Convex Optimization and an Introduction to Congestion Control

Lecture Notes

Fabian Wirth

August 29, 2012

Contents

Introduction

Convex optimization is a varied field which has in recent years seen enormous success because on one hand many practical optimization problems are in fact convex and on the other hand the concepts developed in the field allow for elegant results as well as powerful algorithms. Convex optimization problems occur in many fields, such as communication and networks, estimation and statistical signal processing, and control systems. We will sometimes touch on such problems but the main goal in the discussion of convex optimization is to provide the necessary tools for the discussion of congestion control techniques.

Of course, not all interesting or important optimization problems are convex. The methods used for nonconvex problems, however, are of quite different nature and we will not discuss this direction in these notes. Sometimes some concepts are of such generality that we will briefly discuss the general concepts, but this mainly happens in order to be able to point out how much more can be said in the convex case.

How to use these notes

These notes are a supplement to the course "Convex Optimization and Congestion Control" given within the framework of the Telecommunication Graduate Initiative. On the one hand they contain most of the material discussed in the course, with the possible exception of the odd example. On the other hand they supply proofs in those cases, where maybe time was not sufficient during the course to completely discuss proofs in detail. There is extensive literature on convex analysis and convex optimization and in writing these notes I have extensively drawn on [6, 10, 4]. There are other notable texts in this area and it is certainly advisable to have a look at [1, 11] to obtain an impression of the modern developments of the theory.

For the short discussion of congestion control material from [9, 3, 8, 7] has been used.

The notes cover the material discussed during the course. There are two exceptions to this rule: First, I sometimes found it reasonable to present a more general approach to the theory than what is really presented in class. I hope that for some it will prove insightful to have a view which is a bit broader. For those who find this intimidating, please just ignore the extra content and translate everything back to the discussion given during the lectures.

As the name suggests the course will mainly treat the general theory of convex optimization and discuss some applications of this theory in the area of congestion control. As there will be other courses within TGI focusing more on congestion control, the partition of time will be slanted towards the general discussion of convexity and convex optimization. There will be 10 lectures of two and a half hours each. The content of these lectures will be

- Lectures 1, 2, 3: Convexity and Convex Functions

- Lectures 4 and 5: Convex Optimization
- Lectures 6 and 7: Numerical Methods
- Lecture 8: Congestion Control
- Lecture 9: Utility Based Congestion Control

This plan has been made at the time of writing and is up for constant reconsideration as we see how we get along with the material.

My sincere thanks go to the organizers of the TGI, who have done an enormous job in putting together the whole initiative and in organizing everything so smoothly. I am very grateful for the opportunity to give this course.

Chapter 1

Convex Sets and Convex Functions

1.1 Convex Sets

In this chapter we study the concepts that are the building blocks for convex optimization: convex set and functions. It turns that the seemingly simple requirement of convexity imposes strong properties on sets as well as on functions. For sets the concept of separation is vital in deriving properties in the theory of convexity and ultimately in the area of convex optimization. Convex functions on the other hand are by the simple requirement of convexity almost differentiable in the sense that the subgradient is a well defined and well behaved object. From the point of view of optimization it is of course of utmost importance that local minima are automatically global. This is one of the reasons why in optimization theory the dividing line between simple and hard is along the line of convex versus nonconvex; whereas someone coming from dynamical systems might expect it to be between linear and nonlinear.

A convex set C is a set with the property that the line segment between any two points contained in C is a subset of C. So in principle all that is necessary to define convex sets is a vector space structure over $\mathbb{R}$. Our usage of convex sets will be restricted either to $\mathbb{R}^n$ or to the space of square integrable functions $L^2(J, \mathbb{R}^n)$ on an interval $J \subset \mathbb{R}$. We will therefore present some of the basic results in real separable Hilbert space X. Recall that a real *Hilbert space* is a complete vector space over $\mathbb{R}$ endowed with an *inner product* $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ with the properties

(i) $\langle \alpha_1 x_1 + \alpha_2 x_2, y \rangle = \alpha_1 \langle x_1, y \rangle + \alpha_2 \langle x_2, y \rangle$, for all $\alpha_1, \alpha_2 \in \mathbb{R}, x_1, x_2, y \in X$,

(ii) $\langle x, y \rangle = \langle y, x \rangle$, for all $x, y \in X$.

The space X is called separable, if there exists a countable dense set, or equivalently, if there exists a countable orthonormal basis for H. If the reader feels uncomfortable with such statements, then please always think of $\mathbb{R}^n$ with the *standard inner product*

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{n} x_i y_i \tag{1.1.1}$$

as a prime example of a Hilbert space.

A further example of a Hilbert space that will be important to us is the *space of symmetric matrices*

$$\mathcal{H}_n := \{ Q \in \mathbb{R}^{n \times n} ;\ Q^T = Q \} . \tag{1.1.2}$$

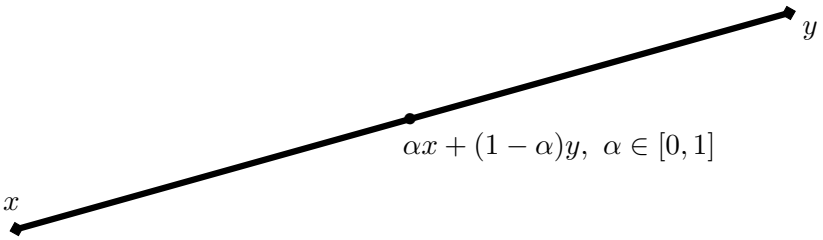

Figure 1.1: The line segment between x and y

This space becomes a Hilbert space together with the *Frobenius norm*

$$\|Q\|_F := \sum_{i,j=1}^{n} |q_{ij}|^2 \tag{1.1.3}$$

and the associated inner product

$$\langle P, Q \rangle = \operatorname{trace} PQ\,. \tag{1.1.4}$$

The definition at the beginning of the whole theory presented here is the following.

Definition 1.1.1 *Let X be a Hilbert space, then $C \subset X$ is called* convex *if for all $x, y \in C$, $0 \leq \lambda \leq 1$ it holds that*

$$\lambda x + (1-\lambda)y = y + \lambda(x-y) \ \in C\,.$$

In other words, C is convex, if for any $x, y \in C$ the *line segment*

$$[x, y] := \{\lambda x + (1-\lambda)y\,;\ \lambda \in [0,1]\}$$

is contained in C, see also Figure 1.1.

Convex sets come in many guises. Then can be open, closed or neither of the two, they can be bounded or unbounded, with smooth boundary or with corner points. A small zoo is depicted in Figure 1.2

On the other hand there are obviously nonconvex sets as the one shown in Figure 1.3. The line leaves the set, which excluded in the concept of convexity.

There are important classes of convex sets and we will briefly discuss some of them

(i) affine sets

(ii) hyperplanes

(iii) half-spaces

(iv) convex polytopes

(v) Euclidean balls and ellipsoids

(vi) convex cones

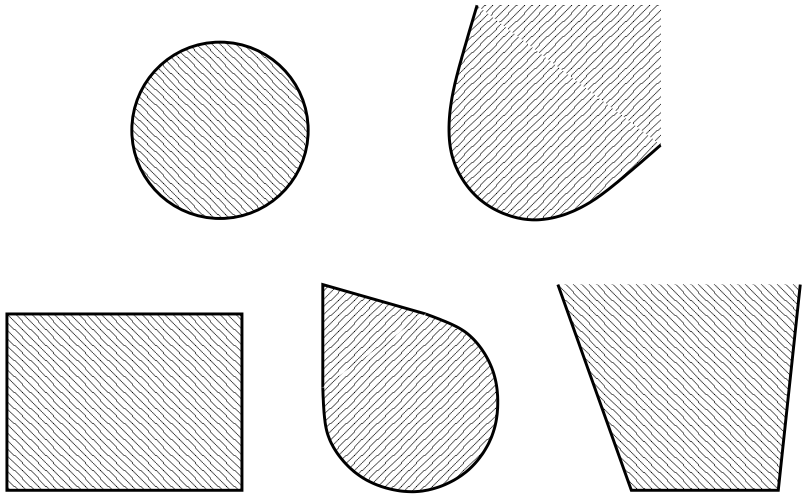

Figure 1.2: Convex sets

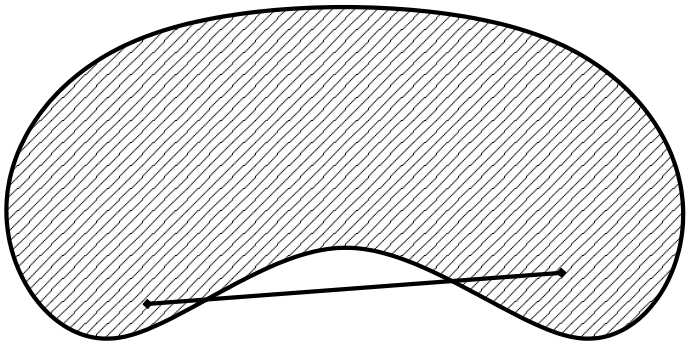

Figure 1.3: Definitely not a convex set.

Affine Sets

An *affine set* in a Hilbert space X is a linear subspace shifted by a constant vector. Thus a set of the form

$$\{b + v\,;\; v \in V\} \tag{1.1.5}$$

for some $b \in X$ and some linear subspace $V \subset H$. In $\mathbb{R}^n$ it is very often convenient to think of an affine space as the image of a linear map $F \in \mathbb{R}^{n\times n}$ shifted by a vector b, so that we arrive at a description

$$M = \{Fx + b\,;\; x \in \mathbb{R}^n\}\,. \tag{1.1.6}$$

The following obvious statement will be useful later.

Lemma 1.1.2 *Provided it is nonempty, the intersection of affine sets is an affine set.*

Hyperplanes

A special case of affine sets are *hyperplanes*, which are closed affine subspaces of codimension 1. In terms of the representation (1.1.5) this means for hyperplanes, that V is such that there exists a vector $z \notin V$ such that every vector $x \in X$ has a unique representation

$x = \gamma z + v$ for suitable and unique $\gamma \in \mathbb{R}$ and $v \in V$. In terms of the representation (1.1.6) for $\mathbb{R}^n$ this means that F has rank $n-1$. A hyperplane H has a unique representation in terms of a vector $q \neq 0$ orthogonal to H. In this case the inner product $\langle x, q\rangle$ is constant over H and we can write a further description of a hyperplane as

$$H = \{x \in X\,;\ \langle x, q\rangle = c\} \tag{1.1.7}$$

for a suitable constant $c \in \mathbb{R}$.

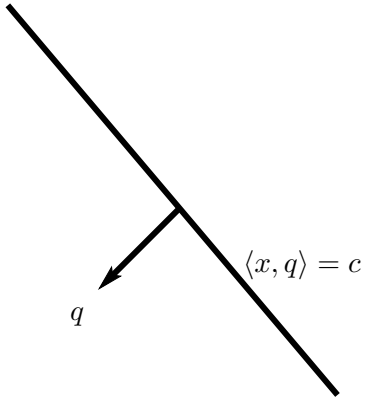

Figure 1.4: Hyperplane

Half-Spaces

A hyperplane separates a Hilbert space into two parts. The corresponding negative and positive *half-spaces*

$$H_+ := \{x \in X\,;\ \langle x, q\rangle > c\}$$
$$H_- := \{x \in X\,;\ \langle x, q\rangle < c\}\,,$$

which lie to either side of the hyperplane. The half-spaces described just now are open half-spaces. Depending on the circumstances it can also be convenient to consider closed half space, in which case the strict inequalities in the definition of H_+, H_- are simply changed to nonstrict ones. We will denote the closed half-spaces by $\overline{H}_+$, resp. $\overline{H}_-$. If we want to emphasize the defining data of a hyperplane we will write $H_+(q, c)$, etc.

Convex Polytopes

Convex polytopes are given as finite intersections of half-spaces. Thus a *convex polytope* is described by pairs $(q_i, c_i) \in X \times \mathbb{R}, q_i \neq 0, i = 1, \ldots, k,\ k \geq 2$. The set defined by these data is then given by

$$P := \{x \in X\,;\ \langle x, q_i\rangle \leq c_i, i = 1, \ldots, k\}\,. \tag{1.1.8}$$

In Figure 1.6 a bounded convex polytope is depicted. But note that in (1.1.8) nothing ensures that P is bounded.

Euclidean Balls and Ellipsoids

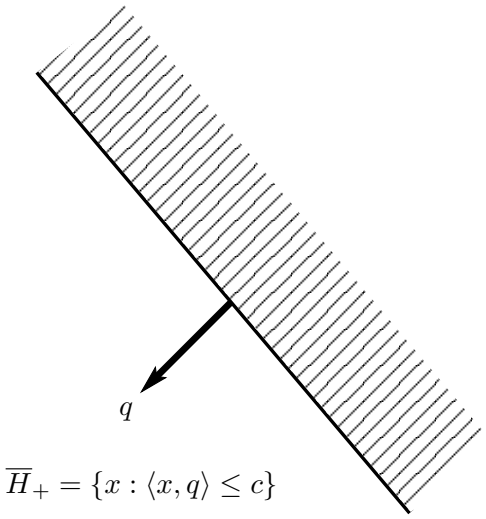

Figure 1.5: Half space

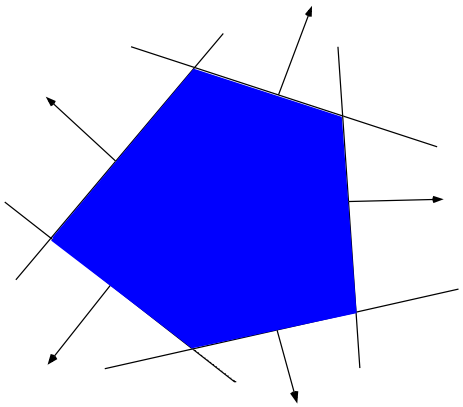

Figure 1.6: Convex Polytope

We will now speak of sets in $\mathbb{R}^n$ although they natural generalizations to general Hilbert spaces. The open ball of radius ϵ around a point $x^* \in X$ is defined by

$$B_\varepsilon(x^*) := \{x \in X\,;\ \|x - x^*\|_2 < \varepsilon\}\,.$$

It is easily seen to be convex.

An *ellipsoid* centered at a point x^* is given by a positive semidefinite matrix[1] $P \geq 0$ and a value $\varepsilon > 0$. The ellipsoid is the given by

$$\mathcal{E} := \{x \in \mathbb{R}^n\,;\ (x - x^*)^T P(x - x^*) < \varepsilon\}\,.$$

Again it is easy to see that this set is convex. It is bounded, if P is positive definite.

[1] When we speak of positive definite, or positive semidefinite matrices, we will always tacitly assume that the matrix in question is symmetric.

Convex Cones

Definition 1.1.3 (Convex Cones)

(i) A cone C *in* $\mathbb{R}^n$ *is a set with the property*

$$x \in C \;\Rightarrow rx \in C \quad \text{for all } r > 0.$$

(ii) A cone is called pointed, *if it does not contain a whole line.*

(iii) A convex cone *is a cone that is a convex set.*

Proposition 1.1.4 *A cone C is convex if and only if*

$$C = C + C\,.$$

Proof. Assume that $C = C + C$ and let $x, y \in C, \lambda \in (0,1)$. Then $\lambda x, (1-\lambda)y \in C$, as C is a cone. It follows from the assumption that $\lambda x + (1-\lambda)y \in C + C = C$. Thus C is convex.

To prove the converse, first note that for any cone we have $C \subset C + C$, as $x = 1/2x + 1/2x$.

We thus need to prove that $C + C \subset C$, if C is convex. So let $x, y \in C$. Then

$$x + y = \frac{1}{2}(2x + 2y) \in C\,,$$

by the cone property. This proves the assertion. ∎

Important cones are the *positive orthant*

$$\mathbb{R}^n_+ := \{x \in \mathbb{R}^n\,;\; x_i \geq 0, i = 1, \ldots, n\}\,, \tag{1.1.9}$$

or an *ice cream cone* as depicted in Figure 1.7. Sometimes the terminology ice cream cone is reserved for the specific cone the *Lorentz cone*. In $\mathbb{R}^n$ this cone is defined by

$$L^n := \left\{ x \in \mathbb{R}^n\,;\; x_n \geq \sqrt{x_1^2 + \ldots + x_{n-1}^2} \right\}\,, \tag{1.1.10}$$

in which case the cone is aligned with the axis x_k (and in $\mathbb{R}^3$ the ice cream is less likely to drop out of the cone).

In the space of symmetric matrices the set of positive semidefinite matrices form a cone that is of interest in many applications.

Convex cones are of interest in general, as they define an orders in their respective space as follows. Let K be a convex cone in a Hilbert space X, then the *partial order generated by K* is defined by

$$x \leq_K y \;:\Leftrightarrow\; y \in x + K\,. \tag{1.1.11}$$

Note that in order to prove *transitivity* of the order we need that K is convex. Indeed, transitivity of the order means, if $x \leq_K y$, and $y \leq_K z$, then it should follow that $x \leq_K z$. Otherwise we would not speak of an order. So given $y \in x + K, z \in y + K$, we can write $y = x + k_1, z = y + k_2$, so

$$z = x + k_1 + k_2\,.$$

To conclude that $z \in x + K$ we therefore have to know that $k_1 + k_2 \in K$ and this is true if the cone is convex by Proposition 1.1.4.

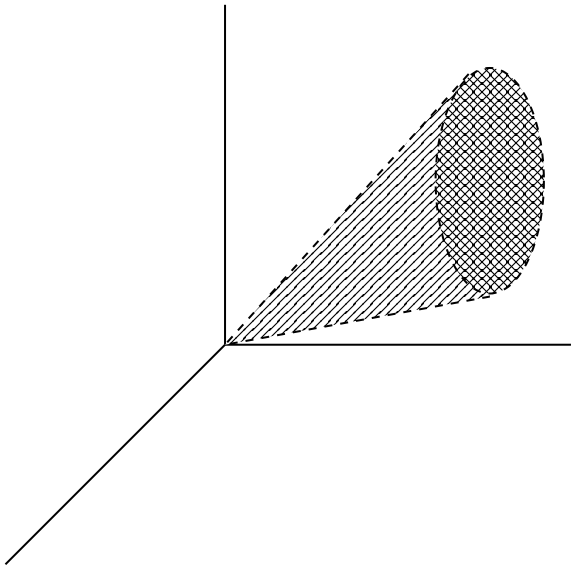

Figure 1.7: The ice cream cone.

1.1.1 Operations Preserving Convexity

It is of interest to characterize some operations on sets that preserve convexity. So that we have some tools to work with convex sets.

Lemma 1.1.5 *Let X be a Hilbert space, and $C_1, C_2 \subset X$ be convex. Then*

(i) $C_1 \cap C_2$ is convex,

(ii) $\operatorname{cl} C_1$ is convex, $\operatorname{int} C_1$ is convex,

(iii) The Minkowski sum $C_1 + C_2 := \{x + y \mid x \in C_1, y \in C_2\}$ *is convex.*

Proof. (i) If $x, y \in C_1 \cap C_2$, then for $\lambda \in [0,1]$ we have $\lambda x + (1-\lambda)y \in C_i$, $i = 1, 2$ because both C_1 and C_2 are convex. This shows that $\lambda x + (1-\lambda)y \in C_1 \cap C_2$.

(ii) If x, y are elements of the closure $\operatorname{cl} C_1$, then by definition there are sequences $x_k \in C_1, y_k \in C_2$ such that $x_k \to x, y_k \to y$. Fix $\alpha \in [0,1]$, then $\alpha x_k + (1-\alpha)y_k \in C_1$, as C_1 is convex. It follows by convergence that $\alpha x_k + (1-\alpha)y_k \to \alpha x + (1-\alpha)y \in \operatorname{cl} C$. Thus $\operatorname{cl} C_1$ is convex.

If $x, y \in \operatorname{int} C_1$, then there are $\varepsilon_1, \varepsilon_2$ such that $B_{\varepsilon_1}(x) \subset C_1, B_{\varepsilon_2}(y) \subset C_1$. For $\alpha \in [0,1]$ it follows that $\alpha B_{\varepsilon_1}(x) + (1-\alpha)B_{\varepsilon_2}(y) \subset C_1$. This defines an open neighborhood of $\alpha x + (1-\alpha)y$ contained in C_1, thus $\alpha x + (1-\alpha)y \in \operatorname{int} C_1$.

(iii) For $x, y \in C_1 + C_2$ we can by definition find $x_1, y_1 \in C_1, x_2, y_2 \in C_2$ such that $x_1 + x_2 = x, y_1 + y_2 = y$. Then for $\lambda \in [0,1]$ we have

$$\lambda x + (1-\lambda)y = \lambda x_1 + (1-\lambda)y_1 + \lambda x_2 + (1-\lambda)x_2 \in C_1 + C_2\,,$$

as desired. ∎

It is easy to see that the proof of Lemma 1.1.5 (i) extends to arbitrary intersections of convex sets. Note that here we are tacitly calling the empty set convex. This is justified by the definition, but it is only done out of convenience, so that in statements like

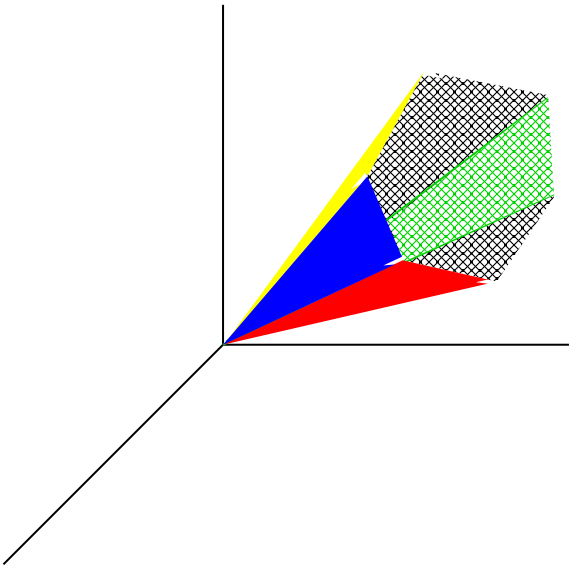

Figure 1.8: A polytopic cone.

Lemma 1.1.5 (i), we do not have to discuss the case that the intersection is empty separately.

A further property that is frequently useful is that convexity is preserved under affine maps, and also under the consideration of the preimage under an affine map.

Proposition 1.1.6 *Let $A \in \mathbb{R}^{n\times m}, b \in \mathbb{R}^n$.*

(i) If $C \subset \mathbb{R}^m$ is convex then the image of C under the map $x \mapsto Ax + b$ is convex, i.e.

$$AC + b := \{Ax + b\,;\; x \in C\} \quad \text{is convex}.$$

(ii) If $D \subset \mathbb{R}^n$ is convex, then the preimage of D under the map $x \mapsto Ax+b$ is convex, i.e.

$$A^{-1}(D - b) := \{x \in \mathbb{R}^m\,;\; \exists y \in D \text{ such that } Ax + b = y\} \quad \text{is convex}.$$

Proof. (i) Let $Ax_1 + b, Ax_2 + b \in AC + b$ and fix $\alpha \in [0, 1]$. Then

$$\alpha(Ax_1 + b) + (1 - \alpha)(Ax_2 + b) = A(\alpha x_1 + (1 - \alpha)x_2) + b\,.$$

As C is convex we have $\alpha x_1 + (1 - \alpha)x_2 \in C$ and so the convex combination on the right is an element of $AC + b$.

(ii) Let $x_1, x_2 \in A^{-1}(D - b)$ and fix $\alpha \in [0, 1]$. Then there exist $y_1, y_2 \in D$ such that $Ax_i + b = y_i, i = 1, 2$. Thus

$$D \ni \alpha y_1 + (1 - \alpha)y_2 = \alpha(Ax_1 + b) + (1 - \alpha)(Ax_2 + b) = A(\alpha x_1 + (1 - \alpha)x_2) + b\,.$$

This shows that $\alpha x_1 + (1 - \alpha)x_2$ is mapped to D. ∎

In Figure 1.1.1 we see an example how the Minkowski sum of two convex sets might look.

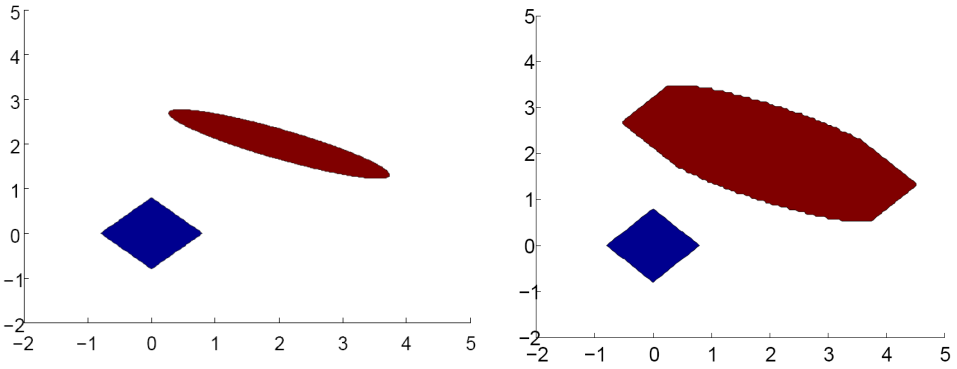

Figure 1.9: Left: two convex sets; right: one of the convex sets and the Minkowski sum of the original two sets.

Definition 1.1.7 *Let X be a Hilbert space, and $M \subset X$. The* convex hull *of M is defined by*

$$\operatorname{conv} M := \left\{ \sum_{k=1}^{m} \alpha_k x_k \mid m \in \mathbb{N}, \alpha_k \geq 0, \sum_{k=0}^{m} \alpha_k = 1, x_k \in M \text{ for } k = 1, \ldots, m \right\} .$$

The *convex hull* is not a misnomer as it is really a convex set. Indeed, if $y, z \in \operatorname{conv} X$ and $\lambda \in [0,1]$, then we have for suitable $\alpha_k \geq 0, \sum_{k=0}^{m} \alpha_k = 1, x_k \in X$ and $\alpha'_k \geq 0, \sum_{k=0}^{m'} \alpha'_k = 1, x'_k \in X$ that

$$y = \sum_{k=1}^{m} \alpha_k x_k \,, \quad z = \sum_{k=1}^{m'} \alpha'_k x'_k \,.$$

Then $1 = \lambda + (1-\lambda) = \sum_{k=1}^{m} \lambda\alpha_k + \sum_{k=1}^{m'} (1-\lambda)\alpha'_k$, the new coefficients are of course all in $[0,1]$ and by definition

$$\lambda y + (1-\lambda) z = \sum_{k=1}^{m} \lambda\alpha_k x_k + \sum_{k=1}^{m'} (1-\lambda)\alpha'_k x'_k \in \operatorname{conv} X \,.$$

So that $\operatorname{conv} X$ is convex.

One of the fundamental results in finite-dimensional convexity theory is *Carathéodory's Theorem*, which reduce the complexity in describing the convex hull of a set in $\mathbb{R}^n$.

Theorem 1.1.8 (Carathéodory) *Let $M \subset \mathbb{R}^n$ then for every $x \in \operatorname{conv} M$ there exist $x_0, \ldots, x_n \in M$, such that*

$$x = \sum_{k=0}^{n} \alpha_k x_k \,, \text{ where } \quad \alpha_k \geq 0, \sum_{k=0}^{n} \alpha_k = 1 \,.$$

Proof. Let $\emptyset \neq M \subset \mathbb{R}^n$ and $x \in \operatorname{conv} M$. Then by definition there exist $N \in \mathbb{N}, x_1, \ldots, x_N \in M$ and $\alpha_k \geq 0, \sum_{k=0}^{N} \alpha_k = 1$ such that

$$x = \sum_{k=0}^{N} \alpha_k x_k \,.$$

We may assume that $\alpha_k > 0$, $k = 0, \ldots, N$ as otherwise we could reduce the length of the convex combination. If $N \leq n$ there is nothing to show. So assume $N > n$. In this case there exists a solution $\beta = (\beta_0, \ldots, \beta_N) \neq 0$ to the linear equation

$$\sum_{k=0}^{N} \beta_k x_k = 0\,, \text{ and } \quad \sum_{k=0}^{N} \beta_k = 0\,. \tag{1.1.12}$$

The solution exists because we have at least $n+2$ variables β_k and only $n+1$ equations, given by the n rows of the x_k and the extra condition that the β_k sum to zero. Note that with β also $\gamma\beta$ is a solution of (1.1.12) for arbitrary $\gamma \in \mathbb{R}$ and of course some of the β_k have to be negative. So we may pick $\gamma > 0$ small enough so that

$$\alpha_k + \gamma\beta_k \geq 0\,, \quad \text{for all } k = 0, \ldots, N$$

with equality in at least one index k. After reordering we may assume that this index is N, so that $\alpha_N + \gamma\beta_N = 0$. Then we have $\alpha_k + \gamma\beta_k \geq 0$, $k = 0 \ldots, N-1$ and

$$1 = \sum_{k=0}^{N} \alpha_k = \sum_{k=0}^{N} \alpha_k + \gamma\beta_k = \sum_{k=0}^{N-1} \alpha_k + \gamma\beta_k\,.$$

The last equation implies in particular, that $1 \geq \alpha_k + \gamma\beta_k$, $k = 0 \ldots, N-1$. Also we have

$$x = \sum_{k=0}^{N} \alpha_k x_k = \sum_{k=0}^{N} (\alpha_k + \gamma\beta_k)\, x_k = \sum_{k=0}^{N-1} (\alpha_k + \gamma\beta_k)\,.$$

Thus we can represent x by a convex combination of $N' = N-1$ elements of M. This procedure can be repeated as long as $N' > n$, so that we arrive at as convex combination using $n+1$ elements of M as desired. ∎

In general, it is not true even in $\mathbb{R}^n$ that the convex hull of a closed set is again closed. As an example let $X = [0, \infty) \times \{0\} \cup \{(x, 1/x) \mid x > 0\}$. It is easy to see that the convex hull of this set is

$$\operatorname{conv} X = \{(x, y) \mid x > 0, 0 \leq y \leq 1/x\} \cup \{(0, 0)\}\,.$$

And this set is not closed because the positive y-axis is "missing". If we add the assumption of boundedness then the statement it true for $\mathbb{R}^n$ (but still false, in infinite dimensions).

