An Introduction to Duality in Convex Optimization

Stephan Wolf
Betreuer: Stephan M. Günther, M.Sc.
Hauptseminar Innovative Internettechnologien und Mobilkommunikation SS2011
Lehrstuhl Netzarchitekturen und Netzdienste
Fakultät für Informatik, Technische Universität München
Email: wolfst@in.tum.de

ABSTRACT
This paper provides a short introduction to the Lagrangian duality in convex optimization. At first the topic is motivated by outlining the importance of convex optimization. After that mathematical optimization classes such as convex, linear and non-convex optimization, are defined. Later the Lagrangian duality is introduced. Weak and strong duality are explained and optimality conditions, such as the complementary slackness and Karush-Kuhn-Tucker conditions are presented. Finally, three different examples illustrate the power of the Lagrangian duality. They are solved by using the optimality conditions previously introduced.

The main basis of this paper is the excellent book about convex optimization [5] of Stephen Boyd and Lieven Vandenberghe.

Keywords
mathematical optimization problem, convex optimization, linear optimization, Lagrangian duality, Lagrange function, dual problem, primal problem, strong duality, weak duality, Slater's condition, complementary slackness, Karush-Kuhn-Tucker conditions, constrained least squares problem, water filling algorithm

1. MOTIVATION
Convex optimization[1] is very important in practice. Applications are numerous. Important areas are for example automatic control systems, estimation and signal processing, communications and networks, electronic circuit design, data analysis and modelling, statistics and finance (see [5], p. xi). Furthermore linear optimization, which is a subclass of convex optimization, bases mainly on the theory of convex optimization.

One advantage of these convex optimization problems is that there exists methods to solve them very reliably and efficiently, whereas there are no such methods for the general non-linear problem so far. One example is the interior-point method, which can be used to solve general convex optimization problems. However, its reliability and efficiency are still an active topic of research, but it is likely that these difficulties will be overcome within a few years. (See [5], p.8).

Another even more important advantage is the associated dual problem. Each convex optimization problem can be transformed to a dual problem, which provides another perspective and mathematical point of application. With the dual problem it is often possible to determine the solution of the primal problem analytically or alternatively to determine efficiently a lower bound for the solution (even of non-convex problems). Furthermore the dual theory is a source of interesting interpretations, which can be the basis of efficient and distributed solution methods.

Therefore, when tackling optimization problems, it is advisable to be able to use the powerful tool of Lagrangian duality. This paper offers an introduction to this topic by outlining the basics and illustrating these by three examples.

The following section presents an overview over the different optimization classes and explains the difference of convex and linear optimization. After that, Lagrangian duality is introduced and intuitively derived. Furthermore, weak and strong duality are explained and Slater's condition, which guarantees strong duality for convex optimization, is described. The Section 4 introduces optimality conditions, concretely the complementary slackness condition and the Karush-Kuhn-Tucker conditions. They will be used to demonstrate the power of duality to solve convex optimization problems by the dual in Section 5. In particular, duality is used to solve a constrained least squares problem and to derive the water-filling method. At the end, a conclusion is drawn and further literature hints are presented.

2. OPTIMIZATION PROBLEMS
There are different kinds of mathematical optimization problems, for example non-convex, convex and linear as well as constrained and unconstrained optimization problems. These classes do not only differ in their definition, but also in their solvability. The more specific requirements for an optimization class are, the easier it is usually to solve.

2.1 The general optimization problem
The standard form of a *mathematical optimization problem* or just *optimization problem* consists of an *optimization variable* $x = (x_1, ..., x_n)$ and an *objective function* $f_0 : \mathbb{R}^n \mapsto \mathbb{R}$. Furthermore there are *inequality constraint functions* $f_i : \mathbb{R}^n \mapsto \mathbb{R}$ and *equality constraint functions* $h_i : \mathbb{R}^n \mapsto \mathbb{R}$, which constrain the solution.

[1] The definition of convex optimization problems and convexity itself can be found in Section 2.3

doi: 10.2313/NET-2011-07-2_20

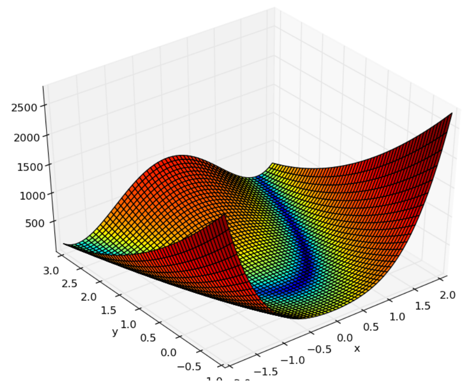

Figure 1: The non-convex Rosenbrock function $f(x, y) = (1 - x)^2 + 100(y - x^2)^2$.

The standard form of the problem is:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, ..., m \\ & h_i(x) = 0, \quad i = 1, ..., p \end{array}$$

The problem is to find x such that the *objective function* f_0 is minimized while satisfying the inequality and equality constraints. If the problem has no constraints, it is called *unconstrained*.

The set which the objective and constraint functions are defined for is called the *domain* and is defined as:

$$\mathbb{D} = \bigcap_{i=0}^{m} \mathbf{dom} f_i \cap \bigcap_{i=0}^{p} \mathbf{dom} h_i$$

A point $x \in \mathbb{D}$ is *feasible* if it satisfies the constraints. The problem itself is feasible if there exists at least one feasible point. All feasible points form the *feasible set* or *constraint set*.

A vector $x^\star = (x_1, ..., x_n)$ which is feasible and minimizes the objective function is called *optimal* or *solution*. Its corresponding value is called *optimal value* $p^\star$ and is defined as:

$$\inf\{f_0(x) | f_i(x) \leq 0, i = 1, ..., m \wedge h_j(x) = 0, j = 1, ..., p\}$$

By definition $p^\star$ can be $\pm\infty$. $p^\star$ is $+\infty$ if the problem is infeasible and $-\infty$ if the problem is *unbounded below*, that means that there are feasible points x_k with $f_0(x_k) \to -\infty$ for $k \to \infty$.

