac{\epsilon}{1+C_B}, \kappa_s \Delta_k\right) > 0 \qquad (11.74)$$ thm-tr-convergence-sta

holds for all $k \in \mathcal{J}$ because of (11.73). Therefore, since f is bounded from below, we have

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{J}}} \Delta_k = 0$$

and furthermore for $k \in \mathcal{J}$ sufficiently large

$$\Delta_k \leq \frac{2}{\epsilon\eta_s\kappa_s}(f_k - f_{k+1}).$$

We can therefore deduce that

$$\|x_{t_i} - x_{\ell_i}\| \leq \sum_{\substack{j=t_i \\ j\in\mathcal{J}}}^{\ell_i - 1} \|x_j - x_{j+1}\| \leq \sum_{\substack{j=t_i \\ j\in\mathcal{J}}}^{\ell_i - 1} \Delta_j \leq \frac{2}{\epsilon\eta_s\kappa_s}(f_{t_i} - f_{\ell_i})\,.$$

However, since (f_k) is monotonic, the right-hand side must converge to zero for $i \to \infty$. Thus $\|x_{t_i} - x_{\ell_i}\|$ converges to zero and since f twice continuously differentiable, this holds also for $\|g_{t_i} - g_{\ell_i}\|$. This, however, contradicts the definitions of the sequences (t_i) and (ℓ_i), thus no such subsequence (t_i) can exist. □

11.2.9 A Related Alternative: Adaptive Cubic Regularisation

A new related alternative to trust-region methods for unconstrained optimization is the so-called adaptive cubic regularisation that is discussed in detail in [**?**]. The approach is based on the approximate global minimization of a local cubic regularisation of the objective function. Suppose the Hessian $\nabla_{xx}f(x)$ of the objective function f is globally Lipschitz continuous on $\mathbb{R}^n$ with Lipschitz constant L. Then

$$f(x_k + s) \leq m_k^C(s) := f_k + g_k^T s + \frac{1}{2}s^T H(x_k)s + \frac{1}{6}L\|s\|^3 \qquad \text{for all} \quad s \in \mathbb{R}^n\,.$$

Introducing a dynamic positive parameter σ_k instead of the Lipschitz constant L and further allow for a symmetric approximation B_k to the local Hessian $H_k := \nabla_{xx}f(x_k)$we obtain the cubic model function

$$m_k^c(s) := f_k + g_k^T s + \frac{1}{2}s^T B_k s + \frac{1}{3}\,\sigma_k\,\|s\|^3 \qquad (11.75)$$ cubic-model

The rules for updating the parameter σ_k follows the update rules of the trust-region radius. This approach is justified by analogy to trust-region methods, since σ_k might be regarded as the reciprocal of the trust-region radius.

Furthermore, the step s_k is only required to decrease the model as good as that provided by a suitable Cauchy point.

The resulting algorithm is very similar to the basic trust-region algorithm Algorithm 1.

Algorithm 3: Adaptive Regularisation using Cubics (ARC)

Choose an initial vector x_0, an initial σ_0 and update parameters $0 < \eta_s \leq \eta_v < 1$, $\gamma_2 \geq \gamma_1 > 1$;;

for $k = 0, 1, 2, \ldots$ **do**

tr-alg3 (1) Compute (approximately) the solution s_k of

$$\min_{s \in \mathbb{R}^n} \; m_k^c(s)$$

such that $m_k^c(s_k) \leq m_k^c(s_k^C)$, where the Cauchy point s_k^C is determined by

$$s_k^C = -\alpha_k^C g_k \qquad \text{with} \qquad \alpha_k^C = \arg\min_{\alpha \in \mathbb{R}^+} \; m_k^c(-\alpha g_k)$$

;

tr-alg4 (2) Compute ρ_k given by (11.55) (with $m_k(s_k)$ replaced by $m_k^c(s_k)$).;

tr-alg5 (3) Update x_k and σ_k: ;

if $\rho_k > \eta_v$ *(very successful iteration)* **then**
Set $x_{k+1} = x_k + s_k$ and $\sigma_{k+1} \in [0, \sigma_k]$.

else if $\rho_k \geq \eta_s$ *(successful iteration)* **then**
Set $x_{k+1} = x_k + s_k$ and $\sigma_{k+1} \in [\sigma_k, \gamma_1 \sigma_k]$.

else
Set $x_{k+1} = x_k$ and $\sigma_{k+1} \in [\gamma_1 \sigma_k, \gamma_2 \sigma_k]$.

acr-alg

Remark 11.18.

1. *If the for the current iterate $\rho_k < \eta_s$ then, as for the basic trust-region method, the reduction in the objective function is regarded as insufficient and the weight σ_k is increased with the intention to implicitly*

reducing the size of the step in the next iteration. In this way the updating rules for σ_k mimmick those ones of changing the trust-region radius.

2. *As for the definition of the trust-region, the ℓ_2-norm in the definition of the model function $m_k^c(s)$ can be replaced by a more general norm on $\mathbb{R}^n$ of the form $\|s\|_M := \sqrt{s^T M s}$, where M is a given symmetric positive definite matrix.*

3. *The regularisation term $\|s\|^3$ may also be replaced by a term of the form $\|s\|^p$, for some $p > 2$.*

11.2.10 Global Convergence

In the following we will again make use of the set of successful iterations

$$\mathcal{S} := \{k \geq 0 \ : \ k \text{ successful or very successful}\}\,.$$

At first, we derive a guaranteed lower bound on the decrease in f, similar to that one for the trust-region method

acr-lem0 **Lemma 11.19.** *Suppose that the step s_k satisfies $m_k^c(s_k) \leq m_k^c(s_k^c)$. Then for $k \geq 0$, we have that*

$$f_k - m_k^c(s_k) \geq f_k - m_k^c(s_k^C) \geq \frac{\|g_k\|^2}{6\sqrt{2}\,\max(1+\|B_k\|, 2\sqrt{\sigma_k\,\|g_k\|})} = \frac{\|g_k\|}{6\sqrt{2}} \min\left(\frac{\|g_k\|}{1+\|B_k\|}, \frac{1}{2}\sqrt{\frac{\|g_k\|}{\sigma_k}}\right).$$

Proof. See [**?**]. □

Furthermore, in the following we will make the assumption that there exists a constant $C_B \geq 0$ such that

$$\|B_k\| \leq C_B \qquad \text{for all} \quad k \geq 0\,. \tag{11.76}$$ acr-AM1

The following auxiliary Lemma is needed to prove the global convergence results.

acr-lem1 **Lemma 11.20.** *Suppose (11.76) holds for some $C_B \geq 0$ and that $\mathcal{I}$ is an infinite index set such that*

$$\|g_k\| \geq \varepsilon, \quad \text{for all} \quad k \in \mathcal{I} \quad \text{and some} \quad \varepsilon > 0, \quad \text{and} \quad \sqrt{\frac{\|g_k\|}{\sigma_k}} \to 0, \ \text{as } k \to \infty, \quad k \in \mathcal{I}. \tag{11.77}$$ acr-lem1-1

Then

$$\|s_k\| \leq 3\sqrt{\frac{\|g_k\|}{\sigma_k}}, \qquad \textit{for all} \quad k \in \mathcal{I} \quad \textit{sufficiently large.} \tag{11.78}$$ acr-lem1-2

Additionally, if

$$x_k \to x^*, \ \textit{as } k \to \infty, \qquad k \in \mathcal{I}, \ \textit{for some } x^* \in \mathbb{R}^n, \tag{11.79}$$ acr-lem1-3

then each iteration $k \in \mathcal{I}$ that is sufficiently large is very successful, and

$$\sigma_{k+1} \leq \sigma_k, \qquad \textit{for all} \quad k \in \mathcal{I} \quad \textit{sufficiently large.} \tag{11.80}$$ acr-lem1-4

Proof. See [**?**] □

Provided the algorithm produces only fintely many successful iterations, all subsequent iterates are stationary points.

acr-lem2 **Lemma 11.21.** *Suppose (11.76) holds for some $C_B \geq 0$. Moreover, assume that there are only finitely many successful iterations. Then $x_k = x^*$ for all sufficiently large k and $\nabla f(x^*) = 0$.*

Proof. See [**?**]. □

As next, we prove that if the objective function is bounded from below, then Algorithm 3 either produces finitely many successful iteration or it produces an infinite sequence that contains a subsequence $(x_k)_{k\in\mathcal{K}}$ with $\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} g(x_k) = 0$.

acr-thm1 **Theorem 11.22.** *Suppose (11.76) holds for some $C_B \geq 0$. Moreover, assume that f is bounded from below.*

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{11.81}$$ acr-thm1-1

Proof. If there exist only finitely many successful iterations, then the result is true by Lemma 11.21. Therefore we assume that Algorithm 3 produces infinitely many successful iterations. Moreover, we assume that (11.81) does not hold. Hence there exists an $\varepsilon > 0$ such that

$$\|g_k\| \geq \varepsilon \qquad \text{for all} \quad k \geq 0. \tag{11.82}$$ acr-thm1-2

First we will prove that

$$\sum_{k\in\mathcal{S}}^{\infty} \sqrt{\frac{\|g_k\|}{\sigma_k}} < +\infty. \tag{11.83}$$ acr-thm1-3

It follows from Lemma 11.19, (11.76) and (11.82) that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \eta_s(f(x_k) - m_k^c(s_k)) \\ &\geq \frac{\eta_s \varepsilon}{6\sqrt{2}} \min\left(\frac{\epsilon}{1+C_B}, \frac{1}{2}\sqrt{\frac{\|g_k\|}{\sigma_k}}\right) \end{aligned}$$

for all $k \in \mathcal{S}$. However, since the sequence $(f(x_k))$ is monotonically decreasing and f is supposed to bounded from below, it is convergent and thus the minimum on the right-hand side will be attained at the second argument for sufficiently large $k \in \mathcal{S}$. Hence,

$$f(x_k) - f(x_{k+1}) \geq \frac{\eta_s \varepsilon}{12\sqrt{2}} \sqrt{\frac{\|g_k\|}{\sigma_k}}$$

for sufficiently large $k \in \mathcal{S}$. Summing up over all sufficiently large iterations (i.e. choosing k_0 suffienciently large) yields

$$f(x_{k_0}) - f(x_{j+1}) = \sum_{\substack{k=k_0 \\ k\in\mathcal{S}}}^{j} (f(x_k) - f(x_{k+1})) \geq \frac{\eta_s \varepsilon}{12\sqrt{2}} \sum_{\substack{k=k_0 \\ k\in\mathcal{S}}}^{j} \sqrt{\frac{\|g_k\|}{\sigma_k}}, \quad (11.84)$$

for any $j \in \mathcal{S}$. Letting $j \to \infty$ then since $(f(x_k))$ is convergent, (11.83) holds true and

$$\sqrt{\frac{\|g_k\|}{\sigma_k}} \to 0, \qquad k \to \infty, \qquad k \in \mathcal{S} \quad (11.85)$$ acr-thm1-3+

Next we prove that the sequence of iterates (x_k) is Cauchy sequence. By (11.82) and (11.83) the set $\mathcal{S}$ satisfies the conditions of the set $\mathcal{I}$ in Lemma 11.20 and therefore

$$\begin{aligned} \|x_{l+r} - x_l\| &\leq \sum_{k=l}^{l+r-1} \|x_{k+1} - x_k\| = \sum_{k=l}^{l+r-1} \|s_k\| \\ &\leq 3 \sum_{k=l}^{l+r-1} \sqrt{\frac{\|g_k\|}{\sigma_k}} \to 0 \qquad \text{for} \quad l \to \infty (l, r \geq 0)\,. \end{aligned}$$

Since (x_k) is Cauchy sequence $x_k \to x^*$ for some limit point $x^* \in \mathbb{R}^n$. Thus again by Lemma 11.20 all $k \in \mathcal{S}$ are very successful.

Suppose all sufficiently large iterations k are successful, i.e. $k \in \mathcal{S}$ for all $k > k_1$ for some $k_1 \geq 0$. Then Lemma 11.20 implies that $\sigma_{k+1} \leq \sigma_k$ for

all $k > k_1$ and so (σ_k) is bounded above, which contradicts (11.83). Thus (11.82) cannot hold.

It therefore remains to prove that $k \in \mathcal{S}$ for all $k > k_1$ for some $k_1 \geq 0$. Assume that this is not the case. However, since we assumed that $\mathcal{S}$ is infinite, there exits an (infinite) sequence (k_i) of very successful iterations such that $k_i - 1$ is unsuccessful for all $i \geq 0$. Then by our update rules $\sigma_{k_i} \leq \gamma_2 \sigma_{k_i - 1}$ for all i. Moreover, since $k_i - 1$ is unsuccessful, we have that $g_{k_i} = g_{k_i - 1}$ for all i. Thus by (11.85)

$$\sqrt{\frac{\|g_{k_i-1}\|}{\sigma_{k_i-1}}} \to 0, \qquad i \to \infty, \qquad k \in \mathcal{S}$$

Let $\mathcal{I} = \{k_i - 1 \ : \ i \geq 0\}$, then $\mathcal{I}$ satisfies the conditions of Lemma 11.20 and thus $k_i - 1$ is very successful for sufficiently large $i \geq 0$. This, however, contradicts our assumption that $k_i - 1$ is unsuccessful for all i. □

Under the additional assumption that the gradient is uniformly continuous on the sequence of iterates (x_k), it can be shown that the whole sequence of gradients converges to zero.

acr-cor1 **Corollary 11.23.** *In addition to the condition of Theorem 11.22, assume that*

$$\|g_{k_i} - g_{l_i}\| \to 0 \qquad \textit{whenever} \qquad \|x_{k_i} - x_{l_i}\| \to 0, \quad i \to \infty, \tag{11.86}$$ acr-AF2

then

$$\lim_{k \to \infty} \|g_k\| = 0. \tag{11.87}$$ acr-cor1-1

Proof. See [**?**] □

11.2.11 Optimality Conditions for the Minimizer

Next we prove a result concerning the necessary and sufficient optimality conditions for the global minimizer of the cubic model function $m_k^c(s)$ which is very similar to Theorem 11.58, the corresponding result for trust-region subproblem.

In order to state the optimality conditions for a global minimizer s^*, we need the derivatives of $m_k^c(s)$. These may be expressed as

$$\begin{aligned} \nabla_s m_k^c(s) &= g_k + B_k s + \lambda s \\ \nabla_{ss} m_k^c(s) &= B_k + \lambda I + \lambda \left(\frac{s}{\|s\|} \right) \left(\frac{s}{\|s\|} \right)^T, \end{aligned}$$

where $\lambda = \sigma_k \|s\|$.

acr-thm2 **Theorem 11.24.** *Any s^* is a global minimizer of $m_k^c(s)$ over $\mathbb{R}^n$ if and only if it satisfies the system of equations*

$$(B_k + \lambda^* I)s^* = -g_k \tag{11.88}$$ acr-thm2-1

where $\lambda^* = \sigma_k \|s^*\|$ and $B_k + \lambda^* I$ is positive semidefinite. If $B_k + \lambda^* I$ is positive definite, s^* is unique.

Proof. First, suppose that s^* is a global minimizer of $m_k^c(s)$, then

$$\nabla_s m_k^c(s^*) = g_k + B_k s^* + \lambda s^* = 0,$$

which proves (11.88). Moreover, by the second order necessary conditions we have that

$$w^T \left(B_k + \lambda I + \lambda \left(\frac{s}{\|s\|} \right) \left(\frac{s}{\|s\|} \right)^T \right) w \geq 0 \tag{11.89}$$ acr-thm2-2

for all $w \in real^n$. If $s^* = 0$, then (11.89) is equivalent to $\lambda^* = 0$ and B_k being positive semidefinite. Thus we only need to consider minimizers that satisfy $s^* \neq 0$.

Moreover, for all vectors $w \in \mathbb{R}^n$ with $w^T s^* = 0$ (11.89) becomes

$$w^T \left(B_k + \lambda I \right) w \geq 0$$

Hence, suppose that $w^T s^* \neq 0$. Then the line $s^* + \alpha w$ intersects the ball of radius $\|s^*\|$ twice, at s^* and another point $u^* \neq s^*$, such that $\|u^*\| = \|s^*\|$. Let $w = u^* - s^*$, then (since s^* is supposed to be a global minimizer of $m_k^c(s)$) we have

$$\begin{aligned} 0 &\leq m_k^c(u^*) - m_k^c(s^*) \\ &= g_k^T(u^* - s^*) + \frac{1}{2}(u^*)^T B_k u^* - \frac{1}{2}(s^*)^T B_k s^* + \frac{\sigma_k}{3}(\|u^*\|^3 - \|s^*\|^3) \\ &= g_k^T(u^* - s^*) + \frac{1}{2}(u^*)^T B_k u^* - \frac{1}{2}(s^*)^T B_k s^*. \end{aligned}$$

However, by (11.88) we have

$$g_k^T(u^* - s^*) = (s^* - u^*)^T B_k s^* + \lambda^*(s^* - u^*)^T s^*.$$

and since $\|u^*\| = \|s^*\|$

$$(s^* - u^*)^T s^* = \frac{1}{2}(s^*)^T s^* + \frac{1}{2}(u^*)^T u^* - (u^*)^T s^* = \frac{1}{2}(w^*)^T w^*$$

Together with the above inequality this gives

$$\begin{aligned} 0 &\leq \frac{1}{2}\lambda^*(w^*)^T w^* + \frac{1}{2}(u^*)^T B_k u^* - \frac{1}{2}(s^*)^T B_k s^* - (u^*)^T B_k s^* \\ &= \frac{1}{2}(w^*)^T (B_k + \lambda^* I) w^* \end{aligned}$$

thus

$$w^T(B_k + \lambda^* I)w \geq 0 \qquad \text{for all} \quad w \text{ with } w^T s^* \neq 0\,.$$

The uniqueness of s^* in the case that $B_k + \lambda^* I$ is positive definite follows directly from (11.88).

For the reverse direction see [**?**].

□

Remark 11.25.
Note that this result and its proof are very similar to Theorem 11.8. Moreover, if the global solution of the trust-region subproblem satisfies $\|s^*\| = \Delta_k$, then comparing the Theorem results we get $\sigma_k = \lambda^*/\Delta_k$. In view of this property we can interprete the parameter σ_k in Algorithm 3 as inversely proportional to the trust-region radius.

11.2.12 Optimality Conditions

Basic Optimality Conditions for Nonlinear Programming

Consider the general Nonlinear Program (NLP)

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & h(x) = 0 \\ & g(x) \geq 0\,, \end{aligned} \tag{11.90}$$

nlp

where $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^q$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are here assumed to be twice continuously differentiable functions. The following definitions [**?**] are essential in the theory of Nonlinear Programming.

The general stationarity condition for an ordinary NLP of the form (11.90) is

$$\nabla f(x^*)^T d \geq 0 \qquad \forall\, d \in \mathcal{T}(\mathcal{X}, x^*)\,, \tag{11.91}$$

nlpstat-condition

where $\mathcal{X} := \{x \in \mathbb{R}^n : h(x) = 0,\, g(x) \geq 0\}$ denotes the feasible region and $\mathcal{T}(\mathcal{X}, x^*)$ denotes the tangent cone that is defined as follows [**?**]:

tangent-cone

Definition 11.26. *Let $\mathcal{M} \subseteq \mathbb{R}^\ell$ denote a nonempty set and let $x \in \mathcal{M}$. The tangent cone of $\mathcal{M}$ at x is defined by*

$$\mathcal{T}(\mathcal{M}, x) \;=\; \Big\{ d \in \mathbb{R}^\ell \;\Big|\; \exists\, (x^k) \subset \mathcal{M}, \\ \exists\, (\eta_k) \subset \mathbb{R}, \eta_k \searrow 0 :\; x^k \to x \;\; \textit{and} \;\; (x^k - x)/\eta_k \to d \Big\}.$$

Moreover the corresponding normal cone of $\mathcal{M}$ at x is

$$\mathcal{N}(\mathcal{M}, x) = (\mathcal{T}(\mathcal{M}, x))^\circ .$$

Condition (11.91) represents the fact that there exists no feasible descent direction at a local optimum x^* and it is equivalent to

$$-\nabla f(x^*) \;\in\; \mathcal{N}(\mathcal{X}, x^*) . \tag{11.92}$$

nlp-stat-condition-dua

As these two stationarity conditions are difficult to verify they are in particular not well practicable for numerical purposes. Some constraint qualifications (CQ) are therefore typically used to guarantee that the unwieldy tangent cone $\mathcal{T}(\mathcal{X}, x^*)$ can be replaced by the linearized tangent cone

$$\mathcal{T}_{lin}(\mathcal{X}, x^*) := \{ d \in \mathbb{R}^\ell | \nabla h_j(x^*)^T\, d = 0,\; \forall\, j \in I_h(x^*),\; \nabla g(x^*)^T\, d \geq 0,\; \forall\, j \in I_g(x^*) \} ,$$

where

$$I_h(x) = \{ i \in \{1, .., q\} \;:\; h_i(x) = 0 \} , \\ I_g(x) = \{ i \in \{1, .., m\} \;: g_i(x) = 0 \} ,$$

denote the sets of the active constraints in x.

One of the most basic constraint qualification is the so-called *Abadie Constraint Qualification* [**?**].

acq

Definition 11.27. *Let $x^* \in \mathcal{X}$, then x^* is said to satisfy the Abadie Constraint Qualification (ACQ), if $\mathcal{T}(\mathcal{X}, x^*) = \mathcal{T}_{lin}(\mathcal{X}, x^*)$.*

Suppose x^* satisfies the ACQ, then (11.92) can be replaced by

$$-\nabla f(x^*) \;\in\; (\mathcal{T}_{lin}(\mathcal{X}, x^*))^\circ$$

which can then by the Farkas Lemma (Lemma 2.27 in [**?**]) proved to be equal to the KKT-conditions (see Definition 11.30), which mostly form the basis of solution methods and software for NLPs. Since the ACQ is difficult to verify, often some stronger constraint qualifications are used that imply the ACQ and hence the admissibility of the KKT-conditions. Two basic regularity assumptions concerning the feasible region of the NLP that imply the ACQ are:

nlp-licq **Definition 11.28.** *Let x^* be feasible for (11.90), then x^* is said to satisfy the Linear Independence Constraint Qualification (LICQ), if the family*

$$\begin{array}{ll} \nabla h_i(x^*) & i \in \{1, \ldots, q\}, \\ \nabla g_j(x^*) & j \in I_g(x^*) \end{array}$$

is linear independent.

nlp-mfcq **Definition 11.29.** *Let x^* be feasible for (11.90), then x^* is said to satisfy the Mangasarian-Fromowitz Constraint Qualification (MFCQ), if*

1. *the family $\nabla h_i(x^*)$ $i = 1, \ldots, q$ is linear independent and*
2. *there exists a vector $d \in \mathbb{R}^n$ that satisfies the conditions $\nabla g_j(x^*)^T d > 0$ for all $j \in I_g(x^*)$ and $\nabla h_i(x^*)^T d = 0$ for all $i \in \{1, \ldots, q\}$*

It can be proved (see for example [**?**]), that these two constraint qualifications satisfy the implications LICQ $\Rightarrow$ MFCQ $\Rightarrow$ ACQ.

Next we define the *Karush-Kuhn-Tucker (KKT-) conditions* that form a necessary optimality condition [**?**] for Nonlinear Programming problems.

def-nlp-kkt **Definition 11.30.** *Let $x^* \in \mathbb{R}^n$. We call the conditions*

$$\begin{array}{rcll} \nabla f(x^*) - \nabla g(x^*)\lambda^* - \nabla h(x^*)\mu^* & = & 0 & \\ h(x^*) & = & 0 & \\ g(x^*) & \geq & 0 & \\ \lambda^* & \geq & 0 & \\ g_i(x^*)\lambda_i^* & = & 0 & \qquad i = 1, \ldots, m \end{array} \tag{11.93}$$

nlp-kkt

Karush-Kuhn-Tucker (KKT-) conditions. Moreover, if there exist $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^q$, such that (x^*, λ^*, μ^*) satisfies (11.93), then we call x^* a stationary point of (11.90) and the vectors λ^* and μ^* Lagrange multipliers of x^*.

Suppose that a local solution x^* of (11.90) satisfies either LICQ or MFCQ, then the existence of vectors $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^q$, such that (x^*, λ^*, μ^*) satisfies the KKT-conditions form a necessary optimality condition.

thm-nlp-kkt **Theorem 11.31.** *Let $x^* \in \mathbb{R}^n$ be a local solution of (11.90). If x^* satisfies either LICQ or MFCQ, then there exist vectors $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^q$, such that (11.93) is satisfied.*

Proof. See for example Theorem 2.39 and 2.41 in [**?**]. □

Define the sets

$$I_g^+(x,\lambda) = \{i \in \{1,\dots,m\} \,:\, g_i(x) = 0,\ \lambda_i > 0\}\,,$$
$$I_g^0(x,\lambda) = \{i \in \{1,\dots,m\} \,:\, g_i(x) = 0,\ \lambda_i = 0\}$$

and

$$\begin{aligned}\mathcal{S}(x,\lambda) \;=\; \{\quad & d \in \mathbb{R}^n\backslash\{0\} | \\ & \nabla h_i(x)^T d = 0, \quad i \in \{1,\dots,q\}, \\ & \nabla g_j(x)^T d = 0, \quad j \in I_g^+(x,\lambda), \\ & \nabla g_j(x)^T d \geq 0, \quad j \in I_g^0(x,\lambda)\,\}\end{aligned}$$

and let

$$\mathcal{L}(x,\lambda,\mu) = f(x) - \sum_{j=1}^{m} \lambda_j g_j(x) - \sum_{i=1}^{q} \mu_i h_i(x) \tag{11.94}$$

nlp-lagrange

denote the *Lagrangian function* of (11.90), then we can define a standard Second Order Sufficient Condition (SOSC) for x^* to be a local solution of (11.90).

def-nlp-sosc

Definition 11.32. *Let x^* be a stationary point of (11.90) with multipliers λ^* and μ^* and suppose that*

$$d^T \nabla^2_{xx} \mathcal{L}(x^*,\lambda^*,\mu^*)\, d > 0 \quad \forall\, d \in \mathcal{S}(x^*,\lambda^*)\,, \tag{11.95}$$

nlp-sosc

then x^* is said to satisfy the Second Order Sufficient Condition (SOSC) for (11.90).

nlp-sosc-thm

Theorem 11.33. *Let (x^*,λ^*,μ^*) satisfy the KKT-conditions and the SOSC, then x^* is a strict local solution of (11.90).*

Proof. See for example Theorem 2.55 in [**?**]. □

11.3 Sequential Quadratic Programming (SQP)

Sequential Quadratic Programming (SQP) methods are the basis of some of the most effective modern nonlinear programming solvers, as for example `SNOPT, filterSQP, DONLP2` (see also *http://www-neos.mcs.anl.gov/neos*). As the name already reveals, these methods are based on the successive solution of quadratic programs (QP). These QPs form an approximation of the NLP in a current iterate x^k. In each outer iteration of the SQP method the solution of such a QP yield a new search direction d and associated multipliers λ_{qp} and μ_{qp}. A new solution estimate is then obtained by setting $x_{k+1} = x_k + d$ and taking λ_{qp} and μ_{qp} as new multiplier estimates λ_{k+1} and μ_{k+1}, respectively. In general, the quadratic subprograms are solved either by an active set strategy or by Interior Point Methods.