Lemma 1.1.9 *Let $M \subset \mathbb{R}^n$.*

(i) If M is compact then $\operatorname{conv} M$ *is compact.*

(ii) If M is bounded then $\operatorname{cl}\operatorname{conv} M = \operatorname{conv}\operatorname{cl} M$.

Proof.

(i) Let A denote the set

$$A := \{(\alpha_0, \ldots, \alpha_n) \mid \alpha_k \geq 0, \sum_{k=0}^{n} \alpha_k = 1\}\,.$$

By Carathéodory's theorem the set $\operatorname{conv} M$ is the image of $A \times M^{n+1}$ under the function

$$g : (\alpha_0, \ldots, \alpha_n, x_0, \ldots, x_n) \mapsto \sum_{k=0}^{n} \alpha_k x_k \,.$$

Now the set $A \times M^{n+1}$ is compact by assumption thus its image under a continuous map is again compact. This proves the assertion.

(ii) To begin with we have $\operatorname{cl}\operatorname{conv} M \subset \operatorname{cl}\operatorname{conv}\operatorname{cl} M = \operatorname{conv}\operatorname{cl} M$ by part (i). Now if $x \in \operatorname{conv}\operatorname{cl} M$, then (using the notation of part (i)) $x = g(\alpha_0, , \ldots, \alpha_n, x_0, \ldots, x_n)$, where $x_0, \ldots, x_n \in \operatorname{cl} M$. Now we may take sequences $x_{0k}, \ldots, x_{nk}$ in M converging to (respectively) $x_0, \ldots, x_n$. Then $x_k := g(\alpha_0, , \ldots, \alpha_n, x_{0k}, \ldots, x_{nk}) \in \operatorname{conv} M$ and $x_k \to x$ by continuity of g, so that $x \in \operatorname{cl}\operatorname{conv} M$.

■

1.1.2 Dimension and Relative Interior

Given a convex set $K \subset H$ we can consider the set of affine spaces containing K. Amongst these there is a minimal one, as the intersection of affine sets is again an affine set. The smallest affine set has the representation

$$\operatorname{aff} K = \left\{ \sum_{i=1}^{k} \alpha_i x_i \,;\; x_i \in K, \alpha_i \in \mathbb{R}, \sum_{i=1}^{k} \alpha_i = 1, k = 1, 2, \ldots \right\} .$$

Definition 1.1.10 (Dimension of a Convex Set) *Let $K \subset H$ be convex. The dimension of K is the* dimension *of the smallest affine space containing K.*

Given K and $\operatorname{aff} K$, the smallest affine space containing K, the *relative interior* of K is defined as the interior of K relative to $\operatorname{aff} K$.

Definition 1.1.11 (Relative Interior) *Let $K \subset \mathbb{R}^n$ be convex. The* relative interior $\operatorname{ri} K$ *of K is the interior of K with respect to the smallest affine subspace of $\mathbb{R}^n$ containing K. That is*

$$x \in \operatorname{ri} K :\Leftrightarrow \exists\ \varepsilon > 0 : B_\varepsilon(x) \cap \operatorname{aff} K \subset K \,.$$

1.2 Distance to Convex Sets

Theorem 1.2.1 *Let X be a Hilbert space and M a linear subspace of X and $x \in X$ be arbitrary. If there is a vector $m_0 \in M$ such that $\|x - m_0\| \leq \|x - m\|$ for all $m \in M$, then m_0 is unique. In this case the vector m_0 is characterized by*

$$\langle x - m_0, m \rangle = 0 \,, \qquad \forall m \in M$$

Proof. We first show that if a minimizing vector m_0 exists, then $x - m_0$ is orthogonal to M. Assume this is not the case, so that $\langle x - m_0, m \rangle = \delta \neq 0$ for some $m \in M$ with $\|m\| = 1$. Then we have

$$\begin{aligned} \|x - m_0 - \delta m\|^2 &= \|x - m_0\|^2 - 2\langle x - m_0, \delta m \rangle + |\delta|^2 && (1.2.1) \\ &= \|x - m_0\|^2 - |\delta|^2 < \|x - m_0\|^2 \,, && (1.2.2) \end{aligned}$$

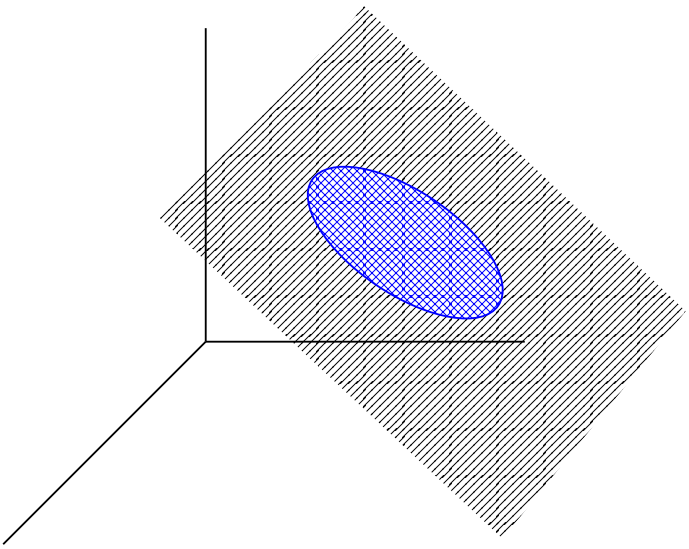

Figure 1.10: Relative interior of a 2-dimensional convex set in $\mathbb{R}^3$.

so that m_0 is not minimizing.

Let now $x - m_0$ be orthogonal to M. Then for $m \in M$ we have

$$\|x - m\|^2 = \|x - m_0 + m_0 - m\|^2 =$$

$$\|x - m_0\|^2 + 2\langle x - m_0, m_0 - m\rangle + \|m_0 - m\|^2 .$$

As the middle term on the right hand side is zero by assumption this implies $\|x - m\|^2 > \|x - m_0\|^2$ for $m \neq m_0$, so that m_0 is the unique minimizing vector. ∎

Note that the assumption in the previous Theorem 1.2.1 concerning the existence of a minimizing element is only relevant in infinite dimensions, where subspaces need not be closed. In $\mathbb{R}^n$ it is clear that a minimizing element exists.

1.3 Separation

Separation is one of the main tools in the analysis of convex sets. We will see later that many arguments for the existence or nonexistence of certain object rely on a separation argument. We will therefore now discuss the basic notions of this concept.

Theorem 1.3.1 *Let X be a Hilbert space, $x \in X$ and assume that $K \subset X$ is closed and convex. Then there is a unique $y_0 \in K$ such that*

$$\|x - y_0\| \leq \|x - y\| \qquad \textit{for all } y \in K .$$

Furthermore, y_0 is characterized by

$$\langle x - y_0, y - y_0\rangle \leq 0 \qquad \textit{for all } y \in K . \tag{1.3.1}$$

The proof may seem somewhat complicated, if we are just working in $\mathbb{R}^n$, where a compactness argument would be used. In a general Hilbert space however, closed bounded

sets are not necessarily compact. So that we need a different and more lengthy approach to prove the existence of a minimizer.

Proof. To show existence of a minimizing y_0 let $\{y_k\}_{k\in\mathbb{N}}$ be a sequence in K such that

$$\|x - y_k\| \to \delta := \inf_{y\in K} \|x - y\| .$$

By the parallelogram law we have for all $k, l \in \mathbb{N}$ that

$$\|y_k - y_l\|^2 = 2\|y_k - x\|^2 + 2\|y_l - x\|^2 - 4\left\|x - \frac{y_k - y_l}{2}\right\|^2 .$$

As K is convex $\frac{1}{2}(y_k - y_l) \in K$, so that the last term on the right hand side is bounded below by δ. Hence

$$\|y_k - y_l\|^2 \leq \|y_k - x\|^2 + 2\|y_l - x\|^2 - 4\delta^2 .$$

This expression tends to zero for k and l tending to ∞. Hence $\{y_k\}$ is a Cauchy sequence with limit y_0 that is contained on K as K is closed. By continuity $\|x - y_0\| = \delta$ as desired.

In the next step let y_0 be a point in K closest to x. We have to show that

$$\langle x - y_0, y - y_0\rangle \leq 0 \qquad \text{for all } y \in K .$$

Assume this is not the case, so that for some $z \in K$ we have $\langle x - y_0, z - y_0\rangle = \delta > 0$. As K is convex, we have that $(1 - \alpha)y_0 + \alpha z \in K$, $\alpha \in [0, 1]$. Then we obtain

$$\|x - (1 - \alpha)y_0 - \alpha z\|^2 = \|x - y_0 - \alpha(z - y_0)\|^2 =$$

$$\|x - y_0\|^2 - 2\alpha\langle x - y_0, z - y_0\rangle + \alpha^2\|z - y_0\|^2 ,$$

which is strictly smaller than $\|x - y_0\|^2$ for small $\alpha > 0$. This contradicts the minimality of y_0.

Conversely, if $\langle x - y_0, y - y_0\rangle \leq 0$ for some $y_0 \in K$ and all $y \in K$, then for $y_0 \neq y$

$$\|x - y\|^2 = \|x - y_0 + y_0 - y\|^2 = \|x - y_0\|^2 + 2\langle x - y_0, y_0 - y\rangle + \|y_0 - y\|^2 > \|x - y_0\|^2 ,$$

which shows that y_0 is the unique minimizing point in K. ∎

By the previous Theorem 1.3.1 we may for a closed convex set K define the projection onto K by

$$\pi_K(x) := y \in K , \text{ where } \|x - y\| = \operatorname{dist}(x, K) .$$

It is useful to note the following Lipschitz property of π_K.

Proposition 1.3.2 *Let X be a Hilbert space and $K \subset X$ be closed and convex. Then the projection onto K π_K is Lipschitz continuous with constant $L = 1$.*

Proof. Fix $x, y \in X$ and denote $z := \pi_K(x) - \pi_K(y)$. We may assume that $z \neq 0$ as otherwise the claim is trivial. Let $c_1 := \langle \pi_K(x), z\rangle$ and $c_2 := \langle \pi_K(y), z\rangle$. Note that $c_1 > c_2$ and consider the hyperplanes

$$H_1 := \{p \in X \mid \langle p, z\rangle = c_1\} ,$$

$$H_2 := \{p \in X \mid \langle p, z\rangle = c_2\} .$$

By definition we have $\operatorname{dist}(H_1, H_2) = \|z\|$. By the characterization (1.3.1) we have

$$\langle x - \pi_K(x), \pi_K(y) - \pi_K(x)\rangle \leq 0\,.$$

and hence

$$\langle x, \pi_K(y) - \pi_K(x)\rangle \leq \langle \pi_K(x), \pi_K(y) - \pi_K(x)\rangle = -c_1\,,$$

and similarly

$$\langle y, \pi_K(x) - \pi_K(y)\rangle \leq \langle \pi_K(y), \pi_K(x) - \pi_K(y)\rangle = c_2\,,$$

This shows that x lies to the right of H_1, that is $\langle x, z\rangle \geq c_1$ and y lies to the left of H_2, that is $\langle y, z\rangle \leq c_2$. This immediately implies that

$$\|x - y\| \geq \operatorname{dist}(H_1, H_2) = \|z\| = \|\pi_K(x) - \pi_K(y)\|\,,$$

as desired. ∎

From the previous results we now arrive at the first *separation principle* for convex sets.

Corollary 1.3.3 *Let X be a Hilbert space, $K \subset X$ be closed and convex and $x \notin K$. Then there exists $q \in X$, $c_1 > c_2 \in \mathbb{R}$ such that*

$$\langle q, x\rangle = c_1$$

and

$$\langle q, y\rangle \leq c_2 \qquad \textit{for all } y \in K\,.$$

Proof. Let y_0 be the point in K closest to x and define $q := x - y_0$. Then

$$\langle q, x\rangle \leq \langle q, y_0\rangle \qquad \text{for all } y \in K\,,$$

and

$$\langle q, x\rangle = \|q\|^2 + \langle q, y_0\rangle\,.$$

Thus the assertion follows with $c_1 := \|q\|^2 + \langle q, y_0\rangle > \langle q, y_0\rangle = c_2$. ∎

Corollary 1.3.4 *Let X be a Hilbert space, $K \subset X$ be closed and convex. Then K is equal to the intersection of the closed half spaces containing K.*

Proof. A closed half space in X is defined by a nonzero vector $n \in X$ and a number $c \in \mathbb{R}$ by

$$H_-(n, c) := \{x \in X \mid \langle x, q\rangle \leq c\}\,,$$

By the previous Corollary 1.3.3 for any $x \in X$, $x \notin K$, there exists a half space $H_-(n, c)$ with $x \notin H_-(n, c)$, $K \subset H_-(n, c)$. Thus

$$K = \bigcap H_-(n, c)\,,$$

where the intersection is over all closed half-spaces containing K. ∎

An interesting further interpretation of the convex hull of a set is that the convex hull of a set X is the smallest convex set containing X. This is the content of the following result.

Lemma 1.3.5 *Let X be a Hilbert space, and $M \subset X$ then*

$$\operatorname{conv} M = \bigcap C\,, \tag{1.3.2}$$

where the intersection is taken over all convex sets $C \supset M$.

Proof. We have seen that $\operatorname{conv} M$ is convex and clearly $M \subset \operatorname{conv} M$, as we may choose $m = 1$ in Definition 1.1.7. This implies that $\operatorname{conv} M$ is one of the sets appearing in the intersection on the right hand side of (1.3.2). Hence

$$\operatorname{conv} M \supset \bigcap_{C\text{convex}, M\subset C} C\,.$$

Conversely, if $x \notin \operatorname{conv} M$ then we may separate x from M by a hyperplane. That is, there exist n, c such that $\langle x, n\rangle = c$ and

$$\langle y, n\rangle < c\,, \qquad \text{for all } y \in \operatorname{conv} M\,.$$

This shows that $M \subset \operatorname{conv} M \subset H_-$, where H_- is the half space defined by n, c. But then for any $x \notin \operatorname{conv} M$, we can construct a convex set (even a half space) H, such that $M \subset H$ and $x \notin H$. This shows

$$\operatorname{conv} M \subset \bigcap_{C\text{convex}, M\subset C} C\,.$$

■

A nice application of the previous result is the following.

Lemma 1.3.6 *Let X be a Hilbert space, and $M \subset X$ then*

$$\operatorname{conv} B_\varepsilon(M) = B_\varepsilon(\operatorname{conv} M)\,.$$

Proof. By Lemma 1.1.5 the set $B_\varepsilon(\operatorname{conv} X) = \operatorname{conv} X + \varepsilon\, B$ is convex. Furthermore $B_\varepsilon(X) \subset B_\varepsilon(\operatorname{conv} X)$ and this implies that $\operatorname{conv} B_\varepsilon(X) \subset B_\varepsilon(\operatorname{conv} X)$ by Lemma 1.3.5.

To prove the converse inclusion let $x + y \in B_\varepsilon(\operatorname{conv} X)$ where $x \in \operatorname{conv} X$ and $y \in \varepsilon\, B$. Then by definition x is a convex combination of elements $x_k, k = 1, \dots m$ in X of the form

$$x = \sum_{k=1}^{m} \alpha_k x_k\,, \text{ where } \alpha_k \geq 0, \sum_{k=1}^{m} \alpha_k = 1\,,$$

Now $x_k + y \in B_\varepsilon(X)$ and hence

$$x + y = \sum_{k=1}^{m} \alpha_k (x_k + y) \in \operatorname{conv} B_\varepsilon(X)\,.$$

This shows the assertion. ■

It is a basic fact of the theory of compact sets, that descending intersections of compact sets are nonempty. We will now show that this operation commutes with taking the convex hull.

Lemma 1.3.7 *Let $M_\varepsilon, \varepsilon > 0$ be a descending family of compact subsets of $\mathbb{R}^n$ as $\varepsilon \to 0$, i.e. $M_\varepsilon \subset M_\delta$ if $\varepsilon < \delta$ then*

$$\operatorname{conv} \bigcap_\varepsilon M_\varepsilon = \bigcap_\varepsilon \operatorname{conv} M_\varepsilon \,.$$

Proof. By $\bigcap_\varepsilon M_\varepsilon \subset \bigcap_\varepsilon \operatorname{conv} M_\varepsilon$ and Lemma 1.3.5 it is clear that $\operatorname{conv} \bigcap_\varepsilon M_\varepsilon \subset \bigcap_\varepsilon \operatorname{conv} M_\varepsilon$.

To prove the converse direction let $x \in \bigcap_\varepsilon \operatorname{conv} M_\varepsilon$ then by Carathéodory's theorem there exist for every ε scalars $\alpha_{0\varepsilon}, \ldots, \alpha_{n\varepsilon} \geq 0, \sum_{k=0}^n \alpha_{k\varepsilon} = 1$ and points $x_{0\varepsilon}, \ldots, x_{n\varepsilon} \in M_\varepsilon$ such that $x = \sum_{k=0}^n \alpha_{k\varepsilon} x_{k\varepsilon}$. By compactness, we may choose a sequence $\varepsilon_l \to 0$ such that $\alpha_{k\varepsilon_l} \to \alpha_k$ and $x_{k\varepsilon_l} \to x_k$ as $l \to \infty$, for $k = 0, \ldots, n$.

It follows that $x_k \in \bigcap_\varepsilon M_\varepsilon$, $k = 0, \ldots, n^2$. and hence $x = \sum_{k=0}^n \alpha_k x_k \in \operatorname{conv} \bigcap_\varepsilon M_\varepsilon$. This shows the assertion. ∎

The main separation result for convex sets concerns the separation of two disjoint convex sets. This result is useful in many proofs and we will make frequent use of it.

Theorem 1.3.8 (Separation of Convex Sets) *Let X be a Hilbert space and $C_1, C_2 \subset X$ be convex such that*

$$C_1 \cap C_2 = \emptyset \,.$$

Then there exists a hyperplane H given by $n \neq 0$, $c \in \mathbb{R}$ such that

$$C_1 \subset \overline{H}_+ = \{x \in X \,;\, \langle x, n\rangle \geq c\}$$

and

$$C_2 \subset \overline{H}_- = \{x \in X \,;\, \langle x, n\rangle \leq c\} \,.$$

One might wonder, why it is not always possible to separate two convex sets by a hyperplane H given by $n \neq 0$, $c \in \mathbb{R}$ such that

$$\overline{H}_+ = \{x \in X \,;\, \langle x, n\rangle \geq c\}$$

contains one of the sets C_1, C_2 and

$$H_- = \{x \in X \,;\, \langle x, n\rangle < c\}$$

contains the other one. The reason for this is that there are strange examples such as the following one.

Example 1.3.9 Consider the two sets $C_1, C_2 \subset \mathbb{R}^2$ defined by

$$C_1 := \{(x_1, x_2) \,;\, x_2 > 0\} \cup \{(x_1, x_2) \,;\, x_1 \geq 0, x_2 = 0\}$$

and

$$C_1 := \{(x_1, x_2) \,;\, x_2 < 0\} \cup \{(x_1, x_2) \,;\, x_1 < 0, x_2 = 0\} \,.$$

Clearly, $C_1 \cap C_2 = \emptyset$. Also the two sets are convex; they are depicted in Figure 1.11. The full line on the x_1 axis belongs to C_1, while the dashed line on the x_1 axis belongs to x_2. The only possible choice for a hyperplane separating the relative interiors of C_1 and C_2 is in fact the x_1 axis. But this necessarily intersects both C_1 and C_2, so that we really need nonstrict inequalities for both sides of the hyperplane.

[2]This may be seen as follows. Fix $r > 0$. $x_{1\varepsilon_l} \in \overline{B}_r(x1)$ for all l large enough. Hence $\overline{B}_r(x1) \cap M_\varepsilon$ defines a descending chain of nonempty compact sets so that $\bigcap \varepsilon \overline{B}_r(x1) \cap M_\varepsilon = \overline{B}_r(x1) \cap \bigcap_\varepsilon M_\varepsilon \neq \emptyset$. As $r > 0$ was arbitrary it follows that $x_1 \in \bigcap_\varepsilon M$ by compactness.

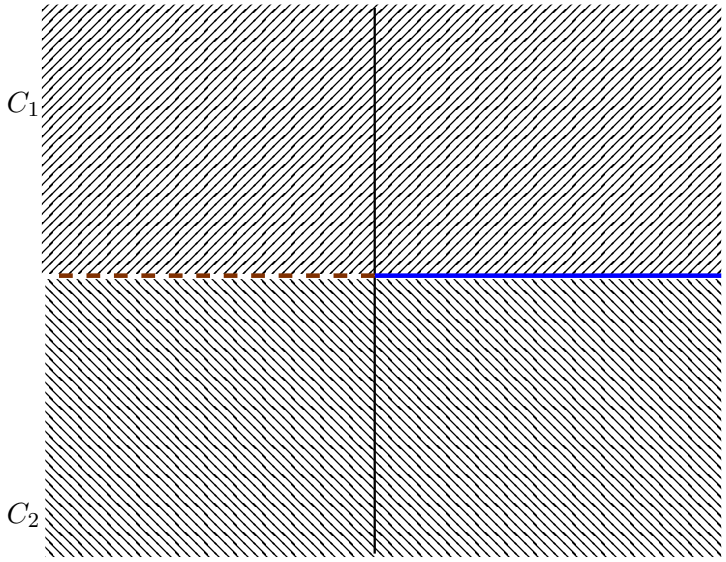

Figure 1.11: Convex sets described in Example 1.3.9.

Remark 1.3.10 It can be proved in very general situations that given two convex sets C_1, C_2 in a vector space X it is always possible to find two convex sets A_1, A_2 such that

(i) $A_1 \cup A_2 = X$, i.e. the two convex sets A_1, A_2 form a partition of the state space,

(ii) $C_1 \subset A_1$ and $C_2 \subset A_2$.

The way of proving this in a very abstract setting is via Zorn's lemma. So it is a non-constructive way of arriving at the result. Also we have to presuppose that the axiom of choice holds. No such assumption is necessary for the constructions we discuss in these notes.

We will not prove Theorem 1.3.8 in full generality but restrict ourselves to the case that the two convex sets C_1, C_2 are a positive distance apart. We say that two convex sets can be *strongly separated*, if there exists a hyperplane $H(q, c)$ and an $\varepsilon > 0$ such that

$$C_1 \subset \overline{H}_{+,\varepsilon} = \{x \in X\,;\ \langle x, q\rangle \geq c + \varepsilon\}$$

and

$$C_2 \subset \overline{H}_{-,\varepsilon} = \{x \in X\,;\ \langle x, q\rangle \leq c - \varepsilon\}\,.$$

We then have the following result, see also Figure 1.12.

Theorem 1.3.11 (Strong Separation of Convex Sets) *Two convex sets $C_1, C_2 \subset \mathbb{R}^n$ can be strongly separated if and only if*

$$\operatorname{dist}(C_1, C_2) := \inf\{\|x - y\|\,;\ x \in C_1, y \in C_2\} > 0\,.$$

Proof. It is clear that if the convex sets C_1, C_2 can be strongly separated, then $\operatorname{dist}(C_1, C_2) \geq \operatorname{dist}(C_1, H) + \operatorname{dist}(C_2, H) \geq 2\varepsilon/\|q\|^2$.

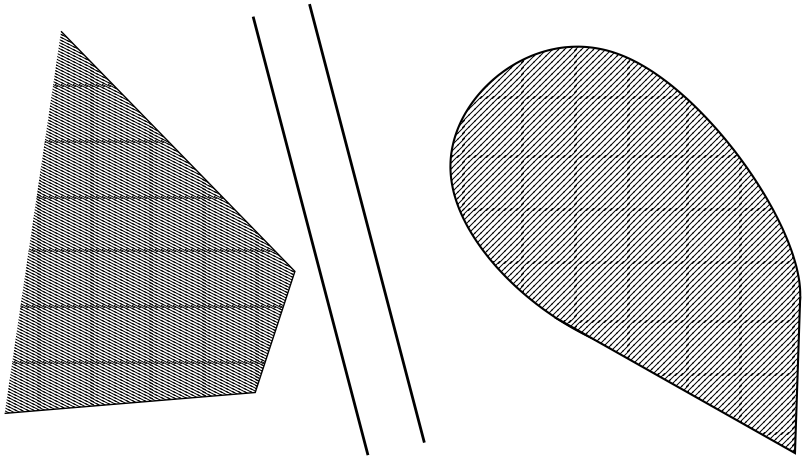

Figure 1.12: Strong separation means separation by two hyperplanes.

To prove the converse statement let $x_1 \in \operatorname{cl} C_1, x_2 \in \operatorname{cl} C_2$ be points minimizing the distance $\operatorname{dist}(C_1, C_2)$ and define $q := x_1 - x_2 \neq 0$. Defining

$$H_- = \{x \in X\,;\ \langle x, q\rangle \leq \langle x_2, q\rangle\} \quad H_+ = \{x \in X\,;\ \langle x, q\rangle \geq \langle x_1, q\rangle\}$$

we claim that $C_1 \subset H_+$, and $C_2 \subset H_-$. This proves the theorem because $\langle x_1, q\rangle > \langle x_2, q\rangle$ by

$$\langle x_1, q\rangle - \langle x_2, q\rangle = \langle q, q\rangle > 0\,.$$

The remaining claim follows from an application of Theorem 1.3.1. Indeed, by definition x_1 is the point in C_1 minimizing the distance to x_2. Thus from (1.3.1) we have

$$\langle x_2 - x_1, y - x_1\rangle \leq 0\,, \qquad \text{for all } y \in C_1\,.$$

And hence

$$\langle x_1, q\rangle \leq \langle y, q\rangle\,, \qquad \text{for all } y \in C_1\,.$$

By a similar argument we obtain

$$\langle y, q\rangle \leq \langle x_2, q\rangle\,, \qquad \text{for all } y \in C_2\,.$$

This shows the assertion. ∎

A further concept closely related to separating hyperplanes is that of *supporting hyperplanes*.

Definition 1.3.12 (Supporting hyperplane) *Let C be a convex set in a Hilbert space X and x be a boundary point of C. A hyperplane H is said to be supporting in x (with respect to the convex set C), if H is represented by $n \in X, n \neq 0$, $c \in \mathbb{R}$ and*

(i) $x \in H$, i.e. $\langle x, q\rangle = c$,

(ii) $C \subset \overline{H}_+$, or $C \subset \overline{H}_-$.