The other problem classes are subclasses of the general optimization problem. The main difference is the class of the objective and constraint functions. Figure 1 shows the Rosenbrock function.[2] It is a non-convex function, which is used as performance test for optimization algorithms for non-convex problems.

[2]The plot bases on a script from [1]

2.2 The linear optimization problem

The problem is called a *linear program*, if the objective function f_0 and the inequality and equality constraints $f_1, ..., f_m$, $h_1, ..., h_p$ are linear, that means that they fullfill the following equation for all $x, y \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$:

$$f_i(\alpha x + \beta y) = \alpha f_i(x) + \beta f_i(y)$$

One example for a two dimensional linear function is shown in Figure 2.

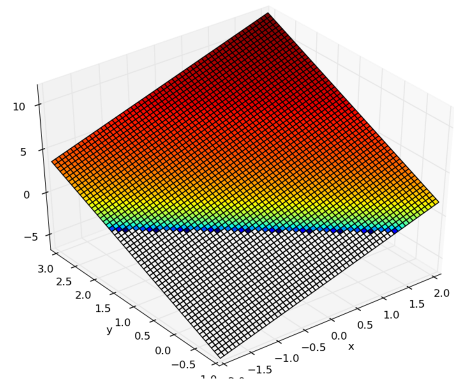

Figure 2: A linear function $f(x, y) = 3x + 2.5y$.

A linear program can also be written as:

$$\begin{array}{ll} \text{minimize} & c^T x + d \\ \text{subject to} & Gx \preceq q \\ & Ax = b \end{array}$$

The matrices $G \in \mathbb{R}^{m \times n}$ and $A \in \mathbb{R}^{p \times n}$ specify the linear inequality and equality constraints and the vectors c and $d \in \mathbb{R}^n$ parameterize the objective function. The vector d can be left out, as it does not influence the feasible set and the solution $x^\star$ (see [5], p. 146). Therefore the vector d is ignored in other definitions.

As the negation of a linear function $-f(x)$ is also linear, a linear maximization problem can be easily transformed to a linear minimization problem. For example, if the objective function $c^T x + d$ should be maximized, one can solve the problem by minimizing the objective function $-c^T x - d$. That is the reason why linear maximization problems are also linear programs.

If at least one constraint or the objective function is not linear, then the problem is called a *non-linear program*.

2.3 The convex optimization problem

The requirement for convex optimization problems is that the equality constraints are still linear but the inequality constraints and the objective function have to be convex, that means they must fulfill the following inequality for all $x, y \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$, with $\alpha + \beta = 0, \alpha, \beta \geq 0$:

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

As one can see, this requirement is less restrictive as the previous requirement for linear programs, where equality is required. Consequently the linear programs can be seen as

doi: 10.2313/NET-2011-07-2_20

a subclass of the convex optimization problems and the theory of convex optimization can be also applied to linear programs.

Figure 3 illustrates a convex function. The intuitive characteristics of such functions is that if one connects two points, the inner line segment always lies above the graph.

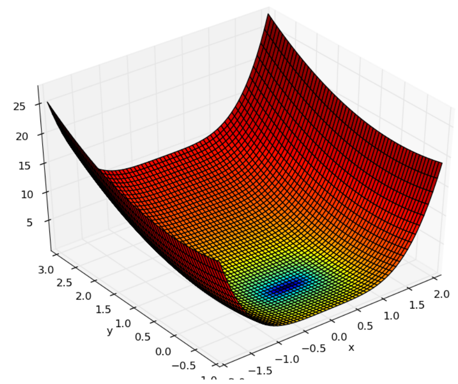

Figure 3: The convex function $f(x, y) = x^4 + y^2$.

3. THE LAGRANGE DUAL PROBLEM

Optimization problems can be transformed to their dual problems, called Lagrange dual problems, which help to solve the main problem. First, with the dual problem one can determine lower bounds for the optimal value of the original problem. Second, under certain conditions, the solutions of both problems are equal. In this case the dual problem often offers an easier and analytical way to the solution.

3.1 Lagrangian function

Let us take the general optimization problem of the standard form, of which we do not know anything about the convexity or linearity of the constraint or objective functions:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le b_i, \quad i = 1, ..., m \\ & h_i(x) = 0, \quad i = 1, ..., p \end{array}$$

We define the *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \mapsto \mathbb{R}$ of the problem as sum of the objective function and a weighted sum of the constraint functions:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

The domain of the dual problem is equal to the domain of the primal problem times the domain of the parameters:

$$\mathbf{dom}\, L = \mathbb{D} \times \mathbb{R}^m_{0,+} \times \mathbb{R}^p$$

λ_i is called the *Lagrange multiplier* of the i-th inequality constraint $f_i(x) \le 0$ and accordingly ν_i is called the *Lagrange multiplier* of the i-th equality constraint $h_i(x) = 0$. The vectors λ and ν are referred to as the *dual variables* or *Lagrange multiplier vectors.*

In addition to that, the *Lagrange dual function* (or just *dual function*) $g : \mathbb{R}^m \times \mathbb{R}^p \mapsto \mathbb{R}_{0,+}$ is the infimum of the Lagrangian over x (for all $\lambda \in \mathbb{R}^m$,$\nu \in \mathbb{R}^p$)

$$g(\lambda, \nu) = \inf_{x \in \mathbb{D}} L(x, \lambda, \nu)$$

If there is no lower bound of the Lagrangian, its dual function takes on the value $-\infty$. The main advantage of the Lagrangian dual function is, that it is concave even if the problem is not convex. The reason for this is that the dual function is the pointwise infimum of a family of linear functions of (λ, ν) (see [5], p. 216).