11.3.1 Lagrange-Newton Method

We start our discussion with the Newton method applied to the KKT-conditions of the equality constrained nonlinear optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & h(x) = 0\,. \end{aligned} \tag{11.96}$$

ecnlp

The KKT-conditions of (11.96) are

$$\begin{aligned} \nabla_x \mathcal{L}(x^*, \mu^*) = \nabla f(x^*) - \nabla h(x^*)\mu^* &= 0 \\ h(x^*) &= 0\,. \end{aligned} \tag{11.97}$$

ecnlp-kkt

This is a system in $n+q$ unknowns and $n+q$ equalities. Define the nonlinear function $F(x, \mu)$ by

$$F(x,\mu) := \begin{pmatrix} \nabla_x \mathcal{L}(x,\mu) \\ h(x) \end{pmatrix} \tag{11.98}$$

lnewton-func

then (x^*, μ^*) is a KKT-tupel of (11.96) if and only if it is a root of $F(x, \mu)$. Probably the most famous method to find solutions of $F(x, \mu) = 0$ is Newtons method. However, to be able to apply Newtons method, we need the continuously differentiability of F. We therefore assume that f and h are twice continuously differentiable, since then

$$F'(x,\mu) = \begin{pmatrix} \nabla^2_{xx}\mathcal{L}(x,\mu) & \nabla h(x) \\ \nabla h(x)^T & 0 \end{pmatrix}$$

exists and is continuous. Consider a current iterate (x_k, μ_k), then the Newton step $s_k = (s_k^x, s_k^\mu)$ is determined by the Newton equation

$$F'(x_k, \mu_k) s_k = -F(x_k, \mu_k) \tag{11.99}$$

newton

and the resulting Lagrange-Newton algorithm is

11.3.2 Local Convergence

Since the local convergence properties of the Lagrange-Newton algorithm are determined by those ones of the Newton method applied to arbitrary nonlinear equations $F(x) = 0$, we first review the local convergence properties of the general Newton method.

newton-conv

Theorem 11.34. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and let $\bar{x}$ be such that $F(\bar{x}) = 0$ and $F'(\bar{x})$ is regular. Then there exist a $\delta > 0$ such that for all $x_0 \in B_\delta(\bar{x})$ it holds*

Algorithm 4: Lagrange-Newton Algorithm

lnewton1 1 *Choose initial values for* $x_0 \in \mathbb{R}^n$ *and* $\mu_0 \in \mathbb{R}^q$.;

repeat

lnewton2 2 *Compute the solution* s^k *of*

$$\begin{pmatrix} \nabla^2_{xx}\mathcal{L}(x_k,\mu_k) & \nabla h(x_k) \\ \nabla h(x_k)^T & 0 \end{pmatrix} \begin{pmatrix} s^x_k \\ s^\mu_k \end{pmatrix} = - \begin{pmatrix} \nabla_x \mathcal{L}(x_k,\mu_k) \\ h(x_k) \end{pmatrix}$$

lnewton3 3 *Update the iterate and the multiplier:*;

$x_{k+1} \leftarrow x^k + s^x_k$;
$\mu_{k+1} \leftarrow \mu_k + s^\mu_k$;

lnewton4 4 $k \leftarrow k+1$;

ln-alg **until** (x_k, μ_k) satisfies the KKT conditions.;

1. *the Newton method is well-defined and produces a sequence* (x_k) *that converges to* $\bar{x}$.

2. *The rate of convergence of* (x_k) *is q-superlinear, i.e.*

$$\|x_{k+1} - \bar{x}\|_2 = o(\|x_k - \bar{x}\|_2) \qquad \text{for} \quad k \to \infty\,.$$

3. *If in addition* F' *is locally Lipschitz continuous, then the rate of convergence is even q-quadratic, i.e.*

$$\|x_{k+1} - \bar{x}\|_2 = O(\|x_k - \bar{x}\|_2^2) \qquad \text{for} \quad k \to \infty\,.$$

Proof. See for example [**?**] pages 234-239 □

The next Lemma states conditions such that $F(x,\mu)$ defined by (11.98) is regular in a solution (x^*, μ^*).

ln-lem **Lemma 11.35.** *Let* f *and* h *be twice continuously differentiable and let* (x^*, μ^*) *be such that*

1. $\nabla h_1(x^*), \ldots, \nabla h_q(x^*)$ *are linearly independent, i.e. the LICQ holds in* x^*

2. $s^T \nabla^2_{xx}\mathcal{L}(x^*\mu^*)s > 0$ *holds for all* $s \in \mathbb{R}^n \setminus \{0\}$ *with* $\nabla h(x)^T s = 0$, *i.e. the SOSC holds in* x^*.

Then $F'(x^*, \mu^*)$ is regular.

See for example [**?**] Satz 5.28 As a conclusion, we obtain the following theorem as local convergence result for the Lagrange-Newton Algorithm.

ln-thm **Theorem 11.36.** *Let f and h be twice continuously differentiable and let (x^*, μ^*) be a KKT-pair such that the LICQ and the SOSC are satisfied in x^*, then there exists a $\delta > 0$ such that for all $(x_0, \mu_0) \in B_\delta(x^*, \mu^*)$ it holds*

1. *the Lagrange-Newton algorithm either stops with $(x_k, \mu_k) = (x^*, \mu^*)$ or it produces a sequence (x_k, μ_k) that converges to q-superlinearly to (x^*, μ^*).*
2. *If in addition $\nabla^2 f$ and $\nabla^2 h_j$ are locally Lipschitz continuous in $B_\delta(x^*)$, then the rate of convergence is even q-quadratic.*

Proof. The first part of the Theorem directly follows from Theorem 11.34 and Lemma 11.35. Concerning the second part, it remains to prove that the Lipschitz continuity of $\nabla^2 f$ and $\nabla^2 h_j$ implies the Lipschitz continuity of $F'(x, \mu)$ on $B_\delta(x^*, \mu^*)$. However, it holds

$$\begin{aligned}
\|F'(x,\mu) - F'(x',\mu')\|_2 &\leq (L_f + L_h)\|x - x'\|_2 + \sum_{j=1}^{q} \|\mu_j \nabla^2 h_j(x) - \mu'_j \nabla^2 h_j(x')\|_2 \\
&\leq (L_f + L_h)\|x - x'\|_2 + \sum_{j=1}^{q} |\mu_j| \|\nabla^2 h_j(x) - \nabla^2 h_j(x')\|_2 \\
&\qquad + \sum_{j=1}^{q} |\mu_j - \mu'_j| \|\nabla^2 h_j(x')\|_2 \\
&\leq L \, \|(x,\mu) - (x',\mu')\|_2
\end{aligned}$$

□

The Lagrange-Newton method can be extended to general nonlinear optimization problems (NLP), if one replaces the inequality conditions and the complementarity condition corresponding to the inequality constraints in the KKT conditions with an equality condition. This can be done using a so-called NCP-Function. These are special functions $\phi : \mathbb{R}^2 \to \mathbb{R}$ that satisfy the condition

$$\phi(a, b) = 0 \qquad \Longleftrightarrow \qquad a \geq 0, \quad b \geq 0, \quad \text{and} \qquad ab = 0.$$

The Minimumfunction $\phi(a,b) = \min(a,b)$ is one example of an NCP-function. However, most of the NCP-functions have the property that they are not everywhere continuously differentiable (in particular not in $(a,b) = (0,0)$). Thus in order to apply the Newton method to the resulting equation one needs to undertake further considerations to solve this problem.

11.3.3 Local SQP Method for Equality Constrained NLPs

Consider the Newton equation (11.99) of the Lagrange-Newton method, which can also be formulated by

$$\begin{aligned} \nabla^2_{xx}\mathcal{L}(x_k,\mu_k)s^x_k + \nabla h(x_k)s^\mu_k &= -\nabla_x \mathcal{L}(x_k,\mu_k) \\ \nabla h(x_k)^T s^x_k &= -h(x_k)\,. \end{aligned} \tag{11.100}$$

ecsqp-1

However, if we substitute $\mu_{qp} = \mu_k + s^\mu_k$ in (11.100), then it becomes

$$\begin{aligned} \nabla^2_{xx}\mathcal{L}(x_k,\mu_k)s^x_k + \nabla h(x_k)\mu_{qp} &= -\nabla f(x_k) \\ \nabla h(x_k)^T s^x_k &= -h(x_k)\,. \end{aligned} \tag{11.101}$$

ecsqp-2

which are the KKT-conditions of the equality constrained quadratic program (QP)

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \quad & g_k^T d + \tfrac{1}{2} d^T B_k d \\ \text{subject to} \quad & h(x_k) + \nabla h(x_k)^T d = 0 \end{aligned} \tag{11.102}$$

ecqp

where B_k denotes the matrix $\nabla^2_{xx}\mathcal{L}(x_k,\mu_k)$ itself or an approximation of it.

Next, we state conditions for x_k being a local solution of (11.96) in terms of the solutions d_k and μ_{qp} of (11.102) with $B_k = \nabla^2_{xx}\mathcal{L}(x_k,\mu_k)$.

ecsqp-thm

Theorem 11.37. *Let f and h be twice continuously differentiable and let $x_k \in \mathbb{R}^n$ and $\mu_{qp} \in \mathbb{R}^q$. Then the following two statements are equivalent:*

1. *(x_k, μ_{qp}) is a KKT-pair of (11.96) that satisfies the SOSC.*
2. *$d_k = 0$ is an isolated (strict) local minimum of (11.102) and μ_{qp} is a suitable Lagrange multiplier.*

Proof. (1)⇒(2):
Assume that (x_k, μ_{qp}) is a KKT-pair of (11.96) that satisfies the SOSC. The Lagrangian function of (11.102) is

$$\mathcal{L}^{qp}_k(d,\mu_{qp}) = g_k^T d + \frac{1}{2} d^T B_k d + \mu_{qp}^T (h(x_k) + \nabla h(x_k)^T d)\,.$$

Hence, for the derivatives of $\mathcal{L}_k^{qp}$ we get

$$\begin{aligned}\nabla_d \mathcal{L}_k^{qp}(d, \mu_{qp}) &= g_k + B_k d + \nabla h(x_k)\mu_{qp} = \nabla_x \mathcal{L}(x_k, \mu_{qp}) + B_k d \\ \nabla_{dd}^2 \mathcal{L}_k^{qp}(d, \mu_{qp}) &= B_k = \nabla_{xx}^2 \mathcal{L}(x_k, \mu_{qp}),\end{aligned}$$

such that $(d_k, \mu_{qp}) = (0, \mu_{qp})$ satisfies $\nabla_d \mathcal{L}(x_k, \mu_{qp}) = 0$ and

$$s^T \nabla_{dd}^2 \mathcal{L}_k^{qp}(d, \mu_{qp}) s = s^T \nabla_{xx}^2 \mathcal{L}(x_k, \mu_{qp}) s > 0 \qquad \forall s \in \mathbb{R}^n \backslash \{0\} \;\text{ with }\; \nabla h(x_k)^T s = 0\,.$$

Therefore $(d_k, \mu_{qp}) = (0, \mu_{qp})$ is a strict local minimizer of (11.102), as it satisfies the second order sufficient conditions for (11.102).
(2)⇒(1):
Suppose $d_k = 0$ is an isolated (strict) local minimum of (11.102) and μ_{qp} is a suitable Lagrange multiplier. Then $d_k = 0$ is feasible for (11.102) and thus $h(x_k) = 0$ and x_k is feasible for (11.96). Furthermore, because

$$0 = \nabla_d \mathcal{L}_k^{qp}(0, \mu_{qp}) = \nabla_x \mathcal{L}(x_k, \mu_{qp})$$

(x_k, μ_{qp}) is a KKT-pair of (11.96). To prove the SOSC, let $s \in \mathbb{R}^n \setminus \{0\}$ be in the nullspace of $\nabla h(x_k)$ and define the quadratic function

$$\phi(t) = g_k^T(ts) + \frac{t^2}{2} s^T B_k s\,.$$

Then, since $d_k = 0$ is a strict local minimizer of (11.102), $t = 0$ must be a minimum of a convex function $\phi(t)$. Therefore,

$$0 < \phi''(0) = s^T B_k s = s^T \nabla_{xx}^2 \mathcal{L}(x_k, \mu_{qp}) s$$

and (x_k, μ_{qp}) satisfies the SOSC for (11.96). In particular, x_k is a strict local minimum of (11.96). □

We before we end this section with a reduced SQP algorithm, here we first state the basic local SQP algorithm for equality constraint NLPs.

11.3.4 Reduced SQP Method

One possibility to solve the quadratic programs (11.102) is a so-called null-space approach, where instead of searching a solution in $\mathbb{R}^n$ and introducing linear equality constraints one reduces the problem to the null-space of $A_k^T := \nabla h(x_k)^T$. In doing so, one reduces the dimension of the problem and one gets rid of the equality constraints.

Algorithm 5: Local SQP Algorithm for Equality Constrained NLPs

sqp1 1 *Choose initial values for* $x_0 \in \mathbb{R}^n$, $\mu_0 \in \mathbb{R}^q$.;

repeat

sqp1 2 *Compute local minimizer* d_k *of (11.102) and an associated multiplier* μ_{qp}.;

sqp2 3 *Update the iterate* x_k *and the multiplier:*;

$$x_{k+1} = x_k + d_k \qquad \text{and} \qquad \mu_{k+1} = \mu_{qp}$$

sqp2 4 $k \leftarrow k + 1$;

local-ecnlp-sqp **until** (x_k, μ_k) satisfies the KKT conditions.;

Suppose B_k is positive definite on the null-space of A_k^T and A_k has full column rank. Let u_k be any solution of

$$A_k^T u_k = -h_k \,,$$

where $h_k := h(x_k)$ and let $Z_k \in \mathbb{R}^{n \times r}$ with $r = n - q$ be a matrix whose columns form a basis of the null-space of A_k^T. Then

$$h_k + A_k^T d = 0 \qquad \Longleftrightarrow \qquad \exists v \in \mathbb{R}^r \,:\, d = u_k + Z_k v \,.$$

Moreover, the vector v is uniquely given by

$$v = (Z_k^T Z_k)^{-1} Z_k^T (d - u_k) \,.$$

If we substitute $d = u_k + Z_k v$ in (11.96) we obtain the equivalent reduced problem

$$\min_{v \in \mathbb{R}^r} q_k^r(v) \qquad \text{with} \qquad q_k^r(v) := q_k(u_k + Z_k v) \,, \tag{11.103}$$

redqp

where $q_k(d) = g_k^T d + \frac{1}{2} d^T B_k d$.

redqp-lem **Lemma 11.38.** *The point* $d_k \in \mathbb{R}^n$ *is a solution of (11.102) if and only if*

$$v_k = (Z_k^T Z_k)^{-1} Z_k^T (d_k - u_k)$$

is a solution of (11.103).

Proof. Let d_k be a solution of (11.102) and let $v_k = (Z_k^T Z_k)^{-1} Z_k^T (d_k - u_k)$. Then we have

$$d_k = u_k + Z_k v_k.$$

Moreover, for any $v \in \mathbb{R}^r$ the vector $d = u_k + Z_k v$ is feasible for (11.102), since $h_k + A_k^T d = 0$. Therefore, v_k is a local solution of (11.103), since

$$q_k^r(v) = q_k(d) \geq q_k(d_k) = q_k^r(v_k)\,.$$

On the other hand, if v_k is a solution of (11.103) and suppose $d_k = u_k + Z_k v_k$, then $h_k + A_k^T d_k = 0$. Let

$$v = (Z_k^T Z_k)^{-1} Z_k^T (d - u_k)$$

for any $d \in \mathbb{R}^n$. Then $d = u_k + Z_k v$ and it follows that d_k is a local solution of (11.102), since

$$q_k(d) = q_k^r(v) \geq q_k^r(v_k) = q_k(d_k)\,.$$

□

Remark 11.39. *The first and second order derivative of the reduced function $q_k^r(v)$ are given by*

$$\nabla q_k^r(v) = Z_k^T \nabla q_k(u_k + Z_k v) = Z_k^T (g_k + B_k(u_k + Z_k v))\,, \qquad \nabla^2 q_k^r(v) = Z_k^T\, B_k Z_k\,. \tag{11.104}$$

deriv-red

Hence the Hessian of $q_k^r(v)$ is positive definite and the solution v_k of (11.103) can be determined by the equation $\nabla q_k^r(v) = 0$. Moreover, by Lemma 11.38 the solution of (11.102) is then $d_k = u_k + Z_k v_k$, where u_k solves $A_k^T u_k = -h_k$, and the corresponding Lagrange multipier μ_{qp} is the solution of the multiplier rule $B_k d_k + A_k \mu_{qp} = -g_k$.

Next, note that since $q_k^r(v)$ is a quadratic function, the second order Taylor approximation is exact and therefore

$$q_k^r(v) = q_k(u_k) + (Z_k^T (g_k + B_k\, u_k))^T v + \frac{1}{2} v^T Z_k^T\, B_k Z_k v\,.$$

Furthermore, if the KKT-pair (x^*, μ^*) satisfies the SOSC, then for any $\bar{Z}$ that consists of a basis of the nullspace of $\nabla h(x^*)^T$, then the reduced Hessian of the Lagrangian $\bar{Z}^T \nabla^2_{xx} \mathcal{L}(x^*, \mu^*) \bar{Z}$ is positive definite, such that for any (x_k, μ_k) that is close enough to (x^*, μ^*), a positive definite approximation M_k

is assumed to be a better approximation of $Z_k^T \mathcal{L}(x_k, \mu_k) Z_k$ then of $\mathcal{L}(x_k, \mu_k)$. Therefore it seems to be appropriate to minimize

$$\tilde{q}_k^r(v) = q_k(u_k) + (Z_k^T(g_k + B_k\, u_k))^T v + \frac{1}{2} v^T M_k v$$

instead of the exact reduced function $q_k^r(v)$. However, we still have to deal with the so-called cross term $c_k = Z_k^T\, B_k\, u_k$ that involves the matrix B_k. In the following, we make the following assumptions:

$$\begin{aligned} \text{rank}\nabla h(x^*) &= q && (11.105)\\ \|M_k^{-1}\|, \|B_k\| &\leq C_H && (11.106)\\ \|Z_k - \bar{Z}\|_2 &= O(\|x_k - x^*\|_2)\,, && (11.107) \end{aligned}$$

for some constant $C_H \geq 0$. Then $u_k \in \mathbb{R}^q$ can be chosen such that

$$\|u_k\|_2 = O(\|h_k\|_2)\,. \qquad (11.108) \quad \boxed{\texttt{uk-cond}}$$

Therefore $\|Z_k^T\, B_k\, u_k\|_2 = O(\|h_k\|_2)$ and thus in the following we will assume that the Cross-Term c_k satisfies

$$\|c_k\|_2 = O(\|h_k\|_2)\,. \qquad (11.109) \quad \boxed{\texttt{ck-cond}}$$

Since the constant term $q_k(u_k)$ is irrelevant for the minimization of $\tilde{q}_k^r$ we consider the minimzation problem

$$\min_{v \in \mathbb{R}^r} \hat{q}_k^r(v) \qquad \text{with} \qquad \hat{q}_k^r(v) := (Z_k^T g_k + c_k)^T v + \frac{1}{2} v^T M_k v\,, \qquad (11.110) \quad \boxed{\texttt{red-qp}}$$

which can exactly be solved by

$$M_k v_k = -(Z_k^T g_k + c_k)\,.$$

Since (11.110) is an unconstrained optimization, we do not have a direct condition to determine the associate multiplier. However, we can again use the multiplier rule of the original problem, which leads to the equation

$$Y_k^T A_k \mu_{k+1} = -Y_k^T (g_k + B_k d_k)\,,$$

where the columns of Y_k consist of a basis of the range of A_k (e.g. $Y_k = A_k$). On the right-hand side, though, we obtain again a term involving the matrix B_k that we would like to ommit. However, as it can be shown that this term is of the order $O(\|x_k - x^*\|_2)$ and we suppose that x_k is close to x^* it is admissible to ommit this term. Up to now, the new multiplier μ_{k+1}

depends on the "old" values g_k and A_k although the new iterate x_{k+1} is already available (and therefore g_{k+} and A_{k+1}), thus we could instead use these values in our update of μ_{k+1}. Combining both alternatives yields the update

$$Y_{k+1}^T A_{k+1} \mu_{k+1} = -Y_{k+1} g_{k+1} \,. \tag{11.111}$$

red-qp-mult

Together this yields the following reduced local SQP Method. Finally, a

Algorithm 6: Local Reduced SQP Algorithm

sqp1 1 *Choose initial values for $x^0 \in \mathbb{R}^n$, $c_0 \in \mathbb{R}^r$ and a positive definite matrix M_0;*

for $k = 0, 1, 2, \ldots$ **do**

alg:lrsqp-1 2 *Compute multiplier μ_k according to (11.111)* ;

alg:lrsqp-2 3 **if** *(x_k, μ_k) satisfies the KKT conditions.* **then** STOP.

alg:lrsqp-3 4 *Determine Z_k, whose columns consists of a basis of A_k^T* ;

alg:lrsqp-4 5 *Compute u_k that solves $A_k^T u_k = -h_k$.* ;

alg:lrsqp-5 6 *Compute v_k as a solution of (11.110).* ;

alg:lrsqp-6 7 *Set $d_k = u_k + Z_k v_k$ and $x_{k+1} = x_k + d_k$.* ;

alg:lrsqp-7 8 *Update the matrix M_k and the cross term c_k that satisfies (11.109).*;

alg:local-red-sqp

local convergence theorem for Algorithm 6 is :

local-red-sqp-thm

Theorem 11.40. *Let (x^*, μ^*) be a KKT-pair of (11.96) and suppose the LICQ and the SOSC holds for (11.96) in (x^*, μ^*). Consider Algorithm 6 and let $M_k = Z_k^T \nabla^2 \mathcal{L}(x_k, \mu_k) Z_k$. Moreover, assume that the conditions (11.107), (11.108) and (11.109) are satisfied. Then there exists a $\delta > 0$ such that for all $x_0 \in B_\delta(x^*)$ the Algorithm produces a sequence (x_k) that converges 2-step q-superlinearly to x^*, i.e.*

$$\|x_{k+1} - x^*\|_2 = o(\|x_{k-1} - x^*\|_2) \qquad \text{for} \quad k \to \infty \,.$$

Furthermore, the sequence (μ_k) converges 2-step r-superlinearly to μ^*. If $\nabla^2 f$ and $\nabla^2 h_j$ are Lipschitz continuous in $B_\delta(x^*)$, then the rate of convergence changes to 2-step q-quadratically and 2-step r-quadratically for (x_k) and μ_k, respectively.

11.3.5 Local SQP Method for general NLPs

In this section we will now extend the local SQP method of the previous section to general NLPs of the form (11.90). This can easily be done by

extending the equality constraint, quadratic program (11.102). We just add the linearized inequality constraints which then yields the general quadratic subproblem

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \quad & g_k^T d + \tfrac{1}{2} d^T B_k d \\ \text{subject to} \quad & h(x_k) + \nabla h(x_k)^T d = 0 \\ & g(x_k) + \nabla g(x_k)^T d \geq 0 \end{aligned} \tag{11.112}$$

qp-sqp

As a general local SQP algorithm we then obtain Algorithm 7 A main local

Algorithm 7: Local SQP Algorithm

sqp1 1 *Choose initial values for* $x_0 \in \mathbb{R}^n$, $\lambda_0 \in \mathbb{R}^m$ *and* $\mu_0 \in \mathbb{R}^q$.;

repeat

sqp1 2 *Compute a local minimizer* d_k *of (11.112) that is closest to the origin and compute associated multipliers* λ_{qp}^k *and* μ_{qp}^k.;

sqp2 3 *Update the iterate and the multipliers:*;

$x_{k+1} \leftarrow x_k + d_k$;
$\lambda_{k+1} \leftarrow \lambda_{qp}^k$;
$\mu_{k+1} \leftarrow \mu_{qp}^k$;

sqp2 4 $k \leftarrow k+1$;

until (x_k, λ_k, μ_k) satisfies the KKT conditions.;

alg:genlocal-sqp

convergence result for Algorithm 7 (Theorem 15.2.2 in [**?**]) is:

sqp-allg-conv

Theorem 11.41. *Suppose that the second derivatives of* f, g *and* h *exist and are Lipschitz continuous in some neighbourhood* Ω *of a stationary point* x^* *of (11.90) with multipliers* λ^* *and* μ^*. *Assume that the LICQ and the SOSC hold in* (x^*, λ^*, μ^*) *and furthermore that* $\lambda_j^* \neq 0$ *for all* $j \in I_g(x^*)$. *Then the following holds:*

1. *Consider any sequence* (λ_k, μ_k) *converging to* (λ^*, μ^*). *Then there exists a neighbourhood* $\mathcal{X} \subset \Omega$ *of* x^* *for which the sequence* (x^k) *generated by Algorithm 3.1 converges q-superlinearly to* x^* *from any starting point* $x_0 \in \mathcal{X}$. *Furthermore, if*

$$\|(\lambda_k, \mu_k) - (\lambda^*, \mu^*)\| = O(\|x_k - x^*\|),$$

then the convergence is q-quadratic.

2. *Let (x_k) and (d_k) be the sequences generated by Algorithm 3.1 and let $(\lambda_{k+1}, \mu_{k+1})$ be the Lagrange multipliers associated with d_k. Then there is a neighbourhood $\mathcal{X} \subset \Omega$ of x^* and another neighbourhood $\mathcal{Y}$ of (λ^*, μ^*) for which the sequence $((x_k, \lambda_k, \mu_k))$ converges q-quadratically to (x^*, λ^*, μ^*) from any starting point $((x_0, \lambda_0, \mu_0)) \in \mathcal{X} \times \mathcal{Y}$.*
3. *In either case, the set of constraints that are active at x^* are precisely those that are active for the quadratic subproblem (11.112) at d_k for large enough k.*

Remark 11.42. *A similar local convergence result (q-superlinear/quadratic convergence of the KKT-sequence (x_k, λ_k, μ_k)) can be shown by the Newton method applied to the nonlinear system of equations:*

$$F(x, \lambda, \mu) := \begin{pmatrix} \nabla_x \mathcal{L}(x, \lambda, \mu) \\ h(x) \\ \Phi(-\lambda, g(x)) \end{pmatrix} = 0\,, \tag{11.113}$$

sqp-prf-func

where $\Phi(-\lambda, g(x)) := (\phi(\lambda_1, g_1(x)), \ldots, \phi(\lambda_m, g_m(x)))$ and $\phi(a, b)$ denotes an NCP-function (see Remark in the previous section), e.g. $\phi(a, b) = \min(a, b)$ - for the proof see for example [?] p.246-249.