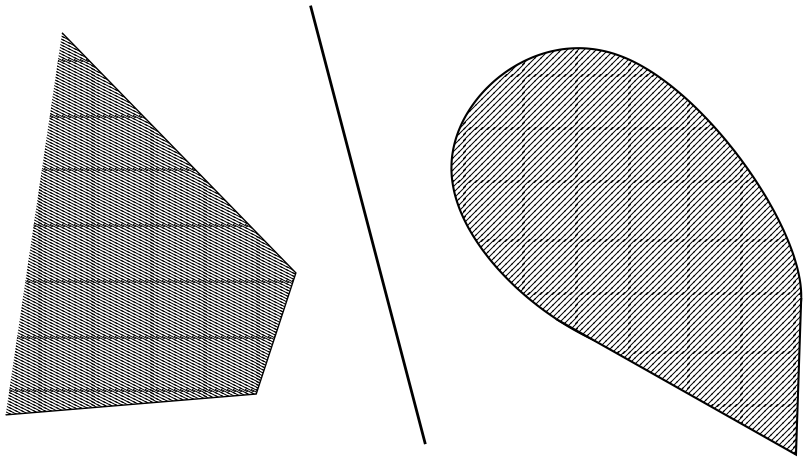

Figure 1.13: Separating Hyperplane

Note that for a convex set $C \neq X$ a supporting hyperplane exists in every boundary point x of C. This is a consequence of the Separation Theorem 1.3.8. If $\operatorname{int} C \neq \emptyset$, then we can consider the convex sets $\operatorname{int} C$ and x. These can be separated by a hyperplane H. Then necessarily $x \in H$ and $\operatorname{int} C \cap H = \emptyset$. If C is of lower dimension, we could simply use a hyperplane containing C and this would be supporting by definition. This, however, can be considered cheating. A less trivial way of doing it, would be to construct a supporting hyperplane at x in the affine hull $\operatorname{aff} C$, where we can again take recourse to Theorem 1.3.8. This hyperplane can then be extended to the hyperplane of the whole space by adding the orthogonal complement of $\operatorname{aff} C$ to it.

The supporting hyperplane in a point need not be unique, as is shown in Figure 1.14.

1.4 Faces, Extreme Points and Recession Cones

In the description of convex sets it is useful to consider those points which cannot be described as convex combinations of other points in C.

Definition 1.4.1 (Faces) *Let $C \subset H$ be convex. A convex set $C' \subset C$ is a* face *of C, if for every segment $[x, y] \subset C$ such that*

$$\operatorname{ri}[x, y] \cap C' \neq \emptyset$$

we have

$$[x, y] \subset C' .$$

A face C' is called exposed *if there exists a hyperplane $H(q, c)$ such that $C' \subset H$ and $(C \setminus C') \subset H_+(q, c)$*[3]*.*

Note that $\emptyset$ and C itself are (pretty boring) faces of C. A particular case of faces are the zero-dimensional faces.

[3]Note: we are considering the open half-space $H_+(q, c)$.

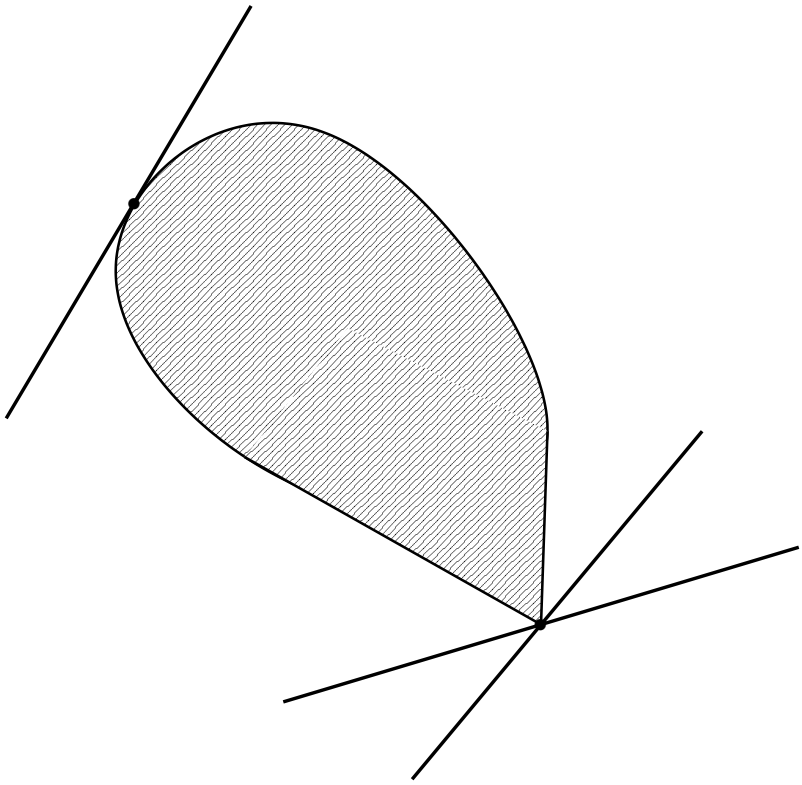

Figure 1.14: Supporting Hyperplane

Definition 1.4.2 (Extreme Points) *Let H be a Hilbert space and $C \subset H$ be convex. A point $x \in \operatorname{cl} C$ is an* extreme point *of C, if from the condition*

$$x = \lambda y + (1-\lambda)z\,, \quad y, z \in \operatorname{cl} C, \lambda \in (0,1) \tag{1.4.1}$$

it follows that $x = y = z$.

Briefly put, a point $x \in C$ is an extreme point if it not possible to represent x as a convex combination of two other points contained in x. Exposed points are of course extreme points, but the converse is not true.Recall that x is an exposed point, if there is a supporting hyperplane through x only containing x. Figure 1.15 shows some (four actually) extreme points that are not exposed. The figure represents the convex hull of two circles of equal diameter. For instance at the lower end of the circle on the right, any hyperplane supporting this point has to contain the straight line.

Note that the set of extreme points of a closed set need not be closed. Indeed, compactness of a convex set C does not imply that the set of extreme points of this set is compact. An example to this effect can be constructed as follows.

Example 1.4.3 In this example we construct a compact set, that could be a candidate of a set of extreme points, but when taking the convex hull we see that this is not the case. Consider the following set in $\mathbb{R}^3$.

$$Z := \{(0,0,1),(0,0,-1)\} \cup \{(x,y,0)\,;\ (x-1)^2 + y^2 = 1\}\,.$$

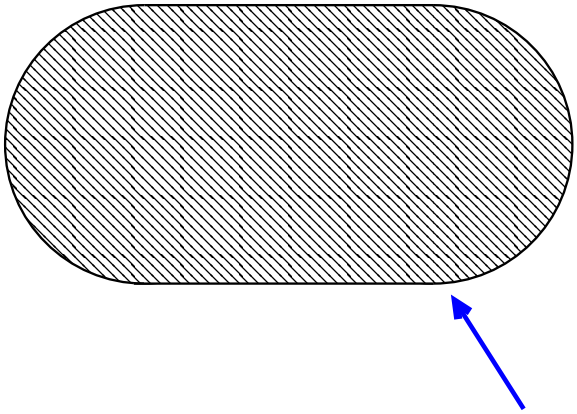

Figure 1.15: Extreme points that are not exposed.

Thus Z consists of two points on the z-axis and a circle in the plane $z = 0$ centered at $(1, 0)$ with radius equal to 1. If we now consider the convex hull $\operatorname{conv} Z$, then clearly $\operatorname{ext} \operatorname{conv} Z \subset Z$.

It is clear that the points $(0, 0, 1)$ and $(0, 0, -1)$ are extremal points of $\operatorname{conv} Z$. So are the points on the circle, except for the point $(0, 0, 0)$, which lies on the circle $(x-1)^2+y^2 = 1, z = 0$, but which can also be represented in the form

$$(0, 0, 0) = \frac{1}{2}(0, 0, 1) + \frac{1}{2}(0, 0, -1)\,.$$

Hence the origin is not an extreme point of $\operatorname{conv} Z$ and we see that indeed

$$\operatorname{ext} \operatorname{conv} Z = Z \setminus \{(0, 0, 0)\}\,.$$

This shows that the set of extreme points of Z is not closed and hence not compact.

Some convex sets do not have extreme points. Hyperplanes and half-spaces are examples in point. For bounded convex sets, however, extreme points represent a convex set, because we have the following theorem.

Theorem 1.4.4 *Any closed bounded convex set in $\mathbb{R}^n$ is the convex hull of its extreme points.*

Proof. This is actually a corollary of the more general statement of Theorem 1.4.9 below. ∎

To obtain complete descriptions of convex sets also in the unbounded case we need the notion of the *recession cone* and extreme direction.

Definition 1.4.5 (Recession Cone) *Let $C \subset X$ be a nonempty convex set. We say C recedes in the direction $y \in X$ if*

$$x + \lambda y \in C\,, \qquad \text{for all } x \in C, \lambda \geq 0\,.$$

The recession cone 0^+C of C is the set

$$0^+C := \{y \in X\,;\ C \text{ recedes in direction } y\}\,.$$

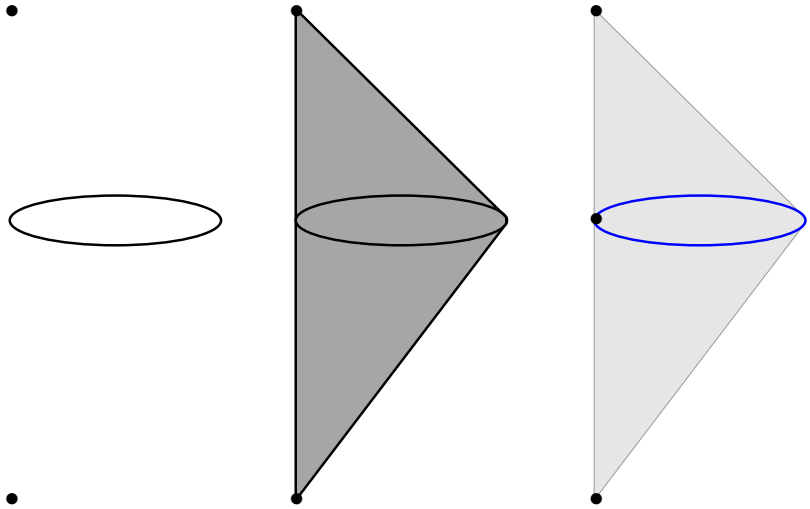

Figure 1.16: The set of extreme points of a compact set need not be compact. To the left we see a compact set in $\mathbb{R}^3$, in the middle its convex hull: two finite slanted ice cream cones put on top of each other. The middle point on the right is on the line from the top to the bottom point, so not an extreme point.

Lemma 1.4.6 *Let $C \subset X$ be a nonempty convex set. The recession cone 0^+C is a convex cone. It satisfies*

$$0^+C = \{y \in X\,;\; C + y \subset C\}\,.$$

Proof. We begin by proving the equality. It is clear that if C recedes in the direction y, then $C + y \subset C$. Conversely, if y is an element in the set on the right hand side we obtain by iteration that $C + ky \subset C + (k-1)y \subset \ldots \subset C + y \subset C$ for all $k \in \mathbb{N}$. BY convexity it follows that in fact $C + \lambda y \subset C$ for $\lambda \geq 0$, as we can write any real number $\lambda \geq 0$ as $\alpha k + (1-\alpha)k'$ for suitable nonnegative integers k and $\alpha \in [0,1]$.

To show convexity let y_1, y_2 two element of the set on the right hand side and $x \in C$. Then for $y = \lambda y_1 + (1-\lambda)y_2, \lambda \in [0,1]$ we obtain

$$x + y = \lambda(x + y_1) + (1-\lambda)(x + y_2)$$

which is an element of C, as $y_1, y_2 \in 0^+C$ and C is convex. This concludes the proof. ∎

We note the following properties of the recession cone.

Proposition 1.4.7 *(i) If C is a closed convex set containing the origin, then*

$$0^+C = \bigcap_{\varepsilon > 0} \varepsilon C = \{y\,;\; \varepsilon^{-1} y \in C\,, \quad \text{for all } \varepsilon > 0\}\,.$$

(ii) If $C_i, i \in I$ is a family of convex sets, such that its intersection is nonempty, then

$$0^+\left(\bigcap_{i \in I} C_i\right) = \bigcap_{i \in I} 0^+ C_i\,.$$

(iii) A nonempty closed and convex set is bounded if and only if $0^+C = \{0\}$.

If we consider convex cones, the notion of an extreme point is not very useful, as only 0 qualifies as an extreme point by the cone property. For cones we therefore consider *extreme rays*, which are define as one-dimensional faces of the cone containing the origin in their closure and which are not linear subspaces. In other words extreme rays are half-lines in the cone that are faces. Again exposed rays are extreme rays but the converse need not be true.

More generally we introduce the concept of an extreme direction of a convex set.

Definition 1.4.8 (Extreme Direction) *Let $C \subset H$ be a nonempty convex set. We say that y is an* extreme direction *of C, if there exists an $x \in C$ such that*

$$x + \{\lambda y\,;\ \lambda \geq 0\}$$

is a face of C.

It can be shown that any extreme direction of C is also an extreme direction of the recession cone 0^+C. The same statement is true for exposed directions. The converse is not true. For instance the set

$$C = \{(x, y) \in \mathbb{R}^2\,;\ x^2 \geq y\}$$

does not have extreme directions as the boundary of the set is "curved", and so the only faces (except the trivial ones, $\emptyset$, C) are points. The recession cone of C, however, is given by

$$O^+C = \operatorname{cone}\{(0, 1)\}\,.$$

If we have a set of points $M_1 \subset H$ and a set of extreme directions M_2 , then we define the *convex hull* of $M = M_1 \cup M_2$ as the smallest convex set containing M_1 and receding into every direction contained in M_2. Similar to the case of the convex hull of a set of points it can be shown that

$$\operatorname{conv} M = \left\{ \sum_{i=1}^{k} \lambda_i x_i\,;\ k \in \mathbb{N}, \sum_{i=1}^{k} \lambda_i = 1, \lambda_i \geq 0, x_i \in M_1 \text{ or } \alpha x_i \in M_2 \text{ for some } \alpha > 0 \right\}.$$

In other words, if we define the set of rays defined by M_2 as

$$\hat{M}_2 := \{\alpha x_2\,;\ x_2 \in M_2, \alpha \geq 0\}$$

then

$$\operatorname{conv} M = \operatorname{conv}\left(M_1 + \hat{M}_2\right). \tag{1.4.2}$$

Also for the convex hull of sets of points and sets of directions Carathéodory's theorem applies: In $\mathbb{R}^n$ it is sufficient to consider the convex combination of at most $n+1$ elements in M_1 and the set of rays $\hat{M}_2$. This is immediate from (1.4.2), but note that we have not proved this equality here.

With the previous remarks we finally arrive at a more general representation theorem for convex sets.

Theorem 1.4.9 *Let $C \subset \mathbb{R}^n$ be a closed convex set containing no lines. Let S be the set of extreme points and extreme directions of C. Then*

$$C = \operatorname{conv} S\,.$$

1.5 Duality for Convex Sets

As a first glimpse into duality theory, which will prove to be of vital importance in the theory of convexity, let us define the *dual cone* corresponding to a cone.

Definition 1.5.1 (Dual Cone) *Let $K \subset \mathbb{R}^n$ be a cone. The* dual cone K^* *is defined by*

$$K^* := \{y \in \mathbb{R}^n\,;\ \langle x, y\rangle \geq 0 \text{ for all } x \in K\}\,. \tag{1.5.1}$$

Thus $y \in K^*$ if y is orthogonal to a hyperplane that supports the cone K at the origin.

1.6 Convex Functions

We now consider functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. The *domain* of such a function is defined as the set, where the function has finite values, i.e.

$$\operatorname{dom} f := \{x \in \mathbb{R}^n\,;\ f(x) \in \mathbb{R}\}\,.$$

With this notation we may as well consider the finite function

$$f : \operatorname{dom} f \subset \mathbb{R}^n \to \mathbb{R}\,, \tag{1.6.1}$$

but it will in some cases be convenient to admit the value ∞. The important definition is now the concept of convexity for functions.

Definition 1.6.1 (Convex Function) *A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called* convex, *if*

(i) $\operatorname{dom} f$ *is convex,*

(ii) for all $x, y \in \operatorname{dom} f$, $\lambda \in [0,1]$ we have

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y)\,. \tag{1.6.2}$$

The function f is called strictly convex *if in addition the inequality in* (1.6.2) *is strict, whenever $x \neq y$, $\lambda \in (0,1)$.*

The function f is *concave* if $-f$ is convex and *strictly concave* if $-f$ is strictly convex. An easy characterization of convexity is in terms of the *epigraph* of f defined by

$$\operatorname{epi} f := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}\,;\ x \in \operatorname{dom} f\,, \quad y \geq f(x)\}\,. \tag{1.6.3}$$

Lemma 1.6.2 *A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex if and only if* $\operatorname{epi} f$ *is convex.*

Proof. Obvious. ∎

In case that f is differentiable, there are easy criteria for convexity.

Proposition 1.6.3 *Consider $f : \operatorname{dom} f \subset \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and assume that* $\operatorname{dom} f = \operatorname{ri} \operatorname{dom} f$ *is convex.*

(i) If f is continuously differentiable on $\operatorname{dom} f$*, then f is convex if and only if*

$$f(x) + \langle y - x, \nabla f(x)\rangle \leq f(y)\,, \qquad \text{for all } x, y \in \operatorname{dom} f\,. \tag{1.6.4}$$

If the inequality in (1.6.4) *is strict whenever $x \neq y$, then f is strictly convex.*

(ii) Assume f is twice continuously differentiable. Then f is convex if and only if the Hessian *of f satisfies*

$$Hf(x) \geq 0\,, \qquad \textit{for all } x \in \operatorname{dom} f\,. \tag{1.6.5}$$

If the Hessian is positive definite everywhere on $\operatorname{dom} f$*, then f is strictly convex.*

Proof. (i) Assume that (1.6.4) is satisfied. Fix $x, y \in \operatorname{dom} f$ and $\lambda \in [0,1]$. Let $z = \lambda x + (1-\lambda)y$. Then using (1.6.4) for the differential in z we obtain

$$\begin{aligned}\lambda f(x) + (1-\lambda) f(y) \geq \\ \lambda f(z) + \langle \lambda(x-z), \nabla f(z)\rangle + (1-\lambda) f(z) + \langle (1-\lambda)(y-z), \nabla f(z)\rangle \\ = f(z) + \langle \lambda(1-\lambda)(x-y) + (1-\lambda)\lambda(y-x), \nabla f(z)\rangle = f(z)\,. \end{aligned} \tag{1.6.6}$$

This shows that f is convex. By the same argument we obtain strict convexity if the inequality in (1.6.4) is strict.

Conversely, assume (1.6.4) is not satisfied. By the assumption of differentiability we have that

$$\langle y - x, \nabla f(x)\rangle \;=\; \lim_{h \searrow 0} \frac{f(x + h(y-x)) - f(x)}{h}\,.$$

Then by our assumption we have for h sufficiently small that

$$f(x) + \frac{f(x + h(y-x)) - f(x)}{h} > f(y)$$

or equivalently

$$f(x + h(y-x)) > f(x) + h(f(y) - f(x))\,.$$

As we may assume that $h \in (0,1)$, this shows that f is not convex.

(ii) The condition that the Hessian is positive definite is equivalent to $(x-y)^\top H(x)(x-y) \geq 0$ for all $x, y \in \operatorname{dom} f$. If we consider the real map $g : \alpha \mapsto f(x + \alpha(y-x))$, then

$$g''(\alpha) = (x-y)^\top H(x + \alpha(y-x))(x-y)\,.$$

It is thus sufficient to prove the statement for one-dimensional maps. So assume $x \in \mathbb{R}$ and let $f : \operatorname{dom} f \to \mathbb{R}$ be such that $f''(x) \geq 0$ for all $x \in \operatorname{dom} f$. This implies that f' is strictly increasing in x. Consequently, for all $x \leq y \in \operatorname{dom} f$ we have

$$f(y) = f(x) + \int_x^y f'(z)dz \geq f(x) + f'(x)(y-x)\,.$$

This is just condition (1.6.4) in the one-dimensional case. The case $y \leq x$ follows exactly the same way. ∎

Note that $\langle y - x, \nabla f(x)\rangle$ is simply the directional derivative of f in the point x in direction $y - x$. So one interpretation of (1.6.4) is that the first order Taylor expansion from x in the direction of y always underestimates $f(y)$. A further interpretation is in terms of the epigraph. If we rearrange (1.6.4) we obtain the equivalent condition that for all $x, y \in \operatorname{dom} f$ it should hold that

$$\langle y, \nabla f(x)\rangle - f(y) \leq \langle x, \nabla f(x)\rangle - f(x)\,.$$

Thinking in terms of the epigraph of f we can write this as

$$\left\langle \begin{pmatrix} y \\ f(y) \end{pmatrix}, \begin{pmatrix} \nabla f(x) \\ -1 \end{pmatrix} \right\rangle \leq \left\langle \begin{pmatrix} x \\ f(x) \end{pmatrix}, \begin{pmatrix} \nabla f(x) \\ -1 \end{pmatrix} \right\rangle \tag{1.6.7}$$

And we see that in fact

$$n := \begin{pmatrix} \nabla f(x) \\ -1 \end{pmatrix}, \quad c := \left\langle \begin{pmatrix} x \\ f(x) \end{pmatrix}, \begin{pmatrix} \nabla f(x) \\ -1 \end{pmatrix} \right\rangle$$

defines a supporting hyperplane of the epigraph in the point $(x, f(x))$ as shown in Figure 1.17.

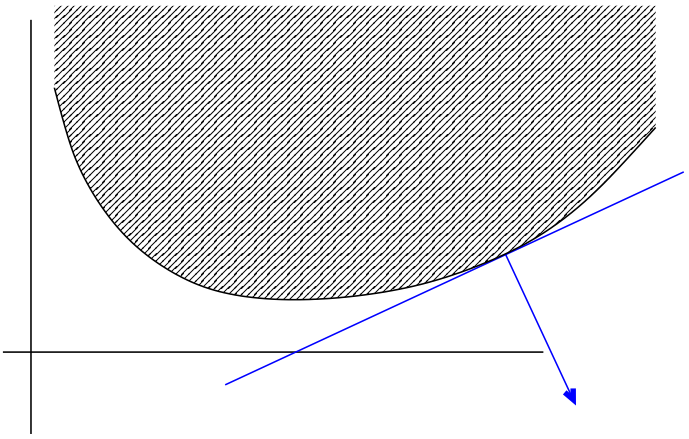

Figure 1.17: Supporting hyperplane of the epigraph.

The conditions for operations on convex sets that preserve convexity immediately leads to operations on functions that preserve convexity.

Lemma 1.6.4 *Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. Then*

(i) $f_1 + f_2$ is convex.

(ii) $\max\{f_1, f_2\} : x \mapsto \max\{f_1(x), f_2(x)\}$ is convex.

Proof. (i) can be checked by direct calculation.
The second claim is a consequence of $\operatorname{epi} \max\{f_1, f_2\} = \operatorname{epi} f_1 \cap \operatorname{epi} f_2$. ∎
A consequence of the fact that a closed convex set is the intersection of all closed half spaces containing it is the following maximization property of convex functions.

Proposition 1.6.5 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex, then for $x \in \operatorname{ri} \operatorname{dom} f$ we have*

$$f(x) = \max\{\langle x, c\rangle \,;\, c \in \mathbb{R}^n \text{ such that } \langle \cdot, c\rangle \leq f(\cdot)\}\,. \tag{1.6.8}$$

We now turn to the regularity properties of convex functions. It is clear that convex functions need not be continuously differentiable, as for instance the maximum of two differentiable functions is frequently not differentiable everywhere.

As it turns out, convexity is strong enough to prove continuity and even Lipschitz continuity. As the proof[4] of Lipschitz continuity uses continuity, we will prove continuity first.

Let $U \subset \mathbb{R}^n$ be a nonempty set. Recall that a function $f : U \to \mathbb{R}$ is called *locally Lipschitz continuous*, if for every $x \in U$ there is an $\varepsilon > 0$ and a constant L such that for all $y \in U \cap B_\varepsilon(x)$ it holds that

$$|f(x) - f(y)| \leq L\|x - y\| .$$

If the constant L can be chosen so that the previous inequality holds for all $x, y \in U$, then f is called globally Lipschitz continuous on U.

For our proof of continuity we need the following increase property of convex functions.

Lemma 1.6.6 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and $x \in \operatorname{dom} f$. Then for any $y \in \mathbb{R}^n, 0 < m \leq l, 0 < h \leq k$ we have*

$$\frac{f(x) - f(x - ly)}{l} \;\leq\; \frac{f(x) - f(x - my)}{m} \;\leq\; \frac{f(x + hy) - f(x)}{h} \;\leq\; \frac{f(x + ky) - f(x)}{k} .$$

Proof. If some of the points $x - ly, x - my, x + hy, x + ky \notin \operatorname{dom} f$, then the corresponding value of f is infinity and the inequalities are trivially true. Otherwise we may rearrange the inequalities equivalently, to obtain the conditions for convexity of f as

$$f(x - my) \leq \frac{l - m}{l} f(x) + \frac{m}{l} f(x - ly) , \quad f(x) \leq \frac{h}{h + m} f(x - my) + \frac{m}{h + m} f(x + hy) ,$$
$$f(x + hy) \leq \frac{k - h}{h} f(x) + \frac{h}{k} f(x + ky) .$$

This shows the assertion. ∎

Theorem 1.6.7 (Continuity of Convex Functions) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then at every point $x \in \operatorname{ri} \operatorname{dom} f$, the function f is continuous with respect to* $\operatorname{aff} \operatorname{dom} f$.

Note that the restriction to the affine hull of the domain is necessary, because if the domain s of lower dimension then a discontinuity appears immediately by moving slightly outside of the domain, at which point the value of the function jumps to ∞. Thus the theorem states that in the directions that keep the state in the domain, the function f changes continuously.

Proof. Let $x \in \operatorname{ri} \operatorname{dom} f$. Pick a closed ball B_∞ with respect to the infinity norm (a closed n-cube) centered at x, such that

$$R := B_\infty \cap \operatorname{aff} \operatorname{dom} f \subset \operatorname{ri} \operatorname{dom} f .$$

By construction P is a polyhedron, and so has finitely many extreme points $z_1, \ldots, z_k$. Any $z \in P$ may be expressed by a convex combination of the extreme points $z = \lambda_1 z_1 + \ldots + \lambda_k z_k$, with $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$. By convexity we have for all $z \in P$

$$f(z) = f(\lambda_1 z_1 + \ldots + \lambda_k z_k) \leq \sum_{i=1}^{k} \lambda_i f(z_i) \leq \max_i f(z_i) .$$

[4] That is, the proof we are presenting here

Thus f is bounded over the polytope P by a constant $c > 0$. Now let $U := P - x$. Then $U = -U$ is a closed neighborhood of 0 the linear subspace parallel to aff dom f. To show continuity we need to show that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $y \in U, h < \delta$ implies $|f(x + hy) - f(x)| < \varepsilon$. Using the rightmost inequality of Lemma 1.6.6 with $k = 1$ we see that

$$f(x + hy) - f(x) \leq h(c - f(x)), \quad \text{for all } h \in [0, 1].$$

Similarly, as $x - y \in P$, we may use the second inequality of Lemma 1.6.6 with y replaced by $-y$, m replaced by h, and h replaced by 1 to obtain

$$f(x) - f(x + hy) \leq h(c - f(x)).$$

Combining these inequalities we arrive at $|f(x) - f(x + hy)| \leq h(c - f(x))$, which implies continuity. Note that $c - f(x) \geq 0$ as the is the maximum of f over $x + U$. ∎

With the help of the previous theorem we can prove an even stronger regularity condition on the relative interior of the domain of a convex function.

Theorem 1.6.8 *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then f is locally Lipschitz continuous on* ri dom f.

Proof. Let $M \subset \operatorname{ri} \operatorname{dom} f$ be compact. Let Z be the set

$$Z := \{z \in \mathbb{R}^n ;\ \|z\| = 1, \text{for all } x \in \operatorname{dom} f\colon\ x + z \in \operatorname{aff} \operatorname{dom} f\}.$$

In other words, Z is the set of vectors of length 1 in the linear subspace of directions in aff dom f. As M is compact and ri dom f is open in aff dom f if follows that we may choose $h > 0$ small enough so that

$$M + hZ \subset \operatorname{ri} \operatorname{dom} f.$$

Let $x, y \in K$. Setting $z := (y - x)/\|y - x\| \in Z$, we obtain from Lemma 1.6.6 that

$$\frac{f(x) - f(x - hz)}{h} \leq \frac{f(y) - f(x)}{\|y - x\|} \leq \frac{f(y + hz) - f(y)}{h}.$$

The claim is thus proved if we can find an upper bound for

$$\frac{|f(x) - f(x - hz)|}{h}, \quad x \in M, z \in Z. \tag{1.6.9}$$

Note that this would supply bounds for both sides of the inequality as $z \in Z$ implies $-z \in Z$. As $h > 0$ is fixed, the function in (1.6.9) is continuous in (x, z) by Theorem 1.6.7. As a continuous function it attains it maximum over the compact set $M \times Z$. This proves the assertion. ∎

The previous result has far-reaching consequences. Lipschitz continuous functions are differentiable almost everywhere by Rademacher's theorem. For a convex function defined on a lower dimensional set, this of course only makes sense in the directions aligned with aff dom f, but still restricted to that linear space convex functions are differentiable almost everywhere on their domain of definition.