The basic idea behind Lagrangian duality is to take the constraints and put them into the objective function. The most intuitive way would be to rewrite the problem as the following unconstrained problem:

$$\text{minimize } l(x) = f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) + \sum_{i=1}^{p} I_0(h_i(x))$$

Here I_- and I_0 ($\mathbb{R} \mapsto \mathbb{R}$) are the indicator functions of non-positive reals and 0 respectively:

$$I_-(u) = \begin{cases} 0 & u \le 0 \\ \infty & u > 0 \end{cases} \qquad I_0(u) = \begin{cases} 0 & u = 0 \\ \infty & u \neq 0 \end{cases}$$

These indicator functions express our displeasure with previously infeasible points. If a point was previously infeasible, that means at least one constraint was violated, then at least one indicator function takes the value ∞ and prohibits that point from being a solution. However, this method is really brutal and causes discontinuity at the edges of the feasible set. This discontinuity is not desired as we want to use analytical techniques to solve the problem. So it is advisable to find another solution which offers a smoother transition.

In Lagrangian duality, these indicator functions are replaced by linear functions which approximate the hard indicator functions. Concretely, $I_-(u)$ is replaced by $\lambda_i u$ ($\lambda_i \ge 0$) and $I_0(u)$ is replaced by $\nu_i u$ (here the domain of ν_i is not restricted). When the inequality constraint $f_i(x)$ is 0 then our displeasure is 0. However, when the inequality constraint is greater than zero, our displeasure is finite, but depends on "how" much the constraint is violated (remind $\lambda_i \ge 0$). On the other side, our pleasure grows when the constraint is "more" fulfilled, i.e. it has more margin.

Clearly this approximation is rather poor, but it is ensured that the linear functions underestimate the indicator functions since $\lambda_i u \le I_-(u)$ and $\nu_i u \le I_0(u)$ for all $u \in \mathbb{R}$. As a result, the dual function is always a lower bound for the optimal value of the original function, i.e. for any $\lambda \succeq 0$ and any ν holds:

$$g(\lambda, \nu) \le p^\star$$

This can be easily proven. Let $\tilde{x}$ be a feasible point, then $f_i(\tilde{x}) \le 0$ and $h_i(\tilde{x}) = 0$. Consequently:

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \le 0$$

As a result the inequality follows:

$$g(\lambda, \nu) = \inf_{x \in \mathbb{D}} L(x, \lambda, \nu) \le L(\tilde{x}, \lambda, \nu) \le f_0(\tilde{x})$$

doi: 10.2313/NET-2011-07-2_20

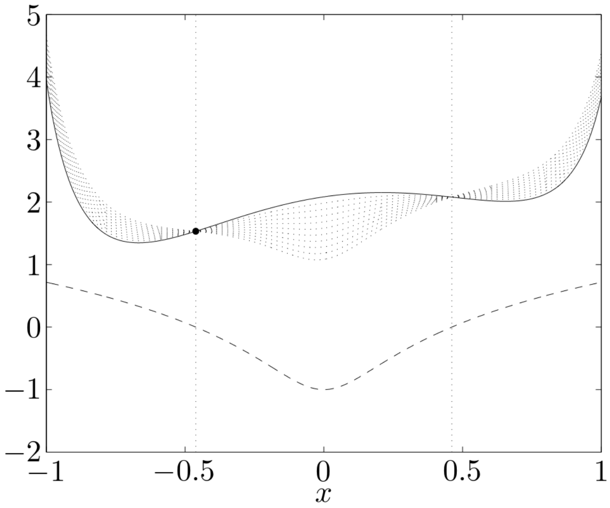

Figure 4: Illustration of the lower bound (from [5], p. 217)

Figure 4 illustrates this. The solid curve represents the objective function f_0 and the dashed curve shows the constraint function f_1. The feasible set is characterized by $f_1(x) \leq 0$ and here it is the interval $[-0.46, 0.46]$, which is indicated by the two dotted vertical lines. The circle shows the optimal point $(x^\star, p^\star) = (-0.46, 1, 54)$ and the dotted curves show $L(x, \lambda)$ for $\lambda = 0.1, 0.2, ..., 1.0$. As we see, $L(x, \lambda) \leq f_0(x)$ holds for the feasible set and $\lambda \geq 0$. Consequently, each minimum value of $L(x, \lambda)$ is less or equal to $p^\star$.

However, when $g(\lambda, \nu) = -\infty$ then the inequality is useless. The lower bound for $p^\star$ only makes sense if $\lambda \geq 0$ and $(\lambda, \nu) \in \mathbf{dom}\, g$, which means $g(\lambda, \nu) > -\infty$. We call such a pair (λ, ν) *dual feasible.*

The challenge is to find the best lower bound, which leads to the following optimization problem, called the *Lagrangian dual problem* (whereas the original problem is often referred to as *primal problem*):

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

We define $(\lambda^\star, \nu^\star)$, which is one solution to this problem, as *dual optimal* or *optimal Lagrange multipliers*. As the dual objective function is concave (even if the original problem is not) and the constraints are convex, one can solve the problem by minimizing $-g(\lambda, \nu)$, which is consequently convex. Therefore the dual problem is equivalent to a convex minimization problem.

3.2 Weak duality

After estimating the optimal value of the dual problem $d^\star$, we have by definition, the best lower bound for the optimal value of the primal problem $p^\star$, which can be found using Lagrange duality:

$$d^\star \leq p^\star$$

This inequality also applies if the original problem is not convex and is called *weak duality.*

It also holds when $p^\star$ and $d^\star$ are infinite. If the original problem is unbounded below, this means $p^\star = -\infty$, then the optimal value of the Lagrange dual problem $d^\star$ is consequently also $-\infty$ and the dual problem is *infeasible.* Whereas when the dual problem is unbounded above, this means $d^\star = \infty$, then $p^\star = \infty$ and the primal problem is *infeasible.*

The difference $p^\star - d^\star$ is an important value as it characterizes the gap between the optimal value of the primal problem and its best lower bound. Accordingly it is called the *duality gap* and as a result of the previous inequality it is always non-negative.