11.3.6 Elastic Mode

One question that we have not considered so far concerns the feasible sets of the QP subproblems. First, if the NLP is convex, then we can prove that the feasible sets of (11.112) are nonempty.

lem:sqp-feasible

Lemma 11.43. *Suppose the feasible set of (11.90) is nonempty and the functions h_i are affine-linear $(i = 1, \ldots, q)$ and the functions g_j are concave $(j = 1, \ldots, m)$ then the feasible set of the quadratic subproblems (11.112) are also nonempty.*

Proof. Let $\tilde{x}$ be a feasible point of (11.90), then define $d_k = \tilde{x} - x_k$. Then since all $-g_j$ are assumed to be convex, it holds

$$-g_j(x_k) - \nabla g_j(x_k)^T d_k = -g_j(x_k) - \nabla g_j(x_k)^T(\tilde{x} - x_k) \leq -g_j(\tilde{x}) \leq 0\,,$$

for all $(j = 1, \ldots, m)$. Moreover, since all h_i are affine

$$h_i(x_k) + \nabla h_i(x_k)^T d_k = h_i(x_k) + \nabla h_i(x_k)^T(\tilde{x} - x_k) = h_i(\tilde{x}) = 0\,,$$

for all $(i = 1, \ldots, q)$. Hence d_k is feasible for (11.112). □

However, in the general, nonconvex case, the quadratic subproblems might have an empty feasible set. One workaround to this problem is to consider the bounds of the feasible sets to be elastic by means of introducing auxiliary (slack) variables. Moreover, we append a sort of penalty term to the objective function, such that the relaxation of the constraints is only as large as is needed. The modified quadratic subproblem then takes the form

$$\begin{aligned} \min_{d,\xi,\eta^+,\eta^-} \quad & g_k^T d + \tfrac{1}{2} d^T B_k d + \alpha \left(\sum_{i=1}^m \xi_i + \sum_{j=1}^q \eta_j^+ + \sum_{j=1}^q \eta_j^- \right) \\ \text{subject to} \quad & h(x_k) + \nabla h(x_k)^T d = \eta^+ - \eta^- \\ & g(x_k) + \nabla g(x_k)^T d \geq -\xi \\ & \xi,\, \eta^+,\, \eta^- \geq 0\,. \end{aligned} \tag{11.114}$$

qp-sqp-mod

Some properties of the modified subproblem are summarized in the following lemma.

Lemma 11.44.

1. *The feasible set of the QP (11.114) is nonempty. Moreover, if B_k is symmetric and positive definite, then (11.114) has a solution.*

2. *The vector $d \in \mathbb{R}^n$ is feasible for (11.112) if and only if $(d, 0, 0, 0) \in \mathbb{R}^{n+2m+q}$ is feasible for (11.114).*

3. *Let B_k be symmetric and positive definite. If $d \in \mathbb{R}^n$ is a solution of (11.112) with multipliers $(\lambda_{k+1}, \mu_{k+1})$ and it holds*

$$\alpha \geq \max\{\lambda_{1,k+1}, ..., \lambda_{m,k+1}, |\mu_{1,k+1}|, ..., |\mu_{q,k+1}|\}\,,$$

then $(d, 0, 0, 0)$ is a solution of (11.114). Conversely, if $(d, 0, 0, 0)$ is a solution of (11.114), the $d \in \mathbb{R}^n$ is a solution of (11.112) with multipliers $(\lambda_{k+1}, \mu_{k+1})$ that satisfy the inequality.

Proof. See Exercise or [**?**] Lemma 5.41. □

In Subsection 11.3.9 we will discuss an associated algorithm of this modification in combination with a line-search approach as globalization strategy.

11.3.7 Globalized SQP Methods

As we have seen in the previous section, the local SQP algorithm can only be guaranteed to yield a convergent sequence of iterates, if the initial point $(x_0,\, \lambda_0,\, \mu_0)$ is close enough to a solution of the NLP (see Theorem 11.41),

most SQP algorithms incorporate a globalisation strategy. The most popular approaches to promote global convergence of an SQP algorithm concern line-search methods applied to a suitable penalty or merit function, trust-region approaches or most recently filter methods. In this section, we will briefly discuss all these globalization strategies.

11.3.8 Penalty Methods

Before we start our discussion of a globalized SQP method that uses a penalty function, we first briefly review some properties of penalty function methods.

Penalty methods are based on so-called penalty functions. Instead of dealing with the constraints separately, one adds a penalty term, that depends on a penalty parameter α, to the original objective function which then yields the penalty function $P(x;\alpha)$ and minimizes the unconstrained problem

$$\min_{x\in\mathbb{R}^n} P(x;\alpha)\,. \tag{11.115}$$

penalty

The question that arises with this approach how to choose the penalty term, i.e. which type of function should we use and how large do we have to choose α, such that the solution of (11.115) corresponds to the solution of the original, constrained problem. A special class of penalty functions are the so-called exact penalty functions. Consider the general constrained optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{s.t.} \quad x\in X\,, \tag{11.116}$$

genproblem

with $X=\{x\in\mathbb{R}^n \,:\, h(x)=0,\, g(x)\le 0\}$, then the definition of an exact penalty function for (11.116) is as follows.

exact-penalty

Definition 11.45. *A penalty function of the form*

$$P_r(x;\alpha) := f(x) + \alpha\, r(x)$$

where $r:\mathbb{R}^n\to\mathbb{R}$ denotes a continuous function that satisfies

$$r(x)\ge 0 \qquad \forall\, x\in\mathbb{R}^n \qquad \textit{and} \qquad r(x)=0 \quad\Leftrightarrow\quad x\in X\,,$$

is called exact in a local minimizer x^* of (11.116) if there exist a finite parameter $\bar\alpha>0$ such that x^* is also a local minimizer of $P_r(x;\alpha)$ for all $\alpha\ge\bar\alpha$.

Unfortunately, it can be proved that if $P_r(x;\alpha)$ is exact, the function $r(x)$ is in general nondifferentiable in x^* (which poses diffiulties to the direct application of unconstrained optimization algorithms).

thm:exact-penalty1

Theorem 11.46. *Let x^* be a local minimum of (11.116) with $\nabla f(x^*) \neq 0$. Suppose that the penalty function $P_r(x;\alpha)$ is exact in x^*. Then $r(x)$ is not differentiable in x^*.*

Proof. Exercise or [**?**] Lemma 5.9. □

One class of exact (nondifferentiable) penalty function for (11.116) can be obtained using the ℓ_q-norm as penalty term:

$$r_q(x) := \|(g(x)_+, h(x))\|_q \qquad \text{with} \quad (g_i(x))_+ := \max(0, g_i(x))\,. \tag{11.117}$$

lq-function

The presumably most popular, classical penalty function is the exact ℓ_1-penalty function

$$P_1(x;\alpha) := f(x) + \alpha \sum_{i=1}^{q} |h_i(x)| + \alpha \sum_{j=1}^{m} \max(0, g_j(x))\,. \tag{11.118}$$

l1-penalty

The next result proves the exactness of this penalty function for convex NLPs in solutions x^* that satisfy some CQ such that the KKT conditions are necessary.

thm:exact-penalty2

Theorem 11.47. *Let (x^*, λ^*, μ^*) be a KKT point of (11.116), where f and g_i $(i = 1, .., m)$ are supposed to be convex, and h_j $(j = 1, ..., q)$ to be affine linear. Then $P_1(x;\alpha)$ defined in (11.118) is exact in x^*.*

Proof. Since each KKT-point of a convex NLP is a saddlepoint of the Lagrangian function, we have

$$\mathcal{L}(x^*, \lambda^*, \mu^*) \leq \mathcal{L}(x, \lambda^*, \mu^*) \qquad \forall\, x \in \mathbb{R}^n\,.$$

Set $\bar{\alpha} := \|(\lambda^*, \mu^*)\|_\infty$ and let $\alpha \geq \bar{\alpha}$. Then

$$\begin{aligned}
P_1(x^*, \alpha) &= f(x^*) + \alpha \sum_{i=1}^{q} |h_i(x^*)| + \alpha \sum_{j=1}^{m} \max(0, g_i(x^*)) \\
&= f(x^*) \\
&= f(x^*) + \sum_{i=1}^{q} \mu_i^* \, h_i(x^*) + \sum_{j=1}^{m} \lambda_j^* g_j(x^*) \\
&\leq f(x) + \sum_{i=1}^{q} \mu_i^* \, h_i(x) + \sum_{j=1}^{m} \lambda_j^* g_j(x) \\
&\leq f(x) + \sum_{i=1}^{q} \mu_i^* \, |h_i(x)| + \sum_{j=1}^{m} \lambda_j^* \max(0, g_j(x)) \\
&\leq f(x) + \bar{\alpha} \sum_{i=1}^{q} |h_i(x)| + \bar{\alpha} \sum_{j=1}^{m} \max(0, g_i(x)) \\
&\leq f(x) + \alpha \sum_{i=1}^{q} |h_i(x)| + \alpha \sum_{j=1}^{m} \max(0, g_i(x)) \\
&= P_1(x, \alpha) \, .
\end{aligned}$$

Hence x^* is in fact a global minimizer of $P_1(x, \alpha)$ for all $\alpha \geq \bar{\alpha}$. □

A more general result concerning the exactness of ℓ_q-penalty functions is the following, which can be found in [**?**].

thm:exact-penalty3

Theorem 11.48. *Let x^* be a strict local minimum of (11.116) that satisfies the MFCQ. Then for every $q \in [1, \infty]$ there exist an $\bar{\alpha}_q > 0$ such that x^* is a local minimum of $P_q(x, \alpha)$ for all $\alpha \geq \bar{\alpha}_q$, i.e. $P_q(x, \alpha)$ is exact in x^*.*

sqp-line

11.3.9 Line-Search Approach

In this section we will now use the exact ℓ_1-penalty function $P_1(x; \alpha)$ in order to globalize our local SQP method (in this subsection applied to (11.116)). We will see that under suitable conditions the solutions d_k of the SQP subproblems are descent directions $P_1(x; \alpha)$. Hence we can use an Armijo-like stepsize $\sigma_k > 0$ and obtain a descent method applied to the penalty function.

However, in the previous subsection we have seen that $P_1(x; \alpha)$ is in general not differentiable everywhere. Hence, we first have to consider how to decide if d is a descent direction for the penalty function, before we can apply some sort of descent algorithm to (11.115).

thm:global-sqp1

Theorem 11.49. *The directional derivative of the exact ℓ_1-penalty function $P_1(x;\alpha)$ (defined by (11.118)) in some point x in the direction d is given by*

$$\begin{aligned} P_1'(x;\alpha;d) &= \nabla f(x)^T d + \alpha \sum_{i:g_i(x)>0} \nabla g_i(x)^T d + \alpha \sum_{i:g_i(x)=0} \max(0, \nabla g_i(x)^T d) \\ &\quad + \alpha \sum_{j:h_j(x)>0} \nabla h_j(x)^T d - \alpha \sum_{j:h_j(x)<0} \nabla h_j(x)^T d + \alpha \sum_{j:h_j(x)=0} |\nabla h_j(x)^T d| \,. \end{aligned}$$

Proof. See [**?**] p.250-252. □

This result can now be used to prove that the SQP direction d_k (the solution of the subproblems) is a descent direction of $P_1(x;\alpha)$.

thm:global-sqp2

Theorem 11.50. *Let $d_k \neq 0$ be the solution of the quadratic subproblem*

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \quad & g_k^T d + \tfrac{1}{2} d^T B_k d \\ \textit{subject to} \quad & h(x_k) + \nabla h(x_k)^T d = 0 \\ & g(x_k) + \nabla g(x_k)^T d \leq 0 \end{aligned} \tag{11.119}$$

glob-sqp-qp

with a symmetric and positive definite matrix B_k and let λ_{k+1} and μ_{k+1} be the associated multipliers. Moreover, assume that the penalty parameter satisfies $\alpha \geq \|(\lambda_{k+1}, \mu_{k+1})\|_\infty$. Then

$$P_1'(x_k;\alpha;d_k) \leq -d_k^T B_k d_k < 0\,,$$

i.e. d_k is a descent direction of $P_1'(x_k;\alpha;d_k)$ in x_k.

Proof. Since $(d_k, \lambda_{k+1}, \mu_{k+1})$ satisfies the KKT-conditions of the quadratic subproblem, it holds

$$\lambda_{i,k+1}(g_i(x_k) + \nabla g_i(x_k)^T d_k) = 0$$

for all $i = 1, ..., m$ adding this term to the directional derivative and substitute

$$\nabla h_j(x_k)^T d_k = -h_j(x_k)$$

we obtain

$$\begin{aligned}
P_1'(x_k;\alpha;d_k) &= \nabla f(x_k)^T d_k + \sum_{i=1}^{m} \lambda_{i,k+1} \nabla g_i(x_k)^T d_k + \sum_{i=1}^{m} \lambda_{i,k+1} g_i(x_k) \\
&\quad + \alpha \sum_{i:g_i(x_k)>0} \nabla g_i(x_k)^T d_k + \alpha \sum_{i:g_i(x_k)=0} \max(0, \nabla g_i(x_k)^T d_k) \\
&\quad - \alpha \sum_{j:h_j(x_k)>0} h_j(x) + \alpha \sum_{j:h_j(x_k)<0} h_j(x) \\
&\leq \nabla f(x_k)^T d_k + \sum_{j=1}^{m} \lambda_{k+1} \nabla g_i(x_k)^T d_k + \sum_{j=1}^{m} \lambda_{k+1}\, g_i(x_k) \\
&\quad - \alpha \sum_{i:g_i(x_k)>0} g_i(x_k) - \alpha \sum_{j:h_j(x_k)>0} h_j(x_k) + \alpha \sum_{j:h_j(x_k)<0} h_j(x_k)
\end{aligned}$$

since by $g(x_k)+\nabla g(x_k)^T d_k \leq 0$ it follows that $\max(0, \nabla g_i(x)^T d_k) = 0$ for all i with $g_i(x_k) = 0$. Moreover, by the multiplier rule of the KKT conditions for the quadratic subproblem, we can substitute $\nabla f(x_k)^T d_k$

$$\nabla f(x_k)^T d_k = -d_k^T B_k d_k - \sum_{i=1}^{m} \lambda_{i,k+1} \nabla g_i(x_k)^T d_k - \sum_{j=1}^{q} \mu_{i,k+1} \nabla h_j(x_k)^T d_k$$

which yields

$$\begin{aligned}
P_1'(x_k;\alpha;d_k) &\leq -d_k^T B_k d_k - \sum_{i=1}^{m} \lambda_{i,k+1} \nabla g_i(x_k)^T d_k - \sum_{j=1}^{q} \mu_{i,k+1} \nabla h_j(x_k)^T d_k \\
&\quad + \sum_{j=1}^{m} \lambda_{k+1} \nabla g_i(x_k)^T d_k + \sum_{j=1}^{m} \lambda_{k+1}\, g_i(x_k) \\
&\quad - \alpha \sum_{i:g_i(x_k)>0} g_i(x_k) - \alpha \sum_{j:h_j(x_k)>0} h_j(x_k) + \alpha \sum_{j:h_j(x_k)<0} h_j(x_k) \\
&\leq -d_k^T B_k d_k + \sum_{i:g_i(x_k)>0} (\lambda_{i,k+1} - \alpha) g_i(x_k) + \sum_{i:g_i(x_k)\leq 0} \lambda_{i,k+1} g_i(x_k) \\
&\quad + \sum_{j:h_j(x_k)>0} (\mu_{i,k+1} - \alpha) h_j(x_k) + \sum_{j:h_j(x_k)>0} (\mu_{i,k+1} + \alpha) h_j(x_k) \\
&\leq -d_k^T B_k d_k\,,
\end{aligned}$$

since $\lambda_{i,k+1} g_i(x_k) \leq 0$ for all i with $g_i(x_k) \leq 0$ and by $\alpha \geq \bar{\alpha} \geq \|(\lambda_{k+1}, \mu_{k+1})\|_\infty$ it follows that the remaining terms of $P_1'(x_k; \alpha; d_k)$ are all nonpositive. □

Theorem 11.50 implies that we can use d_k as a descent direction for $P_1(x; \alpha)$. In combination with a stepsize algorithm, we therefore obtain a globally convergent algorithm. Furthermore, if the iterates $(x_k \lambda_k, \mu_k)$ are close enough to the solution $(x^* \lambda^*, \mu^*)$ the local convergence properties of the local SQP algorithm come into play. The resulting globally convergent algorithm is displayed in Algorithm 8.

Algorithm 8: Global Line-Search SQP Algorithm

gsqp1 1 *Choose initial values for* $(x_0, \lambda_0, \mu_0) \in \mathbb{R}^{n+m+q}$*, a positive definite matrix* B_0 *and parameter* $\beta \in (0,1)$,$\sigma \in (0,1)$.;

for $k = 0, 1, 2, \ldots$ **do**

alg:gsqp-2 2 **if** (x_k, λ_k, μ_k) *satisfies the KKT conditions.* **then**
STOP.

alg:gsqp-3 3 *Compute a solution* d_k *of (11.119) and associated multipliers* λ_{k+1}, μ_{k+1}. ;

alg:gsqp-4 4 **if** $d_k = 0$ **then**
STOP.

alg:gsqp-5 5 *Compute stepsize* $t_k = \max\{\beta^\ell : \ell = 0, 1, 2, ...\}$ *such that*

$$P_1(x_k + t_k d_k; \alpha) - P_1(x_k; \alpha) \leq \sigma\, t_k\, P_1'(x_k; \alpha; d_k)\,.$$

;

alg:gsqp-6 6 *Set* $x_{k+1} = x_k + t_k d_k$. ;

alg:gsqp-7 7 *Update the matrix* B_k.;

alg:global-sqp

Remark 11.51. *In a practical method, one would also include an update strategy for the penalty parameter* α.

Similarly, we can globalize the modified SQP method that uses the elastic mode. As a convergence result we obtain for a corresponding algorithm:

thm:global-sqp3 **Theorem 11.52.** *Let* (x_k) *be the sequence produced by the global modified SQP algorithm. Suppose the symmetric matrices* B_k *satisfy*

$$c_1 \|d\|^2 \leq d^T B_k d \leq c_2 \|d\|^2 \qquad \forall\, d \in \mathbb{R}^n,\, k \in \mathbb{N}$$

for some constants $c_1, c_2 > 0$. *Then, every accumulation point of* (x_k) *is a stationary point of* $P_1(\cdot; \alpha)$.

Proof. See [**?**]. □

11.3.10 The Maratos Effect

In the following are concerned with a numerical effect that is due to N. Maratos. Before, we just stated that in a local neighbourhood of the solution, the local convergence properties of the SQP method come into play. This, however, is only true if for sufficiently large $k \in \mathbb{N}$ always the stepsize $t_k = 1$ is chosen. In particular, we therefore need that

$$P_1(x_k + d_k; \alpha) < P_1(x_k; \alpha)\,.$$

Unfortunately, N. Maratos showed in his PhD-thesis in 1978 that this is not always true.

Consider e.g. the following example:

$$\min f(x,y) = 2(x^2+y^2-1) - x \qquad \text{subject to} h(x,y) = x^2+y^2-1 = 0$$

As derivatives we then have:

$$\nabla f(x,y) = \begin{pmatrix} 4x-1 \\ 4y \end{pmatrix}, \qquad \nabla h(x,y) = 2\begin{pmatrix} x \\ y \end{pmatrix} \tag{11.120}$$

$$\nabla^2 f(x,y) = 4I \qquad \nabla^2 h(x,y) = 2I, \qquad \Rightarrow \nabla^2 \mathcal{L}(x,y,\mu) = (4+2\mu)I. \tag{11.121}$$

Moreover, for all feasible $(x,y) \in \mathbb{R}^2$ we know that $\|(x,y)\| = 1$ such that

$$f(x,y) = 2h(x,y) - x = -x \begin{cases} > -1 & \text{for} (x,y) \neq (1,0)^T \\ = -1 & \text{for} (x,y) = (1,0)^T \end{cases}$$

and therefore the solution is $(x^*, y^*) = (1,0)^T$ with $f(x^*, y^*) = -1$. As Lagrange multiplier, we obtain $\mu^* = -\frac{3}{2}$. Now, consider a feasible point $(x_k, y_k) \neq (\pm 1, 0)^T$ and $\mu_k < -1$. Let d_k be the solution of the SQP subproblem. Then there exist a multiplier μ_{qp}, such that the KKT-conditions of the subproblem are satisfied. Assume that $d_k = 0$, then it follows by

$$0 = \nabla f(x_k,y_k) + \nabla^2 \mathcal{L}(x_k,y_k,\mu_k) d_k + \mu_{qp}^k \nabla h(x_k,y_k) = (4+2\mu_{qp}) \begin{pmatrix} x_k \\ y_k \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

that $(x_k, y_k)^T = \alpha(1,0)^T$, which cannot be true since we chose a feasible point $(x_k, y_k) \neq (\pm 1, 0)^T$ and thus $d_k \neq 0$. Since f is quadratic, the second order Taylor approximation is exact which yields

$$f((x_k,y_k)+d_k) - f(x_k,y_k) = \nabla f(x_k,y_k)^T d_k + \frac{1}{2} d_k^T \nabla^2 f(x_k,y_k) d_k = \nabla f(x_k,y_k)^T d_k + 2\|d_k\|^2.$$

By the KKT-conditions, we can substitute $\nabla f(x_k, y_k)$ by its representation in terms of $\nabla^2 \mathcal{L}(x_k, y_k, \mu_k)$ and $\nabla h(x_k, y_k)$, i.e. we obtain

$$\begin{aligned} f((x_k, y_k) + d_k) - f(x_k, y_k) &= -(\mu_{qp} \nabla h(x_k, y_k) + \nabla^2 \mathcal{L}(x_k, y_k, \mu_k) d_k)^T d_k + 2\|d_k\|^2 \\ &= -\mu_{qp} \nabla h(x_k, y_k)^T d_k - (4 + 2\mu_k)\|d_k\|^2 + 2\|d_k\|^2 \\ &= -\mu_{qp}\, h(x_k, y_k) - 2(1 + \mu_k)\|d_k\|^2 \,. \end{aligned}$$

Because $h(x_k, y_k) = 0, \mu_k < -1$ and $d_k \neq 0$, it follows that$f(x_k + d_k) - f(x_k) > 0$. As h is also quadratic the second order Taylor approximation is here also exact and we have

$$\begin{aligned} |h((x_k, y_k) + d_k)| - |h(x_k, y_k)| &= |h((x_k, y_k) + d_k)| \geq h((x_k, y_k) + d_k) \\ &= h(x_k, y_k) + \nabla h(x_k, y_k)^T d_k + \frac{1}{2} d_k^T \nabla^2 h(x_k, y_k) d_k = \|d_k\|^2 > 0 \,. \end{aligned}$$

and therefore we finally get

$$P_1((x_k, y_k) + d_k; \alpha) - P_1((x_k, y_k); \alpha) = f((x_k, y_k) + d_k) - f(x_k, y_k) + \alpha(|h((x_k, y_k) + d_k)| - |h(x_k, y_k)|) > 0 \,.$$

A way out that is often used in practical SQP algorithms is to apply a so-called Second Order Correction (or SOC) step, which is defined by

$$d_{SOC} := -\nabla h(x_k)(\nabla h(x_k)^T \nabla h(x_k))^{-1} h(x_k + d_k)$$

This additional step is relatively small (in comparison to d_k) but improves the feasibility of the new iterate $x_k + d_k + d_{SOC}$ significantly. In particular, we have

$$h(x_k + d_k + d_{SOC}) = O(\|d_k\|^3) \qquad \text{compared to} \qquad h(x_k + d_k) = O(\|d_k\|^2) \,.$$

Under some suitable assumptions, the new step $d_k + d_{SOC}$ produces an Armijo step size $t_k = 1$ in the region of local convergence, such that by the small size of the correction step d_{SOC} the fast local convergence is again attained.

11.3.11 Trust-Region Method

In the following we will briefly discuss trust-region SQP methods. One main disadvantage of line-search SQP methods (see Theorem 11.50 and Theorem 11.52) is the need for positive definite matrices B_k. In Chapter 1, where we discussed the general trust-region methods, we noticed that this assumption is not essential for this class of methods. Hence, it is a natural idea to search suitable trust-region adaptions of the local SQP method.

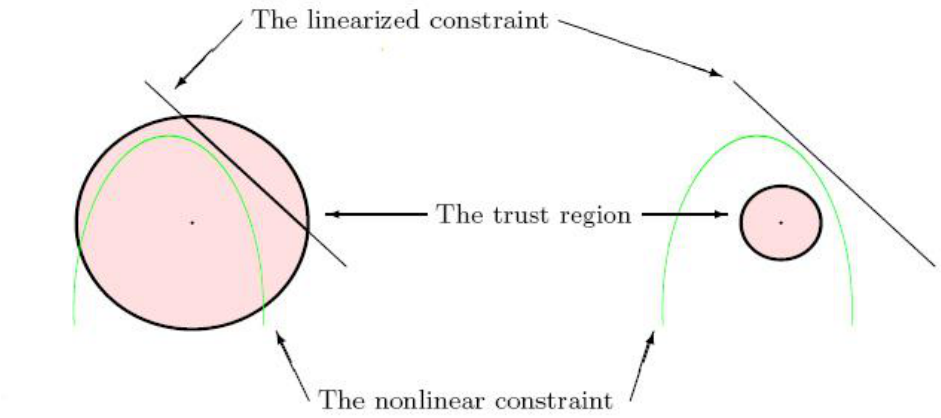


Figure 29: The intersection of the linearization of a nonlinear constraint and a spherical trust-region. In the left figure, the trust-region is large enough such that there exists an intersection (and hence a nonempty feasible for the trust-region QP). In the right figure, this is not the case, hence the QP is inconsistent. The Figure is taken from [**?**]. fig:tr-sqp1

One obvious trust-region generalization of the basic SQP subproblem (applied to (11.96)) is

$$\min_{d\in\mathbb{R}^n} g_k^T d + \frac{1}{2} d^T B_k d \qquad \text{subject to} \quad A_k^T d = -h_k\,, \quad \|d\| \le \Delta_k\,. \tag{11.122}$$ simple-tr-sqp

But there exists one critical difficulty with this approach. If the trust-region is too small and $h_k \neq 0$, then this SQP subproblem does not have any feasible points d (see also Figure 29). The critical radius Δ_{crit} is given by

$$\Delta_{crit} = \min \|d\| \qquad \text{subject to} \quad A_k^T d = -h_k\,.$$

We are therefore looking for some suitable alternatives that ommit this problem. Here we present two alternatives, the first one is the Sℓ_pQP method, the second one is a composite-step method.

11.3.12 Sℓ_pQP Method

In this approach, we try to minimize the exact ℓ_p-penalty function of the QP subproblem. Hence instead of (11.122) we consider the problem

$$\min_{d\in\mathbb{R}^n} \tilde{P}_p(d;\alpha) := f_k + g_k^T d + \frac{1}{2} d^T B_k d + \alpha \|h_k + A_k^T d\|_p \qquad \text{subject to} \quad \|d\| \le \Delta_k \,. \tag{11.123}$$

slpqp-1

This problem is always feasible and as long as α and Δ_k are large enough and $h_k + A_k d = 0$ is consistent, the solution of (11.123) is the SQP direction d_k. Furthermore, if we choose a polyhedral norm, i.e. $\|\cdot\|_1$ or $\|\cdot\|_\infty$, then (11.123) is again equivalent to a quadratic subproblem (see Exercise or [**?**]).