We note the following further facts of interest. The proof is beyond the scope of these notes.

Theorem 1.6.9 (Continuous Differentiability) *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. If f is differentiable on an open set $U \subset \operatorname{dom} f$, then it is continuously differentiable on U.*

Theorem 1.6.10 (Alexandrov's Theorem) *If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex, then it is almost everywhere twice differentiable on* $\operatorname{ri} \operatorname{dom} f$.

1.7 Subgradients

Convex functions need not be differentiable everywhere, but as they locally Lipschitz by Theorem 1.6.8, they are differentiable almost everywhere by Rademacher's theorem. It is then of interest to have generalized differentiability properties everywhere. The key idea here is that as the epigraph of convex functions is convex, it has supporting hyperplanes everywhere. The data describing these hyperplanes are the subgradients. The idea of the definition can be seen as a generalization of convexity for differentiable functions we have seen in Proposition 1.6.3. Instead of using (1.6.4) as a property that can be proved for the gradient we use it as the defining property of the subgradient.

Definition 1.7.1 (Subgradients) *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and $x \in \operatorname{dom} f$. A vector $p \in \mathbb{R}^n$ is called a subgradient vector of f in x, if* $\operatorname{aff} \operatorname{dom} f$ *recedes in direction p and*

$$f(x) + \langle y - x, p \rangle \leq f(y)\,, \qquad \text{for all } y \in \operatorname{dom} f\,. \tag{1.7.1}$$

The subgradient of f at x is now a set defined by

$$\partial f(x) := \{p \in \mathbb{R}^n\,;\ p \text{ satisfies } (1.7.1)\}\,. \tag{1.7.2}$$

Note that the definition of the supporting hyperplanes of the epigraph just used (1.6.4). Thus the subgradients p are precisely the vectors for which $(p^\top, -1)^\top$ defines a supporting hyperplane of the epigraph. The condition that $\operatorname{aff} \operatorname{dom} f$ recedes in direction p is only necessary, if $\operatorname{dom} f$ is a lower dimensional set. It is automatic, if $\operatorname{dom} f$ has interior points.

At boundary points of $\operatorname{dom} f$ (where the boundary has to be understood with respect to affine space generated by $\operatorname{dom} f$) it might occur that the subgradient is empty. For every point in $\operatorname{ri} \operatorname{dom} f$ the subgradient of f is a nonempty set.

As with the ordinary gradient is useful to consider rules for the calculation of the subgradient.

Lemma 1.7.2 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. Then for all $x \in \operatorname{ri} \operatorname{dom} f$,*

(i) the subgradient $\partial f(x)$ is bounded, closed, convex and nonempty.

(ii) if f is differentiable at x, then the subgradient contains the gradient of f at x as its only element, i.e.,

$$\partial f(x) = \{\nabla f(x)\}$$

Note that as f is only defined on $\operatorname{dom} f$, differentiability of f only makes sense for the directions in which f is defined. The gradient thus has to be understood with respect to the variables y such that $x + hy \in \operatorname{dom} f$ for y small enough.

Proof. In the proof we will assume that $\operatorname{dom} f$ is n-dimensional, or otherwise use new variables representing $\operatorname{dom} f$ in lower dimensions.

(i) It is clear from the defining property eq:subgraddef of the subgradient that this is a condition that is closed and convex in p. This shows $\partial f(x)$ is closed and convex. Then for $x \in \operatorname{int}\operatorname{dom} f$ there exists a supporting hyperplane in $\mathbb{R}^{n+1}$ to the epigraph at the point $(x, f(x))$. The vector orthogonal to the supporting hyperplane cannot be of the form $(p, 0)$ because x is an interior point of the domain of definition and so $\langle p, y - x\rangle$ can have any sign for y in a neighborhood of x. Thus after rescaling the hyperplane is given by a vector of the form $(p, -1)$, and then $p \in \partial f(x)$.

As convex functions are locally Lipschitz continuous we may for every $x \in \operatorname{ri}\operatorname{dom} f$ choose a Lipschitz constant L valid in a bounded neighborhood U of x relative to $\operatorname{aff}\operatorname{dom} f$. This implies

$$\langle y - x, p\rangle \leq |f(y) - f(x)| \leq L\|y - x\|$$

for all $p \in \partial f(x), y \in U$. If $\partial f(x)$ is unbounded, the left hand side of this inequality can be made arbitrarily large, while the right hand side is bounded. This contradiction shows that $\partial f(x)$ is bounded.

(ii) If f is differentiable at x, then it follows from (1.6.4), that $\nabla f(x) \in \partial f(x)$. If there are two distinct elements $p_1 \neq p_2 \in \partial f(x)$, then we have by rearranging (1.6.4)

$$\langle y, p_i\rangle \leq \frac{f(x + hy) - f(x)}{h}, \quad \text{for all } y \in \mathbb{R}^n \text{ and } h > 0 \text{ small enough.} \tag{1.7.3}$$

As $p_1 \neq p_2$ we may choose $y \in \mathbb{R}^n$ such that $\langle y, p_1\rangle < \langle y, p_2\rangle$. Then $p_1 \neq \nabla f(x)$ because otherwise for this choice of y the right hand side of (1.7.3) has to converge to $\langle \nabla f(x), y\rangle = \langle p_1, y\rangle$ but the lower bound in (1.7.3) prevents this. By exchanging the role of p_1 and p_2, we also see, that $p_2 \neq \nabla f(x)$. ∎

Before presenting some rules for the calculus with subgradients, it will be useful to present some general considerations concerning the map

$$x \mapsto \partial f(x), \quad x \in \operatorname{ri}\operatorname{dom} f,$$

where f is some convex function. As the subgradient of a convex function at a point is in general not unique this map is what is called a *set-valued map*. One of the notions of regularity of such maps is that of upper semicontinuity.

Definition 1.7.3 (Upper Semicontinuity of Set-Valued Maps) *Let $D \subset \mathbb{R}^n$ be a nonempty set and consider a set-valued map*

$$F : D \to \{ \text{ nonempty, compact subsets of } \mathbb{R}^m \}.$$

The map F is called upper semicontinuous *at $x \in D$, if for every sequence $x_k \to x$ in D and every convergent sequence $\{y_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^m$ satisfying*

$$y_k \in F(x_k), \quad \text{for all } k,$$

we have $\lim_{k\to\infty} y_k \in F(x)$.

The set-valued map is called upper semicontinuous, if it is upper semicontinuous at every point.

Proposition 1.7.4 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex, then the set-valued map*

$$x \mapsto \partial f(x)$$

is upper semicontinuous on $\operatorname{ri}\operatorname{dom} f$.

Proof. We have already seen that $\partial f(x)$ is nonempty, compact and convex on $\operatorname{ri}\operatorname{dom} f$. If $x_k \to x \in \operatorname{ri}\operatorname{dom} f$ and $p_k \to p$ with $p_k \in \partial f(x_k)$, then we have for all k

$$f(x_k) + \langle y - x_k, p_k \rangle \leq f(y)\,, \qquad \text{for all } y \in \operatorname{dom} f\,.$$

As f is continuous on $\operatorname{ri}\operatorname{dom} f$, and $x_k \to x, p_k \to p$, the inequality holds in the limit, so that $p \in \partial f(x)$. This shows upper semicontinuity. ∎

To arrive at the main result concerning the description of the subgradient we need the following simple observation.

Lemma 1.7.5 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. If for $a < 0 < b$*

$$\{x + tq\,;\ t \in [a, b]\}$$

is a line segment contained in $\operatorname{ri}\operatorname{dom} f$, and $p \in \partial f(x)$, then $\langle p, q\rangle$ is a subgradient vector of the convex function

$$g : t \mapsto f(x + tq)\,, \quad t \in [a, b]$$

at $t = 0$.

Proof. We have by definition of the subgradient, that

$$g(0) + t\langle p, q\rangle = f(x) + \langle p, tq\rangle \leq f(x + tq) = g(t)\,.$$

This shows that $\langle p, q\rangle \in \partial g(0)$. ∎

Using the fact that f is differentiable almost everywhere, we now arrive at a new description of the subgradient.

Theorem 1.7.6 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and $x \in \operatorname{ri}\operatorname{dom} f$. Then*

$$\partial f(x) = \operatorname{conv}\left\{p \in \mathbb{R}^n\,;\ \exists x_k \in \operatorname{ri}\operatorname{dom} f, x_k \to x, p = \lim_{k\to\infty} \nabla f(x_k)\right\}\,. \tag{1.7.4}$$

Note that in formulating the previous description we tacitly assume that $\nabla f(x_k)$ exists, if we write it down.

Proof. Denote

$$\mathcal{Q} := \left\{p \in \mathbb{R}^n\,;\ \exists x_k \in \operatorname{ri}\operatorname{dom} f, x_k \to x, p = \lim_{k\to\infty} \nabla f(x_k)\right\}\,.$$

As the subgradient is an upper semicontinuous set-valued map, the limits in $\mathcal{Q}$ are all elements of the subgradient $\partial f(x)$. As the subgradient is convex we obtain that $\operatorname{conv}\mathcal{Q}$ is contained in $\partial f(x)$. In particular, $\mathcal{Q}$ is bounded. Also by construction $\mathcal{Q}$ is a closed set, hence $\operatorname{conv}\mathcal{Q}$ is closed and bounded by Lemma 1.1.9.

It remains to prove that

$$\partial f(x) \subset \operatorname{conv}\mathcal{Q}\,. \tag{1.7.5}$$

Assume this is not the case. As $\partial f(x)$ and $\operatorname{conv}\mathcal{Q}$ are both closed we may pick $\varepsilon > 0$ small enough so that there is a $p \in \partial f(x) \setminus \operatorname{conv}\operatorname{cl} B_\varepsilon(\mathcal{Q})$. By assumption we may assume that for $\delta > 0$ small enough, for all points of differentiability $z \in \operatorname{cl} B_\delta(x)$ of f we have

$$\nabla f(z) \in \operatorname{cl} B_\varepsilon(\mathcal{Q})\,.$$

We may the apply the separation principle and choose $q \in \mathbb{R}^n, q \neq 0, c \in \mathbb{R}$ such that

$$\langle p, q\rangle = c \quad \text{and } \langle \tilde{p}, q\rangle < c - 2 \quad \text{for all } \tilde{p} \in B_\varepsilon(\mathcal{Q})\,.$$

We now perturb q slightly, so that the points of differentiability of f are dense on a line segment

$$[x - Tq, x + Tq]\,.$$

for some $T > 0$. For this new value of q (which we again call q) we may assume that

$$\langle p, q\rangle = c \quad \text{and } \langle \tilde{p}, q\rangle < c - 1 \quad \text{for all } \tilde{p} \in B_\varepsilon(\mathcal{Q})\,.$$

By Lemma 1.7.5 the value c is a subgradient of the map

$$t \mapsto f(x + tq)\,,$$

at the point $t = 0$. This implies by definition

$$f(x) + ct \leq f(x + tq)\,, \qquad \text{for all } t \in [-T, T]\,. \tag{1.7.6}$$

On the other hand we have for all points of differentiability of the form $x + tq \in B_\delta(x)$, that

$$\langle \nabla f(x + tq), q\rangle \leq c - 1\,.$$

If we fix $t_1 > 0$ small enough and at a point of differentiability, and $t_2 > 0$ small enough, we have using the Taylor formula,

$$\begin{aligned} f(x + (t_1 + t_2)q) &= f(x + t_1 q) + \langle \nabla f(x + t_1 q), q\rangle\ t_2 + o(t_2) \\ &\leq f(x + t_1 q) + (c - 1)t_2 + o(t_2) < f(x + t_1 q) + ct_2\,. \end{aligned} \tag{1.7.7}$$

We will bring the two statements (1.7.6) and (1.7.7) to a contradiction. First note that we obtain a convex combination by

$$\frac{t_2}{t_1 + t_2}\ x + \frac{t_1}{t_1 + t_2}\ (x + (t_1 + t_2)q) = x + t_1 q\,.$$

Using both (1.7.6) and (1.7.7) we arrive at

$$\begin{aligned} f(x + t_1 q) &= \left(\frac{t_2}{t_1 + t_2} + \frac{t_1}{t_1 + t_2}\right) f(x + t_1 q) \\ &> \frac{t_2}{t_1 + t_2}\,(f(x) + ct_1) + \frac{t_1}{t_1 + t_2}\,(f(x + (t_1 + t_2)q) - ct_2) \\ &= \frac{t_2}{t_1 + t_2} f(x) + \frac{t_1}{t_1 + t_2} f(x + (t_1 + t_2)q)\,. \end{aligned}$$

This inequality contradicts the convexity of f. This contradiction completes the proof. ∎

Proposition 1.7.7 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and $\lambda > 0$, then*

$$\partial(\lambda f)(x) = \lambda \partial f(x)\,, \quad \forall x \in \operatorname{dom} f\,.$$

Proof. A straightforward calculation. ∎

Theorem 1.7.8 *Let $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, $i = 1, \dots, k$ be convex and define the convex function*

$$F(x) := \max\{f_1(x), \dots, f_k(x)\}\,.$$

For $x \in \cap_{i=1}^k \operatorname{ri}\operatorname{dom} f_i$ we have

$$\partial F(x) = \partial\left(\max\{f_1, \dots, f_k\}\right)(x) = \operatorname{conv} \bigcup_{f_i(x)=F(x)} \partial f_i(x)\,.$$

Proof. Assume that for $x \in \operatorname{ri}\operatorname{dom} F$ and an $i_0 \in \{1, \dots k\}$ we have $F(x) = f_{i_0}(x)$. Let $p \in \partial f_{i_0}(x)$. Then we have for all $y \in \operatorname{dom} F \subset \operatorname{dom} f_{i_0}$ that

$$F(x) + \langle y - x, p\rangle = f_{i_0}(x) + \langle y - x, p\rangle \le f_{i_0}(y) \le F(y)\,,$$

where in the last step we have used the maximization in the definition of F. This shows

$$\partial F(x) \supset \bigcup_{f_i(x)=F(x)} \partial f_i(x)\,,$$

and as $F(x)$ is convex by Lemma 1.7.2 we obtain

$$\partial F(x) \supset \operatorname{conv} \bigcup_{f_i(x)=F(x)} \partial f_i(x)\,.$$

Again by Lemma 1.7.2 and applying Lemma 1.1.9 the sets on the left and the right of this inclusion are both compact.

The remainder of the proof is similar to the proof of Theorem 1.7.6 and we try to be brief.

If $f_i(x) < F(x)$, then by continuity, this inequality holds in a neighborhood of x. Restricting ourselves to this neighborhood and discarding the maps which are not involved in the maximization, we may assume that

$$F(x) = f_1(x) = \ldots = f_k(x)\,.$$

Assume there exists $p \in \partial F(x)$ and $p \notin \operatorname{conv} \cup_{i=1}^k \partial f_i(x)$. We may the apply the separation principle and choose $q \in \mathbb{R}^n, q \neq 0, c \in \mathbb{R}$ such that

$$\langle p, q\rangle = c \quad \text{and } \langle \tilde{p}, q\rangle < c - 2 \quad \text{for all } \tilde{p} \in \cup_{i=1}^k \partial f_i(x)\,.$$

By Lemma 1.7.5 the value c is a subgradient of the map

$$t \mapsto F(x + tq)\,,$$

at the point $t = 0$. This implies by definition

$$F(x) + ct \le F(x + tq)\,, \quad \text{for all } t \in [-T, T]\,. \tag{1.7.8}$$

On the other hand we have for all $\tilde{p} \in \cup_{i=1}^k \partial f_i(x)$, that

$$\langle \tilde{p}, q\rangle \le c - 2\,.$$

This implies for all $i = 1, \ldots, k$ that

$$f_i(x) + (c-1)t > f_i(x+tq)\,, \qquad \text{for all } t \in (0, T]\,.$$

This implies

$$F(x) + (c-1)t > F(x+tq)\,, \qquad \text{for all } t \in (0, T]\,.$$

The latter condition clearly contradicts (1.7.8) and this contradiction completes the proof. ■

Theorem 1.7.9 (Moreau, Rockafellar) *If f, g are both convex and bounded on $D \subset \mathbb{R}^n$, D convex, then for all $x \in \operatorname{ri} D$ we have*

$$\partial(f+g)(x) = \partial f(x) + \partial g(x)\,, \tag{1.7.9}$$

where the sum on the right is the Minkowski sum of the (convex) subgradients of f and g at x.

Proof. One direction is obvious: if $p_1 \in \partial f(x), p_2 \in \partial g(x)$, then

$$f(x) + \langle y - x, p_1 \rangle \leq f(y)$$

and

$$g(x) + \langle y - x, p_2 \rangle \leq g(y)$$

where both inequalities hold for all $y \in \operatorname{dom} f \cap \operatorname{dom} g$. Summing these two inequalities we see that $p_1 + p_2 \in \partial(f+g)(x)$. The converse direction is the one that is more tricky can be proved using a separation argument. However, we can go back to an application of Theorem 1.7.6. The equality is true on the dense set of full measure, where both f and g are differentiable, by standard calculus. The result then follows because for two closed sets we have $\operatorname{conv}(C_1 + C_2) = \operatorname{conv} C_1 + \operatorname{conv} C_2$ and we may apply this equality to (1.7.4) to obtain the desired result. ■

Note that we cannot really expect a *product rule*, as the product of two convex functions has no reason to be convex. Consider for instance the product of the functions $f(x) = x, g(x) = x^2$.

1.8 Optimality

As we are interested in optimizing convex functions we now collect some properties of minima. Recall that for a function $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$, the point x^* is a *local minimum* of f, if there exists an open neighborhood U of x^* such that

$$f(x^*) \leq f(x)\,, \quad \forall x \in U \cap D\,. \tag{1.8.1}$$

The local minimum is called *strict*, if the inequality in (1.8.1) is strict for all $x^* \neq x \in U \cap D$. The minimum is *global*, if $U = \mathbb{R}^n$.

A point x^* is a *local (strict, global) maximum*, of f, it x^* is a local (strict, global) minimum of $-f$.

Proposition 1.8.1 (Minima of Convex Functions) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then*

(i) every local minimum is a global minimum;

(ii) the set of global minima of f is convex;

(iii) if f is strictly convex, there exists at most one minimum, which is then necessarily global.

Proof. (i) Assume the assertion is false, so that there exists a local minimum x^* that is not global. Hence, there exists a y such that

$$f(x) > f(y)\,.$$

By convexity it follows that for all $\lambda \in (0,1]$

$$f(\lambda y + (1-\lambda)x^*) \leq \lambda f(y) + (1-\lambda) f(x^*) < f(x^*)\,. \tag{1.8.2}$$

In any neighborhood U of x^* there exist points of the form $\lambda y + (1-\lambda)x$, if $\lambda > 0$ is chosen small enough. Thus (1.8.2) contradicts the local minimality of x^*.

(ii) If x, y are both minima, then using minimality and then convexity we have

$$f(x) \leq f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda) f(x) = f(x)\,. \tag{1.8.3}$$

This shows that $f(x) = f(\lambda y + (1-\lambda)x)$, so that all convex combinations of x and y are minima.

(iii) If there are two minima x, y then the chain of inequalities (1.8.3) consists in fact of equalities. For strictly convex functions this cannot occur. Hence, there is at most one minimum. ∎

Theorem 1.8.2 (Subgradient Characterization) *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. Then x^* is a minimum of f, if and only if*

$$0 \in \partial f(x^*)\,.$$

Proof. If $0 \in \partial f(x^*)$ then by definition

$$f(x^*) \leq f(x)\,, \quad \forall\, x \in \operatorname{dom} f\,.$$

This shows that x^* is a minimum. Conversely, if x^* is a minimum, then the previous inequality holds, so that by definition $0 \in \partial f(x^*)$. This shows the assertion. ∎

If we consider the case of restricted optima, we obtain a slight generalization of the previous result.

Proposition 1.8.3 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and let $C \subset \operatorname{dom} f$ be convex. Then $x \in C$ is a minimal point of f with respect to C if and only if there exists $p \in \partial f(x)$ such that*

$$\langle y - x, p\rangle \geq 0\,, \qquad \textit{for all } y \in C\,.$$

Proof. If p exists then we have by the definition of the subgradient for all $y \in \operatorname{dom} f$ (and so for all y in C) that

$$f(x) + \langle y - x, p\rangle \leq f(y)\,.$$

As $\langle y - x, p\rangle \geq 0$ for $y \in C$ this implies $f(x) \leq f(y)$, so that x is a minimum point.

Conversely, assume that $f(x) \leq f(y)$ for all $y \in C$ and assume there is a $\bar{y} \in C$ such that

$$\langle \bar{y} - x, p \rangle \leq c < 0$$

for all $p \in \partial f(x)$. Note that the expression is bounded away from 0, as $\partial f(x)$ is compact. If we consider the map $g := t \mapsto f(x + t(\bar{y} - x))$, then we have by Lemma 1.7.5 for $t > 0$ sufficiently small that

$$f(x) = g(0) > g(0) + \frac{c}{2}\, t > g(t) = f(x + t(\bar{y} - x))\,.$$

This shows that x is not a minimal point of C. ∎

Note that if $x \in \operatorname{ri} C$ the previous condition implies that there is a $p \in \partial f(x)$ that is orthogonal to $\operatorname{aff} C$. If x is on the boundary of C with respect to $\operatorname{aff} C$, then C lies to the positive side of the hyperplane defined by p and the constant $c = \langle x, p \rangle$. Thus $-p$ defines a supporting hyperplane of the constraint set C.

The other type of extremum to be considered a maxima. Here the situation is not as nice as local maxima may exist and the set of maxima has no reason to be convex. On the other hand it is of interest to point out, that maxima will occur at extreme points, if a set is given as the convex hull of its extreme points.

Theorem 1.8.4 (Maxima of Convex Functions) *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and let $C \subset \operatorname{dom} f$ be convex.*

(i) If there exists an $x^* \in \operatorname{ri} C$ such that

$$f(x^*) = \sup\{f(x)\,;\ x \in C\} \tag{1.8.4}$$

then f is constant on C.

(ii) Let $W \subset C$ be the set of points x^* satisfying (1.8.4)*. Then W is a union of faces.*

(iii) Let $M \subset \operatorname{dom} f$, then

$$\sup\{f(x)\,;\ x \in M\} = \sup\{f(x)\,;\ x \in \operatorname{conv} M\}\,.$$

Proof. (i) Let $x \in C$. As $x^* \in \operatorname{ri} C$ we may extend the line segment $[x, x^*]$ a bit further and remain in $\operatorname{ri} C$. In other words, there exists a $\lambda > 0$ such that

$$y = (1 - \lambda)x + \lambda x^* \in C\,.$$

This implies that we have the convex combination

$$x^* = (1 - 1/\lambda)x + 1/\lambda y\,.$$

By convexity of f we have

$$f(x^*) \leq (1 - 1/\lambda)f(x) + 1/\lambda f(y) \leq f(x^*)\,,$$

where in the last step we have used $f(x) \leq f(x^*), f(y) \leq f(x^*)$. As we have equality throughout it follows that $f(x) = f(x^*), f(y) = f(x^*)$.

(ii) If $W = \emptyset$ there is nothing to show. Let $x \in W$. Then there is a unique face C' such that $x \in \operatorname{ri} C'$. By (i) f is constant on C'. This shows the assertion.
(iii) For convex functions level sets of the form $f^{-1}(-\infty, c) = \{x\,;\ f(x) < c\}$ are convex. Thus $M \subset f^{-1}(\infty, c)$ if and only if $\operatorname{conv} M \subset f^{-1}(-\infty, c)$. The assertion now follows. ∎

We now discuss some important cases of convex functions; these are

(i) affine functions

(ii) quadratic functions

(iii) norms

(iv) convex functions of affine functions.

Affine Functions

Linear functions $f : \mathbb{R}^n \to R$ are given by vectors $n \in \mathbb{R}^n$, a constant $c \in \mathbb{R}$ and are the form $x \mapsto \langle x, n\rangle$. The epigraph of such a function is then just $\operatorname{epi} f = \overline{H}_+(n, c)$. The supporting hyperplane of the epigraph is then exactly the hyperplane defined by n, c so that $\partial f(x) = \{n\}$ for all $x \in \mathbb{R}^n$. In infinite dimensions the same reasoning applies, if we restrict our attention to *continuous* affine functions.

Quadratic Functions

Quadratic functions are given by symmetric matrices $P \in \mathcal{H}_n$, $n \in \mathbb{R}^n$, $c \in \mathbb{R}$. They are of the form

$$x \mapsto x^\top P x + \langle x, q\rangle + c\,.$$

Such functions are convex if and only if the Hessian is positive semidefinite everywhere, which is equivalent to $\Pi \geq 0$. As the function is differentiable, the subgradient is given by the gradient and is of the form

$$\partial f(x) = \left\{ \frac{1}{2} x^\top P + n \right\}.$$

Norms For norms we explicitly restrict our attention to $\mathbb{R}^n$, as in finite dimensions all norms are equivalent.

The definition given for a *norm* is usually the following.

Definition 1.8.5 (Norm) *A norm on $\mathbb{R}^n$ is a function $v : \mathbb{R}^n \to [0, \infty)$ satisfying for all $x, y \in \mathbb{R}^n, \lambda \in \mathbb{R}$:*

(i) (positive definiteness): $v(x) \geq 0$, and $v(x) = 0 \Leftrightarrow x = 0$.

(ii) (positive homogeneity): $v(\lambda x) = |\lambda| v(x)$,

(iii) (triangle inequality): $v(x + y) \leq v(x) + v(y)$.

Indeed the combination of *homogeneity* and the *triangle inequality* implies that a norm is a convex function. To see this note that for arbitrary $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$ we have by first using the triangle inequality and then homogeneity that

$$v(\lambda x + (1 - \lambda)y) \leq v(\lambda x) + v((1 - \lambda)y) = \lambda v(x) + (1 - \lambda) v(y)\,.$$

Conversely, we see that from convexity and positive homogeneity the triangle inequality follows. So norms could as well be defined by the three requirements (i) positive definiteness, (ii) positive homogeneity, (iii) convexity.

As norms are convex the respective closed unit ball

$$\overline{B} := \{x \in \mathbb{R}^n \,;\; v(x) \leq 1\}$$

is a convex set. Note that $B = -B$, so B is symmetric, by positive homogeneity. In fact, there is a one-to-one correspondence between norms and closed symmetric convex sets $\overline{B}$, containing 0 in its interior. Any such set defines a norm via the *Minkowski function* given by

$$m(x) := \inf\{\lambda > 0 \,;\; \lambda^{-1}x \in \overline{B}\}\,.$$

Recall that for a fixed norm v on $\mathbb{R}^n$ the *dual norm* is defined by

$$v^*(x) := \max\{\langle l, x\rangle \mid v(l) \leq 1\}\,. \tag{1.8.5}$$

This definition may be understood as follows. If $x = 0$ the maximization yields the value 0. For $x \neq 0$ for all l in the unit ball of the norm v we have that

$$\langle l, x\rangle \leq v^*(x) \tag{1.8.6}$$

and for a particular l^* which is maximizing in the definition (1.8.5) equality holds. This means that the unit ball

$$\bar{B}_1^v(0) := \{l \in \mathbb{R}^n \,;\; v(l) \leq 1\} \subset \overline{H}_-(x, v^*(x))\,, \tag{1.8.7}$$

i.e. the closed unit ball of the norm v lies in the closed negative hyperplane defined by the vector x and the value $v^*(x)$. In the points l^* in which the maximum is attained we have a supporting hyperplane given as $H(x, v^*(x))$. This situation is depicted for two different vectors and the $\|\cdot\|_1$ norm in Figure 1.18 and for an elliptic norm in Figure 1.19.

We note that (1.8.6) holds in particular for all vectors l with $v^*(l) = 1$. By scalar multiplication we obtain that for all $x, l \in \mathbb{R}^n$

$$\langle l, x\rangle \leq v^*(x)v(l)\,. \tag{1.8.8}$$

Lemma 1.8.6 *Let v be a norm on $\mathbb{R}^n$, then $v = (v^*)^*$.*

Proof. For the origin there is nothing to show. So pick $l \in \mathbb{R}^n$ with $(v^*)^*(l) = 1$. By definition we have

$$(v^*)^*(l) = \max\{\langle l, x\rangle \mid v^*(x) \leq 1\} \tag{1.8.9}$$

So if we pick a maximizing x, which then necessarily has norm $v^*(x) = 1$, we obtain

$$1 = (v^*)^*(l)\; v^*(x) = \langle l, x\rangle \leq v^*(x)v(l)\,.$$

Where in the last step we have used (1.8.8). This shows that $v \geq (v^*)^*$ or equivalently, that the unit ball of v is a subset of the unit ball of $(v^*)^*$.