Although the weak duality does not enable us to find the exact solution of the primal problem, it is useful in practice. The main advantage is that the dual problem is a concave maximization problem and therefore one can efficiently calculate a lower bound, as it can be easily transformed to a convex minimization problem. (see [5], p.226).

In [5] this is demonstrated by the two-way partitioning problem. Given a set of n elements, the task is to find a partition which minimizes costs. The costs are specified by a matrix W. If two elements i and j are in one partition, then they cause the cost $w_{i,j}$ and, if they are in different partitions, they cause the cost $-w_{i,j}$.

The problem can be described as a non-convex problem:

$$\begin{array}{lll} \text{minimize} & x^T W x & \\ \text{subject to} & x_i^2 = 1 & i = 1, ..., n \end{array}$$

The components x_i of the vector $x \in \mathbb{R}^n$ are restricted to -1 and $+1$ by the equality constraint and define whether the object i is in partition 1 or 2. The matrix $W \in \mathbb{R}^{n \times n}$ specifies the corresponding costs as stated before, and consequently $x^T W x$ produces the total costs. This problem is hard to solve, as the complexity rises exponentially with n.

Fortunately it can be transformed to a dual problem:

$$\begin{array}{ll} \text{maximize} & -1^T \nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0 \end{array}$$

diag creates a $n \times n$ matrix with the components of the vector on the diagonal. For a more detailed description of the derivation of the dual problem see [5], p. 219f.

This problem can be solved efficiently by semidefinite programming and delivers a useful lower bound for the hard primal problem.

3.3 Strong duality

Strong duality is even more useful. By definition, *strong duality* means that the duality gap is zero, i.e. that the optimal value of the dual problem is equal to the optimal value of the primal problem:

$$d^\star = p^\star$$

Whereas weak duality always holds, strong duality only holds

doi: 10.2313/NET-2011-07-2_20

under certain conditions. For convex problems, strong duality is mostly achieved. But to be precise, convex problems must also satisfy other conditions which are called *constraint qualifications*.

Slater's constraint qualification

One very simple and widespread example for a constraint qualification is *Slater's condition*:

$$\exists x \in \textbf{relint}\,\mathbb{D} : \forall i = 1, ..., m : f_i(x) < 0 \land Ax = b$$

This means, if one can find a point which is *strictly feasible* and the problem is convex, then strong duality applies. For clearness, **relint** $\mathbb{D}$ denotes the relative interior of $\mathbb{D}$, which means intuitively all interior points of the set and not the points on the edge.

For convex problems with linear inequality constraints, there also exists a *refined Slater's condition*. Given the first k constraint functions are linear, then strong duality also applies under the following condition:

$$\exists x \in \textbf{relint}\mathbb{D} : Ax = b \land$$
$$\forall i = 1, ..., k : f_i(x) \le 0 \land \forall i = k, ..., m : f_i(x) < 0$$

This means, strict inequality is only required for nonlinear constraint functions. As a result, the refined Slater's condition reduces to feasibility if all equality and inequality constraints are linear and the domain of the objective function f_0 is open. (See [5], p. 227)

In addition to that, Slater's condition also implies that not only strong duality holds for convex problems, but also guarantees that the dual optimal value is attained if $d^\star > -\infty$, i.e. that there exists a dual feasible point $(\lambda^\star, \nu^\star)$ with $g(\lambda^\star, \nu^\star) = d^\star = p^\star$ (See [5], p.227; [4] p.90 "dual attainment theorem")

In practice, real problems usually fulfill Slater's condition. In engineering for example, given an inequality constraint which limits the force usually satisfies Slater's condition. If Slater's condition would not apply, then this would imply, for example for the inequality constraint $F < 100$, that it is possible to have a force which is 99,9999 Newton, but it is impossible to have a force that is 100 Newton. This differentiation usually does not make sense in practice.

For a proof that Slater's condition implies strong duality see [5], §5.3.2 p.234ff

4. OPTIMALITY CONDITIONS

One motivation of the Lagrangian duality was that it provides a theoretical anchor which helps solving the problem. Above all, optimality conditions are often used to determine the solution of the primal problem by solving the dual problem analytically. As an example, the complementary slackness condition is now presented, which will be later used to solve the constrained least-square problem. Furthermore the more powerful but also more complex Karush-Kuhn-Tucker conditions are described[3] and will be used to derive the water-filling method used in information theory or the convex quadratic minimization problem.

[3] which also contain the complementary slackness condition

4.1 Complementary slackness

The *complementary slackness* condition is an optimality condition. That means, an optimal value must satisfy this condition if strong duality holds.

If $x^\star$ is the primal optimal and $(\lambda^\star, \nu^\star)$ the dual optimal, then we can state:

$$\begin{aligned} f_0(x^\star) &= g(\lambda^\star, \nu^\star) && (1) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^\star f_i(x) + \sum_{i=1}^{p} \nu_i^\star h_i(x) \right) && (2) \\ &\le f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star h_i(x^\star) && (3) \\ &\le f_0(x^\star) && (4) \end{aligned}$$

The first line results from strong duality. The second line is the definition of the dual function and the third line holds, since the infimum of the dual function over x is less or equal to its value at $x = x^\star$. As $x^\star$ is feasible, $f_i(x^\star) \le 0$ holds for $i = 1, ..., m$ and $h_i(x^\star) = 0$ holds for $i = 1, ..., p$. In addition to that, λ_i is always non-negative and consequently the final inequality follows.

As a result, $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$. However it does not have to be the only minimizer. The Lagrangian $L(x, \lambda^\star, \nu^\star)$ can also have other minimizers (see [5], p. 243).