R. Fletcher proposed the use of the ℓ_1-penalty function, i.e. the Sℓ_1QP method in a paper in 1981 (see also [**?**] or [**?**]) as trust-region variant of the basic SQP method. The corresponding extension to the general NLP then yields the subproblems

$$\min_{d\in\mathbb{R}^n} \tilde{P}_1(d;\alpha) := \quad f_k + g_k^T d + \frac{1}{2} d^T B_k d + \alpha \|h_k + A_k^T d\|_1 + \alpha \sum_{j=1}^{m} |(g_j(x_k) + \nabla g_j(x_k)^T d)_+|$$

$$\text{subject to} \qquad \|d\| \le \Delta_k \,.$$

11.3.13 Composite-step Method

Another possibility concerns the separation of the step computation into two stages. First, we compute a normal step n_k which aims to improve the feasibility of the current iterate, i.e. it moves x_k in the direction of the feasible set (according to the linearized constraints and the trust-region, see also Figure 30). The target of the second step t_k, the tangential step, is then to reduce the value of the objective function within the trust region and without loosing the feasibility that we obtained by the normal step n_k, i.e. we get the condition

$$h_k + A_k^T(n_k + t_k) = h_k + A_k n_k \qquad \Rightarrow \qquad A_k^T t_k = 0$$

11.3.14 Filter Methods

One of the first filter methods was implemented in the SQP solver `filterSQP`, which was developed by Fletcher and Leyffer [**?**]. In `filterSQP` the SQP method is combined with a trust-region and a filter approach. Filter methods provide an alternative to penalty function methods to promote global convergence as they allow the full Newton step and one does not need to find

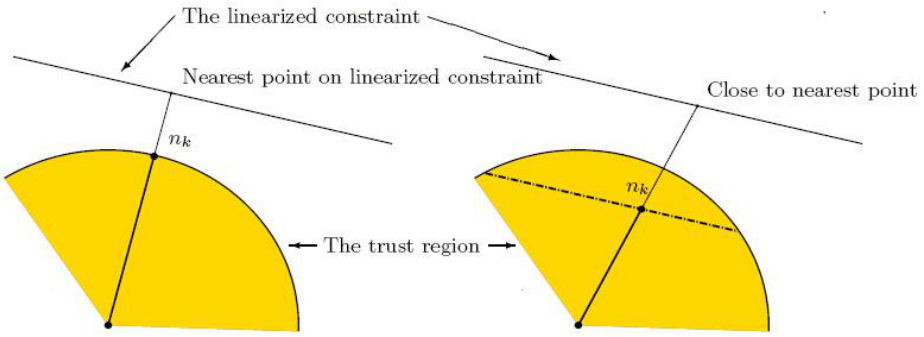


Figure 30: The left-hand figure shows the largest possible normal step, the right-hand figure illustrates a shorter normal step n, and the freedom this then allows for the tangential step - any point on the dotted line is a potential tangential step. The Figure is taken from [**?**]. fig:tr-sqp2

a suitable penalty parameter [**?**]. The difference of a filter method compared to a penalty function method can briefly be explained as follows:

Solving a Nonlinear Programming Problem of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c(x) \geq 0 \end{array}$$

comprises two targets: the minimization of the objective function $f(x)$ and of the constraint violation, which could be measured for example by $h(x) := \| (c(x))^- \|$. Using a penalty function these two objectives are combined into one single function and the second one is weighted with an increasing penalty parameter as feasibility has to be achieved in the solution. Instead of combining it, filter methods treat the NLP as a *biobjective optimization problem* [**?**], where either the objective function value $f(x)$ or some measure of the infeasibility of x has to be decreased sufficiently, compared to a test set of previously determined iterates called the *filter*.

Next, we explain the software package `filterSQP` and the filter method used therein more explicitly. `filterSQP` solves NLPs of the form [**?**]

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & lb_x \leq x \leq ub_x & (11.124) \\ & lb_c \leq c(x) \leq ub_c & \end{array}$$ nlp-filter

by solving a sequence of quadratic approximations of (11.124) in the current iterate x_k within a trust-region that is determined by the condition $\|d\|_\infty \leq \rho$, with ρ denoting the trust-region radius. The QPs thus have the form

$$\begin{aligned} \text{minimize} \quad & q_k(d) \\ \text{subject to} \quad & lb_x \leq x_k + d \leq ub_x \\ & lb_c \leq c(x_k) + \nabla c(x_k)^T d \leq ub_c \\ & \|d\|_\infty \leq \rho\,, \end{aligned} \tag{11.125}$$

qp-filter

where $q_k(d) := \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2_{xx} \mathcal{L}(x_k, \lambda_k) d$ corresponds to the quadratic approximation of the Lagrangian function $\mathcal{L}(x, \lambda)$ of (11.124). In contrast to other solver `filterSQP` uses the exact Hessian $\nabla^2_{xx} \mathcal{L}(x_k, \lambda_k)$. The QPs (11.125) are solved by `bqpd`, which is a robust QP solver that is based on a null-space active set method (for more information see [**?**]).

The solution d_k of (11.125) gives a next trial iterate $x^{k+1} = x_k + d_k$ and it is tested, if x^{k+1} can be accepted by the filter. The filter consists of a list of pairs $(f(x^\ell), h(x^\ell))$ of previous iterates x^ℓ, that are not dominated by any other pair. The concept of domination was adopted from multiobjective optimization and is defined in [**?**] as follows:

Definition 11.53. *A pair $(f(x_\ell), h(x_\ell))$ is said to dominate another pair $(f(x_k), h(x_k))$ if and only if both $f(x_\ell) \leq f(x_k)$ and $h(x_\ell) \leq h(x_k)$.*

Concerning the basic filter SQP algorithm (Algorithm 1 in [**?**]), the trial iterate x_{k+1} will be accepted by the filter, if the pair $(f(x_{k+1}),\ h(x_{k+1}))$ is not dominated by any other pair of the current filter. Algorithmic extensions of the basic filter SQP algorithm that are incorporated in `filterSQP` concern a *Second Order Correction (SOC) step*, an upper bound on the constraint violation, the elemination of *blocking entries* from the filter, a *sufficient reduction test* and a *North-West* and *South-East* corner rule. However, we will not further discuss these extensions here but refer the interested reader to [**?**].

If the trial point x_{k+1} is accepted by the filter, it is chosen to be the new iterate. The pair $(f(x_{k+1}),\ h(x_{k+1}))$ is then added to the filter and pairs $(f(x_\ell),\ h(x_\ell))$ that are dominated by $(f(x_{k+1}), h(x_{k+1}))$ are removed from the filter. If x_{k+1} is rejected by the filter, then the trail step d_k is discarded, the trust-region radius is reduced and the QP (11.125) is solved again.

Reducing the trust-region radius, though, might cause an infeasible QP, if the current iterate is not feasible for (11.124). Therefore, `filterSQP` incorporates a *feasibility restoration phase*, which aims to minimize the constraint

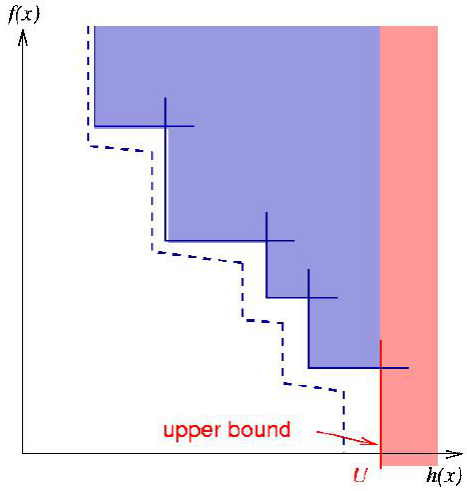


Figure 31: A typical filter, all pairs $(f(x), h(x))$ that are below and left of the envelope (dashed line) are exceptable to the filter. Figure taken from [**?**]

violation by applying a trust region SQP method to solve the problem

$$\begin{aligned} &\text{minimize} && \textstyle\sum_{j\in\mathcal{J}}(c_j(x))^+ \\ &\text{subject to} && c_j(x) \le 0 \qquad j \in \mathcal{J}^\perp\,, \end{aligned} \tag{11.126}$$

restoration-filter

where the sets $\mathcal{J}$ and $\mathcal{J}^\perp$ partition the nonlinear constraints into those ones that cannot be satisfied for the current QP (that is $c_j(x_k) + \nabla c_j(x_k)^T d > 0, \quad j \in \mathcal{J}$) and those that can be satisfied. For more details about the restoration phase and `filterSQP` in general, we refer to [**?**] and [**?**].

11.3.15 Convergence

It can be shown (see e.g. in [**?**, **?**]), that under the assumptions that the iterates x_k lie in a compact set X, the functions f and c are twice continuously differentiable the filter methods and the Hessians H_k remain bounded, filter methods have the following global convergence properties :

1. The restoration phase fails to find a filter acceptable point for which the QP is consistent for some $\rho \ge \underline{\rho}$

Algorithm 1: SQP Filter Method

x_0, $k \leftarrow 0$, $\mathcal{F}_0 \leftarrow \{U, -\infty\}$, `optimal` $\leftarrow$ `false`

while `not optimal` **do**
- reset the trust-region radius: $\rho_k \geq \underline{\rho}$
- `terminate` $\leftarrow$ `false`
- **repeat**
  - solve the QP (2.1) for a step s
  - **if** $s = 0$ **then**
    - `optimal` $\leftarrow$ `true`; STOP
  - **if** QP (2.1) incompatible **then**
    - add (h_k, f_k) to $\mathcal{F}_k$
    - enter restoration phase
  - **else**
    - **if** $x_k^+ := x_k + s$ not acceptable **then**
      - reduce trust-region $\rho_k \leftarrow \rho_k/2$
    - **else**
      - `terminate` $\leftarrow$ `true`
- **until** `terminate`
- update the filter $\mathcal{F}_{k+1}$
- set $x_{k+1} \leftarrow x_k + s$ and $k \leftarrow k + 1$

Figure 32: `filterSQP` Algorithm taken from [**?**].

2. The algorithm terminates with a first-order stationary point x^*, i.e. x^* satisfies the KKT-condition.

3. There exists a feasible accumulation point that is either stationary or the MFCQ fails to hold.

Concerning the fast local convergence properties of filter-SQP methods, S. Ulbrich proved q-quadratic convergence in [**?**] under the assumption of the LICQ and the SOSC.

11.4 Interior-Point Methods (IPM)

In the first two sections of this chapter we will focus on the nonlinear inequality constrained optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & g(x) \geq 0\,. \end{aligned} \tag{11.127}$$

`icnlp`

Hence the feasible region is given by $\mathcal{X} = \{x \in \mathbb{R}^n : g(x) \geq 0\}$. Moreover, we define the set of "strictly feasible" points $\text{strict}(\mathcal{X}) := \{x \in \mathbb{R}^n : g(x) > 0\}$. Note, that this set differs from the set of points that lie in the interior of the feasible set, i.e. in $\text{int}(\mathcal{X})$ (consider e.g. the example $g_1(x) = x^2$, $m = 1$).

Just as a reminder, the KKT-conditions of (11.127) are given by

$$\begin{aligned} \nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) - \nabla g(x^*)\lambda^* &= 0 \\ g(x^*) &\geq 0 \\ \lambda^* &\geq 0 \\ g_i(x^*)\lambda_i^* &= 0 \qquad i = 1, \ldots, m \end{aligned} \tag{11.128}$$

ecnlp-kkt

11.4.1 Barrier Methods

We first start with the theoretical and historical background of classical interior point methods. These are given by barrier methods. The idea to solve constrained optimization problems by transforming them into an unconstrained problem via penalty or barrier functions was already considered in the 1960th. However, they became regarded as unreliable due to several drawbacks.

11.4.2 The Barrier Function

In contrast to penalty methods, where a solution of the constrained problem is found by solving a sequence of unconstrained problems where infeasibility of a current iterate x_k is penalized via a penalty term, barrier methods use a barrier function $B(x, \pi)$, that consists of the objective function $f(x)$ and an interior function (or barrier term) that prevents the algorithm to produce infeasible iterates, i.e. the sequence (x_k) stays feasible during the whole solution process. Moreover, inside the feasible region, a suitable barrier function should resemble the original objective function. Desirable properties of the interior function $I(x)$ are thus:

1. $I(x)$ depends only on the constraint functions.
2. $I(x)$ preserves the continuity properties of $g(x)$ at all points in $\text{int}\mathcal{X}$.
3. For any sequence of points in $\text{int}\mathcal{X}$ converging to a point on the boundary of the feasible region, $I(x) \to +\infty$.

Two suitable examples for $I(x)$ are

- the inverse interior function

$$I_{\text{inv}}(x) := \sum_{j=1}^{m} \frac{1}{g_i(x)}$$

- and the logarithmic interior function

$$I_{\log}(x) := -\sum_{j=1}^{m} \ln(g_i(x)).$$

We continue the discussion with the logarithmic interior function, since the use of the corresponding logarithmic barrier function dominates the literature (due to its close connection to perturbed KKT systems, as we will see later on, and other reasons). The logarithmic barrier function is a composite function that is based on the logarithmic interior function and defined as follows:

$$B(x, \pi) := f(x) + \pi I_{\log}(x) = f(x) - \pi \sum_{j=1}^{m} \ln(g_i(x)), \qquad (11.129)$$ barrierfct

where $\pi > 0$ is a scalar barrier parameter. Notice, that the smoothness properties of f, and g_i $(i = 1, \ldots, m)$ are retained, as long as x is strictly feasible (which, as already said, is not exactly the same as $x \in \text{int}(\mathcal{X})$). An obvious basic interior point algorithm for (11.127) that is based on the barrier method is given by Algorithm 9 (see [**?**])

Algorithm 9: Barrier Algorithm

bm1 1 *Choose initial values for $x_0 \in$ strict$(\mathcal{X})$ and $\pi_0 > 0$.*;

repeat

bm2 2 *Compute a solution x_k of the uncontrained problem*

$$\min_{x \in \mathbb{R}^n} B(x, \pi_k)$$

bm3 3 *Update the barrier parameter π_k such that*

$$\pi_k > \pi_{k+1} > 0$$

(e.g. $\pi_{k+1} = 0.1\,\pi_k$ or $\pi_{k+1} = \pi_k^2$) ;

bm4 4 $k \leftarrow k + 1$;

alg-ipm **until** x_k is an .;

A basic convergence result (cf. [**?**]) is as follows.

Theorem 11.54. *Suppose that f and g are twice contiuously differentiable. Moreover, let $\lambda_{k,i} := \pi_k / g_i(x_k)$ and assume that*

$$\|\nabla_x B(x_k, \pi_k)\|_2 \leq \varepsilon_k$$

where $\varepsilon_k \to 0$ as $k \to \infty$ and that x_k converges to x^*, at which the LICQ holds for (11.127). Then x^* satisfies the KKT-conditions for (11.127) and (λ_k) converges to the (unique) associated Lagrange multiplier λ^*.

Proof. Define the set of inactive constraints $I_g^\perp := \{1, \ldots, m\} \setminus I_g(x^*)$ and let the subscripts $_{I_g}$ and $_{I_g^\perp}$ denote the rows of the matrices/vectors whose indices belong to $I_g(x^*)$ and $I_g^\perp$, respectively. Let $\lambda_{k,i} := \pi_k / g_i(x_k)$ $(i = 1, \ldots, m)$ and let

$$A^+_{I_g}(x) = (\nabla g(x)^T_{I_g} \nabla g(x)_{I_g})^{-1} \nabla g(x)^T_{I_g}$$

be the left generalized inverse of $g(x)_{I_g}$ (which exists due to the LICQ and the continuity of g in a neighbourhood of x^*). Define

$$(\lambda^*)_{I_g} = A^+_{I_g}(x^*) \nabla f(x^*)$$

and $(\lambda^*)_{I_g^\perp} = 0$. If $I_g^\perp \neq \emptyset$, then for sufficiently large k we have

$$\|(\lambda_k)_{I_g^\perp}\|_2 \leq \frac{2\,\pi_k \sqrt{|I_g^\perp|}}{\min_{i \in I_g^\perp} g_i(x^*)}. \tag{11.130}$$ thm1ipm

Since,

$$\varepsilon_k \geq \|\nabla_x B(x_k, \pi_k)\|_2 = \|\nabla f(x_k) - \nabla g(x_k) \lambda_k\|_2,$$

we furthermore get

$$\begin{aligned} \|\nabla f(x_k) - \nabla g(x_k)_{I_g} (\lambda_k)_{I_g}\|_2 &\leq \|\nabla f(x_k) - \nabla g(x_k)\lambda_k\|_2 + \|\nabla g(x_k)_{I_g^\perp} (\lambda_k)_{I_g^\perp}\|_2 \\ &\leq \varepsilon_k + \|\nabla g(x_k)_{I_g^\perp}\|_2 \, \tfrac{2\,\pi_k \sqrt{|I_g^\perp|}}{\min_{i \in I_g^\perp} g_i(x^*)} =: \tilde{\varepsilon}_k \,. \end{aligned}$$

(11.131) thm1ipm

Note, that for $k \to \infty$ the scalar $\tilde{\varepsilon}_k$ tends to zero. By the inequality (11.131), we obtain

$$\begin{aligned} \|A^+_{I_g}(x_k) \nabla f(x_k) - (\lambda_k)_{I_g}\|_2 &= \|A^+_{I_g}(x_k)(\nabla f(x_k) - \nabla g(x_k)^T_{I_g} \lambda_k)_{I_g})\|_2 \\ &\leq 2\|A^+_{I_g}(x^*)\|_2 \, \tilde{\varepsilon}_k. \end{aligned}$$

And therefore

$$\begin{aligned} \|(\lambda_k)_{I_g} - (\lambda^*)_{I_g}\|_2 &\leq \|(\lambda_k)_{I_g} - A^+_{I_g}(x_k) \nabla f(x_k)\|_2 + \|A^+_{I_g}(x^*) \nabla f(x^*) - A^+_{I_g}(x_k) \nabla f(x_k)\|_2 \\ &\leq 2\|A^+_{I_g}(x^*)\|_2 \, \tilde{\varepsilon}_k + \bar{\varepsilon}_k, \end{aligned}$$

where due to the continuity properties of f and g the sequence $(\bar{\varepsilon}_k)$ converges to zero for $k \to \infty$. This proves that the sequence (λ_k) converges to λ^*. In addition, continuity of the gradients and (11.131) implies that

$$\nabla f(x^*) - \nabla g(x^*)\lambda^* = 0.$$

The fact that x_k is strictly feasible for all k implies that $g_i(x^*) \geq 0$ for all $i \in \{1, \ldots, m\}$ and moreover $\lambda^* \geq 0$. Finally, the complementarity condition is directly satisfied by the definition of λ^*. □

1. Another local convergence result concerning a sequence of minimizers of the logarithmic barrier function, where the existence of at least one convergent subsequence of (x_k) is shown, can be found in [**?**] (Theorem 3.10).

2. In Algorithm 9 we still have flexibility in the method that we use to solve the inner problems. This, however, must be done with special care, since the Hessians of $B(x, \pi)$ with respect to x are increasingly ill-conditioned (as the condition number of $\nabla_{xx} B(x, \pi)$ can be proved to be $O(1/\pi_k)$ - see also [**?**] and [**?**]for a further discussion).

3. The simple choice of x_k as a starting point for the next inner optimization (e.g. Newton's method) seems to be poor (see [**?**] for a detailed discussion). So again this issue has to be handled with special care.

11.4.3 The Central Path

The central path $x(\pi)$ (also known as the barrier trajectory) is an important concept in the context of interior point methods. It consists of a sequence of barrier minimizers x_π. It can be proved, that under suitable assumptions, the path $x(\pi)$ is differentiable and converges to the minimizer x^*. In the following theorem, we summarize some main properties of the central path.

centralpath

Theorem 11.55. *Let x^* be a local minimum of (11.127) and assume that the following conditions hold:*

a) x^* is a KKT-point, i.e. the multiplier set $\mathcal{M}_\lambda(x^*)$ of Lagrange multipliers for x^* defined by

$$\mathcal{M}_\lambda(x^*) := \{\lambda \in \mathbb{R}^m \ : \ \nabla f(x^*) = \nabla g(x^*)\lambda, \quad \lambda \geq 0 \quad \textit{and} \quad g_i(x^*)\lambda_i = 0 \quad \forall\, i = 1, \ldots, m\}$$

is not empty;

b) the MFCQ holds in x^*;

c) there exists a scalar $\beta > 0$ such that

$$s^T \nabla_{xx} \mathcal{L}(x^*, \lambda) s \geq \beta \|s\|^2, \quad \forall\, \lambda \in \mathcal{M}_\lambda(x^*) \quad \textit{and}$$
$$\forall\, s \in \{p \in \mathbb{R}^n \setminus \{0\} \,:\, \nabla f(x^*)^T p = 0,\ (\nabla g(x^*))^T_{I_g(x^*)} p \geq 0\}\,.$$

Assume that the logarithmic barrier method is applied in which π_k converges monotonically to zero as $k \to \infty$. Then it holds:

1. *there is at least one subsequence of unconstrained minimizers of the barrier function $B(x, \pi_k)$ that converges to x^*;*
2. *let (x_k) denote such a convergent sequence, then the sequence (λ_k) of barrier multipliers defined as $\lambda_{k,i} = \pi_k / g_i(x_k)$ is bounded;*
3. $\lim_{k\to\infty} \lambda_k =: \lambda^* \in \mathcal{M}_\lambda(x^*)$;

   If, in addition, there exits a vector $\lambda \in \mathcal{M}_\lambda(x^*)$ that satisfies $\lambda_i > 0$ for all $i \in I_g(x^*)$ (i.e. strict complementarity), then it also holds
4. $\lambda_i^* > 0$ *for all* $i \in I_g(x^*)$;
5. *for suffienciently large k, $\nabla_{xx} B(x_k, \pi_k)$ is positive definite;*
6. *a unique, continuously differentiable vector function $x(\pi)$ of unconstrained minimizers of $B(x, \pi)$ exists for positive π in a neighbourhood of $\pi = 0$; and*
7. $\lim_{\pi \to 0^+} x(\pi) = x^*$.

Proof. See Theorem 3.12 and its proof in [**?**]. □

In addition, it can be proved that the order of convergence of (x_k) is directly connected to the order of convergence of (π_k).

thmipm-2 **Theorem 11.56.** *Under the assumptions a)-c) of Theorem 11.55 and the additional assumption of strict complementarity at x^*, there exist $\kappa_l > 0$ and $\kappa_u > 0$ such that*

$$\kappa_l \pi_k \quad \leq \quad \|x_k - x^*\| \quad \leq \quad \kappa_u \pi_k\,.$$

The strict complementarity is a vital assumption in Theorem 11.56. It can be shown [**?**], that in the absence of strict complementarity, i.e. if there exists at least one multiplier $\bar{\lambda} \in \mathcal{M}_\lambda(x^*)$ with at least one index $i \in I_g(x^*)$ with $\bar{\lambda}_i = 0$, the guaranteed convergence of the sequence (x_k) is considerably worse and given by

$$\tilde{\kappa}_l \|x_k - x^*\|^2 \quad \leq \quad \pi_k \quad \leq \quad \tilde{\kappa}_u \|x_k - x^*\|^2\,.$$

11.4.4 Perturbed KKT-System

Next, we consider the connection between minimizers of $B(x, \pi)$ and local constrained minimizers of (11.127). First, note that if x_π is a local (unconstrained) minimizer of the barrier function $B(x, \pi)$, then

$$\nabla_x B(x_\pi, \pi) = \nabla f(x_\pi) - \sum_{j=1}^{m} \frac{\pi}{g_j(x_\pi)} \nabla g_j(x_\pi) = 0.$$

Introducing so-called barrier multipliers

$$\lambda_{\pi,i} := \frac{\pi}{g_i(x_\pi)}$$

we can rewrite the first order condition:

$$\nabla f(x_\pi) - \sum_{j=1}^{m} \lambda_{\pi,i} \nabla g_j(x_\pi) = \nabla f(x_\pi) - \nabla g(x_\pi)\lambda_\pi = 0\,,$$

which is exactly the multiplier rule for (11.127). Moreover, it can be proved, that $g(x_\pi) > 0$ for all $\pi > 0$ and therefore $\lambda_\pi > 0$ for all positive barrier parameter π. If we express the definition of λ_π as

$$g_i(x_\pi)\lambda_{\pi,i} = \pi \qquad i = 1, \ldots, m \tag{11.132}$$

pertcomp

then the resemblence to the KKT-system of (11.127) becomes even more apparent. The condition (11.132) corresponds to the perturbed complementarity condition, where zero on the right-hand side is replaced by a positive parameter π that is sequentially driven to zero, i.e. in the limit $\pi \to 0$ not only the multiplier rule and the inequality constraints are satisfied but also the complementarity condition.

Writing both conditions as a nonlinear system of $n + m$ equations, we get

$$F_\pi(x, \lambda) := \begin{pmatrix} \nabla f(x) - \nabla g(x)\lambda \\ G(x)\lambda - \pi e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{11.133}$$

primaldual1

where $G(x) :=$diag$(g_i(x))$ and $e = (1, 1, \ldots, 1)$ (compare $F_\pi(x, \lambda)$ with the original KKT-system for (11.127)). It can easily be seen that the pair (x_π, λ_π), as we defined it before, satisfies $F_\pi(x, \lambda) = 0$. Conversely, any pair that is a root of $F_\pi(x, \lambda)$, is a stationary point of the barrier function $B(x, \pi)$. However, the nonnegativity of $g_i(x)$ and λ_i is not represented in the condition $F_\pi(x, \lambda) = 0$, nor are any second order information used or satisfied for any root of $F_\pi(x, \lambda)$.