On the other hand if we consider l such that $v(l) = 1$, then by (1.8.7) there exists a supporting hyperplane of $\overline{B}^v$ in l and by rescaling x if necessary we may assume that

$$\bar{B}_1^v(0) \subset \overline{H}_-(\hat{x}, 1)\,, \tag{1.8.10}$$

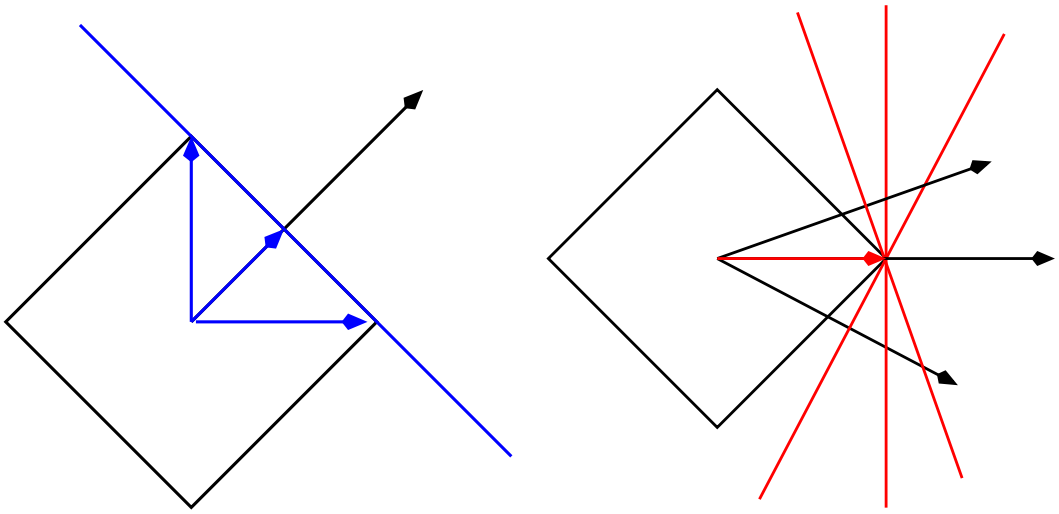

Figure 1.18: Vectors and dual vectors for $\|\cdot\|_1$. To the left the hyperplane defined by the vector $x = (1,1)^\top$ intersects the closed unit ball in a line segment. All the vectors in that segment are maximizing for (1.8.5). Three of these are indicated. To the right we see three different vectors pointing to the right. The supporting hyperplanes defined by these vectors all support the unit ball in $(1,0)^\top$. This vector is thus the maximizing vector for (1.8.5) in all three cases.

which implies $v^*(\hat{x}) = 1$. As the hyperplane is supporting in l we have that

$$v(l) = 1 = \langle l, \hat{x}\rangle \leq (v^*)^*(l)\,,$$

so that we obtain $v \leq (v^*)^*$. This completes the proof. ∎

Given a norm v and its dual v^* a pair of vectors $l, x \in \mathbb{R}^n$ is called *dual pair*, if $\langle l, x\rangle = v(x)v^*(l)$. In terms of supporting hyperplanes this means that for dual pairs (x, l), x defines a supporting hyperplane

A vector l is called dual to $x \in \mathbb{R}^n$, if $v^*(l) \leq 1$ and $\langle x, l\rangle = v(x)$.

For further details on dual norms we refer to [5].

The interest in the dual norm is that is describes the subgradient vectors of a norm.

Proposition 1.8.7 *Let v be a norm on $\mathbb{R}^n$. Then for all $x \in \mathbb{R}^n$*

$$\partial v(x) = \{p \in \mathbb{R}^n\,;\; v^*(p) \leq 1, \langle p, x\rangle = v(x)\}\,. \tag{1.8.11}$$

We note the easy relation that if $p \in \{p \in \mathbb{R}^n\,;\; v^*(p) \leq 1, \langle p, x\rangle = v(x)\}$ then we have for all $y \in \mathbb{R}^n$ that

$$v(x) + \langle y - x, p\rangle = v(x) - v(x) + \langle y, p\rangle \leq (v^*)^*(y) = v(y)\,.$$

This shows that p is a subgradient of v at x. However, it is more convenient to base the proof on a more general principle.

Proof. We know that

$$v(x) = \max\{\langle x, l\rangle\,;\; v^*(l) \leq 1\}\,.$$

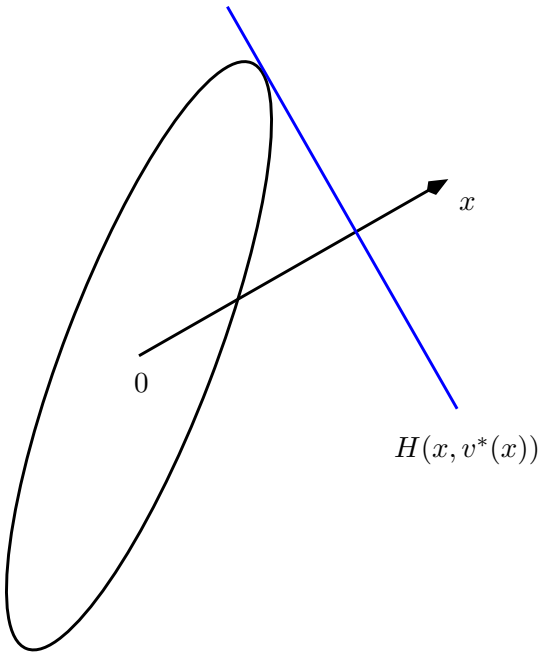

Figure 1.19: Dual vector for an elliptic norm. The norm ball lies to the negative side of the hyperplane $H(x, v^*(x))$. The maximizing vector in the norm ball is the point where hyperplane and norm ball meet.

That is we obtain v as a maximum of a set of linear functions. The subgradient of the function $x \mapsto \langle x, l\rangle$ is the vector l. By Theorem 1.7.8 the subgradient of the function defined by the maximization procedure is thus

$$\partial v(x) = \operatorname{conv}\{p \in \mathbb{R}^n\,;\; v^*(p) \leq 1, \langle p, x\rangle = v(x)\}\,.$$

As the set on the right is already convex it is not necessary to take the convex hull and we obtain the desired equality. ∎

For some specific cases it is of interest to describe the norm and its dual.

- the Euclidean norm $\|\cdot\|_2$: Here it is easy to see that the maximization of

  $$\|x\|_2^* = \max\{\langle l, x\rangle \mid \|l\|_2 \leq 1\}$$

  results in $\|x\|_2^* = \|x\|_2$. Dual pairs are always given by vectors and nonnegative scalar multiples thereof. As the norm is differentiable away from 0, it is a straightforward calculation that the subgradient vector of the Euclidean norm in $x \neq 0$ is given by

  $$\frac{x}{\|x\|_2}\,.$$

- Elliptic norms: for positive definite matrices $P \in \mathcal{H}_n$ we define the norm

  $$\|x\|_P := \left(x^T P x\right)^{1/2} = \|P^{1/2}x\|_2\,,$$

where $P^{1/2}$ is chosen to be positive definite so that $P^{1/2}P^{1/2} = P$. If we compute the dual norm we obtain

$$\begin{aligned} \|x\|_P^* &= \max\{\langle l, x\rangle \mid \|l\|_P \leq 1\} &&= \max\{\langle l, x\rangle \mid \|P^{1/2}l\|_2 \leq 1\} \\ &= \max\{\langle P^{-1/2}l, x\rangle \mid \|l\|_2 \leq 1\} &&= \max\{\langle l, P^{-1/2}x\rangle \mid \|l\|_2 \leq 1\} \\ &= \|P^{-1/2}x\|_2 = \left(x^T P^{-1} x\right)^{1/2} . \end{aligned}$$

Note that in the last step we have used the fact that we know that the dual norm of the Euclidean norm is again the Euclidean norm. Given $x \in \mathbb{R}^n$ the only vectors l such that (x, l) is a dual pair are scalar multiples of $l = Px$. To see this note that in this case

$$\langle x, l\rangle = \langle x, Px\rangle = \langle x, Px\rangle^{1/2}\langle Px, P^{-1}Px\rangle^{1/2} = \|x\|_P \|Px\|_{P^{-1}} .$$

The unit ball of an elliptic norm and its dual is depicted in Figure 1.20.

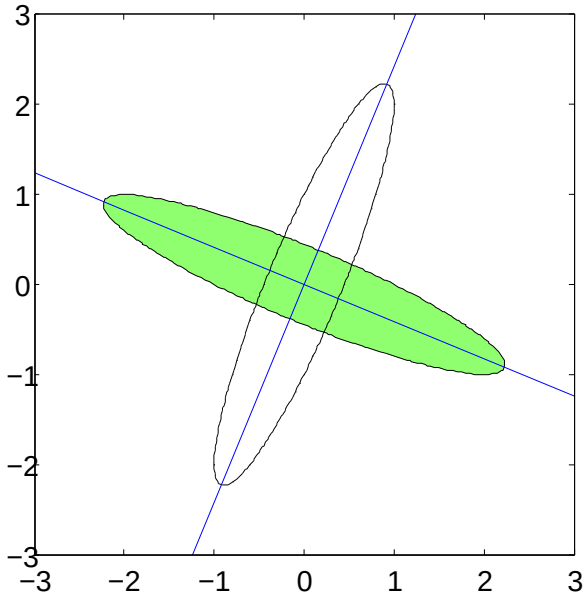

Figure 1.20: A filled elliptic norm ball and the outlines of the unit ball of the dual norm. The lines represent the eigenspaces of the matrix P.

Convex Functions of Affine Functions

If $f : \mathbb{R}^n \to R$ is convex, $A \in \mathbb{R}^{n\times m}, b \in \mathbb{R}^n$, then the map $g(x) = f(Ax + b)$ is again convex. A common application of this fact consists in the considerations of functions of the form $x \mapsto \|Ax - b\|$. It is straightforward to check this fact, because we have

$$\begin{aligned} f(A(\lambda x + (1-\lambda)y) + b) &= f(\lambda(Ax + b) + (1-\lambda)(Ay + b)) \\ &\leq \lambda f(Ax + b) + (1-\lambda) f(Ay + b) . \end{aligned}$$

The epigraph of g is given by

$$\begin{aligned}\operatorname{epi} g &= \{(x,z) \in \mathbb{R}^m \times \mathbb{R}\,;\; z \geq f(Ax+b)\} \\ &= \{(x,z) \in \mathbb{R}^m \times \mathbb{R}\,;\; Ax+b = y, z \geq f(y)\} \\ &= \{(x,z) \in \mathbb{R}^m \times \mathbb{R}\,;\; x \in A^{-1}(y-b), z \geq f(y)\} \\ &= \{(x,z) \in \mathbb{R}^m \times \mathbb{R}\,;\; x \in A^{-1}y, (y+b,z) \in \operatorname{epi} f\}\,.\end{aligned}$$

It follows that $(p,-1)$ defines a supporting hyperplane of $\operatorname{epi} f$ in the point $(y-b,z)$ if and only if $(pA,-1)$ defines a supporting hyperplane of $\operatorname{epi} g$ in $(x = Ay+b, z)$. Consequently, we have the following chain rule like condition for the subgradients

$$\partial g(x) = \partial f(Ax+b)A = \{pA\,;\; p \in \partial f(Ax+b)\}\,.$$

Note that in the previous characterization we have to interpret p as a row vector.

Chapter 2

Convex Optimization

2.1 Optimization Problems

The desire to optimize a choice of a variable to attain a goal pervades modern life. In applications reaching from the banal to fundamental optimal cost, fuel consumption, time spent or a multitude of other criteria are of utmost importance. The idea to use calculus for identifying optima goes back at least to Fermat and Lagrange, while Newton and Gauss proposed iterative methods for finding optimal points. In applications these iterative methods are now of predominant importance as direct computation of optima is in general hard.

The methods and their analysis usually on criteria for optimality and we will begin by deriving such criteria.

If we are thinking about minimization problems the function f to be optimized is usually called the *cost function*. A general optimization problem would be of the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in C\,, \end{aligned} \tag{2.1.1}$$

$f : U \to \mathbb{R}$ is defined on some domain of definition $U \subset \mathbb{R}^n$ and where $C \subset U$ is a *constraint set*. If $C = U = \mathbb{R}^n$, then the problem is called *unconstrained*.

The optimal value of (2.1.1) is defined by

$$p^* := \inf\{f(x)\,;\ x \in C\}\,. \tag{2.1.2}$$

In this formulation the values $p^* = \infty$ or $p^* = -\infty$ are admitted. In the first case $C = \emptyset$ and the problem is called *infeasible*. If $p^* = -\infty$ then the problem is said to be *unbounded from below*.

In solving a problem of the form (2.1.1) it is usually necessary to arrive at some analytic description of the constraint set. We will assume that C is given by a set of inequalities and equalities. So that C is of the form

$$C = \{x\,;\ g_j(x) \leq 0, j = 1, \ldots, m,\ h_j(x) = 0, j = 1, \ldots, p\}\,.$$

If we have a formulation of this type the functions g_j define the *inequality constraints*, while the functions h_j define the *equality constraints*. A point $x \in \mathbb{R}^n$ is called *feasible*, if $x \in C$ or given the analytic description, if x satisfies all inequality and equality constraints.

Within the analytic description we have just discussed an optimization problem is a problem of the following type

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_j(x) \leq 0 & j = 1, \ldots, m \\ & h_j(x) = 0 & j = 1, \ldots, p\,. \end{array} \tag{2.1.3}$$

The *domain* $\mathcal{D}$ of the optimization problem is the set of points, where the cost function as well as the constraint functions are defined. That is, $\mathcal{D}$ is given by the intersection

$$\mathcal{D} := \operatorname{dom} f \ \cap \ \bigcap_{j=1}^{m} \operatorname{dom} g_j \ \cap \ \bigcap_{j=1}^{p} \operatorname{dom} h_j\,.$$

A point x^* is called an *optimal point* for (2.1.3), if x^* is feasible and $f(x^*) = p^*$.

Of course, by changing the sign of f we can describe equally well a maximization problem of the type

$$\begin{array}{lll} \text{maximize} & -f(x) & \\ \text{subject to} & g_j(x) \leq 0 & i = 1, \ldots, m \\ & h_j(x) = 0 & i = 1, \ldots, p\,. \end{array} \tag{2.1.4}$$

In general, the optimization problem is called *convex optimization problem*, if

(i) the cost function f is convex;

(ii) the feasibility set is convex.

For these notes we will consider a restricted notion of convexity. It is of course, possible to describe convex sets with the help of arbitrary inequality and equality constraints. However, it is very often helpful, if the constraints are given in a way that is also amenable to convex considerations. We will therefore consider convex problems in the standard form given by

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_j(x) \leq 0 & j = 1, \ldots, m \\ & Ax = b\,. & \end{array} \tag{2.1.5}$$

Where for (2.1.5) it is assumed, that f and the $g_j, j = 1, \ldots, m$ are convex and $A \in \mathbb{R}^{p\times n}, b \in \mathbb{R}^p$. As the linear equality constraint is well-defined everywhere, the constraint set in the convex case is given by

$$\mathcal{D} := \operatorname{dom} f \ \cap \ \bigcap_{j=1}^{m} \operatorname{dom} g_j\,.$$

And the *feasible set* set is given by

$$F(g, A, b) := \{x \in \mathbb{R}^n\,;\ g_j(x) \leq 0, j = 1, \ldots, m, Ax = b\}$$

Note that by Proposition 1.8.1 the set of optimal points of a convex optimization problem is convex.

Maximization of a concave function can always be reformulated as a minimization problem for a convex function and the characterizations of optima obtained in Proposition 1.8.1 apply. As we have seen in Theorem 1.8.4 a quite different problem occurs, if we want to maximize convex functions. We will not really treat this problem. But just note that it is computationally less pleasant, as optimal points will usually lie in lower dimensional faces of the domain. And the computation of a local maximum does not guarantee that a global maximum has been found.

In case that the cost function is differentiable, then an easy criterion for optimality is the following.

Linear Optimization Problems

A special case of optimization problems consist of problems where an affine function has to be optimized over a convex polytope. This is a *linear optimization problem* also called a *linear program*. The general problem then has the form

$$\begin{aligned} \text{minimize} \quad & c^\top x + d \\ \text{subject to} \quad & Gx \leq h \\ & Ax = b\,. \end{aligned} \tag{2.1.6}$$

Here $x \in \mathbb{R}^n$. The linear function to be optimized is given by $c \in \mathbb{R}^n$ and $d \in \mathbb{R}$. Of, course the constant d is not really relevant to the solution of this problem, since we know the optimum if we optimize the function $x \mapsto c^\top x$. The inequality constraints are given by a matrix $G \in \mathbb{R}^{m\times n}$ and $h \in \mathbb{R}^m$ and the inequality is to be understood componentwise. The equality constraints are again given by $A \in \mathbb{R}^{p\times n}, b \in \mathbb{R}^p$.

We note the following consequence of analysis of *maximization* of convex problems. As linear functions are concave as well as convex, the minimization problem (2.1.6) is equivalent to the maximization of the convex function $x \mapsto -c^\top x - d$. We thus obtain the following result immediately from Theorem 1.8.4.

Theorem 2.1.1 (Optimal Points of Linear Programs) *Consider the linear program* (2.1.6). *If an optimal point x^* exists, then the optimum is attained on some faces of the feasible set*

$$\{x \in \mathbb{R}^n\,;\, Gx \leq h\,, \quad Ax = b\}\,.$$

If the feasible set is a bounded polytope, then the optimal value of the problem is attained in an extreme point of the feasible set.

Proof. The first statement is an immediate consequence of Theorem 1.8.4 (ii) by considering the equivalent problem of maximizing the convex function $-c^\top x - d$. If the feasible set is a bounded polytope, then any face contains extreme points of the feasible sets and the result follows from Theorem 1.8.4 (i). ∎

The situation of optimizing a linear function over a polytope is depicted in Figure 2.1. Note that the level sets of the affine maps define hyperplanes in the space. The optimal value is obtained at the extreme point which has been enlarged.

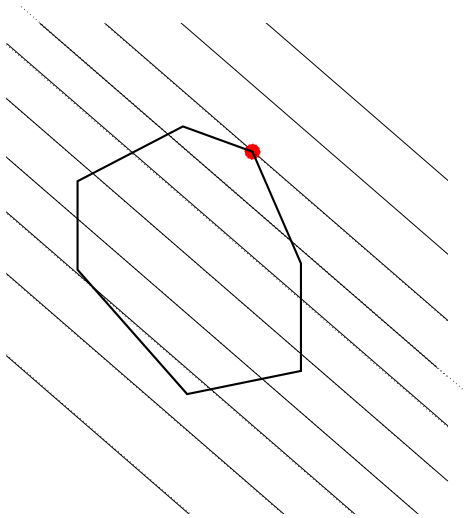

Figure 2.1: Optimal point of a linear program over a bounded polytope.

There are two standard forms of linear programs, consisting on one hand in the formulation

$$\begin{aligned} \text{minimize} \quad & c^\top x + d \\ \text{subject to} \quad & Ax = b \\ & x \geq 0\,. \end{aligned} \tag{2.1.7}$$

and on the other hand in

$$\begin{aligned} \text{minimize} \quad & c^\top x + d \\ \text{subject to} \quad & Ax \leq b \end{aligned} \tag{2.1.8}$$

if there are no equality constraints. It is always possible to transform a general linear optimization problem into the form (2.1.7), resp. (2.1.8) by introducing *slack variables* and positive variables. Slack variables transform the inequality constraint $Gx \leq h$ into the equivalent formulation $Gx + s = h, s \geq 0$, where the new artificial variables s represent the slack in the inequality constraint. On the other hand the variable x can be written as $x = x^+ - x^-, x^+ \geq 0, x^- \geq 0$, which allows to only look for solutions with positive entries.

Quadratic Optimization Problems

A convex optimization problem is called a *quadratic program* if it given by a quadratic cost function, which is to be optimized over a convex polytope. The general form of such a problem is

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x^\top Px + c^\top x + d \\ \text{subject to} & Gx \leq h \\ & Ax = b\,. \end{array} \tag{2.1.9}$$

Here P is a real symmetric positive semidefinite matrix and the remaining data are just as in the case of the general linear optimization problem. If also the constraint set is described by convex quadratic functions, then the resulting problem is called a *quadratically constrained quadratic program.* This type of problem is of the form

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x^\top Px + c^\top x + d \\ \text{subject to} & \frac{1}{2}x^\top Q_i x + q_i^\top x + r_i \leq 0\,, \quad i = 1, \ldots, m\,, \\ & Ax = b\,, \end{array} \tag{2.1.10}$$

where now also the matrices $Q_i \in \mathbb{R}^{n\times n}$ are positive semidefinite, $q_i \in \mathbb{R}^n$, $r_i \in \mathbb{R}$.

Semidefinite Optimization Problems

Semidefinite optimization problems or semidefinite programs are a generalization of linear programs. In these problems constraints are formulated within the space of symmetric matrices $\mathcal{H}_n$. In the following inequalities have to be understood with respect to the notion of positive definite matrices. That is for symmetric matrices , P, Q we have $P \geq Q$ if $P - Q$ is positive semidefinite. This order defines an order on $\mathcal{H}_n$, see also (1.1.2). It is the order generated by the cone of positive semidefinite matrices, cp. (1.1.11). A semidefinite program has the form

$$\begin{array}{rl} \text{minimize} & c^\top x \\ \text{subject to} & x_1F_1 + x_2F_2 + \ldots + x_nF_n + G \leq 0\,, \\ & Ax = b\,. \end{array} \tag{2.1.11}$$

Here $x = \begin{pmatrix} x_1 & \ldots & x_n \end{pmatrix}^\top \in \mathbb{R}^n$, $G, F_1, \ldots, F_n \in \mathcal{H}_n$, $A \in \mathbb{R}^{p\times n}, b \in \mathbb{R}^p$. The problem reduces to a linear program if the matrices $G, F_1, \ldots, F_n$ are all diagonal.

The standard form of a semidefinite program is given in terms of a variable $X \in \mathcal{H}_n$. Similarly to the case of a linear program it can be obtained by introducing slack variables and nonnegative variables.

$$\begin{array}{rl} \text{minimize} & \operatorname{trace} CX \\ \text{subject to} & \operatorname{trace} A_iX = b_i\,, \quad i = 1, \ldots, m\,, \\ & X \geq 0\,, \end{array} \tag{2.1.12}$$

Note that $\operatorname{trace} CX = \sum_{i,j=1}^n c_{ij}x_{ij}$ can represent any linear map from $\mathcal{H}_n$ to $\mathbb{R}$.

2.2 Lagrange Formalism

We consider an optimization problem in standard from given by

$$\begin{aligned} \text{minimize} \quad & f(x) && && (2.2.1)\\ \text{subject to} \quad & g_j(x) \le 0 && i = 1, \dots, m && (2.2.2)\\ & h_j(x) = 0 && i = 1, \dots, p\,. && (2.2.3) \end{aligned}$$

Here $x \in \mathbb{R}^n$, the domain $\mathcal{D}$ is assumed to be nonvoid. The optimal value of the problem is denoted by p^*. For now we do not assume convexity of the problem.

The fundamental idea in the Lagrangian approach is to make the constraints part of the objective function. The Lagrangian associated with (2.2.1) is defined by

$$L(x, \lambda, \nu) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) + \sum_{j=1}^{p} \nu_j h_j(x)\,, \tag{2.2.4}$$

with domain of definition

$$\operatorname{dom} L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p\,.$$

It is standard to call the λ_j the *Lagrange multipliers* associated with the inequality constraint g_j, whereas the ν_j are the Lagrange multipliers corresponding to the equality constraint h_j.

The *Lagrange dual function* is now defined as

$$g_L(\lambda, \nu) := \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)\,. \tag{2.2.5}$$

Note that it is possible, that the value of g_L at a point is $g_L(\lambda, \nu) = -\infty$, as L may be unbounded from below in x.

The first simple observation is the following lower bound property.

Lemma 2.2.1 *For all $\lambda \ge 0$, $\nu \in \mathbb{R}^p$ we have*

$$g_L(\lambda, \nu) \le p^*\,.$$

Proof. If x^* is feasible and $\lambda \ge 0$, then by definition of the constraints we have

$$\sum_{j=1}^{m} \lambda_j g_j(x^*) + \sum_{j=1}^{p} \nu_j h_j(x^*) \le 0\,.$$

Hence

$$L(x^*, \lambda, \nu) = f(x^*) + \sum_{j=1}^{m} \lambda_j g_j(x^*) + \sum_{j=1}^{p} \nu_j h_j(x^*) \le f(x^*)\,.$$

And so

$$g_L(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \le L(x^*, \lambda, \nu) \le f(x^*)\,.$$

As the right hand side is independent of x^* we can minimize this inequality over x^* and obtain $g_L(\lambda, \nu) \le p^*$. ∎

The previous lemma and the observation that the lower bound is independent of x, motivates an optimization over the dual variables λ, ν. This yields the *Lagrange dual problem*

$$\begin{aligned} \text{maximize} \quad & g_L(\lambda, \nu) \\ \text{subject to} \quad & \lambda \geq 0\,. \end{aligned} \tag{2.2.6}$$

The value of this optimization problem is denoted by d^* (as in dual), so that

$$d^* = \sup\{g_L(\lambda, \nu)\,;\ \lambda \in \mathbb{R}^m, \lambda \geq 0, \nu \in \mathbb{R}^p\}\,.$$

Now the use of the symbol p^* for the original, the so-called *primal* problem becomes clear.

Definition 2.2.2 (Duality Gap and Strong Duality) *Consider the primal problem* (2.2.1) *and the associated dual problem* (2.2.6)*. The* duality gap *of the problem is given by the difference*

$$p^* - d^* \geq 0\,.$$

If the duality gap is equal to zero, then we say that strong duality *holds.*

Even for convex problems strong duality need not hold, but very often it does. There are strong conditions which guarantee that such an assumption is true. These are the *Slater conditions*. We now consider the convex optimization problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_j(x) \leq 0 \qquad\qquad j = 1, \ldots, m \\ & Ax = b\,, \end{aligned} \tag{2.2.7}$$

with the standard assumption that $f, g_1, \ldots, g_m$ are convex, $A \in \mathbb{R}^{p\times n}, b \in \mathbb{R}^p$.

Theorem 2.2.3 (Slater's constraint qualification) *Consider the convex optimization problem* (2.2.7)*. If there exists an* $x \in \operatorname{ri}\mathcal{D}$ *such that*

$$Ax = b \quad \text{and} \quad g_j(x) < 0\,, \quad j = 1, \ldots, m \tag{2.2.8}$$

then strong duality holds.

A point $x \in \operatorname{ri}\mathcal{D}$ satisfying (2.2.8) is called a *Slater point*. Such a point is sometimes also called *strongly feasible*.

Remark 2.2.4 There exists a refinement of Slater's condition concerning the inequality constraints g_j. If some of these, say $g_1, \ldots, g_k$, are given by affine functions, then it is sufficient that the point in the relative interior satisfies

$$g_j(x) \leq 0\,, \quad j = 1, \ldots, k, \quad g_j(x) < 0\,, \quad j = k+1, \ldots, m\,.$$

In particular, if all inequality constraints are affine, it is sufficient that there is a point in the relative interior of $\mathcal{D}$.

Proof. We may assume without loss of generality that $\operatorname{rank} A = p$, as otherwise there are redundancies in the equality constraints and we can equivalently reduce the matrix size.

We may also assume that the domain $\mathcal{D}$ has interior points. Otherwise, we may introduce new coordinates for the smallest affine space containing $\mathcal{D}$ and work in these new variables.

The proof relies on an application of the separation properties of convex sets. To this end we introduce the set

$$\mathcal{A} := \{(u, v, t) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}\,;\ \exists x \in \mathcal{D} : g_j(x) \leq u_j, j = 1, \ldots, m; Ax - b = v; f(x) \leq t\}\,.$$

First note that $\mathcal{A}$ is convex. To see this[1], let $(u^1, v^1, t^1), (u^2, v^2, t^2) \in \mathcal{A}$ and denote by x^1, resp. x^2 the values of $x \in \mathcal{D}$ such that the requirements of the construction of the set are met. Fix $\alpha \in (0, 1)$. Then for $j = 1, \ldots, m$ and $x_\alpha = \alpha x^1 + (1-\alpha)x^2$ we have by convexity of g_j

$$g_j(x_\alpha) \leq \alpha g_j(x^1) + (1-\alpha) g_j(x^2) \leq \alpha u_j^1 + (1-\alpha) u_j^2\,.$$

A similar argument applies for $\alpha t^1 + (1-\alpha)t^2$ using the convexity of f and the condition on v^1, v^2 follows form linearity of A. This shows convexity of $\mathcal{A}$.