Second, we can conclude:

$$\sum_{i=1}^{m} \lambda_i f_i(x^\star) = 0$$

And since each summand in this sum is non-positive, it follows:

$$\lambda_i f_i(x^\star) = 0 \qquad \forall i = 1, ..., m$$

This condition is called the *complementary slackness* condition. It states that if the i-th inequality constraint is not *active*, that means $f(x) < 0$, then λ_i must be 0. On the other hand, if $\lambda_i > 0$ then the i-th inequality constraint must be active ($f_i(x) = 0$).

$$\begin{aligned} \lambda_i^\star > 0 \quad &\Rightarrow \quad f_i(x^\star) = 0 \\ f_i(x^\star) < 0 \quad &\Rightarrow \quad \lambda_i^\star = 0 \end{aligned}$$

But, to emphasize it, this requires strong duality.

4.2 Karush-Kuhn-Tucker conditions

The complementary slackness condition is part of the more comprehensive Karush-Kuhn-Tucker optimality conditions. They also require strong duality, but also differentiability of the constraint and objective functions, $f_0, ..., f_m$ and $h_1, ..., h_p$. In return, the Karush-Kuhn-Tucker conditions are more powerful compared to the complementary slackness condition. In addition to that, for convex problems, they are even sufficient and not only necessary.

4.2.1 Non-convex problems

At first we consider nonconvex problems. Let $x^\star$ and $(\lambda^\star, \nu^\star)$ again be the primal and dual optimal points. In order that $x^\star$ minimizes the Lagrangian $L(x, \lambda^\star, \nu^\star)$ its gradient

doi: 10.2313/NET-2011-07-2_20

$\nabla L(x, \lambda^\star, \nu^\star)$ must vanish. This condition is called *stationarity*:

$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star \nabla h_i(x^\star) = 0$$

When summarizing all conditions for the optimal point, we have collected so far, we get:

$$\begin{aligned}
\text{Primal feasibility:}\quad & \forall i = 1, ..., m: & f_i(x^\star) \leq 0 \quad & (1)\\
& \forall i = 1, ..., p: & h_i(x^\star) = 0 \quad & (2)\\
\text{Dual feasibility:}\quad & \forall i = 1, ..., m: & \lambda_i^\star \geq 0 \quad & (3)\\
\text{Compl. slackness:}\quad & \forall i = 1, ..., m: & \lambda_i^\star f_i(x^\star) = 0 \quad & (4)
\end{aligned}$$

Stationarity:

$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star \nabla h_i(x^\star) = 0 \qquad (5)$$

These conditions are called the *Karush-Kuhn-Tucker (KKT)* conditions. Given a problem with differentiable constraint and objective function for which strong duality holds, any pair of primal and dual optimal points must fulfill these conditions. However the KKT conditions are not sufficient for non-convex problems.

4.2.2 Convex problems

For convex problems the KKT-conditions are the same, but they are now also sufficient. That means, given a pair of a primal and dual solution $(\tilde{x}, (\tilde{\lambda}, \tilde{\nu}))$, we can not only check whether this pair is not primal and dual optimal, but we can also check if it is.

Clearly, this means if the inequality constraint and objective functions $f_0, ..., f_m$ are convex and the equality constraint functions $h_1, ..., h_p$ are linear, and the KKT conditions are satisfied by some points $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu})$, then these points are consequently primal and dual optimal points.

Furthermore, if the KKT conditions are satisfied, it implies that the duality gap is zero. The first two KKT conditions (1, 2) state that $\tilde{x}$ is primal feasible. The third condition (3) ensures that the Lagrangian $L(x, \tilde{\lambda}, \tilde{\nu})$ is convex in x, as it consists of a positive sum of convex functions. And as a result of the last condition (5) and the convexity of the Lagrangian, $\tilde{x}$ minimizes $L(x, \tilde{\lambda}, \tilde{\nu})$ over x. Thus:

$$\begin{aligned}
g(\tilde{\lambda}, \tilde{\nu}) &= L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})\\
&= f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(\tilde{x}) + \sum_{i=1}^{p} \tilde{\nu}_i h_i(\tilde{x})\\
&= f_0(\tilde{x})
\end{aligned}$$

The last line results from condition (3) and (4) and thus the duality gap is zero: $p^\star - d^\star = f_0(\tilde{x}) - g(\tilde{\lambda}, \tilde{\nu}) = 0$.

Although we don't need Slater's condition here to prove that the duality gap is zero, we need it for the sufficiency of the optimality criteria. The KKT-conditions depend on the existence of a pair of primal and dual optimal values. However, it is possible, that the primal problem has a primal optimum but the dual problem does not. Consequently, the KKT conditions are useless for finding the primal optimal value, as no pair of primal and dual optimal values can be found and we cannot conclude from a non-existing pair to a non-existing primal optimal. Therefore the KKT conditions alone are not sufficient for determining the primal optimum.

Here Slater's condition helps. When the problem satisfies Slater's condition, then the corresponding dual optimum is always attained. As a result, the existence of a primal optimum requires the existence of a corresponding dual optimum and vice versa. That means, that x is an optimal value of the primal problem only if there exists an (λ, ν) that fulfills the KKT conditions. Then the KKT conditions are necessary and sufficient for optimality (see [5], p.244).

The KKT conditions play an important role in convex optimization. On the one hand, they can be used to solve the problem analytically in special cases, as shown in the example of the derivation of the water filling method. On the other hand, many algorithms solving convex problems are based on the KKT conditions (see [5], p.244).

5. SOLVING THE PRIMAL BY THE DUAL

In this section, we use the optimality criteria to solve optimization problems by considering the dual.

5.1 Equality constrained convex quadratic minimization

At first, let us consider a very simple problem. The task is to minimize a convex quadratic function subject to a set of linear equality constraints (example from [5], p. 244)

$$\begin{array}{ll}
\text{minimize} & f_0(x) = \frac{1}{2} x^T P x + q^T x + r \qquad (P \in \mathbb{S}_+^n)\\
\text{subject to} & Ax = b \text{ (equal to: } Ax - b = 0)
\end{array}$$

Here the objective function is a n-dimensional quadratic function with the parameters q and P, which is a symmetric positive definite $n \times n$ matrix. It is subject to a number of linear equality constraints.