11.4.5 Primal-Dual Interior Methods

Primal-dual interior methods are based on the perturbed KKT-conditions (11.133). They became increasingly popular due to the difficulties that arise when solving the barrier problem with classical unconstrained solution methods (Newton's method). In contrast to these methods that aim for the primal variables x, primal-dual methods are twofold oriented as they aim to find a primal-dual pair (x, λ) that satisfies (11.133), where λ is treated as an independent variable. In this case, strict feasibility of the primal-dual KKT-pair is satisfied, if $g(x) > 0$ and $\lambda > 0$ holds. Applying Newton's method to solve the nonlinear equation (11.133), we have to solve the Newton equation

$$\begin{pmatrix} \nabla_{xx}\mathcal{L}(x,\lambda) & -\nabla g(x) \\ \Lambda\nabla g(x)^T & G(x) \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = - \begin{pmatrix} \nabla f(x) - \nabla g(x)\lambda \\ G(x)(\lambda - \alpha(x,\pi)) \end{pmatrix} \quad (11.134)$$

primaldualnewton

for the vector $(\Delta x, \Delta\lambda)$, where $\alpha_i(x,\pi) := \pi/g_i(x)$. The success of the primal-dual interior point methods is due to their effectiveness in following the central path to the optimum.

thm-ipm3

Theorem 11.57. *Suppose $\tilde{\pi}$ to be a specific value for the barrier parameter π and assume that π is reduced from $\tilde{\pi}$ to $\hat{\pi}$ (i.e. $\tilde{\pi} > \hat{\pi}$). Then the primal-dual direction $(\Delta x, \Delta\lambda)$ of (11.134) for $\pi = \hat{\pi}$ is tangent to the primal-dual trajectory $(x(\pi), \lambda(\pi))$ at $(x,\lambda) = (x(\tilde{\pi}), \lambda(\tilde{\pi}))$.*

Proof. Consider the primal-dual point $(x,\lambda) = (x(\tilde{\pi}), \lambda(\tilde{\pi}))$ that lies on the barrier trajectory and note that it satisfies the equations

$$\begin{aligned} \nabla f(x(\tilde{\pi})) - \nabla g(x(\tilde{\pi}))\lambda(\tilde{\pi}) &= 0\,, \\ \lambda_i(\tilde{\pi}) &= \frac{\tilde{\pi}}{g_i(x(\tilde{\pi}))} \qquad (i = 1,\ldots,m) \\ G(x(\tilde{\pi}))\lambda(\tilde{\pi}) &= \tilde{\pi}\, e \\ \alpha(x(\tilde{\pi}), \hat{\pi}) &= \frac{\hat{\pi}}{\tilde{\pi}}\alpha(x(\tilde{\pi}), \tilde{\pi})\,. \end{aligned}$$

Therefore the new direction $(\Delta x, \Delta\lambda)$ of (11.134) for $\pi = \hat{\pi}$ satisfies

$$\begin{pmatrix} \nabla_{xx}\mathcal{L}(x(\tilde{\pi}),\lambda(\tilde{\pi})) & -\nabla g(x(\tilde{\pi})) \\ \Lambda\nabla g(x(\tilde{\pi}))^T & G(x(\tilde{\pi})) \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = - \begin{pmatrix} 0 \\ (\tilde{\pi} - \hat{\pi})\, e \end{pmatrix}. \quad (11.135)$$

prf-thm-ipm3-1

On the other hand, differentiating the equations for $x(\pi)$ and $\lambda(\pi)$ with respect to π leads to the following equations for the tangent step $(x'(\pi), \lambda'(\pi))$:

$$\begin{pmatrix} \nabla_{xx}\mathcal{L}(x(\pi),\lambda(\pi)) & -\nabla g(x(\pi)) \\ \Lambda(\pi)\nabla g(x(\pi))^T & G(x(\pi)) \end{pmatrix} \begin{pmatrix} x'(\pi) \\ \lambda'(\pi) \end{pmatrix} = \begin{pmatrix} 0 \\ e \end{pmatrix}. \quad (11.136)$$

prf-thm-ipm3-2

Comparing (11.135) and (11.136) we obtain that $\Delta x = (\hat{\pi} - \tilde{\pi})x'(\tilde{\pi})$ and $\Delta\lambda = (\hat{\pi} - \tilde{\pi})\lambda'(\tilde{\pi})$. Hence the new direction is tangent to the primal-dual trajectory at $(x(\pi), \lambda(\pi))$ at $\pi = \tilde{\pi}$. □

By Theorem 11.57, we can assume that the new primal-dual iterate

$$\begin{aligned} x_{k+1} &= x(\tilde{\pi}) + \Delta x = x(\tilde{\pi}) + (\hat{\pi} - \tilde{\pi})\, x'(\tilde{\pi})\, (\approx x(\hat{\pi})\,) \\ \lambda_{k+1} &= \lambda(\tilde{\pi}) + \Delta\lambda = \lambda(\tilde{\pi}) + (\hat{\pi} - \tilde{\pi})\, \lambda'(\tilde{\pi})\, (\approx \lambda(\hat{\pi})\,) \end{aligned}$$

is a good approximation of the next point on the trajectory $(x, \lambda) = (x(\hat{\pi}), \lambda(\hat{\pi}))$. This property does not hold for the classical barrier method [**?**].

11.4.6 Formulation of the Primal-Dual Equations

As the classical barrier methods, primal dual interior point methods have a two-level structure of inner and outer iterations (or minor and major iterations). The inner iterations correspond to the Newton iterations for a given value π, i.e. the solution of (11.134). Under the assumption of strict complementarity and a suitable constraint qualification, these inner iterations converge at a q-quadratic rate [**?**]. Moreover, using suitable termination criteria for the inner loop, the combined sequence of inner iterates converges to the solution x^* of (11.127) q-superlinearly, if π is reduced appropriately. (See also [**?**] for local convergence results of primal-dual interior point methods.)

The key factor for the efficiency of a primal-dual interior point method is an efficient method for solving the linear system (11.134). One common approach is to use block elimination to obtain a smaller condensed system. Eliminating the $(2, 2)$ block of (11.134) yields (remember: $g(x) > 0$)

$$\Delta\lambda = -(\lambda - \alpha(x, \pi)) - G(x)^{-1}\Lambda\nabla g(x)^T\Delta x \qquad (11.137) \quad \boxed{\texttt{ipmelim1}}$$

and thus

$$\nabla_{xx}\mathcal{L}(x, \lambda)\Delta x - \nabla g(x)[-(\lambda - \alpha(x, \pi)) - G(x)^{-1}\Lambda\nabla g(x)^T\Delta x] = -(\nabla f(x) - \nabla g(x)\lambda) \qquad (11.138) \quad \boxed{\texttt{ipmelim2}}$$

or

$$H_c(x, \lambda)\Delta x = [(\nabla_{xx}\mathcal{L}(x, \lambda) + \nabla g(x)G(x)^{-1}\Lambda\nabla g(x)^T]\Delta x = -(\nabla f(x) - \nabla g(x)\alpha(x, \pi))\,. \qquad (11.139) \quad \boxed{\texttt{ipmelim3}}$$

Note that the condensed primal-dual matrix $H_c(x, \lambda)$ is symmetric and equal to the barrier Hessian at any minimizer of the barrier function. Therefore, this matrix is positive definite at point on the trajectory for sufficiently small π (see Theorem 4.2). Moreover, since the right-hand side is equivalent

to the negative gradient od the barrier function, (11.139) resembles the classical Newton barrier equation. Furthermore, like the barrier Hessian, $H_c(x,\lambda)$ becomes increasingly ill-conditioned as $\pi \to 0$. However, this ill-conditioning is usually harmless [**?**] and the system can be solved e.g. using an ordinary Cholesky factorization.

Another strategy to solve (11.134) is to symmetrize and subsequently factorize the system. Multiplying the second block of equations with $\Lambda^{-\frac{1}{2}}$ gives

$$\begin{pmatrix} \nabla_{xx}\mathcal{L}(x,\lambda) & -\nabla g(x)\Lambda^{\frac{1}{2}} \\ \Lambda^{\frac{1}{2}}\nabla g(x)^T & -G(x) \end{pmatrix} \begin{pmatrix} \Delta x \\ -\Lambda^{-\frac{1}{2}}\Delta\lambda \end{pmatrix} = - \begin{pmatrix} \nabla f(x) - \nabla g(x)\lambda \\ \Lambda^{-\frac{1}{2}} G(x)(\lambda - \alpha(x,\pi)) \end{pmatrix}. \tag{11.140}$$

primaldualsymm

In contrast to $H_c(x,\lambda)$, if strict complementarity holds at the solution and the gradients of active constraints are linearly independent, the matrix of (11.140) remains well-conditioned as $\pi \to 0$.

11.4.7 Globalization Strategies

One of the most popular choices to promote global convergence, i.e. convergence from any starting point , is to use a line-search method applied to a penalty or a merit function. In the case of primal-dual interior point methods, a (decrease in a) suitable merit function should encourage the iterates to "move towards the trajectory".

For convex problems, the steplength $t > 0$ is usually be chosen such that the iterates remain strictly feasible, i.e. $g(x + t\Delta x) > 0$ and $\lambda + t\Delta\lambda > 0$, and some norm of the KKT-residual is sufficiently reduced, i.e.

$$\|F_\pi(x + t\Delta x, \lambda + t\Delta\lambda)\| < \sigma \|F_\pi(x,\lambda)\|.$$

For nonconvex problems we consider the merit function

$$M_\pi(x,\lambda) := f(x) - \pi \sum_{j=1}^{m} \ln(g_j(x)) - \pi \sum_{j=1}^{m} \left(\ln\left(\frac{g_j(x)\lambda_j}{\pi}\right) + 1 - \frac{g_j(x)\lambda_j}{\pi} \right),$$

which is the classical barrier function $B(x,\pi)$ augmented by a weighted term that measures the distance of the iterate (x,λ) to the trajectory $(x(\pi),\lambda(\pi))$.

The main property of $M_\pi(x,\lambda)$ is that it is minimized with respect to both variables x and λ at the points $(x(\pi),\lambda(\pi))$ on the trajectory. Hence a decrease of $M_\pi(x,\lambda)$ implies a progress towards the primal-dual trajectory. i.e. a minimizer of $B(x,\pi)$.

At any point $(x(\pi), \lambda(\pi))$ on the trajectory we have $M_\pi(x, \lambda) = B(x, \pi)$, since the additional term vanishes. Moreover, since $(x(\pi), \lambda(\pi))$ is an unconstrained minimizer of $M_\pi(x, \lambda)$, the first and second oder necessary conditions hold in $(x(\pi), \lambda(\pi))$:

i) $\nabla M_\pi(x(\pi), \lambda(\pi)) = 0$

ii) $\nabla^2 M_\pi(x(\pi), \lambda(\pi)) \in \mathbb{R}^{(n+m)\times(n+m)}$ is positive semidefinite.

The minimization of $M_\pi(x, \lambda)$ can for example be done using a line-search algorithm or e.g. a trust region method based on finding an approximate solution of the subproblem

$$\min_{s \in \mathbb{R}^n} \; q(s) = \nabla M_\pi^T s + \frac{1}{2} s^T Q(x, \lambda) s \qquad \text{subject to} \quad \|s\|_T \leq \Delta\,, \quad (11.141) \qquad \boxed{\texttt{pd-tr-sub}}$$

where

$$Q(x, \lambda) = \begin{pmatrix} \nabla_{xx}\mathcal{L}(x, \lambda) + 2\nabla g(x) G(x)^{-1} \Lambda \nabla g(x)^T & -\nabla g(x) \\ \nabla g(x)^T & -\Lambda^{-1} G(x) \end{pmatrix},$$

i.e. $Q(x, \lambda)$ is $\nabla^2 M_\pi(x, \lambda)$ with $\alpha(x, \pi)$ replaced by λ and $\pi\Lambda^{-1}$ replaced by $G(x)$. Furthermore, $\|s\|_T := \sqrt{s^T T s}$ with $T = \text{diag}(I, \Lambda^{-1} G(x))$ (i.e. T is block diagonal).

Another strategy that is used in the solver IPOPT [**?**] to ensure global convergence is based on the concept of a filter.

11.4.8 Treatment of Equality Constraints

In contrast to the previous discussion in this chapter, in this section we will discuss a solution approach for the general NLP

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & h(x) = 0 \\ & g(x) \geq 0\,. \end{aligned} \qquad (11.142) \qquad \boxed{\texttt{ipnlp}}$$

11.4.9 A Barrier-SQP Approach

The motivation for the barrier method was to eliminate the inequality constraints of (11.127) by using a barrier function which implicitly forces the algorithm to produce strictly feasible iterates for (11.127). As an analogue for the mixed constraints, we treat the inequalites of (11.142) as we have

done before for (11.127) and keep only the equality conditions as constraints. This approach then yields the equality constrained problem

$$\begin{array}{rll} (P_\pi) & \min & B(x,\pi) \\ & \text{subject to} & h(x) = 0\,, \end{array} \tag{11.143}$$

ipecnlp

where $B(x,\pi)$ is given by (11.129). Hence to solve (11.142) we need to solve a sequence of equality constrained subproblems (P_{π_k}) for a decreasing sequence of barrier parameter (π_k) converging to zero. As described Section 3.2 these problems can be solved using an SQP approach. The associated KKT-conditions of (11.143) are

$$\begin{aligned} \nabla f(x) - \sum_{j=1}^{m} \frac{\pi}{g_j(x)} \nabla g_j(x) &= \nabla h(x)\mu \\ h(x) &= 0\,. \end{aligned}$$

Again introducing a new Variable λ (the multiplier for the inequality constraints of (11.142)) which are implicitly defined by an additional equation

$$g_i(x)\lambda_i = \pi \qquad i = 1, \ldots, m$$

gives the system

$$F_\pi(x,\lambda,\mu) = \begin{pmatrix} \nabla f(x) - \nabla g(x)\lambda - \nabla h(x)\mu \\ G(x)\lambda - \pi\, e \\ h(x) \end{pmatrix} = 0\,. \tag{11.144}$$

ipF

A special form of the SQP approach that we discussed in Section 3.1 is the Lagrange-Newton method. Here Newton's method is applied to the corresponding nonlinear system of the KKT-conditions of an equality constrained problem.

Applying Newton's method to the system $F_\pi(x,\lambda,\mu) = 0$ yields the Newton equation

$$\begin{pmatrix} \nabla_{xx}\mathcal{L}(x,\lambda,\mu) & -\nabla g(x) & -\nabla h(x) \\ \Lambda \nabla g(x)^T & G(x) & 0 \\ \nabla h(x)^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \\ \Delta\mu \end{pmatrix} = - \begin{pmatrix} \nabla f(x) - \nabla g(x)\lambda - \nabla h(x)\mu \\ G(x)\lambda - \pi\, e \\ h(x) \end{pmatrix} \tag{11.145}$$

ipsqpnewton

This system can e.g. be solved via an iterative method: By the second block of equations, we obtain

$$\Delta\lambda = G(x)^{-1}[-(G(x)\lambda - \pi e) - \Lambda \nabla g(x)^T \Delta x]\,. \tag{11.146}$$

iprhs

If we substitute $\Delta\lambda$ in (11.145) by the right-hand side of (11.146), then we get

$$\begin{pmatrix} \nabla_{xx}\mathcal{L}(x,\lambda,\mu) + \nabla g(x)G(x)^{-1}\Lambda\nabla g(x)^T & \nabla h(x) \\ \nabla h(x)^T & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ -(\mu+\Delta\mu) \end{pmatrix} = -\begin{pmatrix} \nabla f(x) - \nabla g(x)\alpha(x,\pi) \\ h(x) \end{pmatrix} \tag{11.147}$$

ipsqpnewton2

where $\alpha(x,\pi)$ is defined as before. The solution Δx of this system (11.147) solves the quadratic subproblem

$$\begin{aligned} \min \quad & (\nabla f(x) - \nabla g(x)\alpha(x,\pi))^T\Delta x + \tfrac{1}{2}\Delta x^T(\nabla_{xx}\mathcal{L}(x,\lambda,\mu) + \nabla g(x)G(x)^{-1}\Lambda\nabla g(x)^T)\Delta x \\ \text{subject to} \quad & h(x) + \nabla h(x)^T\Delta x = 0\,, \end{aligned} \tag{11.148}$$

A corresponding SQP method can be globalized using a line-search approach applied to an SQP merit function or a trust-region method. The limit points of such derived sequences of iterates satisfy the first and second order necessary condition for a fixed value of π.

11.4.10 A Penalty-Barrier Approach

Combining the barrier approach for the inequality constraints (i.e. $B(x,\pi)$) with the penalty method for the equality conditions (i.e. $P(x;\pi)$), we obtain a penalty-barrier function

$$\Phi_{PB}(x;\pi) := f(x) - \pi\sum_{j=1}^{m}\ln(g_i(x)) + \frac{1}{2\pi}\|h(x)\|_2^2\,.$$

It can be shown, that under suitable assumptions and for π small enough a sequence (x_π) of unconstrained minimizers of $\Phi_{PB}(x,\pi)$ defines a differentiable penalty-barrier path that converges to x^*.

For x_π to be a minimizer of $\Phi_{PB}(x;\pi)$, the first order necessary conditions must hold, i.e. $\nabla\Phi_{PB}(x_\pi;\pi) = 0$. Introducing new variables λ and μ that satisfy the defining conditions

$$G(x)\lambda = \pi\, e \qquad \text{and} \qquad \pi\mu = -h(x)\,,$$

we can rewrite the stationarity condition as the system

$$F_\pi(x,\lambda,\mu) = \begin{pmatrix} \nabla f(x) - \nabla g(x)\lambda - \nabla h(x)\mu \\ G(x)\lambda - \pi\, e \\ h(x) + \pi\mu \end{pmatrix} = 0\,, \tag{11.149}$$

ipF

where λ and μ are the multiplier estimates that converge to the KKT-multipliers λ^* and μ^* for (11.142) as $\pi \to 0$.

Remark 11.58. *Note that in this case the KKT-conditions for (11.142) are perturbed for both the inequalities and the equalities.*

The associated Newton equation is given by

$$\begin{pmatrix} \nabla_{xx}\mathcal{L}(x,\lambda,\mu) & -\nabla g(x) & -\nabla h(x) \\ \Lambda\nabla g(x)^T & G(x) & 0 \\ \nabla h(x)^T & 0 & \pi I \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \\ \Delta\mu \end{pmatrix} = - \begin{pmatrix} \nabla f(x) - \nabla g(x)\lambda - \nabla h(x)\mu \\ G(x)\lambda - \pi\, e \\ h(x) + \pi\mu \end{pmatrix}. \tag{11.150}$$

pbnewton

12 Numerical Methods for Linear Programming and Graph Theory

12.1 The Simplex Method for Linear Programming Problems in Finite Space Dimensions

We consider a linear programming problem given by the following equations. This problem is a special case of convex optimization. Since the cost function is not strictly convex, there is not necessarily a unique minimizer of the problem. Since the constraints are linear the Slater condition is satisfied and the KKT–System is sufficient and necessary for optimality.

$$\begin{aligned} \min c^T x \text{ subject to} \\ G^T x + g^0 \geq \\ H^T x + h^0 = 0 \end{aligned}$$

Here, $c \in \mathbb{R}^n$ and $G \in \mathbb{R}^{n\times m}$ and $H \in \mathbb{R}^{n\times p}$. If we have linearly dependent columns we can always eliminate them in a preprocessing step such that Slater's condition is satisfied.

A typical linear programming problem is solved in standard form which is different from the previous equations in the sense that the inequality constraints are only one–sided box constraints. Every problem of the previous form can be rewritten in standard form by introducing slack variables and thereby extending the size of x. For example, if

$$Bx \leq b \implies (B, Id)(x, y) = b, y \geq 0 \Leftrightarrow Bx + y = b, y \geq 0.$$

Some of the initial variables x might not satisfy the constraint $x \geq 0$. Those variables are splitted as follows

$$x = y^+ - y^-, y^\pm \geq 0.$$

This adds an additional linear equation to the problem. Hence, we will derive the Simplex method for a problem in standard form given by

$$\begin{aligned} c^T x \to \min \qquad & \\ Ax &= b \\ x &\geq 0 \end{aligned}$$

where $A \in \mathbb{R}^{p\times n}$ as full rank $(= p)$ and $b \geq 0$.

From the setting of the problem we immediately conclude hat the feasible set $\mathfrak{S}$ and the set of solutions is convex if they are not empty. The set

$$\mathfrak{S} = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\} \tag{12.151}$$ 040409:1

is called a convex polyeder and if it is additionally compact a convex polytop. Using the notation of the previous paragraphs we have that for our problem

$$\nabla h = A^T, \nabla g = Id, \mathcal{A}_i = \{i : x_i = 0\}.$$

Due to the assumption A^T has full rank p, and therefore p linearly independent rows. If

$$\|\mathcal{A}(x)\| \geq n - p$$

(or the number of inactive indices B_+ is less than p) and if

$$\mathcal{A}_{B^+} = (a^i)_{i\in B^+} \; B^+ = \{i : x_i > 0\}$$

is regular, then we can simplify the problem: we combine the equations $x_i = 0$ with the matrix $Ax = b$ and obtain a system of n equations with a unique solution x. This situation is called edge of the simplex.

Definition 12.1. *$x \in \mathfrak{S}$ is called edge if there exists a set $\tilde{A} \subset \mathcal{A}(x)$ such that $(A^T, (e_i)_{i\in\tilde{A}})$ is invertible.*

Hence, at an edge x we can extend A^T by $n - p$ columns to a regular matrix. This implies that at an edge x we have at least $n-p$ components of x equal to zero or only less than p components strictly larger than zero. We now characterize the edges of the $\mathfrak{S}$ and assume that $A \in R^{p\times n}$ has rank p. Then, we show that an extremal value is attained at the edges of $\mathfrak{S}$. This is the foundation of the simplex method.

Theorem 12.2. *$x \in \mathfrak{S}$ is an edge, if and only if A_{B^+} has full column rank where $B_+ = \{i : x_i > 0\}$.*

Necessarily, we have that $\|B_+\| \leq p$.

Proof. Let x be an edge. Then, there exists $\tilde{A}$ such that (A^T, e_i) is regular. Hence, $\tilde{A}$ contains $n-p$ elements. Let P be a permutation such that $P(e_i)_{i\in\tilde{A}} = (e_i)_{i=p+1,\ldots,n}$. Then, $P(A^T, (e_i)) = (PA^T, P(e_i))$ and since $P(e_i)$ has rank $n-p$, we obtain that the matrix consisting of the first p rows of PA^T (has p columns) is regular. Since $A^T_{B^+}$ is a subset of PA^T $A^T_{B^+}$ has p independent rows or A_{B^+} has p independent columns.

Conversely, let A_{B^+} have full column rank. Then, $r = \|B^+\| \leq p$ and hence $\|\mathcal{A}(x)\| \geq n-p$. Since A has rank p, there exists $s = p-r$ columns of A such that $(A_{B^+}, a_{i_1}, \ldots a_{i_s})$ is regular with $a_{i_j} \in \mathcal{A}(x)$. Denote by $\bar{A} = \mathcal{A}\backslash\{i_1, \ldots, i_s\}$. Then, the matrix $(A^T, (e_i)_{i\in\bar{A}})$ is invertible: for a permutation P such that $PA^T = (A_{B^+}, a_{i_1}, \ldots, a_{i_s})^T$ and $P(e_i)_{i\in\bar{A}} = (e_i)_{\|\bar{A}\|+1,\ldots,n}$ we have that the first p rows of PA^T are invertible. □

We obtain the following conclusions.

- $\mathfrak{S}$ has a most a finite number of edges since the set of subsets of $\{1, \ldots, n\}$ of column numbers of A is finite.
- For every B_+ and x being edge, there exists a set B of $p-$elements, such that A_B is invertible. Since A has rank p, we can expect at most p elements.

The next theorem is the basis of the simplex method.

Theorem 12.3. *x is an edge of $\mathfrak{S}$, if and only if x is an extremal point of $\mathfrak{S}$.*

Proof. Let x be an edge. Then, there exists $\tilde{A} \subset \mathcal{A}(x)$, such that $(A^T, (e_i)_{i\in\tilde{A}})$ is invertible. Assume that x is not an extremal point of $\mathfrak{S}$ and hence $x = \lambda x_1 + (1 - \lambda x_2$ with $x_1 \neq x_2$ and $0 < \lambda < 1$. Hence, $x_i = 0$ and $x_{1,2} \geq 0$ implies $x_{1,i} = x_{2,i} = 0$ and therefore $\mathcal{A}(x) = \mathcal{A}(x_1) = \mathcal{A}(x_2)$. This implies $(A^T, (e_i)_{i\in\tilde{A}})(x_1 - x_2) = 0$ due to the constraints and since this matrix is invertible, we obtain $x_1 = x_2$. Let x be an extremal point of $\mathfrak{S}$. Then, $B = (A^T, (e_i)_{i\in\mathcal{A}(x)}) = (b, 0)^T$. If the rank of B is less than n, then there exists $y \in \mathbb{R}^n$ with $y \neq 0$ and $By = 0$ and therefore $B(x + \lambda y) = (b, 0)^T$. This implies $\mathcal{A} \subset \mathcal{A}(x + \lambda y)$. For $i \notin \mathcal{A}$ we have $x_i > 0$ and hence for δ sufficiently small $x_i \pm \delta y_i > 0$ and therefore $x + \lambda y \in \mathfrak{S}$ for λ sufficiently small. Hence, x cannot be an extremal point. Hence, rank B is equal to n. If rank of B is n and since rank A is p and since the first p columns of B are equal to A we obtain that x is an edge with $\tilde{A} = \mathcal{A}(x)$ and necessarily $\|\mathcal{A}(x)\| = n - p$. □

If $\mathfrak{S}$ is compact, then every point in $\mathfrak{S}$ can be written as a convex combination of the edges. If $\mathfrak{S}$ is not compact, then there exists directions d such that $x + \tau d \in \mathfrak{S}$ for any τ and there again those who are not convex combinations of others. Those directions are called extremal directions and we will latter see that every point in $\mathfrak{S}$ can be written as a convex combination of extremal directions and edges.

Definition 12.4. *$d \neq 0$ is called direction in $\mathfrak{S}$ if for all $x \in \mathfrak{S}$ and $\tau \geq 0$ we have $x + \tau d \in \mathfrak{S}$. d is called extremal direction if d is a direction and if additionally the following implication is true*

$$d = \sigma_1 d_1 + \sigma_2 d_2, d_i \text{ directions }, \sigma_i > 0 \implies d^1 = \beta d^2, \beta > 0.$$

Theorem 12.5. *The set M of all edges of $\mathfrak{S}$ is finite and non–empty. The set of all extremal directions is empty or finite. Then for every $x \in \mathfrak{S}$ we have the representation*

$$x = \sum \alpha_i x_i + \sum \tau_j d_j$$

for x_i being edges, d_j being directions and $\tau_j \geq 0$ and $0 \leq \alpha_i \leq 1$ with $\sum \alpha_i = 1$.

The theorem implies the existence of at least one edge. Furthermore, the cost functional $c^T x$ can be represented with the help of the edges. Hence, we immediately derive the following result.

Theorem 12.6. *Assume that the linear programming problem is in standard form and let $\mathfrak{S}$ be non empty. Then, either $c^T x$ is unbounded on $\mathfrak{S}$ or there exists an edge where $c^T x$ is maximal.*

Proof. Due to the previous theorem we have that either the set of extremal directions is empty or finite. If the set of extremal directions is empty, then the cost functional is given by

$$c^T x = \sum_{i=1}^{s} \alpha_i c^T x_i$$

for the set of edges $x_i, i = 1, \ldots, s$ which is non–empty. Hence, it suffices to compute $i_0 = argmax_{i=1,\ldots,s} c^T x_i$. Clearly, we maximize $c^T x$ by setting $\alpha_{i_0} = 1$. If the set of extremal directions is finite and if there exists a i_0 with $c^T d_{i_0} > 0$ then the problem is unbounded, since we can move along $\tau_{i_0} d_{i_0}$ for $\tau_{i_0} \to \infty$. if all $c^T d_{i_0} \leq 0$, then we maximize $c^T x$ by setting $\tau_i = 0$ and proceed as in the case of no extremal directions. □

Now, for the algorithm we go back to the previous findings: for every edge x we find a set of p elements, such that A_B is invertible and $\{i : x_i > 0\} \subset B$. These sets are called bases and the x_i with $i \in B$ are called variables of the base. A base can be assigned to any edge x. The optimal value will always be attained at an edge, but since it is not unique, it does not necessarily be obtained only(!) at an edge. In principle, one can proceed as follows: consider all sets of bases B and solve $A_B t_B = b$. If additionally $t_B \geq 0$, then $x = (t_B, 0_{i \notin B})$ is a candidate for optimality. Now, compare all results and choose the one with largest possible value of the cost functional. However, the possible set of bases is n over p and this grows like $\exp(p)n^p$. A more systematic approach is the simplex method which proceeds from one edge to the other by using a descent in the cost functional. However, in the worst case it explores all edges and has therefore exponential complexity.