We also point out that if $x \in \mathcal{D}$, then by definition

$$((g_1(x), \ldots, g_j(x))^\top, Ax - b, f(x)) \in \mathcal{A}. \tag{2.2.9}$$

Now define a second set $\mathcal{B}$ by

$$\mathcal{B} = \{(0, 0, s) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}\,;\ s < p^*\}\,.$$

We claim that $\mathcal{A} \cap \mathcal{B} = \emptyset$. This is the case because a point (u, v, t) in the intersection would satisfy $u = 0, v = 0, t < p^*$ as an element of $\mathcal{B}$. As an element of $\mathcal{A}$ there has to exist an $x \in \mathcal{D}$ such that $g_j(x) \leq u_j = 0$, $Ax = b$ and $f(x) \leq t < p^*$. But then x is a feasible point with value less than p^*, which contradicts the definition of the value of the problem.

As $\mathcal{A}$ and $\mathcal{B}$ are convex with empty intersection there exists a separating hyperplane, i.e. there exists $(\tilde{\lambda}, \tilde{\nu}, \mu) \neq 0$ and an $\alpha \in \mathbb{R}$ such that

$$(u, v, t) \in \mathcal{A} \Longrightarrow \tilde{\lambda}^\top u + \tilde{\nu}^\top v + \mu t \geq \alpha \tag{2.2.10}$$

and

$$(u, v, t) \in \mathcal{B} \Longrightarrow \tilde{\lambda}^\top u + \tilde{\nu}^\top v + \mu t \leq \alpha\,. \tag{2.2.11}$$

By definition of $\mathcal{A}$, if $(u, v, t) \in \mathcal{A}$, then for all $r > 0$ also the vector $(u + re_j, v, t) \in \mathcal{A}$. This implies that $\tilde{\lambda}_j \geq 0$ as otherwise, by choosing r large enough, the expression on the right hand side of (2.2.10) is unbounded from below, contradicting the existence of the lower bound α. Hence $\tilde{\lambda} \geq 0$ and by a similar argument $\mu \geq 0$.

By definition of $\mathcal{B}$ (2.2.11) simplifies to $\mu t < \alpha$ for all $t < p^*$ and hence $\mu p^* \leq \alpha$.

Combining these statements with (2.2.9) it follows that for all $x \in \mathcal{D}$ we have

$$\sum_{j=1}^{m} \tilde{\lambda}_j g_j(x) + \tilde{\nu}^\top (Ax - b) + \mu f(x) \geq \alpha \geq \mu p* \ . \tag{2.2.12}$$

[1] Note that the superscripts are used as indices now.

We now consider two cases. If $\mu > 0$ then we can divide (2.2.12) by μ and obtain for all $x \in \mathcal{D}$ that

$$L(x, \tilde{\lambda}/\mu, \tilde{\nu}/\mu) \geq p^* \,.$$

Now weak duality, see Lemma 2.2.1, and minimization over x yields $p^* \geq g(\tilde{\lambda}/\mu, \tilde{\nu}/\mu) \geq p^*$. This shows that strong duality holds and that the optimal value d^* is obtained.

It remains to consider the case $\mu = 0$. It then follows from (2.2.12) that

$$\sum_{j=1}^{m} \tilde{\lambda}_j g_j(x) + \tilde{\nu}^\top (Ax - b) \geq \alpha \geq 0 \,.$$

In particular, this holds for the point x^* satisfying the Slater condition. As this point is feasible by assumption, we obtain

$$\sum_{j=1}^{m} \tilde{\lambda}_j g_j(x^*) \geq 0$$

and as $g_j(x^*) < 0$ by assumption and $\tilde{\lambda} \geq 0$ we see that $\tilde{\lambda} = 0$. As $(\tilde{\lambda}, \tilde{\nu}, \mu)$ were chosen to define a hyperplane, it follows from $\mu = 0, \tilde{\lambda} = 0$ that $\tilde{\nu} \neq 0$. Then (2.2.12) implies that

$$\tilde{\nu}^\top (Ax - b) \geq 0$$

for all $x \in \mathcal{D}$. Furthermore by the rank condition on A we have $\tilde{\nu}^\top A \neq 0$. But the Slater point x^* satisfies $\tilde{\nu}^\top (Ax^* - b) = 0$ and as $x^* \in \operatorname{int} \mathcal{D}$ there are points $x \in D$ with $\tilde{\nu}^\top (Ax - b) < 0$. This contradiction completes the proof. ∎

We now regress for a moment from convex problems and discuss optimization problems of the form (2.2.1) in general. If we assume that primal and dual value are attained, and that there is no duality gap, i.e. they are equal, then we arrive at

$$\begin{aligned} f(x^*) &= g_L(\lambda^*, \nu^*) \\ &= \inf_x \left(f(x) + \sum_{j=1}^{m} \lambda_j^* g_j(x) + \sum_{j=1}^{p} \nu^* h_j(x) \right) \\ &\leq f(x^*) + \sum_{j=1}^{m} \lambda_j^* g_j(x^*) + \sum_{j=1}^{p} \nu^* h_j(x^*) \\ &\leq f(x^*) \,. \end{aligned}$$

The first equality is the statement that the duality gap is zero, followed by the definition of the Lagrangian function g_L. The third inequality is a property of the infimum and finally we use that $\lambda_j^* \geq 0, g_j(x^*) \leq 0$ and $h_j(x^*) = 0$, which follows from the constraints of the primal and dual optimization problem.

The chain of inequalities implies that we have equality throughout. We thus arrive at the following two statements.

Lemma 2.2.5 *Consider the optimization problem* (2.2.7) *and its dual problem. Assume that for these problems the optimal values are attained in x^*, respectively, (λ^*, ν^*) and that strong duality holds. Then x^* minimizes $L(x, \lambda^*, \nu^*)$.*

A further interesting fact concerns the relation of Lagrange multipliers and the behavior of the inequality constraints at the optimal points.

Proposition 2.2.6 *Consider the optimization problem* (2.2.7) *and its dual problem. Assume that for these problems the optimal values are attained in* x^**, respectively,* (λ^*, ν^*) *and that strong duality holds. Then*

$$\lambda_j^* > 0 \;\Rightarrow\; g_j(x^*) = 0\,, \tag{2.2.13}$$

or equivalently

$$g_j(x^*) < 0 \;\Rightarrow\; \lambda_j^* = 0\,. \tag{2.2.14}$$

The conditions (2.2.13), resp. (2.2.14), are called *complementary slackness* conditions. They may be interpreted through the notion of "activity" of a constraint. An inequality constraint given by g_j is said to be active at the optimum point x^*, if in fact $g_j(x^*) = 0$. Otherwise, it is inactive, as in fact the condition $g_j < 0$ poses no hard restriction on the optimal point. Thus Lagrange multipliers are active if the corresponding inequality constraint is inactive or vice versa.

In general the existence of a strictly feasible point as required by Slater's is not very restrictive. We now discuss some cases, in which the condition is easily analyzed.

Linear Optimization Problems

In the case of linear programs in standard form

$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{2.2.15}$$

note that the inequality constraints are of the form $g_i(x) \leq 0$ with $g_i(x) = -x_i$. To formulate the dual problem consider the Lagrangian given by

$$L(x, \lambda, \nu) = c^\top x - \lambda^\top x + \nu^\top (Ax - b) = x^\top (c - \lambda + A^\top \nu) - b^\top \nu\,.$$

As the Lagrange dual function $g_L(\lambda, \nu)$ is obtained by minimizing over x we see that a finite value for this is obtained if and only if $c + \lambda + A^\top \nu = 0$. Otherwise we can make the expression as negative as we wish by choosing appropriate x. Thus

$$g_L(\lambda, \nu) = \begin{cases} -b^\top \nu & \text{if } c - \lambda + A^\top \nu = 0 \\ -\infty & \text{else}\,. \end{cases}$$

If we make the constraints explicit and using the constraint $\lambda \geq 0$ we arrive at the following form of the *dual problem of a linear program*

$$\begin{aligned} \text{maximize} \quad & -b^\top \nu \\ \text{subject to} \quad & A^\top \nu \geq -c\,. \end{aligned} \tag{2.2.16}$$

It is a useful exercise to check by considering the dual problem of this we arrive again at the primal problem.

To study question of strong duality we can refer to the weaker form of the Slater conditions discussed in Remark 2.2.4. As all constraints in the primal problem are affine, the weaker condition is automatically satisfied, if there exists a feasible problem for the primal problem. So if the primal problem is feasible, then strong duality holds. On the other hand the dual problem is again a linear program and all inequality constraints are affine. So if there is a feasible point for the dual problem, Slater's condition is applicable and again strong duality holds. Hence for linear programs the only possibility for strong duality not to hold is that both primal and dual problem are infeasible.

2.3 The KKT conditions

The Karush-Kuhn-Tucker conditions, commonly abbreviated as KKT-conditions, give conditions for optimality that become

In this section we assume that the functions describing the problem are differentiable. Thus $f, g_1, \ldots, g_m, h_1, \ldots, h_p$ are assumed to be $\mathcal{C}^1$. In particular, we assume that the domain of definition of all these functions is open.

For the moment consider the general case, where the functions need not be convex. Assume that we have a primal point x^* and a dual point (λ^*, ν^*) with zero duality gap. By Lemma 2.2.5 x^* minimizes the Lagrangian $L(x, \lambda^*, \nu^*)$ and so the gradient of the Lagrangian with respect to x has to vanish in x^*. We thus obtain as a necessary condition that

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(x^*) = 0\,.$$

Summarizing the conditions that we have so far, we arrive at the KKT-conditions

$$\begin{aligned}
g_j(x^*) &\leq 0, & j &= 1, \ldots, m && (2.3.1)\\
h_j(x^*) &= 0 & j &= 1, \ldots, p && (2.3.2)\\
\lambda_j^* &\geq 0 & j &= 1, \ldots, m && (2.3.3)\\
\lambda_j^* g_j(x^*) &= 0 & j &= 1, \ldots, m && (2.3.4)\\
\nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(x^*) &= 0\,. && && (2.3.5)
\end{aligned}$$

The first two of these conditions state the inequality and equality constraints that an optimal point x^* has to satisfy, see (2.2.1). The third is the inequality constraint of the Lagrangian dual problem (2.2.6). The fourth condition captures the complementary slackness condition, while the final condition is the usual gradient condition for the Lagrangian dual problem.

Note that so far we have not used convexity. For any optimization problem where the data of the problem are differentiable and strong duality holds, the optimal points must satisfy the Karush-Kuhn-Tucker conditions.

If we now turn to convex problems then the KKT conditions are also sufficient for primal and dual optimality.

Theorem 2.3.1 (KKT Conditions for Convex Optimization Problems) *Consider the convex optimization problem* (2.2.7) *and the dual problem* (2.2.6)*. Assume that the cost*

function f and the inequality constraints g_j are all continuously differentiable. The points x^*, and (λ^*, ν^*) are primal resp. dual optimal with zero duality gap, if and only if

$$\begin{aligned} g_j(x^*) &\leq 0, & j &= 1, \dots, m \\ (Ax^* - b)_j &= 0 & j &= 1, \dots, p \\ \lambda_j^* &\geq 0 & j &= 1, \dots, m \\ \lambda_j^* g_j(x^*) &= 0 & j &= 1, \dots, m \\ \nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) + A^\top \nu^* &= 0\,. \end{aligned} \tag{2.3.6}$$

Proof. We have discussed necessity for general differentiable problems, so it only remains to show that the conditions in (2.3.6) are indeed sufficient. The first two conditions state that x^* is a feasible point for the primal problem. As $\lambda^* \geq 0$ it follows that

$$L(x, \lambda^*, \nu^*) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) + \sum_{j=1}^{p} \nu_j (Ax - b)_j$$

is convex as a function of x. The last condition states that the gradient of L with respect to x vanishes in x^*. For convex functions this is a necessary and sufficient condition for attaining the minimum, so that x^* minimizes $L(x, \lambda^*, \nu^*)$. This implies that

$$\begin{aligned} g_L(\lambda^*, \nu^*) &= L(x^*, \lambda^*, \nu^*) \\ &= f(x^*) + \sum_{j=1}^{m} \lambda_j^* g_j(x^*) + \sum_{j=1}^{p} \nu_j^* (Ax^* - b)_j \\ &= f(x^*)\,. \end{aligned}$$

In the last step we used feasibility of x^* so that $Ax^* = b$ as well as the complementarity condition $\lambda_j^* g_j(x^*) = 0$.

In total we see that $g_L(\lambda^*, \nu^*) = f(x^*)$. This shows that both x^* and (λ^*, ν^*) are optimal for the primal, resp. dual problem and that the duality gap is equal to 0. This concludes the proof. ∎

Remark 2.3.2 Note that in the previous proof we have only used differentiability of the objective function and the constraint functions in order to be able to express a necessary and sufficient condition for optimality of the point x^*. For convex functions such a condition is also given by the subgradient even in the non-differentiable case by Theorem 1.8.2. Thus the assumption of differentiability may be dropped and we still we obtain necessary and sufficient conditions for optimality.

Considering this we have also proved the following statement: If we consider the convex optimization problem (2.2.7) and the dual problem (2.2.6), the points x^*, and (λ^*, ν^*) are

primal resp. dual optimal with zero duality gap, if and only if

$$\begin{aligned} g_j(x^*) &\leq 0, & j &= 1,\dots,m \\ (Ax-b)_j &= 0 & j &= 1,\dots,p \\ \lambda_j^* &\geq 0 & j &= 1,\dots,m \\ \lambda_j^* g_j(x^*) &= 0 & j &= 1,\dots,m \\ \partial f(x^*) + \sum_{j=1}^m \lambda_j^* \partial g_j(x^*) + \sum_{j=1}^p \nu_j^* (Ax)_j &\ni 0\,. \end{aligned} \tag{2.3.7}$$

Note that in the last expression we are using the rules for subgradient calculus; in particular, Proposition 1.7.7 and Theorem 1.7.9.

Example 2.3.3 Consider the situation where there are n communication channels. To each of these some power $x_i \geq 0$ can be allocated subject to a bound on the total available power, so that we have to respect

$$\sum_{i=1}^n x_i \leq 1\,.$$

By allocating power x_i to channel i we can create a capacity $\log(x_i + \alpha_i)$, where $\alpha_i > 0$ are some given thresholds.

The problem is then to maximize the total communication capacity given the constraints, so that we arrive formally at

$$\begin{aligned} \text{minimize} \quad & -\sum_{i=1}^n \log(x_i - \alpha_i) \\ \text{subject to} \quad & x \geq 0 \\ & \sum_{i=1}^n x_i = 1\,. \end{aligned} \tag{2.3.8}$$

Note that have opted for the equality constraints in the last condition since it is clear, that if $\sum_{i=1}^n x_i < 1$, we can do better by allocating further power to any channel.

The Lagrange multipliers for the inequality constraint is given by $\lambda \in \mathbb{R}^n$, while the single equality constraint gives rise to a Lagrange multiplier $\nu \in \mathbb{R}$. The KKT conditions are now

$$\begin{aligned} & x^* \geq 0\,, \qquad \sum_{i=1}^n x_i^* = 1\,, \qquad \lambda^* \geq 0\,, \qquad \lambda_i^* x_i^* = 0\,, \quad i = 1,\dots,n \\ & -\frac{1}{\alpha_i + x_i^*} - \lambda_i^* + \nu^* = 0\,, \quad i = 1,\dots,n\,. \end{aligned}$$

The conditions can be solved explicitly to obtain the optimal x^*. First note that we can eliminate λ^* to obtain

$$\begin{aligned} & x^* \geq 0\,, \qquad \sum_{i=1}^n x_i = 1\,, \qquad (\nu^* - 1/(\alpha_i + x_i^*))x_i^* = 0\,, \quad i = 1,\dots,n \\ & \nu^* \geq \frac{1}{\alpha_i + x_i^*}\,, \quad i = 1,\dots,n\,. \end{aligned}$$

We now derive conditions in terms of the real variable ν^* and the given data α_i. If $\nu^* < 1/\alpha_i$ then the last condition implies that $x_i > 0$, so that the complementary slackness condition implies $\nu^* = 1/(\alpha_i + x_i^*)$ or equivalently $x_i^* = 1/\nu^* - \alpha_i$. On the other hand if $\nu^* \geq 1/\alpha_i$ then as $x_i \geq 0$ we have

$$\nu^* \geq \frac{1}{\alpha_i} \geq \frac{1}{\alpha_i + x_i} .$$

If $x_i > 0$ the inequality becomes strict, but this violates the complementary slackness condition, because in this case both factors x_i and $(\nu^* - 1/(\alpha_i + x_i^*)$ are positive. In total we obtain

$$x_i^* = \max\{0, 1/\nu^* - \alpha_i\} . \tag{2.3.9}$$

Adding the assumption that the x_i^* should sum to 1 we obtain the condition

$$\sum_{i=1}^{n} \max\{0, 1/\nu^* - \alpha_i\} = 1 .$$

It is now easy to solve this equation to obtain the optimal value for ν^* from which we derive the optimal value point x^* via (2.3.9).

As we can see the power is first allocated to those channels for which α_i is small, as in this case $1/\nu^* - \alpha_i$ is positive for smaller values of ν^*. An iterative procedure for solving the problem would be to order the channels so that $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n$. We would the partition power equally to the first k channels which all have the same value α_1 until we reach the level of $\alpha_k + 1$, if that is possible. From then on we distribute power equally between channels which are at that point all at the level α_{k+1} until we reach the next level, where a new channel comes into play. This is repeated until the available power has been fully used.

The method is called water-filling because it resembles the filling of a hole of different depth with a given amount of water.

Chapter 3

Numerical Methods

In this section we give a brief overview of numerical methods that are suited to convex optimization problems. We begin by discussing optimization problems without constraints, as some of the basic ideas are most easily presented in this framework. Modifications of the methods presented there may be used to tackle problems with constraints. Finally, we will give an introduction to a successful approach using interior point methods.

3.1 Unconstrained Problems

We now consider convex optimization problems of the form

$$\text{minimize} \quad f(x) \tag{3.1.1}$$

with the assumption that f is twice continuously differentiable and $\operatorname{dom} f \subset \mathbb{R}^n$ is an open set. Unless we are able to compute minimal points of f directly, we are interested in *iterative methods*. That is, algorithms which, given an initial condition x^0, produce a sequence x^k with the property that $x^k \to p^*$ as $k \to \infty$. Of course, in practice we will not be able to actually compute the whole sequence but for k sufficiently large the approximation x^k will very often be good enough.

We will discuss the methods under the following assumptions.

Assumption 3.1.1 *(i) f is twice continuously differentiable and $\operatorname{dom} f \subset \mathbb{R}^n$ is an open set.*

(ii) The sublevel set S corresponding to the initial condition x^0 of the procedure satisfies

$$S := \{x \in \operatorname{dom} f\,;\ f(x) \le f(x^0)\}$$

is closed.

(iii) f is strongly convex *on S, that is, we assume there exists a constant $m > 0$ such that*

$$Hf(x) \ge mI\,, \qquad \text{for all } x \in S\,.$$

Note that the second assumption is automatically satisfied if $\operatorname{dom} f = \mathbb{R}^n$ by continuity of f, or if f has the property that $f(x) \to \infty$ as $x \to \partial \operatorname{dom} f$. Furthermore, note that strong convexity implies strict convexity by Proposition 1.6.3 (ii) but the converse is false as shown by the example $f : x \mapsto x^4$.

3.1.1 Descent Methods

The basic idea of the methods discussed in this section is to derive an algorithm that for each point $x \in S$ defines a direction of descent Δx and a step length $h > 0$ sufficiently large and to consider the iteration

$$x^{k+1} = x^k + h\Delta x^k \,.$$

This approach will be called a descent method, if $f(x^{k+1}) < f(x^k)$ away from the minimum. Using the characterization of minimal points from Proposition 1.8.3 we see that a descent is only possible, if we have

$$\langle \nabla f(x), \Delta x \rangle < 0 \,.$$

Any direction Δx satisfying this equation will be called a *descent direction*. Good choices of descent directions will be discussed in the following. Given a rule for choosing descent directions the generic descent algorithm then has the form

Algorithm 3.1.2
Input *Initial point* $x \in \operatorname{dom} f$.
Repeat
1. *Find descent direction* Δx.
2. *Line search. Find a step length* $h > 0$.
3. *Set* $x := x + h\Delta x$.

Until *stopping criterion is satisfied.*

One of the problems in implementing this algorithms is the determination of the step length h. Once the descent direction has been fixed choosing h reduces to the one-dimensional optimization problem

$$\begin{aligned} \text{minimize} \quad & f(x + h\Delta x) \\ \text{subject to} \quad & h > 0 \,. \end{aligned}$$

In some cases it is possible to solve this minimization problem analytically, or the numerical associated with this problem is low, so that we can use an algorithm for solving this problem exactly (down to machine precision). In this case we speak of *exact line search*.

If this is not possible then *backtracking line search* is one of the many methods proposed for solving convex minimization problems in one dimension. Backtracking is characterized by two constants $\alpha \in (0, 1/2), \beta \in (0, 1)$. It starts with a guess for the step length as $h = 1$ and the iteratively reduces the step length by multiplication with β. The parameter α is used to generate a comparison vector.

Algorithm 3.1.3
Input $x \in \operatorname{dom} f$, *a descent direction* Δx, $\alpha \in (0, 1/2), \beta \in (0, 1)$.
$h := 1$
While $f(t + h\Delta x) > f(x) + \alpha h \langle \nabla f(x), \Delta x \rangle$
Do $h := \beta h$

Note that as Δx is a descent direction, we have $\langle \nabla f(x), \Delta x \rangle < 0$. So the algorithm guarantees that $f(x + h\Delta x) < f(x)$. Also termination is guaranteed, as for h sufficiently small we have by differentiability that

$$f(t + h\Delta x) \approx f(x) + h\langle \nabla f(x) < f(x) + \alpha h \langle \nabla f(x) \,.$$

3.1.2 Steepest Descent

For the sake of speed it is a natural idea to choose as descent direction the direction in which f is decreasing the fastest. That is we would like a direction in which we have a good chance of reducing the value of f. In $\mathbb{R}^n$ there are many ways of measuring what it means to reduce the function f fast. Given an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^n$ a direction of *normalized steepest descent* with respect to that norm is defined as an element of the set

$$\Delta x_{nsd} = \operatorname{argmin}\{\ \langle \nabla f(x), v\rangle\ ;\ \|v\| = 1\}\ . \tag{3.1.2}$$

Note that the minimizing argument argmin is not necessarily unique, so that there may be several directions of steepest descent. Recalling our discussion of dual norms we see that the direction of normalized steepest descent are just the vectors dual to $\nabla f(x)$ multiplied by -1. In particular, when we compare to the definition of the dual norm in (1.8.5), we see that

$$\langle \nabla f(x), \Delta x_{nsd}\rangle = -\|\nabla f(x)\|^*\,.$$

It is therefore common to use a particular renormalization of the normalized steepest descent direction by defining

$$\Delta x_{sd} := \|\nabla f(x)\|^* \Delta x_{nsd}\,.$$

This corresponds to choosing $\Delta x = y$ where y defines a dual pair $(-y, \nabla f(x))$ such that

$$\langle \nabla f(x), y\rangle = -\|y\|\|\nabla f(x)\|^* = -\left(\|\nabla f(x)\|^*\right)^2\ .$$

The revised version of the algorithm is then of the following form.

Algorithm 3.1.4
Input *Initial point* $x \in \operatorname{dom} f$.
Repeat
1. *Compute* Δx_{sd} *as steepest descent direction.*
2. *Line search. Find a step length* $h > 0$ *using exact line search or backtracking.*
3. *Set* $x := x + h\Delta x_{sd}$.

Until *stopping criterion is satisfied.*

In particular, for the Euclidean norm $\|\cdot\|_2$ the dual vector of each nonzero vector x is just the normalized vector $x/\|x\|_2$, as we have

$$\nabla\|x\|_2 = \frac{x}{\|x\|}\,.$$

Thus if we apply steepest descent with respect to the Euclidean norm we arrive at the method called *gradient descent*, where the descent direction is $\Delta x = -\nabla f(x)$, see also Figure 3.1.

The algorithm is then

Algorithm 3.1.5
Input *Initial point* $x \in \operatorname{dom} f$.
Repeat
1. $\Delta x := -\nabla f(x)$.
2. *Line search. Find a step length* $h > 0$ *using exact line search or backtracking.*

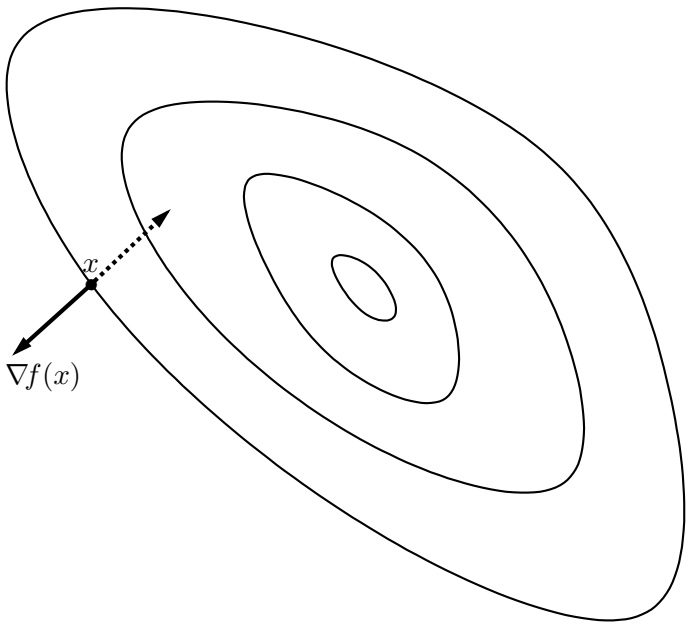

Figure 3.1: Gradient descent

3. Set $x := x + h\Delta x$.

Until *stopping criterion is satisfied.*

For this algorithm the following convergence results may be derived.

Proposition 3.1.6 *Let* f *satisfy Assumption 3.1.1 and assume* $0 < m < M$ *are such that*

$$mI \leq Hf(x) \leq MI\,, \qquad \text{for all } x \in S\,.$$

Then we have

(i) for gradient descent with exact line search that for all $k \geq 0$

$$f(x^k) - p^* \leq \Big(1 - \frac{m}{M}\Big)^k \ \big(f(x^0) - p^*\big)\ .$$

In particular, $f(x^k) - p^* < \varepsilon$ *after at most*

$$\frac{\log((f(x^0) - p^*)/\varepsilon)}{\log(M) - \log(M - m)}$$

steps.

(ii) for gradient descent with backtracking line search that for all $k \geq 0$

$$f(x^k) - p^* \leq c^k \ \big(f(x^0) - p^*\big)$$

with $c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\}$.

In both cases we note that the convergence obtained by the method is linear. Of course, in the case of backtracking line search an estimate for the number of iterations needed to converge down to the error level ε can be derived similarly to the case of exact line search.

If we choose an elliptic norm $\|x\|_P := \left(x^\top Px\right)^{1/2} = \|P^{1/2}x\|_2$ for some positive definite matrix P, we have seen that the dual vectors to x are the scalar multiples of $y = Px$, see p. 42. For the choice of the descent direction this has to be interpreted in the following way: the gradient $\nabla f(x)$ is the vector orthogonal to the level set of f at x. It thus plays the role of the dual vector, i.e. $\nabla f(x) = Px$. The descent direction that leads us down is then obtained by multiplying with P^{-1}.

The algorithm for steepest descent with respect to another elliptic norm is then similar to Algorithm 3.1.5. We only have to replace the choice $\Delta x = -\nabla f(x)$ by

$$\Delta x_{sd} = -P^{-1}\nabla f(x)\,. \tag{3.1.3}$$

Note that this gives the required scaling as we have

$$\langle \nabla f(x), \Delta x_{sd}\rangle = -\langle \nabla f(x), P^{-1}\nabla f(x)\rangle = -\left(\|\nabla f(x)\|_{P^{-1}}\right)^2\,.$$

We note that the choice of a different elliptic norm $\|P^{1/2}x\|_2$ may be interpreted as performing the change of coordinates $\bar{x} = P^{1/2}x$. In this case $\|\bar{x}\|_2 = \|x\|_P$. If we perform this change of variables then we want to minimize the function

$$\tilde{f}(\bar{x}) = f(P^{-1/2}\bar{x}) = f(x)\,.$$

To differentiate the gradient with respect to $\bar{x}$ and x we use the notation $\nabla_{\bar{x}}$, resp. ∇_x. If we apply gradient descent to $\tilde{f}$, then we obtain

$$\nabla_{\bar{x}}\tilde{f}(\bar{x}) = \nabla_{\bar{x}} f(P^{-1/2}\bar{x}) = P^{-1/2}\nabla_x f(x)\,.$$

Thus the gradient descent direction with respect to this new variable corresponds to the direction (in the original variables)

$$\Delta x = P^{-1/2}\left(-P^{-1/2}\nabla_x f(x)\right) = -P^{-1}\nabla_x f(x)\,.$$

Note that this change of variables can have a significant effect on the constant m, M that appear in Proposition 3.1.6. If f is approximated well by its quadratic approximation $\hat{f}(x) = f(x^*) + \langle \nabla f(x^*), x\rangle + \frac{1}{2}x^\top Hf(x^*)x$, then a change of coordinates using the Hessian of f in x^* would make the relative size of m, M similar. In this case the convergence estimates given by Proposition 3.1.6 are much tighter. Of course, we do not know the optimal point x^*, but a very effective method may be interpreted in this way.