As the problem is convex and there are only linear equality constraints, Slater's condition applies.[4] Therefore we can use the KKT conditions to easily determine the solution of this problem. At first we have to satisfy the equality constraint:

$$Ax = b \qquad (1)$$

Second the gradient of the Lagrangian must vanish at the optimum:

$$\begin{aligned}
L(x, \lambda, \nu) &= f_0(x) + \sum_{i=1}^{p} \nu_i h_i(x) && (2)\\
&= \frac{1}{2} x^T P x + q^T x + r + \nu^T (Ax - b) && (3)\\
\nabla L(x^\star, \lambda^\star, \nu^\star) &= \frac{\partial L}{\partial x^T} = P x^\star + q + A \nu^\star \stackrel{!}{=} 0 && (4)
\end{aligned}$$

As a result of 1 and 4, we get the following system of linear equations:

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \cdot \begin{bmatrix} x^\star \\ \nu^\star \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

[4] We assume that the problem is feasible

doi: 10.2313/NET-2011-07-2_20

The solution to this set of $n + m$ equations provides us the optimal primal and dual variables.

5.2 Constrained least squares problem

We consider the problem of finding a linear function $f(x) = mx + t$, which approximates the best a given set of points (x_i, y_i). Best in this case means that the sum of the squares of errors has to be minimized. Furthermore, we consider the constraint $t \leq 0$.

Figure 5 shows a random example of such a constrained least-square problem. The black circles represent the data samples which have to be approximated by a linear function. Without the constraint the solution would be the red function. However, our application domain requires $t \leq 0$, so that this cannot be a solution. The solution to the constraint problem is the blue dashed function.

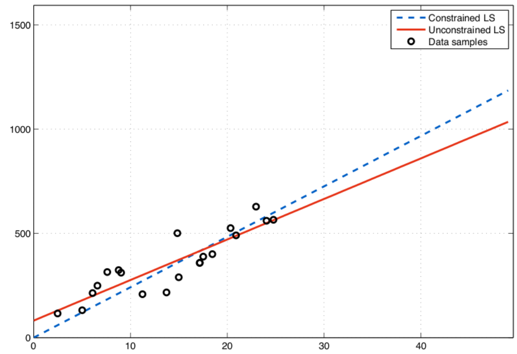

Figure 5: A random constrained least square example.

We can describe the problem with the standard form:

$$\begin{array}{ll} \text{minimize} & f_0(x) = ||Ax - d||_2^2 \\ \text{subject to} & g^T x \leq 0, \qquad g = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{array}$$

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \qquad d = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad x = \begin{bmatrix} m \\ t \end{bmatrix}$$

The parameters m, t of the linear function are the optimization variables. They have to be determined such that the objective function is minimized. The objective function represents the sum of the squares of errors in the Euclidian norm. This can be easily proven:

$$Ax - d = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ t \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} mx_1 + t - y_1 \\ mx_2 + t - y_2 \\ \vdots \\ mx_n + t - y_n \end{bmatrix}$$

$$\begin{aligned} \Rightarrow ||Ax - d||_2^2 &= \sum_{i=1}^{n} \sqrt{(mx_i + t - y_i)^2}^2 \\ &= \sum_{i=1}^{n} (mx_i + t - y_i)^2 \end{aligned}$$

The inequality constraint $t \leq 0$ is expressed by $g^T x \preceq 0$.

For simplicity, it is advisable to replace the norm:

$$\begin{aligned} f_0(x) &= ||Ax - d||_2^2 \\ &= (Ax - d)^T (Ax - d) \\ &= x^T A^T Ax - x^T A^T d - d^T Ax + d^T d \end{aligned}$$

Let us define $B = A^T A$. B plays an important role for the solvability of the problem. If the problem is well formed, then B is invertible.

Fortunately this problem is convex. The sum of squares of errors is apparently convex and the inequality constraint is even linear. Thus Slater's condition is fulfilled and strong duality holds. Therefore it is a good choice to consider the dual function to solve the problem.

The Lagrangian of this problem is:

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

As the problem is convex and meets Slater's condition, the KKT conditions apply and we can determine the infimum of the Lagrangian over x by setting its gradient to 0:

$$\begin{aligned} & g(x, \lambda) = \inf_x L(x, \lambda) \\ \Longleftrightarrow & \nabla L(x, \lambda) = \frac{\partial L}{\partial x^T} = 2Bx - 2A^T d - \lambda g \overset{!}{=} 0 \\ \Longleftrightarrow & x = B^{-1} \left(A^T d + \frac{1}{2} \lambda g \right) \end{aligned}$$

This result can now be inserted into the dual function, which leads to the dual problem:

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \geq 0 \end{array}$$

The solution can be obtained by solving this problem and maximizing the objective function. However, this is not advisable as it leads to cumbersome calculations. An easier approach is to consider optimality conditions. Because of strong duality, we can use the complementary slackness constraint:

$$\lambda^\star > 0 \Rightarrow g^T x^\star = 0$$

If $g^T x^\star < 0$ holds, then the inequality constraint $t \leq 0$ is inactive ($\lambda^\star = 0$), that means it does not determine the solution. Consequently the problem reduces to an unconstrained least squares problem. The second case is that $\lambda^\star > 0$ applies, that means that the constraint is active and determines the solution. Then you can determine λ since $g^T x^\star$ must be 0:

$$\begin{aligned} g^T x^\star = g^T B^{-1} \left(A^T d + \frac{1}{2} \lambda g \right) &\overset{!}{=} 0 \\ \Rightarrow \lambda = -2 \frac{g^T B^{-1} A^T d}{g^T B^{-1} g} & \end{aligned}$$

 doi: 10.2313/NET-2011-07-2_20

Now one can obtain the solution of the constrained problem by determining x by the use of λ.