12.1.1 The Simplex Method

We present the simplex method under the following assumptions and within the following framework.

- We assume that rank $A = p, \mathfrak{S} = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ and $b \geq 0$. Furthermore, we assume that $\mathfrak{S} \neq \emptyset$.
- If x is an edge, then B^+ contains exactly p elements. In general, we only have $\|B^+\| \leq p$ and can extend B^+ until it has p elements. This case is called degenerated (“”entartet“”).
- An initial edge x^0 is known.

Definition 12.7. *Two edges x_1 and x_2 are called neighbors if $\|B(x_1) \cap B_{x_2}\| = p - 1$.*

This implies that only one component changes when moving along neighbors. All others remain positive. Next, we compute cost functional and matrix in terms of a base of an edge. We have $Ax_0 = b$ since x_0 is feasible and write

$$A = (a_1, \ldots, a_n)$$

and B_0 the set of vectors for the edge x_0. Due to our assumption we have that A_{B_0} is regular and we split for a general x

$$A = (A_{B_0}, A_{\bar{B}_0}) \implies x_{B_0} = A_{B_0}^{-1} b - A_{B_0}^{-1} A_{\bar{B}_0} x_{\bar{B}_0}$$

Evaluating the cost functional in the previous expressions yields

$$c^T x_0 = c_{B_0}^T \left(A_{B_0}^{-1} b - A_{B_0}^{-1} A_{\bar{B}_0} x_{\bar{B}_0} \right) + c_{\bar{B}_0}^T x_{\bar{B}_0} = c_{B_0}^T A_{B_0}^{-1} b + d^T x_{\bar{B}_0}$$

for $d = (c_{\bar{B}_0}^T - c_{B_0}^T A_{B_0}^{-1} A_{\bar{B}_0})$. Hence, the cost functional at the edge x_0 can be written as a function of $n-p$ variables. Since $x_{\bar{B}_0} = 0$ at the edge x_0 we obtain the following expressions

$$c^T x_0 = c_{B_0}^T A_{B_0}^{-1} b, \; c^T x = c^T x_0 + d^T x_{\bar{B}_0}$$

with $x_{\bar{B}_0} \geq 0$. This implies in particular the following theorem

Theorem 12.8. *Let x_0 be an edge of the $\mathfrak{S}$ with base B_0 and set $d_{B_0}^0 := 0$. Furthermore, set $d_{\bar{B}_0}^0 = c_{\bar{B}_0} - A_{\bar{B}_0}^T (A_{B_0}^{-1})^T c_{B_0}$.*

x^0 is a solution to the linear programming problem

$$\max c^T x \text{ subject to } Ax = b, x \geq 0$$

if and only if

$$d^0 \leq 0.$$

Given x^0 we have for every x the following two equalities

$$x_{B_0} = A_{B_0}^{-1} b - A_{B_0}^{-1} A_{\bar{B}_0} x_{\bar{B}_0}$$

$$c^T x = c^T x_0 + d_{\bar{B}_0}^T x_{\bar{B}_0}.$$

and the restriction $x \geq 0$. Hence, If $d_l^0 > 0$ for $l \in \bar{B}_0$, then we increase the functional value by moving from x^0 to a new x. At x_0 we have that the lth component $x_{0,l} = 0$. Hence, we obtain for a new point

$$x = x_0 + t^l e_l.$$

for some t^l the functional value $c^T x = c^T x_0 + t^l d_l^0$. However, we can only to that as long as x_{B_0} stays positive. Hence, if $(Id, A_{B_0}^{-1} A_{\bar{B}_0}) e_l \leq 0$ (need to add Id to obtain the correct dimension), then x_{B_0} stays positive for all t^l. Therefore, we obtain an unbounded problem.

Theorem 12.9. *Let x_0 be an edge of the $\mathfrak{S}$ with base B_0 and set $d_{B_0}^0 := 0$. Furthermore, set $d_{\bar{B}_0}^0 = c_{\bar{B}_0} - A_{\bar{B}_0}^T (A_{B_0}^{-1})^T c_{B_0}$.*

If for some $l (\in \bar{B}_0)$ $d_l^0 > 0$ and $(Id, A_{B_0}^{-1} A_{\bar{B}_0}) e_l \leq 0$, then the problem is unbounded.

Sometimes, we also use $(Id, A_{B_0}^{-1}A_{\bar{B}_0})e_l = A_{B_0}^{-1}A_l$ which is true since $l \in \bar{B}_0$. Hence, it remains to discuss the case where for some $k \in B_0$ we have

$$e_k^T(Id, A_{B_0}^{-1}A_{\bar{B}_0})e_l > 0.$$

Now, we cannot increase the component x_l arbitrarily. We have to stop as soon as the kth component of $x_{B_0} = 0$. This implies a bound on t^l which can be computed explicitly by

$$0 \leq x_l \leq \min\{x_{0,k}/t_k^l : t_k^l = e_k^T(Id, A_{B_0}^{-1}A_{\bar{B}_0})e_l > 0, k \in B_0\} = \delta$$

We obtain the new (edge) as

$$x_{1,i} = \begin{pmatrix} \delta & i = l, \\ x_{0,i} - \delta((Id, A_{B_0}^{-1}A_{\bar{B}_0})e_l)_i & i \in B_0\backslash k, \\ 0 & i = k \\ 0 & i \in \bar{B}_0\backslash l \end{pmatrix}$$

This is again an edge with $B_1 = B_0 \cup \{l\}\backslash k$ and this is a neighbor to x_0. We have that A_{B_1} is regular. Here, x_l is base and x_k becomes a non–base variable. This can be efficiently done using Jordan elemination. For numerical stability we need to use the Pivot element for the exchange of rows k and l. The strategy is as follows: Obtain an index l such that $d_l^0 > 0$. We choose l such that

$$l = argmax\{d_j^0 : d_j^0 > 0\}.$$

Choose an index k such that the quotient $x_{k/t_k^l}^0$, $t_k^l = (e_k^T A_{B_0}^{-1}A_{\bar{B}_0}e_l$ is minimal with respect to k. This is as discussed above and is valid in the case of non–degenarcy. In case of degenerate edges we can exchange k and l without changing the cost functional and without moving to a new edge. In this case it is possible that the simplex algorithm cycles and does not converge. There exists strategies to prevent cycling. Under the previous hypotheses this is not possible since every new iterate is an edge different from the current one.

Theorem 12.10. *Under the assumptions of the previous section the simplex method converges in a finite number of steps to an optimal solution if it exists. If $c^T x$ is unbounded on $\mathfrak{S}$ then this is detected within in a finite number of steps.*

Typically, the problem to find an initial edge x_0 of the simplex is difficult. However, the simplex method itself can be used to find an initial edge. To this end we consider the problem for $1 = (1, \ldots, 1)^T \in \mathbb{R}^n$ and

$$\begin{aligned} -1^T y &\to \max \\ Ax + y = b, x \geq 0, y \geq 0, b \geq 0 & \end{aligned}$$

This is a linear programming problem in standard form with $(x, y) \in \mathbb{R}^{2n}$. A initial edge is $x^0 = (0, b)^T$. The edge is not degenerated if $b_i > 0$ for all i. Then, we solve the linear programming problem by the simplex method. If at the solution (x^*, y^*) satisfies $-1^T y < 0$, then $\mathfrak{S} = \{Ax = b, x \geq 0\} = \emptyset$ and the problem is infeasible, since $y \geq 0$ and $-1^T y < 0$ implies that there exists at least one $y_i > 0$ and hence $Ax \neq b$. If $y^* = 0$, then x^* is a feasible solution to $Ax^* = b$. Furthermore, the simplex has terminated at an edge $(x^*, 0)$. Hence, if the base does not contain any y^* variables, then we already have a base for x^* and we can start on the original problem. If the base contains y^* variables, then the edge is degenerated. If rank (A) $= p =$ then we can remove y variables at the expense of x variables without changing the cost functional and proceed as before. For other approaches we refer to the literature.

12.1.2 Dual Problems and Applications

Using the KKT system we can rewrite the (primal) linear programming problem

$$\max c^T x \text{ on } \mathfrak{S} = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$$

as dual problem

$$\min(-\mu)^T b \text{ on } \mathfrak{S}_D = \{\mu \in \mathbb{R}^p : -c \leq A^T \mu\}.$$

For both cases the KKT system is sufficient and necessary and we have for the Lagrangian

$$L = -c^T x - \mu^T (Ax - b) - \xi^T x$$

$$-c - A^T \mu - \xi = 0, \xi \geq 0, Ax = b \Leftrightarrow -c \leq A^T \mu, Ax = b$$

and the latter is the optimality condition for the dual problem. Clearly, the dual problem can be rewritten in standard form and the previous theorem

also covers the existence of an optimal solution to the dual problem. Since the optimal solution for primal and dual problem satisfy the same KKT system, necessarily, the multiplier of the one is the variable of the other. Hence, we have the fundamental theorem of linear programming.

Theorem 12.11. *If the feasible set $\mathfrak{S}$ of the primal problem is not empty and if $c^T x$ is bounded from above on $\mathfrak{S}$, then $c^T x$ attains its maximum on $\mathfrak{S}$. Then, the feasible set $\mathfrak{S}_D$ of the dual problem is not empty and $-\mu^T b$ is bounded from below on $\mathfrak{S}$ and $-\mu^T b$ attains its minimum on $\mathfrak{S}$. The extremal values coincide. The converse is also true.*

12.1.3 Summary of main theoretical results on the simplex method and sketch of the algorithm

- Formulation of the problem.

$$\begin{aligned} & \tilde{c}^T \tilde{x} \to \min \\ & \tilde{B}\tilde{x} \geq \tilde{b} \\ \Leftrightarrow \quad & c^T x \to \min \\ & Ax = b \\ & x \geq 0 \end{aligned}$$

- Basic assumptions

$$x \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, rank(A) = p < n, b \geq 0$$

$$S = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\} \neq \emptyset$$

- Definitions

$$B(x) = \{i : x_i > 0\}, \bar{B} = \{i : x_i = 0\}$$

- **Theoretical background**

  Graphical solution. In the original formulation move the lines l, which are orthogonal to $\tilde{c}$ until they meet the straight lines of the restrictions.

  Mathematical theorem. x is edge, iff x is extrema

  Implications. An algorithm has only to search all edges to find the minimum, iff it exists. There is a condition to determine, if $c^T x$ is unbounded on S.

  Assumptions for the Basic Simplex. $x^0 \in S$ as edge is known. x is arbitrary edge, then $|B(x)| = p$. $\bar{B} = \{1, \ldots, n\} \backslash B$.

- **Termination criteria.**

  (1.1) Let $x \in S$ be an edge and $B = B(x), D = \bar{B}(x)$. Then x is optimal, iff $\lambda = c_D - (A_D)^T A_B^{-T} c_B \geq 0$.

  (1.2) Let $x \in S$ be an edge. Let $l \in B = \bar{B}(x)$. Then $c^T x$ is unbounded on S, iff $\lambda = c_l - (A_{\{l\}})^T A_B^{-T} c_B < 0$ and $A_B^{-1} A_{\{l\}} \leq 0$.

- **Algorithm**

  Description of the primal simplex method

  -1 Reformulate the problem for an edge x to

  $$B = B(x), \bar{B} = \bar{B}(x), x_B = A_B^{-1} b, f(x) = c_B^T A_B^{-1} b + (c_{\bar{B}}^T - c_B^T A_B^{-1} A_{\bar{B}}) x_{\bar{B}}$$

  1 Start with x^0, i.e. $x^0_{\bar{B}(x^0)} = 0$ and calculate $x^0_B, f(x^0)$

  2 Test (1.1) on failure proceed, else x^0 is optimum

  3 Test (1.2) forall $l \in \bar{B}(x^0)$, if fails for at least one l proceed, else $c^T x$ is unbounded on S

  4 We have $\exists l \in \bar{B} : \lambda^0_l < 0$ and $\exists i \in K \subset B : A_i^{-1} A_{\{l\}} > 0$

  5 Calculate $x_B = x^0_B - A_B^{-1} A_{\{l\}} t$, set $x_l = t$ and $x_i = 0, i \in \bar{B} \backslash \{l\}$, where

  $$t = \min_{i \in K} \{x_i / ((A_{B_0}^{-1}) A_{\{l\}})_i : k_i \in B_0, ((A_{B_0}^{-1}) A_{\{l\}})_i > 0\}$$

  6 A x_i is reduced to zero, f is reduced, since $f(x) = f(x^0) + \lambda^0_l t$

  7 Use $x^0 = x$ and restart with (1)

- **Details on the simplex methods** Most of the theory and derivations is taken from [15]. This introduction is based on bases and dictionaries which is more common then the formulation with edges and optimality given in the previous chapter.

- **An example** The main idea of the simplex algorithm is given below.

$$\max c^T x = 0,\ Ax \leq b,\ x \geq 0 \tag{12.152}$$

  For a problem with inequalities we introduce slack variables and rewrite

$$\max c^T x = 0,\ Ax + w = b,\ x \geq 0, w \geq 0 \tag{12.153}$$

Lets assume that $x = 0$ is a feasible solution, i.e. $w \geq 0$ for this choice. Lets assume $c > 0$. Then we have

$$w = b - Ax \tag{12.154}$$

and we call x nonbasic (indepedent) variables and w basic (or depedent variables). The idea is to keep $x_2, \ldots, x_n$ fixed and increase x_1. Then $c^T x$ will increase. However, we can only increase x_1 as long as $w \geq 0$. Since w depends on x it changes and yields bounds for the possible increase or decrease of x_1. After we finished this for x_1 we reformulate the constraints $Ax + w = b$. Since this is a linear system we can do any linear operation without changing the solution. Furthermore we may replace some terms of $c^T x$ by the corresponding expressions in w. The idea is to reformulate the problem in nonbasic and basic variables. Then we proceed as before, i.e. change a nonbasic variable subject to the bounds given by the basic variables. Note that after the first step one slack variable is set to zero deu to the constraints. Hence we have the same setting as before but in different variables. We repeat this process until there is no nonbasic variable in which $\tilde{c}^T \tilde{x}$ increases.

We have the following notations.

Definition 12.12 (Dictionaries, Bases)**.** *Each system of equations encountered along the caculations is called a dictionary.*

The variables depending and appearing on the left are called basic variables.

The variables appearing on the right (indepedent variables) are called nonbasic variables.

General simplex method (primal simplex) The general lp (linear programming) reads

$$\begin{aligned} &\max c^T x \\ &Ax \leq b \\ &x \geq 0 \end{aligned} \tag{12.155}$$

with the following dimensions $c, x \in \mathbb{R}^n, b \in \mathbb{R}^m, A \in \mathbb{R}^{m \times n}$. This problem can be rewritten introducing m slack variables $x_{n+1}, \ldots, x_{n+m}$

as

$$\begin{aligned} &\max \sum_{i=1}^{n} c_i x_i \\ &b_i - \sum_{j=1}^{n} a_{ij} x_j = x_{n+i} \quad i = 1, \ldots, m \\ &x_i \geq 0 \quad i = 1, \ldots, n+m \end{aligned} \tag{12.156}$$

The above setting is the initial dicitonary. As the simplex progresses it moves from one dictionary to another. Each dictionary has m basic variables and n nonbasic variables (independent). Let N denote the set of indices corresponding to the nonbasic variables. Initially, $N = \{1, \ldots, n\}$ and let B denote the nonbasic variables, i.e. $B = \{n+1, \ldots, n+m\}$. Down the road the problem considered reads

$$\begin{aligned} &\max c_0 + \sum_{i \in N} \tilde{c}_i x_i \\ &x_i = \tilde{b}_i - \sum_{j \in B} \tilde{a}_{ij} x_j \quad i \in B \\ &x_i \geq 0 \quad i \in N \cup B \end{aligned} \tag{12.157}$$

At each iteration one variable leaves the set of nonbasic variables N and another enters. The variable going to the basic variables is called *entering* variable. We choose this variable according to the following rule: Choose any variable which coefficient $\tilde{c}_i$ is positive, i.e. pick

$$k \text{ of } \{j \in N : \tilde{c}_j > 0\}. \tag{12.158}$$

If this set is empty we have an optimal solution. The increase of this variable will change the basic variables. Since $x_j = 0$ $j \in N$ except $j = k$ the update is given by

$$x_i = \tilde{b}_i - \tilde{a}_{ik} x_k \quad i \in B \tag{12.159}$$

We must ensure that those $x_i, i \in B$ are nonnegative. Hence we require

$$\tilde{b}_i - \tilde{a}_{ik} x_k \geq 0 \quad i \in B \tag{12.160}$$

We wish to take the largest possible value for the increase of x_k by

$$x_k = \left(\max_{i \in B} \frac{\tilde{a}_{ik}}{\tilde{b}_i} \right)^{-1} \tag{12.161}$$

We discuss the case $0/0$ and $\max < 0$ later. Now we can select the leaving variable (which will be zero afterwards). This variable leaves the set of dependent (basic) variables and enters the set of independent (nonbasic) variables, hence it is *leaving*. It is $l \in B$, s.t. $\frac{\tilde{a}_{ik}}{\tilde{b}_i} \quad i \in B$ is maximal.

After selecting k and l we have to rearrange the current dictionary to obtain the new dictionary. These are matrix operations and is called *pivot*.

We can use different strategies to single out k and l, iff the first choice or the maximization is not unique. We have to consider the degenerated cases $0/0$ and $\max < 0$.

For the matrix versions and the formulation of the steps to calculate refer to the literature. We state only the relation with matrix operations. A dictionary has the property that the basic variables are written as a linear function of the nonbasic variables. We can express this by

$$Ax = b \implies Bx_B + Nx_N = b \tag{12.162}$$

where x_B are the basic variables and x_N are nonbasic variables. Since x_B can be written in terms of x_N we can invert B and see

$$x_B = B^{-1}\left(b - Nx_N\right) = B^{-1}b - B^{-1}Nx_N \tag{12.163}$$

The objective function is written as

$$c^T x = c_B^T B^{-1} b - \left(\left(B^{-1}N\right)^T c_B - c_N\right)^T x_N \tag{12.164}$$

The dictionary can therefore be rewritten in the above formulations.

Comparing with the notation of above we have

$$\begin{aligned} c_0 &= c_B^T B^{-1} b \\ (\tilde{c}_j)_{j\in N} &= c_N - (B^{-1}N)^T c_B \\ (\tilde{b}_i)_{i\in B} &= B^{-1}b \\ (\tilde{a}_{ij})_{i,j\in B} &= B^{-1}N \end{aligned} \tag{12.165}$$

which automatically gives the pivot (e.g. the transformation of one dictionary to another) since in the above formulas we only have to

change the underlying sets of basic and nonbasic variables. Note that B and N corresponds to a matrix partioned like

$$A = [BN] \tag{12.166}$$

The basic solution associated to a dictionary like above is

$$x_N = 0 \; x_B = B^{-1}b \tag{12.167}$$

- **Degeneracy** In the previous section we saw that the algorithm is not well-defined if the denominator is zero. We consider the different cases in 12.161.

- **Case** $\tilde{a}_{ik} = 0$ In this case we can skip the quotient $\tilde{a}_{ik}/\tilde{b}_i$ since there is no restriction on x_i.

- **Case** $0/0$. In this case we define

$$0/0 := 0 \tag{12.168}$$

  and proceed as before. This is a reasonable definition due to the inequality x has to posses.

- **Case** $\tilde{b}_i = 0$. This case can cause difficulties. We call a dictionary *degenerated* if $\tilde{b}_i$ vanish for some $i \in B$. The problems arise when the dictionary produces a *degenerated pivot*, i.e. one of the entering variables is $+\infty$. This can happen, when the numerator is positive (a_{ik}) and the denominator is degenerated. If we have a degenerated pivot we proceed without changing the current nonbasic variable. Usually one obtains then (in the next step) a pivot which is not degenerated. A problem only arise if we cycle, i.e. a sequence of degenerated pivots appear. We have the following result.

  Theorem 12.13. *If the simplex method fails to terminate, then it must cycle.*

  Remark 12.14. *There are choices for the entering and leaving variables s.t. the simplex terminates in all cases.*

- **Case** $\max < 0$**.** In this case the problem is unbounded, since we can increase the nonbasic variable above all bounds without violating any constraint. Hence the maximization problem has no solution.

- Finally, we state the fundamental theorem of linear programming.

  Theorem 12.15. *If there is no optimal solution, then the problem is either unbounded or infeasible. If an optimal solution exists then a basic optimal solution exists.*

- **Dual problems** Introducing dual problems yield upper bounds for the optimal solution. Formally, we multiply every constraint by y_i with $y_i \geq 0$. We sum up the corresponding inequalities and obtain a new functional. We choose the coefficients for y_i at least as large as in the objective function for x_i. This provides lower bounds for y_i. Since then the objective for x_i is bounded above by a sum of y_i we obtain a minimization problem for $\sum \tilde{d}_i y_i$. This problem is called the dual problem.

  Remark 12.16. *In general a dual optimization problem can alos be derived from the Lagrangian function. If this function has a saddle point one can prove KKT. The description of a saddle point can be done as minmax formulation. Evaluating the terms we obtain the dual problem for the Lagrangian multipliers. For more details see [13].*

  Definition 12.17 (Dual problem)**.** *Given a linear programming problem in standard form*

$$\begin{aligned} &\max c^T x \\ &\textit{subject to } Ax \leq b \\ &x \geq 0 \end{aligned} \tag{12.169}$$

  the associated dual problem is

$$\begin{aligned} &\min b^T y \\ &\textit{subject to } y^T A \geq c \\ &y \geq 0 \end{aligned} \tag{12.170}$$

  Straightforward calculations show that the dual of the dual is again the primal. The next theorem is also straight forward.

  Theorem 12.18. *If x is feasible for the primal and y is feasible for the dual problem, then*

$$c^T x \leq b^T y. \tag{12.171}$$

The next theorem is much harder to prove.

Theorem 12.19. *If x is optimal for the primal, then the dual as an optimal solution y s.t.*

$$c^T x = b^T y \tag{12.172}$$

A proof can be found in [15]. Initially in the dictionary we have that the dual dictionary is the negative transposed of the primal one. A simplex step in the primal method is performed. We choose an analogous pivot in the dual (neglecting the rules for pivoting given above). Using the same pivot in both problems we obtain again dictionaries which are negative transposed. Note that the as long as the primal is not optimal the dual solution is infeasible. But at any time the values of the objective function values coincide (between dual and primal).

12.2 Network Flow Problems

This section is mainly taken from [12] and corresponds to the chapter of network flows given by R. K. Ahuja.

We assume the standard knowledge about graphs and nodes. We just state that a tree is a connected acyclic (cycle free) graph. Each tree has at least two leaf nodes. An arc (i, j) is incident to nodes i and j. Adjacence arcs $A(i)$ to a node i are those arcs who emanate from node i. We consider a network $G = (A, V)$ and arc lengths c_{ij} associated to each arc $(i, j) \in A$. By (i, j) we denote an arc from node i to j.

We introduce several algorithms for searching, sorting and finally solving problems defined on networks.

Dijkstra's algorithm

Dijkstra's algorithm finds the shortest path from a source node s to all other nodes.

We assume

$$c_{ij} > 0$$

and s is connected to every node in the network by a path (only a technical assumption).

The algorithm proceeds as follows. Starting at s we move to each node i connected to s and label this node in order of its distance to s. These labels $d(i)$ are permanent, iff we know that it is shortest path from s to i. Otherwise it is temporary. If $(s, j) \notin A$ we set $d(j) = \infty$. Initially all labels are temporary. The minimum temporary label becomes permanet. Then all

arcs in $A(i)$ are scanned and the distance labels are updated. An update of the distance label happens only if the new distance is less than the previous one. Proof of correctness by induction over the set of permanent labels. The algorithm runs in $O(n^2)$. There are improvements possible, for example during selection of the minimum distance.

Maximum Flows

We consider a network with nonnegative capacities u_{ij} for any arc $(i,j) \in A$. We consider a source s and a sink t. We assume for every arc $(i,j) \in A$ (j,i) to be an arc in A, too. This is only technical, since we may set $u_{ji} = 0$. The problem reads

$$\max v \text{ subject to}$$

$$\sum_{j:(i,j)\in A} x_{ij} - \sum_{j:(j,i)\in A} x_{ji} = \begin{cases} v & i = s \\ -v & i = t \\ 0 & \text{else} \end{cases}$$

$$0 \leq x_{ij} \leq u_{ij} \quad (i,j) \in A$$

We introduce the notion of residual networks. Given a flow x on the network the residual capacity r_{ij} consists of two ingredients. First, we have $u_{ij} - x_{ij}$ which is the capacity left on the arc (i,j) and second we have the flow x_{ji} on the arc (j,i) which may be cancelled to increase the flow on (i,j). Hence,

$$r_{ij} = u_{ij} - x_{ij} + x_{ji}$$

We call the network with *positive* residuals the residual graph.

Ford and Fulkerson introduced an algorithm to solve the maxflow problem using residual graphs. The idea is that for a given flow x we look for a path from s to t in the residual network. If such a path exists we augment (increase on $s->t$, decrease on $t->s$) flow on this path (whic implies that the residual decreases) and update the residual graph. We proceed until the network does not contain such a path. In the definition of the residual we therefore may obtain as much as u_{ij} flow. This corresponds to the maximal possible flow. The introducing of backwards flow x_{ji} is necessary in order to handle arcs where a priori is not clear if they distribute flow in the "correct" direction. The algorithm is summarized below.

`Ford/Fulkerson`

0 $x := 0$

1 **while** there is a path P from s to t in $G(x)$ (residual graph) **do**

2 $\Delta := \min\{r_{ij} : (i,j) \in P\}$

3 augment Δ units of flow along P and update $G(x)$

4 **end do**

We have to describe, how to find a path P from s to t in the residual graph $G(x)$ and to prove that the algorithm terminates with maximum flow. The labeling algorithm performs a search of the residual network by fanning out from the source s and building a tree of reachable nodes. It terminates if t is a leaf. Initially all nodes are unlabeled. In the first step the algorithms labels all nodes with $(s,i) \in G(x)$. Then it checks the adjacent arcs $A(i)$ for each node i labeled before and so on. If $t \in A(i)$ for some i it terminates with a direct path from s to t. (Note that we have to add a predecessor index to each labeled node i indicating the node that caused i to be labeled. This allows us to trace back)

In order to prove maximiality we introduce the following notations. A set $Q \subset A$ is a *cutset*, iff $G'(N, A-Q)$ is disconnected and no subset of Q has this property (i.e. Q splits the original network, especially it splits N). A cutset is called $s-t-$*cutset* if s and t belong to different subsets of the nodes denoted by S and $\bar{S}$.