3.1.3 Newton's Method

Newton's method provides a very powerful method because locally near the optimal point the convergence is quadratic, or in other words very fast. The method may be interpreted in the sense that in every point we choose the descent direction, that is given by the steepest descent direction with respect to the elliptic norm given by the Hessian of the current point. At $x \in S$ we thus choose the descent direction

$$\Delta x_N := -H(f)(x)^{-1}\nabla f(x)$$

in correspondence to the choice in (3.1.3).

As $\nabla f(x^k) \to 0$ as $x^k \to x^*$ it is common to use the value of $\|\nabla f(x)\|^2_{H(f)^{-1}(x)} = \nabla f(x)^\top H(f)(x)^{-1}\nabla f(x)$ as a criterion for stopping the algorithm. The algorithm is then as follows.

Algorithm 3.1.7
Input *Initial point $x \in \operatorname{dom} f$ and error bound $\varepsilon > 0$.*
Repeat

1. *Compute the Newton step*
$$\Delta x_N := -H(f)(x)^{-1}\nabla f(x)\,, \quad \lambda^2 = \nabla f(x)^\top H(f)(x)^{-1}\nabla f(x)\,.$$
2. *Stopping criterion.* **Quit** *if $\lambda^2 \leq \varepsilon$.*
3. *Line search. Find a step length $h > 0$ using backtracking.*
4. *Set $x := x + h\Delta x_N$.*

The convergence of the Newton algorithm may be characterized by two phases. Close to the optimal point x^* the Hessian $Hf(x)$ yields a good quadratic approximation of the cost function f and convergence is quadratic. In this regime we speak of the quadratically convergent phase of the Newton algorithm. Further away from the optimal point convergence only linear, as also the Hessian provides only local information and there is no reason to assume that this should result in good descent directions globally. So initially, the Newton algorithm is in the damped Newton phase, in which the approach to the equilibrium is only linear. The following result may be shown.

Theorem 3.1.8 (Convergence of the Newton Algorithm) *Let f satisfy Assumption 3.1.1 and assume $0 < m < M$ are such that*
$$mI \leq Hf(x) \leq MI\,, \quad \text{for all } x \in S\,.$$
Assume furthermore that $Hf(x)$ is Lipschitz continuous with Lipschitz constant L on S. Then there exist constants $0 < \eta < m^2/L$ and $\gamma > 0$ such that

(i) If $\|\nabla f(x^k)\| \geq \eta$ then
$$f(x^{k+1}) - f(x^k) < \gamma\,,$$
(ii) If $\|\nabla f(x^k)\| < \eta$ then the backtracking line search selects $h = 1$ and
$$\frac{L}{2m^2}\|\nabla f(x^{k+1})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^k)\|_2\right)^2\,.$$

Note that the previous estimates for the speed of convergence involve the constants m and L, which might very well be hard to estimate in specific estimations. To overcome problems of this type Nesterov and Nemirovskii introduced the notion of self-concordant functions. A convex function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is called self-concordant, if
$$|f'''(x)| \leq 2f''(x)^{3/2}\,, \quad \text{for all } x \in \operatorname{dom} f\,.$$
A function f defined on $\mathbb{R}^n$ is called self-concordant, if it is self-concordant on every line segment contained in $\operatorname{dom} f$. That is, the one dimensional function obtained by restricting to a line segment should be self-concordant.

It is beyond the scope of these notes to discuss self-concordant functions. But at this point is has to be mentioned that self-concordant functions define a special class of functions for which a complete convergence analysis of Newton's method is possible that does not introduce unknown constants.

3.2 Constrained Problems

We now turn to the discussion of convex optimization problems that have equality constraints. That is we consider problems of the form

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b\,. \end{aligned} \tag{3.2.1}$$

We assume that f is convex and twice continuously differentiable and defined on an open subset $\operatorname{dom} f \subset \mathbb{R}^n$. Also $A \in \mathbb{R}^{p\times n}$ with $\operatorname{rank} A = p < n$. We assume an optimal point x^* exists and as before let $p^* = f(x^*)$. By Remark 2.2.4 the assumptions guarantee that strong duality holds and we can appeal to the KKT conditions to see that $x^* \in \operatorname{dom} f$ is optimal if and only if there exists a $\nu^* \in \mathbb{R}^p$ such that

$$Ax^* = b\,, \quad \text{and} \quad \nabla f(x^*) + A^\top \nu^* = 0\,. \tag{3.2.2}$$

Thus solving the optimization problem (3.2.1) is equivalent to solving the KKT system (3.2.2). We will concentrate on an extension of Newton's method to the solution of constrained problems of the type just described. Other approaches would for example consist of elimination of the equality constraints to rederive an unconstrained problem or by duality methods.

In the interpretation of Newton's method we have seen that the method provides exact solutions to quadratic problems. We follow this reasoning and begin by studying quadratic constrained optimization problems. In a first step we thus consider the problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}x^\top Px + c^\top x + d \\ \text{subject to} \quad & Ax = b\,. \end{aligned} \tag{3.2.3}$$

Here $P \in \mathcal{H}_n$ is assumed to be positive semidefinite, $c \in \mathbb{R}^n$, $d \in \mathbb{R}$. The optimality conditions (3.2.2) become

$$Ax^* = b\,, \quad \text{and } Px^* + c + A^\top \nu^* = 0\,, \tag{3.2.4}$$

or equivalently

$$\begin{bmatrix} P & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix} \tag{3.2.5}$$

This system of linear equations in $n+p$ variables is called the KKT system for the equality constrained system (3.2.3). The coefficient matrix is called the KKT matrix.

If the KKT matrix is invertible, then there exists a unique primal-dual solution (x^*, ν^*). Otherwise there may an affine subspace of solutions, all of which define optimal pairs. If no solution exists, the quadratic optimization problem is unbounded from below or possibly infeasible.

We now return to the constrained problem (3.2.1) and want to describe a Newton like descent direction for this problem. We assume that we have a feasible initial condition x^0, i.e. $Ax^0 = b$. We then would like of course that feasibility is retained for the next step, i.e. that the descent direction satisfies $A\Delta x = 0$.

We now replace the objective function f in (3.2.1) by its second order Taylor approximation in $\bar{x}$ to obtain the approximate problem

$$\begin{aligned} \text{minimize} \quad & f(\bar{x}) + \nabla f(\bar{x})^\top x + \frac{1}{2} x^\top H f(\bar{x}) x \\ \text{subject to} \quad & A(\bar{x} + x) = b\,. \end{aligned} \tag{3.2.6}$$

This is a problem of the type we have solved in (3.2.3). Assuming the corresponding KKT-matrix is nonsingular, we define the Newton step Δx_{Nc} at a feasible point x as the solution of

$$\begin{bmatrix} Hf(x) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{Nc} \\ \nu^* \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix} \tag{3.2.7}$$

If the KKT matrix is singular the Newton step is not defined. Note that if the problem is quadratic the Newton step solves the problem exactly in one step. This is similar to the unconstrained case where the same phenomenon could be observed. Also the conditions in (3.2.7) imply in particular, that $A\Delta x_{Nc} = 0$, so that the Newton step preserves feasibility.

The convergence analysis of the constrained Newton method is similar to the unconstrained case: there is an initial damped phase in which convergence is slow and once we are sufficiently close to the optimum, the quadratic approximation will pay of and result in quadratic convergence of the method.

3.3 Interior Point Methods

We now discuss numerical methods for the solution of convex optimization problems of the general form

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_j(x) \leq 0 \qquad\qquad j = 1, \ldots, m \\ & Ax = b\,. \end{aligned} \tag{3.3.1}$$

We assume that the cost function f and the inequality constraints g_j are twice continuously differentiable and also for $A \in \mathbb{R}^{p \times n}$ has rank $A = p$. Furthermore, we assume that Slater's conditions are satisfied, so that there exist strictly feasible points for the problem. In particular an optimal point x^* exists and strong duality holds. Also there exists an optimal point (λ^*, ν^*) satisfying the KKT conditions

$$\begin{aligned} g_j(x^*) &\leq 0, && j = 1, \ldots, m \\ Ax^* - b &= 0 \\ \lambda_j^* &\geq 0 && j = 1, \ldots, m \\ \lambda_j^* g_j(x^*) &= 0 && j = 1, \ldots, m \\ \nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) + A^\top \nu^* &= 0\,. \end{aligned} \tag{3.3.2}$$

In this text we follow [4] and concentrate on the barrier method as a particular interior

point method. To this end we first notice that (3.3.1) may be reformulated equivalently as

$$\begin{aligned} \text{minimize} \quad & f(x) + \sum_{j=1}^{m} I_-(g_j(x)) \\ \text{subject to} \quad & Ax = b\,, \end{aligned}$$

where I_- is the indicator function of the nonpositive reals defined by

$$I_-(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \infty & \text{if } x > 0\,. \end{cases}$$

This new formulation has no inequality constraints, and it is differentiable on the interior of the domain of the new objective function. But the differential of 0 provides no information and at the boundary of the domain differentiability is a problem. In the next step we approximate I_- by a differentiable function only defined on the set of strictly feasible points. This smooth approximation is given by

$$\hat{I}_t(x) = -\frac{1}{t}\log(-x)\,, \quad x \in (-\infty, 0)\,.$$

Here $t > 0$ is a parameter that control the difference between the approximation $\hat{I}_t(x)$ and $I_-(x)$. Versions of $\hat{I}_t$ for a few choices of t are shown in Figure 3.2.

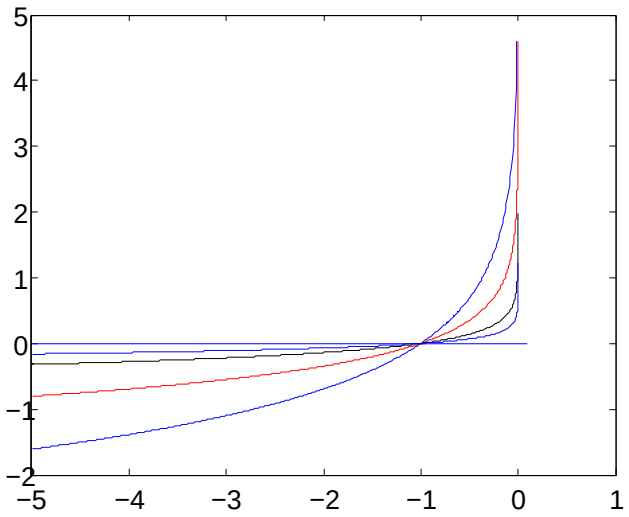

Figure 3.2: Plots of $\hat{I}_t(x)$

With this approximation we arrive at the family of optimization problems

$$\begin{aligned} \text{minimize} \quad & f(x) + \sum_{j=1}^{m} \hat{I}_t(g_j(x)) \\ \text{subject to} \quad & Ax = b\,, \end{aligned} \tag{3.3.3}$$

or equivalently

$$\begin{aligned} \text{minimize} \quad & f(x) + \sum_{j=1}^{m} -\frac{1}{t}\log(-g_j(x)) \\ \text{subject to} \quad & Ax = b\,. \end{aligned} \tag{3.3.4}$$

This defines a differentiable convex optimization problem as $\hat{I}_t(x)$ is convex, increasing in x and differentiable. Under appropriate conditions we can apply Newton's method to the solution of this equality constrained optimization problem. The function

$$\phi(x) := -\sum_{j=1}^{m} \log(-g_j(x))$$

is called the *logarithmic barrier function.* The domain of this function is the set of strictly feasible points. The minimum for the approximative problem will lie within the set of strictly feasible points as $\phi(x) \to \infty$ as x approaches an active inequality constraint, i.e. as $g_j(x) \to 0$. The function thus imposes a barrier for any optimization algorithm to stay away from the boundary of the set of strictly feasible points. As t increases, this barrier is relaxed, but it is present for any $t > 0$.

The idea of interior point methods is now to solve the approximation problems and increase t in the course of the algorithm to obtain better approximations of the original problem. In this context the concept of a central path is important. We consider the equivalent problem to (3.3.4) obtained by multiplying the objective function by t. We thus consider

$$\begin{aligned} \text{minimize} \quad & tf(x) + \phi(x) \\ \text{subject to} \quad & Ax = b\,. \end{aligned} \tag{3.3.5}$$

Note that the set of optimal points is not changed by this operation.

Definition 3.3.1 (Central Path) *Consider the family of optimization problems* (3.3.5) *for $t > 0$. Assume that for each $t > 0$ there is a unique optimal solution $x^*(t)$. The* central path *is defined as the set of optimal points for the problem with parameter t, i.e.*

$$\{x^*(t)\,;\; t > 0\}\,.$$

The points $x^*(t)$ are called central points.

The KKT conditions for the central points are of reduced complexity, since there are no Lagrange multipliers λ necessary for formulating the dual problem. Note that also for problem (3.3.5) strong duality holds, as we assumed this in the beginning for the original problem. Appealing to the KKT conditions, the central points are characterized by the necessary and sufficient conditions that they are strictly feasible, i.e.

$$Ax^*(t) = b\,, \quad g_j(x^*(t)) < 0\,, \quad j = 1, \ldots, m$$

and there exists a $\nu^* \in \mathbb{R}^p$ such that

$$\begin{aligned} 0 &= t\nabla f(x^*(t)) + \nabla\phi(x^*(t)) + A^\top \nu^* \\ &= t\nabla f(x^*(t)) + \sum_{j=1}^{m} \frac{1}{-g_j(x^*(t))}\nabla g_j(x^*(t)) + A^\top \nu^*\,. \end{aligned} \tag{3.3.6}$$

From (3.3.6) we obtain information about dual points for the original problem (3.3.1). Indeed every central point gives rise to a feasible point for the dual problem corresponding to (3.3.1) and thus yields a lower bound for the optimal value p^* of (3.3.1). Define

$$\lambda_j^*(t) = -\frac{1}{tg_j(x^*(t))} > 0\,, \tag{3.3.7}$$

where we have used strict feasibility of $x^*(t)$. We obtain that $(\lambda^*(t), \nu^*/t)$ is feasible for the dual problem. Dividing (3.3.6) by t we obtain

$$\nabla f(x^*(t)) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*(t)) + \frac{1}{t} A^\top \nu^* = 0\,.$$

This shows that $x^*(t)$ is a minimizer for the Lagrangian function $L(x, \lambda^*, \nu^*/t)$ and so

$$\begin{aligned} g_L(\lambda^*, \nu^*/t) &= f(x^*(t)) + \sum_{j=1}^{m} \lambda_j^* g_j(x^*(t)) + \frac{1}{t}(\nu^*)^\top (Ax^*(t) - b) \\ &= f(x^*(t)) - \frac{m}{t}\,, \end{aligned}$$

where we have used that $x^*(t)$ is a primal feasible point and the definition of λ^* in (3.3.7).

In particular, the duality gap between $x^*(t)$ and $g_L(\lambda^*, \nu^*/t)$ is m/t, so that we can conclude

$$f(x^*(t)) - p^* \leq \frac{m}{t}\,.$$

This shows that $f(x^*(t)) \to p^*$ as $t \to \infty$.

A further useful interpretation of the characterization of central points obtained in (3.3.6) is in terms of the KKT conditions. Indeed $x^*(t)$ is characterized by the conditions that there exist (λ^*, ν^*) such that

$$\begin{aligned} g_j(x^*(t)) &\leq 0, & j = 1, \ldots, m \\ Ax^*(t) - b &= 0 \\ \lambda_j^* &\geq 0 & j = 1, \ldots, m \\ -\lambda_j^* g_j(x^*) &= \frac{1}{t} & j = 1, \ldots, m \\ \nabla f(x^*(t)) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*(t)) + A^\top \nu^* &= 0\,. \end{aligned} \tag{3.3.8}$$

When comparing this to the KKT conditions in (3.3.2), we see that the central points are approximations of solutions of the KKT conditions and this approximation improves as $t \to \infty$.

The barrier methods now consists of solving for $x^*(t)$ and updating t in a systematic manner. The resulting algorithm consists of an outer iteration that consists of the computation of an optimal point $x^*(t)$ for a particular parameter t. This optimization problem, which is an equality constrained convex optimization problem, is approximately solved using a Newton algorithm. The iteration steps of the Newton algorithm are called the inner iterations.

Algorithm 3.3.2
Input *Strictly feasible point* $x \in \operatorname{dom} f$, $t^0 > 0$, $\mu > 1$ *and error bound* $\varepsilon > 0$.
Repeat

1. *Centering step:*
   Compute $x^*(t)$ *by minimizing* $tf(x) + \phi(x)$ *subject to* $Ax = b$
2. *Update:* $x := x^*(t)$
3. *Stopping criterion:* **Quit** *if* $m/t \leq \varepsilon$.
4. *Increase* t*:* $t := \mu t$.

For the implementation of the algorithm the parameters $\mu > 1$ and the initial t^0 have to be chosen. If μ is close to 1, then the problems obtained for successive t^k, t^{k+1} are close to one another, so that the previous solution $x^*(t^k)$ may be expected to be a good initial guess for the next problem. Thus the inner Newton algorithm may be expected to converge quickly. This is at the expense of many outer approximations as t increases rather slowly. If μ is rather large the situation is reversed and we expect slower convergence in the inner iteration but rapid progress in the growth of t. This tradeoff has to be dealt with based on the particular problem at hand.

Similarly, the choice of t^0 is crucial. If t^0 is already chosen to be quite large, then very likely the resulting optimization problem is numerically unpleasant and the initial Newton algorithm may take considerable time. On the other hand small t^0 may result in many outer steps again reducing performance. A possible choice is suggested by the optimality condition (3.3.6). The quantity

$$\inf_{\nu} \left\| t\nabla f(x^0) + \nabla\phi(x^0) + A^\top \nu \right\|_2$$

may be interpreted as a measure of distance of x^0 to $x^*(t)$. By solving a least squares problem, t^0 can be chosen so as to minimize this number.

Chapter 4

Congestion Control

4.1 Introduction

In the final chapter of these notes we discuss congestion control in the context of convex optimization. Congestion control refers to a the problem of regulating internet traffic in order to avoid overload of the physical network. In the context of these notes a communication network consists of a number of sources and sinks connected together via links and routers.

In TCP/IP the data is partitioned into packets. Each packet has a header containing some data necessary for routing, a checksum and a sequence order. It is the goal of TCP to transmit all data reliably across the network.

Due to network congestion, traffic load, and other physical effects, IP packets can be lost, be delivered in the wrong order or even be duplicated. It is the goal of TCP to detect any problem with package delivery and to retransmit data in case that some packets appear to have been lost. The fundamental reasoning in TCP is that when a problem has been spotted, the likely cause of this is congestion in the network. The response to this is to reduce the window size of the TCP flow.

A TCP session is initiated by establishing the connection. This happens via a three-way handshake in which request for data is sent and also a random number as a starting point for the numbering of packets. The answer consists of an acknowledgment and this is again acknowledged by a third message, so that in the end both sides have received an acknowledgment. In the modeling this phase is usually ignored. We will also ignore the *slow start* phase of TCP in which initially TCP quite aggressively seeks to gain bandwidth.

Our modeling approach will focus on the *congestion avoidance phase* of TCP, which is characterized by an *Additive-Increase Multiplicative-Decrease (AIMD)* congestion control algorithm. Also we concentrate on routers that employ drop-tail queueing and *Additive-Increase Multiplicative-Decrease (AIMD)* congestion control algorithms. It is shown that the theory of nonnegative matrices may be employed to model such networks. In particular, important network properties such as: (i) fairness; (ii) rate of convergence; and (iii) throughput; can be characterised by certain non-negative matrices. We demonstrate that these results can be used to develop tools for analysing the behaviour of *AIMD* communication networks.

We assume that the links can be modelled as a constant propagation delay together with a queue, that the queue is operating according to a drop-tail discipline, and that all of the sources are operating a *Additive-Increase Multiplicative Decrease* (AIMD) -like

congestion control algorithm. AIMD congestion control operates a window based congestion control strategy. Each source maintains an internal variable $cwnd_i$ (the window size) which tracks the number of sent unacknowledged packets that can be in transit at any time, i.e. the number of packets in flight. On safe receipt of data packets the destination sends acknowledgement (ACK) packets to inform the source. When the window size is exhausted, the source must wait for an ACK before sending a new packet. Congestion control is achieved by dynamically adapting the window size according to an additive-increase multiplicative-decrease law. Roughly speaking, the idea is for a source to probe the network for spare capacity by increasing the rate at which packets are inserted into the network, and to rapidly decrease the number of packets transmitted through the network when congestion is detected through the loss of data packets.

In more detail, the source increments $cwnd_i(t)$ by a fixed amount α_i upon receipt of each ACK. On detecting packet loss, the variable $cwnd_i(t)$ is reduced in multiplicative fashion to $\beta_i cwnd_i(t)$. We shall see that the *AIMD* paradigm with drop-tail queuing gives rise to networks whose dynamics can be accurately modeled as a positive linear system. While we are ultimately interested in general communication networks, for reasons of exposition it is useful to begin our discussion with a description of networks in which packet drops are synchronised (i.e. every source sees a drop at each congestion event). We show that many of the properties of communication networks that are of interest to network designers can be characterised by properties of a square matrix whose dimension is equal to the number of sources in the network. The approach may be extended to a model of unsynchronised networks. Even though the mathematical details are more involved, many of the qualitative characteristics of synchronised networks carry over to the non-synchronised case if interpreted in a stochastic fashion.

4.1.1 Synchronised communication networks

We begin our discussion by considering communication networks for which the following assumptions are valid: (i) at congestion every source experiences a packet drop; and (ii) each source has the same round-trip-time (RTT)[1]. In this case an exact model of the network dynamics can be derived as follows. Let $w_i(k)$ denote the congestion window size of source i immediately before the k'th network congestion event is detected by the source. Over the k'th congestion epoch three important events can be discerned: $t_a(k), t_b(k)$ and $t_c(k)$; as depicted in Figure 4.1. The time $t_a(k)$ denotes the instant at which the number of unacknowledged packets in flight equals $\beta_i w_i(k)$ where β_i is the multiplicative decrease factor associated with the $i'th$ flow (recall that after each congestion event the i'th sources decreases its number of packets in flight by a factor of $(1-\beta_i)$; $t_b(k)$ is the time at which the bottleneck queue is full; and $t_c(k)$ is the time at which packet drop is detected by the sources, where time is measured in units of RTT[2]. It follows from the definition of the *AIMD* algorithm that the window evolution is completely defined over all time instants by knowledge of the $w_i(k)$ and the event times $t_a(k), t_b(k)$ and $t_c(k)$ of each congestion epoch. We therefore only need to investigate the behaviour of these quantities.

We assume that each source is informed of congestion one RTT after the queue at the

[1] One RTT is the time between sending a packet and receiving the corresponding acknowledgement when there are no packet drops.

[2] Note that measuring time in units of RTT results in a linear rate of increase for each of the congestion window variables between congestion events.

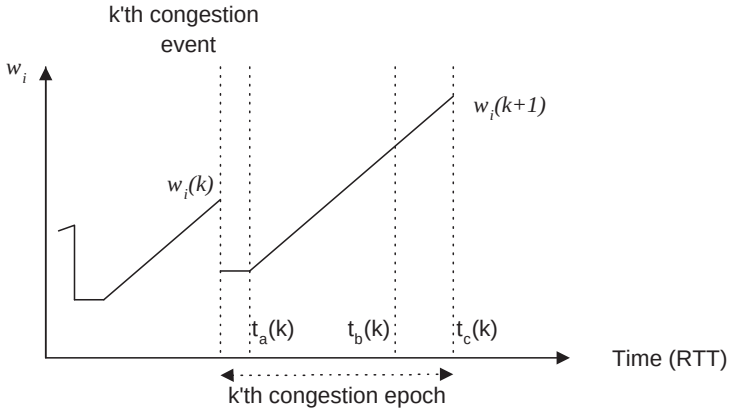

Figure 4.1: Evolution of window size

bottleneck link becomes full; that is $t_c(k) - t_b(k) = 1$. Also,

$$w_i(k) \geq 0, \sum_{i=1}^{n} w_i(k) = P + \sum_{i=1}^{n} \alpha_i, \ \forall k > 0, \tag{4.1.1}$$

where P is the maximum number of packets which can be in transit in the network at any time; P is usually equal to $q_{max} + BT_d$ where q_{max} is the maximum queue length of the congested link, B is the service rate of the congested link in packets per second and T_d is the round-trip time when the queue is empty. At the $(k+1)$th congestion event

$$w_i(k+1) \quad = \quad \beta_i w_i(k) + \alpha_i [t_c(k) - t_a(k)]. \tag{4.1.2}$$

It follows from (4.1.1) and (4.1.2) that

$$t_c(k) - t_a(k) \quad = \quad \frac{1}{\sum_{i=1}^{n} \alpha_i} [P - \sum_{i=1}^{n} \beta_i w_i(k)] + 1. \tag{4.1.3}$$

Hence, it follows that

$$w_i(k+1) = \beta_i w_i(k) + \frac{\alpha_i}{\sum_{j=1}^{n} \alpha_j} [\sum_{j=1}^{n} (1 - \beta_j) w_j(k)], \tag{4.1.4}$$

and that the dynamics an entire network of such sources is given by

$$W(k+1) = AW(k), \tag{4.1.5}$$

where $W^T(k) = [w_1(k), \cdots, w_n(k)]$, and where with $\alpha^T = [\alpha_1 \ldots \alpha_n]$ we have

$$\begin{aligned} A \quad = \quad & \text{diag } (\beta_1, \beta_2, \ldots, \beta_n) \\ & + \frac{1}{\sum_{j=1}^{n} \alpha_j} \alpha \left[\begin{array}{ccc} 1 - \beta_1 & \cdots & 1 - \beta_n \end{array} \right]. \end{aligned} \tag{4.1.6}$$

and the initial condition $W(0)$ is subject to the constraint (4.1.1).

The matrix A is a positive matrix (all the entries are positive real numbers) and it follows that the synchronised network (4.1.5) is a positive linear system [2]. Many results are known for positive matrices and we exploit some of these to characterise the properties of synchronised communication networks. In particular, from the viewpoint of designing communication networks the following properties are important: (i) network fairness; (ii) network convergence and responsiveness; and (iii) network throughput. While there are many interpretations of network fairness, in this paper we concentrate on window fairness. Roughly speaking, window or pipe fairness refers to a steady state situation where n sources operating $AIMD$ algorithms have an equal number of packets P/n in flight at each congestion event; convergence refers to the existence of a unique fixed point to which the network dynamics converge; responsiveness refers to the rate at which the network converges to the fixed point; and throughput efficiency refers to the objective that the network operates at close to the bottleneck-link capacity. These properties can be deduced from the network matrix A.

Theorem 4.1.1 *Let A be defined as in Equation (4.1.6). Then A is a column stochastic matrix with Perron eigenvector*

$$x_p^T = \begin{bmatrix} \frac{\alpha_1}{1-\beta_1} & \cdots & \frac{\alpha_n}{1-\beta_n} \end{bmatrix}$$

The eigenvalue 1 is simple and all other eigenvalues are smaller in magnitude. Further, the network converges to a unique stationary point $W_{ss} = \Theta x_p$, where Θ is a positive constant such that the constraint (4.1.1) is satisfied; $\lim_{k\to\infty} W(k) = W_{ss}$; and the rate of convergence of the network to W_{ss} is bounded by the second largest eigenvalue of A.

Proof. It is straightforward calculation that A is column-stochastic and that x_p as defined above is an eigenvector corresponding to the eigenvalue $\lambda = 1$. As A is a matrix with positive entries, it follows that all other eigenvalues are less than 1 in modulus. This is a consequence of the Perron-Frobenius theorem.

Denote $e = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^\top$. If the initial condition $w(0)$ satisfies the constraint

$$C = \sum_{i=1}^{n} w_i(0) = e^\top w(0)\,,$$

then as A is column stochastic we have

$$e^\top w(1) = e^\top A w(0) = e^\top w(0) = C\,.$$

By iteration we obtain that $e^\top w(k) = C$ for all $k \geq 0$. This in combination with the properties of the eigenvalues implies that $A^k x \to \Theta x_p$, where Θ is chosen such that $e^\top \Theta x_p = C$. ∎

The following results may be deduced from the above.