To put it all in a nutshell, here is the complete solution:

$$x = (A^T A)^{-1}(A^T d + \frac{1}{2}\lambda g),$$

$$\text{with } \lambda = \begin{cases} -2\frac{g^T B^{-1} A^T d}{g^T B^{-1} g} & \text{if } g^T x \leq 0 \, active \\ 0 & otherwise \end{cases}$$

Whether the constraint is active or inactive depends on the problem. By determining the solution of the unconstrained problem ($\lambda = 0$), one can easily check if $t \leq 0$. If this is the case, then the problem is solved. Otherwise the solution of the constrained problem with the active inequality constraint has to be calculated.

5.3 Water-filling

Let us consider a practical problem from information theory. It's about capacity optimization in multiple-input and multiple-output (MIMO) communication systems. These systems are characterized by the use of multiple antennas on the transmitter and receiver side to improve performance. A common problem is to allocate the power available to the transmitter, so that the overall throughput is maximized. For a detailed description see [10].

Allocating power to a transmitter increases its throughput, as it increases its signal-to-noise ratio. However, it depends on the specific transmitter and its damping how useful an allocation is. According to [11] when allocating the power x_i to the channel i of the n communication channels, the mutual information transmitted by the MIMO system can be calculated as followed:[5]

$$I = \sum_{i=1}^{n} \log_2(1 + \frac{x_i}{\varrho^2}\lambda_i)$$

Here ϱ^2 is the mean-square error of the noise and x_i is the power assigned. Furthermore λ_i describes the damping and its value is between 0 and 1. The formula can be derived from Shannon.

For simplicity we rewrite the formula for the information throughput.

$$\begin{aligned}
I &= \sum_{i=1}^{n} \log_2\left(1 + \frac{x_i}{\varrho^2}\lambda_i\right) \\
&= \sum_{i=1}^{n} \log_2\left(\frac{1}{\varrho^2 \frac{1}{\lambda_i}}\left(\frac{\varrho^2}{\lambda_i} + x_i\right)\right) \\
&= \sum_{i=1}^{n} \left(\log_2\left(\frac{\varrho^2}{\lambda_i} + x_i\right) - \log_2\left(\varrho^2 \frac{1}{\lambda_i}\right)\right) \\
&= \sum_{i=1}^{n} \left(\log_2\left(\frac{\varrho^2}{\lambda_i} + x_i\right)\right) - \sum_{i=1}^{n} \left(\log_2\left(\varrho^2 \frac{1}{\lambda_i}\right)\right) \\
&= \sum_{i=1}^{n} \left(\log_2(\alpha_i + x_i)\right) - c
\end{aligned}$$

[5]We assume that there is no interference between channels

Now we can derive a simple optimization problem. Given a set of n communication channels, we want to allocate power to these communication channels in order to maximize the total communication rate. We define x_i as the transmitter power allocated to the i-th communication channel. Its resulting communication rate is $log_2(\alpha_i + x_i)$. Note that α_i is always positive. Furthermore we limit the total amount of power by 1, i.e. 100%. As the objective function $\sum_{i=1}^{n} \log_2(\alpha_i + x_i)$ is concave, we can transform this problem to a convex minimization problem by taking the negation of the objective function (see [5], p.254):

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^{n} log_2(\alpha_i + x_i) \\ \text{subject to} & x \succeq 0, \qquad 1^T x = 1 \end{array}$$

As we have to derive the objective function later, we replace the dual logarithm by the natural logarithm.

$$\begin{aligned}
f_0(x) &= -\sum_{i=1}^{n} \log_2(\alpha_i + x_i) \\
&= -\sum_{i=1}^{n} \frac{\ln(\alpha_i + x_i)}{\ln(2)} \\
&= -\frac{1}{\ln(2)} \sum_{i=1}^{n} \ln(\alpha_i + x_i)
\end{aligned}$$

A positive factor in front of the objective function doesn't change its solution and convexity. Therefore we leave it out and consider the following problem.

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^{n} \ln(\alpha_i + x_i) \\ \text{subject to} & x \succeq 0, \qquad 1^T x = 1 \end{array}$$

Apparently this problem satisfies again Slater's condition. Therefore we can again use the KKT conditions to determine the solution. First we have to determine the Lagrangian:[6]

$$L(x, \lambda, \nu) = -\sum_{i=1}^{n} \ln(\alpha_i + x_i) - \lambda^T x + \nu(1^T x - 1)$$

Then we can apply the KKT conditions to find the optimum.

$$x^\star \succeq 0, \qquad 1^T x^\star = 1, \qquad \lambda^\star \succeq 0, \qquad \lambda_i^\star x_i^\star = 0 \quad (i = 1, ..., n)$$

$$\nabla L(x^\star, \lambda^\star, \nu^\star) = \frac{\partial L}{\partial x} = 0$$

For the gradient, we get:

$$\begin{aligned}
&\frac{\partial L(x, \lambda_i, \nu)}{\partial x_i} \\
&= \frac{\partial}{\partial x_i}\left(\sum_{i=1}^{n}(-\ln(\alpha_i + x_i) - \lambda_i x_i) + \nu\left(\sum_{i=1}^{n}(x_i) - 1\right)\right) \\
&\Leftrightarrow \frac{-1}{\alpha_i + x_i^\star} - \lambda_i^\star + \nu^\star = 0 \\
&\Leftrightarrow \lambda_i^\star = \nu^\star - \frac{1}{\alpha_i + x_i^\star}
\end{aligned}$$

[6]Note that the minus before λ arises because the inequality constraint not a less-equal constraint. Furthermore the equality constraint is also not in the standard form. The 1 has to be taken to the other side.

doi: 10.2313/NET-2011-07-2_20

To solve these equations in order to find $x^\star$, $\lambda^\star$ and $\nu^\star$, we can start by eliminating $\lambda^\star$, which acts as slack variable.

$$x \succeq 0, \qquad 1^T x = 1,$$

$$x_i^\star \left(\nu^\star - \frac{1}{\alpha_i + x_i^\star} \right) = 0 \quad (i = 1, ..., n)$$

$$\nu^\star \geq \frac{1}{\alpha_i + x_i^\star} \quad (i = 1, ..., n)$$

Let us assume $\nu^\star < \frac{1}{\alpha_i}$, then $x_i^\star$ must be positive as a result of the last inequality. Consequently, the third condition implies $\nu^\star = \frac{1}{\alpha_i + x_i^\star}$, which leads to $x_i^\star = \frac{1}{\nu^\star} - \alpha_i$.