For a given flow x we define the flow across a $s-t$-cutset by

$$F_x(S,\bar{S}) = \sum_{i\in S}\sum_{j\in\bar{S}} x_{ij} - \sum_{j\in\bar{S}}\sum_{i\in S} x_{ji} \tag{12.173}$$

The capacity of a $s-t-$cutset is defined as

$$C(S,\bar{S}) = \sum_{i\in S}\sum_{j\in\bar{S}} u_{ij} \tag{12.174}$$

We note that $v = F_x(S,\bar{S}) \leq C(S,\bar{S})$. This implies that $\max_x F_x(S,\bar{S}) \leq C(S,\bar{S})$. However, an even stronger result is obtained and known as max-flow min-cut theorem, namely

$$v = \max_x F_x(S,\bar{S}) = \min_{S,\bar{S}} C(S,\bar{S}) \tag{12.175}$$

Assume a maximal flow x. Define S the set of labeled nodes in $G(x)$ determined by the labelling algorithm. Since x is maximal $s \in S$ and $t \in \bar{S}$. There are no more paths to augment we have obtained a cut $(S,\bar{S})$. Furthermore $r_{ij} = 0$ for $i \in S, j \in \bar{S}$ since we cannot augment flow any more. $r_{ij} = 0$ implies $x_{ij} = u_{ij}$ and $x_{ji} = 0$. This implies $v = F_x(S,\bar{S}) = C(S,\bar{S})$ and since v is a lower bound for $C(S,\bar{S})$ we obtain equality.

Theorem 12.20 (Theorem 4.1 in [12]). *The maximum flow from s to t equals the minimum capacity of all $s-t-$cutsets.*

If the Ford/Fulkerson algorithm terminates we a $s-t-$cutset at hand. We already have seen, that if x is maximal the algorithm terminates. Each labeling iteration of the algorithm scans any node at most once, inspecting each arc $A(i)$. If all arc capacities are integral and bounded by U the capacity of a cutset $(s, N-\{s\})$ is at most nU. Since the labelling increases the flow at least by one unit in any run, it terminates with at most nU iterations.

There are a lot of different algorithms and theories known for min cost flow problems. To limit the representation we only discuss results and algorithms related to the Ford/Fulkerson algorithm. There are faster and more efficient algorithms known. The research on min cost flow problems is more elaboarted than on max flow problems. The max flow problem is a part of the min cost flow theory. It is easily obtained by setting the costs to $c = (0, \ldots, 0)^T, b = 0$ and introducing an edge x_{ts} with costs $c_{ts} = -1$ and unlimited capacity. The max flow problem can be transformed to this problem by adding the arc (t, s) with unlimited capacity and maximizing x_{ts}.

We introduce the negative cycle algorithm and the augmenting cycle property to solve the following problem for fixed c, u and v.

$$\begin{aligned} &\min c^T x \text{ subject to} \\ \sum_{j:(i,j)\in A} x_{ij} - \sum_{j:(j,i)\in A} x_{ji} &= b(i) \\ 0 \le x_{ij} &\le u_{ij} \quad (i,j) \in A \end{aligned}$$

13 Numerical methods for Optimization in Infinite Space Dimensions

13.1 Preliminiary discussion

We minimize a function $f : X \to \mathbb{R}$ on a feasible set $X_{ad} \subset X$ and a Banach space X. We denote by $f'(x) \in X' \equiv X^*$ the first derivative of f with respect to x which is in general an element in the linear space $L(X; \mathbb{R}) = X^*$. X^* is the space of linear functionals on X.

All algorithms will generate a sequence of trial points $x_k \in X$ which should have the following properties.

1. Stationary points: We assume that we have first order necessary conditions. Every point satisfying these conditions is called a stationary point.

2. Global convergence:

   We introduce a so–called stationary measure $\Sigma : X \to \mathbb{R}^+$ such that $\Sigma(w) = 0$ if w is stationary and $\Sigma(w) > 0$ else. An example in the case of unconstrained minimization is $\Sigma(w) = \|f'(w)\|_{X^*}$.

   Given a sequence x^k of trial points we say that it is globally convergent, if one of the following properties is satisfied

   (a) Every accumulation point of x^k is a stationary point

   (b) For some continuous stationary measure we have

   $$\lim \Sigma(x^k) = 0.$$

   (c) There exists an accumulation point such that x^k is stationary

   (d) For the continuous stationary measure Σ we have

   $$\liminf \Sigma(x^k) = 0$$

   Exercise 13.1. *Show that (b) implies (a) and (c) implies (d).*

3. Local convergence and rate of convergence.

   Let $\bar{x}$ be a stationary point. If there exists $\delta > 0$ such that for alll $x^0 \in X$ with $\|x^0 - \bar{x}\| \leq \delta$ and $x^k \to \bar{x}$ we have that

   $$\|x^{k+1} - \bar{x}\|_X = o(\|x^k - \bar{x}\|)$$

   we call the sequence superlinearly convergent.

   If we have

   $$\|x^{k+1} - \bar{x}\|_X = o(\|x^k - \bar{x}\|^2)$$

   we call the sequence quadratically convergent.

13.2 Descent Methods in Hilbert Spaces

We consider unconstrained problems in a Hilbert space X first[11]

$$\min_x f(x) \tag{13.1}$$

[11] Many results are also true for Banach spaces but the presentation simplifies in a Hilbert space since the product $<,>_{X,X'}$ is given by the scalar product on X due to the Riesz representation theorem. Hence, in a Hilbert space the product $f'(x)\xi$ for $\xi \in X$ is well - defined and equivalent to $< f'(x), \xi >$.

for $f : X \to \mathbb{R}$ a given functional. We want to construct a sequence of approximations x_i such that $x_i \to x^*$ and x^* is a local minimum where

$$f'(x^*) = 0.$$

Hence, the stationary measure is $\Sigma(w) = \|f'(w)\|$. We consider gradient based descent methods of the following general type

$$x_{i+1} = x_i + \tau_i s_i \tag{13.2}$$

where $\tau_i \in \mathbb{R}^+$ a suitable stepsize parameter and $s_i \in X$ is a descent direction defined below.

descent direction

Definition 13.2 (Descent directions)**.** *Given a Hilbert space X and $f : X \to \mathbb{R}$ twice Frechet differentiable. We say $s \in X$ is a descent direction for f, iff*

$$f'(x)s < 0 \tag{13.3}$$

The motivation is as follows: Consider the function $\phi : \mathbb{R} \to \mathbb{R}$ given by

$$\phi(t) = f(x_i + ts).$$

If s is a descent direction then

$$\phi'(0) < 0$$

and therefore f is decreasing along s_i. The descent might be very small, but we have additionally the estimate by Cauchy-Schwarz inequality

$$\phi'(0) = f'(x_i)s \geq -\|f'(x_i)\|\|s\|.$$

Hence, one additionally can require for a descent direction to satisfy a angular condition as follows for some $\eta \in (0, 1)$

$$f'(x)s < -\eta\|f'(x)\|\|s\|. \tag{13.4}$$

It remains to choose τ, for example according to the Goldstein-Armijo rule.

goldstein-armijo

Definition 13.3. *Let $f : X \to \mathbb{R}$ be twice Frechet differentiable and X be a Hilbert space; let $s \in X, 0 < \beta < \alpha < 1$ be given. We say $\tau \in \mathbb{R}^+$ satisfies the Goldstein-Armijo rule at a point $x \in X$, iff*

$$\alpha\tau\nabla f(x)s \leq f(x + \tau s) - f(x) \leq \beta\tau\nabla f(x)s \tag{13.5}$$

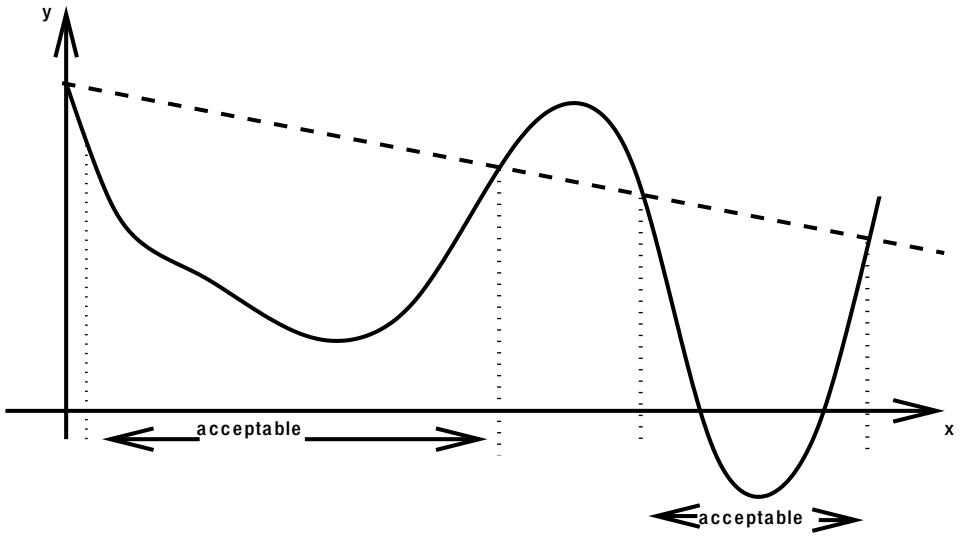


Figure 33: Geometrical interpretation of the Goldstein-Armijo rule. The graph $\tau \to f(x + \tau s)$ has to be inbetween the triangle spanned by the straight lines $\tau \to \alpha \nabla f(x)s + f(x)$ and $\tau \to \beta \nabla f(x)s + f(x)$. fig01

A geometrical interpretation of this rule is given in Figure 33.

The Goldstein-Armijo rule is also known in the case $\alpha = 0$, i.e. as one-sided estimate on the decrease of f. In this case we have to gurantee that τ_i does not tend to zero for $i \to \infty$. This will be discussed in Theorem 13.6.

The Goldstein-Armijo linesearch starts with an suitable large τ and decreases τ successively until the above equation is satisfied. This is important to prevent a sequence τ_i with $\tau_i \to 0$ for $i \to \infty$. We prove convergence under the assumption 13.3.

thm-goldstein **Theorem 13.4.** *Let $f : X \to \mathbb{R}$ be twice continuously Frechet differentiable. Let f and ∇f be weakly lower semicontinuous on the Hilbert space X. Let the sets*

$$\{x \in X : f(x) \leq M\} \tag{13.6}$$

be bounded in X for each $M \in \mathbb{R}$ and empty for M sufficiently small. We assume that $\tau_k \leq c$ and τ_k is choosen such that the Goldstein-Armijo rule is satisfied. Then, the sequence x_k generated by the descent method with descent direction $s_i := -\nabla f(x_i)$ has a weakly convergent subsequence. Then the limit point is a stationary point.

Example 13.5. *The assumption on the upper bound of τ_k is a technical assumptions only and can easily satisfied by for example $c := 1$. The boundedness of the defined set gurantees that there exists a weakly convergent subsequence. To obtain estimates on f we need that $\lim_k f(x_k)$ can be expressed in terms $f(\lim x_k)$. If we require f to be weakly lower semicontinuous, then there exists a solution and the problem is well-posed.*

Proof. First, note that f is bounded from below since the above set is empty for M sufficiently small. Second, note that s_i is a descent direction. By Taylor's formula we obtain for arbitrary $x_0 \in X$.

$$f(x_{i+1}) = f(x_i + \tau_i s_i) \leq f(x_i) \cdots \leq f(x_0) \tag{13.7}$$

Hence, $\{x_i\}$ is bounded by the assumption on the above set for $M \geq f(x_0)$. Then x_k has a weakly convergent subsequence which is also denoted by x_k. Let the limit be x^*. By definition of the iterates we know that

$$x_{k+1} - x_k = \tau_k s_k = -\tau_k \nabla f(x_k) \tag{13.8}$$

Hence,

$$\textstyle\sum_{k=0}^N \|x_{k+1} - x_k\|^2 = \sum_{k=0}^N -\tau_k \nabla f(x_k) \cdot (x_{k+1} - x_k) \leq \tag{13.9}$$

$$\textstyle 1/\beta \sum_{k=0}^N \tau_k \left(f(x_k) - f(x_{k+1})\right) \leq \tag{13.10}$$

$$\textstyle c/\beta \sum_{k=0}^N \left(f(x_k) - f(x_{k+1})\right) = c/\beta(f(x_0) - f(x_{N+1})) \leq p \tag{13.11}$$

where $p := c/\beta(f(x_0) - \inf_x f(x))$. Since p is independent of N we obtain $\sum_{k=0}^\infty \|x_{k+1} - x_k\|^2 \leq p < \infty$ and therefore there exists a subsequence denoted by x_k such that

$$0 \leftarrow \|x_{k+1} - x_k\|^2 = \| - \tau_k \nabla f(x_k)\|^2 \tag{13.12}$$

It remains to prove that there exists a lower bound on τ_k. We note that the whole sequence x_k belongs to a bounded set, see first lines of the proof. Then also the set of all points from the line segement $[x_k, x_{k+1}]$ is bounded and since $D^2 f$ is continuous, we have that $D^2 f(x_k)(\cdot, \cdot)$ is uniformly bounded for all iterates x_k and all line segements $[x_{k+1}, x_k]$. We denote this bound by c and obtain by Taylor's series

$$f(x_{k+1}) - f(x_k) \leq \nabla f(x_k)(x_{k+1} - x_k) + c/2\|x_{k+1} - x_k\|^2 \tag{13.13}$$

By the Goldstein-Armijo rule we have

$$\alpha \tau_k \nabla f(x_k) s_k = \alpha \nabla f(x_k) \tau_k s_k = \alpha \nabla f(x_k)(x_{k+1} - x_k) \leq f(x_{k+1}) - f(x_k) \tag{13.14}$$

Combining both inequalities we observe

$$\alpha\tau_k \nabla f(x_k) s_k \leq \nabla f(x_k)(x_{k+1} - x_k) + c/2 \|x_{k+1} - x_k\|^2 \tag{13.15}$$

$$\Leftrightarrow -\alpha\tau_k \|\nabla f(x_k)\|^2 \leq \tau_k \|\nabla f(x_k)\|^2 + c/2 \| - \tau_k \nabla f(x_k)\|^2 \tag{13.16}$$

$$\Longrightarrow \tau_k \geq \tfrac{2(1-\alpha)}{c} \text{ or } \nabla f(x_k) = 0 \tag{13.17}$$

Hence for $\nabla f(x_k) \neq 0$ the sequence of τ_k is uniformly bounded from below. This implies that $\| - \tau_k \nabla f(x_k)\|^2 \geq c \|\nabla f(x_k)\|$ and hence by the above formulas we conclude $\|\nabla f(x_k)\| \to 0$.

$$\|\nabla f(x^*)\| \leq \lim_k \|\nabla f(x_k)\| = 0 \tag{13.18}$$

since $\|\cdot\|$ and ∇f are weakly lower semicontinuous. □

If we do not reuqire the Goldstein–Armijo rule for selecting the stepsize a different convergence result under additional assumptions on the descent direction can be established.

thm-mu-1 **Theorem 13.6.** *Let f be twice continuosly Frechet-differentiable and bounded from below. Let x_k, s_k, τ_k be generated by the following algorithm (13.2), (13.3) and τ_k such that $f(x_k + \tau_k s_k) < f(x_k)$.*

Assume additionally that

$$\frac{f'(x_k)s_k}{\|s_k\|} \to 0 \implies \|f'(x_k)\| \to 0$$

and

$$f(x_k + \tau_k s_k) - f(x_k) \to 0 \implies \frac{f'(x_k)s_k}{\|s_k\|} \to 0.$$

Then, $\lim f'(x_k) = 0$. In particular, every accumulation of x_k is a stationary point.

Note that there might be many accumulation points and/or minimia depending on the initial datum x_0.

Proof. Let $f^* = \inf f(x_k) > -\infty$ due to the assumption of boundeness. Since $f_k = f(x_k + \tau_k s_k) < f(x_k)$ we have that $f_k \to f^* \in \mathbb{R}$. As in the proof of Theorem 13.4 we consider

$$\infty > f(w_0) - f^* = \sum_{k=0}^{\infty} (f(x_k) - f(x^{k+1})) = \sum |f(x_k + \tau_k s_k) - f(x_k)|.$$

Since the left hand side of the equation is bounded we obtain that the right hand side is absolutely convergent and therefore the terms in the sum generate a sequence tending to zero: $f(x_k + \tau_k s_k) - f(x_k) \to 0$. By assumption this implies

$$\frac{f'(x_k)s_k}{\|s_k\|} \to 0$$

and this in turn

$$\|f'(x_k)\| \to 0.$$

This proves the first part.

Now, consider an accumulation point $\bar{x}$. Then, there exists a subequence x_k such that $x_k \to \bar{x}$. We have $f_k \geq f(\bar{x})$ due to the monotonicity of f_k and the continuity of f. Hence, we apply the first part of the theorem to the subsequence and obtain $\lim f'(x_k) = 0$. Due to the continuity of f' we have $f'(\bar{x}) = 0$. □

Exercise 13.7. *1. Show that the Goldstein–Armijo rule implies the first and second assumption of the previous theorem.*

2. Show that the condition (13.4) implies the first assumption of the previous theorem.

Example 13.8. *We give a short note on the difference to the finite dimensional case in the setting of Theorem 13.4. Let $X = \mathbb{R}^n$ and $f : X \to \mathbb{R}$ be in $C^2(X;\mathbb{R})$. Similar to the above theorem we assume that*

$$\mathcal{L}_{\mathcal{M}} := \{x \in X : f(x) \leq M\} \tag{13.19}$$

is bounded for all $M \in \mathbb{R}$ and empty for M sufficiently small. We conclude from the above that f is bounded from below. By construction we have $f(x_k) \leq f(x_0)$ for all k and therefore there exists M such that the sequence $\{x_k\} \subset L_M$ and hence $\|x_k\|$ is uniformly bounded and by the Theorem of Weierstrass $x_k \to x^*$ strongly. This theorem strongly relies on the finite dimemsion of X. Since $x_k \to x^*$ strongly, we do not need any further assumptions on f or ∇f. Both are assumed to be continuous and hence $\lim_k f(x_k) = f(x^*), \lim_k \nabla f(x_k) = f(x^*)$. Therefore, the proof for the infinite dimensional case carries over to the finite dimensional case without any changes except for the last line: Here we conclude as described previously by the continuity of $\nabla f()$

$$0 \leftarrow \|\nabla f(x_k)\| \to \|\nabla f(x^*)\| \tag{13.20}$$

We now turn to different choices of the descent direction s in (13.2). Of course, the gradient $s_i := -\nabla f(x_i)$ always satisfies (13.2). This method is called **steepest descent method**. This direction also satisfies the angular condition 13.4.

13.2.1 Projection Methods For Constrained Problems

13.2.2 Quasi – Newton Methods

Using the finite dimensional case and a quadratic functional f we derive other algorithms than steepest descent and give motivations for the frequently used Newton-cg or Newton-SOR methods. Let us assume that A is spd. The steepest descent method with exact minimization for a quadratic functional reads

$$x^{k+1} = x^k - \sigma \nabla f(x^k) \tag{13.21a}$$

$$\sigma = \inf f(x^k - \sigma \nabla f(x^k)) \tag{13.21b}$$

$$f(x^k) := \frac{1}{2}(x^k)^T A x^k \tag{13.21c}$$

The discussion remains valid also for $f(x^k) = \frac{1}{2}(x^k)^T A x^k + b^T x^k$. This can be rewriten as

$$x^{k+1} = (I - \sigma A)x^k \tag{13.22}$$

with $\sigma = (x^k)^T A^T A x^k / (x^k)^T A^T A A x^k$. The above iteration is known as scaled Richardson method for iterative solving the linear equation $Ax = 0$. The scaled Richardson method is a fixed point method and hence it is convergent to first-order if the spectral radius $\rho(I - \sigma A) \leq 1$. Note that there exists an optimal σ for this method, namely $\min_\sigma \rho(I - \sigma A)$. To verify, that the steepest descent with exact minimization is well-defined and convergent, we note that $y^T A y / y^T y$ for $y \neq 0$ is the Raleigh coefficient and it holds for positive semidefinite matrices A that $0 \leq \lambda_{\min}(A) \leq y^T A y / y^T y \leq \lambda_{\max}(A)$. Hence, we conclude that $\sigma \geq 1/\lambda_{\max}(A)$. Since the extrem eigenvalues of $(I - \sigma A)$ are given by $1 - \sigma \lambda_{\min,\max}(A)$, we obtain $|1 - \sigma \lambda_{\max}| \leq 1$ and since $0 \leq \lambda_{\min}/\lambda_{\max}$ also $|1 - \sigma \lambda_{\min}| \leq 1$. This yields that the steepest descent with exact minimization is indeed a convergent method (for $A \in \mathbb{R}^{n \times n}$.) As seen in the above calculations, it is not necessary to compute the exact minimization to obtain convergence. Further, the method is only of first-order and for A spd other methods with convergence in finite number of steps (cg) are known. The relationship between iterative methods for solving linear equations and descent methods now yield additional descent methods. We briefly discuss

possible extensions of iterative methods for linear equations and descent algorithms for minimization.

One well-known method is the Jacobi method, i.e., $x^{k+1} = (I - D^{-1}A)x^k$. This corresponds to a descent method with $s_i := -D^{-1}\nabla f(x_i)$ which yields a descent if D^{-1} is positive definite and with $\sigma = 1$. There are relaxed versions possible and teh convergence is guranteed if A is strictly diagonal dominant. For D positive definite, the method belongs to the family of (Quasi-)Netwon methods discussed below.

Another method is the one-step method. Translate in the optimization context we have

$$x^{k+1} = x^k - \sigma e_i \; e_i^T A x^k \geq 0 \tag{13.23}$$

and $\sigma = (x^k)^T A e_i / a_{ii}$. Geometrical this method tries to minimize the functional by successively testing all unit directions. The move is only done, if $f(x^k)$ decreases. This method also applies for general nonlinear f and is very easy to implement. However, the convergence behaviour is very poor. In the case of A spd there exists a proof of convergence with a rate $1/cond(A)$.

Considering again the quadratic case with A positive definite, the method of choice would be the cg algorithm. We now present the general form of a cg-method for nonlinear optimization problems. The proof of convergence is similar to the one given for the gradient method.

1. Given an initial $x^0 \in X$, X Hilbert space and $s^0 = f'(x^0) \in X$. Then for $k = 0, 1, \ldots$
2. $x^{k+1} = x^k - \sigma s^k$ and σ such that sufficient decrease is granted.
3. $d^{k+1} = f'(x^{k+1})$
4. $s^{k+1} = d^{k+1} + (d^{k+1} - d^k, d^k)/\|d^k\| s^k$

Further extensions are methods of the Newton type. Although the computational cost is higher, they are known as the most powerful methods and should be the method of choice. For the abstract theorem refer to Theorem 5.20. In the case of unconstrained optimization one might try to apply Theorem 5.20 to the nonlinear operator equation

$$f'(x) = 0 \in X'$$

The following remarks should be taken into account. First, the good convergence behaviour of the Newton's method is only valid locally. Second, if

we try to embed this method into the above context the correct choice for s would be

$$s^{k+1} = -f''(x^k)^{-1} f'(x^k). \tag{13.24}$$

By the discussion above, the method defined by (13.24) is a descent method, iff $(f'(x^k), f''(x^k)f'(x^k))_{H,H} < 0$. Translating this to the finite dimenaional case, we obtain that $(f'')^{-1}$ needs to be positive definite. This of course is true in the minimum, but might be false in other parts of the domain with the consequence, that the actual s is not a descent direction. Another problem might be to obtain s^{k+1}. In the context of pde-constrained optimization we usually can solve

$$f''(x^k)s^{k+1} = -f'(x^k)$$

by an iterative process. The question therefore occurs, how exact we have to solve this equation to obtain a satisfactory overall performance. We do not discuss further details here and just give two examples of well-known optimzation methods and an abstract convergence theorem.

The Newton-cg method summarizes the discussion above.

1. Given an initial $x^0 \in X$, X Hilbert space. Assume $x^0 \in X$ is close to x^*, the local minimizer. Then for $k = 0, 1, \ldots$
2. Solve $f''(x^k)s^k = -f'(x^k)$ with the cg method. Couple the termination criteria for the cg method to the termination criteria for the overall method.
3. $x^{k+1} = x^k + s^k$

The inexact and Quasi-Newton methods, in particular the BFGS method, is given by the following abtract algorithm. Define $(w \circ z)v := (z, v)_X w$, where $w, v, z \in X$ and X is an Hilbert space.

1. Given an initial $x^0 \in X$, X Hilbert space and $H^0 \in L(X, X)$ symmetric and positive definite[12] Assume $x^0 \in X$ is close to x^*, the local minimizer. Then for $k = 0, 1, \ldots$
2. Solve $H^k s^k = -f'(x^k)$
3. $x^{k+1} = x^k + s^k$

[12]Note that an operator $A : X \to X$, X Hilbertspace, is called symmetric iff $(y, Ax) = (x, Ay)$ forall $x, y \in X$. The operator is called positive definite, iff $(x, Ax) \geq 0$ for all $x \in X$.

4. Update H^k by the BFGS update formula. Let $y := f'(x^{k+1}) - f'(x^k)$.

$$H^{k+1} = H^k + \frac{y \circ y}{(y, s^k)} - \frac{H^k s^k \circ H^k s^k}{(H^k s^k, s^k)}$$

Why using this update formula? Consider the finite-dimensional case, then $(w \circ z)v = (\sum_i z_i v_i w_j)_j = (wz^T)v$. The matrix wz^T has rank one. One can prove that the above is the only rank one update, which keeps the symmetry and the positive definitness, i.e. H^{k+1} is spd. Further, there exists an explicit update formula for the inverse of H^{k+1} in terms of y, H^k and s^k. The update is only valid if (y, s^k) is positive.

Finally, we present a convergence theorem (local result) due to Griewank[8].

Theorem 13.9. *Assume $x^* \in X$ is a local minimum, f is twice continuously Frechet differentiable and satisfies the second order optimality conditions*

$$f'(x^*) = 0 \tag{13.25a}$$

$$C\|x^*\| \geq f''(x^*) \geq c\|x^*\| \tag{13.25b}$$

Further, assume that $\|H^0 - f''(x^*)\|_{L(X,X)}$ and $\|x^0 - x^*\|$ are sufficiently small. Then the update formula is well-defined, i.e. H^k are spd and the method is linear(!) convergent. If additionally $H^0 - f''(x^*)$ is compact [13], then the method is superlinear convergent.

13.3 Augmented Lagrangian Methods

The reference for this part is the thesis of Maruhn [11] that is based on the algorithm by Sachs and Sartenaer.

The general setting is X, Y are Hilbert spaces and $f : X \to \mathbb{R}$ and $c : X \to Y$ are twice continuously Frechet differentiable maps. We want to find a minimizer of the following problem

$$\min f(x) \text{ subject to } c(x) = 0 \;\; x \in X \tag{13.26}$$

The Lagrange multiplier theorem states that if $c'(x^*)$ is surjective and x^* is a minimizer, then there exists a unique $\lambda^* \in Y$ such that the following equality holds

$$\nabla f(x^*) + c'(x^*)^* \lambda^* = 0 \tag{13.27}$$

[13] $A : X \to X$ is called compact, if $AB_1(0)$ is sequentially compact.