(i) **Fairness:** Window fairness is achieved when the Perron eigenvector x_p is a scalar multiple of the vector $[1, ..., 1]$; that is, when the ratio $\frac{\alpha_i}{1-\beta_i}$ does not depend on i. Further, since it follows for conventional TCP-flows ($\alpha = 1, \beta = 1/2$) that $\alpha = 2(1-\beta)$, any new protocol operating an $AIMD$ variant that satisfies $\alpha_i = 2(1-\beta_i)$ will be TCP-friendly - i.e. fair with legacy TCP flows.

(ii) **Network responsiveness:** The magnitude of the second largest eigenvalue λ_{n-1} of the matrix A bounds the convergence properties of the entire network. It is shown in

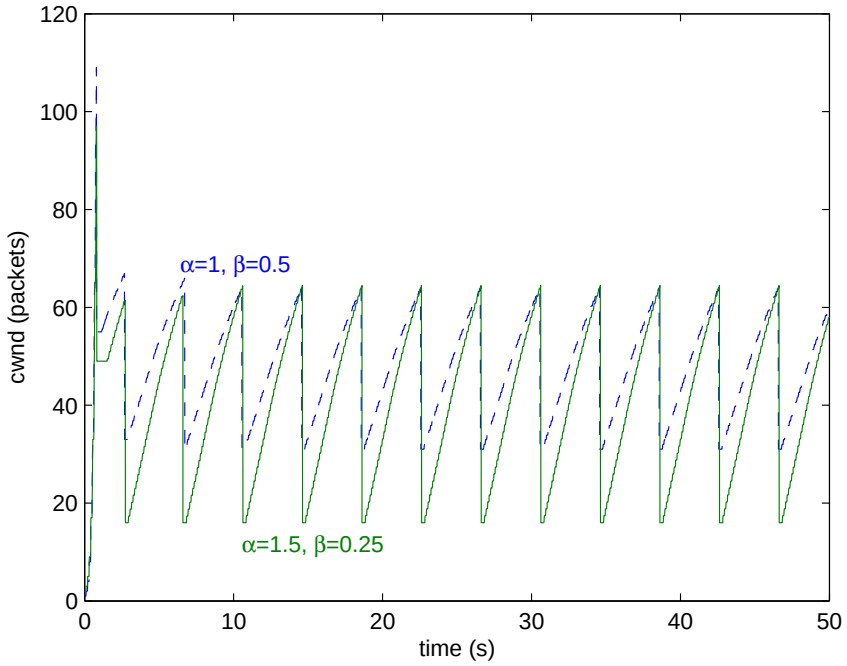

Figure 4.2: Example of window fairness between two TCP sources with different increase and decrease parameters (NS simulation, network parameters: 10Mb bottleneck link, 100ms delay, queue 40 packets.)

[3] that all the eigenvalues of A are real and positive and lie in the interval $[\beta_1, 1]$, where the β_i are ordered as $0 < \beta_1 \leq \beta_2 \leq .... \leq \beta_{n-1} \leq \beta_n < 1$. In particular, the second largest eigenvalue is bounded by $\beta_{n-1} \leq \lambda_{n-1} \leq \beta_n$. Consequently, fast convergence to the equilibrium state (the Perron eigenvector) is guaranteed if the largest backoff factor in the network is small. Further, we show in [3] that the network rise-time when measured in number of congestion epochs is bounded by $n_r = \log(0.95)/\log(\lambda_{n-1})$. In the special case when $\beta_i = 0.5$ for all i, $n_r \approx 4$; see for example Figure 3. Note that n_r gives the number of congestion epochs until the network dynamics have converged to 95 % of the final network state: the actual time to reach this state depends on the duration of the congestion epochs which is ultimately dependent on the α_i.

(iii) **Network throughput :** At a congestion event the network bottleneck is operating at link capacity and the total data throughput through the bottleneck link is given by

$$R(k)^- = \frac{\sum_i^n w_i(k)}{T_d + \frac{q_{max}}{B}} \tag{4.1.7}$$

where B is the link capacity, q_{max} is the bottleneck buffer size, T_d is the round-trip-time when the bottleneck queue is empty and $T_d + q_{max}/B$ is the round-trip time when the queue is full. After backoff, the data throughput is given by

$$R(k)^+ = \frac{\sum_i^n \beta_i w_i(k)}{T_d} \tag{4.1.8}$$

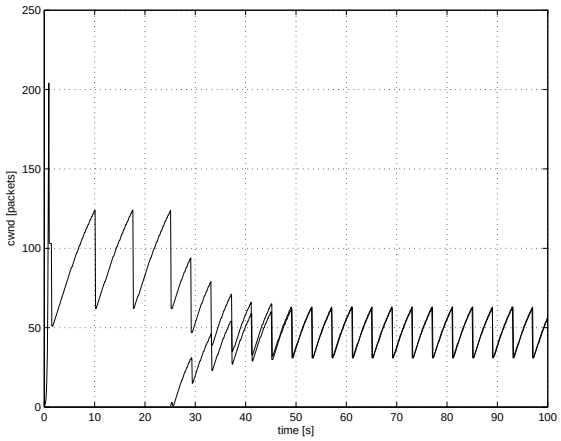

Figure 4.3: NS packet-level simulation ($\alpha_i = 1$, $\beta_i = 0.5$, dumb-bell with 10Mbs bottleneck bandwidth, 100ms propagation delay, 40 packet queue).

under the assumption that the bottleneck buffer empties. It is evident that if the sources backoff too much, data throughput will suffer as the queue remains empty for a period of time and the link operates below its maximum rate. A simple method to ensure maximum throughput is to equate both rates, which may be achieved by the following choice of the β_i:

$$\beta_i = \frac{T_d}{T_d + \frac{q_{max}}{B}} = \frac{RTT_{min}}{RTT_{max}}. \tag{4.1.9}$$

(iv) **Maintaining fairness :** Note that setting $\beta_i = \frac{RTT_{min}}{RTT_{max}}$ requires a corresponding adjustment of α_i if it is not to result in unfairness. Both network fairness and TCP-friendliness are ensured by adjusting α_i according to $\alpha_i = 2(1 - \beta_i)$.

4.2 Utility Based Congestion Control

In internet communication the available bandwidth may be interpreted as a common resource that is used and shared by the users wishing to transmit data across the network.

In resource allocation we consider a utility function that measures the value obtained by a user if a certain amount of the resource is allocated to them.

We consider a network consisting of a set $\mathcal{R}$ of routers or another type of resource. There are n users of the network and each user is using a specific subset of the resources. For the sake of congestion control we can think of the routers that are places along the route of user i. We will identify user i with his subset of used resources, i.e. we will write

$$i \subset \mathcal{R}\,.$$

With this notation the notion that user i makes use of resource $r \in \mathcal{R}$ is denoted by

$$r \in i\,.$$

Note that a resource can be used by several users. The extent to which it is used, or in other words the amount that user i uses of the resources is denoted by x_i. Note that we assume that each user uses the same amount from all resources he is using. This corresponds to the problem of congestion control. The resource might be measured in terms of the data rate, that user i is trying to send across the network. This data rate will then be the same for every router along the way.

The only requirement is that the capacity C_r of resource r is not exceeded. That is, we have the requirement that

$$\sum_{i:r\in i} x_i \leq C_r\,.$$

Definition 4.2.1 (Min-Max Fairness) *An allocation vector $x = (x_1, \ldots, x_n)$ is called* min-max fair*, if*

(i) it satisfies the constraint conditions (4.2.2)*,*

(ii) if y is another vector satisfying (4.2.2) *and for some i we have $y_i > x_i$, then there is an index j such that*

$$y_j < x_j \leq x_i\,.$$

Definition 4.2.2 (Bottlenecks) *Given an allocation vector $x = (x_1, \ldots, x_n)$ a resource r is called a* bottleneck *for user i, if*

(i) full use of the resource, i.e.

$$\sum_{j,r\in j} x_j = C_r\,,$$

(ii) $x_i \geq x_j$ for all users j with $r \in j$.

Lemma 4.2.3 *An allocation vector $x = (x_1, \ldots, x_n)$ is min-max fair if and only if every user has a bottleneck resource.*

Proof. Let $x = (x_1, \ldots, x_n)$ be an allocation such that every user has a bottleneck source and assume (y_i) is a different allocation with $y_i > x_i$. If r is the bottleneck resource for x_i, then

$$y_i + \sum_{j\neq i, j\in r} x_j > \sum_{j,r\in j} x_j = C_r\,.$$

To obtain a feasible allocation it follows that for some $j \in r$ we have $y_j < x_j \leq x_i$. This shows min-max fairness.

Conversely, for a feasible allocation $x = (x_1, \ldots, x_n)$ assume without loss of generality that user 1 does not have a bottleneck router. That is, by definition for all $r \in 1$ we have

$$a_r := \sum_{j,r\in j} x_j < C_r\,,$$

or for some user j_r with $r \in j_r$ we have

$$b_r := x_{j_r} - x_1\,.$$

This implies that

$$c := \min_{r\in 1} \max\{a_r, b_r\} > 0\,.$$

We now can allocate the rate $x_1 + d$ to user 1, then for the routers $r \in 1$ with excess capacity $a_r > 0$, so that the capacity constraint is not violated by this change. For the router were the capacity amount is already met, there is a user j_r with $x_{j_r} \geq x_1 + d$, so the capacity constraint can be met by reducing x_{j_r} to $x_{j_r} - d$. But this shows that x is not min-max fair, as condition (ii) of Definition 4.2.1 does not hold. ∎

An immediate consequence of the previous result is the following observation.

Corollary 4.2.4 *If an allocation vector $x = (x_1, \ldots, x_n)$ is min-max fair then there is a router r such that for all $i \in \{1, \ldots, n\}$ with $r \in i$ we have*

$$x_i = \min\{x_k\,;\ k = 1, \ldots, n\}\,.$$

Proof. Let i be such that

$$x_i = \min\{x_k\,;\ k = 1, \ldots, n\}\,.$$

As the allocation is min-max fair it follows that i has a bottleneck router r. By definition of bottleneck router we have for all j with $r \in j$ that

$$x_i \geq x_j \geq \min\{x_k\,;\ k = 1, \ldots, n\} = x_i\,.$$

This shows that we have equality throughout. This shows the assertion. ∎

We assume from now on that the utility obtained by user i of having x_i bandwidth (or, generally, amount of resource) is given by the utility function U_i. In general, we would like the overall utility to be maximal. The associated optimization problem over all users is then

$$\text{maximize} \quad \sum_{i=1}^{n} U_i(x_i) \tag{4.2.1}$$

$$\text{subject to} \quad \sum_{i:r\in i} x_i \leq C_r \tag{4.2.2}$$

$$x_i \geq 0\,, \quad i = \{1, \ldots, n\}\,.$$

We now assume that the utility functions are strictly concave.

We replace the optimization (4.2.1) by a problem with a barrier function, similar in spirit to the logarithmic barrier function described in Section 3.3. For every $r \in \mathcal{R}$ let $f_r : [0, \infty) \to [0, \infty)$ be a continuous, nod-decreasing function. We assume that

$$\int_0^y f_r(s)ds \to \infty\,, \quad \text{for } y \to \infty\,. \tag{4.2.3}$$

We will use these integrals as a general form of barrier function. To this end we introduce the cost function

$$V(x) = \sum_{i=1}^{n} U_i(x_i) - \sum_{r\in\mathcal{R}} \int_0^{\sum_{i:r\in i} x_i} f_r(s)ds \tag{4.2.4}$$

Lemma 4.2.5 *Assume that for each i the utility function U_i is continuously differentiable, nondecreasing and strictly concave. Assuming* (4.2.3) *the function V defined in* (4.2.4) *is strictly concave.*

Proof. As the sum of the U_i is strictly concave it is sufficient to prove that the integral term convex, as then the subtraction of that term yields a concave contribution.

So consider

$$\sum_{r\in\mathcal{R}} \int_0^{\sum_{i:r\in i} x_i} f_r(s)ds\,.$$

By Lemma 1.6.4 (i) it is sufficient to prove that each summand is convex. So it will be sufficient to see that the function

$$g(x) := \int_0^{\sum_{i=1}^n x_i} f(s)ds$$

is convex. The gradient of this function is given by

$$\nabla g(x) = f\left(\sum_{i=1}^n x_i\right)\begin{bmatrix}1 & \dots & 1\end{bmatrix}^\top\,.$$

Appealing to the criterion of Proposition 1.6.3 (i) we see that

$$g(x) + \langle y-x, \nabla g(x)\rangle = g(x) + f\left(\sum_{i=1}^n x_i\right)\sum_{i=1}^n y_i - x_i = \\ g(x) + \int_{\sum_{i=1}^n x_i}^{\sum_{i=1}^n y_i} f\left(\sum_{i=1}^n x_i\right) ds \le g(y)$$

as f is nondecreasing. This shows convexity of g and this shows convexity of V. ∎

We now add further assumptions that guarantee the existence of unique maxima for the unction V.

Assumption 4.2.6 *For all $i = 1,\dots,n$ it holds that $U_i(x_i) \to -\infty$ as $x_i \to 0$ and $U_i'(x_i) \to 0$ as $x_i \to \infty$.*

We now consider the convex optimization problem

$$\begin{aligned} \text{maximize} \quad & V(x) \\ & x_i \ge 0\,, \quad i = \{1,\dots,n\}\,. \end{aligned} \tag{4.2.5}$$

As V is strictly convex, we know that if a maximal point x^* exists, then it is unique. Assumpotion 4.2.6 guarantees that a maximal point exists. This point is characterized by the condition

$$\nabla V(x^*) = 0 \tag{4.2.6}$$

or equivalently

$$U_i'(x_i^*) - \sum_{r:r\in i} f_r\left(\sum_{j:r\in j} x_j^*\right) = 0\,, \quad i = 1,\dots,n\,. \tag{4.2.7}$$

We now interpret the barrier functions as prices. That is we interpret the value $f_r(y)$ as the price for using resource r at the level y. The assumptions on f then imply that the

price is a continuous in the usage, is non-decreasing as the usage increases. We define the price at router r given a rate by the users j using that router by

$$p_r = f_r(\sum_{j:r\in j} x_j)\,. \tag{4.2.8}$$

Assuming now that the network can provide information to the users of the total price of the routers used along their route. This total price is then given by

$$q_i = \sum_{r\in i} p_r\,.$$

With this notation the optimality condition (4.2.7) can be rewritten as

$$U_i'(x_i^*) - q_i(x^*) = 0\,.$$

We now introduce an algorithm in which users adapt their usage of the resource according to the price information they receive. In continuous time this yields the differential equation

$$\dot{x}_i = k_i(x_i)\left(U_i'(x_i) - q_i(x(t))\right)\,, \tag{4.2.9}$$

where $k_i : [0,\infty) \to (0,\infty)$ is a continuous positive function that regulates the speed of the differential equation.

Using this type of price feedback we obtain asymptotic stability at the maximal point.

Theorem 4.2.7 *Assuming that there is an optimal point x^* for the optimization problem* (4.2.5) *the solution of* (4.2.9) *will converge to x^* for any initial condition $x > 0$.*

Proof. We consider V to define a Lyapunov function for system (4.2.9). In particular, we define the Lyapunov function

$$W(x) = -V(x) + V(x^*)\,.$$

Note that by assumption W has a strict global minimum $W(x^*) = 0$ and that $V(x) \to \infty$ as $x_i \to 0$ for some i. Computing the derivative of V along solutions of (4.2.9) we obtain

$$\begin{aligned}\dot{W}(x) &= \nabla W(x)\left[k_1(x_1)\left(U_1'(x_1) - q_1(x)\right) \quad \dots \quad k_n(x_n)\left(U_n'(x_n) - q_n(x(t))\right)\right]^\top \\ &= \sum_{i=1}^{n} -k_i(x_i)\left(U_i'(x_i) - q_i(x)\right)^2\end{aligned}$$

This expression is negative as long as $x \neq x^*$. This shows the claim. ∎

An alternative view on the congestion control algorithm is obtained if we let routers adjust the prices and the users adjust their rate as a function of the price information obtained. This may be interpreted as a dual version of the congestion control mechanism just discussed. To formulate the problem it is useful to consider the *routing matrix* $R \in \mathbb{R}^{|\mathcal{R}|\times n}$ defined by

$$R_{ir} = \begin{cases} 1 & r \in i \\ 0 & \text{else.} \end{cases}$$

We then have the relations

$$y = Rx$$

where y_r is the total used resource at router r. Also

$$q = R^T p$$

where p is the vector of prices at the routers and q is the vector of total prices for the users. Recall our optimization problem

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n} U_i(x_i) & & \\ \text{subject to} \quad & \sum_{i:r\in i} x_i \le C_r & r \in \mathcal{R} & \qquad (4.2.10) \\ & x_i \ge 0 & i = 1, \ldots, n\,. & \end{aligned}$$

As we are assuming $U_i(x_i) \to -\infty$ for $x_i \to 0$, we can ignore the condition $x_i \ge 0, i = 1, \ldots, n$. The corresponding KKT conditions are

$$\begin{aligned} \sum_{i:r\in i} x_i^* - C_r &\le 0 \quad r \in \mathcal{R} \\ \lambda_r^* &\ge 0 \quad r \in \mathcal{R} \\ \lambda_r^* \left(\sum_{i:r\in i} x_i^* - C_r \right) &= 0 \quad r \in \mathcal{R} \\ U_i'(x_i^*) - \sum_{r\in i} \lambda_r^* &= 0 \quad i = 1, \ldots, n\,. \end{aligned}$$

We interpret the Lagrange multipliers as prices p. In this formulation the KKT conditions become

$$\begin{aligned} y_r - C_r &\le 0 \quad r \in \mathcal{R} && (4.2.11) \\ p_r &\ge 0 \quad r \in \mathcal{R} && (4.2.12) \\ p_r\,(y_r - C_r) &= 0 \quad r \in \mathcal{R} && (4.2.13) \\ U_i'(x_i^*) - q_i &= 0 \quad i = 1, \ldots, n && (4.2.14) \end{aligned}$$

And the last condition may be rewritten as

$$x_i^* = \left(U_i'\right)^{-1}(q_i) \qquad (4.2.15)$$

We now consider the differential equation

$$\dot{p}_r = h_r(p_r) \begin{cases} (y_r - C_r) & \text{if} \quad p_r > 0 \\ \max\{0, (y_r - C_r)\} & \text{if} \quad p_r = 0 \end{cases}, \qquad (4.2.16)$$

where the h_r are positive functions which can be used to adjust the speed of convergence of the system of differential equations. It should be noted that the right hand side of the differential equation has a discontinuity at $p_r = 0$, so the standard theory of ordinary

differential equations does not apply. We do of course hope that setting $\dot{p}_r$ to zero if $p_r = 0$ and the drift $y_r - C_r$ would drive the state p_r to negative values just results in the solution staying put in $p_r = 0$. On the other hand we would like solutions to exist and still be unique also in the points of discontinuity. All of this can be made precise in the context of the theory of discontinuous differential equations. However, it is beyond the scope of these notes to elaborate the corresponding theory.

It is assumed that the total price for each user is transmitted to the user, so that they are able to adjust their rates according to (4.2.15).

Assume R has full row rank, so that $p_1 \neq p_2$ implies $q_1 = R^\top p_1 \neq R^\top p_2 = q_2$. In the following let x^*, p^* denote the unique optimal value of the optimization problem. Denote furthermore $y^* := Rx^*$ and $q^* := R^\top p^*$. Note that the assumption of full row rank of R uniquely determines p^* for $q^* \in \operatorname{Im} R^\top$.

Theorem 4.2.8 (Dual Agorithm) *If R has full row rank, the dual algorithm* (4.2.16) *in combination with the user behavior determined by* (4.2.15) *is globally asymptotically stable in the global optimum of the optimization problem with respect to initial conditions $p > 0$.*

A Lyapunov function is given by

$$W(p) = \sum_{r \in \mathcal{R}} (C_r - y_r^*)\, p_r + \sum_{i=1}^{n} \int_{q^*}^{q_i} \left(x_i^* - \left(U_i'\right)^{-1}(s) \right) ds$$

Proof. Taking the derivative of W along the solutions of (4.2.16) we obtain

$$\dot{W}(p) = \sum_{r \in \mathcal{R}} (C_r - y_r^*)\, \dot{p}_r + \sum_{i=1}^{n} \left(x_i^* - \left(U_i'\right)^{-1}(q_i) \right) \dot{q}_i$$

recalling (4.2.15) we may continue

$$= \sum_{r \in \mathcal{R}} (C_r - y_r^*)\, \dot{p}_r + \sum_{i=1}^{n} (x_i^* - x_i)\, \dot{q}_i$$

and by the full row rank condition $\dot{p}$ is uniquely determined by $\dot{q}$, so that

$$\begin{aligned}
&= \langle C - y^*, \dot{p} \rangle + \langle x^* - x, R^T \dot{p} \rangle = \langle C - y^*, \dot{p} \rangle + \langle R(x^* - x), \dot{p} \rangle \\
&= \langle C - y, \dot{p} \rangle \\
&= - \sum_{r, p_r > 0} h(p_r) |C_r - y_r|^2 - \sum_{r, p_r = 0} h_r(0)(C_r - y_r) \max\{0, (y_r - C_r)\} \leq 0\,,
\end{aligned}$$

where we have used the definition of the differential equation (4.2.16).

In the last step equality can only occur, if for each r we have $y_r = C_r$ or $p_r = 0$ and $C_r \geq y_r$. This means that the inequality constraint (4.2.11) as well as the complementary slackness condition of the KKT conditions, i.e. (4.2.13) is satisfied. Condition (4.2.14) is satisfied automatically, as we assume that users apply the price information as in (4.2.15). The second KKT condition (4.2.12) is satisfied by the definition of the differential equation. In total we see that the asymptotically stable fixed point of (4.2.16) together with (4.2.15) satisfies the KKT conditions and determines thus the unique optimal point of problem (4.2.10). ∎

We now consider the combination of the two approaches. I.e. we combine the dynamic update of the price information as well as of the sending rate. We then arrive at

$$\begin{aligned}\dot{x}_i &= k_i(x_i)\left(U_i'(x_i) - q_i(x(t))\right)\\ \dot{p}_r &= h_r(p_r)\begin{cases}(y_r - C_r) & \text{if} \quad p_r > 0\\ \max\{0, (y_r - C_r)\} & \text{if} \quad p_r = 0\end{cases}\end{aligned}$$

This approach is called the primal-dual algorithm for the utility maximization problem, as the users adjust their rates according to the primal algorithm and at the same time the routers adjust the rates according to the dual approach. Again we have a stability result that may be formulate as follows:

Theorem 4.2.9 (Primal-Dual Agorithm) *The primal-dual algorithm is globally asymptotically stable in (x^*, p^*). A Lyapunov function is given by*

$$W(x) = \sum_{i=1}^{n} \int_{x_i^*}^{x_i} \frac{1}{k_i(s)}\,(s - x_i^*)\,ds + \sum_{r\in\mathcal{R}} \int_{p_r^*}^{p_r} \frac{1}{h_r(s)}\,(s - p_r^*)\,ds\,.$$

So far we have assumed that the price information along the route is transferred back to the users using the acknowledgments. For this first routers have to update their respective price in the header of packets and then this information would have to be sent back. This idea has several drawbacks. For instance the price information would require a significant amount of capacity in the header, also the price update would require some processing of the packets. There is no provision for this overhead in current internet and it is quite unlikely that there ever will be.

A cheaper way of transmitting price information would be to use just one bit in the header of packages. Assume that instead of transferring price information each router sets a "price bit" with the probability of its current price. This assumes of course that prices are in the interval $[0, 1]$. The probability that packet is marked along the route i is then

$$q_i = 1 - \prod_{r:r\in i} (1 - p_r)\,.$$

This information could be sent back cheaply by setting one bit in acknowledgments. The dynamic equations corresponding to this scheme are given by

$$\dot{x}_i = k_i(x_i)\left((1 - q_i)U_i'(x_i) - q_i\right)\,.$$

Also in this case it is possible to show a stability result. However, the asymptotically stable fixed point now has little relation with the optimal points obtained by the utility maximization approach.

Theorem 4.2.10 *The one bit marking controllers are globally asymptotically stable.*

A Lyapunov function is given by the convex function

$$W(x) = -\sum_{i=1}^{n} \int_0^{x_i} \log(1 + U_r'(s))ds - \sum_{r\in\mathcal{R}} \int_0^{y_r} \log(1 - f_r(s))ds\,.$$

4.2.1 A Utility Based View of TCP

In this final section we want to relate the previous models of utility based congestion control to implementations of TCP. Of course, in TCP no price information as such is sent back to the users. But the loss of packages and the frequency with which this occurs can be given a price information. We now consider a differential model for TCP. The variables needed are the following:

- $w_i(t)$: the window size of user i at time t.
- T_i: the round trip time (RTT).
- $q_r(t)$: the fraction of packets lost.
- $x_i(t) = w_i(t)/T_i$: the transmission rate.
- β_i: the decrease factor.

We consider the case of standard TCP, that is, $\alpha = 1$ in terms of the discrete time model we had discussed previously. We now consider a fluid approximation of the dynamics in which it does not really make sense to think of packets send, but rather users are sending continuous pieces of information. Consider user i. Of the rate sent one round trip time previously, thus at time $t - T_i$, a certain amount is acknowledged at time t, which gives the proportion $x_i(t - T_i)(1 - q_i(t))$. For these acknowledged packets the window size is increased by $1/w_i(t)$ for each acknowledgment. On the other hand for unacknowledged packages the window size is decreased by βw_i. With these considerations we arrive at the differential equation

$$\dot{w}_i(t) = \frac{x_i(t - T_i)(1 - q_i(t))}{w_i(t)} - \beta x_i(t - T_i) q_i(t) w_i(t)\,. \tag{4.2.17}$$

If the round trip time T_i is small, then using the approximation $x_i(t) = x_i(t - T_i)$ we obtain the approximate system

$$\begin{aligned}\dot{x}_i &= \frac{1 - q_i(t)}{T_i^2} - \beta x_i^2 q_i(t) \\ &= \left(\beta x_i^2 + \frac{1}{T_i^2}\right)\left(\frac{1}{\beta T_i^2 x_i^2 + 1} - q_i(t)\right)\end{aligned}$$

The latter term we can interpret in terms of the utility approach. The first factor is simply a positive factor, which corresponds to the factor $k_i(x_i)$ in the primal approach. The second factor is the difference of a function of x_i and the fraction of packets lost q_i. We may now interpret this fraction as the total price that user i has to pay along her route. The form of the differential equation then closely resembles the equations for the primal approach for utility based congestion control, cp. (4.2.9).

This corresponds to the utility function given by

$$U_i'(x_i) = \frac{1}{\beta T_i^2 x_i^2 + 1}$$

so that the corresponding utility function is given by

$$U_i(x_i) = \frac{\arctan(x_i T_i \sqrt{\beta})}{\sqrt{\beta T_i}}\,.$$

We note that this function does not satisfy all our assumptions as $U_i(x_i) \not\to -\infty$ as $x_i \to 0$. This assumption however, was only introduced to ensure the existence of optimal points and asymptotically stable fixed points for the primal problem. If the existence of such points can be guaranteed by other means, then these assumptions are of no further relevance.

With this remarks we conclude our brief discussion of congestion control. The main point of the discussion was to provide a brief introduction and also to show how convex optimization may be instrumental in understanding other problems which at first glance do not seem directly related.

Index

Bibliography

[1] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, NJ, 2009.

[2] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*, volume 9 of *Classics in Applied Mathematics*. SIAM Publications, Philadelphia, PA, 1994.

[3] A. Berman, R. Shorten, and D. Leith. Positive matrices associated with synchronised communication networks. *Linear Algebra Appl.*, 393:47–54, 2004.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[5] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[6] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[7] R. N. Shorten, C. King, F. Wirth, and D. Leith. Modelling TCP congestion control dynamics in drop-tail environments. *Automatica*, 43(3):441–449, 2007.

[8] R. N. Shorten, F. Wirth, and D. Leith. A positive systems model of TCP-like congestion control: Asymptotic results. *IEEE/ACM Transactions on Networking*, 14(3):616–629, 2006.

[9] R. Srikant. *Internet congestion control*, volume 14 of *Control theory*. Birkhäuser Boston Inc., Boston, MA, 2004.

[10] J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions*. Springer-Verlag, Berlin, 1970.

[11] H. Wolkowitz, R. Saigal, and L. Vandenberghe, editors. *Handbook if Semidefinite Programming*. Kluwer, Boston, MA, 2000.