On the other hand, if $\nu^\star \geq \frac{1}{\alpha_i}$ holds, $x_i^\star$ cannot be positive, as this would violate the complementary slackness condition: $\nu_i^\star \geq \frac{1}{\alpha_i} > \frac{1}{\alpha_i + x_i^\star}$. Thus, $x_i^\star$ must be non-positive and, because of the first condition, this results in $x_i^\star = 0$.

All in all, we get:

$$x_i^\star = \begin{cases} \frac{1}{\nu^\star} - \alpha_i & \nu^\star < \frac{1}{\alpha_i} \\ 0 & \nu^\star \geq \frac{1}{\alpha_i} \end{cases}$$

This can be summarized to:

$$x_i^\star = \max \left\{ 0, \frac{1}{\nu^\star} - \alpha_i \right\}$$

Taking the second condition into account leads to:

$$\sum_{i=1}^{n} \max \left\{ 0, \frac{1}{\nu^\star} - \alpha_i \right\} = 1$$

The lefthand side can be interpreted as a piecewise linear increasing function of $\frac{1}{\nu^\star}$ with breakpoints at α_i. As a result the equation has a unique solution, which can be easily obtained as the function is monoton increasing.

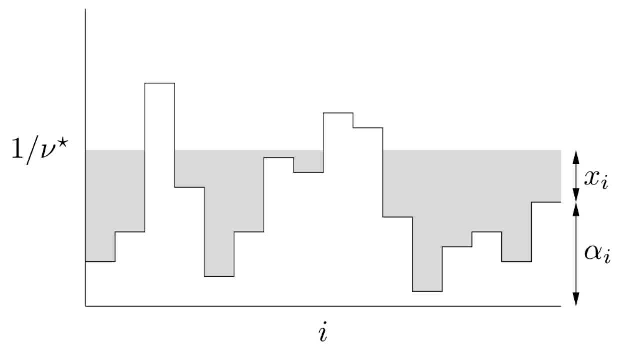

Figure 6: Illustration of the water filling algorithm. The water is shown shadowed and the patches white. (From [5], p.246)

In practice this solution method is known as *water filling* because it has an intuitive interpretation. the default transmitter power α_i of each channel is represented by the height of the patch i in figure 6. We flood the region with water to a depth of $\frac{1}{\nu^\star}$ and calculate the amount of used water: $\sum_{i=1}^{n} \max \left\{0, \frac{1}{\nu^\star}\right\}$. Then we increase or decrease the water level until the amount of used water is equal to 1. As a result the water level above patch i denotes the optimal value $x_i^\star$.

6. CONCLUSION

Convex optimization problems are prevalent in practice. Fortunately, many practical problems meet the requirements for strong duality. Here, the theory of Lagrangian duality offers a powerful tool to determine the exact solution of convex problems. By describing the constraints within the objective function, it enables us to tackle the problem analytically. Especially optimality conditions can be very useful for determining the solution of the primal problem, as demonstrated in the examples.

Furthermore, the theory of strong duality is in particular the basis for efficient and distributed algorithms for convex problems. As the last example shows, the insights gained from duality can be easily transformed to algorithms. Although this example was very problem specific, it is also possible to come up with more general algorithmic approaches for convex problems, for example by considering the ϵ-suboptimality (see [5], p. 241f). In addition to that, strong duality offers the opportunity for perturbation and sensitivity analysis (see [5], p. 249ff).

Besides the theory of strong duality, Lagrangian duality has also a lot of applications. First it can be used to determine lower bounds for non-convex problems. The main advantage of this is, no matter how complex the primal problem is, the dual problem is a concave maximization problem and therefore easy to solve. So duality enables us to determine efficiently a lower bound. Furthermore it can be applied to determine feasibility of a system of equalities or inequalities (see [5], p.258ff).

All in all, duality is a comprehensive theory with a lot of applications and totally worth a deeper look. For more intensive reading, I can recommend the following books: [3], [5], [9] as well as [6] and [7] cover the Lagrangian duality in detail. For a German book about optimization in general, I recommend [8]. In [2] numerous applications of convex optimizations can be found.

7. REFERENCES

[1] Rosenbrock function. Website, April 2011. http://en.wikipedia.org/w/index.php?title=Rosenbrock_function&oldid=426154114 (May 25th 2011).

[2] A. Ben-Tal and A. S. Nemirovskiaei. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[3] D. Bertsekas, A. Nedi, and A. Ozdaglar. Convex Analysis and Optimization. 2003.

[4] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*, volume 3. Springer Verlag, 2000.

[5] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.

[6] J.-B. L. C. Hiriart-Urruty. *Convex Analysis and Minimization Algorithms I.* Springer, 1993.

[7] C. Hiriart-Urruty, Jean-Baptiste Lemarechal. *Convex Analysis and Minimization Algorithms II.* Springer,

1993.
[8] F. Jarre and J. Stoer. *Optimierung*. Springer, 2003.
[9] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.
[10] G. Scutari, D. P. Palomar, and S. Barbarossa. The MIMO Iterative Waterfilling Algorithm. *IEEE Transactions on Signal Processing*, 57(5):1917–1935, 2009.
[11] M. Wennström, M. Helin, A. Rydberg, and T. Öberg. On the Optimality and Performance of Transmit and Receive Space Diversity in MIMO Channels. In *IEE Technical Seminar on MIMO Communication Systems: From Concept to Implementation*, London, 2001.

doi: 10.2313/NET-2011-07-2_20