This is the necessary condition for the Lagrange functional

$$L(x, \lambda) = f(x) + (\lambda, c(x)) \tag{13.28}$$

The formula gives the motivation for augmented Lagrange methods. If we would knew λ^* then we can solve (13.27) by for example Newton's method or any other nonlinear solver. But for $\lambda \neq \lambda^*$ equation (13.27) does not incorporate the information $c(x) = 0$. A simple idea to improve this situation is to consider an augemented Lagrangian function which penalizes the constraint violation. This approach is also used in general penalty algorithms discussed below. The new objective function is

$$\Phi(x, \lambda, r) = f(x) + < \lambda, c(x) > + \frac{1}{2r}\|c(x)\|^2 \tag{13.29}$$

An algorithm for solving (13.27) tries to minimize Φ (unconstrained minizimation) and succesively updates λ and r.

Next we discuss various possibilites for updating λ. Obviously λ^* has to satisfy (13.27). Hence a choice for updating λ would be

$$\lambda_{k+1} := \text{argmin}_\lambda \|\nabla f(x_k) + c'(x_k)^* \lambda\| \tag{13.30}$$

The minimization can be computed exactly and is given by

$$\lambda_{k+1} = -[c'(x_k)^*]^\# \nabla f(x_k) \tag{13.31}$$

where $A^\#$ denotes the pseudo-inverse of the operator A. Unfortunately this operator might be very expensive to compute. Furthermore, since $\|x_k - x^*\|$ might be large, there is no need for the exact value of λ_{k+1}.

Sachs proposed the following update rule and we will later prove convergence for this particular case.

$$\lambda_{k+1} = \lambda_k + \frac{1}{\mu_k} c(x_k) \tag{13.32}$$

For given x, λ we will denote by $\bar{\lambda} := \lambda + \frac{1}{\mu} c(x)$.

We will use the following equality later

$$\nabla_x \Phi(x, \lambda, r) = \nabla f(x) + c'(x)^* \lambda + \frac{1}{r} c'(x)^* c'(x) \tag{13.33}$$

$$\nabla f(x) + c'(x)^* \bar{\lambda} = \nabla_x \Phi(x, \bar{\lambda}(x, \lambda, r)) \tag{13.34}$$

The augmented Lagrange algorithm computes iterates $x_k \in X$ which approximately solve the unconstrained optimization problem

$$\min_x \Phi(x, \lambda_k, r_k) \tag{13.35}$$

for given values λ_k, r_k. Further, we the algorithm gives an update rule for λ_k and r_k. The term "approximately" solve, means

$$\|\nabla_x \Phi(x, \lambda_k, r_k)\| \leq w_k \tag{13.36}$$

for some given tolerance w_k with $w_k \to 0$ for increasing k.

The general algorithm is given by the following steps

1. Given initial guesses for x_0, λ_0 and $r_0 < 1$ and $w_*, \eta_* << 1$ and parameters $\gamma_1, \tau < 1$ and $\gamma_2 > 1$

2. Solve for x_k

$$\min_x \Phi(x, \lambda_k, r_k) \tag{13.37}$$

   in the sense that

$$\|\nabla_x \Phi(x_k, \lambda_k, r_k)\| \leq w_k \tag{13.38}$$

3. Test for convergence: $\|\nabla_x \Phi(x_k, \lambda_k, r_k)\| \leq w_*, \|c(x_k)\| \leq \eta_*$.

4. Update of the multipliers depending on the constraint violation and goto Step 2.

   (a) If $\|c(x_k)\| \leq \gamma_1 \eta_k$ update the Lagrange multiplier
   Choose λ_{k+1} such that $\|\lambda_{k+1} - \bar{\lambda}(x_k, \lambda_k, r_k)\| \leq w_k$
   Let r_k be unchanged, $r_{k+1} = r_k$
   Decrease w_k to $w_{k+1} = r_{k+1} w_k$.
   Decrease η_k to $\eta_{k+1} = \sqrt{r_k}\eta_k$

   (b) If $\|c(x_k)\| \geq \gamma_2 \eta_k$ reduce the penalty parameter
   to $r_{k+1} = \tau r_k$.
   Let λ_k be unchanged, $\lambda_{k+1} = \lambda_k$
   Decrease w_k to $w_{k+1} = r_{k+1}$
   Decrease η_k to $\eta_{k+1} = \sqrt{r_{k+1}}$

   (c) Else do any of the above updates.

We prove the following lemma.

Lemma 13.10.

$$\lim_k w_k = \lim_k \eta_k = 0 \tag{13.39}$$

Proof. By step 4 we observe

$$0 < r_{k+1} \leq r_k \leq \cdots < 1 \tag{13.40}$$

Hence, there exists a limit $r_* \in [0,1]$. For the sequence w_k we note that $0 < w_k < 1$ and by step 4

$$w_{k+1} = r_{k+1} w_k < r_k \tag{13.41a}$$

$$w_{k+1} = r_{k+1} < r_k \tag{13.41b}$$

Hence $w_{k+1} < r_k$. So if $r_* = 0$ then $w_* := \lim_k w_k = 0$. If $r_* > 0$, we first show that $\exists \tilde{k}$ such that $r_k = r_{\tilde{k}}$ for all $k \geq \tilde{k}$. Having this result at hand we conclude that Step 4a is executed for all $k \geq \tilde{k}$ and hence $w_{k+1} = r_{\tilde{k}} w_k$ for all $k \geq \tilde{k}$. We obtain $w_{k+1} = r_{\tilde{k}}^l w_{\tilde{k}}$ for $k + 1 = l + \tilde{k}$. Therefore, $\lim_k w_k = 0$. It remains to prove that $r_* > 0$ implies $\exists \tilde{k}$ such that $r_k = r_{\tilde{k}}$ for all $k \geq \tilde{k}$. This equivalent to assume that step 4a is executed forall $k \geq \tilde{k}$. Assume this is not the case. Assume that n_l is the sequence of indices of all iterates where step 4b is executed. Then $r_{n_l+1} = \tau r_{n_l} = \tau^{l+1} r_{n_0}$. Since $\tau < 1$ we have $\lim_l r_{n_l+1} = 0$ which contradicts $r_* > 0$.

Analogoulsy one proves $\lim_k \eta_k = 0$. □

We assume the following on the problem to prove global convergence of the augmented Lagrangian method introduced above.

1. The mapping $f : X \to \mathbb{R}$ and $c : X \to Y$ are twice Frechet differentiable.

2. The iterates x_k are enclosed in a compact subset Ω of X. This assumption as posed in a infinite dimensional space is stronger than in finite dimensions (where bounded and closed implies compact).

3. At any limit point x^* of x_k the operator $Dc(x^*)$ is surjective

4. The convergence tolerances are $\eta^* = w^* = 0$. This is a technical assumption to prove the convergence.

Further we prove the convergence result with the slightly different Lagrange update rule in Step 4a.

$$\text{Choose } \lambda_{k+1} \text{ as } \lambda_{k+1} = \bar{\lambda}(x_k, \lambda_k, r_k) \tag{13.42}$$

This change does only affect technical difficulties in the proof. For a general proof see [11] Theorem 3.2.3 page 49.

Theorem 13.11. *Assume that all the above assumptions are fullfiled. Let $x^* \in X$ denote a limit point of the sequence (x_k) and let also x_k denote the sequence converging to x^*. Let $\lambda(x^*)$ be defined by $\lambda(x) = -\left(c'(x)^{\#}\right)\nabla f(x)$, i.e., that is the norm minimizer of $\min_\lambda \|\nabla f(x) + c'(x)^*\lambda\|_X$.*

Furhter, assume that x_k, λ_k, r_k are sequences generated by the algorithm above.

Then there exists positive constants κ_1, κ_2 such that

$$\|\lambda(x_k) - \lambda(x_*)\| \leq \kappa_2 \|x_k - x^*\| \tag{13.43a}$$

$$\|\bar{\lambda}(x_k, \lambda_k, r_k) - \lambda(x_*)\| \leq \kappa_1 w_k + \kappa_2 \|x_k - x^*\| \tag{13.43b}$$

$$\|c(x_k)\| \leq \kappa_1 w_k r_k + r_k \|\lambda_k - \lambda(x^*)\| + \kappa_2 r_k \|x_k - x^*\| \tag{13.43c}$$

Since $x_k \to x^*$ we have $\lambda(x_k) \to \lambda(x^*)$ and $\bar{\lambda} \to \lambda^*$ and by definition $\lambda(x^*) = \lambda^*$. Further, we obtain

$$\lim_k \nabla_x \Phi(x_k, \lambda_k, r_k) = \nabla L(x^*, \lambda^*) = 0 \tag{13.44}$$

Proof. By the previous lemma we have $\lim_k w_k = \lim_k \eta_k = 0$. First, we conclude that $c'(x)(\cdot)$ is surjective in a neighbourhood of x^*. This is due to the assumptions 1 and 3. Since $x_k \to x^*$ we know that for all k sufficiently large $c'(x_k)(\cdot)$ is surjective. We denote the subsequence again by x_k. Hence, $c'(x_k)^{\#}$ exists. By theorem we conclude that $c'(\cdot)()^{\#}$ is Lipshitz continuous in some neighbourhood of x^*. Therefore, $c'(\cdot)^{\#}$ is bounded and converges to $c'(x^*)()^{\#}$. Hence we deal with bounded operators c',

$$\| \left(c'(x_k)^{\#}\right)^* \| = \|c'(x_k)^{\#}\| \leq \kappa_1 \tag{13.45}$$

For x_k we know that

$$\|\nabla \Phi(x_k, \lambda_k, r_k)\| = \|\nabla f(x_k) + c'(x_k)^* \bar{\lambda}\| \leq w_k. \tag{13.46}$$

Now we show, that the update $\bar{\lambda}$ is not far from the norm minimizer of the Lagrange function with respect to λ. Since $Id = \left(c'(x_k)c'(x_k)^*\right)^{-1} c'(x_k)c'(x_k)^* = \left(c'(x_k)^{\#}\right)^* c'(x_k)^*$, we have

$$\|\bar{\lambda} - \lambda(x_k)\| = \|Id\ \bar{\lambda} - \left(c'(x_k)^{\#}\right)^* \nabla f(x_k)\| = \tag{13.47}$$

$$\| \left(c'(x_k)^{\#}\right)^* \left(c'(x_k)^* \bar{\lambda} - \nabla f(x_k)\right) \| \leq \kappa_1 w_k \tag{13.48}$$

It remains to show that $\lambda(x_k) \to \lambda^* \equiv \lambda(x^*)$ and that the constraint violation tends to zero. First, note that $\nabla f(\cdot)$ is continuous and hence bounded

for k sufficiently large. Furthermore, by theorem we obtain that $\nabla f(\cdot)$ is Lipshitz.

$$\|\lambda(x_k) - \lambda^*\| \leq \| \left(c'(x_k)^\#\right)^* \nabla f(x_k) - \left(c'(x^*)^\#\right)^* \nabla f(x^*)\|$$
$$\leq \| \left(c'(x_k)^\#\right)^* - \left(c'(x^*)^\#\right)^* \|\|\nabla f(x_k)\| + \| \left(c'(x^*)^\#\right)^* \|\|\nabla f(x_k) - \nabla f(x^*)\|$$
$$\leq \kappa_2 \|x_k - x^*\|$$

Next we prove that $\nabla_x \Phi$ vanishes at (x^*, λ^*).

$$\nabla_x \Phi(x_k, \lambda_k, r_k) = \nabla f(x_k) + c'(x_k)^* \bar{\lambda} \to \nabla_x L(x^*, \lambda^*) \tag{13.49}$$

since c', f are continuous and $\bar{\lambda}$ is convergent. Since $\lim w_k = 0$ we have

$$0 = \lim_k \nabla \Phi(x_k, \lambda_k, r_k) = \nabla L(x^*, \lambda^*) \tag{13.50}$$

and the necessary first order conditions are satisfied at (x^*, λ^*). Finally,

$$c(x_k) = r_k(\bar{\lambda} - \lambda_k) = r_k(\bar{\lambda} - \lambda^*) + r_k(\lambda^* - \lambda_k). \tag{13.51}$$

Now the modified update rule for λ_k comes into play: $\lambda_k = \bar{\lambda}(x_{k-1}, \lambda_{k-1}, r_{k-1}$. Then we can conclude that $c(x_k) \to 0$ and therefore (x^*, λ^*) is a Kuhn-Tucker point. For different update rules one has proceed differently from here one. We already know that $\bar{\lambda} \to \lambda^*$. Therefore, further technical lemmas are needed to prove a convergence for an update rule like step 4a. $\square$

13.4 Penalty algorithms

We consider again the equality constrained problem and try to transform this in an unconstrained problem. We discuss the conditions necessary for this approximation. We discuss the setting in Hilbert spaces which will be clear in the subsequent section.

The problem reads

$$\min f(x) \text{ subject to } g(x) = 0 \tag{13.52}$$

and we assume that f is a real valued functional on the Hilbert space X, $g : X \to Y$ is an operator mapping from X into the Hilbert space Y. We use the notation $< \cdot, \cdot >$ for the inner product on a Hilbert space. Using the previous results the first order necessary optimality conditions for the minimization problem state, that if x_0 is a regular point of $g(x)$ and a local minimum of f, then there exists $y_0^* \in Y^*$ such that

$$\nabla f(x_0) + < y_0^*, \nabla g(x_0) >= 0 \tag{13.53}$$

This conditions also holds if X, Y are Banach spaces. But in a Hilbert space we have that Y is isomorph to Y^* and we can reformulate the above and conclude that there exists $y_0 \in Y$ such that

$$\nabla f(x_0) + < y_0, \nabla g(x_0) >= 0 \tag{13.54}$$

This leads to the definition of the classical penalty function

$$\phi_r(x) := f(x) + \frac{1}{2r} < g(x), g(x) > \tag{13.55}$$

This formulation is not possible for Banach spaces, since $g(x) \in Y \neq Y^*$.

We consider the associated penalized subproblems

$$\min_{x \in X} \phi_r(x) \tag{13.56}$$

The necessary optimality conditions for (13.56) state that if x_r is optimal, then

$$\nabla f(x_r) + \frac{1}{r} < g(x_r), \nabla g(x_r) >= 0 \tag{13.57}$$

which is equivalent to

$$\nabla f(x_r) + < y_r, \nabla g(x_r) >= 0 \tag{13.58}$$

$$r y_r - g(x_r) = 0 \tag{13.59}$$

for $y_r \in Y$. The question is now, which assumptions do we need to conclude that for $r \to 0$, the optimal solutions x_r to (13.56) converge to x_0. Also interesting is to ask, whether we need to solve (13.56) exactly or if a sufficiently close solution x_r will be sufficient.

Theorem 13.12. *Let x_k be a point satisfying*

$$\phi_{r_k}(x_k) \leq \min_{x \in X} \phi_{r_k}(x) + \epsilon_k \tag{13.60}$$

for $r_k \to 0$ and ϵ_k bounded. Further assume that $f(x)$ and $< g(x), g(x) >$ are lower semicontinuous functionals on X.

For any limit point x_0 of the sequence x_k, it holds that x_0 is a global minimum for the constrained optimization problem $\min_x f(x)$ subject to $g(x) = 0$.

Proof.

□Note that this theorems does not gurantee the existence of a limit point x_0. Usually one additionally assumes that the iterates x_k belong to a compact subset of X. See for example thesis of Maruhn, assumptions in the book of Spellucci and so forth.

General reference for abstract penalty functions is Burke, SIAM 1991.

14 Interesting papers and notes

14.1 Zuazua, Controllability of partial differential equations

Example on the pendulum with mass $m = 1$ and gravitational force $g = 1$ the system reads

$$y'' + sin(y) = v.$$

Here, v is the control which is supposed to keep the pendulum near the value $y = \pi$ and where y is the angle of the arm with respect to the vertical axis measured clockwise. The basic idea to design a control is to linearize around the state $y = \pi$. In this case $\sin(y) \approx \pi - y$ and the linearized system in $\phi = y - \pi$ reads

$$\phi'' - \phi = v$$

The goal is to drive ϕ, ϕ' to zero using v. This suggest a feedback law for $\alpha > 0$

$$\phi = -\alpha\phi$$

since when $\phi > 0$ we have $y > \pi$ and action opposite to y is required. However, the resulting equation leads to

$$\phi'' + (\alpha - 1)\phi = 0$$

and the behavior is understood considering its eigenvalues: if $\alpha > 1$ then the roots are complex and the behavior is oscillatory. If $\alpha < 1$ then the roots are real but one is positive and the solution diverges to $\pm\infty$. The desired state is therefore *never* reached! The idea is then to change the feedback to

$$v = -\alpha\phi - \beta\phi'.$$

In this way we may impose exponential decay on ϕ for β sufficiently large. Remark: if we let the control act only at discrete points in time, then even the control $v = \pm 1$ is suitable to stabilize the discrete in time system. This is called bang-bang control and appears from Pontryagins maximum principle.

Existence of minimizers in the case of the direct method of variations. Case 1: full space. If H is Hilbert, $J : H \to \mathbb{R}$ is continuous, convex and coercive, then J attains its minimum. If J is strictly convex, then the minimum is unique.
Case 2:$K \subset H$. K closed convex, K bounded or J coercive, then, there exists a minimum of J over K and the minimum is unique provided that K is convex.

Controllability conditions: Problem is $x' = Ax + Bv, x(0) = x_0$ with $A \in \mathbb{R}^{N\times N}$, $B \in R^{N\times M}$,

- Controllability exists under the Kalman rank condition: $N = rang(B, AB, \ldots, A^{N-1}B)$. (equivalent to the invertibability of the Gramian matrix)
- Controllability can be formulated backwards requiring that the initial 'sees' the terminal state. Kalman rank condition is equivalent to an observability of the state

  $$|\phi^0|^2 \leq C \int_0^T \|B^T\phi\|^2 dt$$

  where ϕ solves the adjoint equation $-\phi' = A^T\phi, \phi(T) = \phi^0$. (equivalent is $B^T\phi = 0$ then $\phi = 0$.)
- Using optimal control

  $$J(\phi^0) = \frac{1}{2}\int_0^T |B^T\phi|^2 dt - (x^1, \phi^0), +(x^0, \phi(0))$$

  is strictly convex in ϕ^0 and continuous and coercive. It allows for a unique minimizer which is the solution to the adjoint system and the control for the forward system is $u = B^T\phi$.

Analogous conditions for solving linear systems: A surjective, iff A^T is injective.

14.2 Singler/Boggard: POD Approach to Control Theory

The basic idea of controling a linear system relies on Laplace transform. In particular, the Laplace transform of $f'(s)$ is $sF(s) - f(0)$. Hence, if we want to control a system of the type

$$x'(t) = Ax(t) + Bu(t)$$

then we obtain after Laplace transformation

$$sX(s) = AX(s) + BU(s).$$

This is an algebraic system and the transfer function is given by

$$G : U \to X$$

as

$$G(U(s)) = (sId - A)^{-1}BU(s).$$

For things depending also on the observation Cx we obtain

$$G = C(sId - A)^{-1}B.$$

Designing controllers can now be done based on approximations of the transferfunction. The hope is to have good transfer functions to have a good control, i.e., assume

$$\|G_r - G\|_\infty \leq \sum_{k>r} \sigma_k.$$

Then, instead of G we compute G_r and obtain a good control by applying the inverse to the current system's state

$$U(s) = G_r^{-1}x(s)$$

and transform with the inverse Laplace transform to obtain the original feedback law.

Different approaches to compute G_r exist: POD and interpolation.

appendix1

A Notation

1. For $h(x) : \mathbb{R}^n \to \mathbb{R}^m$ we denote by $Dh(x) \in \mathbb{R}^{m\times n}$ the derivative of h at x. By $\nabla h(x)$ we denote the transposed of $Dh(x)$, i.e., $\nabla h(x) = Dh(x)^T$.

2. A convex optimization problem means f is convex and $h(x)$ is affine linear.

3. NLO stands for $\min f(x)$ subject to $h(x) = 0$.

4. KKT holds means, there exists λ^* and x^* such that the Karush-Kuhn-Tucker system of equations is satisfied.

5. $f(t) = O(t) :\Leftrightarrow \exists C > 0 \exists t_0 \forall t \leq t_0 : f(t) \leq Ct$.

6. $f(t) = o(t) :\Leftrightarrow \forall C > 0 \exists t_0 \forall t \leq t_0 : f(t) \leq Ct$.

Equality(!) constrained problems, ONLY

1. MFCQ = LICQ

2. x^* local minimum, h affine linear $\implies$ KKT holds

3. x^* local minimum, MFCQ $\implies$ KKT holds, set of λ^* bounded

4. x^* local minimum, LICQ $\implies$ KKT holds, λ^* unique

5. NLO convex, KKT holds $\implies$ x^* is global minimum of NLO.

6. NLO convex, LICQ: KKT holds in (x^*, λ^*) (unique!) $\Leftrightarrow$ x^* is global minimum

7. Slater $\implies$ modified MFCQ

B Fast Facts on Sobolev Spaces

We recommend the book of Adams [1]. Introduction to Sobolev spaces and their properties. Spaces will be introduced as subspaces of $L^q(\Omega)$ where Ω is an open, domain in $\mathbb{R}^n$. Further assumptions will be given below.

If $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a n-tupel of nonnegative integers a_j we call α a multi-index and denote by x^α the polynomial $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ which is of the degree $|\alpha| = \sum_j \alpha_j$. Similarly, we denote by $D_j = \partial/\partial x_j$ and by $D^\alpha =$

$D_1^{\alpha_1} D_2^{\alpha_2} \cdots D_n^{\alpha_n}$, i.e. a differential operator of order n. Further, $D^{(0,\dots,0)}u = u$ and for example

$$D_1^2 D_2^1 u = (\partial x_1)^2 \partial x_2 u = \partial x_1 \partial x_1 \partial x_2 u \tag{2.1a}$$

$$n = 2, \sum_{0 \le |\alpha| \le 2} 1 = \sum_{(0,0),(1,0),(0,1),(2,0),(1,1),(0,2)} 1 = 6 \tag{2.1b}$$

We define the functional $\|\cdot\|_{m,p}$ where m is a non-negative integer and $1 \le p \le \infty$ by

$$\|u\|_{m,p} = \left(\sum_{0 \le |\alpha| \le m} \int_\Omega (D^\alpha u)^p dx \right)^{1/p} \tag{2.2a}$$

$$\|u\|_{m,\infty} = \max_{0 \le |\alpha| \le m} \|D^\alpha u\|_\infty \tag{2.2b}$$

Then the space $H^{m,p}(\Omega)$ is the completion of $\{u \in C^m(\Omega) : \|u\|_{m,p} < \infty\}$ subject to the $\|\cdot\|_{m,p}$ norm. On the other hand we can consider a subspace of $L^p(\Omega)$ with the following properties. We call $(D^\alpha u :=) v \in L^p(\Omega)$ the weak derivative of $u \in L^p(\Omega)$ if

$$\int_\Omega u D^\alpha \phi dx = (-1)^{|\alpha|} \int_\Omega v \phi dx \qquad \forall \phi \in C_0^\infty(\Omega) \tag{2.3}$$

Now, we define $W^{m,p}(\Omega) = \{u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega) \text{ for all } 0 \le |\alpha| \le m.\}$. The relation between these spaces is given by the Theorem of Meyers/Serrin: $H^{m,p}(\Omega) \equiv W^{m,p}(\Omega)$ for $1 \le p < \infty$ for every domain (open, connected) $\Omega \subset \mathbb{R}^n$. The counterexample for $p = \infty$ is the function $\Omega := \{x \in \mathbb{R} : -1 < x < 1\}$ and $u(x) = |x| \in W^{1,\infty}(\Omega)$ and $\notin H^{1,\infty}$. But, $H \subset W$ for all (m,p). In the case of boundary values (i.e. spaces $H_0^{m,p}(\Omega)$) we need additional assumptions on Ω to concluded.

$W^{m,p}(\Omega)$ is a Banach space, separable if $1 \le p < \infty$, reflexive for $1 < p < \infty$ and especially for $p = 2$ it is a Hilbert space with inner product

$$(u,v)_{m,p} = \sum_{0 \le |\alpha| \le m} \int_\Omega D^\alpha u D^\alpha v dx. \tag{2.4}$$

Let $1 \le p, q < \infty$ and $1 - \frac{n}{p} = -\frac{n}{q}$ and $u \in H^{1,p}(\mathbb{R}^n)$. Then we have

$$\|u\|_{L^r(\mathbb{R}^n)} \le C\mu\left(\{u \ne 0\}\right)^{\frac{1}{r}-\frac{1}{q}} \|\nabla u\|_{L^p(\mathbb{R}^n)} \tag{2.5}$$

for all $1 \le r < q$. For Ω bounded with Lipschitz boundary and $p = 2$ we additional obtain

$$\|u\|_{L^2(\Omega)} \tag{2.6}$$

References

Adams [1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, San Francisco, London, 1975.

A02 [2] H. W. Alt, *Lineare Funktionalanalysis*, Springer Verlag, Berlin, Heidelberg, New York, 2002.

AmbrosettiProdi1993 [3] A. Ambrosetti and G. Prodi, *A Primer of Nonlinear Analysis*, Cambridge University Press, Cambridge, New York, Melbourne, 1993.

BressanHJB [4] A. Bressan, *Viscosity solutions of hamilton-jacobi equations and optimal control problems*, in Tutorial, Penn State University, 2006.

Burger2003 [5] M. Burger, *Infinte-dimensional optimization and optimal design*, Lecture Notes, Department of Mathematics, UCLA (2003).

Coron2007 [6] J.-M. Coron, *Control and Nonlinearity*, AMS Publishers, Vol. 136, Providence, Rhode Island, 2007.

Evans1999 [7] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, 1999.

Griewank1987 [8] A. Griewank, *The local convergence of broyden-like methods on lipschitzian problems in hilbert spaces*, SIAM J. Numer. Anal., 24 (1987), p. 684.

Lions1971 [9] J. L. Lions, *Optimal Control of System of Partial Differential Equations*, Springer, New York Heidelberg Berlin, 1971.

Luenberger1969 [10] D. G. Luenberger, *Optimization by vector space methods*, John Wiley&Sons, Inc., New York, London, Sydney, Toronto, 1969.

Maruhn2001 [11] J. H. Maruhn, *An augmented Lagrangian algorithm for optimization with equality constraints in Hilbert spaces*, PhD thesis, Faculty of Virgina Polytechnic Institute and State University, 2001.

NKT89 [12] G. L. Nemhauser, A. H. G. R. Kan, and M. J. Todd, *Optimization*, Elsevier Science Publishers B.V., Amsterdam, New York, Oxford, Tokyo, 1989.

S93 [13] P. Spellucci, *Numerische Verfahren der nichtlinearen Optimierung*, Birkhäuser Verlag, Basel, Boston, Berlin, 1993.

Troeltzsch2002 [14] F. Tröltzsch, *Optimalsteuerung bei partiellen differentialgleichungen*, Lecture Notes, Technische Universität Berlin (2002).

V00 [15] R. J. Vanderbei, *Linear programming*, Kluwer Academic Publishers, Boston, London, Dordrecht, 2000.

Y71 [16] K. Yosida, *Functional Analysis*, Springer Verlag, Berlin, Heidelberg, New York, 1971.